Омский государственный технический университет

GraphiCon 2024

Материалы
34-й Международной конференции
по компьютерной графике и машинному зрению

(Россия, Омск, 17–19 сентября 2024 года)

Научное текстовое локальное электронное издание

Омск
Издательство ОмГТУ
2024

Сведения об издании: 1, 2, 3

© ОмГТУ, 2024

ISBN 978-5-8149-3873-2

УДК 004.92
ББК 32.973.26-018.2
G73

Редакционная коллегия:

К. Л. Панчук, д-р техн. наук, профессор каф. «Инженерная геометрия и САПР» ОмГТУ;

Н. В. Кайгородцева, канд. пед. наук, зав. каф. «Инженерная геометрия и САПР» ОмГТУ;

Е. В. Любчинов, канд. техн. наук, доц. каф. «Инженерная геометрия и САПР» ОмГТУ (отв. редактор)

GraphiCon 2024 : материалы 34-й Междунар. конф. по компьютерной графике и машинному зрению (Россия, Омск, 17–19 сент. 2024 г.) / Ом. гос. техн. ун-т ; ред-кол.: Е. В. Любчинов (отв. ред.) [и др.]. – Омск : Изд-во ОмГТУ, 2024. – 1 CD-ROM (61,8 Мб). – Минимальные систем. требования: процессор с частотой 800 МГц и выше ; 128 Мб RAM и более ; свободное место на жестком диске 300 Мб и более ; Linux / Windows XP и выше ; MacOS X 10.4 и выше ; CD/DVD-ROM-дисковод ; ПО для просмотра pdf-файлов. – Загл. с титул. экрана. – ISBN 978-5-8149-3873-2.

Представлены материалы 34-й Международной конференции GraphiCon 2024, проходившей на базе Омского государственного технического университета. Соорганизатор конференции – Институт прикладной математики им. М. В. Келдыша РАН.

Конференция GraphiCon ведет свою историю с 1991 года и является крупнейшей в России научно-дискуссионной площадкой в области методов и технологий компьютерного анализа изображений, визуальной и когнитивной аналитики, 3D-реконструкции, визуальной навигации и человеко-машинного взаимодействия, виртуальной и дополненной реальности, распознавания образов и др.

Издание адресовано сотрудникам научно-исследовательских и образовательных организаций, специалистам предприятий ИТ-индустрии, аспирантам, студентам, а также широкому кругу читателей.

© ОмГТУ, 2024

Научное текстовое локальное электронное издание

GraphiCon 2024

Материалы
34-й Международной конференции
по компьютерной графике и машинному зрению

(Россия, Омск, 17–19 сентября 2024 года)

Издание поставляется на одном CD-ROM-диске

Воспроизведение издания автоматическое –
без установки на жесткий диск компьютера

Для корректной работы с изданием на компьютере должны быть установлены
CD/DVD-ROM-дисковод и программное обеспечение
для просмотра pdf-файлов

Ответственность за содержание материалов несут авторы

Издается в авторской редакции

Электронный оригинал-макет издания подготовлен
на кафедре «Инженерная геометрия и САПР» ОмГТУ

Для дизайна обложки использованы материалы
из открытых интернет-источников

Иллюстрации для издания предоставлены авторами

Сводный темплан 2024 г.
Подписано к использованию 07.11.2024.
Объем 61,8 Мб. Тираж 8 эл. опт. дисков.

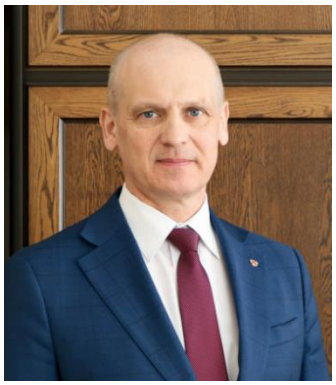
Издательство ОмГТУ. 644050, г. Омск, пр. Мира, 11; т. 8(3812)23-02-12.
Эл. почта: publisher@omgtu.ru, izdomgtu@mail.ru

Организационный комитет конференции ГрафиКон2024
Руководство организационного комитета:

Председатель

Корчагин Павел Александрович

И.о. ректора ОмГТУ, д.т.н., профессор



Со-Председатель

Ложников Павел Сергеевич

Проректор по научной и инновационной деятельности ОмГТУ, д.т.н., доцент

Со-Председатель

Панчук Константин Леонидович

Профессор кафедры «Инженерная геометрия и САПР» ОмГТУ, д.т.н., доцент

Заместитель председателя

Берберова Мария Александровна

Доцент кафедры Промышленной информатики РТУ МИРЭА, заместитель директора АНО «НИЦФТИ», к.т.н.

Члены оргкомитета:

Бабаян Павел Вартанович, к.т.н., доцент, заведующий кафедрой Автоматики и информационных технологий в управлении РГРТУ им. В.Ф. Уткина, Рязань, Россия

Бондарев Александр Евгеньевич, к.ф.-м.н., старший научный сотрудник ИПМ им. М.В. Келдыша РАН, доцент кафедры Высшей математики РТУ МИРЭА, Москва, Россия

Будак Владимир Павлович, д.т.н., профессор, профессор кафедры Светотехники НИУ МЭИ, главный редактор журнала «Светотехника / Light & Engineering», Москва, Россия

Волобой Алексей Геннадьевич, д.ф.-м.н., доцент, ведущий научный сотрудник ИПМ им. М.В. Келдыша РАН, Москва, Россия

Галактионов Владимир Александрович, д.ф.-м.н., профессор, главный научный сотрудник ИПМ им. М.В. Келдыша РАН, Москва, Россия

Дебелов Виктор Алексеевич, д.т.н., доцент, Новосибирск, Россия

Ерёмченко Евгений Николаевич, научный сотрудник МГУ им. М.В. Ломоносова, Москва, Россия

Кайгородцева Наталья Викторовна, к.пед.н., доцент, заведующая кафедрой «Инженерная геометрия и САПР» ОмГТУ, Омск, Россия

Карпов Леонид Евгеньевич, д.т.н., главный научный сотрудник ИСП им. В.П. Иванникова РАН, ответственный секретарь редколлегии журнала «Программирование», доцент кафедры Системного программирования ВМК МГУ им. М.В. Ломоносова, Москва, Россия

Конопацкий Евгений Викторович, д.т.н., доцент, заведующий кафедрой Инженерной графики и информационного моделирования ННГАСУ, Нижний Новгород, Россия

Крылов Андрей Серджевич, д.ф.-м.н., профессор, заведующий лабораторией Математических методов обработки изображений факультета ВМК МГУ им. М.В. Ломоносова, Москва, Россия

Крысова Ирина Викторовна, к.т.н., доцент кафедры «Инженерная геометрия и САПР» ОмГТУ, Омск, Россия

Лагунова Марина Викторовна, д.пед.н., профессор, заместитель заведующего кафедрой Инженерной графики и информационного моделирования ННГАСУ**,** Нижний Новгород, Россия

Подвесовский Александр Георгиевич, к.т.н., доцент, профессор кафедры Информатики и программного обеспечения БГТУ, Брянск, Россия

Ротков Сергей Игоревич, д.т.н., профессор, профессор кафедры Инженерной графики и информационного моделирования ННГАСУ**,** директор АНО «НИЦФТИ», Нижний Новгород, Россия

Толок Алексей Вячеславович, д.т.н., профессор, Главный научный сотрудник ИПУ РАН им. В.А. Трапезникова, заведующий кафедрой инженерной графики МГТУ «СТАНКИН», Москва, Россия

Турлапов Вадим Евгеньевич, д.т.н., доцент, профессор кафедры высокопроизводительных вычислений и системного программирования ННГУ им. Н.И. Лобачевского, Нижний Новгород, Россия

Филинских Александр Дмитриевич, к.т.н., доцент, заведующий кафедрой «Графические информационные системы» НГТУ им. Р.Е. Алексеева, Нижний Новгород, Россия

Фонов Владимир Владимирович, к.т.н., доцент, ООО АСКОН-Интеграция, Омск, Россия

Хранилов Валерий Павлович, д.т.н., профессор, заместитель директора Учебно-научного Института радиоэлектроники и информационных технологий по научной работе НГТУ им. Р.Е. Алексеева, Нижний Новгород, Россия

Программный комитет конференции ГрафиКон2024

Председатель

Галактионов Владимир Александрович

Д.ф.-м.н., профессор, главный научный сотрудник
Института прикладной математики им. М.В. Келдыша РАН

Члены программного комитета:

Алексеев Алексей Кириллович, ОАО «РКК «Энергия», Москва

Бабаян Павел Вартанович, РГРТУ им. В.Ф. Уткина, Рязань (председатель секции)

Барладян Борис Хаимович, ИПМ им. М.В. Келдыша РАН, Москва

Бондарев Александр Евгеньевич, ИПМ им. М.В. Келдыша РАН, РТУ МИРЭА, Москва (председатель секции)

Будак Владимир Павлович, НИУ МЭИ, Москва

Васильев Евгений Павлович, ННГУ им. Н.И. Лобачевского, Нижний Новгород

Ватолин Дмитрий Сергеевич, МГУ им. М.В. Ломоносова, Москва

Волобой Алексей Геннадьевич, ИПМ им. М.В. Келдыша РАН, Москва (председатель секции)

Гетманская Александра Александровна, ННГУ им. Лобачевского, Нижний Новгород

Дебелов Виктор Алексеевич, ИВМиМГ СО РАН, Новосибирск

Ерёмченко Евгений Николаевич, МГУ им. М.В. Ломоносова, Москва (председатель секции)

Ершов Сергей Валентинович, ИПМ им. М.В. Келдыша РАН, Москва

Жданов Андрей Дмитриевич, Университет ИТМО, Санкт-Петербург

Жданов Дмитрий Дмитриевич, Университет ИТМО, Санкт-Петербург

Захарова Алёна Александровна, ИПУ РАН, Москва

Исаев Руслан Александрович, БГТУ, Брянск

Конопацкий Евгений Викторович, ННГАСУ, Нижний Новгород (председатель секции)

Коростелев Дмитрий Александрович, БГТУ, Брянск

Костров Борис Васильевич, РГРТУ им. В.Ф. Уткина, Рязань

Крылов Андрей Серджевич, МГУ им. М.В. Ломоносова, Москва (председатель секции)

Крысова Ирина Викторовна, ОмГТУ, Омск

Курячий Михаил Иванович, ТУСУР, Томск

Лагунова Марина Викторовна, ННГАСУ, Нижний Новгород (председатель секции)

Ложников Павел Сергеевич, ОмГТУ, Омск

Макарова Елена Андреевна, ООО «К-Скай», Петрозаводск

Михайлюк Михаил Васильевич, НИИСИ РАН, Москва

Небаба Степан Геннадьевич, ТПУ, Томск

Подвесовский Александр Георгиевич, БГТУ, Брянск

Поройков Антон Юрьевич, НИУ МЭИ, Москва

Потёмин Игорь Станиславович, Университет ИТМО, Санкт-Петербург

Приоров Андрей Леонидович, ЯрГУ им. П.Г. Демидова, Ярославль

Пчелинцев Яков Антонович, МГУ им. М.В. Ломоносова, Москва

Ротков Сергей Игоревич, ННГАСУ, Нижний Новгород (председатель секции)

Рябинин Константин Валентинович, ПГНИУ, Пермь

Саблина Виктория Александровна, РГРТУ им. В.Ф. Уткина, Рязань

Семенов Виталий Адольфович, ИСП им. В.П. Иванникова РАН, Москва

Скорнякова Надежда Михайловна, НИУ МЭИ, Москва

Сорокин Дмитрий Васильевич, МГУ им. М.В. Ломоносова, Москва

Столбова Ирина Дмитриевна, ПНИПУ, Пермь (председатель секции)

Толок Алексей Вячеславович, ИПУ РАН, МГТУ «СТАНКИН», Москва (председатель секции)

Трубаков Андрей Олегович, БГТУ, Брянск

Турлапов Вадим Евгеньевич, ННГУ им. Н.И. Лобачевского, Нижний Новгород (председатель секции)

Филинских Александр Дмитриевич, НГТУ им. Р.Е. Алексеева, Нижний Новгород (председатель секции)

Фонов Владимир Владимирович, ООО АСКОН-Интеграция, Омск (председатель секции)

Фролов Владимир Александрович, МГУ им. М.В. Ломоносова, Москва

Хвостиков Александр Владимирович, МГУ им. М.В. Ломоносова, Москва

Холопов Иван Сергеевич, ООО «Квантрон Групп», Рязань

Хранилов Валерий Павлович, НГТУ им. Р.Е. Алексеева, Нижний Новгород (председатель секции)

Чуприна Светлана Игоревна, ПГНИУ, Пермь

Шкляр Алексей Викторович, ТПУ, Томск

Якименко Игорь Владимирович, филиал НИУ МЭИ, Смоленск

Организаторы

АНО Научное общество «ГрафиКон»
(организатор)

Омский государственный технический университет
(организатор)

Рекламно-производственная группа «Брендируем.РФ»
(соорганизатор)

Институт прикладной математики им. М.В. Келдыша РАН
(соорганизатор)

АНО «Научно-исследовательский центр физико-технической информатики»
(соорганизатор)

Нижегородский государственный архитектурно-строительный университет
(соорганизатор)

Брянский государственный технический университет
(соорганизатор)

Нижегородский государственный технический университет им. Р.Е. Алексеева
(соорганизатор)

Рязанский государственный радиотехнический университет им. В.Ф. Уткина
(соорганизатор)

МИРЭА - Российский технологический университет
(соорганизатор)

Уважаемые коллеги!

В 2024 году г. Омск впервые принял эстафету организации и проведения конференции ГрафиКон. ГрафиКон – это одна из наиболее известных и авторитетных конференций, проводимых ежегодно в России, начиная с 1991 г. вплоть до настоящего времени. Это также многоплановая научно-дискуссионная площадка, на которой отечественные и зарубежные ученые докладывают и обсуждают результаты своих исследований в областях науки, связанных с анализом изображений, визуализацией, компьютерным зрением и компьютерной графикой, анализом биомедицинских изображений, автоматизированным проектированием, геометрическим и компьютерным моделированием и их приложениями в различных областях науки, практики и образования. Конференция ГрафиКон пользуется большим авторитетом среди ученых и практиков не только в России, но и за ее пределами. Она в течение десятков лет неизменно способствует развитию научных связей участников конференции, объединяет усилия промышленных предприятий, научных организаций и университетов по разработке и внедрению инновационных проектов в различных областях промышленности. Очень важной традицией ГрафиКона является максимальное вовлечение студентов, магистров и аспирантов в работу конференции. Для молодых ученых эта конференция служим стартовой площадкой или очередных этапом на пути апробации и развития своих научных идей.

Ведущей организацией для проведения конференции ГрафиКон-2024 стал Омский государственный технический университет (ОмГТУ). В университете функционирует инновационная структура, способствующая производству новых высокотехнологичных продуктов. Структура включает: 22 научно-образовательных центра, 17 ресурсных центров, 31 научно-исследовательскую лабораторию, 14 студенческих конструкторских бюро, 17 студенческих научно-исследовательских лабораторий, 33 малых инновационных предприятия. Основные направления научно-исследовательской деятельности ОмГТУ включают: прикладную механику; повышение ресурса изделий машиностроения и приборостроения; технологию машиностроения; авиационно-космическую технику; информационные технологии в экономике, науке и образовании; радиотехнические, телекоммуникационные и информационные измерительные системы и системы диагностики; экологию и энергосбережение.

Уважаемые коллеги! От имени организационного и программного комитетов конференции ГрафиКон-2024 желаю всем вам новых творческих успехов, ярких и полезных встреч, бесед и обсуждений во время работы конференции!

Со-председатель организационного комитета
ГрафиКон-2024
доктор технических наук, профессор
кафедры «Инженерная геометрия и САПР»
ОмГТУ
Панчук Константин Леонидович

ОГЛАВЛЕНИЕ

ПЛЕНАРНЫЕ ДОКЛАДЫ

Конференция ГРАФИКОН: путь длиной в полвека 16
Стратегический проект «Суверенные САПР» 19
Использование математических моделей в глубоком обучении 22
Визуализация, построение, и анализ многомерных данных 23
Средства компьютерной графики и вычислительной оптики для решения задач виртуального прототипирования оптических устройств 26

РЕАЛИСТИЧНАЯ КОМПЬЮТЕРНАЯ ГРАФИКА И ВЫЧИСЛИТЕЛЬНАЯ ОПТИКА

Объемно-ориентированная визуализация 3D сцен с глобальным освещением и со множеством динамических точечных источников света в реальном масштабе времени 36
Расширение методов вычисления показателя ослепленности для двумерных изображепний большого динамического диапазона 46
Визуализация дорожных сцен на основе характеристик отражения МКО 55
Отрисовка с антиалиасингом с помощью булевых операций над многоугольниками 67
Применение обратного рендеринга для выбора ракурсов в задаче реконструкции материалов 84
Оптимизация численных методов моделирования переноса излучения в газоразрядной СВЧ ртутно-аргоновой плазме 96
Виртуальная лаборатория анализа музейного освещения 105
Проектирование и виртуальное прототипирование компактной системы дополненной реальности с двумерной дифракционной решеткой 123
Моделирование лесных массивов в системах виртуального окружения с помощью тиражирования облаков точек 136
Анализ производительности методов обхода двухуровневых BVH деревьев в трассировке лучей на графических процессорах 147
Восстановление параметров объёмного рассеяния на основе измерения ДФР 164
Восстановление оптических свойств объектов сцены методом дифференцируемого рендеринга с применением оптимизации выбора наиболее важных точек 179
Сравнительный анализ методов трассировки лучей функций расстояний со знаком 194
Тестирование шейдеров для графической компоненты ОС, используемой на гражданских авиалайнерах 206
Концепция федеративного рендерига 213
Простой метод вычисления цвета краски металлик 225
Использование визуальной оболочки для реконструкции геометрии по набору RGB изображений с помощью дифференцируемого рендеринга 238

НАУЧНАЯ ВИЗУАЛИЗАЦИЯ И ВИЗУАЛЬНАЯ АНАЛИТИКА

Факторы контекстуальной точности в UX-тестах и их комбинаторное влияние на эффективность взаимодействия с интерфейсом 251
Применение оптимизации Гаусса-Ньютона при гильберт-диагностике пламени Бунзена 258

Визуализация поля течения, сопряженных параметров и источников тепла при управлении структурой течения 264
Создание цифровой модели намоточного изделия 271
Роль плотности информационного поля объекта в графических нейросетях 283
Отображение точек многомерного информационного текстового массива на упругую карту для оценки взаиморасположения областей смысловой близости 292
Разработка инструментов визуализации данных на основе предметно-ориентированного моделирования 300
Использование средств компьютерной графики в проектировании освещения метро 315
Разработка и применение высокопроизводительных алгоритмов машинного зрения на ПЛИС 318
Разработка методики визуализации и анализа дефектов в аддитивных технологиях 322
Стохастическая семантика больших данных 327
Нормативное обеспечение и применение методов визуализации данных 347
Сетевая технология программирования визуальных интерфейсов 354
Визуализация влияния вихревых структур на сверхзвуковое обтекание крыла под углом атаки 361
Опыт проведения НИР «Научная визуализация и визуальная аналитика» в РТУ-МИРЭА в 2024 году 368
Инструментальное средство визуализации динамики событий в условиях неполноты картографического материала 373
Применение методов визуального анализа данных для выявления потребности в семантической интеграции данных 389
Сравнительный анализ точности солверов OpenFOAM для задачи обтекания конуса, имеющего сферическое затупление 403
Применение технологий 3D-печати при изготовлении инструментов для микродеформирования листовой заготовки 411
Аддитивная технология изготовления персонализированных каркасов сухих электродов для интерфейсов мозг-компьютер 418
Онтологически управляемый инструментарий интеграции нейроинтерфейсов, дружелюбный к пользователю 428
Ивент-анализ как инструмент социального прогнозирования 438
Автоматизация визуального контроля печатных плат 451

КОМПЬЮТЕРНОЕ ЗРЕНИЕ

Координатная привязка автономного подводного аппарата к объектам подводного добычного комплекса по стереоизображениям 467
Гибридные алгоритмы обнаружения объектов в системах компьютерного зрения реального времени 476
Алгоритмы и программное обеспечение для оценки точности совмещения изображений 484
Применение методов обучения с подкреплением для задачи управления роботом-манипулятором в виртуальной среде 491
Программная технология слежения за подвижным объектом на основе алгоритмов CSRT и SAM 500
Полноростовой сканер для 3D-печати статуэток: статический и динамический подходы 509

Распознавание рукописных текстов отчетов губернаторов Енисейской губернии 19 века 519
Вейвлет-фильтрация зашумленных изображений с использованием их аппроксимаций в субполосах деталей Wavelet-based filtering of noisy images 525
Подход к автономному сопровождению объектов на базе оптических и радио-систем видения 532
Исследование влияния неравносторонних форм апертуры пятикаскадного рекурсивно-сепарабельного цифрового фильтра на отношение сигнал/шум и быстродействие 538
Разработка системы для распознавания людей по изображениям земной поверхности ... 544

СИСТЕМЫ АВТОМАТИЗАЦИИ ПРОЕКТИРОВАНИЯ — РАЗРАБОТКА, ВНЕДРЕНИЕ И ПРИОРИТЕТЫ РАЗВИТИЯ

Технология адаптивного пользовательского интерфейса САПР 552
Реинжиниринг параметрической модели просечно-вытяжного листа с использованием оператора скиннинга поверхности 564
Опыт внедрения и адаптации ИС AISPOL и ГИС в лесоустройстве 573
Оптимизация процесса коммуникации разработчиков 3D-моделей 582
Разработка и внедрение интерактивного электронного технического руководства с интеграцией технологий дополненной реальности для оптимизации производственных процессов 590
Сквозные 3D-технологии в составе комплекса программ в защищенном исполнении «Система полного жизненного цикла «Цифровое предприятие» (СПЖЦ САРУС) 600

ОБРАБОТКА И АНАЛИЗ БИОМЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Прогноз и визуализация трендовых тем в обработке биомедицинских изображений на основе машинного обучения 609
Нейросетевой метод поиска по содержимому для гистологических изображений 618
Исследование эффективности современных архитектур сверточных нейронных сетей в задаче распознавания кожных новообразований 629
Исследование влияния операций предварительной обработки изображений на точность сегментации с помощью сверточных нейронных сетей 636
Цветовые преобразования видеоэндоскопических изображений 641
Методы классификации гистологических изображений на основе сверточных нейронных сетей с фильтром Габора 656
Scale Invariance property of Kolmogorov-Arnold Networks for MRI Gibbs RingingSuppression 669
XAI и слабая аннотация в обеспечении точности и интерпретируемости вложенной классификации опухолей мозга 675

ЦИФРОВАЯ ЗЕМЛЯ И БОЛЬШИЕ ДАННЫЕ

Концептуальная модель экологической цивилизации в цифровой среде Антропоцена 691
Пятна на ясном Солнце или проблемы Цифровой Земли у коллективных субъектов 697
Информационная система биосфероориентрированного эргодизайна среды жизнедеятельности 701
Новый дескриптор на основе частотного анализа для защиты векторных карт 708
Анализ и обработка смоделированных изображений морских навигационных карт 715

Методология визуального моделирования как компонент междисциплинарного исследования геопространства 725
Методология визуального моделирования социотехноприродных процессов 731

ГЕОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ. КОМПЬЮТЕРНАЯ ГРАФИКА

Использование перемещений оси схватоносителя по линейчатой поверхности ограничивающей телесный угол сервиса антропоморфных роботов для преодоления тупиковых ситуаций 735
Кубическая кривая Фергюсона на плоскости 741
Единый конструктивный алгоритм построения окружностей Мальфатти 747
Determining new higher order curves using biquadratic transformation methods 755
Сопряжение набора Безье кривых с анализом точности в САПР «Сарус» 761
Исследование операции «Скругление» в программном обеспечении «Система полного жизненного цикла «САРУС» 766
Оптимизация создания сборочных моделей в СПЖЦ «САРУС» 776
Подготовка кадров для дальнейшего развития геометрического ядра САПР «САРУС» .. 785
Роль системы координат в построении 3D-модели в СПЖЦ «САРУС» 795
Особенности применения булевых операций в программном модуле «Система конструкторского проектирования» ПО «САРУС» 805
Применение методов геометрического моделирования и средств компьютерной графики при формировании геометрического облика вертолета 813
Цифровые инструменты цветового конструирования в компьютерной графике 820
Анализ топологии полиэдров в задаче аппроксимации замкнутых поверхностей полиэдрами с группами конгруэнтных граней 827
Динамическое редактирование геометрических образов с использованием логических инструментов в программной среде Симплекс 837
Имплементация в учебный процесс вуза тестирования нового САПР «САРУС» 843

ГРАФИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ИММЕРСИВНЫЕ ТЕХНОЛОГИИ НА СТАДИЯХ ЖИЗНЕННОГО ЦИКЛА ПРОДУКТА

Интеллектуализация графических и биомеханических интерфейсов управления коллаборативными роботами 855
Применение технологий геометрического моделирования и VR/AR для разработки инструкций по сборке и ремонту производственного оборудования 863
Исследование восприятия графической информации в интерфейсе веб-сайта 870
Исследование влияния композиционного решения на восприятие зрительной информации в кадре 879
Исследование визуального восприятия дискретных цветовых шкал в сложных рельефах 888
Автоматизация в задачах внутрицехового оперативно-календарного планирования производства 895
Исследование методов прикладного программирования Renga API для конечных пользователей информационных моделей 903
Моделирование движения автотранспорта в городской среде со сложным рельефом для определения безопасных режимов движения при различных погодных условиях 911

Разработка 3D-моделей подводного аппарата с помощью технологии реверс-инжиниринг 921
Анализ методов оптимизации трехмерных сцен в webgl 929
Оптимизация текстур в 3D-сценах с учетом расстояния визуализации 934
Создание виртуального стенда естественной циркуляции ФТ-101 для проведения лабораторных работ 941

ПОДГОТОВКА КАДРОВ ВЫСШЕЙ КВАЛИФИКАЦИИ В ОБЛАСТИ ИНЖЕНЕРНОЙ ГЕОМЕТРИИ, КОМПЬЮТЕРНОЙ ГРАФИКИ И ЦИФРОВОЙ ПОДДЕРЖКИ ЖИЗНЕННОГО ЦИКЛА ИЗДЕЛИЙ

Задачи цифровой поддержки проектирования аэрогидродинамики скоростных амфибийных судов 954
Оптимизация геометрической формы башен высотного комплекса при аэродинамическом обтекании 968

ЦИФРОВАЯ ТРАНСФОРМАЦИЯ ОБРАЗОВАНИЯ

Преобразующая сила конвергентных технологий 974
Управление качеством образования в области графики в технических университетах ... 981
Об опыте проведения занятий в рамках проекта «Цифровая кафедра» для студентов инженерного профиля 988
Стратегическое планирование как инструмент методической работы при проектировании курсов графической подготовки в вузах 997
Саморегуляция учебной деятельности: когнитивный подход 1003
Обучение будущих инженеров-строителей в пиринговогом взаимодействии в ходе выполнения расчетно-графических работ 1010
Об организации автоматической проверки инженерно-геометрических заданий на образовательных интернет-платформах 1016
Разработка специализированных модулей САПР бакалаврами Омского Государственного технического университета 1025
Использование протокола WebSocket для технологии “Internet of things” как элемент электронного образования 1030
Средства безопасности и предохранения баз данных в IoT современного вуза 1036
Программный комплекс HiConf для проведения конференций в высших учебных заведениях 1042

ПЛЕНАРНЫЕ ДОКЛАДЫ

Plenary reports

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-16-18

Конференция ГРАФИКОН: путь длиной в полвека
GraphiCon Conference: a Road Half a Century Long

Галактионов Владимир Александрович,
главный научный сотрудник, доктор физ.-мат. наук,
профессор
ИПМ им. М.В. Келдыша РАН

V.A. Galaktionov
The Keldysh Institute of Applied Mathematics RAS, Moscow

Волобой Алексей Геннадьевич,
ведущий научный сотрудник, доктор физ.-мат. наук,
доцент
ИПМ им. М.В. Келдыша РАН

A. G. Voloboy
The Keldysh Institute of Applied Mathematics RAS, Moscow

Аннотация. В этом году в г. Омске на базе Омского государственного технического университета проводится 34-я международная конференция по компьютерной графике и машинному зрению Графикон-2024. Данный доклад – это попытка проанализировать, насколько значимым является такое мероприятие как Графикон, проводимый уже 34-й год подряд, не только для участников конференции, но и для развития современной науки, производства, образования в нашей стране. В докладе приводится история конференции Графикон и анализируется опыт, приобретенный за время ее проведения, начиная с первой в 1991 г. вплоть до наших дней.

***Ключевые слова*:** компьютерная графика, машинное зрение, визуализация, обработка изображений, организация конференции

Аннотация. This year, the 34th international conference on computer graphics and computer vision, Graphicon-2024, is being held at the Omsk State Technical University. This report is an attempt to analyze how significant an event such as Graphicon, held for the 34th year in a row, is for the conference participants and also for the development of modern science, production and education in our country. The report provides the history of the Graphicon conference and analyzes the experience gained during its holding from the first one in 1991 to the present day.

***Ключевые слова*:** computer graphics, computer vision, visualization, image processing, conference organization

I. Предыстория: конференции по машинной графике в СССР

В 60-80-х годах прошлого века в Советском Союзе машинная (компьютерная) графика развивалась в научных центрах в Новосибирске, Москве, Владивостоке, Нижнем Новгороде и других городах. К концу 70-х был накоплен уже достаточный опыт развития методов, алгоритмов и систем машинной графики для проведения своих Всесоюзных конференций по этой проблематике, и они проводились каждые два года. Однако на границе 80-х и 90-х годов в стране произошли драматические изменения. Спрос на исследования и разработки на внутрироссийском рынке упал практически до нуля, и вместе с тем исчезли традиционные (советские) возможности финансирования. Зато открылись возможности международного сотрудничества. Это привело к кардинальному изменению тематики и условий работы, а также требований к НИОКР.

II. Первая конференция Графикон-91

Именно в таких условиях в 1990-м году Академия наук СССР совместно с американской ассоциацией ACM SIGGRAPH при активном участии ИПМ АН СССР, Союза архитекторов СССР и ряда других организаций начали подготовку международной конференции по компьютерной графике ГрафиКон. Конференция состоялась в 1991 г. в Москве под девизом «Компьютерная графика в науке и искусстве» и прошла с большим успехом. В ней принял участие ряд крупнейших ученых и специалистов в этой области из США и других стран, а общее число участников доходило до 3000. Был подготовлен и издан сборник докладов ведущих американских специалистов в различных областях компьютерной графики, представленных на Графикон-91. Они были переведены на русский язык и выдавались участникам конференции при регистрации.

С тех пор конференция стала ежегодной и позволила многим российским специалистам установить контакты и наладить сотрудничество как внутри страны, так и со специалистами других стран. Как показал опыт проведения последующих Графиконов, эти конференции оказали существенное влияние на развитие компьютерной графики и геометрии, машинного зрения и смежных с ними областей в стране. Она способствовала росту и развитию квалифицированного кадрового состава, что привело к получению результатов, конкурентоспособных как в нашей стране, так и на мировом научном рынке. Для некоторых организаторов и участников конференции, включая авторов доклада, первый Графикон существенно повлиял на последующую профессиональную деятельность.

.III. Конференции Графикон в последующие годы

1. *Эволюция тематик программ конференций*

В докладе рассматриваются детали подготовки и проведения конференций Графикон, включая проблемы и сложности, связанные с привлечением ведущих зарубежных специалистов, вопросы рецензирования поступивших докладов, организации школ для молодых ученых, аспирантов и студентов, и пр. Большое внимание уделено эволюции тематики конференций, начиная от 1990-х годов, когда компьютерная графика, как наука об алгоритмах и методах генерации изображений, была базовой тематикой конференции, вплоть до наших дней, когда она перестала быть доминирующей. Развитие технологий, рост производительности процессоров, появление в начале 2000-х годов GPGPU и, тем более, бурное развитие глубокого обучения существенным образом повлияли на изменение программ конференций Графикон. Появились тематики, связанные с вычислительной оптикой, геоинформатикой и цифровой Землей, биометрией и биомедициной, машинным обучением, компьютерной графики в образовании, САПР и ТПП, и др.

В 1990-е годы разделы программы еще, бывало, формировались по поданным статьям, поэтому в какой-то год некоторая тематика могла вообще быть не представлена на Графиконе. В последние годы конференция полностью перешла на формирование программы по секциям, фактически “workshops” в рамках единой конференции, посвященным определенным научным темам. В этом случае в каждой секции ключевую роль играет ее председатель, он и несет полную ответственность за организацию своей секции. Именно председатель формирует пул рецензентов докладов, поступивших на секцию, организует процесс рецензирования и принятия докладов и отвечает за наполнение секции, т.е. приглашает участников и пр. В результате обеспечивается высокий научный уровень рецензирования, а Программный комитет конференции становится практически независимым от места проведения.

Состав и число секций не фиксируется. Таким образом, сохраняя основное направление Графикона, конференция легко и быстро адаптируется к изменениям в тематике. Практически все локальные организаторы добавляют свою секцию (секции) к конференции в соответствии с научным направлением их университетов или кафедр. Некоторые из них впоследствии становятся постоянными.

2. *Расширение географии*

Переход к секционной форме организации, уменьшение иностранного участия и возросший интерес к технологиям, связанным с компьютерной графикой, в регионах привели к новой фазе развития конференции – расширению ее географии. Если первые 22 конференции Графикон, вплоть до 2012 г., проводились лишь в крупных городах и научных центрах, таких как Москва, Санкт-Петербург, Нижний Новгород и Новосибирск, то последующие конференции состоялись также на базе университетов Владивостока, Ростова-на-Дону, Перми, Томска, Брянска, Рязани, и вот теперь Омска.

3. *Молодежная политика*

Устойчивой традицией конференций Графикон является максимальное вовлечение молодежи в работу конференций. По этой причине подавляющее большинство Графиконов состоялось в университетах разных городов страны. Другой традицией конференции, которой следовали в течение многих лет, является наличие в ней наряду с серьезной научной программой существенной образовательной составляющей в форме школы по компьютерной графике, машинному зрению и смежных областей для студентов, аспирантов и молодых специалистов. В результате многие из молодых, начинающих, участников Графикона впоследствии становятся квалифицированными специалистами и представляют на конференцию серьезные доклады.

IV. Заключение

Опыт проведения 33-х предыдущих конференций Графикон показал, что конференция пользуется значительным авторитетом среди специалистов по компьютерной графике, машинному зрению, геометрическому моделированию и других областей, как в нашей стране, так и за ее пределами. Конференция является значимой для нашей страны и оказывает существенное влияние на подготовку квалифицированных кадров и на развитие отечественной профильной науки и технологий.

УДК 004.925.8
DOI: 10.25206/978-5-8149-3873-2-2024-19-21

Стратегический проект «Суверенные САПР»
Sovereign CAD Strategic Project

Конопацкий Евгений Викторович,
доктор техн. наук, доцент
ННГАСУ

E. V. Konopatskiy
Nizhny Novgorod State University of Architecture and Civil Engineering, Nizhny Novgorod, Russia

Ротков Сергей Игоревич,
доктор техн. наук, профессор
ННГАСУ

S. I. Rotkov
Nizhny Novgorod State University of Architecture and Civil Engineering, Nizhny Novgorod, Russia

Лагунова Марина Викторовна,
доктор пед. наук, профессор
ННГАСУ

M. V. Lagunova
Nizhny Novgorod State University of Architecture and Civil Engineering, Nizhny Novgorod, Russia

Аннотация. Доклад посвящён стратегическому проекту развития ННГАСУ «Суверенные САПР» в рамках программы «Приоритет – 2030». Целью проекта является реализация передовых исследований и инноваций, обеспечивающих трансформацию цифровых решений строительной отрасли, её суверенитет и преимущества на международном уровне. Проект включает комплекс фундаментальных и прикладных исследований, направленных на разработку, развитие и внедрение математического аппарата, обеспечивающего создание высокопроизводительного отечественного комплексного программного обеспечения в области ТИМ и САПР нового поколения. Концептуальную основу математического аппарата составляет понятие геометрического тела в виде выделенной части заполненного пространства.

***Ключевые слова*:** САПР, твердотельное моделирование, граничная модель, точечное исчисление, параллельные вычисления, геометрическое ядро.

Аннотация. The report is devoted to the strategic development project of NNSAGU “Sovereign CAD” within the framework of the program “Priority – 2030”. The aim of the project is to realize advanced research and innovations that ensure the transformation of digital solutions of the construction industry, its sovereignty and advantages at the international level. The project includes a set of fundamental and applied research aimed at the design, development and implementation of mathematical apparatus that ensures the creation of high-performance domestic integrated software in the field of TIM and CAD of new generation. The conceptual basis of the mathematical apparatus is the concept of a geometric solid in the form of a selected part of the filled space.

***Ключевые слова*:** CAD, solid-state modeling, boundary model, point calculus, parallel computing, geometric core.

I. Введение

Существующие отечественные и зарубежные САПР хорошо подходят для решения инженерных задач с ограниченным количеством объектов и пока не могут обеспечить достаточную производительность для создания полноценных цифровых двойников, включающих большие и гипербольшие массивы трёхмерных элементов. Например, создания цифровых двойников на уровне микрорайона, района, города или региона. Другим недостатком современных САПР является отсутствие единого комплексного программного обеспечения, объединяющего возможности геометрического моделирования (CAD) и расчётных САПР (CAE), что приводит к необходимости импорта геометрических моделей объектов в расчётные комплексы и обеспечивает наличие следующей проблемы, связанной с недостаточной интероперабельностью отечественных и зарубежных САПР. Уход с отечественного рынка зарубежных вендоров САПР с одной стороны создаёт значительные хотя и временные проблемы для отечественного промышленного сектора, с другой стороны – это уникальная возможность для российских ИТ-компаний, усилиями которых можно создать САПР нового поколения, основанный на лучших научных разработках отечественных и зарубежных учёных.

Основой всех современных САПР является геометрическое ядро и математический аппарат, на основе которого оно реализовано. Перечисленные выше этапы эволюции во многом достигли своего идеологического предела, направленного на использование вычислительных мощностей современных компьютерных систем. И для обеспечения следующего эволюционного шага необходимо вернуться к истокам, т.е. к разработке и развитию нового математического аппарата, использование которого может обеспечить прорывные решения для создания высокопроизводительных САПР нового поколения.

II. Цели и задачи стратегического проекта «Суверенные САПР»

Целью стратегического проекта ННГАСУ в рамках программы «Приоритет – 2030», который получил название «Суверенные САПР» является реализация передовых исследований и инноваций, обеспечивающих трансформацию цифровых решений строительной отрасли, её суверенитет и преимущества на международном уровне. Следует отметить, что в данном случае была использована отраслевая привязка университета, но вместе с тем предлагаемый математический аппарат является инвариантным по отношению к отраслям промышленности и может быть реализован в любых САПР.

Задачи стратегического проекта:

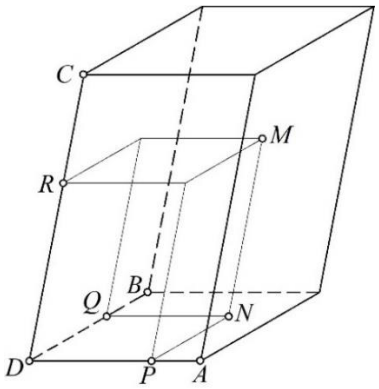
1. Фундаментальные и прикладные исследования для разработки, развития и внедрения математического аппарата, обеспечивающего создание единого высокопроизводительного отечественного комплексного программного обеспечения в области ТИМ и САПР нового поколения.
2. Фундаментальные и прикладные исследования для разработки и внедрения математического аппарата, обеспечивающего создание полноценных геометрических моделей изотропных и анизотропных тел с функционально-управляемым распределением плотности точек.
3. Разработка новой технологии генерации полноценных объёмных изображений в трёхмерном пространстве.

III. Концептуальный подход к твердотельному моделированию

Под геометрическим телом понимается геометрическое множество точек, у которого количество текущих параметров соответствует размерности пространства. Тогда одномерным геометрическим телом будет отрезок прямой – выделенная часть одномерного пространства, двумерным телом – выделенная часть плоскости и т.д. А нульмерным геометрическим телом является точка – уникальный геометрический объект, который не имеет никаких метрических характеристик и вместе с тем любой геометрический объект, какой бы сложный он ни был можно представить организованным множеством точек. Как наш организм состоит из атомов, так и любой геометрический объект, любого пространства, можно описать организованным множеством точек. Необходим только идеологически подходящий математический аппарат для его аналитического описания. И такой математический аппарат, получивший название «Точечное исчисление», был разработан профессором Балюбой И.Г. и его учениками. Балюба И.Г. использовал его для моделирования кривых, поверхностей и гиперповерхностей многомерного пространства. И уже в трудах его учеников точечное исчисление было эффективно использовано для моделирования геометрических тел, как выделенной части пространства, заполненного организованным множеством точек.

IV. Пример определения твердотельной модели параллелепипеда

Геометрическая схема и точечное уравнение определения твердотельной модели параллелепипеда.



$$M = Au + Bv + Cw + D(1-u-v-w).$$

$$\Updownarrow$$

$$\begin{cases} x_M = (x_A - x_D)u + (x_B - x_D)v + (x_C - x_D)w + x_D \\ y_M = (y_A - y_D)u + (y_B - y_D)v + (y_C - y_D)w + y_D, \\ z_M = (z_A - z_D)u + (z_B - z_D)v + (z_C - z_D)w + z_D \end{cases}$$

где $u = \frac{DP}{DA}$, $v = \frac{DQ}{DB}$, $w = \frac{DR}{DC}$.

V. Выводы и заключение

Основные преимущества предложенного подхода и перспективы его дальнейшего использования в САПР:

1. Разработка новых методов хранения геометро-графической информации, основанных на использовании точечных уравнений, инвариантных относительно параллельного проецирования.
2. Отсутствие необходимости согласования геометрической информации в процессе взаимодействия между CAD и FEA системами.
3. Реализация параллельных вычислений по данным на уровне математического аппарата «Точечное исчисление».
4. Разработка новых методов расчёта напряжённо-деформированного состояния твёрдых тел, основанные на функционально-управляемой анизотропии и альтернативные по отношению к методу конечных элементов.
5. Замена воксельных 3D моделей векторными.
6. Разработка новых высокопроизводительных методов рендеринга изображений.
7. Разработка новой технологии генерации полноценных объёмных изображений в трёхмерном пространстве.

УДК 510
DOI: 10.25206/978-5-8149-3873-2-2024-22-22

Использование математических моделей в глубоком обучении

Крылов Андрей Серджевич,
доктор физ.-мат. наук, профессор
Лаборатория математических методов обработки изображений, факультет вычислительной математики и кибернетики, МГУ имени М. В. Ломоносова, Москва, Россия

Andrey S. Krylov
Laboratory of Mathematical Methods of Image Processing, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia

Часто, при решении практических задач, с которыми уже не один год боролись ученые-прикладники, возникает следующая ситуация: берется набор данных (например, снимков с медицинского прибора, с имеющимися результатами диагностики врачами), на этих данных обучается какая-нибудь нейронная сеть (трансформер, …). И на данных, слабо отличающихся от тренировочных, эта сеть очень успешно решает задачу. Намного лучше, чем это получалось у “классических” алгоритмов без машинного обучения. Но, в реальности, устойчивости у нейросетевого решения нет. Для аналогичных данных сеть дает неправильные ответы (“шаг-влево, шаг-вправо” – и все становится очень плохо). Более того любую сеть можно заставить ошибаться очень маленькими изменениями в данных. Этого нет у “классических” алгоритмов – они более устойчивы (на самом деле из-за мало-параметричности используемых моделей). В связи с этим, очень перспективна идея объединить нейронные сети, позволяющие делать тонкую настройку на специфику конкретных данных, с алгоритмами, созданными ранее и основанными на математических моделях, например, удовлетворяющих законам сохранения. Одному из направлений по созданию таких гибридных алгоритмов и посвящен данный доклад.

Последние несколько лет характеризуются очень быстрым развитием нейросетевых методов, использующих, в том числе, математические модели, построенные для описания анализируемого процесса или объекта. Примером являются нейронные сети, учитывающие физику процесса PINN (Physics-Informed Neural Networks), которые позволяют строить нейросетевое решения для нелинейных уравнений в частных производных. При этом, при обучении, в функции потерь учитывается требование удовлетворения дифференциального уравнения на сетке и его начальных и краевых условий. В докладе будет проведен анализ такого подхода, его достоинств и недостатков.

Более перспективным по мнению многих авторов является новый подход, основанный на использовании так называемых нейронных операторов. Наиболее часто используемыми на практике являются нейронный оператор Фурье FNO и нейронный оператор DeepONet. Нейронные операторы основываются на использовании теоремы об аппроксимации персептроном произвольного непрерывного нелинейного оператора с любой заданной точностью. Отметим, что обычно нейронные операторы применяются для моделирования операторов, основанных на использовании дифференциальных уравнений. Однако, на практике они могут быть эффективны и для решения широкого круга задач, для которых в предыдущие годы были предложены нелинейные операторы (алгоритмы), решающие задачу, например в компьютерном зрении, анализе изображений и т. д.

В докладе будут приведены практические примеры использования PINN и FNO в задачах анализа результатов диагностики материалов, решения задачи численного дифференцирования зашумленных данных и подавления осцилляций Гиббса на изображениях, возникающих на медицинских МРТ снимках.

В заключении, будет также обсужден метод построения нейронных операторов, возникший в этом году, и основанный на использовании теоремы о представлении Колмогорова-Арнольда.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-23-25

Визуализация, построение, и анализ многомерных данных
Visualization, constructing, and analysis of multidimensional data

Бондарев Александр Евгеньевич,
старший научный сотрудник, кандидат физ.-мат. наук
ИПМ им. М.В. Келдыша РАН

A.E. Bondarev
Keldysh Institute of Applied Mathematics RAS, Moscow

Аннотация. В настоящее время изучение многомерных данных является одним из важнейших направлений анализа данных. Однако многомерные решения могут быть получены в результате параметрического решения задач математического моделирования. Для построения подобных решений используется технология обобщенного вычислительного эксперимента. Задачи математического моделирования рассматриваются в пространстве определяющих параметров, где определяющими параметрами могут выступать характерные числа рассматриваемой конкретной задачи, геометрические характеристики, управляющие параметры численного метода. Для каждого параметра проводится разбиение в определенном диапазоне, и задача решается в каждой точке подобного разбиения. В результате применения подобного подхода получается решение не для одной, отдельно взятой задачи, а для класса задач, заданного диапазонами изменения определяющих параметров. Рассматриваются вопросы применения параллельных технологий для реализации такого подхода. Представлены примеры реализации построения и анализа многомерных решений для различных типов задач.

***Ключевые слова*:** многомерные данные, обобщенный вычислительный эксперимент, пространство определяющих параметров, параллельные технологии

Аннотация. Nowadays, the study of multidimensional data is one of the most important areas of data analysis. However, multidimensional solutions can be obtained as a result of parametric solution of mathematical modeling problems. The technology of generalized computational experiment is used to construct such solutions. Mathematical modeling problems are considered in the space of defining parameters, where the defining parameters can be characteristic numbers of a particular problem under consideration, geometric characteristics, and control parameters of the numerical method. For each parameter a partitioning in a certain range is carried out, and the problem is solved at each point of such partitioning. As a result of application of such an approach, the solution is obtained not for a single, separately taken problem, but for a class of problems specified by the ranges of changes in the determining parameters. The questions of application of parallel technologies for realization of such approach are considered. Examples of implementation of construction and analysis of multidimensional solutions for different types of problems are presented.

***Ключевые слова*:** multidimensional data, generalized computational experiment, defining parameter space, parallel technologies

I. Введение

Задолго до наступления компьютерной эры основным источником информации в задачах газовой динамики был физический эксперимент. Именно эксперимент позволял получать необходимые сведения о течениях и их свойствах, получать наглядное представление о картине течения и устанавливать соотношения между газодинамическими величинами, характерные для этой картины. Главной целью физического эксперимента всегда было не моделирование самого физического явления, а выяснение обстоятельств, при которых оно происходит, т. е. получение зависимости появления явления от определяющих параметров задачи, таких как числа Маха, Рейнольдса, Прандтля и т. д., и геометрических параметров задачи. Фактически установление таких физических законов для ударных волн, отрывных течений, характерных конфигураций обтекаемых тел было главной задачей механики жидкости и газа. Появление вычислительной техники позволило решать задачи математического моделирования течений, что резко сократило необходимость проведения масштабных

физических экспериментов. Появление высокопроизводительной вычислительной техники, позволяющей проводить параллельные вычисления, обеспечило возможность параллельного расчета одной и той же задачи с разными входными данными в многозадачном режиме. Это позволяет в настоящее время построить и провести обобщенный вычислительный эксперимент. Основные результаты изложены в [1-7].

II. Обобщенный вычислительный эксперимент

Обобщенный вычислительный эксперимент является вычислительной технологией, основанной на синтезе решения задач математического моделирования, параллельных технологий и визуального анализа многомерных данных. Построение обобщенного вычислительного эксперимент начинается с построения разбиения каждого из определяющих параметров задачи в пределах определенного диапазона. Таким образом, формируется сеточная декомпозиция для некоторого многомерного параллелепипеда, составленного из определяющих параметров рассматриваемой задачи газовой динамики. Для каждой точки этой сетки задача рассчитывается в пространстве определяющих параметров. Формально это можно записать следующим образом.

Предположим, что существует надежный численный метод решения двумерных и трехмерных нестационарных задач вычислительной газовой динамики. Тогда мы можем получить численное решение $F(x, y, z, t, A_1, \dots, A_N)$для любой точки пространства расчетной области, где x ,y, z — пространственные координаты, t — время, $A_1, \dots, A_N$— определяющие параметры задачи. В качестве определяющих параметров задачи будем иметь в виду характерные числа, описывающие свойства рассматриваемого течения, такие как числа Маха, Рейнольдса, Прандтля , Струхаля и т. д., и характерные геометрические параметры. Каждый из характерных параметров ограничен в определенном диапазоне $A_i^{min} \le A_i \le A_i^{max}$, $i = 1, \dots, N$.

Разделим каждый из параметров A_iна k -1 частей, так что для каждого параметра получим разбиение, состоящее из k точек. Объем N -мерного пространства, образованного набором определяющих параметров, A_iзаполняется набором k^Nточек.

Обозначив точку из данного множества через $(A_1^*, \dots, A_N^*)$, приходим к тому, что для каждой точки совокупности необходимо получить численное решение газодинамической задачи $F(, x, y, z, t, A_1^*, \dots, A_N^*)$.

Легко видеть, что это потребует решения k^N газодинамических задач, что невозможно без использования параллельных вычислений в многозадачном режиме.

III. Результаты применения обобщенного вычислительного эксперимента

Первый пример – задача нестационарного взаимодействия сверхзвукового вязкого потока со струйным препятствием [1-3]. Препятствие возникает из-за спутной недорасширенной струи, истекающей из сопла. Сопло помещено во внешний сверхзвуковой вязкий поток. Расширяющаяся струя распространяется по внешней поверхности сопла, создавая препятствие во внешнем поле течения. Зависящее от времени управляющее воздействие (скорость роста степени нерасчетности в недорасширенной струе) позволяет изменять пространственно-временную структуру поля течения. Новая пространственно-временная структура представляет собой особый режим течения, при котором струя распространяется против потока по внешней стенке сопла. Мы рассматриваем решающую скорость роста степени нерасчетности струи в качестве управляющего параметра. Основной целью исследования является оценка и определение зависимости управляющего параметра от четырех характерных параметров задачи – чисел Маха, Рейнольдса, Прандтля и Струхаля . Эти параметры изменяются в определенных диапазонах, создавая четырехмерное пространство.

В результате применения подхода получен пятимерный массив данных, переменными которого являются четыре характерных параметра M_∞, Re_∞, Pr , Sh_∞ и критическая скорость V^* . Для полученных данных построено наглядное представление данных в главных компонентах. Представление позволяет предположить, что точки объема данных можно грубо аппроксимировать параметрической плоскостью. После определения коэффициентов для плоскости и обратного преобразования к исходным переменным получаем искомую зависимость $V^* = F(M_\infty, Re_\infty, Pr, Sh_\infty)$ в аналитическом виде. Размерность рассматриваемого четырехмерного массива может быть уменьшена до трех, поскольку характеристический параметр Re_∞ оказывает на решение очень малое влияние. Поэтому окончательный вид для V^* можно записать следующим образом: $V^* = -0{,}1\, M_\infty + 0{,}115\, Пр + 0{,}24\, Ш_\infty$.

Второй пример применения общего численного эксперимента посвящен задаче оптимизации. В примере представлен поиск оптимальной формы трехмерного лопастного узла энергетической установки [5]. Данный эксперимент основан на разработанной вычислительной технологии расчета силовых нагрузок на трехмерный узел лопастей энергетической установки в ветровом потоке. Расчет для различных сочетаний основных геометрических параметров узла с использованием параллельных вычислений позволяет найти оптимальную

форму узла по его мощностным характеристикам. Создана виртуальная экспериментальная установка для моделирования обтекания энергетической установки на основе решения уравнений Навье – Стокса. Проведены расчеты по определению оптимальной формы лопастного узла с учетом ограничений на его конструкцию и проведен всесторонний анализ результатов с использованием предложенной процедуры оптимизации. Решение задачи оптимизации основано на параметризации конструкции по трем основным параметрам. Варьировались три параметра – два угла наклона лопастей и ширина лопастей. На дискретном наборе значений этих параметров находятся максимумы двух целевых функций — величины полной аэродинамической силы и величины вращающего момента, — определяющих аэродинамическое качество силовой установки.

Применение построения обобщенного вычислительного эксперимента нашло еще одно важное применение – в задачах верификации при сравнительной оценке точности различных солверов на референтном решении. В качестве референтных решений рассматривались задачи о косом скачке под углом атаки, о волне разрежения, об обтекании конуса под углом атаки, об обтекании затупленных тел. Частично результаты приведены в [6-7]. . Таким образом обобщенный вычислительный эксперимент может быть применен для сравнительной оценки точности численных методов.

IV. Выводы и заключение

Представленная в статье концепция обобщенного вычислительного эксперимента имеет широкий спектр возможных приложений. Прежде всего, для задач вычислительной гидродинамики такой подход позволяет получить решение не только одной, отдельно взятой задачи, но и целого класса задач, определенных в некотором диапазоне комплекса определяющих параметров. Следует отметить, что применение подхода позволяет проводить разведочные расчеты на грубых сетках для класса задач с последующим уточнением для наборов определяющих параметров, представляющих особый интерес.

Список литературы

[1] A.E. Bondarev, “Analysis of Space-Time Flow Structures by Optimization and Visualization Methods” // Transactions on Computational Science XIX, LNCS 7870, 158-168 (2013).

[2] A.E. Bondarev and V.A. Galaktionov, “Parametric Optimizing Analysis of Unsteady Structures and Visualization of Multidimensional Data” // International Journal of Modeling, Simulation and Scientific Computing, V.04, supp01, (2013) DOI 10.1142/S1793962313410043.

[3] A.E. Bondarev and V.A. Galaktionov, “Analysis of Space-Time Structures Appearance for Non-Stationary CFD Problems” // Procedia Computer Science, 51, 1801–1810 (2015).

[4] A.E. Bondarev, Bondarev A.E. On the Construction of the Generalized Numerical Experiment in Fluid Dynamics // Mathematica Montisnigri, Vol. XLII, 2018, p. 52-64. https://www.lppm3.ru/files/journal/XLII/MathMontXLII-Bondarev.pdf

[5} Andreev S.V. et al., “A Computational Technology for Constructing the Optimal Shape of a Power Plant Blade Assembly Taking into Account Structural Constraints”, Programming and Computer Software, 43(6), 345-352 (2017). DOI: 10.1134/S0361768817060020

[6] Alexander E. Bondarev and Artem E. Kuvshinnikov, “Analysis of the Accuracy of OpenFOAM Solvers for the Problem of Supersonic Flow Around a Cone”, Lecture Notes in Computer Science (LNCS), 10862, 221–230, (2018). DOI: 10.1007/978-3-319-93713-7_18

[7] A.E. Bondarev, A.E. Kuvshinnikov. Analysis and Visualization of the Computational Experiments Results on the Comparative Assessment of OpenFOAM Solvers Accuracy for a Rarefaction Wave Problem (2021). Scientific Visualization 13.3: 34 - 46, DOI: 10.26583/sv.13.3.04

УДК 004.946
DOI: 10.25206/978-5-8149-3873-2-2024-26-34

Средства компьютерной графики и вычислительной оптики для решения задач виртуального прототипирования оптических устройств
Computer graphics and computational optics tools for solving problems of virtual prototyping of optical devices

Д.Д. Жданов
Университет ИТМО, Санкт-Петербург, Россия
кандидат физ.-мат. наук, доцент,
ведущий научный сотрудник факультета программной инженерии и компьютерной техники

D.D. Zhdanov
ITMO University, Saint-Peterburg, Russia

Аннотация. Проектирование и создание современных оптических устройств освещения и наблюдения является сложной и трудоемкой задачей. Проектирование оптических систем данного типа основывается, как правило, на методах вычислительной оптики, метриками которой являются аберрации и передаточные функции. Эти метрики не имеют реального визуального представления, демонстрирующего работу этих устройств в условиях реальной эксплуатации. В работе представлено, как методы компьютерной графики позволяют создать визуальные метрики работы современных оптических устройств и приводятся примеры результатов виртуального прототипирования различного рода оптических систем.

***Ключевые слова*:** компьютерная графика, вычислительная оптика, виртуальное прототипирование, стохастическая трассировка лучей, методы Монте-Карло, оптические системы наблюдения, осветительная оптика.

Аннотация. Designing and creating modern optical lighting and surveillance devices is a difficult and time-consuming task. The design of optical systems of this type is based, as a rule, on the methods of computational optics, the metrics of which are aberrations and transfer functions. These metrics do not have a real visual representation demonstrating the operation of these devices in real-world operation. The paper presents how computer graphics methods make it possible to create visual metrics of the work of these devices and provides examples of the results of virtual prototyping of various kinds of optical systems.

***Ключевые слова*:** computer graphics, computational optics, virtual prototyping, stochastic ray tracing, Monte-Carlo methods, optical observation systems, illumination systems optical.

I. Введение

Методы проектирования оптических систем возникли задолго до появления первых компьютеров. Несмотря на то, что теоретические основы, заложенные более ста лет назад, имели приближенное решение (например, теория аберраций третьего порядка), методы анализа и проектирования оптических систем, построенные на этой теоретической базе, нашли применение в современных системах компьютерного моделирования и проектирования оптических систем. Однако, в последние десятилетия был осуществлен прорыв в области технологий создания новой элементной базы для производства оптико-электронных устройств, что в свою очередь привело к созданию принципиально новых видов устройств освещения и наблюдения. Поскольку любое оптическое устройство освещения или наблюдения ориентировано на восприятие светового излучения человеческим глазом или некоторым электронным устройством, возникает вопрос создания новой метрики для оценки качества работы данных устройств, особенно в условиях их реальной эксплуатации. Эту метрику могут предоставить современные решения, предлагаемые компьютерной графикой, основанной на физически корректных законах распространения светового излучения, которые позволяют создать реалистичный образ виртуальной модели реального мира. Синтез методов, алгоритмов и программных решений компьютерной графики и вычислительной оптики позволил создать компьютерную систему виртуального прототипирования оптических систем освещения и наблюдения в условиях их реальной эксплуатации. Система виртуального

прототипирования дает возможность существенно снизить издержки на проектирование устройств данного типа, поскольку большая часть проблем, связанных с эксплуатацией этих устройств, может быть решена на стадии виртуального прототипирования.

II. Постановка задачи

Основными задачами, поставленными в данной работе, являются, во-первых, формирование новых визуальных и численных критериев, позволяющих оценить эксплуатационные характеристики оптических устройств освещения и наблюдения в условиях, близких к условиям их реальной работы, и, во-вторых, алгоритмические и программные решения, позволяющие физически корректно формировать виртуальные прототипы оптических устройств данного типа. Визуальный пример виртуального прототипа работы автомобильного дисплея при дневном и салонном освещении показан на рис. 1.

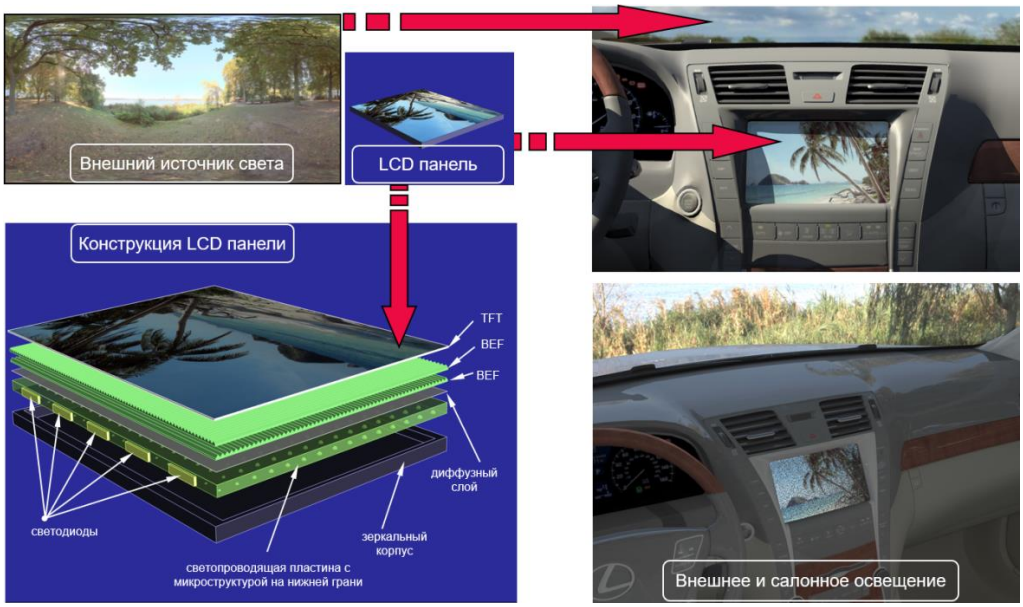


Рис. 1. Виртуального прототип работы автомобильного дисплея при дневном и салонном освещении

III. Теория

Для физически корректного компьютерного моделирования распространения светового излучения в оптически сложной сцене решается уравнение глобального освещения [1]. Данное решение может осуществляться методами прямой трассировки световых лучей, трассировки путей, двунаправленной трассировки лучей с прогрессивными фотонными картами и рядом других методов.

$$\frac{L(\lambda,\vec{x}_o,\vec{\Theta}_x)}{n^2(\lambda,\vec{x}_o)} = \tau(\lambda,\vec{x},\vec{\Theta}_x,\vec{x}_o)\frac{L_e(\lambda,\vec{x},\vec{\Theta}_x)}{n^2(\lambda,\vec{x})} + \int_0^{4\pi}\tau(\lambda,\vec{y},\vec{\Theta}_y,\vec{x})\frac{L(\lambda,\vec{y},\vec{\Theta}_y)}{n^2(\lambda,\vec{y})} f_r(\lambda,\vec{x},\vec{\Theta}_y,\vec{\Theta}_x)\left|\vec{\Theta}_y\cdot\vec{\mathrm{N}}_x\right| d\omega_y \quad (1)$$

Для ускорения расчетов был предложен ряд оригинальных алгоритмов, позволяющих оптимальным образом находить точки сбора яркости вторичного и каустического освещений, оптимизировать процесс расчета яркости каустического освещения, оптимизировать доступ к фотонным картам, использовать многоуровневую схему параллельных и распределенных вычислений, оптимизированную для работы на многопроцессорных рабочих станциях, выполнять конфиденциальные расчеты на федеративной основе без доступа к данным заказчика. Кроме того, для анализа рассеянного света в оптических устройствах была разработана концепция косвенной освещенности, позволяющая визуализировать источники возникновения рассеянного света на элементах данного устройства, а для анализа сложных сцен предложена концепция мультикритерия, позволяющая выполнять одновременный расчет неограниченного числа изображений, сформированных траекториями лучей, удовлетворяющих определенным критериям.

Для интеграции методов компьютерной графики с методами вычислительной оптики был разработан ряд методов, позволяющих выполнять физически корректное и эффективное моделирование распространения светового излучения в оптических системах [2]. Во-первых, метод определения взаимодействия между различными объектами оптической системы как то: оптический контакт или нулевой зазор (при этом нулевой зазор позволяет задать между контактирующими поверхностями ряд дополнительных поверхностей). Во-вторых, метод построения и интерполяции входных и выходных зрачков оптической системы, что на порядки повышает процесс трассировки лучей в этой системе. В-третьих, метод квази-детерминистический двухуровневой

трассировки лучей в оптической системе, повышающий стабильность и скорость вычислений в оптических системах с поверхностями свободной формы, заданными, например, полиномами Якоби высокой степени (2).

$$z(\rho) = \frac{c\rho^2}{1+\sqrt{1-\varepsilon c^2\rho^2}} + D_{con}\left(\frac{\rho}{\rho_{max}}\right) (2)$$

$$\rho^2 = x^2 + y^2, D_{con}(u) = u^4 \sum_{m=0}^{M} a_m Q_m^{con}(u^2), Q_m^{con}(u) = P_m^{(0,4)}(2u-1)$$

IV. Результаты экспериментов

В работе представлены результаты виртуального прототипирования различного рода устройств изображающей и осветительной оптики. На рис. 2 представлен результат виртуального прототипирования широкоугольных оптических систем. На левой части рисунка представлено изображение, формируемое объективом, в котором исправлена дисторсия, но имеется падение освещенности изображения по полю зрения, а на правой части рисунка представлено изображение, формируемое объективом, имеющим постоянную освещенность по полю зрения и неисправленную дисторсию.

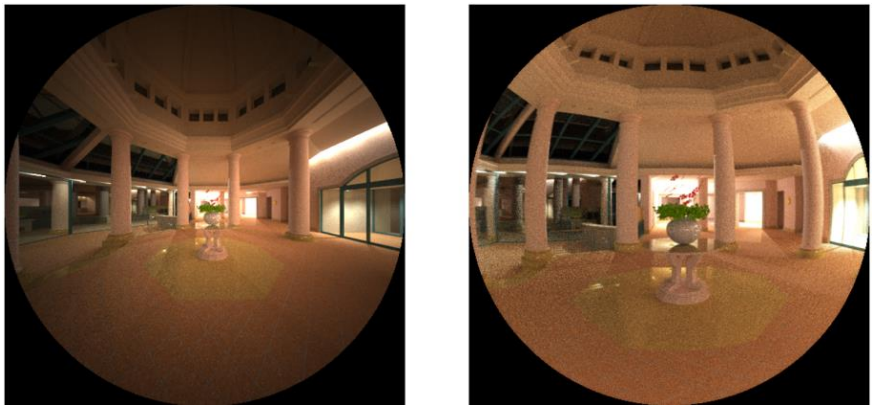


Рис. 2. Изображение, формируемое объективом, в котором исправлена дисторсия (слева); изображение, формируемое объективом, имеющим постоянную освещенность по полю зрения (справа)

На рис. 3 представлен результат виртуального прототипирования бликов в фотографическом объективе от ярких источников в поле зрения (яркий блик от купола церкви создает ряд дополнительных бликов).

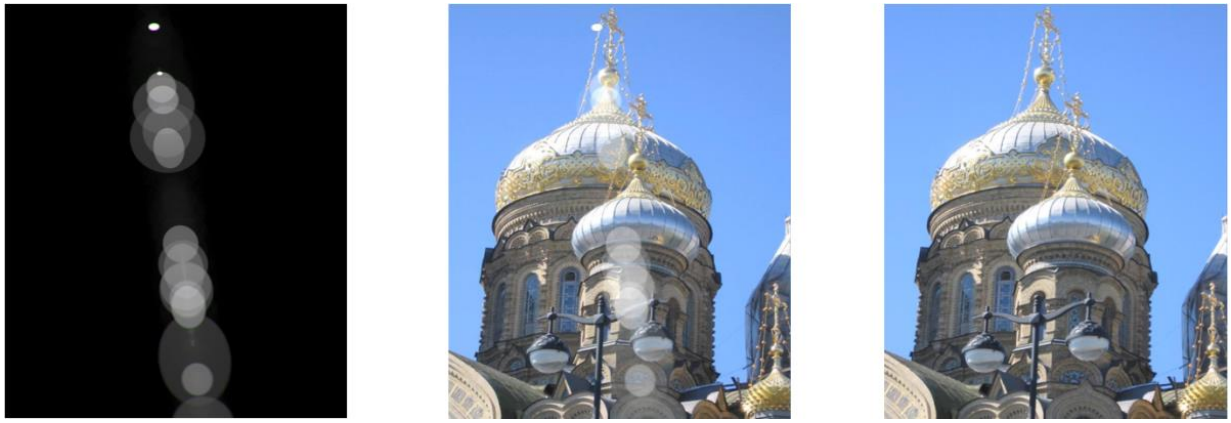


Рис. 3. Блики, формируемые фотографическим объективом. Справа налево: изображение без бликов, изображение с бликами, «чистые» блики без изображений

На рис.4 представлен результат виртуального прототипирования паразитной засветки в телескопическом объективе, вызванной рассеиванием света на торцах линз и элементах конструкции. Правое верхнее изображение есть результат каустического освещения, правое нижнее – вторичного. Оба изображения были получены в рамках одного вычислительного процесса.

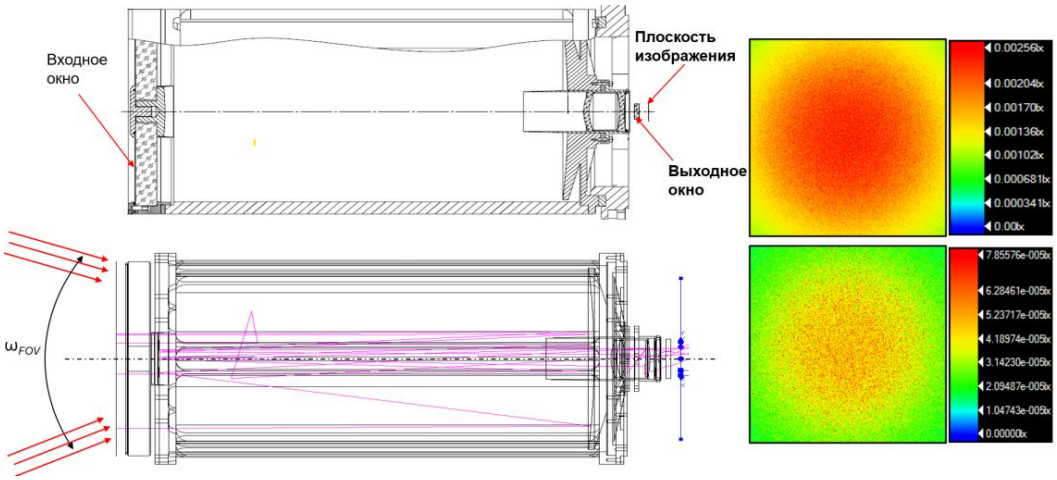


Рис. 4. Паразитная засветка в телескопическом объективе, вызванная рассеиванием света на торцах линз и элементах конструкции

Для анализа рассеянного света была разработана концепция косвенной освещенности, т. е. паразитной освещенности изображения, отнесенной к поверхностям оптического устройства, создающим эту освещенность. Рис. 5 демонстрирует результат расчета косвенной освещенности в телескопическом объективе. Зеленым светом показана косвенная освещенность. Большая яркость косвенной освещенности на элементе оптического устройства соответствует большей паразитной засветке, вызываемой данным элементом оптического устройства.

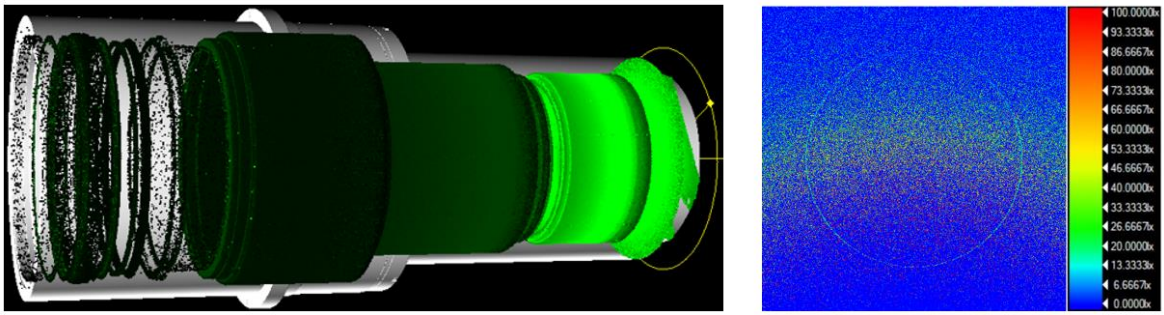


Рис. 5. Косвенная освещенность на элементах конструкции объектива и соответствующая паразитная освещенность изображения

На рис. 6 представлена оптическая схема навигационной системы дополненной реальности вертолета. Основной задачей виртуального прототипирования являлась оценка видимости навигационной информации и минимизация яркого пятна, вызванного эффектом обратной проекции.

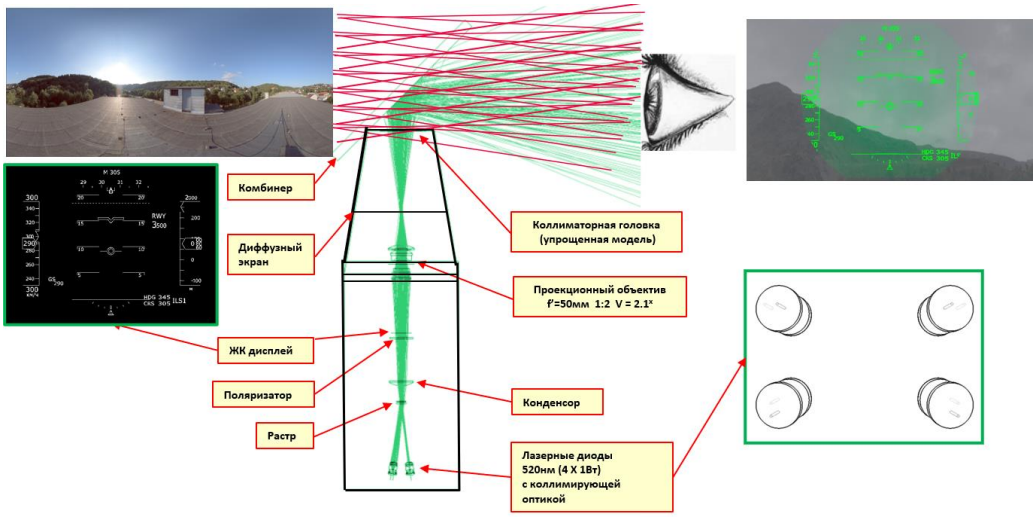


Рис. 6. Оптическая схема навигационной системы дополненной реальности вертолета

На рис. 7 представлен результат виртуального прототипирования данной навигационной системы дополненной реальности при различных уровнях средней яркости внешнего освещения и яркости виртуального изображения.

Рис. 7. Результат виртуального прототипирования навигационной системы дополненной реальности вертолета. Сверху вниз, слева направо средняя яркость наблюдаемого объекта, яркость дисплея: Lav =2227 кд/м^2, Ld =10000 кд/м^2, Lav =2227 кд/м^2, Ld =1600 кд/м^2, Lav =2389 кд/м^2, Ld =10000 кд/м^2, Lav =2389 кд/м^2, Ld =1600 кд/м^2, Lav =2.6 кд/м^2, Ld =200 кд/м^2, Lav =2.6 кд/м^2, Ld =6 кд/м^2

Следующий пример демонстрирует возможность физически корректного виртуального прототипирования очков дополненной реальности, построенных по схеме светопроводящей пластины с одномерной дифракционной решеткой, вводящей изображение с проектора, и двумерной дифракционной решеткой, формирующей изображение у наблюдателя. Особенностью данной схемы является то, что она формирует два изображения: одно изображение в сторону наблюдателя, а другое перевернутое изображение в сторону, противоположную от наблюдателя. Оптическая схема устройства и ход лучей показаны на рис. 8.

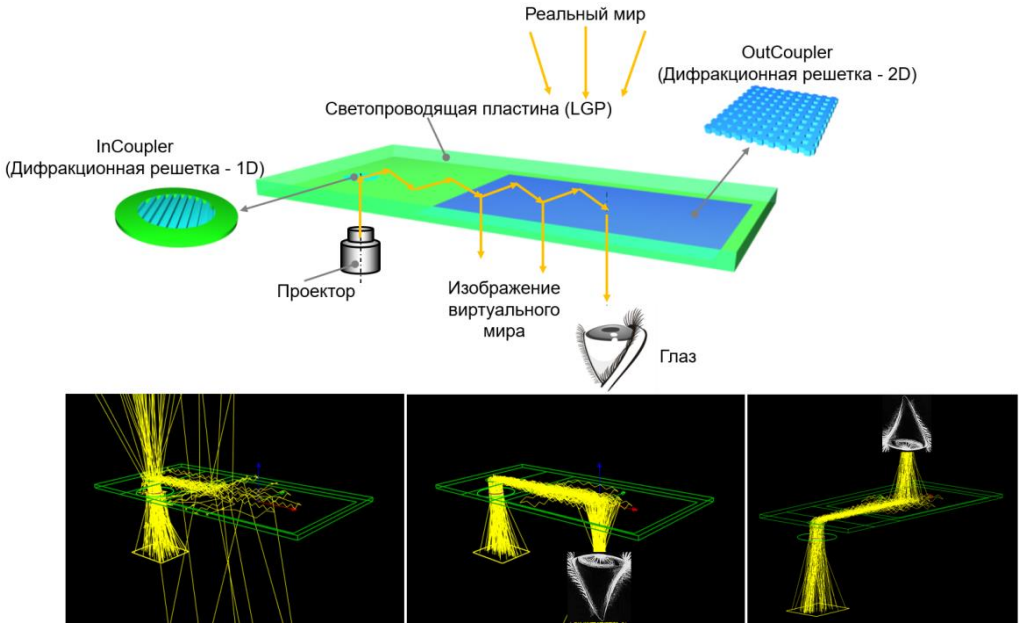


Рис. 8. Оптическая схема очков дополненной реальности с двумерной дифракционной решеткой и ход лучей в этой схеме (слева направо: все лучи, лучи в сторону наблюдателя, лучи в сторону, противоположную от наблюдателя)

Поскольку параметры дифракционных решеток и светопроводящей пластины не были оптимизированы, то изображения в пределах области наблюдения изображения (i-box системы) имеют непостоянную яркость (на рис. 9 девять изображений расположены в следующем порядке: слева направо – от проектора и сверху вниз – поперек основного хода лучей). Два правых изображения на рис. 9 демонстрируют результат визуализации для

центральной области i-box в сторону наблюдателя (верхнее изображение) и в сторону, противоположную от наблюдателя (нижнее изображение).

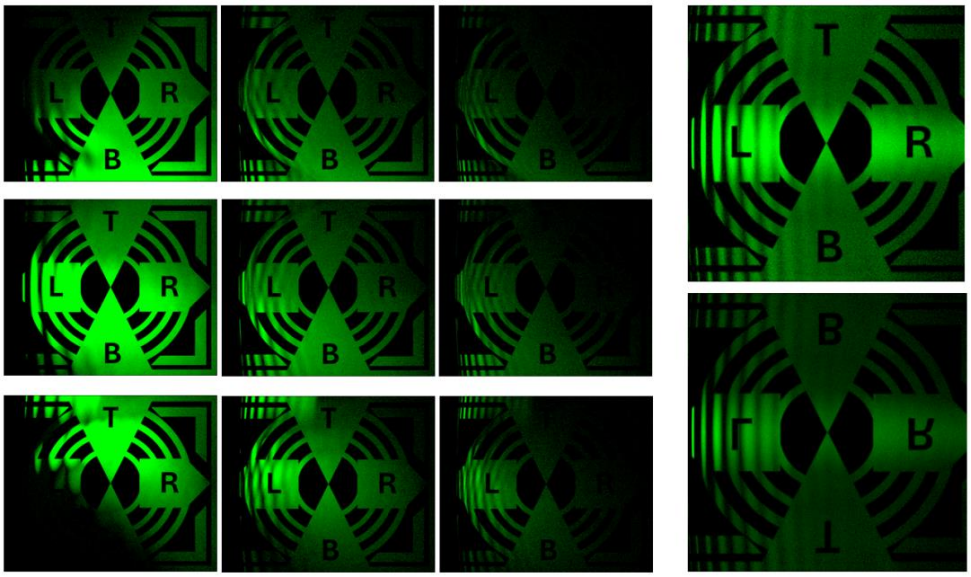


Рис. 9. Изображения, формируемые системой дополненной реальности в девяти зонах i-box (слева) и в центральной зоне с двух сторон наблюдения (справа)

Системы дополненной и смешенной реальностей могут быть также построены по «видеопрозрачной» схеме. В этом случае изображение реального мира транслируется с помощью видеокамер и накладываются на изображение реального мира. Следующий виртуальный прототип оценивает возможность построения данной системы с использованием одной камеры, расположенной в центре устройства, карты глубин, формируемой, например, ToF системой, и системой трекинга зрачка, см. рис. 10.

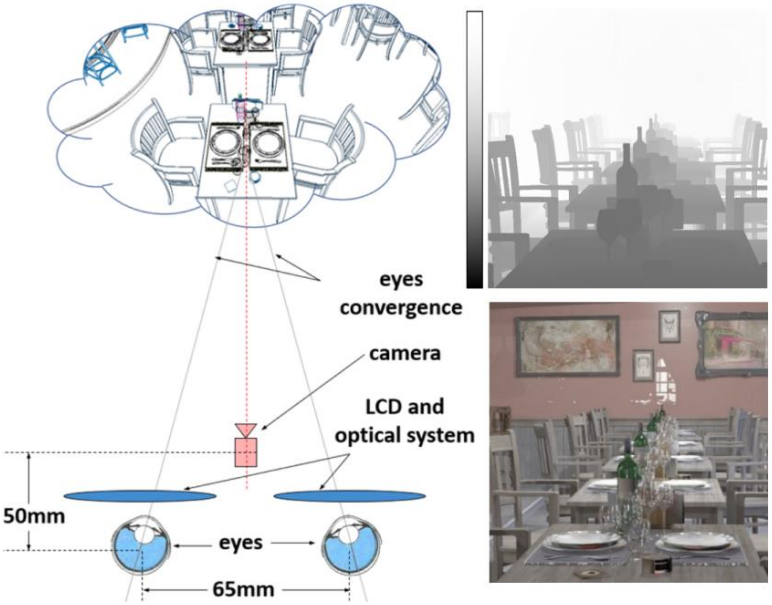


Рис. 10. Оптическая схема видеопрозрачной системы смешанной реальности (слева) и карта глубин и изображение, формируемые камерой, (справа)

Для данной системы был построен виртуальный прототип и проведена оценка качества стереоизображения, восстановленного по одной камере, карте глубин и системе трекинга зрачка. Рис. 11 демонстрирует группу изображений, полученных для различных условий конвергенции глаз наблюдателя, отслеживаемых системой трекинга зрачка.

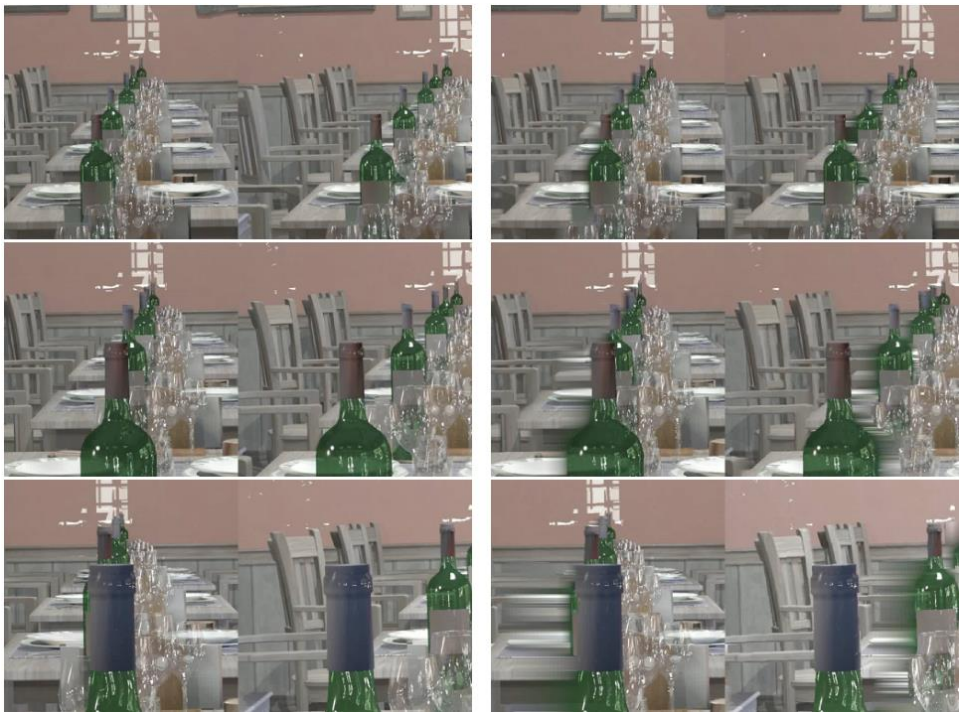


Рис. 11. Ожидаемые изображения для левого и правого глаза (слева) и соответствующие восстановленные изображения по одной камере (справа). Точка конвергенции глаз 2 метра, 1 метр и 0.5 метра (сверху вниз)

Изображение вне плоского физического экрана может быть построено с помощью так называемого Mid-Air дисплея. Было выполнено виртуальное прототипирования дисплея, формирующего изображение в воздухе перед специальным растровым элементом. Оптическая схема данного устройства представлена на рис.12.

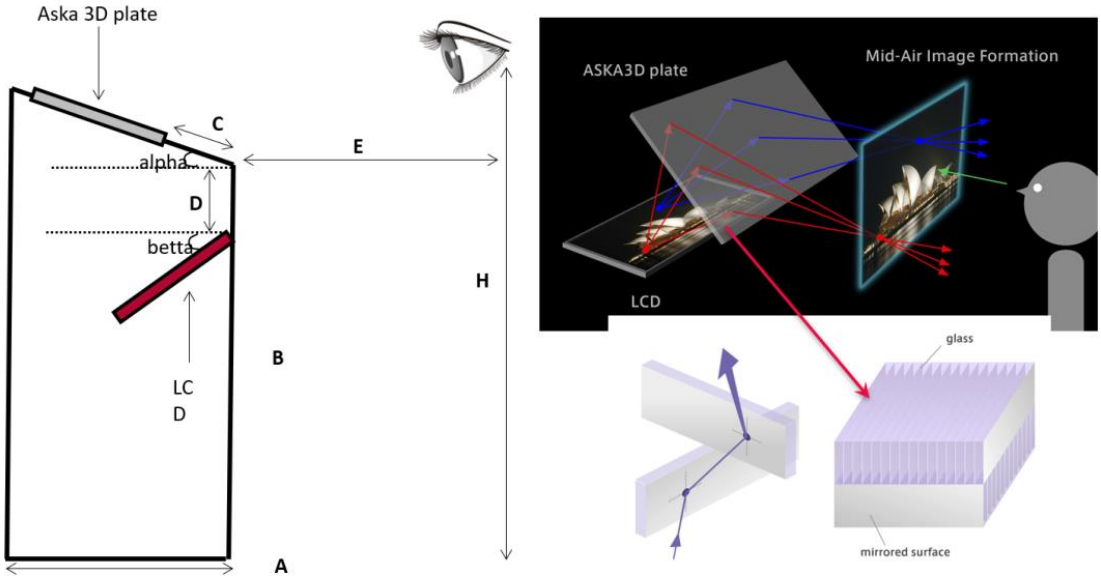


Рис. 12. Оптическая схема Mid-Air дисплея

Задача виртуального прототипирования заключалась в исследовании возможности увеличения размера ячейки растра и уменьшении общих его габаритов (размера D и углов alpha и betta). Результаты виртуального прототипирования представлены на рис. 13.

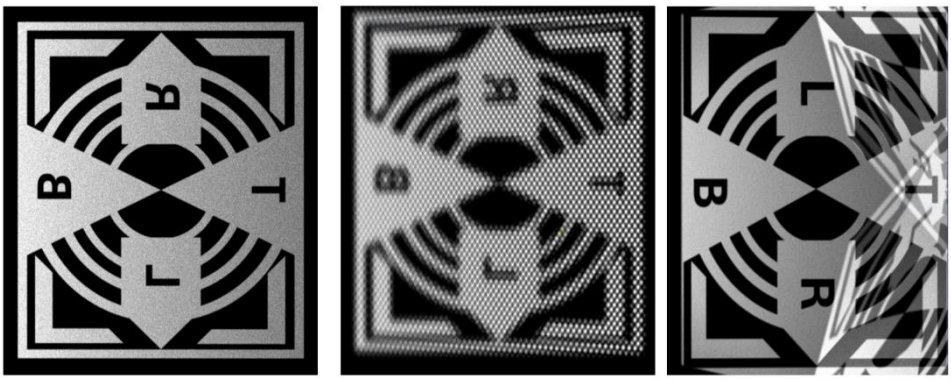


Рис. 13. Результат виртуального прототипирования. Слева направо: исходная схема, размер ячейки растра увеличен в 10 раз на исходной схеме, размер D уменьшен в 5 раз на исходной схеме

Последний пример виртуального прототипирования – это система обеззараживания вдыхаемого воздуха за счет облучения его ультрафиолетом в специальной камере. Схема работы устройства представлена на рис. 14.

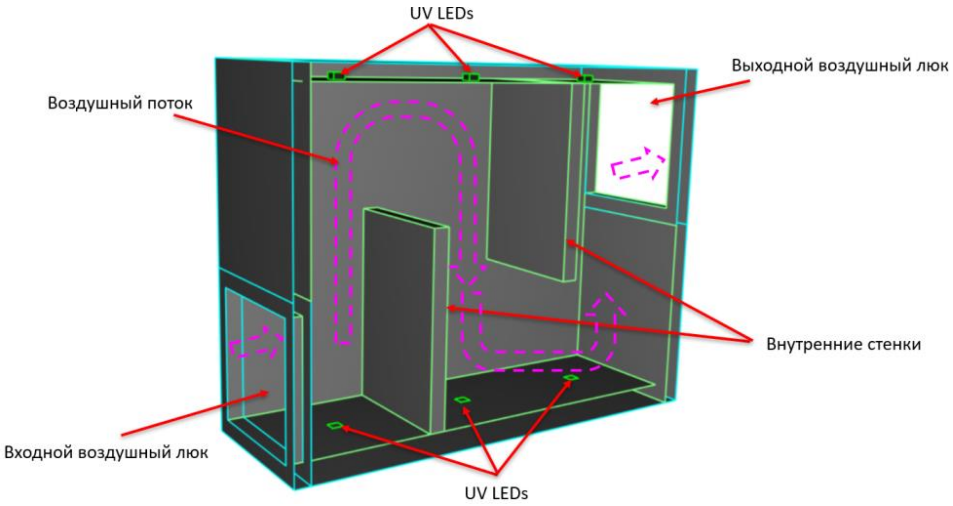


Рис. 14. Принцип работы системы обеззараживания воздуха

Результатом виртуального прототипирования являлось распределение светового потока в объеме камеры. Размер камеры составлял 70x110x10 мм, размер пространственной ячейки, в которой оценивался световой поток 2x2x2 мм. Результаты моделирования для трех сечений камеры (поперек ее 10 мм толщины) представлены в виде тепловых карт на рис. 15.

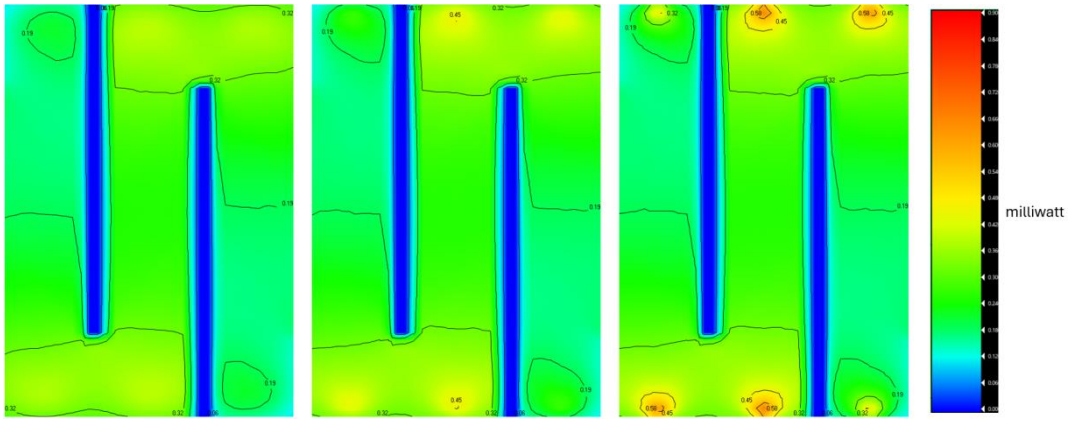


Рис. 15. Тепловые карты распределения светового потока (слева направо) для сечений 9 мм, 5 мм и 1 мм

V. Обсуждение результатов

Виртуальное прототипирование может стать заменой реального макетирования при условии, что, во-первых, его методы и алгоритмы обеспечивают физически корректное распространение светового излучения в

пространстве сцены, включая взаимодействие светового излучения с границами ее объектов. Во-вторых, модель сцены физически корректна и может обеспечить корректный результат взаимодействия светового излучения с ее объектами.

В данном приближении методы и алгоритмы распространения светового излучения ограничены лучевой теорией, поэтому ряд дифракционных и интерференционных явлений остаются за рамками виртуального прототипа. Однако, для большинства практических задач данное ограничение не принципиально, тем более, что ряд явлений интерференции и дифракции может быть локализован на границах объектов и стать частью сложных моделей сцены.

Модель сцены является наиболее сложной частью системы виртуального прототипирования, что связано с ее многомерностью и сложностью получения исходных данных для ее построения. Поэтому система виртуального прототипирования, как правило, ограничена параметрами модели сцены, которые снижают размерность исходных данных, например, поверхностные свойства объекта сцены сводятся к однородной двунаправленной функции рассеивания.

VI. Выводы и заключение

В работе продемонстрирована важность использования компьютерной системы виртуального прототипирования для решения задач анализа и проектирования широкого круга сложных оптических устройств освещения и наблюдения. Приведенные примеры показывают, что компьютерная система виртуального прототипирования может быть внедрена не только в процесс проектирования, но и частично заменить реальное макетирование и испытание оптических устройств.

Источник финансирования. Благодарности

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, грант № 22-11-00145.

Список литературы

1. Фролов В.А., Волобой А.Г., Ершов С.В., Галактионов В.А. Современное состояние методов расчёта глобальной освещенности в задачах реалистичной компьютерной графики. Фролов В.А., and. // Труды Института системного программирования РАН – 2021. vol. 33. no. 2, P. 7-48.

2. Mikhail A. Gan, Dmitriy D. Zhdanov, Vadim V. Novoselskiy, Sergey I. Ustinov, Alexander O. Fedorov, and Igor S. Potyemin DEMOS: state-of-the-art application software for design, evaluation, and modeling of optical systems" // Optical Engineering 31(4), (1 April 1992).

Секция 1

РЕАЛИСТИЧНАЯ КОМПЬЮТЕРНАЯ ГРАФИКА И ВЫЧИСЛИТЕЛЬНАЯ ОПТИКА

Realistic Computer Graphics and Computational Optics

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 004.925.3
DOI: 10.25206/978-5-8149-3873-2-2024-36-45

Объемно-ориентированная визуализация 3D сцен с глобальным освещением и со множеством динамических точечных источников света в реальном масштабе времени
Volume-oriented visualization of 3D scenes with global illumination and with a variety of dynamic point light sources in real time

С. И. Вяткин [1], Б. С. Долговесов [1]
[1]*Институт автоматики и электрометрии СО РАН, Новосибирск, Российская Федерация*

S.I. Vyatkin [1], B.S. Dolgovesov [1]
[1]*Institute of Automation and Electrometry, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russian Federation*

Аннотация. При визуализации сцен с глобальным освещением в режиме реального времени с большим количеством источников света необходимо создание большого количества карт теней для проверки видимости при виртуальном точечном освещении. Поэтому есть два пути: либо ограничить их количество, либо применить небольшие, частичные карты теней с допущением визуальной погрешности. В результате это может снизить точность непрямого освещения или исключить рендеринг полностью динамичных сцен. С другой стороны, методы трассировки лучей и излучательности могут генерировать физически точные изображения, однако их скорость рендеринга слишком мала для приложений реального времени. Мы предлагаем объемно-ориентированный подход к визуализации сцен с глобальным освещением в реальном масштабе времени. Подход сочетает в себе преимущества метода с множеством источников света и вокселизации. Мы используем эффективное разреженное воксельное восьмеричное дерево для проверки видимости вместо дорогостоящих теневых карт в методах с большим количеством источников света. В результате предложенный метод позволяет визуализировать сцены с большим количеством источников света в реальном времени, с поддержкой тысячи динамически генерируемых источников света.

Ключевые слова: объемно-ориентированная визуализация, вокселизация, воксел, разреженное воксельное восьмеричное дерево, виртуальный точечный источник света, глобальное освещение

Abstract. When visualizing scenes with global illumination in real time with a large number of light sources, it is necessary to create a large number of shadow maps to check visibility in virtual point lighting. Therefore, there are two ways: either limit their number, or apply small, partial shadow maps with the assumption of visual error. As a result, this can reduce the accuracy of indirect lighting or eliminate rendering of fully dynamic scenes. On the other hand, ray tracing and emissivity techniques can generate physically accurate images, but their rendering speed is too slow for real-time applications. We offer a volume-oriented approach to visualizing scenes with global illumination in real time. The approach combines the advantages of the method with multiple light sources and voxelization. We use an efficient sparse voxel octal tree to check visibility instead of expensive shadow maps in methods with a large number of light sources. As a result, the proposed method allows you to visualize scenes with a large number of light sources in real time, with support for thousands of dynamically generated light sources.

Keywords: volume-oriented visualization, voxelization, voxel, sparse voxel octal tree, virtual point light, global illumination.

I. Введение

Фотореализм становится актуальным для графических приложений в реальном масштабе времени. Для получения фотореализма важным является моделирование переноса света, результат которого называется глобальным освещением [1]. В связи с этим возникает необходимость разработки эффективных алгоритмов глобального освещения, ориентированных на приложения реального времени. Так, в работе [2] представлен интерактивный алгоритм с одним отражением и непрямым освещением, который учитывает непрямую видимость в полностью динамичной сцене. В [3] представлена реализация прогрессивного фотонного отображения с применением аппаратного обеспечения трассировки лучей. В статье [4] представлена программная платформа, которую можно использовать в процессе изучения методов глобального освещения в

реальном времени. Одним из недостатков алгоритмов с большим количеством источников света является то, что они требуют создания большого количества карт теней для проверки видимости. Это значительно снижает их эффективность [5]. С другой стороны, если упростить карты теней для ускорения вычислений, они становятся неточными для непрямых теней.

Карты окружения используются для представления и вычисления освещенности в виртуальных сценах. Однако их оценка и выборка из них являются дорогостоящими, что ограничивает их применимость для рендеринга в реальном времени. Были рассмотрены следующие сопутствующие работы. В работе [6] предлагается дифференцируемый метод глубокого обучения, способный сжимать и обучаться выборке из единой карты среды. Для решения проблемы рендеринга в реальном времени глобального освещения используются масштабируемые иерархии сферических гармоник. В статье [7] описан метод, позволяющий вычислять световые датчики и преобразовывать их в коэффициенты сферических гармоник, которые используются в качестве анизотропных изменяющихся во времени источников света.

Виртуальные точечные источники света (VPLs) впервые были представлены в статье [8]. Идея заключалась в создании большого количества точечных источников света, каждый из которых имеет карту теней, для имитации непрямого освещения. Однако проблема заключается в необходимости рендеринга всех карт теней для каждого VPL, в результате снижается производительность. В работе [9] описывается создании небольших, частичных карт теней для каждого VPL кадра, с относительно небольшой визуальной погрешностью. Для этого создается точечное представление геометрии сцены, и для каждой небольшой карты теней выбирается подмножество точек. Затем это подмножество преобразуется в текстуру. Таким образом, создается грубое приближение карт теней, которые затем поэтапно улучшаются. Создание карт теней за один проход на одном и том же ресурсе текстур можно сделать очень быстро. Эти небольшие карты теней затем используются в качестве карт теней для VPL. Однако плохая выборка VPL с одним отражением приводит к появлению артефактов и незатененных поверхностей сцены. Для построения надежных путей переноса света в работе [10] была изменена стратегия выборки VPL на двунаправленную трассировку траектории. Используются глобальные пучки лучей, которые представляют собой группы параллельных лучей, определяющих геометрию сцены. Однако глобальные пучки лучей создают дополнительные трудности, поскольку каждый раз, когда используется VPL, полигоны сцены должны быть растеризованы заново, чтобы создать соответствующий глобальный пучок лучей. Следовательно, хотя эта технология позволяет получать фотореалистичные изображения для сложных сцен, она не работает с высокой частотой кадров. Таким образом, виртуальные точечные источники света (VPL) - это эффективный подход к определению непрямых источников света и оценке результирующего непрямого освещения. Тем не менее, VPL страдает от нежелательных артефактов. В следующих подходах решается эта проблема другими способами.

Виртуальные сферические источники света (VSL) позволяют избежать артефактов, если рассматривать сферы вместо точек [11]. Однако оценка интеграла освещенности с помощью метода Монте-Карло приводит к появлению шума на конечном изображении. В работе [12] предлагается HVL как расширение VSL в рамках сферических гармоник (SH), с определением замкнутой формы оценки интегрального освещения. Предлагается эффективное SH-прогнозирование вклада сферических источников света. В статье [13] представлен метод, называемый Virtual Planar Area Lights, для создания рассеянного непрямого освещения с одним отражением для динамических сцен и источников света. В отличие от методов, основанных на VPL, которые извлекают точечные источники света из отражающих карт теней, метод соответствует отражающим картам теней многими плоскими областями, которые непосредственно используются в качестве виртуальных источников света для создания непрямого освещения. В [14] описан метод рендеринга с динамическим освещением, камерами, материалами и объектами. Рендеринг включает в себя различные важные глобальные эффекты освещения, такие как рассеянное освещение, динамическое затенение и непрямое освещение. В статье [15] предлагается метод точечных стохастических lightcuts для рендеринга высококачественных эффектов освещения с множеством источников света. В обзорной статье [16] рассматриваются основные принципы построения иерархий ограничивающих объемов для ускорения поиска, а также методы построения и обхода объема. Быстрые алгоритмы глобального освещения основаны на программной платформе Virtual Point Lights (VPLs). Она разделяет освещение на два этапа: сначала распределяет освещение по сцене и сохраняет его в виртуальных источниках света, затем собирает освещение из этих виртуальных источников света. В работе [17] используют адаптивное завершение матрицы для аппроксимации информации о видимости после начального шага кластеризации. Вычисление происходит в два этапа: сначала распределяется излучение по сцене и сохраняется в виде виртуальных источников света, затем собирается освещение из этих виртуальных источников света. Чтобы ускорить выполнение второго этапа, виртуальные источники света и точки приема сгруппированы иерархически с использованием многомерных световых рамок. Узким местом является вычисление видимости между кластерами виртуальных источников света и точек приема. Алгоритмы завершения матрицы полностью восстанавливают матрицу низкого ранга из

неполного набора выбранных элементов. Восстанавливается информация о видимости, с использованием 10-20% выборок для большинства сцен, и она объединяется с информацией о затенении, вычисляемой отдельно, параллельно на графическом процессоре. В статье [18] предлагается упрощенный подход к расчету глобального освещения, основанный на гибридном алгоритме трассировки конусов, с помощью которого можно визуализировать сцены в реальном времени. По сравнению с методом VCT с 3D текстурами (voxel cone trace - voxel-based cone ray tracing), потребление памяти может быть снижено до 80%. В работе [19] предлагается расширение метода рендеринга множества источников света в реальном времени, характеризующееся интеграцией фотометрических данных в процессе распределения виртуальных точечных источников света. Метод основывается на определении источников света и создании виртуальных точечных источников освещения на описании, представленном в стандартном формате IES, созданном Обществом инженеров по освещению (IES). В [20] предлагается подход трассировки лучей на основе VPL, который полностью выполняется на графическом процессоре и обеспечивает интерактивную частоту от 3 до 25 кадров в секунду при рендеринге сцен с глобальным освещением. Подход базируется на трассировке лучей, используется кластеризация как точек затенения, так и VPL.

В нашей работе предложено объединение пространственной вокселизации и технологии виртуальных точечных источников света, реализован данный подход на GPU, что в совокупности привело к скорости визуализации, сопоставимой с реальным временем (20-30 кадров в секунду). Однако вокселизация с использованием обычных сеток потребляет много памяти графического процессора. Чтобы решить эту проблему и связанные с ней артефакты, применен быстрый динамический метод генерации SVO на GPU для высококачественного рендеринга [21]. SVO — это разреженное восьмеричное дерево сцены.

Наш основной вклад заключается в следующем:

– Альтернативный метод глобального освещения с большим количеством источников света в реальном времени, который может поддерживать тысячи динамически генерируемых VPL.

– Тесты видимости VPL на основе разреженного воксельного октодерева (SVO) в интерактивном режиме без использования дорогостоящих вычислительных карт теней.

II. Постановка задачи

Хотя алгоритмы глобального освещения, такие как трассировка лучей и методы излучательности (radiosity), могут генерировать физически точные изображения, их скорость рендеринга слишком мала для приложений реального времени. Была поставлена задача разработки гибридного метода глобального освещения в реальном времени, который сочетает в себе преимущества метода с множеством источников света и вокселизации с использованием эффективного разреженного воксельного октодерева (SVO) для проверки видимости вместо дорогостоящих теневых карт. При этом метод должен обладать высокой масштабируемостью и поддерживать тысячу виртуальных точечных источников света, генерируемых "на лету".

Кроме VPL в сценах должны присутствовать несколько источников света spotlight, которые будут выступать в качестве главных источников освещения. Они будут имитировать, например, солнечный свет, проникающий через открытое окно. Источники света в компьютерной графике обычно ограничиваются упрощенными приближениями к реальным источникам света. Точечный свет (point lights), направленный свет (directional lights) и прожекторы (spotlights) отображаются только с помощью положения в окружающем пространстве, без каких-либо измерений и с помощью простых типов излучения света. Единственный тип виртуального источника света, который имеет реальные размеры, - это точечный свет. Реальные источники света могут иметь сложные характеристики излучения. Светоизлучатели могут иметь разные размеры и излучать свет с разной интенсивностью в разных направлениях излучения. Для реализма освещения важна «маска», присваиваемая точечным источникам света для неравномерного распределения интенсивности излучения света вокруг точечного источника света. Этот процесс приводит к более реалистичному отображению поверхностей, расположенных вблизи источника излучения, на которые в первую очередь влияет составляющая прямого освещения.

Целью представленной работы является создание объемно-ориентированного метода визуализации 3D сцен с глобальным освещением в реальном масштабе времени с динамическими точечными источниками света, количество и яркость которых может меняться "на лету". Перед разработкой предлагаемого метода было взято во внимание то, что для глобального освещения в реальном масштабе времени дискретизация исходной сцены в вокселы имеет ряд преимуществ:

- Отсутствует зависимость от сложности геометрии.
- Существует много эффективных методов преобразования вокселизированных сцен в структуру разреженного воксельного восьмеричного дерева.

- Вычисления пересечений лучей, и проверка видимости воксельных структур данных выполняются очень быстро.
- Имеются высококачественные методы сглаживания воксельных данных.

В данной статье представлен метод реального времени, который сочетает в себе преимущества метода с множеством источников света и вокселизации. Для проверки видимости вместо дорогостоящих теневых карт используется эффективное разреженное воксельное восьмеричное дерево. Метод ориентирован на большое количество источников света (512-1024 VPLs). Карты окружения используются для представления и вычисления освещенности в виртуальных сценах. Однако их оценка и выборка из них являются дорогостоящими, что ограничивает их применимость к рендерингу в реальном времени.

III. Теория

1. Функционально заданные объекты

Базы данных хранятся в компактном виде [22]. Для описания объектов в данной работе используются функции отклонения (второго порядка) от базовой квадрики. Функция задается алгебраическим неравенством второй степени с тремя неизвестными *x*,*y*,*z* в виде $F(x,y,z) \geq 0$. Рассматриваются поверхности как замкнутые подмножества евклидова пространства, определяемые описывающей функцией $F(x,y,z) \geq 0$, где *F* - непрерывная вещественная функция; *x*,*y*,*z* - задаваемая координатными переменными точка в трехмерном евклидовом пространстве (E^3). Поверхности строятся с помощью квадрик и представляются композицией базовой квадрики и возмущений:

$$F'(x,y,z) = F(x,y,z) + \sum_{i=1}^{N} f_i R_i(x,y,z), \quad (1)$$

где *fi*- форм-фактор; $R_i(x,y,z)$ – возмущение:

$$R_i(x,y,z) = \begin{cases} Q_i^3(x,y,z), \text{ если } Q_i(x,y,z) \geq 0 \\ 0, \text{ если } Q_i(x,y,z) < 0 \end{cases}, \quad (2)$$

где $Q_i(x,y,z)$ – возмущающая квадрика.

Для формирования моделей сложных объектов на базе функций возмущения используются теоретико-множественные операции. Перед визуализацией в режиме реального времени функционально заданные объекты триангулируются [23]. Далее алгоритм работает в реальном масштабе времени.

2. Вокселизация 3D сцен

Преобразуем сцену в набор вокселей, содержащих свойства материала (например, цвет) и нормали поверхностей. Для вокселей рассчитывается входящее и исходящее излучение от источников освещения, а результат записывается в 3D текстуру. На рис. 1 показаны основные этапы вычисления для создания массива вокселей. Эта техника работает за один проход GPU.

Вычисляется ограничивающий куб сцены. Устанавливается разрешение вокселизированной сцены. Выставляется ортографическая камера над сценой так, чтобы уместить ограничивающий куб. Далее происходит рендеринг каждого объекта сцены.

Рис. 1. Структура вычислительного процесса вокселизации объекта

В геометрическом шейдере происходит обработка треугольников объекта. С помощью нормали треугольника выбирается ось, в направлении которой площадь проекции этого треугольника максимальна. В зависимости от оси поворачиваем треугольник наибольшей проекцией к растеризатору, что обеспечивает большее число вокселей.

В пиксельном шейдере каждый фрагмент будет являться вокселем. Создаём структуру Voxel с координатами, цветом и нормалью, и добавляем его в массив вокселей.

На этом этапе используем такие уровни детальности (LOD) объектов, чтобы не иметь проблем с отсутствующими вокселями и со слиянием вокселей смежных треугольников.

Для оптимизации расхода памяти применяется алгоритм упаковки данных. В предлагаемой реализации используется разреженное восьмеричное дерево сцены — Sparse Voxel Octree (SVO). Каждый узел такого дерева разбивает подпространство сцены на восемь равных частей. Разреженное дерево при этом не хранит информацию о пространстве, которое не занято. Будем пользоваться этой структурой для поиска вокселей по их координатам в пространстве. Вокселизированная сцена хранится в виде массива вокселей. На GPU этот массив реализован через StructuredBuffer / RWStructuredBuffer.

Опишем поэтапно алгоритм создания восьмеричного дерева. Под адресом узла в SVO далее подразумевается индекс узла, который преобразуется в 2D-координаты текстуры. С помощью текущего значения счётчика узлов вычисляются адреса, в которых будут располагаться дочерние узлы выделенного узла. Счётчик узлов увеличивается на количество дочерних узлов (восемь). Адреса дочерних узлов записываются в поля текущего узла. В поле флага текущего узла записывается, что он выделен. Адрес текущего узла записывается в родительское поле дочерних узлов.

Алгоритм создания восьмеричного дерева высотой N выглядит следующим образом:

1. Выделяется корневой узел. Текущий уровень дерева = 1. Счётчик узлов = 1.
2. Перебираются воксели. Для каждого воксела определяется узел (подпространство) на текущем уровне дерева, которому он принадлежит. Отмечаем такие узлы флагом, что необходимо выделение. Этот этап распараллеливается на вершинном шейдере с помощью Input-Assembler без буферов, используя только количество вокселей и атрибут SV_VertexID для адресации по массиву вокселей.
3. Для каждого узла текущего уровня дерева проверяется флаг. Если необходимо, выделяется узел. Этот этап также распараллеливается на вершинном шейдере.
4. Для каждого узла текущего уровня дерева прописываются в дочерних узлах адреса соседей.
5. Текущий уровень дерева: повторяются шаги 2-5, пока текущий уровень дерева < N.
6. На последнем уровне вместо выделения текущего узла записывается индекс воксела в массиве вокселей. Таким образом, узлы последнего уровня будут содержать до 8 индексов.

Построив такое восьмеричное дерево, легко можно найти воксел сцены по координатам в пространстве, спускаясь вниз по SVO.

После создания восьмеричного дерева вокселизированной сцены используем эту информацию для сохранения освещённости, переотражённой вокселями от источников освещения. Освещённость сохраняется в

виде 3D текстуры формата R8G8B8A8. Это позволяет использовать трилинейную интерполяцию GPU при сэмплировании текстуры, и в результате получить более гладкое итоговое изображение. Такая текстура называется буфером блоков (brick buffer), потому что она состоит из блоков, расположенных в соответствии с SVO. Блок — это другое представление группы вокселей в пространстве. Группы из 2x2x2 вокселей, индексы которых расположены в листьях SVO, отображаются на блоки вокселей 3x3x3 из 3D текстуры.

Таким образом, имеется воксельное представление данных. SVO используется в качестве структуры пространственных данных для быстрых тестов видимости виртуальных источников света. Виртуальные точечные источники света (VPL) - это эффективный подход к определению источников непрямого света и оценке результирующего непрямого освещения. Структура SVO сохраняется в памяти GPU. Позже структура SVO считывается несколькими потоками GPU одновременно во время тестов видимости VPL. Чтобы использовать кэши групп потоков GPU, группы из восьми узлов дерева размещаются вместе.

3. Глобальное освещение

После реализации SVO создаем геометрический буфер (G-buffer), который хранит информацию о местоположении, нормалях и материалах для прямого и непрямого освещения. Затем G-буфер разделяется для выборки с чередованием VPL. Для каждого источника освещения сцены мы визуализируем сцену в режиме освещения и создаем карту отражающих теней, в которую заносим положение, информацию о нормалях, необходимые для выборки значимости VPL (importance sampling）[24].

Прямое освещение рассчитывается с использованием стандартных карт теней, а вклады всех источников освещения сцен накапливаются в буфере прямого освещения.

Косвенное освещение рассчитывается путем накопления вкладов источников света в буфер косвенного освещения. Чтобы ускорить процесс непрямого освещения, каждому затеняемому пикселю присваивается только подмножество VPL с использованием предварительно сгенерированного чередующегося шаблона выборки.

Тест на видимость выполняется путем испускания теневых лучей от пикселя к подмножеству VPL. При этом выполняется переход по лучу SVO для определения - закрыт ли теневой луч. Для согласованности кэша графического процессора, группируем пиксели, которые используют одно и то же подмножество VPL, разбивая G-буфер на фрагменты размером 4x4. По завершении работы буфер непрямого освещения фильтруется и объединяется с буфером прямого освещения для создания окончательного изображения. На рис. 2 показаны основные шаги глобального освещения с большим количеством источников света с использованием разреженного восьмеричного дерева. Перечислим эти шаги:

Шаг-1: Вокселизация сцены и генерация SVO

Шаг-2: Генерация G-буфера. Используются материалы, отображаемые в виде текстур с плавающей запятой, положения в мировой системе координат, нормали и диффузные материалы. Буферы положения и нормали разделены для расчета непрямого освещения.

Шаг-3: Генерация отражающих карт теней. Сгенерированы с точки зрения источников света сцены.

Шаг-4: Создание VPL с первым отражением. Создаются путем выборки отражающих карт теней.

Шаг-5: Генерация прямого освещения с использованием источников освещения сцены и соответствующих карт теней.

Шаг-6: Генерация непрямого освещения, рассчитанного с разделением G-буфера и SVO ray-marching. Ray marching («маршрирование лучей») — это достаточно новая техника, используемая для рендеринга сцен реального времени. Она особенно интересна тем, что полностью вычисляется в шейдере экранного пространства. Мы «шагаем» точкой вдоль луча, пока не найдём, где точка пересекает объект. Сэмплирование этой точки вдоль луча является относительно простой и малозатратной операцией, и практичной для реального времени.

Шаг-7: Слияние и фильтрация. Для получения окончательного результата непрямого освещения выполняются слияние и фильтрация.

Шаг-8: Для получения окончательного изображения применяется отображение тонов HDR. В видео карте используется высококачественный пиксельный шэйдинг HDR.

Для расширения этапа выборки VPL, добавляем второе и третье отражения. В сложных сценах, второе и третье отражения VPL необходимы для освещения сложной геометрии сцен. Для создания буфера прямого освещения применяется стандартное затенение с помощью карт теней освещения сцены. Для непрямого освещения используются накопленные VPL, которые влияют на фрагмент затенения, при этом сохраняются результаты в буфере непрямого освещения.

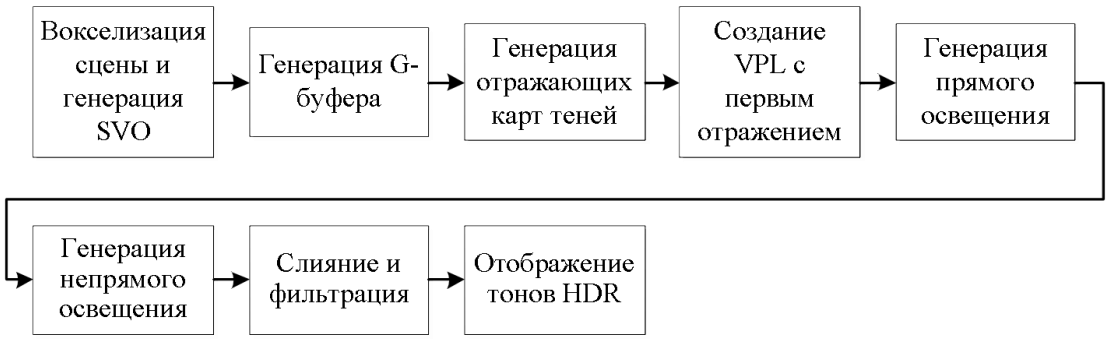


Рис. 2. Основные шаги вычисления освещения

Последовательность операций для реализации алгоритма ray-marching показана на рис. 3.

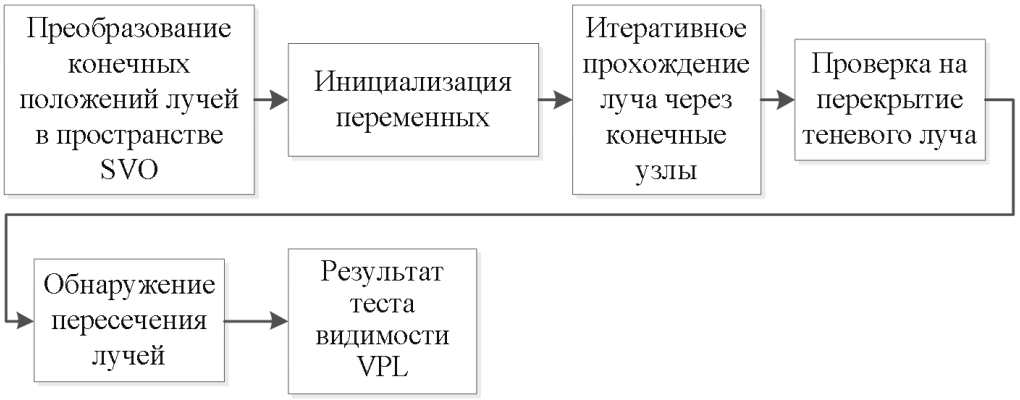


Рис. 3. Основные шаги алгоритма ray-marching

Алгоритм ray-marching использует конечные положения лучей в мировом пространстве и корень SVO в качестве входных данных. Сначала преобразуются конечные положения лучей в пространство SVO. Инициализируются следующие переменные: sceneMaxT - это расстояние между конечными позициями луча. Оно используется в качестве условия завершения перемещения луча, когда луч проходит через конечные узлы. Переменная попадания указывает, попадает ли луч в конечный узел, содержащий геометрическую информацию, или нет. Поскольку используется SVO ray marching в первую очередь для проверки видимости VPL, всякий раз, когда луч попадает в геометрию, алгоритм завершает работу и немедленно возвращается.

При итеративном прохождении луча через конечные узлы делается проверка, не происходит ли перекрытия теневого луча. В нашей реализации используется octree nodes AABB, который легко и эффективно реализован на графическом процессоре. Реализация основана на алгоритме обнаружения пересечения лучей и ограничивающего параллелепипеда, выровненного по координатным осям AABB.

Поскольку G-буфер разделен на фрагменты размером 4x4 с чередующимся шаблоном выборки, все фрагменты внутри одного и того же фрагмента используют точно такое же подмножество VPL. Таким образом, доступ к данным на узлах VPL и SVO обеспечивает хорошую локальность. Это значительно повышает производительность ray-marching теневых лучей.

IV. Результаты тестирования

Тестирование производилось на процессорах Intel Core i7 3820 и GPU NVIDIA GeForce GTX 980. Для тестирования использовались несколько сцен, четыре из них показаны на рис. 4, 5, 6 и рис. 7. Разрешение изображения для всех примеров рендеринга составляет 1024 x 768. Все сцены полностью динамичные.

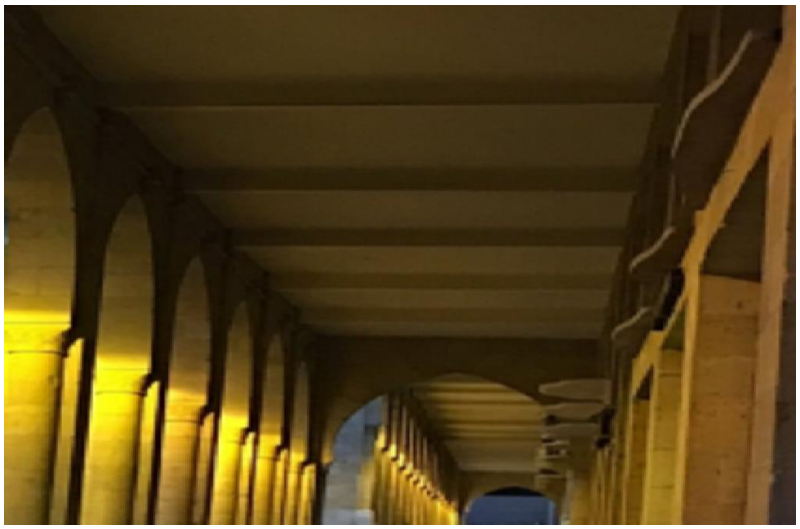


Рис. 4. Сцена со множественными источниками света (1024 VPL)

Рис. 5. Сцена с глобальным освещением со множественными источниками света (1024 VPL)

Рис. 6. Сцена с глобальным освещением со множественными источниками света (1024 VPL)

Рис. 7. Сцена с глобальным освещением со множественными источниками света (512 VPL)

G-буфер, буфер прямого освещения и буфер непрямого освещения имеют то же разрешение, что и вторичный буфер. Для чередующейся выборки VPL G-буфер разбивается на фрагменты размером 4x4. Таким образом, набор VPL разделен на 16 подмножеств, каждое из которых соответствует отдельной ячейке G-буфера. Количество VPL варьируется, и наблюдается их влияние на скорость рендеринга и качество изображения. Время в миллисекундах, затраченное на основные этапы работы метода, суммировано в табл. 1.

ТАБЛИЦА 1

РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

Сцена	Рис. 4	Рис. 5	Рис. 6	Рис. 7
Кол-во функционально заданных объектов	142	84	80	67
Кол-во треугольников	253K	196K	180K	121K
Разрешение SVO	256^3	256^3	256^3	256^3
Вокселизация сцены и генерация SVO (мс)	13,6	6,2	6,0	5,1
Генерация отражающих карт теней (мс)	0,6	0,4	0,3	0,2
Генерация VPL (мс)	0,02	0,02	0,02	0,02
Прямое освещение (мс)	0,3	0,2	0,1	0,09
Непрямое освещение (мс)	36	33	32	28
Кол-во кадров в сек.	20	25	26	30

Для непрямого освещения с одним отражением для некоторых сцен (рис. 7) не достаточно для освещения всех областей сцены. В областях, где находится недостаточное количество VPL, могут появляться артефакты.

V. Заключение

Разработан объемно-ориентированный метод визуализации трехмерных сцен с глобальным освещением в режиме реального времени. В методе используется разреженное воксельное восьмеричное дерево и ray-marching алгоритм проверки видимости с большим количеством источников света. Метод эффективен для выборки VPL со вторым и третьим отражениями, особенно для сложных сцен при сохранении частоты кадров в реальном времени, и поддерживает полностью динамичные сцены.

Тем не менее метод не так эффективен, как быстрые карты теней для вычисления непрямого освещения с одним отражением. Метод имеет ограничения, присущие визуализации в реальном времени. Например, глянцевые поверхности невозможно реконструировать без достаточного количества VPL, равномерно распределенных по всей сцене. Применение чередующейся выборки к поверхностям сцены, не освещенным непосредственно, приводит к заметным артефактам при плохом распределении VPL.

Будущей работой будет разработка надежного VPL распределения и метода сбора. Это можно будет сделать, используя преимущества структуры данных SVO. Например, разделение теневых лучей на пакеты на основе VPL, которые находятся в одном и том же телесном угле, улучшит согласованность доступа к узлу восьмеричного дерева. Другая оптимизация может быть реализована для глянцевых отражений, метод можно расширить, чтобы

получать геометрические данные о местоположении попадания луча в геометрию. Для поддержки крупномасштабных сцен планируется реализация каскадной схемы управления сеткой SVO с помощью каскадных объемов распространения света.

Источник финансирования

Работа выполнена при поддержке Министерства науки и высшего образования в рамках выполнения работ по Государственному заданию № 124041700102-4 в ИАиЭ СО РАН.

Список литературы

1. Encyclopedia of Color Science and Technology / T. Ritschel [and others]. Springer Cham, 2023. 1613 p.

2. Zhang B., Oh K. Indirect illumination with efficient monte carlo integration and denoising // Multimedia Tools and Applications. 2021. Vol. 80. № 2. P. 1–19.

3. Kern R., Brull F., Grosch T. Accelerating Photon Mapping for Hardware-based Ray Tracing // Journal of Computer Graphics Techniques. 2023. Vol. 12. № 1. P. 1-28.

4. A Framework for Learning and Rapid Implementation of Real-Time Global Illumination Methods / C. Lambru [and others] // Applied Sciences. 2022. Vol. 12. № 11. P. 1–18.

5. Encyclopedia of Computer Graphics and Games / T. Wang [and others]. Springer Cham, 2024. 1000 p.

6. NEnv: Neural Environment Maps for Global Illumination / C.R. Pardo, J. Fabre, E. Garces, J.L. Moreno // Computer Graphics Forum. 2023. Vol.42. № 4. P. 1–14.

7. Metzgar J., Semwal S.K. Scalable Spherical Harmonics Hierarchies // Journal of WSCG. 2020. Vol. 20. № 1-2. P. 116–125.

8. Keller A. Instant radiosity. // In SIGGRAPH ’97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. 1997. NY. USA. P. 49-56.

9. Imperfect shadow maps for efficient computation of indirect illumination / T. Ritschel [and others] // ACM Transactions on Graphics. 2008. Vol. 27. № 5. P. 1-8.

10. Tokuyoshi Y., Ogaki S.: Real-time bidirectional path tracing via rasterization // In: Proceedings of the ACM Symposium on Interactive 3D Graphics and Games. 2012. P. 183–190.

11. Tokuyoshi Y. Virtual spherical gaussian lights for real-time glossy indirect illumination // Computer Graphics Forum.2015. Vol. 34. P. 89–98.

12. Harmonics Virtual Lights: Fast Projection of Luminance Field on Spherical Harmonics for Efficient Rendering / P. Mezieres, F. Desrichard, D. Vanderhaeghe, M. Paulin // Computer Graphics Forum. 2022. Vol.41. № 2. P. 1–13.

13. Real-time indirect illumination by virtual planar area lights / W. Zeng [and others] // Computers & Graphics. 2021. Vol. 97. № 4. P. 183-190.

14. NeLT: Object-oriented Neural Light Transfer / C. Zheng [and others] // ACM Transactions on Graphics. 2023. Vol. 42. № 5. P. 1–16.

15. Foveated Stochastic Lightcuts / X. Shi [and others] // IEEE Transactions on Visualization and Computer Graphics. 2022. Vol. 28. № 11. P. 3684-3693.

16. Survey on Bounding Volume Hierarchies for Ray Tracing / D. Meister [and others] // Computer Graphics Forum. 2021. Vol. 40. № 2. P. 683-712.

17. Wang S., Holzschuch N. Adaptive Matrix Completion for Fast Visibility Computations with Many Lights Rendering // Computer Graphics Forum. 2020. Vol. 39. № 4. P. 47-58.

18. Liu T., Gao J., Lei Z. An Approach to Global Illumination Calculation Based on Hybrid Cone Tracing // IEEE Access. 2020. Vol. 8. P. 92061 - 92071.

19. Implementing Many-Lights Rendering with IES-Based Lights / D. Gadia [and others] // Applied Sciences. 2024. Vol. 14. № 3. P. 1-15.

20. Santos L.P., Marques R., Colom A. Interactive VPL-based global illumination in the GPU using fuzzy clustering // Computers & Graphics. 2022. Vol. 108. № 4. P. 74-85.

21. Interactive indirect illumination using voxel cone tracing / C. Crassin [and others] // Computer Graphics Forum. 2011. Vol. 30. P. 1921-1930.

22. Вяткин С. И., Долговесов Б. С. Метод сжатия геометрических данных с применением функций возмущения // Автометрия. 2018. Т.54. №4. С. 18- 25.

23. Vyatkin S. I. Polygonization method for functionally defined objects // International Journal of Automation, Control and Intelligent Systems. 2015. Vol. 1, № 1. P. 1−8.

24. Wavelet importance sampling: efficiently evaluating products of complex functions / P. Clarberg [and others] // ACM Transactions on Graphics. 2005. Vol. 24. P. 1166-1175.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-46-54

Расширение методов вычисления показателя ослепленности для двумерных изображенний большого динамического диапазона
Enhancement of glare rating calculation methods for 2-dimensional high-dynamic range images

Е.Д. Бирюков, М. С. Копылов
Институт прикладной математики им. М.В. Келдыша РАН, Москва, Россия

E.D. Birukov1, M.S. Kopylov1
[1]Keldysh Institute of Applied Mathematics of RAS, Moscow, Russia

Аннотация. Эта статья описывает различные методики улучшения работы алгоритмов вычисления показателя ослепленности. Вычисление показателя ослепленности с помощью формул, работающих с точечными источниками света, у которых известны координаты положения в трехмерном пространстве, не всегда является возможным. В некоторых случаях в качестве входных данных по тем или иным причинам невозможно использовать готовую трехмерную сцену с известными пространственными координатами источников света, точки расположения наблюдателя, и положения других объектов. Вместо этого требуется использовать единственное двумерное изображение, на котором не заданы явно координаты источников света и каких-либо других объектов. Точечные источники света при этом необходимо заменить бликовыми областями, т.е. группами соседних пикселей изображения, удовлетворяющих некоторым условиям. Дополнительной проблемой является использование источников света с неоднородным пространственным распределением яркости, к которым относятся, в первую очередь, светодиодные светильники, а также дневного освещения. В настоящей работе предлагаются некоторые улучшения существующих методов вычисления показателя ослепленности для двумерных изображений. В частности, предлагается вычисление индекса Гута с использованием средневзвешенного суммирования по яркости пикселей. Такой подход дает более адекватные результаты, в первую очередь, при использовании неоднородных источников света. Также предлагается использование дополнительной попиксельной маски для объединения или исключения некоторых бликовых областей. Такой подход также применяется для неоднородных источников света, а также позволяет точнее обрабатывать бликовые области сложной формы.

***Ключевые слова*:** показатель ослепленности, объединенный показатель дискомфорта, показатель ослепленности для двумерных изображений, индекс Гута, бликовые области, широкоугольные камеры, изображения большого динамического диапазона.

Annotation. This article describes different methods of improving glare rating calculation algorithms. Calculation of the glare rating using formulas that work with point light sources whose three-dimensional coordinates are known is not always possible. In some cases usage of a ready-made three-dimensional scene with known spatial coordinates of light sources, the observer's location, and the positions of other objects as input data, is impossible for different reasons. Instead, it is necessary to use a single two-dimensional image where the coordinates of light sources and any other objects are not explicitly specified. In this case, point light sources must be replaced with glare areas, i.e. groups of adjacent image pixels that satisfy certain conditions. An additional problem is usage of light sources with non-uniform spatial brightness distribution, which primarily include LED lamps, as well as daylight. This article proposes some improvements to the existing methods for calculating the glare rating for two-dimensional images. In particular, it is proposed to calculate the Guth position index using a weighted average sum of pixel brightness. This approach gives more adequate results, first of all, when using non-uniform light sources. It is also proposed to use an additional per-pixel mask to combine or exclude some glare areas. This approach is also used for non-uniform light sources, and also provides more accurate processing of glare areas of complex shape.

***Keywords*:** glare rating, uniform glare rating (UGR), glare rating for two-dimensional images, Guth position index, glare regions, fisheye cameras high-dynamic range images.

I. Введение

Блики яркого света представляют собой значительный профессиональный риск в рабочих средах, обусловливая необходимость их прогнозирования и оценки как одного из важнейших аспектов качества осветительных установок. Начиная с первой половины прошлого века исследователи стремились количественно

определить дискомфортный уровень бликов, однако до сих пор физиологические и психофизические механизмы данной проблемы остаются полностью не изученными.

Для оценки различных аспектов дискомфорта от бликов были разработаны различные варианты показателя ослепленности. Оценка бликов от окон может быть выполнена с помощью показателя ослепленности дневным светом (DGI), предложенной Хопкинсоном [1]. Винольд и Кристоффersен расширили DGI, разработав показатель вероятности ослепления дневным светом (DGP) [2]. Для количественной оценки дискомфорта от бликов при внутреннем освещении Международная комиссия по освещению (МКО) предложила Объединенный показатель дискомфорта (Unified Glare Rating, UGR). В Северной Америке Гутом был разработан альтернативный показатель, известный как вероятность визуального комфорта (VCP) [3]. Более ранними показателями ослепленности для закрытых помещений являются британский показатель ослепленности (BGI) и показатель ослепленности МКО (CGI). В Европе UGR является общепринятым методом оценки дискомфорта от ослепления при внутреннем освещении и включена в европейский стандарт по освещению рабочих мест в помещениях EN 12464-1 [4].

Расчет большинства вышеупомянутых видов показателя ослепленности основан на распределении силы света, которое аппроксимирует любой неравномерный источник света в виде однородного точечного источника. Поскольку неоднородные источники света не были приняты во внимание при разработке традиционных разновидностей показателя ослепленности, точность определения дискомфортных бликов от неоднородных источников света с использованием современных вариантов показателя ослепленности все еще обсуждается. Например, светодиодные светильники часто демонстрируют неравномерное распределение освещенности с областями высокой освещенности, перемежающимися с относительно темными зонами. С ростом доли светодиодных светильников на рынке освещения интерьеров становится крайне важной более точная оценка дискомфортных бликов для этих неоднородных источников света.

Настоящая работа описывает реализацию алгоритмов подсчёта показателя ослепленности, которые позволяют устранить некоторые ограничения традиционных разновидностей такого показателя и повысить точность их оценки. Конечная цель работы заключается во внедрении разработанных алгоритмов в систему оптического моделирования в виде специального расширения. Данное расширение позволяет рассчитывать необходимые варианты показателя ослепленности как для результатов фотореалистичного рендеринга, так и для реальных изображений, полученных с камер.

II. Обзон методов расчета показателя ослепленности для трехмерных сцен

Количественное значение показателя ослепленности описывает субъективное ощущение дискомфорта от блика, выраженное с помощью численной шкалы. Но для таких субъективных ощущений до сих пор не придумано универсальной шкалы, пригодной для всех случаев. Первая попытка оценки психологического уровня восприятия бликов была сделана Хопкинсоном в 1940 году. Он предложил четыре уровня восприятия бликов: «ощущаемый», «приемлемый», «некомфортный» и «невыносимый». Позже, в 1972 году, были приняты промежуточные критерии, и это привело к появлению 9-уровневой шкалы, которая широко используется для оценки дискомфорта от бликов внутри помещений [5].

Существуют несколько различных формул, которые описывают субъективную степень дискомфорта от бликов. Большинство из них включает в себя ту или иную комбинацию одинаковых параметров: яркость каждого конкретного источника света, яркость фона, телесный угол, стягиваемый источником света, угол между осевым направлением от наблюдателя и направлением от наблюдателя до источника света [6, 7]. Одними из наиболее популярных являются Объединенный показатель дискомфорта (Unified Glare Rating, UGR); показатель блескости (Glare Rating, GR); показатель вероятности ослепления дневным светом (Daylight Glare Probability, DGP) и показатель предсказания ощущения от дневного света (Predicted Glare Sensation Vote, PGSV). Показатель блескости (GR) применяется для анализа изображений открытого пространства, в то время как остальные методы применяются в основном для анализа изображений внутри помещений с различными типами освещения (естественный свет и источники искусственного освещения). В этой статье мы подробно остановимся на объединенном показателе дискомфорта и показателе блескости, которые используются в разработанной нами системе обработки изображений.

4. Объединенный показатель дискомфорта

Объединенный показатель дискомфорта применяется в основном для сцен, представляющих собой некоторое закрытое помещение. На таких сценах легко выделить источники искусственного освещения, а также окна, которые фактически тоже являются источниками света. Стандартная эмпирическая формула для вычисления

показателя ослепленности, предусматривающая использование точечных источников света, была предложена МКО [8]. Она выглядит следующим образом:

$$UGR = 8\log_{10}\frac{0.25}{L_b}\cdot\sum_{i=1}^{n}\frac{L_s^2\omega_s}{P^2}. \quad (3)$$

Здесь Lb – фоновая яркость; Ls – яркость каждого источника света в сцене; ω_s – телесный угол (в стерадианах) бликового региона (считается относительно положения камеры); P – индекс Гута. Телесный угол в данной формуле должен удовлетворять условию $0.0003 < \omega_s < 0.1$. В настоящее время объединенный показатель дискомфорта широко используется для определения дискомфорта от бликов, вызванных источниками света внутри помещений, и включена в европейский национальный стандарт освещения рабочих помещений (CEN 2012).

Хотя вычисление объединенного показателя дискомфорта требует яркость источника света на входе, традиционно применяется преобразование из яркости в интентивность дальнего поля [9]. Однако, некоторые исследователи отметили, что при использовании распределения интенсивности дальнего поля неоднородные источники света не отличаются от однородных [10]. При использовании светодиодов, источники с неоднородным пространственным распределением яркости стали более распространенными, и исследования показали важность неоднородности при подсчете дискомфорта от бликов. Это послужило толчком к целому ряду исследований, целью которых была разработка методов вычисления показателя ослепленности для неоднородных источников света [9].

5. Показатель блескости

Показатель блескости применяется для сцен, представляющих собой открытое пространство. Эмпирическая формула, предложенная МКО в 1994 году для показателя блескости, следующая:

$$GR = 27 + 24\log_{10}\left(\frac{L_{vl}}{L_{ve}^{0.9}}\right). \quad (4)$$

Здесь L_{vl} – вуалирующая яркость всех источников света в сцене, а L_{ve} – вуалирующая яркость фона. Вуалирующая яркость источников света вычисляется по следующей формуле:

$$L_{vl} = 10 * \sum_{i=1}^{n}\frac{E_{eye\,i}}{\Theta_i^2} \quad (5)$$

где $E_{eye\,i}$ – освещенность в камере на плоскости, перпендикулярной линии обзора, от i-го источника света, а Θ_i – угол между осевой линией камеры и направлением лучей света от i-го источника (в градусах от 1.5° до 60°).

III. Вычисление показателя ослепленности для двумерных изображений

1. Переход от точечных источников света к бликовым областям

Формулы, для объединенного показателя дискомфорта и показателя блескости, приведенные выше, работают с точечными источниками света, координаты которых в трехмерном пространстве заведомо известны. Но исходные данные для показателя ослепленности далеко не всегда выражены в форме трехмерной сцены с известными координатами источников света. Зачастую показатель ослепленности необходимо рассчитать для обычного двумерного изображения, на котором источники света не выделены явным образом, и их координаты неизвестны.

Использование двумерных изображений требуется, в первую очередь, в тех случаях, когда показатель ослепленности необходимо расчитать для уже построенных зданий и помещений в этих зданиях. Несмотря на возможность построения трехмерной модели этих помещений по чертежам (либо конвертации существующих моделей в формат, пригодный для вычисления показателя ослепленности), реально существуюие помещения имеют свои особенности в виде материалов, физические свойста которых отличаются от заданных в проекте; различных жалюзи и штор, закрывающих окна; растений, которые перекрывают часть поступающего дневного света, и других факторов. В таких случаях ранее требовалось тщательное изучение и измерение параметров всех этих дополнительных факторов для последующих вычислений, либо непосредственное измерение освещенности в большом количестве точек [11, 12]. Однако, все эти трудоемкие операции можно заменить вычислением показателей ослепленности по двумерным изображениям, сделанным с помощью фотокамер.

С данными в виде двумерных изображений работает в том числе наша реализация алгоритмов вычисления показателя ослепленности, встроенная в программу обработки изображений большого динамического диапазона. Для двумерных изображений требуются иные способы задания источников света. Как правило, при работе с такими изображениями вместо точечных источников света используются так называемые бликовые области, т.е. области из соседних (находящихся вплотную друг к другу) пикселей, яркость которых удовлетворяет заданным условиям. Подобная методика также используется, например, в [13, 14, 15], а также во многих других работах.

Следует заметить, что алгоритмы вычисления показателя ослепленности на двумерных изображениях работают только с изображениями большого динамического диапазона, поскольку такие изображения хранят данные реальных физических величин (яркости и освещенности), и их динамический диапазон достаточен для выделения источников света на остальной части изображения. Эти методы можно применить и при вычислении показателя вероятности ослепления дневным светом, используя его полную формулу, а не упрощенный вариант, что повышает точность вычислений.

Для задания бликовой области на изображении чаще всего используются три метода [15]. Все эти методы предполагают отнесение к бликовой области группы соседних пикселей, яркость которых превышает некоторый определенный предел. Однако этот предел задается различными способами. В первом методе к бликовым областям относятся пиксели, яркость которых превышает некоторую константу (фиксированную либо задаваемую пользователем). Во втором методе критерием является превышение яркости пикселя над средней яркостью всего изображения в определенное количество раз. Наконец, третий метод является разновидностью второго, но критерием в нем является превышение над средней яркостью не всего изображения, а определенной его части, заранее заданной пользователем.

В нашей работе используется разновидность первого метода, когда яркость пикселя должна превышать некоторую величину, задаваемую пользователем. Однако отличительной особенностью нашего метода является использование дополнительного верхнего предела, такого, что все пиксели с яркостью, превышающей его величину, не относятся к бликовым областям. Эта величина также указывается пользователем. Такой подход позволяет более гибко обрабатывать изображения и выключать из каждого конкретного расчета некоторые лишние области, например, для того, чтобы не учитывать какой-то определенный источник света.

2. Нормировка параметров источника света по пикселям бликовой области

Основной проблемой при переходе к двумерным изображениям является вычисление яркости источника света, а также других величин, используемых в формулах, упомянутых в предыдущем разделе, в частности, индекса Гута. В работах [15, 16, 17] уже были описаны методы вычисления этих величин для двумерных изображений. В простейшем случае для этого используется средняя яркость всех пикселей, которая вычисляется по следующей формуле:

$$\bar{\bar{L}} = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{8}$$

где N – общее количество пикселей в области, а L_i – яркость каждого пикселя.

Но во многих случаях простейшей формулы не достаточно для корректной обработки изображений. Вместо этого требуется нормировка яркостей пикселей с некоторым нормировочным множителем в зависимости от расположения пикселя относительно осевой линии изображения и других параметров.

В частности, в работе [18] рассматривается вычисление показателя ослепленности применительно к изображениям с камеры «рыбий глаз». Для подобных изображений простейшая формула средней яркости не работает, поскольку она дает различающиеся результаты для разных типов проекции и углов обзора. Поэтому для таких изображений требуется формула, учитывающая угол между направлением на пиксель из камеры и осевой линией камеры (называемый также углом падения луча). Она выглядит следующим образом:

$$\bar{L}_e = \frac{\sum_{i=1}^{N}\left(L_i\frac{\sin\theta_i}{\theta_i}\right)}{\sum_{i=1}^{N}\left(\frac{\sin\theta_i}{\theta_i}\right)} \tag{9}$$

Здесь θ_i – угол падения луча.

Особенностью же нашей работы является возможнось использования изображений с самых разных камер, в том числе таких изображений, где углы падения лучей в каждом конкретном пикселе неизвестны. Поэтому нормировка по углам падения луча в нашем случае неприменима. Мы используем нормировку по площади пикселей:

$$L_s = \frac{\sum_i L_i S_i}{\sum_i S_i} \tag{10}$$

Фактически такая нормировка соответствует простому среднему арифметическому яркостей пикселей, однако она позволяет корректно обрабатывать случаи, когда площади пикселей отличаются. Поскольку данная формула некорректно работает с большими углами обзора, то для изображений сферической проекции, которые зачастую получаются при использовании камер «рыбий глаз», угол обзора которых известен, вычисление показателя ослепленности производится только для углов падения лучей, меньших 90 градусов.

Наиболее важным изменением, примененным при разработке нашего метода, было использование нормировки по яркости пикселей при вычислении индекса Гута. Такой подход обусловлен частыми случаями неравномерного распределения яркости внутри бликовой зоны, и в том числе позволяет лучше обрабатывать

источники света с подобным распределением яркости (включая светодиодные светильники), описанные в предыдущих разделах. Бликовая зона может содержать как яркие, так и темные пиксели, а следовательно, простое среднее арифметическое, как и нормировка по площади, не будет отражать реальное визуальное восприятие бликов человеком. Даже если в бликовую область включаются только пиксели, яркость которых превышает заданный предел, внутри этой области яркость может быть распределена очень неравномерно. Зачастую бликовая область состоит из небольшого количества наиболее ярких пикселей, и большого ореола вокруг них. При этом пороговое значение для перехода от самого блика к его ореолу будет разным для каждого источника света, что мешает просто повысить порог яркости бликовой области. В таком случае следует в первую очередь учитывать наиболее яркие пиксели каждой бликовой области, весовой коэффициент которых будет выше, чем для остальных пикселей.

Формула для вычисления индекса Гута с учетом нормировки по яркости выглядит следующим образом:

$$P = \frac{\sum_i P_i L_i}{\sum_i L_i} \tag{11}$$

где P_i – индекс Гута для соответствующего пикселя, а L_i – яркость данного пикселя.

Эффект от нормировки по яркости наиболее заметен для бликовых зон, содержащих ярко выраженные светлые пиксели. Пример подобной зоны, характеризующейся высокой пространственной неоднородностью яркости, которая проявляется в резких перепадах между соседними пикселями, показан на рис 1.

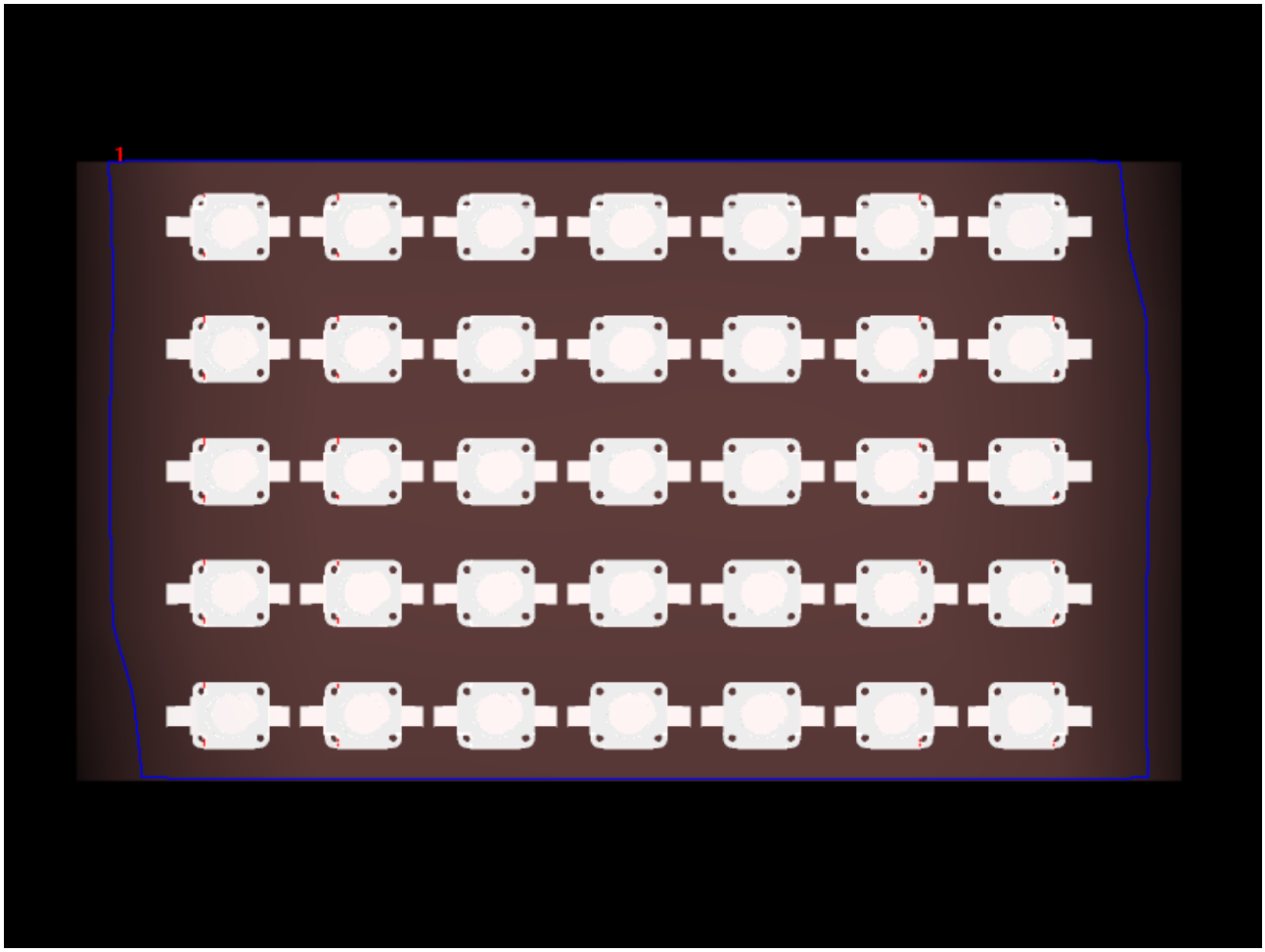


Рис. 1. Бликовая зона от светодиодного источника света имеющая неравномерное распределение яркости.

В случаях аналогичных этому, использование нормировки по яркости позволяет получить более точную оценку индекса Гута, а следовательно, более адекватно отразить восприятие бликовой зоны отражая ее неравномерность, как это видно из результатов показанных на рис.2 и рис. 3.

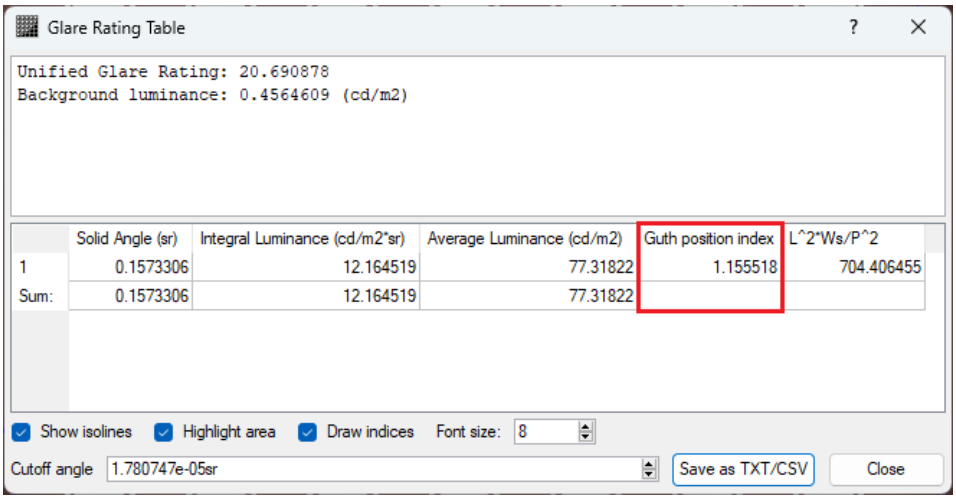


Рис. 2. Результаты подсчёта индекса Гута, нормированного по яркости.

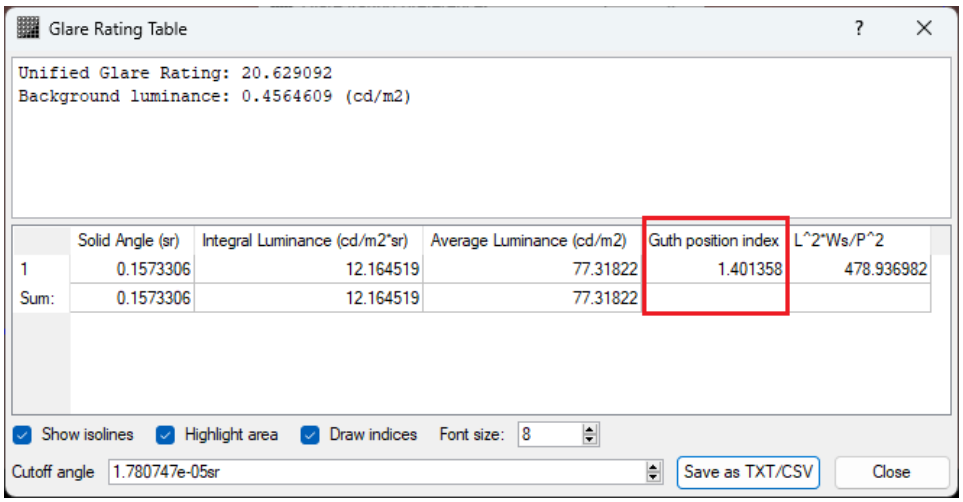


Рис. 3. Результаты подсчёта индекса Гута без нормировки по яркости (с нормировкой по площади).

IV. Способы уточнения границ бликовых областей

При расчете показателя ослепленности для двумерных изображений и представлении источников света как бликовых областей возникают дополнительные затруднения. В первую очередь, они связаны с наличием темных полос на освещенных областях, которые приводят к разделению фактически единой освещенной области на несколько отдельных областей. Примером такого случая является окно, на котором установлены жалюзи или подобные затеняющие элементы. Работа описанного выше алгоритма расчета показателя ослепленности приведет к тому, что такое окно будет расцениваться как большое количество отдельных областей, что является неудобным для сбора статистики.

Для устранения данной проблемы чаще всего пользователю предоставляется возможность задать некоторый интервал в пикселях между соседними яркими регионами, такой, что такие регионы, расположенные друг от друга на расстоянии не более данного интервала, будут считаться единым регионом. При этом следует обратить внимание, что сами темные промежутки не включаются в расчет показателя ослепленности, то есть яркость этих пикселей не учитывается в статистике. Подобный метод применяется, в частности, в работе [14].

Вышеописанный метод используется и в нашей реализации алгоритма расчета показателя ослепленности. Но если в работе [14] по умолчанию используется некоторый ненулевой интервал, то в нашей реализации интервал по умолчанию отсутствует. Это дает возможность лучше определять отдельные источники света на изображении при первоначальном запуске, так как основным применением наших алгоритмов вычисления показателя ослепленности являются сцены с отдельными точечными источниками света, такими, как светодиодная подсветка.

Работа нашего алгоритма объединения соседних областей с примером области, для которой необходимо применение интервала между соседними областями и соответствующим пользовательским интерфейсом отображены на рис. 4.

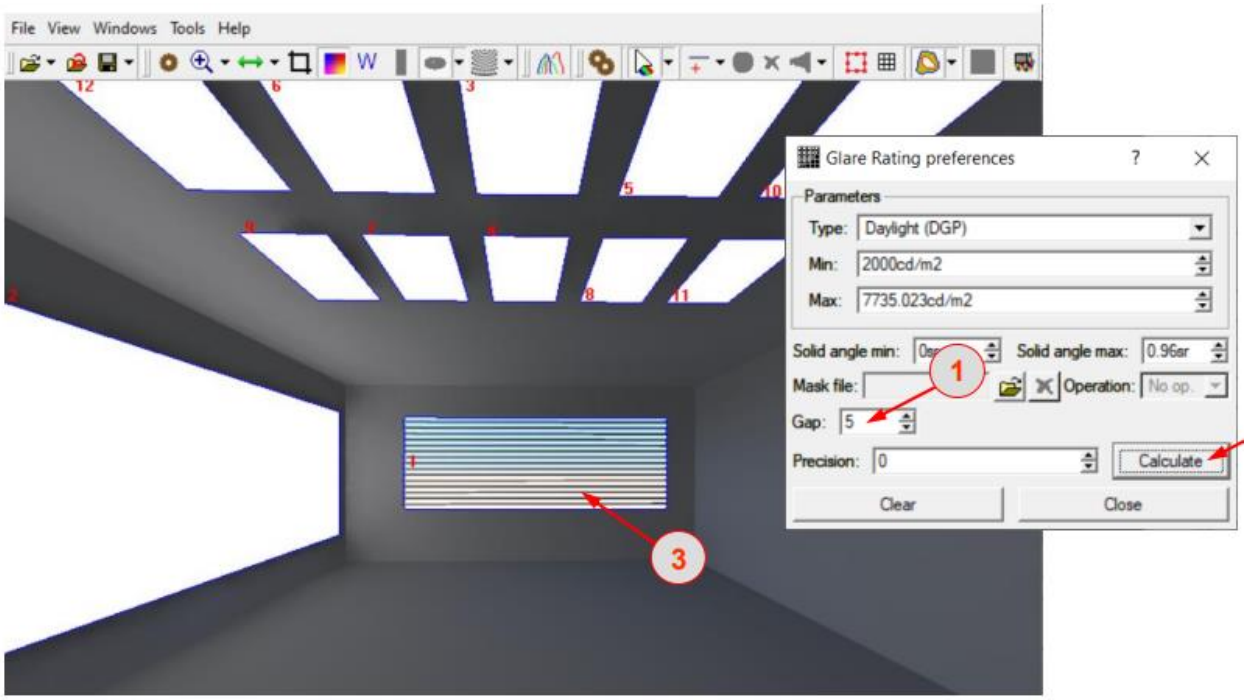


Рис. 4. Пользовательский интерфейс и результат работы алгоритма расчета показателя ослепленности с объединением близкорасположенных областей

В некоторых случаях также необходимо применить некоторую фильтрацию регионов, включаемых в расчет показателя ослепленности. Например, можно выключить какие-либо регионы из расчета показателя ослепленности, либо добавить дополнительные регионы, яркость пикселей которых выходит за пределы заданного диапазона. Для подобных случаев в нашей реализации расчета показателя ослепленности применяется оригинальный метод. Он заключается в использовании дополнительного изображения-маски. Данное изображение является монохромным, и фактически представляет собой матрицу булевых значений, которые могут использоваться для определения областей, участвующих в вычислении показателя ослепленности, самостоятельно либо совместно с основным изображением.

При выборе маски необходимо задать тип логической операции, который будет определять ее взаимодействие с основным изображением. Предусмотрены следующие типы операций: отсутствие операции (использование только маски без основного изображения), логическое ИЛИ, логическое И. Операция «логическое И» предусматривает, что для попадания пикселя исходного изображения в статистику показателя ослепленности необходимо, чтобы одновременно выполнялись два условия: попадание яркости пикселя исходного изображения в требуемый диапазон и логическая единица (белый цвет) в соответствующем пикселе на изображении-маске. Таким образом возможно выключить некоторые области из статистики показателя ослепленности. Пример работы данной операции можно посмотреть на рис. 5.

Операция «логическое ИЛИ» предусматривает, что пиксель исходного изображения учитывается в статистике показателя ослепленности в том случае, если выполняется хотя бы одно из вышеописанных условий. Это обеспечивает возможность дополнительно включать некоторые регионы в расчет показателя.

Операцию «логическое ИЛИ» можно использовать для объединения близкорасположенных ярких регионов с темными промежутками вместо предыдущего метода (задание интервала). Однако следует учитывать, что в отличие от метода задания интервала в данном случае темные промежутки также будут включены в расчет показателя ослепленности.

V. Заключение

Различные варианты показателя ослепленности широко применяются в настоящее время для расчетов в области архитектуры, эргономики офисных и производственных помещений, градостроительстве и в других случаях. Вычисление показателя ослепленности приобретает особую важность при использовании источников света с неоднородным распределением яркости, таких, как повсеместно распространенные в последнее время светодиодные светильники, а также при использовании дневного света в помещениях.

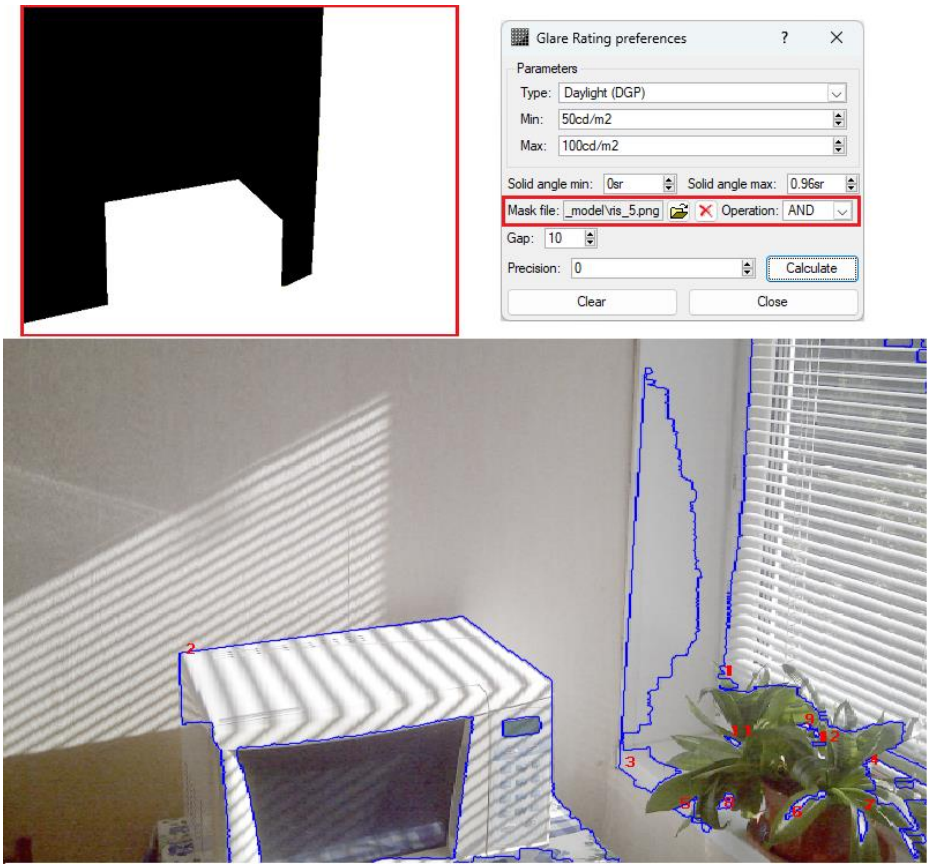


Рис. 5. Работа алгоритма с применением маски с операцией «логическое И»

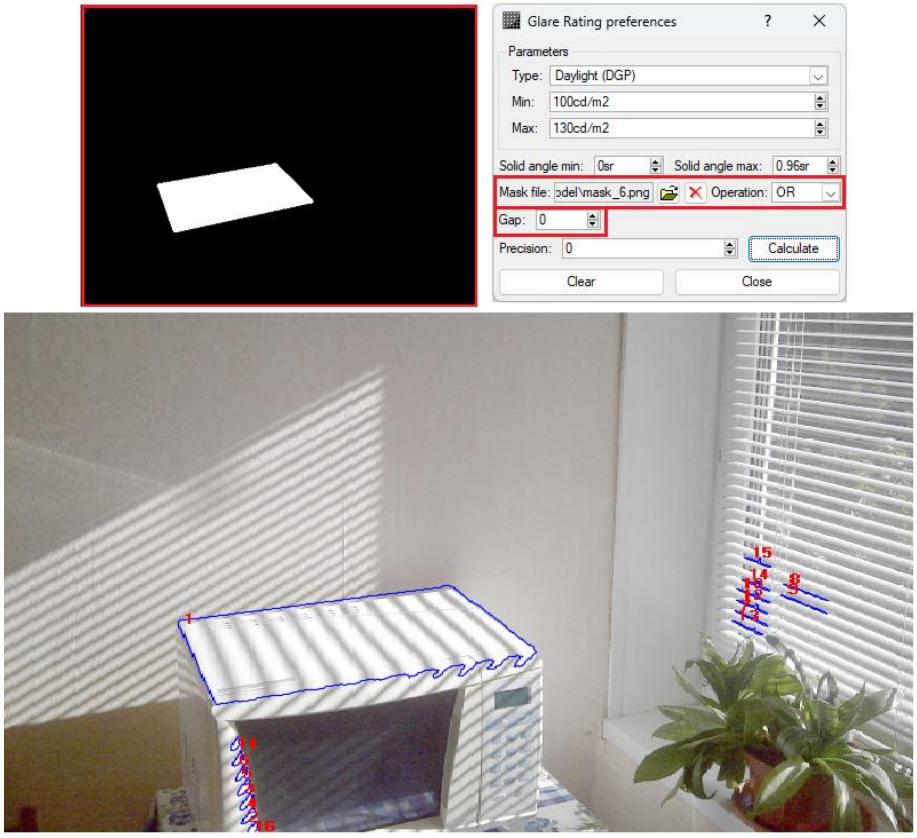


Рис. 6. Работа алгоритма с применением маски с операцией «логическое ИЛИ»

Использование методов, работающих с двумерными изображениями, существенно расширяет область применения показателя ослепленности и облегчает работу по подготовке исходных данных для ее вычисления. Вместо сложных измерений взаимного положения наблюдателя, всех источников света и других объектов, либо построения трехмерной модели помещения со всеми вышеуказанными элементами, достаточно получить качественную фотографию в формате изображения с большим динамическим диапазоном. Такую возможность предоставляют многие современные фотокамеры.

Разработанные нами усовершенствования метода вычисления показателя ослепленности для двумерных изображений, такие, как нормировка значений индекса Гута для пикселей по яркости этих пикселей, а также применение дополнительного изображения-маски для объединения соседних бликовых областей либо исключения лишних областей, позволяют еще более упростить процесс вычисления и улучшить его результаты. В частности, они позволяют лучше обрабатывать источники света с неоднородным распределением яркости. Практическая реализация алгоритмов вычисления показателя ослепленности, описанных в данной статье, в настоящее время успешно применяются в разработанном нами комплексе оптического моделирования.

Список литературы

[1] Hopkinson R. G. Glare from daylighting in buildings //Applied ergonomics. – 1972. – Т. 3. – №. 4. – С. 206-215.

[2] Wienold J., Christoffersen J. Evaluation methods and development of a new glare prediction model for daylight environments with the use of CCD cameras //Energy and buildings. – 2006. – Т. 38. – №. 7. – С. 743-757.

[3] Rubiño M. et al. Discomfort glare indices: a comparative study //Applied optics. – 1994. – Т. 33. – №. 34. – С. 8001-8008.

[4] Light and Lighting. Lighting of Work Places. Part 1. Indoor Work Places. European Standard. EN 12464-1, European Committee for Standardization (CEN), Brussels, 2011.

[5] Sawicki D., Wolska A. The unified semantic glare scale for GR and UGR indexes //2016 IEEE Lighting Conference of the Visegrad Countries (Lumen V4). – IEEE, 2016. – С. 1-6.

[6] M. Rea, editor. IES Lighting Handbook, Reference Volume 3, page 79. Illuminating Engineering Society of North America, New York, 1993.

[7] Discomfort glare in the interior working environment, 1983. Publication CIE, No. 55 (TC-3.4), Commission Internati´onale de l'´Eclairage (CIE), Vienna, Austria, 43 pp.

[8] CIE, Discomfort Glare in the Interior Lighting, Commission Internationale de lE clairage (CIE), Technical committee TC-3.13, Division 4, Interior Environment and Lighting Design, Vienna Austria, 1992.

[9] Spieringhs R. M. et al. Exploring the applicability of the unified glare rating for an outdoor non-uniform residential luminaire //Sustainability. – 2022. – Т. 14. – №. 20. – С. 13199.

[10] Scheir G. H. et al. Calculation of the Unified Glare Rating based on luminance maps for uniform and non-uniform light sources //Building and Environment. – 2015. – Т. 84. – С. 60-67.

[11] Coyne S. et al. The use of high dynamic range luminance mapping in the assessment, understanding and defining of visual issues in post occupancy building assessments //International conference on energy efficiency in commercial buildings (IEECB'08). Frankfurt, Germany, 10e11 April. – 2008.

[12] Hirning M., Coyne S., Cowling I. The use of luminance mapping in developing discomfort glare research

[13] Funke C., Schierz C. Extension of the unified glare rating formula for non-uniform LED luminaires //Proc. 12th Lux Junior Conf. Ilmenau. – 2015. – С. 80-81.

[14] Wienold J. Dynamic simulation of blind control strategies for visual comfort and energy balance analysis. – 2007.

[15] Pierson C., Wienold J., Bodart M. Daylight discomfort glare evaluation with Evalglare: influence of parameters and methods on the accuracy of discomfort glare prediction //Buildings. – 2018. – Т. 8. – №. 8. – С. 94.

[16] Dubois M. C. Impact of shading devices on daylight quality in offices //Simulations with Radiance. – 2001.

[17] Shin J. Y., Yun G. Y., Kim J. T. View types and luminance effects on discomfort glare assessment from windows //Energy and Buildings. – 2012. – Т. 46. – С. 139-145.

[18] Hirning, Michael B. The application of luminance mapping to discomfort glare: a modified glare index for green buildings. Diss. Queensland University of Technology, 2014.

//Journal of Light & Visual Environment. – 2010. – Т. 34. – №. 2. – С. 101-104.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-55-66

Визуализация дорожных сцен на основе характеристик отражения МКО
Visualization of road scenes on the base of CIE reflection characteristics

Д. Д. Жданов1, В. Г. Соколов1, И. С. Потемин1, А. Г. Волобой2
1Университет ИТМО, Санкт-Петербург, Российская Федерация
2Институт прикладной математики им. Келдыша, Москва, Российская Федерация

D. D. Zhdanov1, V. G. Sokolov1, I. P. Potemin1, A. G. Voloboy2
[1]ITMO University, St. Peterburg, Russian Federation
[2]Keldysh Institute of Applied Mathematics RAS, Moskov, Russian Federation

Аннотация. Построение реалистичных изображений дорожных сцен, физически корректное проектирование систем освещения дорог требует точного задания оптических свойств всех элементов сцены и особенно дорожного покрытия. Отражательные свойства дорожных покрытий сложны и зависят от множества факторов: материала, из которого они изготовлены (тип асфальта, бетона), погодных условий (сухое или мокрое дорожное покрытие). Такие отражательные свойства, заданные в угловом пространстве, можно определить с помощью двунаправленной функции отражения (ДФР). Однако для полноценных и точных измерений ДФР требуется сложное оборудование, и такие измерения предпочтительно проводить на месте, что ещё более усложняет задачу получения этих свойств. При этом существует стандартный набор отражательных свойств, представленный спецификацией Международной комиссии по освещению (МКО), так называемые таблицы отражения, так называемые «r-таблицы». Свойства, заданные «r-таблицами», определяют отражение поверхности дороги в очень ограниченном угловом пространстве и могут быть источником погрешностей при визуализации дорожных сцен. В данной статье показан подход к преобразованию «r-таблиц» в более общий формат двунаправленной функции отражения, используемый в коммерческом оптическом программном обеспечении, представлены изображения дорожных сцен, синтезированных как на основе r-таблиц, так и на основе полностью измеренной двунаправленной функции рассеяния, дано сравнение обоих подходов к описанию отражательных свойств дороги.

***Ключевые слова*:** Дорожное покрытие, Дорожные сцены, Двунаправленная Функция Отражения, ДФР, r-таблица, Асфальт, Бетон, Международная комиссия по освещению, МКО.

Аннотация. The generation of realistic images for road scenes, physically correct design of illumination systems for roads requires the precise specification of optical properties for all scene elements and especially road surface. The reflection properties of road surfaces are complex and depend on a lot factors: material they are made (kind of asphalt, concrete), weather conditions (dry or wet road surface). Such reflective properties specified in while angular space can be determined with Bi-Directional Reflectance Distribution Function (BRDF). However, whole, and accurate BSDF measurements require complex equipment, and it is preferable to launch such measurements on-site. At the same time there are standard set of reflective properties presented in specification by International Commission on Illumination (CIE), so called reflection tables, so called “r-tables”. The properties defined by “r-tables” determine road surface reflection on very limited angular space and can be source incorrectness for visualization of road scenes. In the given paper it is showed the approach to transform CIE “r-tables” into more general BRDF format acceptable with commercial optical software, images of road scenes are generated as on the base of “r-tables” as well on entirely measured BRDF. It is presented the comparison of both approaches to describe road reflectance properties

***Ключевые слова*:** Road surface, Road scenes, Bi-Directional Scattering Distribution Function, BRDF, r-tables, Asphalt, Concrete, International Commission on Illumination, CIE.

I. Введение

Точное оптическое моделирование дорожных сцен требует физически корректной спецификации отражательных свойств дорожного покрытия. Обычно эти свойства задаются двунаправленной функцией отражения (ДФО) [1-7]. В большинстве случаев ДФО определяется с помощью измерений. Для измерения ДФО дорожного покрытия во всем угловом пространстве необходимо применение специальных измерительных приборов, которые достаточно дороги и не всегда доступны. Кроме того, форма ДФО дорожного покрытия достаточно сложна (см. рис. 1) и зависит от состояния дорожного покрытия (сухости, различной степени влажности, износа), что требует проведения измерений с высоким угловым разрешением, особенно для мокрых

покрытий. Другая проблема заключается в том, что ДФО дорожного покрытия, должна определяться в направлениях наблюдения, близких к поверхности, а точное измерение в направлениях, близких к горизонту, представляет собой очень сложную задачу. Учитывая упомянутые проблемы, в данной статье предлагается использовать стандартные характеристики отражения, так называемые «r-таблицы», представленные в CIE 144–2001 [8], подробности см. в следующей главе. Эти характеристики отражения можно интерпретировать как ДФО, определённой в ограниченном угловом пространстве. Согласно МКО, эти данные могут использоваться при дорожном освещении, дневном свете и освещении фар транспортных средств, однако в случае использования этих данных при визуализации дорожных сцен требуется расширение «r-таблиц» данными для дополнительных направлений наблюдения. В работе представлено исследование использования «r-таблиц» без каких-либо дополнительных измерений в задачах визуализации, рассматривается алгоритм преобразования «r-таблиц» в более общий формат ДФР, поддерживаемой коммерческим программным комплексом Lumicept [9]. В статье показано сравнение изображений дорожных сцен, синтезированных на основе измеренных ДФО реального дорожного покрытия с их аналогами, полученными на основе «r-таблиц».

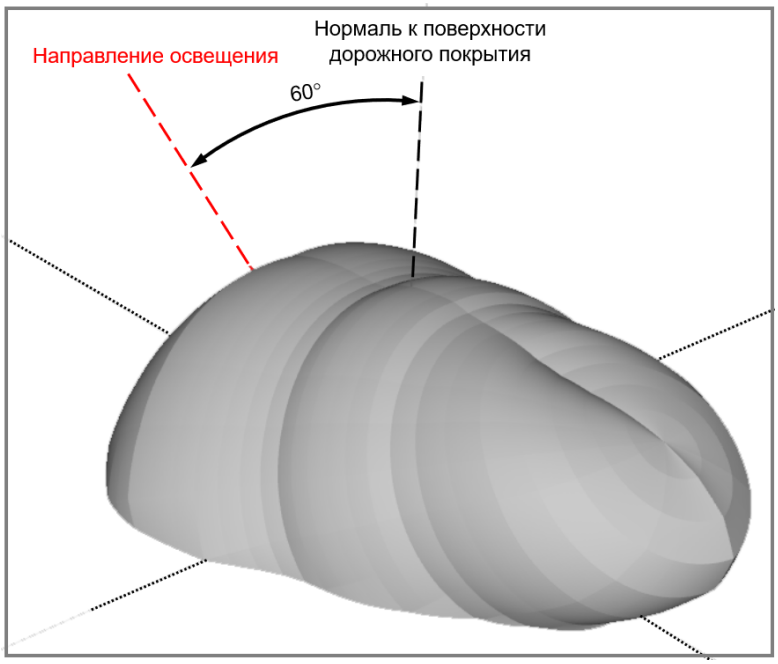


Рис. 1. Трехмерная визуализация ДФО асфальта

На рис. 1 представлены ДФО, измеренная для реального сухого дорожного асфальтового покрытия. Это трехмерная визуализация углового распределения коэффициента яркости (=ДФО) при наклонном параллельном освещении под углом 60° относительно нормали к поверхности дороги.

II. «r-таблица» - характеристика отражения МКО

«r-таблицы» согласно МКО (CIE 144–2001) [8] представляют ДФО как распределение коэффициента (показателя) яркости в табличном формате. Коэффициент яркости q элемента на поверхности проезжей части представляет собой отношение яркости L дороги в заданном направлении к освещённости дороги E. Единица измерения: cd * m^{-2} * lx^{-1}. Система координат для «r-таблиц» представлена на рис. 2.

Система координат «r-таблиц» определяет отражательную способность дорожного покрытия по трём углам: α, β, γ, где: α – угол между направлением наблюдения и горизонтальной плоскостью дороги; β – угол азимута, измеренный между двумя вертикальными плоскостями, содержащими направления освещения и наблюдения; γ — угол падения, измеренный от вертикали (нормали к поверхности дороги) к направлению освещения. Угол γ изменяется в пределах от 0° до близкого к 90°, а угол β от 0° до 180°. Предполагается, что отражательная способность дорожного покрытия имеет плоскую симметрию. Для движения автомобиля представляет интерес угол наблюдения α в диапазоне от 0.5° до 2°. Измерения традиционно проводились при α= 1° [9]. Итак, «r-таблица» – это ДФО, определяющая отражательную способность в ограниченном угловом пространстве с помощью двух углов β, γ. Формат «r-таблицы» представлен в таблице 1.

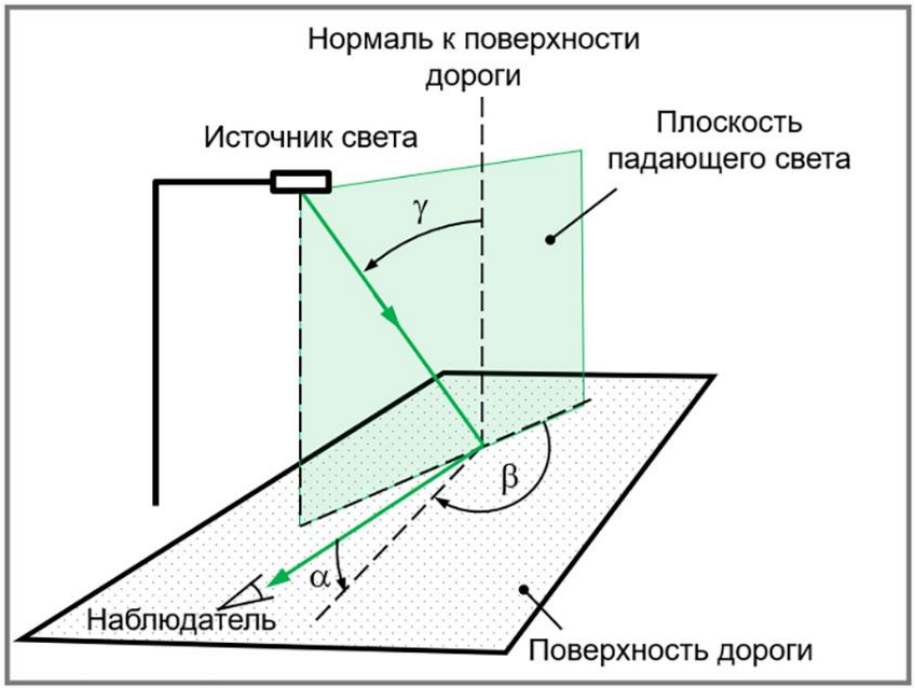


Рис. 2. Координатная система ДФО задаваемой «r-таблицей»

ТАБЛИЦА 1
ФОРМАТ «R-ТАБЛИЦЫ»

	β(град.)								
tg(γ)	**0**	**2**	**5**	**10**	**15**	**20**	**150**	**165**	**180**
0	770	770	770	770	770	770	770	770	770
0.25	710	708	703	710	712	710	724	719	723
0.5	586	582	587	581	581	576	587	581	589
0.75	468	467	465	455	457	446	437	438	445
1	378	372	373	363	347	331	318	323	329
1.25	308	304	305	285	270	244	237	238	245
1.5	258	254	251	229	203	178	177	179	184
1.75	217	214	205	182	153	129	134	137	138
…									
8	34	23	11	5	4	3	0	0	0
8.5	32	21	9	5	4	3	0	0	0
9	29	19	8	4	3	3	0	0	0
9.5	27	17	7	4	3	3	0	0	0
10	26	16	6	3	3	3	0	0	0
10.5	25	16	6	3	2	1	0	0	0
11	23	15	6	3	2	1	0	0	0
11.5	22	14	6	3	2	0	0	0	0
12	21	14	5	3	2	0	0	0	0

Таблица 1 представляет собой стандартный формат, который принят большинством компьютерных программ, предназначенных для проектирования дорожного освещения. В следующих главах статьи рассматриваются более общие задачи, связанные с визуализацией дорожных сцен. Числовые значения в таблице: r = q cos(γ) 10000.

III. Преобразование «r-таблиц» в ДФО

При использовании «r-таблиц» для визуализации дорожных сцен необходимо преобразовать их в более общий формат ДФО, поддерживаемый оптическими программами, предназначенными для визуализации (обладающими технологией рендеринга). В статье используется программный комплекс Lumicept [10], поэтому нам необходимо

преобразовать «r-таблицу» в формат ДФО, поддерживаемый в Lumicept. Система координат такого ДФО представлена на рис. 3. ДФО в Lumicept представлена в табличном формате набором двухмерных таблиц. Два угла ψ и σ определяют направление падающего света. Угол σ определяет наклон падающего света от нормали к образцу (в нашем случае дорожному покрытию). Это аналог угла γ в системе координат «r-таблицы», см. рис. 2. Азимутальным углом ψ, определяющим поворот направления освещения вокруг нормали к поверхности, можно пренебречь, поскольку мы считаем, что отражательная способность дорожного покрытия изотропна. Два угла φ и θ определяют направление наблюдения. Специфика системы координат ДФО в Lumicept заключается в отсчёте направления наблюдения от направления зеркального отражения. Угол θ – это угол между направлением зеркального отражения и направлением наблюдения, а угол φ – между двумя плоскостями: плоскостью падающего света и плоскостью, образованной двумя линиями – направлением зеркального отражения и направлением наблюдения.

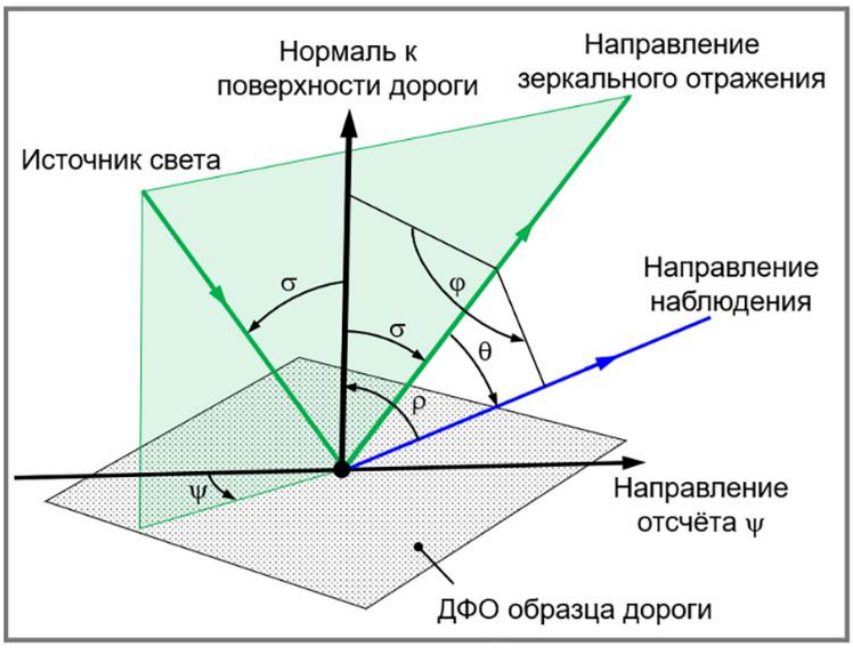


Рис. 3. Координатная система ДФО в Lumicept [10]

Система координат ДФО в Lumicept представляет собой типичную геоцентрическую систему координат, полюса которой проходят через направление зеркального отражения. Обратите внимание, что эта система координат очень удобна для ДФО, имеющего максимумы вдоль направления зеркального отражения, и не оптимальна для описания отражательных свойств дороги, так как здесь более важны направления наблюдения вблизи поверхности, но это частная техническая проблема, не влияющая на общие цели исследования, представленные в статье. Ещё одним отличием между форматами «r-таблиц» и ДФО в Lumicept являются единицы измерения. В «r-таблицах» задаётся коэффициент яркости, приведённый к освещённости, а ДФО в Lumicept представляет собой угловое распределение коэффициента яркости - отношение яркости, излучаемой образцом покрытия дороги к яркости идеального рассеивателя.

Для экспорта «r-таблиц» в ДФО поддерживаемую в Lumicept нам нужно выразить углы σ, φ, θ (см. рис. 3) через углы «r-таблицы» β, γ (см. рис. 2). Для лучшего понимания процесса преобразования см. рис. 4, где все углы связаны между собой.

Для удобства сформируем систему координат XYZ таким образом, чтобы ось Z соответствовала направлению зеркального отражения, а ось Y была перпендикулярна плоскости падения света. В этой системе координат орты вектора наблюдения (o_x, o_y, o_z) можно определить с помощью углов φ, θ по формулам (1, 2, 3):

$$o_x = -sin(\theta) \cdot cos\ (\varphi) \quad (1)$$

$$o_y = -sin(\theta) \cdot sin\ (\varphi) \quad (2)$$

$$o_z = -cos(\theta) \quad (3)$$

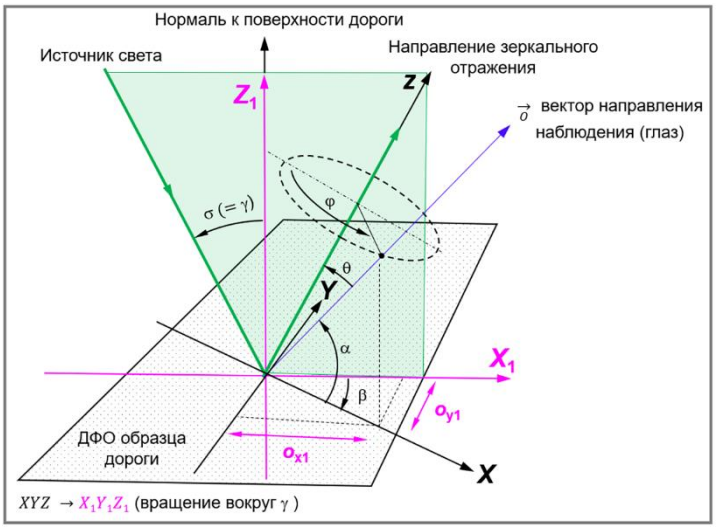


Рис. 4. Преобразование r-таблицы в ДФО Lumicept

Учитывая, что угол падения имеет одинаковое значение как в «r-таблицах», так и в ДФО Lumicept (σ = γ), орты вектора наблюдения (o_{x1}, o_{y1}, o_{z1}) в системе координат $X_1Y_1Z_1$, повёрнутого на угол σ вокруг начальной оси X, можно получить с помощью формулы (4, 5, 6):

$$o_{x1} = cos\,(\sigma)\cdot o_x + sin(\sigma)\cdot o_z \quad (4)$$

$$o_{y1} = o_y \quad (5)$$

$$o_{z1} = -sin(\sigma)\cdot o_x + cos(\sigma)\cdot o_z \quad (6)$$

Наконец, углы α, β могут быть выражены формулами (7, 8):

$$\beta = atan2\left(\frac{o_{y1}}{o_{x1}}\right) \quad (7)$$

$$\alpha = acos\left(\frac{\sqrt{o_{x1}^2+o_{y1}^2}}{\sqrt{o_{x1}^2+o_{y1}^2+o_{z1}^2}}\right) \quad (8)$$

«r-таблица» определена только для одного α = 1°, поэтому нет полных данных для определения ДФО на всем диапазоне α от 0° до 90°. Однако задание ДФО на основе одного и того же значения α для всего диапазона от 0° до 90° приведёт к превышению общего коэффициента отражения выше допустимого значения 100%. Чтобы избежать этой проблемы, предлагается компромиссное решение, а именно, вычислять ДФО только для ограниченного диапазона α вблизи горизонта дорожного покрытия, около 10° - 15°, а для остальной части диапазона до 90° отражение считается идеальными Ламбертовким [11], иными словами коэффициент яркости ДФО будет постоянным, взятым от одного направления с углами: γ = 0° и β = 0° (см. рис. 5).

На рис. 5 показано, как рассчитывается ДФО в зависимости от направления наблюдения внутри или снаружи конуса определяемого углом «среза», угол конуса = 90° – угол «среза»: по «r-таблице», если направление внутри конуса, внутри конуса ДФО определяется постоянным коэффициентом яркости, взятым от направления наблюдения для углов β = γ = 0°. В статье использован угол «среза» = 10°.

Следующим шагом в спецификации ДФО является преобразование значений r в коэффициент яркости (LF), используемый в формате Lumicept ДФО и определяемый по формуле (9):

$$LF = \frac{L_r}{L_e} \quad (9)$$

где L_r — яркость образца (дорожного покрытия), а L_e — яркость идеального (ламбертовского) рассеивателя («эталона»). Спецификация МКО определяет коэффициент яркости q по формуле (10):

$$q = \frac{L_r}{E_e} \quad (10)$$

где L_r — яркость образца (дорожного покрытия) и E_e освещённость. Значения в «r-таблицах связаны» с q как r = q · $\cos^3(\gamma)$ 10000, таким образом, мы можем определить формулы преобразования по формуле (11):

$$q = \frac{L_r}{E_r} = \frac{L_r}{\pi \cdot L_r} = \frac{LF}{\pi} \Rightarrow LF = \frac{\pi \cdot r}{cos^3(\gamma) \cdot 10000} \quad (11)$$

Итоговый подход к преобразованию «r-таблиц» в ДФО Lumicept представлен на рис. 6.

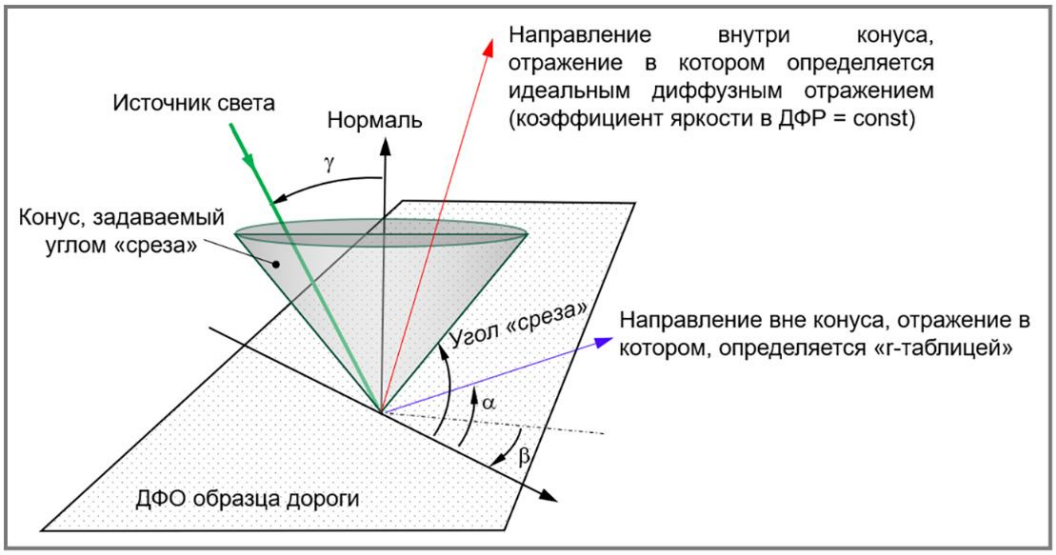


Рис. 5 Угол «среза» определяющий метод ДФР определения

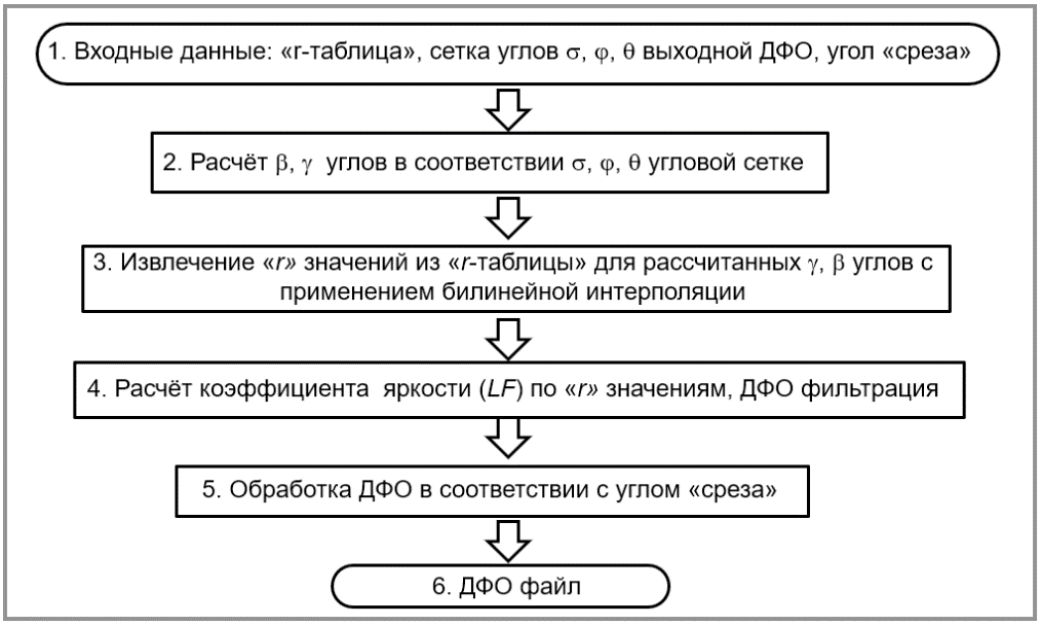


Рис.6 Схема преобразования «r-таблицы» в ДФО

Конвертер для расчёта ДФО реализован с помощью скрипта Python. Конверсия начинается с ввода данных: загрузки «r-таблицы», ввода сетки углов σ, φ, θ выходной ДФО и угла «среза». Затем на втором этапе углы σ, φ, θ преобразуются в γ и β с помощью формул (1–8). Следующим шагом является извлечение значений *r* из «r-таблицы» для рассчитанных γ, β посредством билинейной интерполяции [12]. Четвёртый шаг в алгоритме — расчёт коэффициента яркости, используемого в формате ДФО Lumicept, из значений r по формуле (11), далее ДФО фильтруется, что будет объяснено в следующей главе. Затем, на пятом этапе, распределение коэффициента яркости обрабатывается в соответствии с углом «среза», значения ДФО для направлений, находящихся внутри конуса (см. рис. 5), заменяются на постоянное значение коэффициента яркости. И последний шаг схемы — сохранение файла с ДФО в соответствии с форматом Lumicept [10].

IV. Визуализация дорожной сцены

Для проверки полученных ДФО была разработана простая сцена, которая содержит набор плоских поверхностей, имитирующих дорожное покрытие с разметкой (см. рис 7). Дорожное покрытие освещается точечным источником света. Размеры всех элементов (дорога, высота источников света над дорогой), индикатриса излучения источников света также приближены к реалистичным дорожным сценам. Основное упрощение сцены — отсутствие фоновых объектов, например стен туннеля. Таким образом, отсутствуют переотражения света, и можно использовать простые средства визуализации с высокой скоростью вычислений, например, обратную трассировку лучей [13–15]. Условия наблюдения (положение камеры) также задаются согласно реальным условиям, которые определены для реального водителя.

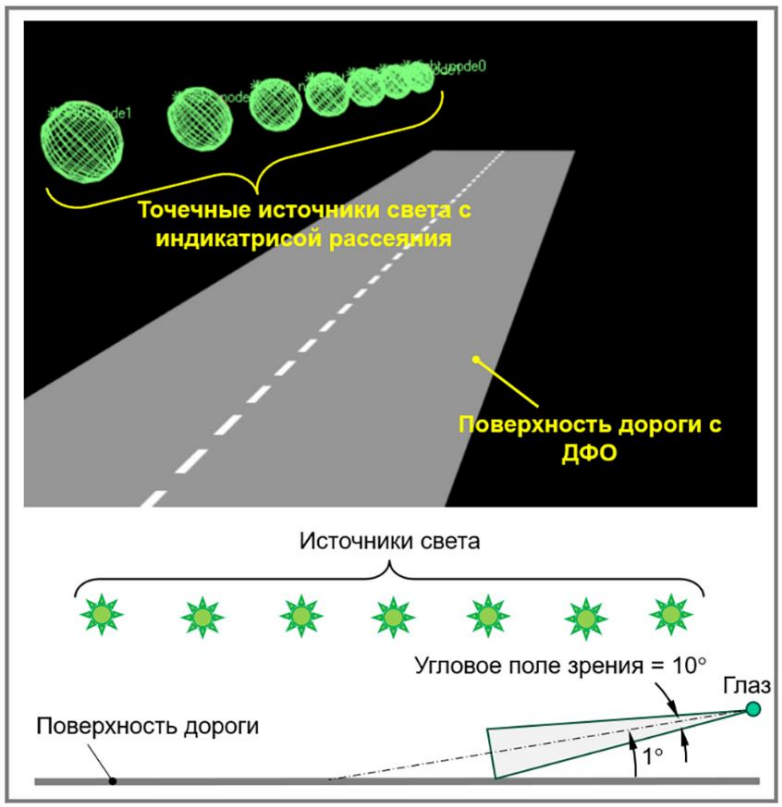


Рис.7 Упрощённая модель дороги с освещением

Следует отметить, что ДФО определяется в табличной форме, это означает, что коэффициент яркости отражённого света указан для дискретной сетки для углов σ, φ, θ (см. рис. 3). Следовательно. точность ДФО будет зависеть от разрешения угловой сетки, шагов между соседними значениями для этих углов. Для нахождения оптимального разрешения сетки генерируется набор изображений с разными шагами для углов σ, φ, θ (см. рис 8). Шаги для этих углов обозначаются как Δσ, Δφ, Δθ. Для первоначального исследования была выбрана «Table C1», см. приложение к CIE 144:2001 [8].

Анализируя рис. 8, видно, что изображения, созданные с грубой сеткой для углов σ, φ, θ с шагом = 10°, 5° (изображения А и Б), имеют явные дефекты (рваные линии), тогда как изображения В и Д, полученные для ДФО, с шагом = 2°, 1° имеют лучшее качество без заметных артефактов. Эксперименты показывают, что дальнейшее увеличение углового разрешения не приводит к заметной разнице в качестве изображения. Примечание. ДФО с подробной детальной угловой сеткой имеет довольно большой размер файла, например, файл ДФО для углового шага = 1° имеет размер более 20 МБ. Чтобы решить проблему с большими данными, можно применить фильтрацию к ДФО, сгенерированной с грубым угловым шагом. Это исследование было проведено с применением фильтрации Гаусса, которая была запущена для ДФО, с угловым шагом для всех углов σ, φ, θ = 5°. Готовое решение было взято из библиотеки SciPy [16]. Изображения на рис. 9 представляют результаты фильтрации.

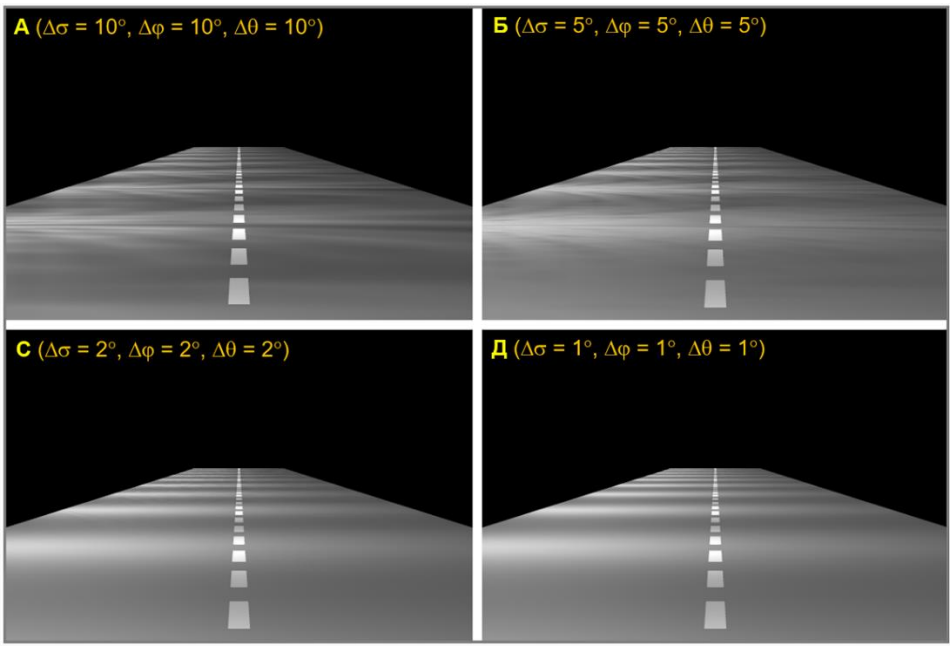


Рис.8 Визуализация дороги с ДФО, рассчитанной с разным разрешением угловой сетки

Рис.9 Влияние фильтрации ДФО на качество изображения

Изображение А на рис. 9 синтезировано для нефильтрованной ДФО. Б и С представляют аналогичные изображения для дорожного покрытия ДФО с Гаусс-фильтрацией для разных значений «sigma». Параметр «sigma» представляет собой стандартное отклонение для ядра Гаусса [16]. Чем больше это значение, тем выше уровень фильтрации (сглаживания). Как видно из изображений Б и С на рис. 9 фильтрация устраняет артефакты на изображениях. Таким образом, фильтрацию можно применять, если важен размер ДФО, например, при работе с огромной базой данных. Обратите внимание, что в следующем исследовании фильтрация не будет применяться для сохранения чистоты экспериментов.

Следующий пример дорожной сцены, представленный на рис. 10, является более реалистичным.

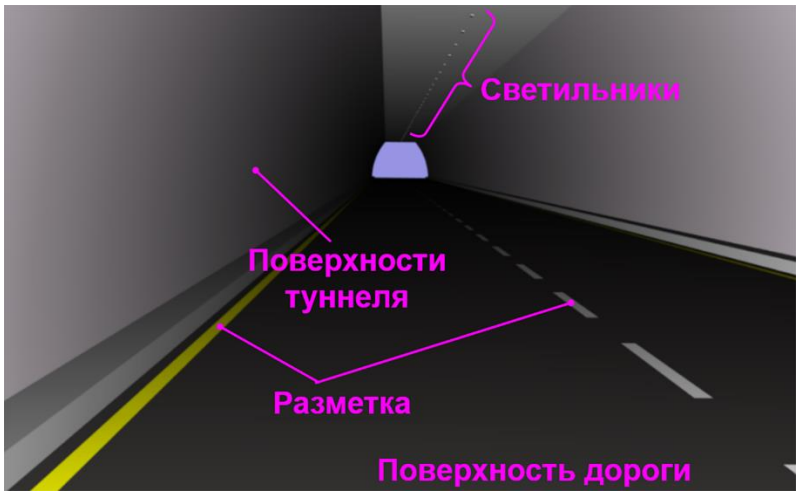


Рис.10 Пример дорожной сцены с туннелем

В модели дорожной сцены, представленной на рис. 10, представлены дополнительные поверхности, имитирующие туннель, кроме того, эта модель содержит более сложные и реалистичные светильники с пространственным излучением (не точечными источниками света как в предыдущем упрощённом примере). Оптические свойства поверхностей туннеля определены измеренными ДФО В такой сцене возникают переотражения света, что требует применения более продвинутых методов расчёта при синтезе реалистичных изображений. В этом исследовании был использован усовершенствованный гибридный трассировщик лучей, так называемый «Path Tracing» [17-20], доступный в Lumicept [10], который позволяет физически корректно моделировать такие сцены. На рис. 11 представлены изображения, сгенерированные для ДФО, рассчитанных для двух стандартных «r-таблиц», представленных в CIE 144–2001: «Table C1» и «Table W1» приложения [8].

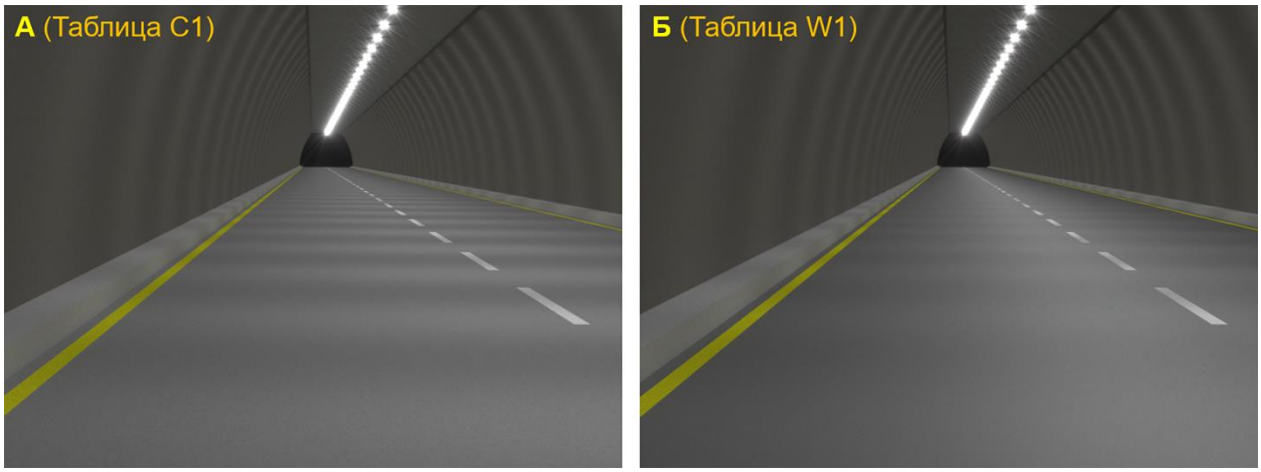


Рис.11 Визуализация дорожной сцены для «r-таблиц»: А – «Table C1» и Б – «Table W1»

Изображения на рисунке 11 показывают визуализацию более реалистичной сцены по сравнению с простым примером, представленным на рис. 7–9. Очевидно, что в этой модели имеются переотражения света между поверхностью дороги и другими геометрическими элементами модели под направлениями, которые не описываются данными «r-таблиц» и ДФО рассчитанными этих таблиц. Аналогичная проблема возникнет и для прямого отражённого света, если увеличить поле зрения (FOV). При этом интересно оценить возможную разницу изображений, полученных для полного ДФО дорожного покрытия, измеренного для всего углового пространства, по сравнению с ДФО, полученного из «r-таблиц», где измерения выполняются только вблизи горизонта (α = 0.5°–2°, см. рис. 2). Эти исследования представлены в следующей главе.

V. Применимость ДФО, полученных на основе «r-таблиц»

Для проверки корректности использования ДФО дорожного покрытия, полученного на основе «r-таблиц» для задач визуализации, сравним изображения, созданные для полного измеренного ДФО реального асфальтного покрытия, с ДФО на основе «r-таблицы», извлечённой из этого полного измеренного ДФО. Схематически этот подход показан на рис. 12.

tg(γ) \ β(°)	0	2	5	10	15	20	25	30	35	40	4
0	614	614	614	614	614	614	614	614	614	614	6
0.25	524	524	524	518	509	509	503	501	501	498	4
0.5	425	425	425	424	421	426	426	419	427	427	4
0.75	302	302	302	304	304	305	307	307	309	311	3
1	203	203	204	205	205	209	209	215	212	216	2
1.25	156	156	158	149	172	152	159	153	152	159	1
1.5	130	130	131	118	140	114	130	115	115	119	1

Рис.12 Схема проверки ДФО, сгенерированной на основе «r-таблицы»

Идея этой проверки состоит в том, чтобы извлечь «r-таблицу» из полного измеренного ДФО дорожного покрытия, а затем с помощью преобразования, описанного ранее, получить из этой «r-таблицы» ограниченный ДФО. Извлечение «r-таблицы» из полного ДФО осуществляется с помощью тех же формул (1–8), которые задают связь между сеткой r-таблицы для углов γ, α, β и сеткой для углов ДФО (σ, φ , θ). Проблема заключается в том, какой подход использовать для нахождения более точных значений углов по этим сеткам. В области такого

преобразования было проведено несколько различных экспериментов, включая линейную 2D- и 3D-интерполяцию. Оптимальные результаты были получены при простом поиске ближайших значений между обеими угловыми сетками. Для этой цели было использовано готовое решение — функция KDTree из библиотеки SciPy [21]. Этот подход дал минимальную погрешность, которая оценивалась при преобразовании из стандартных «r-таблиц» в ограниченные ДФО и обратно в «r-таблицы». Значения «*r*» можно получить из распределения углового коэффициента яркости, представленного в ДФО, по формуле (12), которая следует из формулы (11).

$$r = \frac{\pi \cdot LF \cdot cos^3(\gamma) \cdot 10000}{\pi \cdot} \tag{11}$$

После преобразования полного измеренного ДФО асфальтовой поверхности в ограниченный ДФО на основе «r-таблицы» можно визуализировать дорожную сцену, представленную рис. 10 для обеих этих ДФО и сравнить их. Далее этот ограниченный ДФО называется просто «r-таблицей». Результаты этой визуализации представлены на рис. 13, где представлены изображения рендеринга для различных полей зрения (FOV) от 5° до 30°. Наблюдатель размещается на высоте 1.5 м от поверхности дороги, точка (цель) наблюдения размещается на расстоянии 100 м от наблюдателя на поверхности дороги. Другими словами, заданы вполне реалистичные условия наблюдения. Левая часть на рис. 13 представляет изображение, синтезированное для полного измеренного ДФО, а правая часть представляет изображение, синтезированное для ДФО, полученное из «r-таблицы». Левая и правая части изображения разделены пурпурной линией для лучшего понимания.

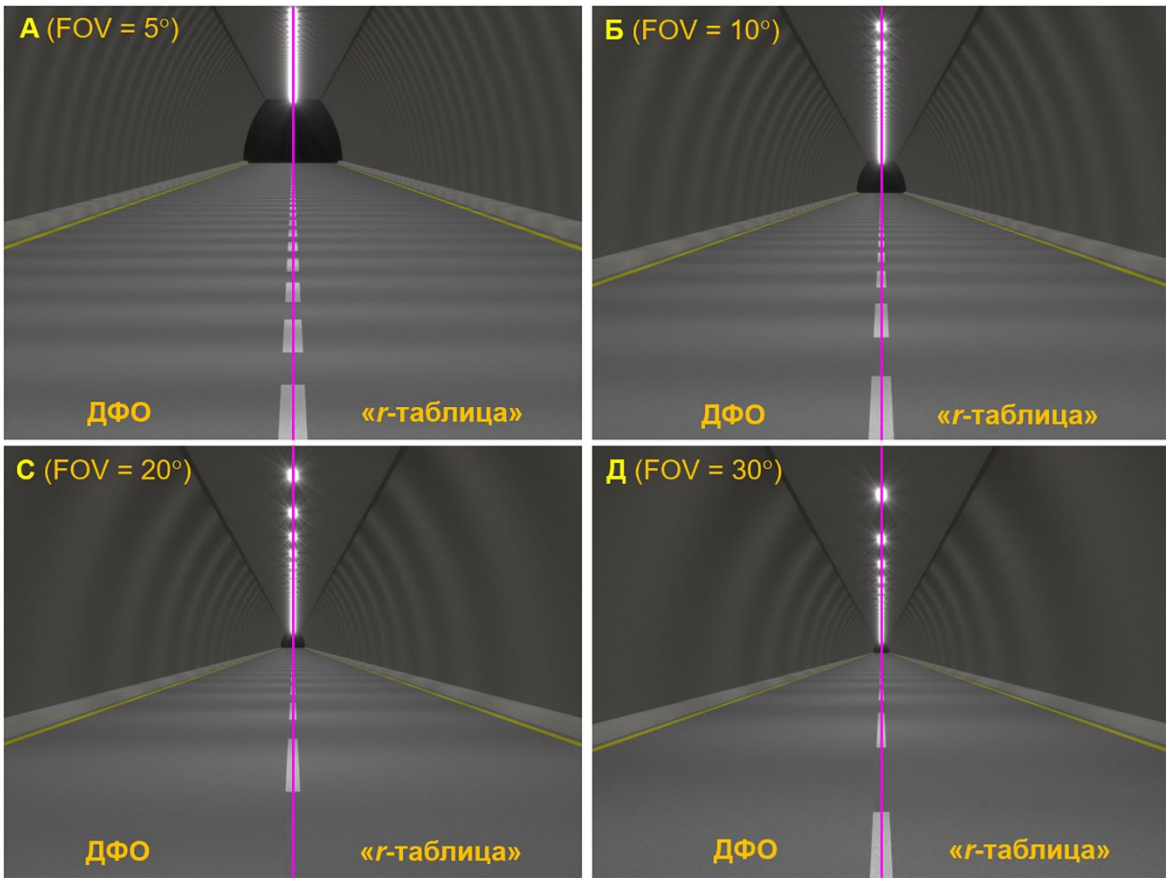


Рис.13. Сравнение изображений, созданных для реального измеренного ДФО с ДФО, полученным из r-таблицы. Изображения А, Б, С, Д различаются полем зрения (FOV)

Из сравнения изображений рис. 13 видно, что они практически неотличимы. Чтобы увидеть разницу, они были преобразованы в искусственные, см. рис. 14., где представлено распределения яркости в искусственных цветах. Красный цвет показывает максимальную яркость, а синий – нулевую яркость. Эти изображения (первый ряд) подобны изображениям на рис. 13, где используется наложение изображений. В левой половине представлено изображение, полученное на основе полного измеренного ДФО, а в правой – на основе «r-таблицы». Обратите внимание, мы не можем даже увидеть центральную границу между ними. Второй нижний ряд на рис. 14 представляет собой результат вычитания яркости между двумя изображениями. Цвет цветовой

шкалы такой же, как и для изображений верхнего ряда. Обратите внимание, что изображения разности яркостей практически синие (разница в яркости = 0), лишь незначительная разница наблюдается в зоне дорожной разметки и в дальней зоне дороги.

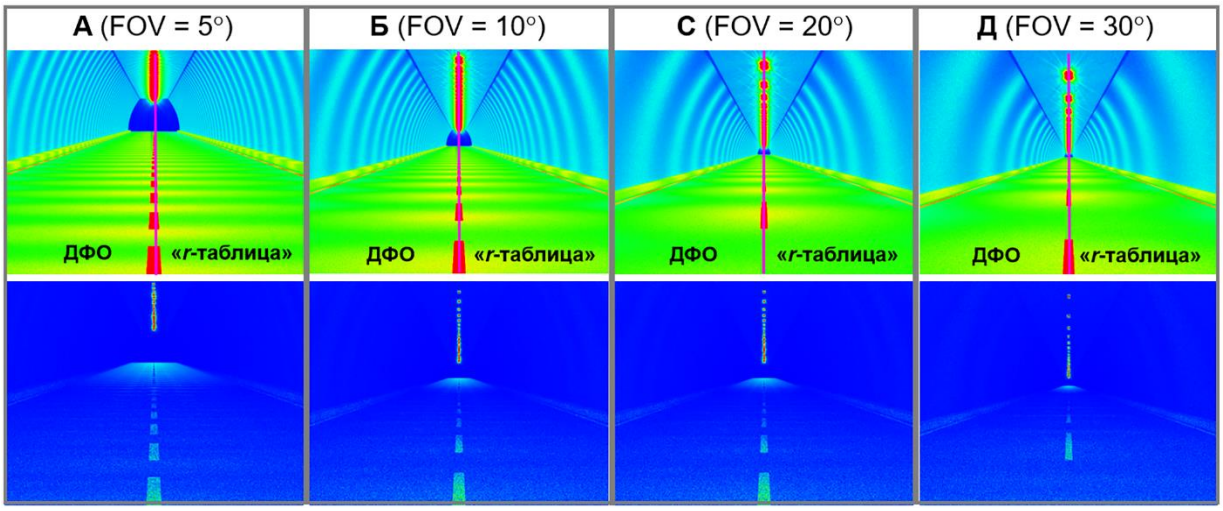


Рис.14. Сравнение изображений в искусственных цветах, созданных для реального измеренного ДФО с ДФО, полученным из r-таблицы. Изображения А, Б, С, Д различаются полем зрения (FOV)

Численные различия в распределении яркости для разных полей зрения, рассчитанные для полного измеренного ДФО в сравнении с ДФО, извлечённым из «r-таблицы», представлены в таблице 2 в форме относительного среднеквадратического отклонения (RMS).

ТАБЛИЦА 2
СРЕДНЕКВАДРАТИЧНОЕ ОТКЛОНЕНИЕ (RMS) МУЖДУ ИЗМЕРЕННЫМ ДФО И ДФО «R-ТАБЛИЦЫ»

FOV	5°	10°	20°	30°
RMS	1.9%	2.0%	2.6%	3.0%

Относительное среднеквадратичное отклонение рассчитывали по формуле (12):

$$RMS_{relative} = \sqrt{\frac{\sum_i^n \left(\frac{L_{1i}-L_{2i}}{0.5\cdot(L_{1i}+L_{2i})}\right)^2}{n}} \cdot 100\% \qquad (12)$$

Где L_{1i} – яркость, рассчитанная для реального измеренного ДФО для i пикселя изображения, L_{2i} – яркость, рассчитанная для ДФО «r-таблицы» для i пикселя изображения, n – сгенерированное количество точек (пикселей) в распределении яркости (общее количество пикселей в изображении).

Как видно из таблицы 2, RMS незначительно увеличивается с увеличением поля зрения (FOV), однако даже для максимального поля зрения = 30° RMS не превышает 5%.

VI. Выводы и заключение

Рассмотрена возможность применения «r-таблиц» МКО для визуализации дорожных сцен. Исследования проводились на основе программного пакета коммерческого освещения Lumicept [9]. Данные «r-таблицы» были успешно преобразованы в широко распространённый формат ДФО и поддерживаемый в Lumicept, что позволило синтезировать изображения для нескольких моделей дорожных сцен. Проверка проводилась путём сравнения реального измеренного ДФО дорожного покрытия с ограниченным ДФО, извлечённым из «r-таблицы», полученной из измеренного ДФО. Исследования статьи показали, что, несмотря на предупреждения МКО о возможной необходимости расширения «r-таблиц» дополнительными измерениями в более широком угловом пространстве для использования их в задачах визуализации, эти данные могут быть применены к приложениям, связанным с визуализацией дорожных сцен, по крайней мере, при условиях наблюдения, близких к реальным, когда взгляд водителя направлен на дорогу. Соответствие изображений, полученных на основе «r-таблиц» против изображений, полученных на основе полных измеренных ДФО практически идеальное. В будущем будет полезно

расширить эти исследования, включив в них различные свойства асфальта, включая мокрые поверхности, а также различные оптические свойства поверхностей туннеля и более широкие условия наблюдения.

Источник финансирования. Благодарности

Работа выполнена при поддержке Российского научного фонда (Разработка методов, алгоритмов и вычислительной архитектуры реконструкции геометрии реального окружения для построения системы смешанной реальности, свободной от дискомфорта зрительного восприятия, №. 22-11-00145)

Список литературы

(исключаются учебники, учебно-методические пособия, рукописи (диссертации, авторефераты))

1. Kim D., Park K., Kim K., Seo M., H.Lee (2009). High-dynamic-range camera-based bidirectional reflectance distribution function measurement system for isotropic materials. Optical Engineering - OPT ENG. 48. 10.1117/1.3223247.
2. Torrance, K.; Sparrow, E. (1967). "Theory for Off-Specular Reflection from Roughened Surfaces". Journal of the Optical Society of America. 57 (9): 1105–1114. doi:10.1364/JOSA.57.001105
3. Rusinkiewicz. S. "A Survey of BRDF Representation for Computer Graphics". Retrieved 2007-09-05.
4. Marschner S. R., Westin S.H, Lafortune E.P.F, Torrance K.E., Greenberg D.P. (1999) Image-Based BRDF Measurement Including Human Skin. In: Lischinski D., Larson G.W. (eds) Rendering Techniques’ 99. Eurographics. Springer, Vienna
5. Torrance K. and Sparrow E. Theory for Off-Specular Reflection from Roughened Surfaces. J. Optical Soc. America, vol. 57., pp. 1105—1114, 1976.
6. Ward, J. Gregory “Measuring and modeling anisotropic reflection”. Proceedings of SIGGRAPH. (1992), pp. 265—272. doi:10.1145/133994.134078.
7. James F. Blinn. Models of light reflection for computer synthesized pictures // Proc. 4th annual conference on computer graphics and interactive techniques: journal. 1977. pp. 192. doi:10.1145/563858.563893
8. CIE 144:2001. Road Surface and marking reflection characteristics, ISBN: 978 3 901906 12 1, URL: https://cie.co.at/publications/road-surface-and-road-marking-reflection-characteristics
9. 21. Muzet V, Bernasconi J, Iacomussi P, et al. Review of road surface photometry methods and devices – Proposal for new measurement geometries. Lighting Research & Technology. 2021;53(3):213-229. doi:10.1177/1477153520958454
10. Lumicept – Hybrid Light Simulation Software, 2024. URL: https://integra.jp/en/products/lumicept.
11. Ikeuchi, Katsushi, "Lambertian Reflectance". Encyclopedia of Computer Vision. Springer. pp. 441–443. doi:10.1007/978-0-387-31439-6_534. ISBN 978-0-387-30771-8, (2014).
12. Bi-linear interpolation, Python, SciPy library v. 1.12.0 Manual, scipy.interpolate.griddata, URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.griddata.html.
13. Arvo J., “Backward ray tracing,” Dev. Ray Tracing (1986).
14. Guan Y., Sang X, Xing S., Li Y., Chen Y., Chen D., Yang L., and Yan B., "Backward ray tracing based high-speed visual simulation for light field display and experimental verification," Opt. Express 27, 29309-29318 (2019).
15. Anderson D. C. and Cychosz J., “An introduction to ray tracing,” Image Vis. Comput. 8(2), 169 (2003)
16. Gaussian filter, Python, SciPy library v. 1.12.0 Manual, scipy.ndimage.gaussian_filter, URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.gaussian_filter.html.
17. Lafortune E. P., Willems Y. D., “Bi-directional path tracing”, Proceedings of Third International Conference on Computational Graphics and Visualization Techniques (Compugraphics '93), p.145-153, Alvor, Portugal, December 1993
18. Jensen H. W. Global illumination using photon maps // Eurographics workshop on Rendering techniques. – Springer, Vienna, 1996. – Pp. 21-30.
19. He H., Wang T., Xu Q., Xing Y. Multi-core parallel of photon mapping //Visual Information Communication. – Springer, Boston, MA, 2009. – Pp. 365-374.
20. Zhdanov A., Zhdanov D., Biryukov E., The realistic rendering with forward and backward photon mapping, Preprints of the Keldysh Institute of Applied Mathematics, 2020, 077, 22 pp., DOI: https://doi.org/10.20948/prepr-2020-77 .
21. KDTree, Python, SciPy library v. 1.12.0 Manual, scipy.spatial.KDTree, URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-67-83

Отрисовка с антиалиасингом с помощью булевых операций над многоугольниками
Antialiased Rendering Using Boolean Operations on Polygons

К. Л. Богомолов, В. А. Фролов
Институт прикладной математики им. М.В. Келдыша РАН, Москва, Россия

K. L. Bogomolov, V. A. Frolov
Keldysh Institute of Applied Mathematics of RAS, Moscow, Russia

Аннотация. Устранение ступенчатости (антиалиасинг) путем повышения частоты сэмплирования ограничено в точности: всегда можно найти сцену, содержащую в некоторых пикселях намного больше треугольников, чем сэмплируемых точек. Мы предлагаем алгоритм устранения ступенчатости для полигональных сцен, вычисляющий интеграл функции изображения по каждому пикселю, и учитывающий все видимые треугольники, попадающие в пиксель. Алгоритм строит видимую на экране сцену в векторном формате с помощью хранилища многоугольников, и затем растеризует эти многоугольники, смешивая цвета в каждом пикселе согласно площади пересечения многоугольника и пикселя. Реализация с использованием целочисленной арифметики обеспечивает корректную работу и отсутствие накопления погрешностей даже на сложных сценах с очень мелкими треугольниками. Определение видимой геометрии сцены основывается на использовании булева вычитания многоугольников. Данная работа демонстрирует принципиальную возможность реализации отрисовки с указанными характеристиками.

***Ключевые слова*:** антиалиасинг, определение видимости, пересечение многоугольников, вычитание многоугольников.

Abstract. Antialiasing by increasing the frequency of sampling is limited in precision: it is always possible to find a scene which contains many more triangles than multi-sampling points in some of the pixels. We propose an antialiasing algorithm for polygonal scenes that computes the integral of the picture function over each pixel and considers all visible triangles present in the pixel. The algorithm constructs the scene visible on the screen in a vectorized format using a storage of polygons, and then rasterizes these polygons by mixing colors at each pixel according to the area of intersection of polygon and pixel. An implementation that uses integer arithmetic ensures correct execution and absence of error accumulation even on complex scenes with very small triangles. The determination of visible geometry in a scene is based on using Boolean subtraction of polygons. The present work shows that it is in principle possible to implement rendering with specified properties.

***Keywords*:** antialiasing, visibility, polygon intersection, polygon subtraction.

I. Введение

Алиасинг является хорошо известным нежелательным эффектом при отрисовке трехмерных и двумерных изображений в двумерный растр. Алиасинг может возникать как в отрисовке в реальном времени, так и не в реальном времени. Устранение алиасинга в отрисовке в реальном времени является, разумеется, более сложной проблемой; это объясняется двумя причинами. Первая причина очевидна: в реальном времени нужно успевать отрисовать каждый кадр за долю секунды. Вторая причина заключается в том, что в реальном времени хочется дать пользователю возможность максимально свободно управлять камерой. Например, он может рассмотреть вблизи кустик травы, состоящий из тысяч треугольников, а потом плавно переместить камеру на общий вид зеленого луга; желательно, чтобы при этом не возникал алиасинг и не происходило некрасивого переключения уровней детализации моделей. При отрисовке не в реальном времени можно просто избегать таких сложных ситуаций.

Прогресс аппаратуры для отрисовки в реальном времени сталкивается с необходимостью повышения качества антиалиасинга. Например, если мы отрисовываем кадр из 8 миллионов пикселей, и используем мультисэмплинг с 8 точками на пиксель, то нет смысла поддерживать сцены с количеством треугольников, намного превышающим 64 миллиона, так как на результирующее изображение сцены могут повлиять не более 64 миллионов треугольников. Кроме того, использование фиксированного количества точек мультисэмплинга в пикселе является существенным ограничением, так как можно придумать сцены, в которых во многие пиксели

попадает намного больше треугольников, как в упомянутом выше случае с не упрощаемой травой. Желательно, чтобы алгоритм антиалиасинга учитывал все треугольники, попадающие в каждый пиксель.

Изложим рассматриваемую проблему более формально. Алиасинг появляется из-за того, что функция изображения, зависящая от вещественных координат на экране монитора, переводится в растр, состоящий из конечного числа пикселей, что и порождает ошибку дискретизации. Для того, чтобы уменьшить эту ошибку, не увеличивая разрешения изображения, необходимо осуществлять дискретизацию этой функции специальным образом – нужно вычислить интеграл функции по площади пикселя (формула 1).

$$I_{ij} = \int_{i \le x \le i+1,\ j \le y \le j+1} P(x,y)\ dx\,dy \qquad (1)$$

Ситуация осложняется тем, что функция изображения может содержать разрывы на краях объектов, элементы, имеющие размер намного меньше 1 пикселя растра, и, возможно, зоны с регулярно повторяющимися элементами, например досками забора. Поэтому ключевой вопрос антиалиасинга состоит в том, как вычислить этот интеграл точнее и за приемлемое время. Зачастую (что далеко не лучший способ) это делается при помощи метода Монте-Карло – усреднения некоторого числа точечных выборок (сэмплов) внутри пикселя. При этом процесс генерации выборок обычно называется сэмплированием. Самый простой способ вычисления интеграла — сэмплирование в центрах пикселей. Этот способ, очевидно, приводит к алиасингу: края объектов обрываются ступенчатой линией, маленькие элементы могут попасть или не попасть в центр пикселя, что приводит к мельтешению, а регулярно повторяющиеся элементы в результате наложения на регулярную сетку сэмплирования могут привести к муару.

Один из практичных и очевидных способов антиалиасинга – повышение частоты сэмплирования. Super-Sampling Anti-Aliasing (SSAA) сэмплирует сцену в нескольких точках каждого результирующего пикселя и усредняет полученные цвета. Multi-Sampling Anti-Aliasing (MSAA) является оптимизацией SSAA, в которой цвет каждого сэмплируемого треугольника в пикселе вычисляется только один раз, вне зависимости от того, сколько сэмплов на него попадает. Эти методы легко подходят для аппаратной реализации и приводят к хорошим результатам в современных приложениях. SSAA, хотя и дороже MSAA, помогает улучшить ситуацию с алиасингом шейдерных эффектов, таких, как блик. Однако, как мы уже упоминали, сколько бы сэмплов мы ни брали в каждом пикселе, всегда можно придумать сцену, содержащую пиксели с бóльшим количеством геометрических элементов. Рост числа треугольников в сцене не будет приводить к повышению качества изображения в этом случае.

Известен и другой подход к антиалиасингу – это замена сэмплирования сцены в точках на свертку с функцией фильтра:

$$W(x,y) = \int_{\Omega} P(s,t)\ F(s-x,\ t-y)\,ds\,dt. \qquad (2)$$

Здесь W, P и F – это, соответственно, отфильтрованный результат, исходное изображение и фильтр. Интегрирование ведется по области Ω, в которой сдвинутая функция фильтра не равна 0. Результирующая функция W, вычисленная в центре каждого пикселя, записывается в растр. Самым очевидным фильтром является «box» фильтр: 1 внутри квадрата пикселя и 0 вне его; в этом случае мы просто вычисляем в каждом пикселе интеграл (1). Возможны и более сложные фильтры, приводящие к сглаживанию изображения в диапазоне нескольких пикселей, например, гауссова функция, срезаемая в 0 на небольшом расстоянии от начала координат. Эти фильтры приводят к субъективному улучшению или ухудшению качества изображения: например, они могут сгладить высококонтрастные повторяющиеся полоски, что предотвратит появление муара. Как замечено в работе [1], разница в эффекте разных фильтров будет скорее количественной, чем качественной. При повышении разрешения результаты работы фильтров станут схожими, в то время как сэмплирование без антиалиасинга будет продолжать создавать артефакты в результирующем изображении.

Мы предлагаем алгоритм антиалиасинга, вычисляющий интеграл (1). Алгоритм применяется к сценам из непрозрачных треугольников; поддержка полупрозрачных треугольников может быть добавлена относительно несложно. Алгоритм состоит из двух основных стадий: построение векторного хранилища всех видимых на экране многоугольников (вместе с их исходными треугольниками) и растеризация этих многоугольников со смешиванием цветов в каждом пикселе согласно площади пересечения многоугольника и пикселя. Вычисление цвета каждого треугольника в пикселе производится шейдером, который предоставляется пользователем. Сложность этого шейдера может варьироваться; на момент написания там реализована интерполяция для Гуро-шейдинга. Теоретически, в этом шейдере можно реализовать, например, трассировку лучей, что могло бы заодно обеспечить поддержку полупрозрачных и зеркальных поверхностей.

Программа реализована на центральном процессоре; значительное внимание было уделено использованию эффективных структур данных, дающих возможность увеличения сложности сцен с сохранением разумного

времени выполнения. Производительность сейчас далека от реального времени: отрисовка 1 миллиона треугольников занимает порядка 10 минут. Но мы надеемся, что изложенные идеи могут оказаться полезными для будущего прогресса аппаратуры, а также для отрисовки не в реальном времени.

Мы считаем, что существенными достижениями данной работы являются: (1) алгоритм антиалиасинга, учитывающий все треугольники, попадающие в каждый пиксель, и реализующий box фильтр; (2) использование целочисленной арифметики, обеспечивающей отсутствие накопления погрешности и стабильную работу на сколь угодно сложных сценах; (3) реализация алгоритмов булевых операций над многоугольниками и структуры данных BoxQuadtree, которые могут оказаться полезными и в других приложениях.

Статья организована следующим образом. В разделе II представлен обзор близкой по теме литературы. В начале раздела III приведены две стадии основного алгоритма, которые изложены подробно в подразделах III.1 и III.2. В подразделах раздела IV описаны в общих чертах вспомогательные алгоритмы, вызываемые из основного: булевы операции над многоугольниками, точная рациональная арифметика и хранилище BoxQuadtree. Результаты отрисовки с разными типами антиалиасинга на нескольких сценах разной сложности приведены в разделе V. Заключение располагается в разделе VI.

II. Обзор литературы

По-видимому, первый алгоритм антиалиасинга с вычислением интеграла (1) предложил Edwin Catmull в работе [1]. В качестве приложения представлено использование в двумерной анимации.

В публикациях [2], [3], [4] предложен подход, названный Non-Sampled Anti-Aliasing (NSAA). Это вычисление свертки спроецированной сцены с радиально симметричной функцией фильтра. Функция задается в виде аппроксимирующего многочлена. Интеграл выражен аналитически с помощью трансцендентных функций, и вычисляется точно в пределах возможностей стандартной машинной системы чисел. Авторы сообщают о реализации со впечатляющими скоростными характеристиками благодаря параллелизму на подходящих графических процессорах.

NSAA, так же как [1] и наш подход, вычисляет интеграл (2) с некоторой функцией фильтра без сэмплирования. Данный подход опирается на радиально симметричные фильтр-функции, что является существенным ограничением. Например, используемый фильтр не может быть box функцией. Используются сглаживающие по нескольким соседним пикселям фильтры, например, гауссова функция, отсекаемая в 0 на расстоянии 2.3 пикселя. Это приводит к легкому размытию изображения; это размытие может хорошо смотреться вблизи (рисунок 11); также оно устраняет контрастные повторяющиеся элементы, что предотвращает появление муара в растровом результате. Но вдалеке сглаживание сцены по нескольким пикселям приводит к ощущению, что мы смотрим через расфокусированный оптический прибор (рисунок 12). Мы полагаем, что box фильтр является во многих случаях предпочтительным благодаря большей четкости получаемого изображения. Как уже было замечено, качество изображения вблизи будет улучшаться с любым фильтром при повышении разрешения, а муар будет пропадать (мы проверили это, отрисовав тестовую сцену с кругами из работы [4] с box фильтром в разрешении 8192x8192 вместо 1024x1024, и получили более четкие края в нашем методе, чем в [4]).

Еще одним потенциальным преимуществом нашего подхода перед NSAA являются точные вычисления при определении видимости (используется арифметика с высокой, но ограниченной точностью). Мы не пользуемся эпсилон-подходами для обеспечения устойчивости, благодаря чему мы можем поддерживать сцены с огромным количеством очень мелких треугольников, а также гарантируем корректное исполнение на любой входной сцене, что важно для практических приложений.

Недавняя публикация [5] развивает идею A-буфера, предложенную в 1984 году в [6]. Это практичные подходы, пытающиеся повысить эффективное разрешение отрисовки без прямого использования MSAA. В каждом пикселе вводятся, например, 32 точки, и для каждого треугольника, пересекающегося с пикселем, вычисляется 32-битная маска, показывающая в какие точки попал этот треугольник. Затем все попавшие в пиксель треугольники комбинируются согласно их маскам, глубинам и т. п. Авторы стремятся приблизиться к качеству 256-кратного MSAA, в то время как [1], [4] и наш поход обеспечивают точное вычисление результата антиалиасинга.

Похожий подход применяется в методе DCAA (Decoupled Coverage Anti-Aliasing) [7]. Этот метод использует спроецированные в пространство экрана рёбра треугольника для того, чтобы детектировать возможную границу поверхности. Далее он генерирует битовую маску (64 бит) для каждого треугольника, попадающего в пиксель, чтобы сохранить для субпикселя т. н. «карту покрытия» (coverage map) пикселя треугольником в разрешении 8x8 точек. То есть представляет субпиксель при помощи нескольких значений и битовой маски – карты покрытия. DCAA можно в некоторой степени считать предшественником метода из работы [5]; в целом он имеет те же недостатки.

В мире отрисовки не в реальном времени известна архитектура Reyes [8]. В качестве антиалиасинга в ней изначально использовался A-буфер [6]. Но уже в 1987 году, по-видимому, используется подход с 16 точками сэмплирования в каждом пикселе со случайно возмущенными позициями [8]. Особенностью архитектуры является использование в качестве входной геометрии произвольных поверхностей, таких, как бикубические патчи. Эти поверхности перед отрисовкой разделяются на микрополигоны размером приблизительно 1/4 пикселя, что хорошо соответствует используемому антиалиасингу. Если целый элемент поверхности имеет экранный размер намного меньше пикселя до разбиения на микрополигоны, то вероятны проблемы с алиасингом; но в практических приложениях Reyes можно просто избегать таких ситуаций.

Антиалиасинг является одним из возможных приложений метода, представленного в статье [9]. Так же, как и в нашем подходе, для определения видимой геометрии используется векторное хранилище. В то время как в нашем алгоритме хранятся общие многоугольники (возможно, невыпуклые и с дырами), в [9] используются только выпуклые многоугольники. При вычитании треугольника они разбивают получающийся общий многоугольник на выпуклые части. Для надежности вычислений координаты вершин хранятся в узлах целочисленной сетки; координаты точек пересечения округляются к ближайшим целым. Так как в результате округления многоугольник может испортиться (например, стать невыпуклым), они заменяют каждый получающийся многоугольник его выпуклой оболочкой. Это оригинальное решение проще и, наверное, быстрее нашего, так как не нужно производить булевы операции над общими многоугольниками. В нашем подходе определение видимой геометрии производится без погрешностей в промежуточных вычислениях; разница с алгоритмом из [9], возможно, будет проявляться только на пикселях, в которые попал миллиард (или еще больше) треугольников, но этот риск в подходе [9] может присутствовать.

Алгоритмы булевых операций над многоугольниками, которыми мы пользуемся, сходны в общих чертах с теми, которые предложили Weiler и Atherton в [10]. Мы добавили стадию реконструкции контуров из [11], что помогает обеспечить корректную работу в ситуациях, где несколько контуров одного многоугольника касаются в общей вершине. Также, нам пришлось продумать некоторые детали реализации, например, как должны обрабатываться многоугольники с совпадающими частями границ.

III. Предложенный алгоритм

Предлагаемый алгоритм преобразует трехмерную сцену из непрозрачных треугольников в растр, содержащий в каждом пикселе значение интеграла (1) (учитывающего площади видимых фрагментов всех треугольников внутри пикселя).

Идея алгоритма состоит в том, чтобы для каждого пикселя определить фрагменты треугольников (многоугольники), видимые в нём, и затем сложить цвета этих многоугольников, умноженные на их площади (размер пикселя считаем равным 1). Цвет треугольника в пикселе мы считаем постоянным. Это упрощение нам кажется приемлемым несмотря на то, что в реальности результат работы пиксель-шейдера является непостоянной и даже нелинейной функцией.

Алгоритм состоит из 2 стадий:

1. построение векторного хранилища, состоящего из видимых на экране входных треугольников, каждый из которых также содержит непустое множество своих видимых фрагментов (многоугольников);
2. растеризация каждого видимого многоугольника с добавлением в растр вычисленного в исходном треугольнике цвета каждого затронутого пикселя, умноженного на площадь пересечения многоугольника и пикселя.

6. Построение векторного хранилища

На данной стадии решается задача удаления невидимых поверхностей в сцене. Решение основывается на следующем наблюдении: если 2 треугольника в трехмерном пространстве не имеют общих внутренних точек, то их проекции на экране либо так же не пересекаются внутренними точками (рисунок 1 (a), (b)), либо накладываются так, что можно выбрать один треугольник, находящийся поверх другого (рисунок 1 (c)). Теперь, если с каждым треугольником связано множество его видимых на экране фрагментов (рисунок 2), то вычитание этих множеств в соответствующем порядке решит задачу удаления невидимых поверхностей для данной пары треугольников. Не ограничивая общности предположим, что мы добавляем треугольники в сцену по одному, и при этом известно, что треугольники не пересекаются внутренними точками в пространстве. Тогда для каждого нового треугольника можно определять, с какими из уже добавленных треугольников он пересекается на экране, и для каждого пересечения определять, какой из треугольников загораживает другой, и осуществлять вычитание множеств видимых фрагментов. Если необходима поддержка треугольников, пересекающихся внутренними точками в трехмерном пространстве и не лежащих в одной плоскости (то есть протыкающих друг друга, рисунок

1 (d)), то можно разреза́ть добавляемый треугольник по каждому из уже существующих протыкающих треугольников (линия разреза определяется пунктирным отрезком на рисунке).

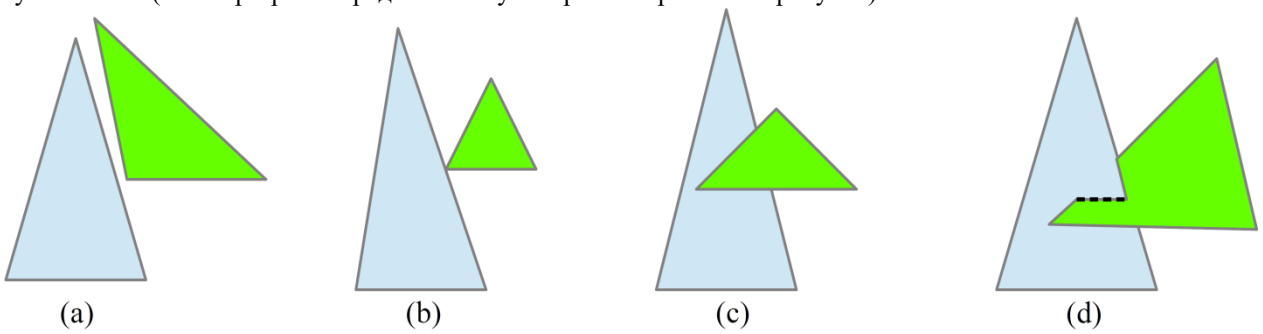


Рис. 1. Взаимное расположение 2 треугольников: (a), (b) – не пересекаются внутренними точками ни в пространстве, ни на экране; (c) – не пересекаются внутренними точками в пространстве, но имеют общие внутренние точки на экране; (d) – пересекаются внутренними точками в пространстве

Крайне желательно, чтобы задача определения видимых фрагментов решалась без существенной потери точности. Дело в том, что если координаты видимых фрагментов вычисляются с округлением, то между этими фрагментами могут образовываться щели. Через эти щели могут быть видны другие фрагменты, которые должны быть полностью скрытыми, и это приводит к усложнению содержимого векторного хранилища и понижению качества результирующего изображения. Решением этой проблемы является использование точных вычислений на данном этапе алгоритма. Координаты входных вершин «защелкиваются» к точкам целочисленной сетки, имеющей малый шаг, и затем вычисления производятся в целочисленной арифметике с использованием чисел с подходящей битовой шириной для каждого конкретного выражения. (Альтернативным решением этой проблемы могут быть эпсилон-подходы (например, в работе [3]), но для них трудно доказать устойчивость на произвольных входных сценах, и, кроме того, они могут влиять на качество результата, например, в ситуации, где на входе присутствует большое количество очень мелких треугольников.) Насколько нам известно, по похожей причине растеризация в аппаратуре и в таких библиотеках, как OpenGL, всегда реализуется в арифметике с фиксированной точкой.

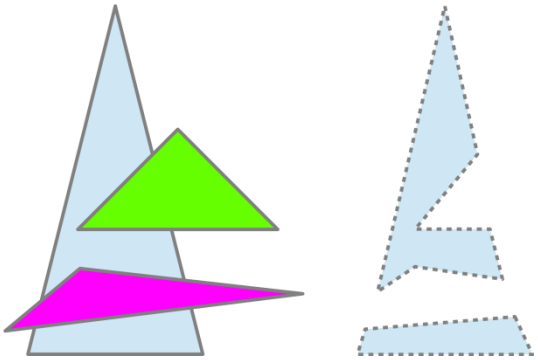


Рис. 2. Треугольник, частично загороженный на экране двумя треугольниками (слева), и его множество видимых фрагментов (справа)

Результатом вычитания двух треугольников является многоугольник, поэтому множества видимых фрагментов состоят из многоугольников. В принципе, можно всегда работать с треугольниками: при каждом вычитании можно генерировать триангулированный результат и хранить его. Но в этом случае мы будем терять точность, т. к. координаты вершин новых треугольников являются выражениями, зависящими от координат входных треугольников, и глубина зависимости может расти неограниченно по мере того, как мы добавляем новые треугольники. Проводить вычисления без округлений становится затруднительно, а округления приведут к образованию нежелательных щелей, как было описано выше. В отличие от треугольников, для многоугольников булевы операции можно реализовать так, чтобы глубина зависимости выражений оставалась ограниченной, что позволяет проводить вычисления точно.

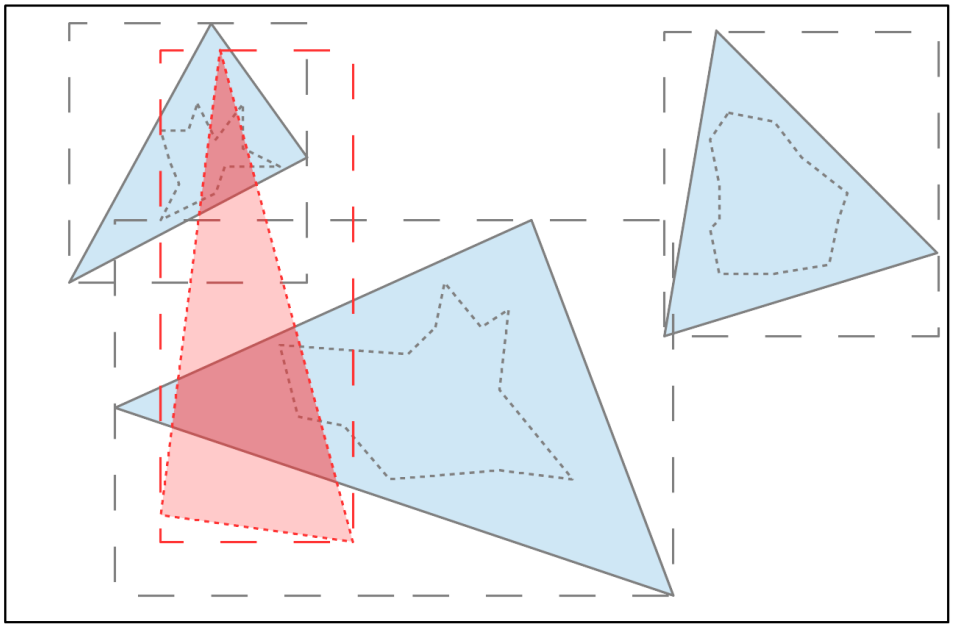


Рис. 3. Векторное хранилище с тремя из хранящихся в нём треугольниками (голубого цвета), с их видимыми множествами (мелкий пунктир) и описанными прямоугольниками (крупный пунктир). Красный треугольник – добавляемый в хранилище

Теперь мы можем привести алгоритм построения векторного хранилища видимой геометрии. В начале это хранилище инициализируется как пустое множество. Затем для каждого добавляемого треугольника выполняются следующие шаги.

1. Проецирование и обрезка по полю зрения.
2. Защелкивание координат вершин к целочисленной сетке (она имеет шаг $1 / 2^{45}$ размера пикселя).
3. Если треугольник не ориентирован против часовой стрелки на экране, он пропускается.
4. Из хранилища производится выборка уже добавленных треугольников, потенциально пересекающихся с добавляемым (по описанным прямоугольникам: на рисунке 3 выберутся 2 голубых треугольника слева).
5. Добавляемый треугольник разрезается на треугольные фрагменты по всем выбранным треугольникам, которые протыкают его в трехмерном пространстве (см. рисунок 1 (d)). Если разрезание произошло, то для каждого фрагмента выполняются шаги 2–4, затем 6 и далее.
6. У добавляемого треугольника инициализируется множество его видимых фрагментов (состоящее из одного многоугольника, равного добавляемому треугольнику).
7. Следующие два шага выполняются для каждого треугольника из выборки.
8. Вычисляется зона наложения с добавляемым треугольником (см. рисунок 1 (c)) и определяется, какой из двух треугольников ближе к зрителю (если вычисления в арифметике с плавающей точкой на этом этапе показали, что наложения нет, считаем, что уже хранящийся треугольник ближе к зрителю).
9. Производится вычитание множеств видимых многоугольников: из того, что дальше от зрителя, вычитается то, что ближе к зрителю. Если видимое множество треугольника, находящегося в хранилище, становится пустым, то этот треугольник удаляется.
10. Конец цикла.
11. Если видимое множество добавляемого треугольника не пусто, то он добавляется в хранилище.

Шаги 1, 5 и 8 производятся в арифметике с плавающей точкой; они не критичны для устойчивости алгоритма.

В шаге 2 степень 45 выбрана из следующих соображений. Координаты исходных вершин, поступающие на вход булевых операций над многоугольниками, являются 64-битными беззнаковыми целыми числами. Если младшие 45 бит этих чисел представляют положение вершины внутри пикселя, то на номер пикселя остается 19 бит. Два старших бита должны быть нулями из-за особенностей текущей реализации. Остающиеся 17 бит позволяют использовать максимальное разрешение большее, чем 100K пикселей в каждом направлении.

Шаг 3 предполагает, что входная сцена состоит из односторонних треугольников. Одинаковая ориентированность контуров на экране важна для алгоритмов булевых операций над многоугольниками, которыми мы пользуемся.

Если требуется поддержка полупрозрачных входных треугольников, то этот алгоритм несколько усложнится. Скорее всего, каждый элемент в хранилище видимых непрозрачных треугольников должен будет содержать список накладывающихся на него полупрозрачных треугольников, чтобы при растеризации можно было вычислять цвет в каждом пикселе с учетом всех слоев полупрозрачностей. На момент написания это не реализовано.

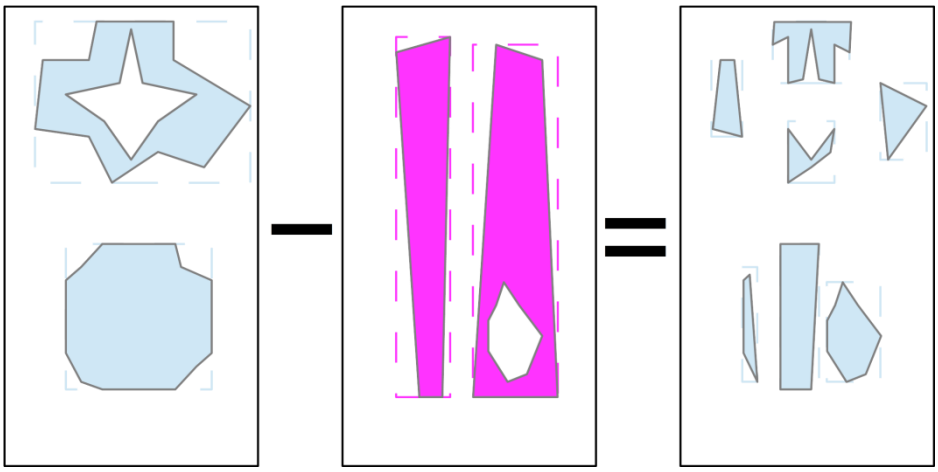


Рис. 4. Вычитание двух множеств многоугольников. Каждое множество хранится в структуре BoxQuadtree (хранилище треугольников на рисунке 3 также является BoxQuadtree, только с другим типом элемента)

В шаге 9 алгоритма производится обращение к вспомогательным алгоритмам, описанным в разделе IV. Каждое множество видимых многоугольников представляет собой структуру данных BoxQuadtree (секция IV.3): набор многоугольников с описанными прямоугольниками, которые хранятся в квадро-дереве. Многоугольники из одного BoxQuadtree могут контактировать только в общих вершинах, то есть они не могут накладываться своими территориями или отрезками границ. Под вычитанием двух BoxQuadtree подразумевается булево вычитание их многоугольников как множеств точек на плоскости (точнее, замыкание результата такого вычитания, так как получающиеся многоугольники содержат свои границы). Это вычитание производится в 2 подсистемах: BoxQuadtree производит определение накладывающихся пар описанных прямоугольников из двух деревьев, и для каждой пары вызывается подсистема вычитания 2 многоугольников (потенциально невыпуклых и содержащих дыры, секция IV.1), вычисляющая, строго говоря, замыкание результата этого вычитания. Пример вычитания двух множеств многоугольников приведен на рисунке 4.

Шаг 9 алгоритма можно модифицировать: вычитать не множество видимых многоугольников, а исходный треугольник. Это потенциально может уменьшить количество вычислений, так как загораживающая геометрия вычитается одним куском, а не фрагментами, как может получаться, если вычитаемый треугольник сам частично загорожен другими треугольниками. Также, это позволило бы уменьшить общую сложность системы, так как подсистема вычитания множеств многоугольников была бы больше не нужна (достаточно уметь вычитать лишь один треугольник). Но эксперименты на сложной сцене показали, что такая модификация приводит к замедлению: похоже, что подсистема вычитания хорошо справляется со своей задачей оптимизации.

7. Растеризация многоугольников и смешивание цветов

После добавления всех входных треугольников в векторное хранилище необходимо перевести его в растр. Для каждого треугольника в хранилище мы проходим по всем многоугольникам в его видимом множестве и производим их растеризацию. Эта задача нетривиальна: многоугольник является сложным объектом (он может быть невыпуклым и содержать дыры), и результатом растеризации для каждого пикселя должен стать набор многоугольников, являющийся пересечением растеризуемого многоугольника и пикселя. Желательно также, чтобы обнаруживались блоки пикселей, целиком содержащиеся в растеризуемом многоугольнике, чтобы их можно было напрямую залить цветами исходного треугольника без смешивания.

Следует заметить, что для воображаемых очень сложных сцен, состоящих из треугольников намного меньших пикселя, для большинства многоугольников растеризация тривиальна. Достаточно определить, что описанный прямоугольник многоугольника целиком содержится в пикселе, и выдать в этом случае в качестве результата растеризации сам многоугольник. Поэтому, хотя приводимый ниже метод растеризации довольно сложен, его производительность не является очень важной в дальней перспективе.

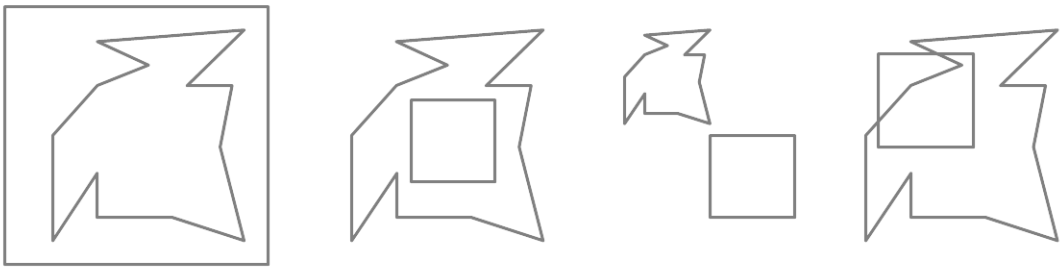


Рис. 5. Четыре возможных взаимных положения многоугольника и квадрата

Нам понадобятся 2 функции из подсистемы булевых операций над многоугольниками.

Функция PolygonAndSquareSituation производит упрощенное пересечение многоугольника и квадрата. В качестве результата она выдает только информацию об их взаимном положении (одно из 4 значений, в слегка упрощенном виде): многоугольник содержится в квадрате, квадрат содержится в многоугольнике, квадрат и многоугольник не имеют общих внутренних точек, или и квадрат, и многоугольник имеют как общие, так и не общие внутренние точки (рисунок 5).

Функция PolygonPixelIntersection возвращает набор многоугольников пересечения входного многоугольника и квадрата пикселя.

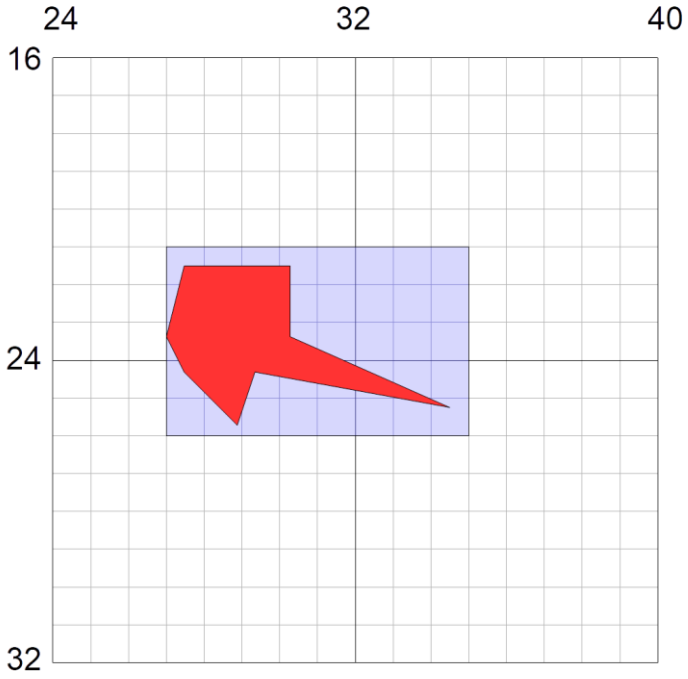


Рис. 6. Многоугольник с описанным прямоугольником размером 8x5 пикселей и 4 квадрата со стороной 2^3 пикселя, в которых содержится этот прямоугольник

Мы растеризуем многоугольник рекурсивным способом, подобно построению квадро-дерева, но без явного создания дерева в памяти. Каждый вызов рекурсивной функции рассматривает квадрат пикселей, имеющий размер – степень 2 и минимальные координаты, кратные этой степени 2.

В начале мы рассматриваем описанный прямоугольник растеризуемого многоугольника (расширенный до границ пикселей) и определяем степень 2, соответствующую большему из размеров прямоугольника (так, чтобы этот размер прямоугольника был не больше размера квадрата, но больше половины размера квадрата). Затем мы определяем от 1 до 4 квадратов с данной степенью 2, в которых содержится прямоугольник (рисунок 6).

Затем мы запускаем рекурсию. Если рассматривается квадрат размером 1 пиксель, то рекурсия достигает дна. В этом случае вызывается PolygonPixelIntersection и пользователю через указатель на callback-функцию возвращается набор многоугольников пересечения (если он не пуст). Если пиксель целиком содержался в многоугольнике, то мог быть вызван callback для заливки квадрата пикселей размером 1x1. Если же рассматривается квадрат больший, чем 1 пиксель, то вызывается PolygonAndSquareSituation и, в зависимости от

результата, вызывается callback для заливки квадрата пикселей, целиком содержащегося в многоугольнике, или рекурсия в квадраты меньшего размера, или рекурсия прекращается. Рисунок 7 иллюстрирует этот процесс.

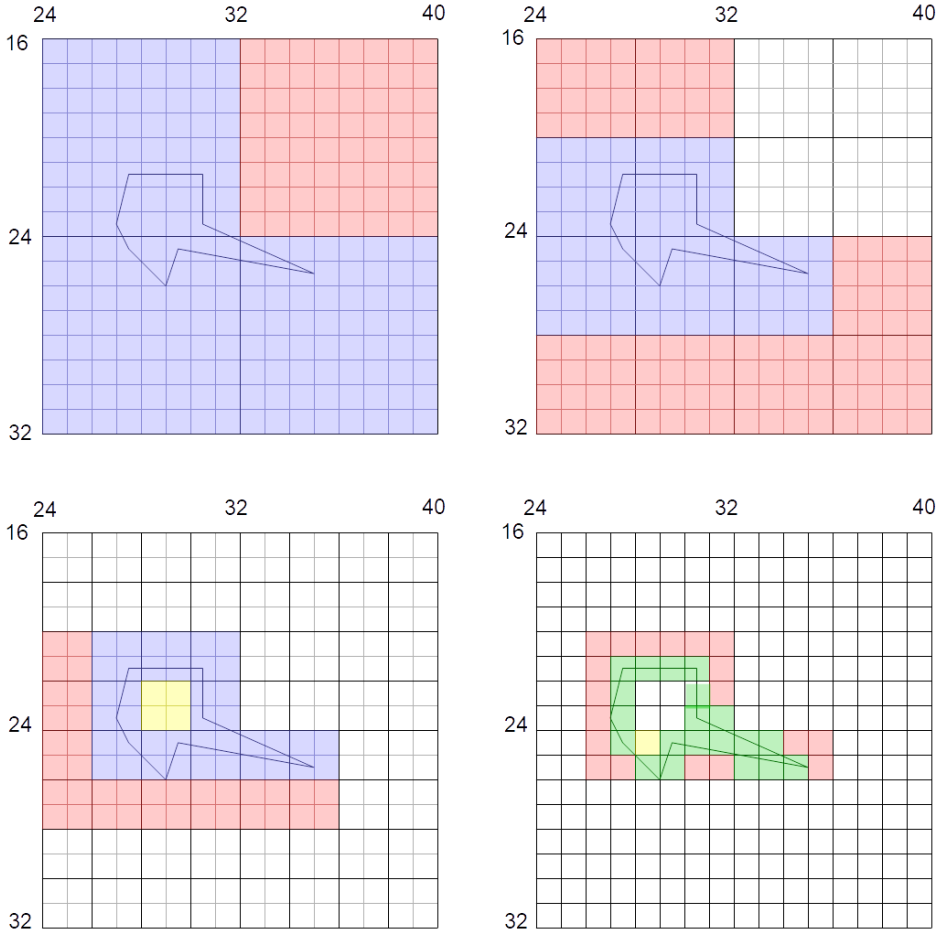


Рис. 7. Результаты рекурсивной растеризации на разных глубинах рекурсии, с квадратами от 8x8 до 1x1. Синий цвет означает спуск на следующий уровень рекурсии. Иначе рекурсия терминируется; при этом красный цвет означает отсутствие дополнительных действий; зеленый – вызов пользовательской функции, в которую передается набор многоугольников пересечения растеризуемого многоугольника и пикселя; желтый – вызов пользовательской функции заливки квадрата пикселей, целиком содержащегося в многоугольнике

Нам осталось описать, как производится смешивание цветов в растре. В начале растр заполняется черным цветом (нулями). Если при растеризации вызывается функция заливки целиком содержащегося в многоугольнике квадрата, то в пиксели просто записываются цвета, возвращаемые пиксель-шейдером исходного треугольника, вычисляемые в центрах пикселей. Если же вызывается функция для одного частично занятого пикселя (зеленый случай на рисунке 7), то в ней вычисляется суммарная площадь всех переданных в нее многоугольных фрагментов, которая затем умножается на цвет, вычисленный шейдером в центре пересечения исходного треугольника и пикселя, и прибавляется к значению, уже записанному в растре.

Нужно еще указать, с какой точностью производятся вычисления для смешивания цветов. Так как процедуры булевых операций над многоугольниками производятся без округлений, вершины многоугольников хранятся в виде рациональных выражений с числителем и знаменателем, состоящими из нескольких 32-битных цифр. Нам нужно вычислить площади многоугольников с такими вершинами. Это не требуется делать с высокой точностью, так как цвет из пиксель-шейдера, на который умножится суммарная площадь, уже может содержать значительную ошибку. Поэтому, в принципе, эти вычисления можно реализовать в стандартной машинной арифметике с плавающей точкой. Однако в сложных ситуациях в 1 пиксель будет попадать такое большое количество многоугольников, что ошибки округления промежуточных вычислений накопятся и приведут к значительному отклонению результата антиалиасинга. Поэтому мы реализовали вычисление площадей, умножение на цвет и суммирование в растре в целочисленной арифметике с округлением, и выбрали битовую ширину чисел так, чтобы гарантировать, что в результирующий цвет пикселя из-за округлений вносится ошибка не более 1 в сколь угодно сложном случае, возможном на 64-битном компьютере. Это относительно несложно, и позволяет нам не беспокоиться по поводу округлений.

IV. Вспомогательные алгоритмы

Здесь описаны вспомогательные алгоритмы, реализованные в нашей программе. Они вызываются из основного алгоритма, которому посвящен раздел III.

1. Булевы операции над многоугольниками

Нам нужны вычитание и пересечение пары многоугольников. Предлагаемый алгоритм комбинирует идеи из работ [10] и [11]. Наша реализация содержит многочисленные, существенные детали, которые, возможно, достойны отдельной публикации. Здесь представлен только общий обзор нашей работы.

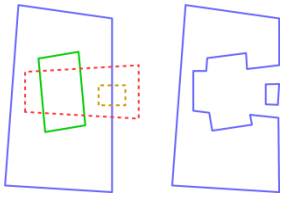


Рис. 8. Слева: многоугольник *B* (синий с зеленой дырой) и многоугольник *A* (красный с желтой дырой); справа: результат вычитания (два многоугольника)

Многоугольники задаются контурами: внешним контуром и, возможно, контурами дыр. Контуры являются простыми, то есть единственными пересечениями между отрезками одного контура являются контакты с соседями по контуру в начале и в конце каждого отрезка. Внешний контур ориентирован против часовой стрелки, а дыры ориентированы по часовой стрелке. Дыры могут касаться внешнего контура или между собой только в общих вершинах.

Точка считается принадлежащей многоугольнику, если она принадлежит внешнему контуру или конечной области плоскости, ограниченной этим контуром, но не принадлежит открытой конечной области, ограниченной дырой. Булевы операции над многоугольниками определяются как регуляризованные булевы операции (такие как объединение, пересечение, вычитание) над множествами точек этих многоугольников. Под регуляризацией результата поточечной булевой операции понимается замыкание внутренней области этого результата; это нужно для того, чтобы, например, результат вычитания содержал все свои граничные точки, а результат пересечения двух многоугольников, только касающихся общим ребром, был бы пустым множеством, а не вырожденным «многоугольником» из одного отрезка. При таком определении результаты нужных нам булевых операций над многоугольниками являются наборами многоугольников, или пустыми множествами.

Дальше мы приведем общие идеи алгоритма вычитания двух многоугольников: *B* минус *A*. Чтобы получить ответ, достаточно вычесть из *B* многоугольник *A* без дыр (обозначим его A_O), а затем присоединить к полученному списку многоугольников результаты пересечения *B* с каждой из дыр *A* (рисунок 8). Вычитание *B* и A_O производится с помощью следующих общих шагов.

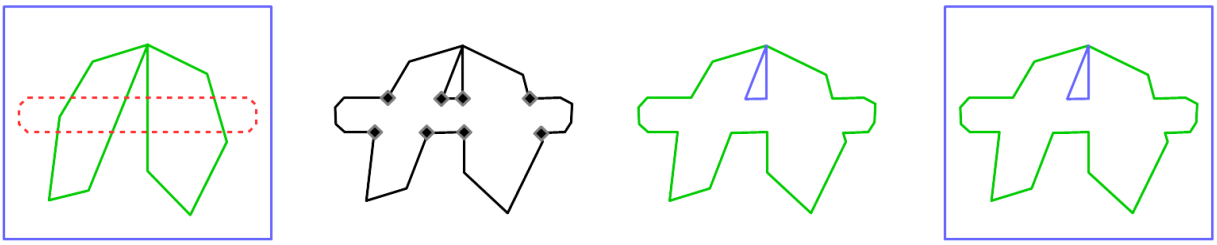


Рис. 9. Процесс вычитания двух многоугольников; (a): многоугольник *B* (синий с 2 зелеными дырами, касающимися в общей вершине) и многоугольник *A* (красный); (b): цепи отрезков – результат шага 3 алгоритма (8 цепей, вершина сверху является общей для 2 цепей); (c): восстановленные контуры – результат шага 4 алгоритма (1 дыра (зеленая) и 1 внешний контур (синий)); (d): результат вычитания – после шага 5 алгоритма (2 синих многоугольника, один из которых содержит зеленую дыру)

1. Отрезки обоих многоугольников подразделяются таким образом, чтобы любые 2 результирующие отрезка либо не пересекались, либо контактировали в начальной или конечной точке, либо совпадали.
2. Определяются ситуации входа и выхода контура A_O в и из *B*. Эти входы и выходы могут происходить как на внешней границе, так и в дырах *B*. Важным моментом здесь является то, что отрезки A_O, совпадающие с отрезками *B*, считаются находящимися вне *B*, или, другими словами, лежащими «справа» от отрезков *B*. Ситуации входа и выхода должны быть привязаны к отрезкам, а не к вершинам, так как

вершины могут быть общими для нескольких контуров, а вход и выход относятся к конкретным парам контуров.

3. Для каждой ситуации входа производится построение исходящих цепей отрезков B и A_O до первой встретившейся ситуации выхода (рисунок 9 (b)). Цепи отрезков A_O строятся с изменением направления на обратное.
4. Производится восстановление замкнутых контуров из множества сгенерированных цепей отрезков (рисунок 9 (c)). Для этого мы берем любую цепь и находим другую цепь, которую можно присоединить к концу данной. Затем мы идем вдоль присоединенной цепи и проверяем, не пересекается ли она в данной вершине с первой цепью. Если такое пересечение найдено, то мы можем выдать готовый замкнутый контур и продолжить процесс с тем, что осталось в цепях после выемки построенного контура. Если же мы дошли до конца присоединенной цепи, то мы считаем эти 2 цепи единой цепью и опять находим одну цепь, которую можно присоединить к концу, и повторяем процесс. В конце концов мы израсходуем все отрезки, и получим набор результирующих контуров (здесь, как и в других частях данного алгоритма, существенно, что мы проводим вычисления без округлений).

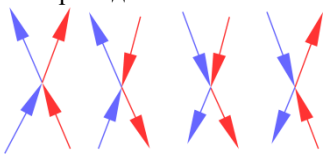


Рис. 10. Определение типа соседства 2 контуров по ситуации в общей вершине. Если известно, что контуры не пересекаются (т. е. нет входов и выходов), и если оба контура ориентированы против часовой стрелки, то их типы соседства будут, слева направо: синий содержится в красном, синий вне красного, синий содержит красный. Четвертая ситуация возможна только в том случае, если у контуров есть входы и выходы

5. Для правильной интерпретации типов полученных контуров (внешний или дыра какого-либо многоугольника) важно использовать информацию о типе соседства 2 контуров. Эти типы соседства изображены на рисунке 5: контуры могут содержаться один в другом, или быть разрозненными, или пересекаться. Нам важны типы соседства A_O с внешним контуром B и его дырами. Если мы встречали ситуацию входа или выхода между этими контурами, то их соседство – пересекающееся. Если входов и выходов не было, но были общие вершины, то тип соседства можно определить по расположениям и направлениям отрезков при этих вершинах (рисунок 10). Если же контуры вообще не контактировали, то мы можем воспользоваться известным алгоритмом определения, находится ли вершина одного контура внутри или снаружи другого контура. Теперь, вооружившись полученной информацией, мы можем дать окончательный ответ о результате вычитания B и A_O. Возможны следующие ситуации: B стирается полностью, B остается без изменений, B получает новую дыру (которая, возможно, «схлопывает» некоторые из существующих дыр) (рисунок 9 (d)), или вместо B получается набор новых многоугольников (в этом случае все сгенерированные в пункте 4 контуры – внешние, а не дыры). Наконец, может быть необходимым перенести оставшиеся в B дыры в новые многоугольники.
6. Если какой-либо из входных многоугольников планируется использовать в дальнейших операциях, то производится удаление новых вершин, установленных на его отрезки в шаге 1 (иначе эти вершины могут накапливаться и существенно замедлять программу).

Важным моментом для программной реализации этого алгоритма является использование исходных отрезков. Можно видеть, что каждый отрезок, генерируемый нашим алгоритмом, является фрагментом исходного отрезка, принадлежавшего изначально поступившему на вход треугольнику. Вершины этого исходного отрезка целочисленны (они были защелкнуты к имеющей малый шаг сетке). Благодаря тому, что каждый генерируемый отрезок хранит указатель на исходный отрезок, при вычислении координат точки пересечения двух отрезков можно воспользоваться формулой для точки пересечения двух прямых, проходящих через пары целочисленных точек. Это позволяет хранить координаты всех вершин как рациональные числа с фиксированной битовой шириной числителя и знаменателя, и избегнуть роста глубины зависимости выражений по мере усложнения отрисовываемой сцены.

2. Точная рациональная арифметика

Для того, чтобы алгоритмы булевых операций над многоугольниками работали надежно, необходимо, чтобы геометрические вычисления проводились без потери точности. Как уже обсуждалось, координаты вершин, изначально подающихся на вход алгоритмов, являются 64-битными беззнаковыми целыми. Координаты вершин пересечения являются рациональными числами. Мы приведем некоторые подробности.

Пусть A и B – двумерные векторы. Мы будем обозначать с помощью выражения $[A, B]$ скалярную величину $A_xB_y - A_yB_x$.

Пусть одна прямая задана разными точками A и B, а другая прямая – разными точками C и D. Если выражение $[B - A, D - C]$ не равно нулю, то прямые не являются совпадающими или параллельными, и пересекаются в одной точке. Координаты этой точки могут быть вычислены по следующим формулам.

$$V = A + (B - A)\frac{[C - A, D - C]}{[B - A, D - C]} = \frac{A[B - A, D - C] + (B - A)[C - A.D - C]}{[B - A, D - C]}$$

Таким образом, если координаты точек A, B, C, D являются целыми, V может быть точно представлена двумя рациональными числами, имеющими одинаковые знаменатели.

Далее мы объясним, как эти рациональные числа вычисляются и хранятся в программе на языке Си. Так как Си не предоставляет доступа к флагу переноса после сложения или вычитания, и к верхней половине результата умножения, мы будем производить вычисления над 32-битными фрагментами чисел, но в 64-битном режиме, чтобы получить доступ к верхним битам результатов.

Компоненты векторов A, B, C, D являются беззнаковыми 64-битными целыми. Их разности, например, $(B - A)_x$, также являются беззнаковыми 64-битными, но с дополнительным 1-битным флагом − знаком числа. Произведение 2 разностей помещается в две 64-битные величины плюс бит знака. Сумма или разность произведений разностей (например, число $[B - A, D - C]$) требует 129 бит плюс знак, и хранится как 5 беззнаковых 32-битных величин плюс флаг знака; это знаменатель позиции вершины. Далее нам нужны произведения чисел последнего типа и 64-битных величин. Можно вычислить, проведя операции над максимально возможными величинами, что такие произведения требуют 193 бита плюс знак, и сумма этих произведений требует 194 бита плюс знак, так что эти числа хранятся как массивы из 7 беззнаковых 32-битных целых, плюс флаг знака. Это является форматом хранения числителей позиции вершины.

Похожие рассуждения могут быть проведены для более сложных выражений, которые могут потребоваться в программе, например, для сравнений координат вершин. Сложность этих выражений ограничена, так что мы сможем обойтись операциями сложения, вычитания и умножения над массивами целых чисел фиксированной длины. Также потребуется деление (самая сложная операция), например, для вычисления целых частей позиций вершин для построения описанных прямоугольников вокруг отрезков и контуров; оно также было реализовано.

3. Структура данных BoxQuadtree

В приведенных выше алгоритмах есть несколько мест, в которых производятся операции над группами двумерных геометрических объектов. Например, нам нужно уметь находить все пересечения отрезков из двух многоугольников. Может также потребоваться вычитать попарно многоугольники из двух множеств многоугольников. После построения нового результирующего многоугольника желательно уметь находить все потенциально попадающие в него дыры. Для ускорения этих операций желательно уметь группировать находящиеся рядом геометрические объекты. Мы предлагаем использовать для этого квадро-дерево; одним из достоинств такой структуры данных является автоматическая адаптация к входным объектам разного размера.

Мы реализовали универсальную структуру данных BoxQuadtree. Объекты, которые могут храниться в этой структуре, должны быть представлены их описанными прямоугольниками, ориентированными по осям координат. Прямоугольники хранятся в виде минимальных и максимальных беззнаковых 64-битных целых координат по осям x и y. Эти координаты соответствуют сетке с фиксированным шагом, которую мы используем для координат входных вершин. Каждая пара целых чисел (i, j), принадлежащая прямоугольнику, представляет следующую область плоскости:

$$i \le x < i + 1,$$
$$j \le x < j + 1.$$

Такая организация обеспечивает то, что каждая точка плоскости в пределах границ входной сетки принадлежит в точности одной ячейке, используемой прямоугольниками, и эта ячейка имеет индексы, равные целым частям координат точки. Для любого отрезка или контура мы можем однозначно определить минимальный прямоугольник, содержащий его, и затем использовать этот прямоугольник на начальном этапе поиска пересечений: отрезки или контуры могут пересекаться только если их прямоугольники имеют непустую зону наложения.

Организация структуры BoxQuadtree достаточно проста. Она работает только с 64-битными беззнаковыми целыми описанных прямоугольников. Каждый узел квадро-дерева представляет квадратную область; подузлы этого узла, если они присутствуют, представляют 4 равные квадратные четверти этой области. Размер квадрата узла всегда является степенью двойки, и показатель этой степени называется уровнем данного узла. Наименьший уровень равен 0; эти узлы размером 1x1 не могут иметь подузлы. Позиция узла определяется наименьшими

целыми координатами его квадрата. Эти координаты должны быть делимыми на размер квадрата; это означает, что двоичные представления этих координат имеют нули во всех битах с индексами меньшими, чем уровень.

В узлах хранятся описанные прямоугольники элементов. Каждый прямоугольник хранится в 1, 2 или 4 узлах, имеющих одинаковый уровень. Этот уровень l определяется по максимуму из ширины и высоты прямоугольника d согласно формуле

$$2^{l-1} < d \leq 2^l.$$

Это обеспечивает, что прямоугольник занимает не более 2 узлов в каждом направлении, и не более 4 узлов всего.

Определение структур BoxQuadtree на языке Си приведено в приложении 1.

В главной Си-структуре BoxQuadtree хранится не 1, а 4 корня квадро-деревьев, т. е. лес из не более чем 4 деревьев, занимающий квадрат с определенными минимальными координатами. Это нужно для того, чтобы обойти требование выравненности координат узлов по размерам узла. Если бы использовался один корень, то маленький прямоугольник, пересекающий границу между двумя узлами высокого уровня, приводил бы к тому, что уровень корня был без необходимости высок.

Структура узла хранит двусвязный список звеньев, соответствующих хранимым элементам, и указатели на подузлы и на вышестоящий узел, если они присутствуют. Структура звена хранит указатели на следующий и предыдущий элементы в двусвязном списке, и указатели на элемент и на узел. Структура элемента содержит описанный прямоугольник, указатель на полезные данные элемента (предоставленные пользователем) и 4 указателя на звенья. Из этих 4 указателей, только указатель с индексом 0 используется всегда, так как элемент всегда находится в одном из узлов. Если элемент находится в 2 узлах, то будут использоваться указатели на звенья с индексами 0 и 1 или 0 и 2; если элемент находится в 4 узлах, то будут использоваться все 4 указателя. Координаты узлов, в которые попадает элемент, легко вычисляются по его прямоугольнику.

Эта организация обеспечивает быстрый обход, добавление и удаление элементов в BoxQuadtree. Координаты и размеры узлов не хранятся, так как их легко вычислять в рекурсивных операциях на деревьях.

В структурах элементов также предусмотрены 2 флага, показывающие, что элемент уже был пройден операцией, или уже произвел свой обход противоположного BoxQuadtree в операциях, обрабатывающих два BoxQuadtree. Эти флаги нужны, так как элементы могут содержаться в нескольких узлах, и поэтому могут быть пройдены более, чем 1 раз в процессе операции. Чтобы не было необходимости сбрасывать эти флаги после операций, они определены как 64-битные величины, которые хранят номер операции. Когда необходимо сбросить все эти флаги, достаточно лишь увеличить номер операции, хранимый в главной структуре BoxQuadtree.

Операция запроса у BoxQuadtree всех прямоугольников, пересекающихся с заданным, достаточно проста. Она организована в одной рекурсивной функции и использует одно из полей флагов в структурах элементов.

Операция пересечения двух BoxQuadtree должна сообщить все пересечения между прямоугольниками, хранимыми в двух деревьях, избегая при этом повторных сообщений. Эта операция не производит существенной модификации деревьев. В процессе этой операции необходимо уметь проводить совместный обход двух поддеревьев с корнями в одном и том же квадрате. Этот обход должен быть исчерпывающим для всех элементов, хранимых в данных корнях, чтобы их можно было отметить как совершившие свои обходы (иначе мы рискуем сообщать некоторые пересечения более, чем 1 раз). Поэтому, так как элементы могут храниться более чем в одном узле, мы должны иметь возможность обращаться к соседям данного корня поддерева, чтобы осуществлять рекурсивные запросы в этих соседях элементом, который пересекается с данным корнем соседа. Такая функциональность была реализована.

Нам также может понадобиться операция вычитания двух BoxQuadtree; однако, если произвести модификацию основного алгоритма, описанную в последнем абзаце секции III.1, то, возможно, эта операция будет не нужна. Она более сложна, чем пересечение, так как одно из деревьев может изменяться в процессе операции. Здесь мы больше не требуем, чтобы функция вычитания пары многоугольников вызывалась ровно один раз для каждой пары накладывающихся описанных прямоугольников; причина этого в том, что такое требование привело бы к значительному усложнению программы, что могло бы перевесить получаемые преимущества. В целом, реализация вычитания сходна с реализацией пересечения BoxQuadtree. Здесь не требуется уметь вызывать рекурсию из соседей данного узла, но, из-за необходимости модифицировать дерево, из которого производится вычитание, реализация остается довольно сложной.

Для структуры данных BoxQuadtree были также реализованы несколько более простых операций, таких как добавление и удаление элементов и итерация через всё дерево.

V. Результаты

На рисунках 11, 12, 13 приведены сравнения результатов антиалиасинга разных алгоритмов. Реализация NSAA использует гауссов фильтр с радиусом 2.3 пикселя. Отрисовка производилась в разрешении 1024x1024. В

левых квадратах на рисунках приведен результат работы нашего алгоритма, справа — увеличенный фрагмент с разными алгоритмами антиалиасинга. Мы рекомендуем читателям цифровой версии данной работы рассматривать эти рисунки со значительным увеличением, так как понижение разрешения растрового изображения на экране может вносить существенные артефакты.

Время отрисовки с нашим алгоритмом составляет порядка 10 минут на 1 миллион треугольников (при использовании одного потока вычислений на 64-битном процессоре Intel с тактовой частотой 3.1 ГГц). Мы не приводим времена отрисовки для других алгоритмов антиалиасинга, так как простоты ради мы реализовали их в нашей программе с использованием того же метода определения видимости, что и в нашем алгоритме (хранилище многоугольников), в то время как другие варианты антиалиасинга могут использовать более быстрые методы и исполняться на подходящих графических процессорах, что обеспечивает гораздо более высокую скорость.

В таблице 1 приведены значения PSNR (peak signal-to-noise ratio) относительно результата антиалиасинга с SSAA 16x, формально считающегося эталоном.

ТАБЛИЦА 1
PSNR РАЗНЫХ АЛГОРИТМОВ ОТНОСИТЕЛЬНО SSAA 16X

Тип AA	**PSNR звездочки**	**PSNR дороги**	**PSNR Hairball**
Без AA	36.4	40.7	23.5
MSAA 4x	45.3	46.8	32.1
MSAA 8x	51.1	55.8	36.7
NSAA	38.6	42.7	25.1
Предложенный метод	**54.1**	**60.2**	**39.2**

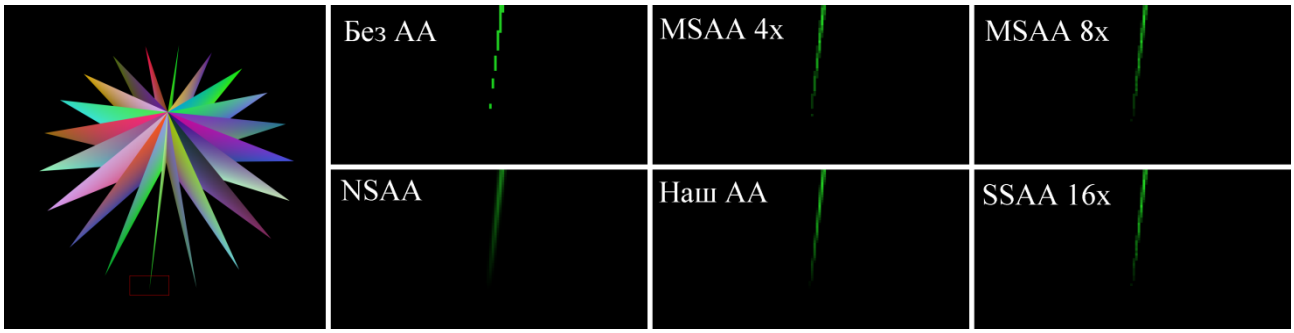


Рис. 11. Звездочка с разными вариантами антиалиасинга

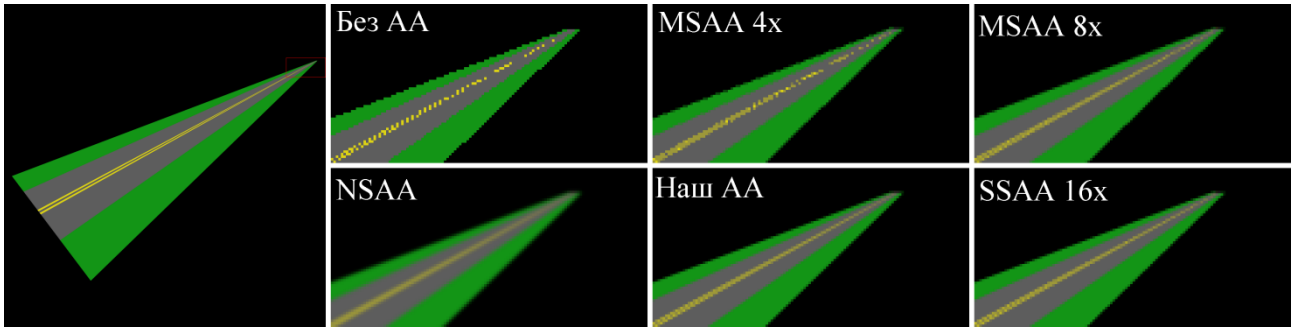


Рис. 12. Дорога с разными вариантами антиалиасинга

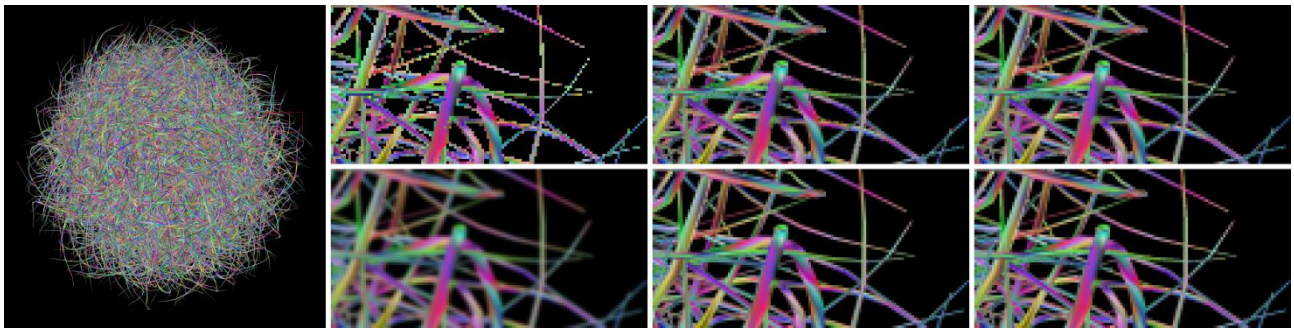


Рис. 13. Модель Hairball с разными вариантами антиалиасинга (расположение вариантов такое же, как на рисунках 11, 12)

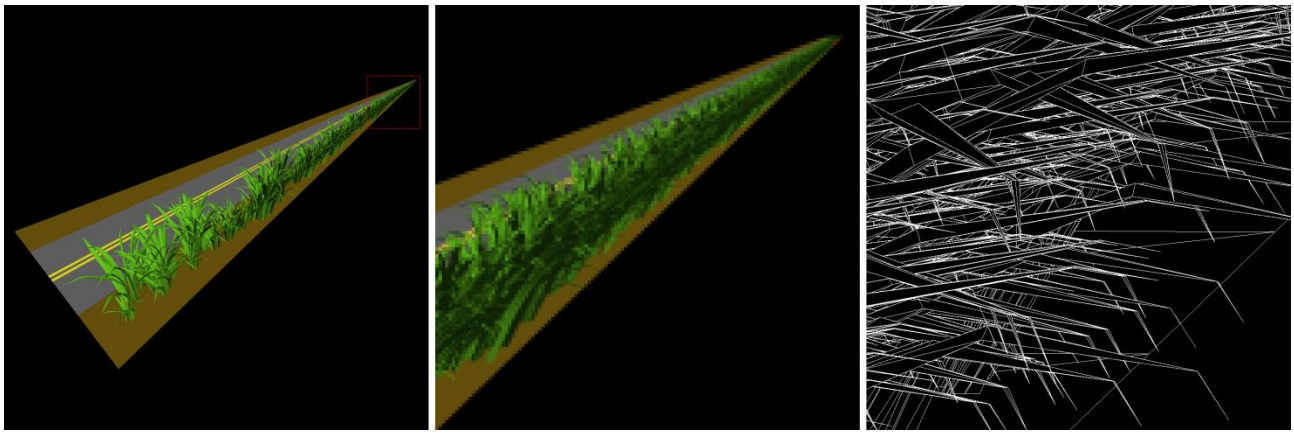


Рис. 14. Сцена с моделями травы с нашим антиалиасингом

Рисунок 14 представляет результат работы нашего алгоритма с уходящей вдаль смоделированной травой. Изображение в центре является увеличением дальнего участка сцены; справа отрисован в контурном режиме самый сложный пиксель сцены. В него попадают 3572 многоугольных фрагмента.

Модели звездочки и дороги являются простейшими сценами, заданными непосредственно в коде программы. Модели Hairball и кустика травы взяты из открытых источников в интернете (авторам спасибо!). Hairball состоит из 2.88 миллиона треугольников. В сцене с травой используются 1800 экземпляров модели общей сложностью около 2 миллионов треугольников.

VI. Заключение

Мы придумали и реализовали алгоритм антиалиасинга, вычисляющий интеграл видимого изображения по каждому пикселю и учитывающий все видимые треугольники, пересекающиеся с данным пикселем. Использование целочисленной арифметики обеспечивает стабильную работу на сколь угодно сложных сценах, помещающихся в память 64-битного компьютера, и предотвращает накопление погрешностей. Подсистема булевых операций над многоугольниками и структура данных BoxQuadtree могут оказаться полезными и в других приложениях.

Производительность нашей программы не очень высока, и на данном этапе ее лучше рассматривать не как реализацию практичного метода, а как демонстрацию того, что антиалиасинг с упомянутыми характеристиками в принципе может быть реализован. Возможностью дальнейшего развития является адаптация к характеристикам развивающейся вычислительной аппаратуры. Для нашей программы было бы весьма полезно увеличение быстродействия оперативной памяти и уменьшение падения производительности при неправильно предсказанных ветвлениях в длинных конвейерах процессоров, так как вероятно, что эти особенности современных компьютеров являются существенными препятствиями для получения высокопроизводительной реализации нашего алгоритма.

Благодарности

Thomas Auzinger помог с реализацией NSAA, за что мы его благодарим. Мы также благодарим Алексея Геннадьевича Волобоя за полезные советы и замечания к данной работе.

Список литературы

1. Catmull E. A Hidden-Surface Algorithm with Antialiasing // SIGGRAPH (New York, NY, USA, 1978): Proc. of SIGGRAPH ’78 / ACM. P. 6–11.

2. Auzinger T., Guthe M., Jeschke S. Analytic Antialiasing of Linear Functions on Polytopes // Comp. Graph. Forum 31 (May 2012). P. 335–344.

3. Auzinger T., Wimmer M., Jeschke S. Analytic Visibility on the GPU // Comp. Graph. Forum 32 (2013). P. 409–418.

4. T. Auzinger, P. Musialski, R. Preiner, M. Wimmer. Non-Sampled Anti-Aliasing // Proceedings Vision, Modeling and Visualization (2013). P. 169–176.

5. C. Crassin, C. Wyman, M. McGuire, A. Lefohn. Correlation-Aware Semi-Analytic Visibility for Antialiased Rendering // HPG ’18: High-Performance Graphics (Vancouver, Canada, 2018): ACM, New York, NY, USA.

6. Carpenter L. The A-buffer, an Antialiased Hidden Surface Method // SIGGRAPH (1984): Proc. of SIGGRAPH ’84 / ACM. P. 103–108.

7. Yuxiang Wang, Chris Wyman, Yong He, and Pradeep Sen. Decoupled coverage anti-aliasing. // In Proceedings of the 7th Conference on High-Performance Graphics (HPG '15). 2015. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/2790060.2790068

8. Cook R.L., Carpenter L., Catmull E. The Reyes Image Rendering Architecture // SIGGRAPH (1987): Proc. of SIGGRAPH 87 / ACM. P. 95–102.

9. Yang Zhou, Lifan Wu, Ravi Ramamoorthi, Ling-Qi Yan. Vectorization for Fast, Analytic and Differentiable Visibility // ACM Transactions on Graphics, Vol. 40, No. 3, Article 27 (2021).

10. Weiler K., Atherton P. Hidden surface removal using polygon area sorting // SIGGRAPH (1977): Proc. of the SIGGRAPH 1977 / P. 214–222.

11. Margalit A., Knott G.D. An algorithm for computing the union, intersection or difference of two polygons // Computers & Graphics. 1989. 13: P. 167–183.

Приложение 1

Здесь приводятся определения структур BoxQuadtree на языке Си (стандарта C90).

```
typedef unsigned long long u64;


typedef struct CBoundingBox
{
    u64 mXMin,    /* the box actually spans [ mXMin, mXMax + 1 ) */
        mYMin,    /* and [ mYMin, mYMax + 1 )                    */
        mXMax,
        mYMax;
}
CBoundingBox;


typedef struct CQuadtreeEntry
{
    CBoundingBox      mBox;
    void              *mpData;

    u64               mTraversedMark,
                      mTraversingMark;

    struct CEntryLink *mLinks[4];
}
CQuadtreeEntry;


typedef struct CEntryLink
{
    struct CEntryLink    *mpNext,
                         *mpPrevious;
```

```
    CQuadtreeEntry       *mpEntry;

    struct CQuadtreeNode *mpNode;
}
CEntryLink;


typedef struct CQuadtreeSubnodes
{
    struct CQuadtreeNode *mNodes[4];
}
CQuadtreeSubnodes;


typedef struct CQuadtreeNode
{
    CEntryLink           *mpEntryLinkFirst,
                         *mpEntryLinkLast;

    CQuadtreeSubnodes    *mpSubnodes; /* NULL indicates a leaf node */

    struct CQuadtreeNode *mpSupernode;
}
CQuadtreeNode;


typedef struct CBoxQuadtree
{
    int                miRootsLevel; /* of a single node of mRoots */

    u64                mTraversingCounter,
                       mTraversedCounter;

    u64                mXPos,
                       mYPos;

    CQuadtreeSubnodes mRoots;
}
CBoxQuadtree;
```

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-84-95

Применение обратного рендеринга для выбора ракурсов в задаче реконструкции материалов
Inverse rendering for viewpoint selection in application to material reconstruction

В. В. Санжаров[3], В. А. Фролов[1] , В. А. Галактионов[2]
[1] Институт перспективных исследований проблем искусственного интеллекта и интеллектуальных систем МГУ имени М.В. Ломоносова, 119191 Москва, Ломоносовский пр-т, д.27, к.1.
[2] ИПМ им. М.В.Келдыша РАН, Миусская пл. 4, Москва, 125047, Россия
[3] МГУ им. М.В. Ломоносова, факультет ВМК, ГСП-1, Ленинские горы, Москва, 119991, Россия

V. V. Sanzharov[3], V. A. Frolov[1] , V.A. Galaktionov[2]
[1]IAI Moscow State University, GSP-1, Lomonosovsky Prospect, 1, 27, Moscow, 119191, Russia
[2]Keldysh Institute of Applied Mathematics RAS, Miusskaya pl. 4, Moscow, 125047, Russia
[3]Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia

Аннотация. Методы выбора точки обзора для 3D сцен имеют различные применения в задачах компьютерного зрения и научной визуализации, для сокращения размеров обучающей выборки при 3D реконструкции свойств материалов. В данной работе предлагается метод для выбора ракурсов (положений камеры) в приложении к задаче реконструкции материалов, основанный на обратном рендеринге с использованием модифицированных оценок качества ракурса в качестве функции потерь. Предложенный метод получает ракурсы только на основе геометрических характеристик 3D объекта и ограничений на положения камеры и не требует наличия некоторого заранее известного набора ракурсов. Предложенный метод протестирован на открытом наборе данных для 3D реконструкции.

***Ключевые слова*:** обратный рендеринг, 3D реконструкция, выбор ракурса, дифференцируемый рендеринг

Аннотация. Viewpoint selection methods for 3D scenes have various applications in computer vision, scientific visualization tasks, and in 3D reconstruction to reduce the training sample size. In this paper, we propose a method for viewpoint selection in application to material reconstruction problems based on inverse rendering using modified viewpoint quality scores as a loss function. The proposed method obtains viewpoints based only on the geometric characteristics of the 3D object and constraints on the camera positions and does not require some pre-known set of viewpoints. The proposed method is tested on an open dataset for 3D reconstruction.

***Ключевые слова*:** inverse rendering, 3D reconstruction, viewpoint selection, differentiable rendering.

I. Введение

1. Задача реконструкции 3D сцены

Задача реконструкции трехмерной сцены состоит в том, что нужно получить её описание в некотором виде, имея на входе набор изображений реальной сцены. На рис.1 показан пример реконструкции моделей материалов по фотографии. Описание 3D сцены в задаче реконструкции может быть задано как в «классическом» для компьютерной графике виде (полигональные сетки для геометрии объектов, текстуры и скалярные значения для моделей материалов и источников света и пр.), так и с использованием таких представлений как поля освещенности [1], т. н. «gaussian splats» [2] и другие. Подобные методы, однако, требуют большого объема памяти для работы и хранения результирующих сцен. Для работы с классическим представлением сцены используют подход обратного рендеринга, в частности, дифференцируемого рендеринга, основанного на градиентной оптимизации [3, 4].

Чтобы точно реконструировать свойства материала некоторого объекта реального мира, обычно необходимо рассматривать параметры модели материала как меняющиеся в пространстве, т. к. поверхность большинства реальных объектов будет обладать различными особенностями (повреждением, неравномерным нанесением материала и пр.). Другими словами, часто требуется реконструировать свойства материалов как текстуры. Чтобы реконструировать текстуру полностью, система обратного рендеринга должна получить на вход набор ракурсов целевого объекта, который бы покрывал все или почти все его части — полигоны его полигональной сетки.

Существуют методы, которые акцентируются на использовании 1 изображения (см. обзор [5]) целевого объекта для реконструкции. Однако, очевидно, что свойства материала на невидимых частях объекта в таком

случае, будут отсутствовать или «додумываться» нейросетью. Например, работа [6] использует диффузионную модель для получения изображений объекта с любого ракурса на основе изображения.

Методы, использующие несколько изображений, обычно рассматривают некоторое произвольное количество изображений сцен, расположенных примерно по полусфере вокруг целевого объекта или полученных с использованием поворотного столика. Это можно увидеть на примере наборов данных для сравнения и тестирования методов реконструкции [7-9]. Однако, использование большого числа изображений для реконструкции приводит к значительным затратам времени на реконструкцию. Методы на основе дифференцируемого рендеринга могут работать несколько часов, – например, авторы [4] говорят о двух часах работы, что подтверждается и проведенными в рамках данной работы экспериментами. Стоит так же отметить, что реконструкция на основе дифференцируемого рендеринга демонстрирует результаты, превосходящие по качеству методы на основе полей освещенности, что подтверждается экспериментальными данными, представленными в [8]. В данной работе рассматривается реконструкция свойств материалов на основе дифференцируемого обратного рендеринга.

Рис. 1 Пример реконструкции материалов. Для каждого из объектов слева – фото, справа – синтезированное изображение с реконструированными текстурами. Фото, 3D модели и HDR-панорамы из набора данных [7].

2. *Задача выбора ракурса и существующие решения*

Задача выбора ракурса встает при необходимости выполнить визуализацию или съемку (фотографирование, сканирование и т.п.) некоторого объекта, так, чтобы результат был наиболее «хорошим» с точки зрения его дальнейшего использования. Например, при подготовке изображений некоторого экспоната для виртуального музея [10] нужно выбрать такие ракурсы объекта, которые в максимально полной мере продемонстрируют его форму и детали наблюдателю.

В работе [11] было предложено использование мер оценки качества ракурса для того, чтобы из заранее известного набора изображений объекта выбрать те, которые будут наиболее информативны с точки зрения задачи реконструкции текстур, и на которых суммарно будут видны все части (полигоны) реконструируемого объекта. Полученные авторами результаты подтверждают эффективность мер оценки качества ракурса для задачи реконструкции.

Область исследований, посвященная выбора ракурса на основе некоторой меры качества является достаточно проработанной. Работы [12, 13] посвящены их обзору существующих оценок качества ракурсов. Авторы этих исследований вводят классификацию, основывающуюся на том, какие данные, используются для вычисления оценок. Исходными данными могут выступать площадь полигонов, силуэт, глубина, кривизна, семантическая информация, салиентность. Наибольшее развитие получили методы, использующие формализм теории информации для вычисления таких величин, как энтропия (например, [10-11, 14-15]) и взаимная информация [16] на основе перечисленных выше геометрических данных. Существуют также подходы на основе глубинного обучения [17, 18].

Одной из проблем является то, что лишь немногие подходы предлагают алгоритм для последовательного выбора нескольких ракурсов, который при этом может быть ограничен в количестве выбираемых ракурсов. Например, в [10] выбор осуществляется по одному ракурсу из квадранта декартовой системы координат, построенной относительно выравненного по осям целевого объекта.

Другая проблема существующих подходов состоит в том, что они отталкиваются от некоторого набора изображений 3D объекта с разных ракурсов, среди которых и происходит выбор с помощью той или иной оценки качества. Такой набор изображений может быть известен заранее [11], или же получен в ходе работы метода путем рендеринга целевого 3D объекта с позиционированием виртуальной камеры в некоторых предопределенных точках. Это могут быть точки, равномерно распределенные по поверхности сферы, построенной вокруг объекта [10, 14-16]. Или же вершины икосаэдра [17], который затем последовательно тесселируется (т. е. также до некоторой аппроксимации сферы). Использование предопределенных ракурсов приводит к тому, что оптимальные положения камеры могут быть упущены. Кроме того, в этом случае

ограничены возможности позиционирования камеры, — невозможно промоделировать все возможные положения и обычно, допускается лишь вращение вокруг целевого объекта на фиксированном расстоянии.

Таким образом, представляет интерес задача выбора ракурсов с произвольными ограничениями на позиционирование камеры и без использования фиксированного набора ракурсов, из которого осуществляется выбор. При выполнении реконструкции для некоторого нового 3D объекта это позволило бы сэкономить время не только на реконструкцию, но и на получение самого набора изображений для реконструкции, т. е. на фотографирование объекта. Кроме того, при обычном подходе к съемке объекта (по полусфере, с помощью поворотного столика) какие-то его части могут так и не попасть в кадр.

II. Постановка задачи

Задачей данного исследования является разработка метода получения набора ракурсов 3D объекта для дальнейшего использования в задаче реконструкции его материалов.

На входе имеется геометрическая модель 3D объекта, заданная полигональной сеткой из n вершин и k треугольников:

$\{V_i\}, V \in \mathbb{R}^3, i = 1 \dots n$ – координаты вершин треугольников,

$\{N_i\}, N \in \mathbb{R}^3, i = 1 \dots n$ – векторы нормалей в вершинах,

$\{T_i\}, T \in \mathbb{R}^2, i = 1 \dots n$ – текстурные координаты в вершинах,

$\{I_i\}, I \in \mathbb{N}, i = 1 \dots k$ индексов вершин I, формирующих треугольники.

Также могут быть заданы ограничения на положение и ориентацию виртуальной камеры.

На выходе метод выдает набор квадратных матриц $\{M_i^{4\times4}\}, i = 1 \dots m$, каждая из которых характеризует положение и ориентацию камеры для выбранного ракурса.

Отметим, что текстурные координаты $\{T_i\}$ могут быть получены для геометрического объекта как автоматически, так и в полуавтоматическом или ручном режиме с использованием 3D редактора. При этом это по сути лишь некоторое, в общем случае произвольное, отображение 3D координат вершин на 2D пространство изображения (текстуры), содержащего пространственно распределенные значения оптических свойств. Таким образом, наличие текстурных координат не нарушает условий исходной задачи поиска ракурсов только по геометрической информации. Если при реконструкции целевого объекта предполагается, что его оптические свойства постоянны для всего объекта, то текстурные координаты не нужны.

III. Теория

1. Общая схема работы

В основе предлагаемого подхода лежит использование дифференцируемого обратного рендеринга, общая схема работы представлена на рис. 2.

Создается 3D сцена, содержащая целевой 3D объект, расположенный в начале координат, начальное положение виртуальной камеры задается некоторым образом с учетом заданных ограничений. Это может быть, например, случайное положение или же положение, выбранное из некоторых предопределенных положений на полусфере. Для целевого 3D объекта также рассчитываются дополнительные геометрические атрибуты, необходимые для расчета оценки качества ракурса.

Далее идет **этап дифференцируемого рендеринга** (т.е. рендеринга с вычислением градиентов по оптимизируемым параметрам – положению и ориентации камеры) созданной 3D сцены. Результатом рендеринга является **визуализация атрибута** (или атрибутов) геометрической модели целевого объекта, значения которого используются для вычисления оценки качества ракурса. Например, в качестве значения пикселя может быть выведено значение площади треугольника целевого объекта, видимого в данном пикселе. Из результирующего изображения вычисляется **оценка качества ракурса**, которая используется для вычисления функции потерь. Значение функции потерь передается оптимизатору и на этом одна итерация процесса обратного рендеринга завершается.

В качестве оптимизируемых параметров используются переменные, на основе которых вычисляется положение и ориентация виртуальной камеры. Например, это могут быть эйлеровы углы или же угол и ось вращения и т.д. Конкретные переменные зависят от моделируемой системы позиционирования камеры. После завершения процесса оптимизации на выходе получается **матрица трансформации** виртуальной камеры.

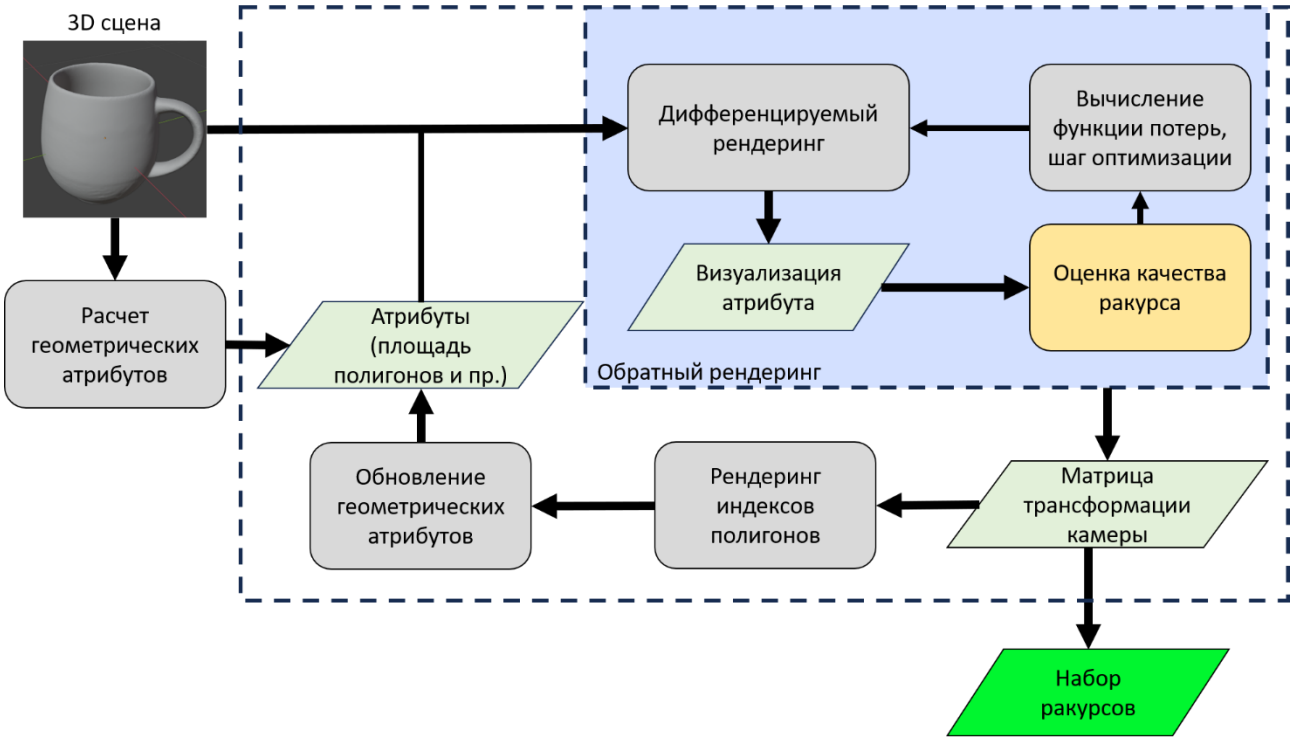


Рис. 2 Общая схема предлагаемого метода

Далее происходит **рендеринг индексов полигонов** для полученного положения камеры, на основе которого происходит **обновление геометрических атрибутов** 3D объекта, т.е. учитывается уже найденный ракурс для поиска следующего. Весь процесс повторяется до достижения некоторого условия остановки, которое будет рассмотрено позже. На выходе получается последовательность матриц преобразования камеры, определяющих найденные ракурсы.

Здесь и далее при рассмотрении предлагаемого метода будем считать, что 3D сцена содержит один целевой объект, т.к. такой сценарий наиболее распространен и представлен в большинстве работ и наборов данных в данной области исследований.. Если необходимо выполнить поиск ракурсов для сцены из нескольких объектов, то они могут быть объединены в один объект, например, путем конкатенации массивов атрибутов вершин и перерасчета массива индексов. В зависимости от используемой рендер-системы это может и не потребоваться, т.к. зачастую для уменьшения числа аллокаций и буферов на GPU такое объединение уже может происходить. Также предложенный метод может быть легко адаптирован и для случая, когда объединение объектов нежелательно. Для этого необходимо:

1. при вычислении оценки качества ракурса суммировать значения оценки по всем объектам;
2. выполнять обновление геометрических атрибутов также для всех объектов;
3. дополнительно синтезировать маски по объектам, которые потребуются для выполнения двух предыдущих пунктов.

2. Оценки качества ракурса

Как было рассмотрено ранее, существует много различных оценок для выбора ракурса. В работе [11], где также решалась задача выбора ракурса для реконструкции материалов объекта, использовались такие меры как энтропия спроецированной площади полигонов VE, взаимная информация ракурсов VMI и энтропия площади текстуры $VE_{texarea}$. При этом в целом лучше всего себя показала оценка VE, оценка $VE_{texarea}$ также показала хорошие результаты с незначительным отставанием и выиграла в некоторых случаях.

Поскольку предлагаемое решение основано на дифференцируемом рендеринге, оценка ракурса для её использования в функции потерь должна быть дифференцируема относительно параметров сцены.

Энтропия спроецированной площади полигонов VE вычисляется следующим образом:

$$VE(v) = -\sum_{i=0}^{N} \frac{a_z(v)}{a_t(v)} \log \frac{a_z(v)}{a_t(v)}$$

где N- число полигонов в 3D модели, $a_z(v)$ – спроецированная площадь полигона z для ракурса v и $a_t(v)$ – спроецированная площадь всей 3D модели для ракурса v.

При этом спроецированные площади полигонов вычисляются в соответствии с [15] как число всех пикселей на синтезированном изображении, принадлежащих конкретному полигону. То есть на этапе рендеринга в буфер кадра записывается индекс полигона, видимого в данном пикселе, после чего происходит подсчет числа пикселей для каждого из значений индекса. Подобный способ расчета, очевидно, приводит к тому, что оценка VE не удовлетворяет требованию дифференцируемости и не может быть использована в составе функции потерь.

Отметим также близкую по смыслу оценку ракурса, исходные данные для которой могут быть рассчитаны заранее, — отношение видимости (visibility ratio), предложенное в [19]:

$$VIS(v) = \frac{\sum_{z \in Z_v} A_z}{A_t} \quad (1)$$

где Z_v- множество полигонов 3D модели, видимых на ракурсе v, A_z – площадь полигона z, A_t – суммарная площадь всех полигонов модели.

Энтропия площади текстуры $VE_{texarea}$ задается как:

$$VE_{texarea}(v) = -\sum_{i=0}^{N_v} \frac{b_z}{b_t} \log \frac{b_z}{b_t} \quad (2)$$

где N_v- число полигонов в 3D модели, видимых на ракурсе v, b_z – площадь текстуры, используемая полигоном z, b_t – площадь текстуры, используемая всем полигонами 3D модели. Площадь текстуры, используемая полигоном, вычисляется как площадь соответствующего полигона в текстурном пространстве с использованием текстурных координат вершин полигона. Сами значения площади b_z и b_t не зависят от ракурса и могут быть рассчитаны заранее и сохранена как атрибут 3D объекта.

Меры VIS и $VE_{texarea}$ также не могут быть вычислены в таком виде для использования в функции потерь, т.к. содержат внутри себя процесс подсчета по видимым полигонам. Поэтому предлагается модификация – вместо суммирования по видимым полигонам производить суммирование по всем видимым пикселям. Зависимость от ракурса в таком случае будет учтена за счет того, что чем больше пикселей занимает полигон, тем больше его вклад в оценку. Использование формализма энтропии в данном случае теряет смысл, т.к. при расчете по всем видимым пикселям, а не полигонам, значение суммы отношений площадей, как в (1), так и в (2), может превысить единицу. Таким образом, выражения (1) и (2) приобретают следующий вид:

$$VIS'(v) = \frac{\sum_{p \in P_v} A_z}{A_t}$$

$$VE_{texarea}'(v) = \frac{\sum_{p \in P_v} b_z}{b_t}$$

где P_v- множество всех пикселей синтезированного изображения для ракурса v, в которых виден целевой 3D объект.

Значения площадей A_z и b_z будем предварительно рассчитывать исходя из пространственных и текстурных координат вершин соответственно, и сохранять в качестве атрибутов 3D модели.

Дополнительно так же рассмотрим наивную оценку качества ракурса, основанную на расчете площади видимой части объекта в пикселях. Т.е. чем больше пикселей в синтезированном изображении занимает целевой объект, тем лучше:

$$VA(v) = \sum_{p \in P_v} 1$$

Будем использовать полученные меры VIS', $VE_{texarea}'$, VA для поиска набора ракурсов. При этом только мера $VE_{texarea}'$ требует наличия у 3D объекта текстурных координат. Если при реконструкции допускается, что оптические свойства постоянны для всего объекта и не требуется использовать текстуру для их задания, то использование данной меры не имеет смысла.

4. Алгоритм выбора набора ракурсов

Поскольку поставленная задача требует получения нескольких ракурсов, необходимо составить алгоритм, который позволит после получения очередного ракурса, оптимального по той или иной оценке качества, получить следующий ракурс с учетом уже полученных. В работах [11, 15] предложен схожий алгоритм для решения этой задачи:

Алгоритм 1. Выбор N ракурсов из известного набора ракурсов

Вход:

V – множество всех ракурсов, *Z* – множество полигонов, видимых на ракурсах *V*, *measure* – используемая оценка качества ракурса

Выход:

M – выбранное подмножество ракурсов

```
1  Q ← ∅, M ← ∅
2  while Q ≠ Z do
3  |  measure_view ← ∅
4  |  forall v ∈ V do
5  |  |  forall z ∈ Z do
6  |  |  |  if z ∉ Q then
7  |  |  |  |  m ← ViewpointQuality(measure, v, z)
8  |  |  |  |  measure_view[v] ← m
9  |  |  |  end
10 |  |  end
11 |  end
12 |  view ← SelectBestView(measure, measure_view)
13 |  forall z ∈ Z do
14 |  |  if visible(z, view) then
15 |  |  |  Q ← z ∪ Q
16 |  |  end
17 |  end
18 |  M ← view ∪ M
19 end
20 return M
```

На вход алгоритм получает список видимых хотя бы на одном из ракурсов полигонов. После чего, для каждого из ракурса вычисляется выбранная оценка качества *measure* с помощью вызова $ViewpointQuality$ для всех видимых на нем непомеченных полигонов *z*. Затем выбирается ракурс с наилучшим значением оценки качества с помощью вызова $SelectBestView$. Полигоны, видимые на выбранном ракурсе, помечаются как посещенные, а сам ракурс добавляется в выходное подмножество *M*. Алгоритм завершает работу, когда были посещены все полигоны из списка видимых хотя бы на одном из ракурсов.

Этот алгоритм не подходит для задачи, решаемой в данной работе, поскольку в нашем случае отсутствует заранее определенный набор ракурсов, из которого осуществляется выбор. И, как следствие, мы не можем гарантировать, что все полигоны геометрической модели будут видимы на каком-то из ракурсов. В оригинальном алгоритме заранее известно, какие полигоны видимы на заданном наборе ракурсов и всегда можно найти такой поднабор ракурсов, который охватит все полигоны видимые на всем наборе. В нашем же случае может оказаться, что, например, геометрическая модель целевого 3D объекта достаточно сложная и детализированная, и какие-то полигоны сокрыты другими частями геометрии 3D объекта или же вообще расположены внутри него. Кроме того, предлагаемый метод подразумевает задание некоторых ограничений на положение и ориентацию камеры. И в рамках данных ограничений какие-то из полигонов могут не попасть в область видимости. В связи с этим предлагается следующий алгоритм для получения N ракурсов:

Алгоритм 2. Выбор N ракурсов на основе геометрической информации

Вход:

scene – описание 3D сцены, *attrib* – имя атрибута 3D модели, используемое для расчета оценки качества ракурса, *measure* – используемая оценка качества ракурса, $internal\_cond$ – условие остановки поиска одного ракурса, *weighted* – использовать или нет взвешивание результата по нормали, δ – изменение суммарного значения атрибута *attrib* по 3D модели, меньше которого процесса поиска ракурсов завершается, $learning\_rates$ – отображение оптимизируемого параметра на величину его темпа обучения.

Выход:

M – список матриц трансформации, описывающих положение камеры для выбранных ракурсов

```
1  c ← 0, idx ← 0, cond ← True, M ← List()
2  while cond do
3  |  cam_vars ← init_cam_vars(idx), loss ← 0, it ← 0
4  |  while internal_cond(it) do
5  |  |  apply_transform(cam_vars)
6  |  |  img ← render(scene)
7  |  |  q ← ViewpointQuality(img, measure)
8  |  |  loss ← 1/q
9  |  |  optimization_step(loss, learning_rates)
```

$it \leftarrow it + 1$
$update\_learning\_rate(learning\_rates, it)$ // опциональное изменение темпа обучения
end
$cam\_vars \leftarrow get\_optimized\_vars()$
$M \leftarrow M + view\_matrix(cam\_vars)$
$aov\_img \leftarrow render\_aov(scene)$
$previous\_total \leftarrow total\_attrib$
$total\_attrib \leftarrow update\_mesh\_attrib(aov\_img, weighted, attrib)$
if $\frac{|previous\_total - total\_attrib|}{init\_total} < \delta$ **then**
$cond \leftarrow False$
end
end
return M

Внутренний цикл алгоритма 2 соответствует блоку «Обратный рендеринг» на рис.2. В качестве условия $internal\_cond$ для завершения процесса поиска одного ракурса может выступать достижение некоторого максимального числа итераций или же, например, изменение функции потерь от шага к шагу. Выбор того или иного варианта и конкретных значений числа итераций или величины изменения зависит от заданных значений темпа обучения $learning\_rates$, которые в свою очередь зависят от способа и ограничений позиционирования камеры. Вызов $render$ соответствует блоку «Дифференцируемый рендеринг» на рис.2, $ViewpointQuality$ – блоку «Оценка качества ракурса», оставшаяся часть внутреннего цикла – блоку «Вычисление функции потерь, шаг оптимизации».

Выбор исходного положения камеры с помощью вызова $init\_cam\_vars$ перед началом поиска очередного ракурса, как было отмечено ранее, может быть осуществлен разными способами и зависит от ограничений и способа задания позиционирования камеры. Основное требование к способу задания исходного положения – обеспечить разные начальные точки для поиска решения (т.е. ракурса) на очередной итерации внешнего цикла. При возможности движения камеры по сфере вокруг объекта может использоваться последовательный выбор точек на этой сфере, максимально отстоящих друг от друга. Например, могут выбраны точки, для которых одна координата равна +/- *R*, а остальные равны 0, где *R* – расстояние от камеры до объекта. Камера при этом всегда «смотрит» на объект.

В алгоритме 2 после получения очередного ракурса и соответствующей матрицы преобразования для виртуальной камеры, производится обновление атрибута $attrib$ 3D модели, используемого для расчета оценки качества ракурса с помощью функции $update\_mesh\_attrib$ (блок «Обновление геометрических атрибутов» на рис.2). Обновление атрибутов происходит только для видимых полигонов, которые получается с помощью визуализации значения индекса полигона aov_img, получаемой специальной процедурой рендеринга $render\_aov$, чему соответствует блок «Рендеринг индексов полигонов» на рис.2.

Чтобы иметь возможность учесть ориентацию полигонов относительно камеры, используется параметр $weighted$. Если он имеет значение *False*, то для всех полигонов, видимых на полученном ракурсе, атрибут $attrib$ получает значение 0. Таким образом, увиденные полигоны просто исключаются из расчета. Если же $weighted$ имеет значение *True*, то значение атрибута $attrib$ уменьшается на текущее значение атрибута взвешенное по углу между вектором нормали $\bar{N}$ к этому полигону и направлением на виртуальную камеру $\bar{V}$:

$$mesh[attrib] -= mesh[attrib] * \max(0, (\bar{N} \cdot \bar{V}))$$

В этом случае полигоны, которые видны из виртуальной камеры под углом не исключаются, а обладают меньшим весом для вычисления оценки качества следующего ракурса.

Именно за счет обновления значений атрибутов ($update\_mesh\_attrib$) ракурс, найденный в результате работы внутреннего цикла, учитывается в общем процессе поиска N ракурсов.

Как было уже сказано, мы не можем гарантировать, видимость всех полигонов на совокупности полученных ракурсов. Поэтому в качестве условия остановки предлагается использовать абсолютное значение уменьшения суммарного значения атрибута $attrib$ относительно предыдущей итерации. Если оно оказывается достаточно мало, например, менее 0.5%, то алгоритм завершает работу.

IV. Результаты экспериментов

1. Исходные данные

Для проведения экспериментальной оценки предложенного метода выбора ракурсов использовались данные из набора данных Stanford-ORB [7]. Данный набор данных содержит:

- фотографии 14 объектов в узком и широком динамическом диапазоне,
- эталонные 3D модели объектов, полученные с помощью 3D сканирования,
- HDR-панорамы освещения,
- текстуры объектов для модели материала Principled BSDF – альбедо, шероховатость и «металличность».

Рис.3 Слева – 3D объекты из набора данных [7], на которых проводилось тестирование. Справа – пример выбранных ракурсов для одного из объектов с использованием предложенного метода.

Текстуры были получены авторами при помощи дифференцируемого рендеринга по фотографиям сделанных в специальных условиях освещения, поэтому нельзя считать их в полной мере истинными, будем называть их «псевдоэталонными».

Часть объектов в наборе данных имеют схожую геометрическую форму (половину набора данных составляет посуда цилиндрической формы), но разные материалы. Поскольку рассматриваемые в работе меры оценки качества ракурса основаны на геометрических характеристиках, из набора данных были отобраны отдельные объекты, отличающиеся именно по форме. Выбранные объекты и пример полученных с помощью предложенного метода ракурсов для одного из них представлены на рис.3.

2. *Общая схема проведения экспериментов*

Описанный алгоритм и тестовая система обратного рендеринга были реализованы на базе рендер-системы Mitsuba3 [20]. Предварительный расчет нужных атрибутов геометрической модели (значений площадей полигонов и площадей текстуры, занимаемых полигонами), обновлений их значений в процессе работы разработанного алгоритма 2 и вычисление оценок качества ракурса были реализованы на GPU с использованием JIT-компилятора Dr.Jit [21].

В качестве базового метода рассматривалась реконструкция произведенная на ракурсах, содержащихся в обучающей выборке, представленной в исходном наборе данных [7] (блок «Обучающая выборка» на рис. 4 и строка «обучающая выборка» в таблицах 1 и 2). Таким образом, левая часть схемы на рис. 4 соответствует базовому методу, а правая – предложенному.

Для каждого из отобранных геометрических объектов эксперименты проводились следующим образом (рис. 4):

1. Вычисление атрибутов геометрического объекта и исполнение алгоритма 2 (на рис.4 отмечено зеленым цветом) для 3-х рассмотренных в разделе 3.2 оценок качества ракурса без взвешивания по косинусу угла между нормалью и направлением на наблюдателя и со взвешиванием, - всего 6 проходов.
2. Прямой рендеринг найденных на предыдущем этапе ракурсов с числом выборок на пиксел равным 2048 и с использованием псевдоэталонных текстур из исходного набора данных.
3. Реконструкция текстур альбедо, шероховатости и металличности для модели материала “principled” рендер-системы Mitsuba3 с помощью обратного рендеринга с использованием изображений с предыдущего шага в качестве эталона.
4. Прямой рендеринг с использованием реконструированных текстур на тестовых ракурсах из исходного набора данных (тестовая выборка 1) и на 16 случайных ракурсах (тестовая выборка 2).

5. Оценка качества реконструкции по метрикам PSNR, SSIM, LPIPS, рассчитанным между изображениями полученными на этапе тестового рендеринга и эталонными изображениями из тестовых выборок по бинарной маске объекта.

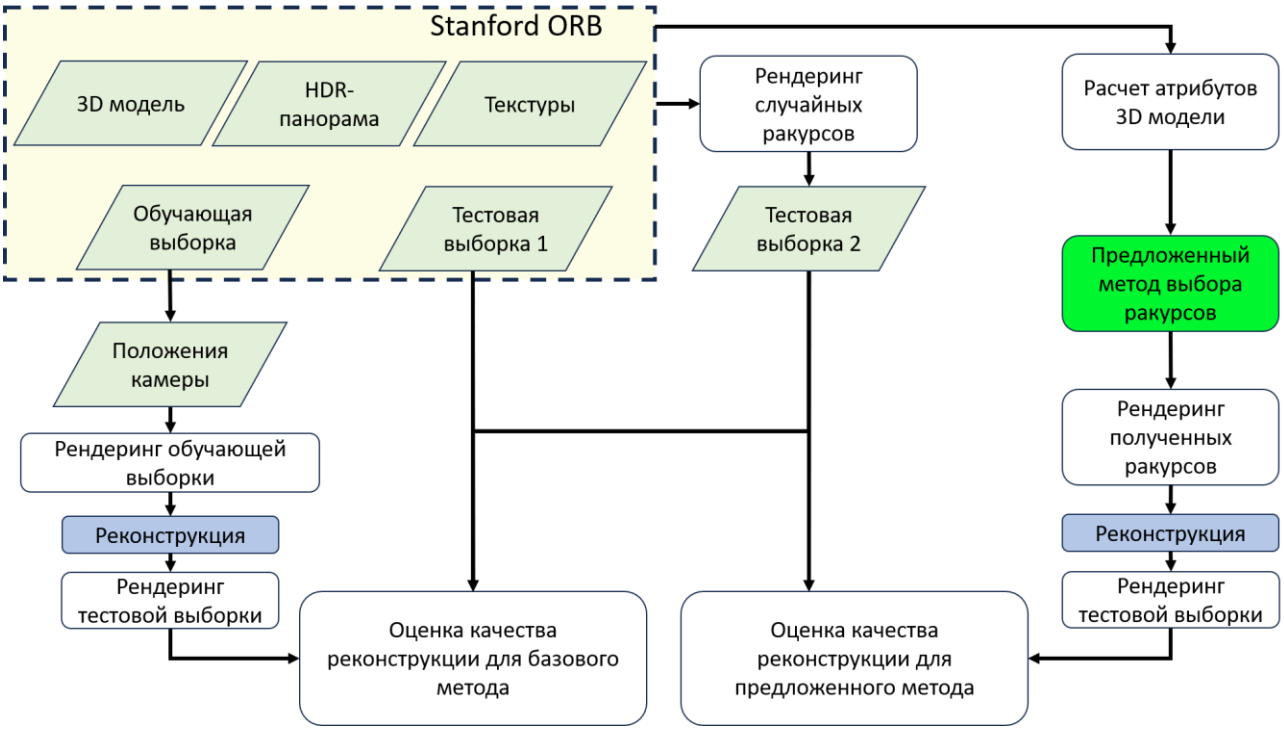


Рис. 4 Схема проведения численных экспериментов.

Также для каждого из отобранных объектов аналогичным образом была проведена реконструкция тех же текстур с использованием обучающих ракурсов из исходного набора данных (фото) и изображений с такими же ракурсами, но полученных путем прямого рендеринга с использованием псевдоэталонных текстур из исходного набора данных. После чего реконструированные таким образом текстуры аналогичным образом использовались для выполнения тестового рендеринга и расчета метрик.

3. *Результаты*

В основной части экспериментов для задания положения камеры допускалось её перемещение по сфере вокруг целевого объекта с возможностью небольшого приближения или удаления. Результаты (таблица 1) получены по описанной в предыдущем подразделе схеме – по алгоритму 2 вычислены ракурсы с помощью той или иной оценки качества (столбец «Способ получения ракурсов») с и без взвешивания по косинусу угла между нормалью и направлением на наблюдателя (что соответствует параметру weighted в алгоритме 2). Далее выполнялась реконструкция свойств материала с использованием изображений целевого объекта на этих ракурсах и оценивалось качество реконструкции. Значения SSIM рассчитывались для изображений с широким динамическим диапазоном, а PSNR и LPIPS для изображений с узким динамическим диапазоном.

Также в составе этих же экспериментов были проведены замеры времени подбора ракурсов по разным оценкам, времени реконструкции и числа подобранных ракурсов (таблица 2).

ТАБЛИЦА 1

Качество реконструкции на ракурсах, полученных разными методами. Для метрик показаны средние значения и стандартное отклонение по всем 3D объектам

Способ получения ракурсов	**Взвешивание**	**Тестовая выборка 1 (фото из Stanford-ORB [7])**			**Тестовая выборка 2 (случайные ракурсы)**		
		PSNR↑	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
Обучающая выборка	-	30.719 ±2.769	0.972 ±0.012	0.053 ±0.025	34.511 ±2.992	0.977 ±0.012	0.048 ±0.026
VA	Да	30.626 ±2.506	0.969 ±0.014	0.058 ±0.029	34.388 ±1.993	0.978 ±0.008	0.047 ±0.020
VA	Нет	30.474	0.968	0.064	33.895	0.977	0.052

		±2.436	±0.015	±0.031	±2.210	±0.009	±0.020
VIS'	Да	**30.854 ±2.458**	**0.972 ±0.013**	**0.055 ±0.022**	34.535 ±1.825	0.979 ±0.009	0.048 ±0.021
VIS'	Нет	30.521 ±2.316	0.970 ±0.014	0.059 ±0.024	33.412 ±2.199	0.975 ±0.010	0.054 ±0.023
$VE_{texarea}{}'$	Да	30.726 ±2.452	**0.972 ±0.013**	**0.055 ±0.022**	**34.855 ±1.973**	**0.980 ±0.009**	**0.046 ±0.019**
$VE_{texarea}{}'$	Нет	30.700 ±2.502	0.971 ±0.013	0.057 ±0.022	34.010 ±2.087	0.977 ±0.012	0.050 ±0.021

ТАБЛИЦА 2

Среднее время реконструкции на ракурсах, полученных разными методами. Замеры производились на GPU Nvidia RTX 2070 Super

Способ получения ракурсов	**Взвеши-вание**	**Число ракурсов**	**Время подбора ракурсов, мин.**	**Время реконструкции, мин.**	**Суммарное время, мин.**
Обучающая выборка	-	59-60	0.00	130.29	130.29
VA	Да	16-20	5.81	24.72	30.54
VA	Нет	13-16	4.29	18.04	22.33
VIS'	Да	12-19	4.64	19.11	23.75
VIS'	Нет	8-10	2.71	11.83	14.53
$VE_{texarea}{}'$	Да	15-21	5.51	23.64	29.14
$VE_{texarea}{}'$	Нет	11-14	4.03	17.40	21.43

На рис.5 представлен пример ракурсов, подобранных предложенным методом, для разных ограничений на положение камеры. Для примера справа на рисунке было разрешено лишь движение камеры вокруг объекта, что эквивалентно вращению объекта на поворотном столике. А для примера слева – движение камеры допускалось по сфере вокруг объекта с небольшим смещением, что эквивалентно произвольному вращению объекта по трем осям координат и также небольшому смещению относительно камеры.

Рис. 5 Пример ракурсов, подобранных для чайника по оценке площади полигонов со взвешиванием по косинуса угла между нормалью и направлением на наблюдателя. Слева - для камеры допускается перемещение по сфере вокруг объекта и небольшое смещение по всем осям координат. Справа – камера двигается вокруг объекта.

V. Обсуждение результатов

Как видно из численных оценок, реконструкция на меньших по размеру наборах ракурсов, полученных с помощью предложенного метода позволяет сохранить качество результатов в сравнении с реконструкцией на обучающей выборке, полученной путем рендеринга 3D объекта с ракурсами из исходного набора данных [7] с использованием псевдоэталонных текстур. Все рассмотренные оценки качества ракурса показывают схожие результаты по качеству реконструкции. Использование взвешивание по косинусу между нормалью к полигону и направлением на наблюдателя ожидаемо приводит к увеличению числа выбранных ракурсов и, соответственно, небольшому приросту по качеству реконструкции.

Наименьшее число ракурсов относительно базового подхода, в среднем до 7.5 раз, предложенный метод выдает при использовании оценки ракурса VIS' на основе площадей полигонов без использования взвешивания.

Время реконструкции в этом случае уменьшается в среднем в ~8.9 раз. Наибольшее количество ракурсов получается при использовании взвешивания для любой из оценок ракурса (однако VIS' в среднем все же выдает меньше ракурсов), которое при этом меньше базовой выборки в 3 раза., чему соответствует уменьшение времени реконструкции в среднем в ~4.2 раза.

Хотя в большинстве случаев наилучшее качество оптимизации текстур соответствует наибольшему числу используемых ракурсов, что логично, тем не менее, ракурсы, полученные с использованием оценки $VE_{texarea}{}'$ в среднем дают большее качество реконструкции, чем простая оценка по числу видимых пикселей объекта VA.

При подборе ракурсов с разными ограничениями на положение камеры (рис. 5) качество реконструкции на тестовой выборке 1 оказывается лучше для случая, когда камера просто движется вокруг объекта (PSNR 29.906±1.065 против 29.674±0.969), т.к. условия съемки в наборе данных [7] такие же. Для тестовой выборки со случайными ракурсами ситуация противоположна (PSNR 34.410±1.572 против 36.769±1.092), что ожидаемо, т.к. в исходном наборе нет ракурсов, где камера смотрит на объект напрямую снизу или сверху. Для движения камеры по сфере вокруг объекта со смещением в процессе обратного рендеринга оптимизировалось 6 чисел – 3 значения углов поворота (по каждой из осей) и 3D вектор смещения. Для движения камеры вокруг объекта оптимизировалось только одно число – угол вращения вокруг заданной оси.

Отметим, что при использовании меньшего числа параметров, определяющих положение камеры (или целевого 3D объекта), необходимо аккуратнее подбирать параметры алгоритма – темп обучения, значение δ условия завершения работы алгоритма 1, инициализация начального положения, использование адаптивного темпа обучения (постепенное уменьшение в процессе обратного рендеринга) становится практически необходимым. Иначе алгоритм выбора ракурса может «увязнуть» в одной позиции, выбирать одинаковые ракурсы и завершить работу, не исследовав всего пространства. Например, на предыдущих шагах были выбраны ракурсы с поворотом в 40° и 60°, и подбор следующего ракурса начался со значения в 50°. Вероятно, что ранее выбранные ракурсы уже покрывают все полигоны объекта и поворот в 50° уже не является хорошим с точки зрения оценки ракурса, но любые шаги в ту или другую сторону от этого значения будут давать ещё худшие ракурсы.

VI. Выводы и заключение

Предложенный метод выбора ракурсов на основе только геометрической информации без использования заранее известного или полученного равномерным позиционированием камеры набора ракурсов показал свою применимость для поставленной задачи. Метод позволяет получить набор ракурсов за ~3-6 минут в зависимости от используемой меры оценки качества ракурса, что значительно меньше, суммарного времени реконструкции материала (на базовом обучающем наборе данных – более двух часов). А меньший размер набора ракурсов позволяет значительно уменьшить время реконструкции.

Метод позволяет задавать различные ограничения на положение камеры при подборе ракурса, в зависимости от планируемых условий съемок реального объекта. Например, возможно моделировать распространенное решение с использованием поворотного столика или же практически произвольное позиционирование камеры вокруг объекта.

В качестве направления дальнейших исследований представляет интерес использованием предложенного метода для сравнения различных условий съемки 3D объектов с точки зрения качества реконструкции. А также исследование других мер оценки качества ракурса для использования с предложенным методом.

Список литературы

1. Mildenhall B. et al. Nerf: Representing scenes as neural radiance fields for view synthesis //Communications of the ACM. – 2021. – Т. 65. – №. 1. – С. 99-106. DOI: 10.1145/3503250

2. Kerbl B. et al. 3d gaussian splatting for real-time radiance field rendering //ACM Transactions on Graphics. – 2023. – Т. 42. – №. 4. – С. 1-14. DOI: 10.1145/3592433

3. Zhao S., Jakob W., Li T. M. Physics-based differentiable rendering: from theory to implementation //ACM siggraph 2020 courses. – 2020. – С. 1-30. DOI:10.1145/3388769.3407454

4. Luan F. et al. Unified shape and svbrdf recovery using differentiable monte carlo rendering //Computer Graphics Forum. – 2021. – Т. 40. – №. 4. – С. 101-113. DOI: 10.1111/cgf.14344

5. Fu K. et al. Single image 3D object reconstruction based on deep learning: A review //Multimedia Tools and Applications. – 2021. – Т. 80. – №. 1. – С. 463-498. DOI: 10.1007/s11042-020-09722-8

6. Melas-Kyriazi L. et al. Realfusion: 360deg reconstruction of any object from a single image //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2023. – С. 8446-8455. DOI: 10.1109/CVPR52729.2023.00816
7. Kuang Z. et al. Stanford-ORB: a real-world 3d object inverse rendering benchmark //Advances in Neural Information Processing Systems. – 2024. – Т. 36. DOI: 10.48550/arXiv.2310.16044
8. Ummenhofer B. et al. Objects With Lighting: A Real-World Dataset for Evaluating Reconstruction and Rendering for Object Relighting //2024 International Conference on 3D Vision (3DV). – IEEE, 2024. – С. 137-147. DOI: 10.1109/3DV62453.2024.00097
9. Voynov O. et al. Multi-sensor large-scale dataset for multi-view 3D reconstruction //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2023. – С. 21392-21403. DOI:10.1109/CVPR52729.2023.02049
10. S. Zeng, G. Geng, M. Zhou, Automatic representative view selection of a 3d cultural relic using depth variation entropy and depth distribution entropy, Entropy 23 (2021) 1561. DOI: 10.3390/e23121561
11. Sanzharov V., Frolov V. Viewpoint selection for texture reconstruction with inverse rendering // GraphiCon 2023: труды 33-й Международной конференции по компьютерной графике и машинному зрению. — Москва: ИПМ им. М.В. Келдыша РАН, 2023. — P. 66–78 DOI: 10.20948/graphicon-2023-66-77
12. Bonaventura X. et al. A survey of viewpoint selection methods for polygonal models //Entropy. – 2018. – Т. 20. – №. 5. – С. 370. DOI: 10.3390/e20050370
13. Zhang Y., Fei G. Overview of 3D scene viewpoints evaluation method //Virtual Reality & Intelligent Hardware. – 2019. – Т. 1. – №. 4. – С. 341-385. DOI: 10.1016/j.vrih.2019.01.001
14. Neumann L. et al. Viewpoint quality: Measures and applications //Proceedings of the 1st Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging. – Aire-la-Vile : The Eurographics Association Press, 2005. – С. 185-192.
15. Vázquez P. P. et al. Automatic view selection using viewpoint entropy and its application to image-based modelling //Computer Graphics Forum. – Oxford, UK and Boston, USA : Blackwell Publishing, Inc, 2003. – Т. 22. – №. 4. – С. 689-700. DOI:10.1111/j.1467-8659.2003.00717.x
16. Feixas M., Sbert M., González F. A unified information-theoretic framework for viewpoint selection and mesh saliency //ACM Transactions on Applied Perception (TAP). – 2009. – Т. 6. – №. 1. – С. 1-23. DOI:10.1145/1462055.1462056
17. Song R. et al. Unsupervised multi-view CNN for salient view selection and 3D interest point detection //International Journal of Computer Vision. – 2022. – Т. 130. – №. 5. – С. 1210-1227. DOI: 10.1007/s11263-022-01592-x
18. Zhou W., Jia J. Training deep convolutional neural networks to acquire the best view of a 3D shape //Multimedia Tools and Applications. – 2020. – Т. 79. – №. 1. – С. 581-601. DOI: 10.1007/s11042-019-08107-w
19. Plemenos D., Benayada M. Intelligent display techniques in scene modelling. new techniques to automatically compute good views //Proceedings of the International Conference GraphiCon'96. – Saint Petersburg (Russia). – 1996.
20. W. Jakob, S. Speierer, N. Roussel, M. Nimier-David, D. Vicini, T. Zeltner, B. Nicolet, M. Crespo, V. Leroy, Z. Zhang, Mitsuba 3 renderer, 2023. URL: https://mitsuba-renderer.org.
21. Jakob W. et al. Dr. JIT: A just-in-time compiler for differentiable rendering //ACM Transactions on Graphics (TOG). – 2022. – Т. 41. – №. 4. – С. 1-19. DOI: 10.1145/3528223.3530099

УДК 533.9.01
DOI: 10.25206/978-5-8149-3873-2-2024-96-104

Оптимизация численных методов моделирования переноса излучения в газоразрядной СВЧ ртутно-аргоновой плазме
Optimization of numerical methods for modeling radiation transfer in a gas-discharge microwave mercury-argon plasma

И. И. Железнов1, В. П. Будак2
1 Национальный исследовательский университет «МЭИ», Москва, Россия
2 Национальный исследовательский университет «МЭИ», Москва, Россия

I. I. Zheleznov1, V. P. Budak2
[1]NRU "MPEI", Moscow, Russia
[2]NRU "MPEI", Moscow, Russia

Аннотация. В данной статье представлен математический анализ механизмов переноса излучения в неравновесной низкотемпературной газоразрядной плазме, выполненный на примере модели ртутно-аргонового СВЧ разряда низкого давления. Авторами предложен метод решения системы кинетических уравнений и переноса излучения, основанный на кубической дискретизации источника излучения, представляющий численное решение уравнения Бибермана-Холстейна. Для проверки метода использована геометрия замкнутого коаксиального цилиндра с точечным источником возбуждения в центре, имитирующая излучение безэлектродной СВЧ бактерицидной лампы. Ключевыми результатами исследования являются получение решения для произвольного контура спектральной линии и учет влияния процессов диффузии, пленения и реабсорбции. Данные результаты способствуют углублению понимания процессов переноса излучения в плазме и открывают новые возможности для дальнейших исследований в этой области.
***Ключевые слова*:** перенос излучения, газоразрядная плазма.

Аннотация. This article presents a mathematical analysis of the mechanisms of radiation transfer in a nonequilibrium low-temperature gas-discharge plasma, performed using the example of a low-pressure mercury-argon microwave discharge model. The authors propose a method for solving a system of kinetic equations and radiation transfer based on cubic discretization of the radiation source, representing a numerical solution of the Biberman-Holstein equation. To test the method, the geometry of a closed coaxial cylinder with a point excitation source in the center was used, simulating the radiation of an electrodeless microwave bactericidal lamp. The key results of the study are to obtain a solution for an arbitrary contour of the spectral line and to take into account the influence of diffusion, capture and reabsorption processes. These results contribute to a deeper understanding of the processes of radiation transfer in plasma and open up new opportunities for further research in this area.

***Ключевые слова*:** radiation transfer, gas discharge plasma.

I. Введение

Концепция моделирования переноса излучения в неравновесных средах, таких как низкотемпературная газоразрядная плазма, является одной из наиболее сложных и многогранных задач в современной физике плазмы и светотехнике. В начале XX века были заложены фундаментальные теоретические основы для моделирования этого процесса, что привело к разработке первых численных методов и алгоритмов для решения уравнений переноса излучения. Одним из ключевых вкладов в эту область стала работа С. Чандрасекара по радиационному переносу, в которой были предложены методы решения уравнений переноса излучения в различных неравновесных средах [1]. Эта работа заложила основу для последующих исследований и стимулировала развитие новых методов и подходов, учитывающих сложные взаимодействия излучения с веществом в условиях, отличных от термодинамического равновесия.

В 1980-х и 1990-х годах развитие вычислительных технологий значительно расширило возможности численного моделирования переноса излучения. В частности, работы [2] по применению метода Монте-Карло к задачам радиационного переноса стали важным этапом в этом направлении. Эти исследования показали, что метод Монте-Карло может эффективно применяться для моделирования сложных процессов переноса излучения в средах с высокой оптической плотностью и сложной геометрией. В то же время дискретно-ординатные методы, представленные в [3], продемонстрировали свою эффективность в решении уравнений переноса излучения в условиях, требующих учета анизотропии рассеяния и пространственной зависимости излучения.

С начала 2000-х годов исследования сосредоточились на повышении точности и эффективности численных методов для решения задач переноса излучения. Были разработаны такие подходы, как адаптивные сетки и методы высокого порядка, которые значительно сократили вычислительные затраты при сохранении высокой точности моделирования [4, 5]. Адаптивные сетки позволяют динамически изменять разрешение в зависимости от локальных особенностей среды, что особенно важно для задач, требующих высокого разрешения в областях с резкими градиентами температуры или плотности. Методы высокого порядка, в свою очередь, существенно снижают численные ошибки при решении уравнений переноса излучения, обеспечивая более точные результаты при тех же вычислительных затратах.

Несмотря на значительные достижения, остаются нерешенные проблемы, связанные с корректным учетом взаимодействия излучения с веществом, особенно в условиях, где неравновесные процессы, такие как диффузия, пленение и реабсорбция, играют ключевую роль. Современные модели часто основываются на приближенных методах и упрощениях, что может приводить к значительным погрешностям в расчетах [6, 7]. В частности, многие модели не учитывают сложные взаимосвязи между различными процессами в неравновесной плазме, что может приводить к искажению интерпретации результатов. Более того, необходимость учета широкого диапазона временных и пространственных масштабов требует разработки новых методов и подходов, способных обеспечить как высокую точность, так и сокращение времени вычислений.

II. Постановка задачи

Таким образом, основная задача настоящей работы состоит в разработке и улучшении численных методов решения уравнения переноса излучения, которые могут учитывать взаимодействия между излучением и веществом в условиях неравновесных процессов, таких как диффузия, пленение и реабсорбция. Важно сократить вычислительные затраты при сохранении высокой точности моделирования и преодолеть ограничения текущих моделей, которые часто опираются на приближенные методы и упрощения.

III. Теория

В качестве объекта исследования выбран ртутно-аргоновый СВЧ разряд низкого давления в цилиндрической геометрии. В условиях локального термодинамического равновесия (ЛТДР) можно принять, что механизм переноса излучения практически не оказывает влияния на изменение распределения температуры внутри цилиндра радиусом R:

$$T(\zeta) = T_w + (T_0 - T_w)(1-\zeta)^{\beta}, \tag{1}$$

где $\zeta = r/R$, $1.5 < \beta < 4$, r - расстояние между частицами, $T_W = 1000$°K – температура на стенке цилиндра, $T_0 =$ 5500 … 6000°K – температура на оси цилиндра.

Данное утверждение справедливо для давлений, соответствующих приблизительно 0.05 Торр для ртути и 0,1…0,7 Торр для аргона, при условии, что радиус разрядной трубки составляет 1…3 см. Отметим, что применительно к смеси легких и тяжелых частиц в газах, ЛТДР предполагает, что температура, давление и прочие термодинамические параметры одинаковы в каждой малой области пространства, даже если система в целом не находится в термодинамическом равновесии [8, 9, 10].

Тогда, уравнение переноса излучения (УПИ) в случае ЛТДР примет вид:

$$(\hat{\mathbf{l}}, \nabla) L(\mathbf{r}, \hat{\mathbf{l}}) = -\kappa(\rho) L(\mathbf{r}, \hat{\mathbf{l}}) + \kappa(\rho) B(\rho) \tag{2}$$

где $L(\mathbf{r}, \hat{\mathbf{l}})$ - спектральная плотность энергетической яркости, ρ – расстояние до точки $\mathbf{r}$ от оси цилиндра в плоскости перпендикулярной оси, $\kappa(\rho)$ – коэффициент поглощения, $B(\rho)$ – объемная плотность излучения черного тела с температурой на расстоянии ρ от оси.

Представим перенос излучения на основе двухуровневой системы. Пусть n_1, n_2 – концентрации атомов на 1 и 2 уровнях.

Тогда в рассматриваемой системе на излучающем уровне 2 происходит образование и разрушение атомов вследствие следующих процессов: столкновения I и II рода с электронами, оптические переходы, которые завершаются или начинаются на этом уровне, а также диффузия атомов к стенкам сосуда.

С учетом вышеизложенного, представим основную систему уравнений в следующей форме [11]:

$$\begin{cases} Z_{21} n_2 = A_{21} n_2 - w_{21} n_2 + w_{12} n_1 + \nabla(D_2 \nabla n_2) + \dfrac{1}{4\pi} \int \oint L_\nu(\mathbf{r}, \hat{\mathbf{l}}) \left[B_{12} n_1 - B_{21} n_2 \right] P(\nu) d\hat{\mathbf{l}} d\nu, \\ (\hat{\mathbf{l}}, \nabla) L_\nu(\mathbf{r}, \hat{\mathbf{l}}) = \kappa_\nu(\mathbf{r}) B_\nu(\mathbf{r}) - \kappa_\nu(\mathbf{r}) L_\nu(\mathbf{r}, \hat{\mathbf{l}}) \end{cases} \tag{3}$$

где Z_{21} - частоты реакций, соответствующие столкновительно-радиационным процессам заселения уровня 2, $P(\nu)$ - функция контура спектральной линии излучения, A_{21}, B_{12} и B_{21} - коэффициенты Эйнштейна, D_2 – коэффициент диффузии атомов сорта 2, $\hat{\mathrm{I}}$ - единичный вектор, w_{12} и w_{21} - вероятности столкновений атомов при ударном возбуждении и тушении. Представим, что произвольная область объёма *V* разбивается на *M* подобъёмов $V = \sum_{i=1}^{M} \Delta V_i$, где ΔV_i – кубические элементы (воксели) декартовой 3D-сетки. В таком случае расстояние между точками **r** и **r**′ будет определяться как:

$$\rho = |\mathbf{r} - \mathbf{r}'| = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2} \cdot \Delta l, \tag{4}$$

где *x,y,z* – целочисленные координаты вокселей, а $\Delta l = \sqrt[3]{\Delta V}$ - длина ребра вокселя. Сопоставим векторам координат (*x,y,z*), (*x*′,*y*′,*z*′) индексы *i* и *k* для каждой пары элементов: (*x,y,z*) → *i*; (*x*′,*y*′,*z*′) → *k*; *i,k* = 1,...,*M*

$$\rho_{ik} = \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2} \cdot \Delta l \tag{5}$$

Установим вспомогательную функцию – оператор (ядро) переноса, которая представляет собой вероятность того, что фотон, излученный в точке **r**' поглотится в точке **r**:

$$\mathrm{K}\left(|\mathbf{r} - \mathbf{r}'|\right) = \frac{1}{4\pi} \int \kappa_\omega(\mathbf{r}) P(\omega) \frac{\exp(-\kappa_\omega(|\mathbf{r} - \mathbf{r}'|)}{|\mathbf{r} - \mathbf{r}'|^2} d\omega, \tag{6}$$

где ω - круговая частота, при которой $P(\omega) = 2\pi P(\nu)$

Основная идея данного подхода заключается в аппроксимации сложных объектов с использованием более простых граничных объемов через разбиение пространства на воксели. Каждый воксель содержит список объектов, находящихся внутри него. При трассировке луча сначала анализируется область начала или конца траектории луча [12]. Если луч пересекает объект из списка объектов в текущем вокселе, пересечения сортируются, и сохраняется ближайшее из них. Если пересечение найдено, дальнейшая проверка объектов становится излишней. В случае, если пересечение не обнаружено или список объектов в вокселе пуст, луч переходит в соседний воксель и сканирует его список объектов. Этот процесс продолжается до тех пор, пока не будет обнаружено пересечение или луч не пройдет через все воксели на своем пути. Такой подход существенно уменьшает количество вычисляемых пересечений, поскольку обрабатываются только те объекты, которые находятся на пути луча, а объекты, удаленные от него, игнорируются.

Размер сетки N_{dim} определяется максимальным числом вокселей вдоль каждой координатной оси (например, x, y, z). Для повышения эффективности вычислений алгоритм трассировки лучей был адаптирован для параллельной обработки на графических процессорах (GPU).

Поскольку каждый поток на графическом процессоре может выполнять только скалярные операции, создаются векторы с координатами каждой пары (r, r_0). Количество обрабатываемых одновременно пар вокселей ограничивается доступной памятью GPU. Далее фрагмент с рассчитанными расстояниями передается на центральный процессор (CPU), где вычисляются соответствующие коэффициенты, которые затем сохраняются в файл.

Для учета всех пересеченных вокселей на пути между элементами *i* и *k* авторами разработан следующий алгоритм трассировки лучей:

1. Устанавливаются начальная и конечная координаты, соответствующие элементам с индексами *i* и *k*;
2. Рассчитываются углы наклона луча $\angle X = \rho_x / \max(\rho_x,\rho_y,\rho_z)$, $\angle Y$ и $\angle Z$;
3. Вычисляются расстояния (h_x,h_y,h_z) из начальной точки до граней пересечения с соседними элементами. Минимальная из этих величин определяет расстояние, которое возможно пройти вдоль луча *k* → *i*, при этом оставаясь в пределах текущего элемента;
4. Рассчитываются инкременты для каждой из компонент $\Delta h_x = \Delta l / \angle X$, Δh_y и Δh_z . К минимальному на данный момент расстоянию min (h_x,h_y,h_z) прибавляется инкремент, а соответствующая координата увеличивается на фиксированный шаг *o*;
5. Осуществляется переход в другой воксель;
6. Пункты 2 − 5 повторяются до тех пор, пока не будет достигнута конечная точка в элементе *i*.

В результате, на пути движения луча образуется массив пересеченных вокселей.

При таком подходе несложно перейти от простейшего двухуровнего случая к многоуровневой системе атома, в данном случае, ртути:

$$A_{s0} n_s - \frac{A_{s0}}{4\pi} \int_\nu \int_V \frac{n_s(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^2} \kappa_\omega \exp\left[-\kappa_\omega |\mathbf{r} - \mathbf{r}'|\right] d^3 r' P(\omega) d\omega = \sum_{s=1}^{6} n_s(\mathbf{r}) Z_{s0} + w_{s0} n_s - w_{0s} n_s + \nabla(D_s \nabla n_s) \tag{7}$$

Сделаем следующую замену переменных:

$$-w_{s0}n_s + w_{0s}n_s - \nabla(D_s\nabla n_s) = F(\mathbf{r}) - \frac{n_s}{\tau_s}, \tag{8}$$

где $F(\mathbf{r})$ – скорость образования резонансных атомов ртути за счет возбуждения электронным ударом и оптических переходов с вышележащих уровней; τ_s – время жизни состояния s, обусловленное тушащими соударениями атомов ртути с электронами.

В результате имеем:

$$A_{s0}\cdot n_s(\mathbf{r}) - A_{s0}\cdot\int_V n_s(\mathbf{r}')\cdot \mathrm{K}\left(\left|\mathbf{r}-\mathbf{r}'\right|\right)d^3r' = \sum_{s=1}^{6} n_s(\mathbf{r})Z_{s0} - F(\mathbf{r}) + \frac{n_s}{\tau_s} = W_s(\mathbf{r}) \tag{9}$$

В случае цилиндрической формы лампы граничные условия, основываясь на [13], можно записать как:

$$\left.\frac{dn_s}{dr}\right|_{r=0} = 0\,,\ n_s(R) = 0\,,\ s = 1,2,...,6\,. \tag{10}$$

где R – радиус трубки.

Тогда, в соответствии со схемой энергетических уровней атома ртути (рис. 1.), представим систему уравнений баланса образования и разрушения возбужденных состояний атома ртути для всех указанных состояний, получаем следующее:

1: 6^3P_0
$$n_e\alpha'_{01}n_0 + n_e\sum_{в=2}^{6}\beta'_{в1}n_в + n_5A_{51} + c_1n_6A_{6^{(1)}1} = n_e\sum_{в=2}^{i}\alpha'_{1в}n_1 + n_e\beta'_{10}n_1 - \nabla(D\nabla n_1) \tag{11}$$

2: 6^3P_1
$$n_e\sum_{н=0}^{1}\alpha'_{н2}n_н + n_e\sum_{в=3}^{6}\beta'_{в2}n_в + n_5A_{52} + c_1n_6A_{6^{(1)}2} + c_2n_6A_{6^{(2)}2} = div\mathcal{E}_{p,254}(\mathbf{r}) + n_e\sum_{в=3}^{i}\alpha'_{2в}n_2 + n_e\sum_{н=0}^{1}\beta'_{2н}n_2 - \nabla(D\nabla n_2) \tag{12}$$

3: 6^3P_2
$$n_e\sum_{н=0}^{2}\alpha'_{н3}n_н + n_e\sum_{в=4}^{6}\beta'_{в3}n_в + n_5A_{53} + c_1n_6A_{6^{(1)}3} + c_2n_6A_{6^{(2)}3} + c_3n_6A_{6^{(3)}3} = n_e\sum_{в=4}^{i}\alpha'_{3в}n_3 + n_e\sum_{н=0}^{2}\beta'_{3н}n_3 - \nabla(D\nabla n_3) \tag{13}$$

4: 6^1P_1
$$n_e\sum_{н=0}^{3}\alpha'_{н4}n_н + n_e\sum_{в=5}^{6}\beta'_{в4}n_в + c_1n_6A_{6^{(1)}4} + c_2n_6A_{6^{(2)}4} = div\mathcal{E}_{p,185}(\mathbf{r}) + n_e\sum_{в=5}^{i}\alpha'_{4в}n_4 + n_e\sum_{н=0}^{3}\beta'_{4н}n_4 - \nabla(D\nabla n_4) \tag{14}$$

5: 7^3S_1
$$n_e\sum_{н=0}^{4}\alpha'_{н5}n_н + n_e\beta'_{65}n_6 = \frac{n_5}{\tau_5} + n_e\sum_{в=6}^{i}\alpha'_{5в}n_5 + n_e\sum_{н=0}^{4}\beta'_{5н}n_5 - \nabla(D\nabla n_5) \tag{15}$$

6: 6^3D
$$n_e\sum_{н=0}^{5}\alpha'_{н6}n_н = \frac{n_6}{\tau_6} + n_e\alpha'_{6i}n_6 + n_e\sum_{н=0}^{5}\beta'_{6н}n_6 - \nabla(D\nabla n_6) \tag{16}$$

Здесь $div\mathcal{E}_{p,s}(\mathbf{r})$ – дивергенция вектора поверхностной плотности потока резонансных фотонов, α'_{ij} – число соударений атомов с электронами, приводящих к переходу с уровня i на уровень j в результате ударов I рода в единице объема в единицу времени в расчете на один электрон; ion – номер уровня ионизации; β'_{ji} – число соударений атомов с электронами, приводящих к переходу с уровня j на уровень i в результате ударов II рода в единице объема в единицу времени в расчете на один электрон. Индекс $6^{(1)}$ (допустим, в $A_{6^{(1)}1}$) соответствует переходу с уровня 6^3D_1, $6^{(2)}$ – с уровня 6^3D_2, $6^{(3)}$ – с уровня 6^3D_3.

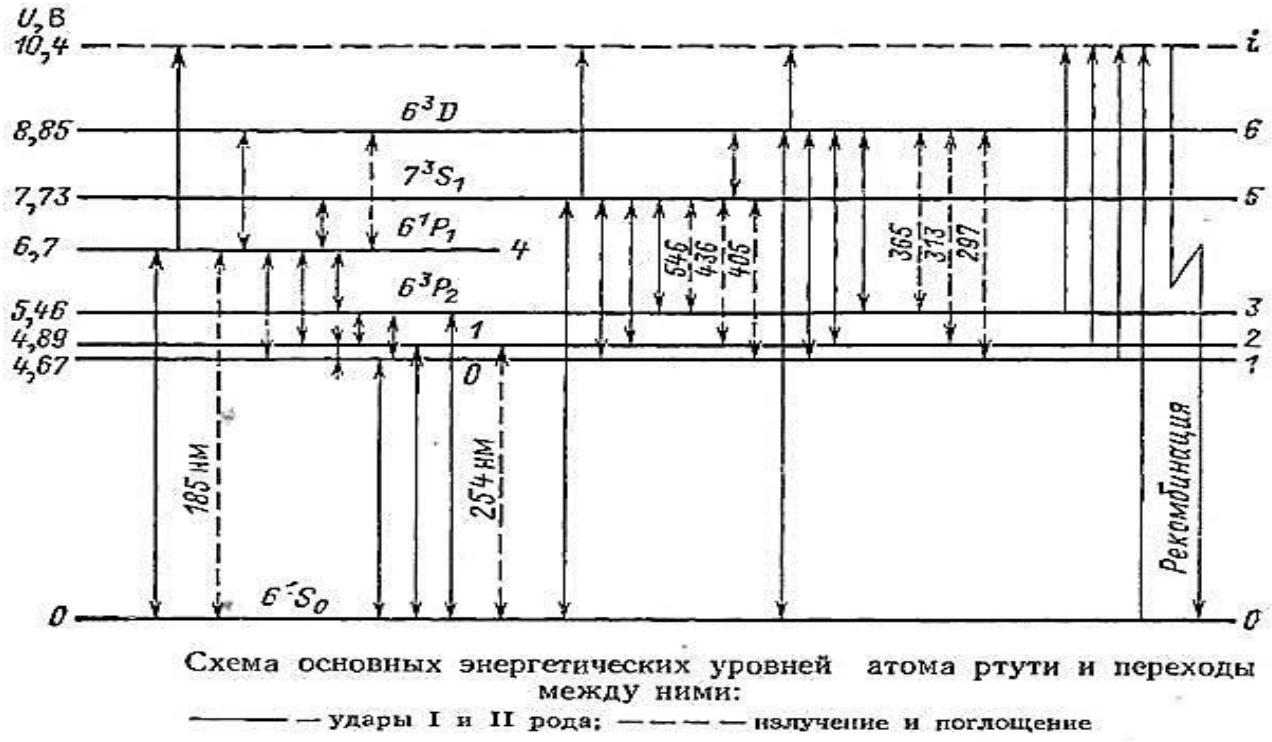


Рис. 1. Схема основных уровней атома ртути

Введем следующие обозначения:

$$c_1 = \frac{n_{6^3D_1}}{n_{6^3D_1} + n_{6^3D_2} + n_{6^3D_3}}, c_2 = \frac{n_{6^3D_2}}{n_{6^3D_1} + n_{6^3D_2} + n_{6^3D_3}}, c_3 = \frac{n_{6^3D_3}}{n_{6^3D_1} + n_{6^3D_2} + n_{6^3D_3}} \tag{17}$$

Соответственно, для уровня 6^3P_1, можно записать:

$$F_{254}(\mathbf{r}) = n_e \sum_{н=0}^{1} \alpha'_{н2} n_н + n_e \sum_{в=3}^{6} \beta'_{в2} n_в + n_5 A_{52} + c_1 n_6 A_{6^{(1)}_2} + c_2 n_6 A_{6^{(2)}_2} \tag{18}$$

$$\tau_{254}(r) = \left[n_e \sum_{в=3}^{i} \alpha'_{2в} + n_e \sum_{н=0}^{1} \beta'_{2н} \right]^{-1} \tag{19}$$

Для решения системы уравнений (11-16) с учетом (17-19) необходимо знание вероятностей спонтанного излучения изолированных атомов ртути. Основываясь на [14, 15, 16], получаем следующее:

$$n_0 + \sum_{s=1}^{6} n_s = \frac{p_{Hg}}{kT} \tag{20}$$

$$\begin{pmatrix} 0 \\ D\Delta n_1 \\ D\Delta n_2 \\ D\Delta n_3 \\ D\Delta n_4 \\ D\Delta n_5 \\ D\Delta n_6 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ Z_{01} & -1/\tau_1 & Z_{21} & Z_{31} & Z_{41} & Z_{51} & Z_{61} \\ Z_{02} & Z_{12} & -1/\tau_2 & Z_{32} & Z_{42} & Z_{52} & Z_{62} \\ Z_{03} & Z_{13} & Z_{23} & -1/\tau_3 & Z_{43} & Z_{53} & Z_{63} \\ Z_{04} & Z_{14} & Z_{24} & Z_{34} & -1/\tau_4 & Z_{54} & Z_{64} \\ Z_{05} & Z_{15} & Z_{25} & Z_{35} & Z_{45} & -1/\tau_5 & Z_{65} \\ Z_{06} & Z_{16} & Z_{26} & Z_{36} & Z_{46} & Z_{56} & -1/\tau_6 \end{pmatrix} \times \begin{pmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \\ n_6 \end{pmatrix} = \begin{pmatrix} p_{Hg}/kT \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{21}$$

IV. Результаты экспериментов

Проведем верификацию предложенного метода. Для этого запишем уравнение, описывающее излучение СВЧ разряда в ртутно-аргоновой плазме:

$$W(r,z) = W_{\max} \cdot \delta(r) \cdot 10^{-\alpha z} \tag{22}$$

где $W_{\max}$ – максимум функции $W(r,z)$, α = const, $\delta(r)$ – функция возбуждения, которая имитирует излучение ТЕМ-волны магнетрона, используемого для зажигания коаксиального СВЧ ртутно-аргонового разряда в

кварцевой цилиндрической трубке длиной 20 см при давлении аргона $p_{Ar} = 0.75$ Торр, и ртути $p_{Hg} = 0.05$ Торр [17].

В ходе анализа была выявлена интересная особенность, связанная с наличием экранирующего препятствия вдоль траекторий полета фотонов. В настоящей работе это учитывалось следующим образом: когда луч пересекает воксель, координаты которого отсутствуют в массиве координат области, то «коэффициент», соответствующий рассматриваемой паре элементов i и k, зануляется.

На основании проведенного математического анализа разработан алгоритм оптимального решения задачи переноса:

1. Инициализация
   - 1.1. Входные параметры: коэффициент Эйнштейна - A_{s0}, объем ИИ - V, коэффициент поглощения в центре резонансной линии κ_0, функция контура спектральной линии $P(\omega)$
   - 1.2. Определение размера сетки N_{dim}
   - 1.3. Построение массива пар элементов
2. Вычисление матричных элементов
   - 2.1. Определение размера подмассива для параллельных вычислений
   - 2.2. Построение подмассива ($\mathbf{r}$, $\mathbf{r}'$)
   - 2.3. Алгоритм трассировки лучей
   - 2.4. Расчет ядра переноса $\mathrm{K}(\rho)$
   - 2.5. Сохранение коэффициентов
   - 2.6. Сбор коэффициентов в матрицу
3. Решение системы уравнений
   - 3.1. Введение источников возбуждения $W_s(\mathbf{r}) = \sum_{s=1}^{6} n_s(\mathbf{r}) Z_{s0} - F(\mathbf{r}) + \frac{n_s}{\tau_s}$
   - 3.2. Решение в виде $A_{s0}x = W_s(\mathbf{r})$, где $x = n_s(\mathbf{r}) - \int_V n_s(\mathbf{r}') \cdot \mathrm{K}\left(|\mathbf{r} - \mathbf{r}'|\right) d^3 r'$

Подчеркнем, что на последнем этапе вычислений задается желаемая форма источников возбуждения $W(\mathbf{r})$, после чего производится решение линейной системы уравнений. В связи с тем, что используется равномерная и декартовая сетка, основная сложность решения, связанная с вычислением пространственных интегралов, устраняется.

Так, для дискретизации ядра переноса $\mathrm{K}(\rho)$ имеем следующее:

$$a_{k,i} = A_{s0} \cdot \begin{cases} \frac{\Delta V}{4\pi} \int_0^{\infty} \frac{P_k(\omega)\kappa_{\omega,i}}{\rho_{k,i}^2} \exp\left(- \sum_{j=l_1^{k,i}}^{l_m^{k,i}} \kappa_{\omega,j} \Delta l \right) d\omega, \; k \neq i \\ 1 + \int_0^{\infty} \left[\int_0^{\Delta r} P_k(\omega)\kappa_{\omega,i} \exp\left(-\kappa_{\omega,i} r\right) dr \right] d\omega, \; k = i \end{cases} \tag{23}$$

Получаем, что выражения справедливы для произвольного контура спектральной линии. Отметим, что когда оптическая плотность $\kappa_0 r > 100$, можно получить асимптотическую форму решения:

$$K(\kappa_0 r) = \frac{1}{8\sqrt{\pi^3 \kappa_0 r^7}} \tag{24}$$

V. Обсуждение результатов

Покажем зависимость полученного решения от размера сетки (рис. 2). Определение оптимального количества точек в сетке основывалось на достижении наилучшего соотношения между временем, требуемым для проведения расчетов, и необходимой степенью детализации для адекватного описания изучаемых процессов.

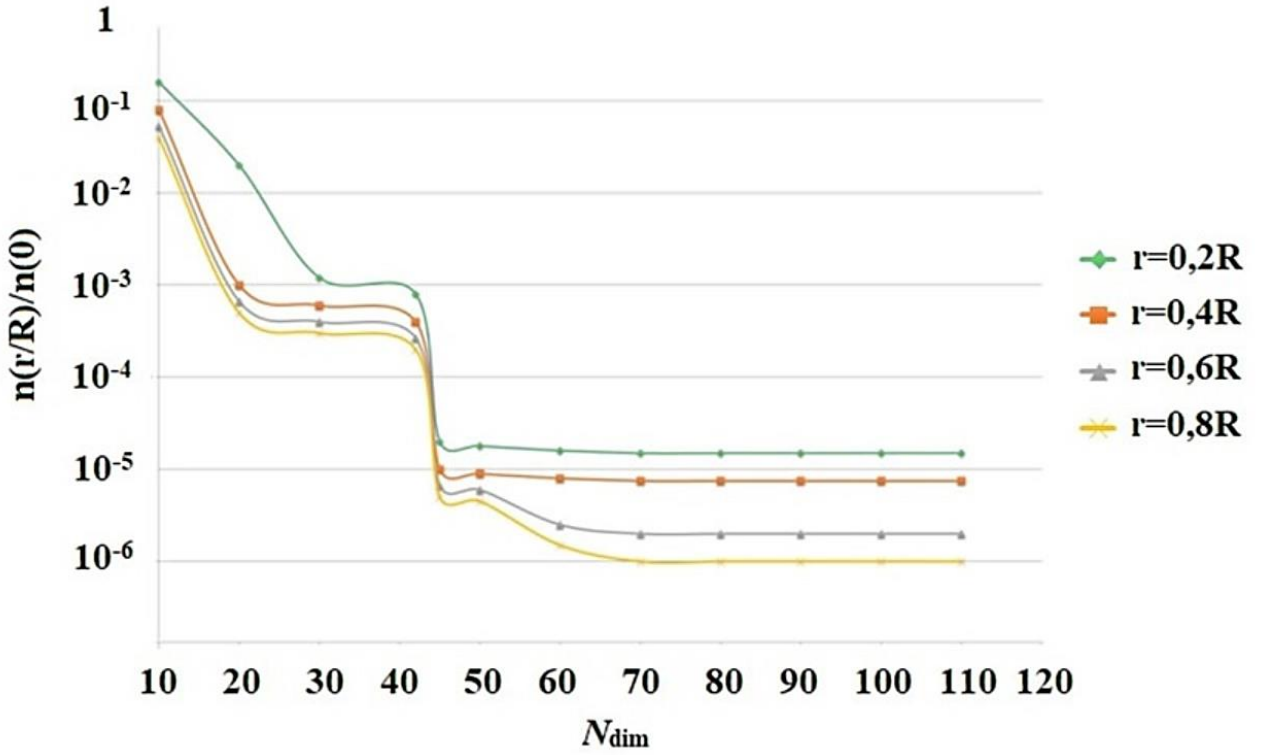


Рис. 2. Относительные значения концентраций как функции N_{dim} в зависимости от различных радиальных позиций

В результате проведенных эмпирических исследований было установлено, что оптимальный размер сетки составляет N_{dim} = 42, при этом коэффициент поглощения κ_0 = 10^6 см$^{-1}$. Таким образом, для каждого элементарного объема (вокселя) в цилиндрической геометрии решение уравнения переноса позволило провести оценку распределения концентрации возбужденных атомов по энергетическим уровням.

Покажем результаты компьютерного моделирования процессов реабсорбции, диффузии и пленения излучения в лампе, полученные в ходе решения системы (21) (рис. 3).

Проведем сравнение решений (21), выполненных по вышеописанному алгоритму с решениями, полученными методами конечных элементов и ступенчатой аппроксимации концентрации (рис. 4). Сравнение проводилось для условий разряда, полученных в ходе экспериментального исследования [18]: $P_1 = 687 Вт \cdot м^{-1}, A_{254} = 0.085 \cdot 10^8 c^{-1}, A_{185} = 6.77 \cdot 10^8 c^{-1}, f = 2.45 ГГц, R = 12 мм, L = 200 мм, p_{Hg} = 0.05 Topp, p_{Ar} = 0.75 Topp, T = 60^{\circ} C$.

Стоит отметить, что все расчеты были реализованы на GPU и проводились на одном и том же компьютере для разных методов.

Из рис. 4. следует, что результаты вычислений показали высокое количественное соответствие. Но, несмотря на это, метод вокселизации требовал значительно меньше времени (~18 мин.) для решения задачи, сохраняя при этом высокую точность вычислений. Для сравнения метод конечных элементов требовал порядка 5 часов, метод ступенчатой аппроксимации около 350 мин.

VI. Выводы и заключение

Разработан метод решения интегрального уравнения Бибермана-Холстейна, который заключается в совместном решении уравнения переноса излучения и системы кинетических уравнений, позволивший учитывать взаимодействие между процессами поглощения и излучения в цилиндрической геометрии, а также различные механизмы уширения резонансных линий. Внедрение предложенного авторами подхода в трехмерных структурах дает возможность корректно моделировать процессы возбуждения атомов и молекул в условиях, приближенных к реальным экспериментальным установкам. Применение метода вокселизации в разработанной модели позволило учитывать такие физические процессы, как реабсорбция, пленение и диффузия излучения в СВЧ ртутно-аргоновом разряде, что ранее было затруднительно с использованием известных моделей. Разработанный эффективный алгоритм параллельных вычислений на графических процессорах (GPU) стал ключевым элементом предложенного метода, обеспечивая высокую производительность и точность расчетов.

Предложенный подход можно в дальнейшем распространить на решение задач по моделированию переноса излучения в таких неравновесных средах, как атмосфера, морская вода, космическая плазма.

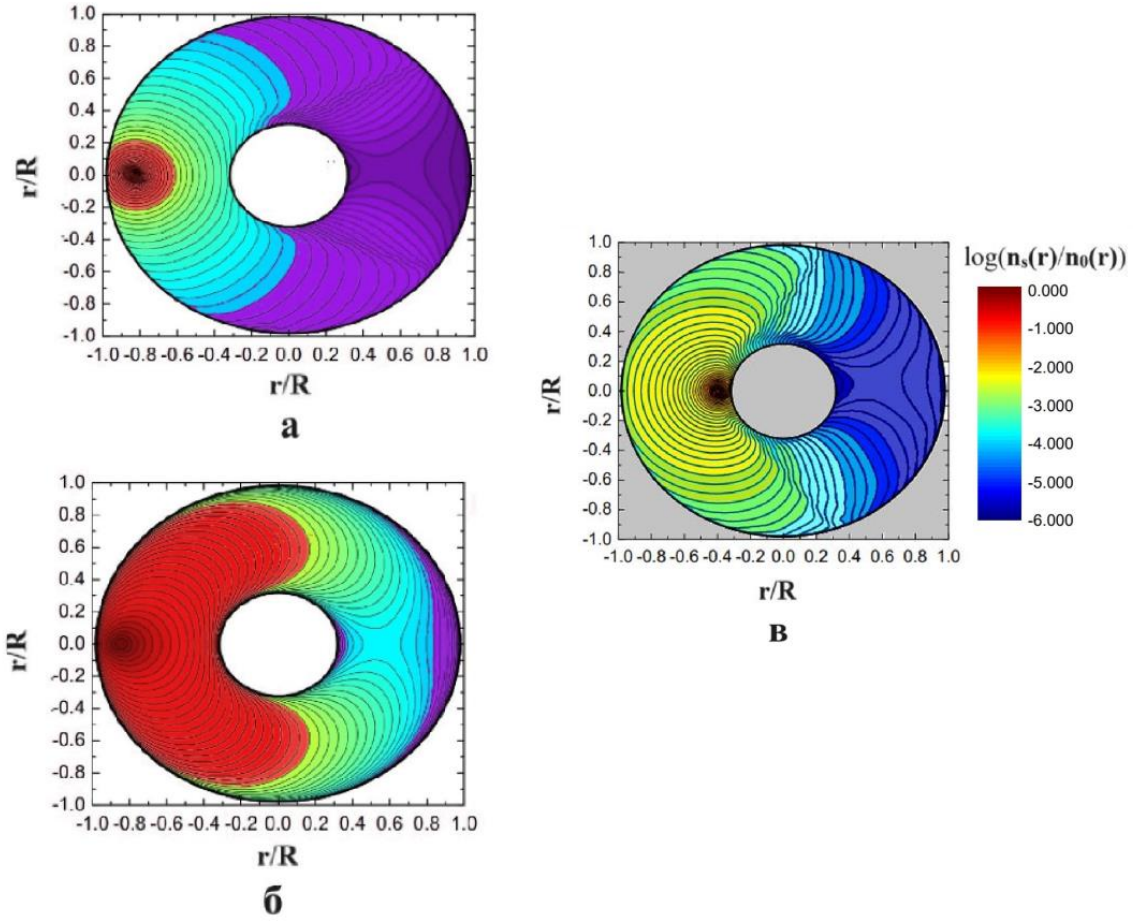


Рис. 3. Результаты моделирования: реабсорбция (а), диффузия (б), пленение (в)

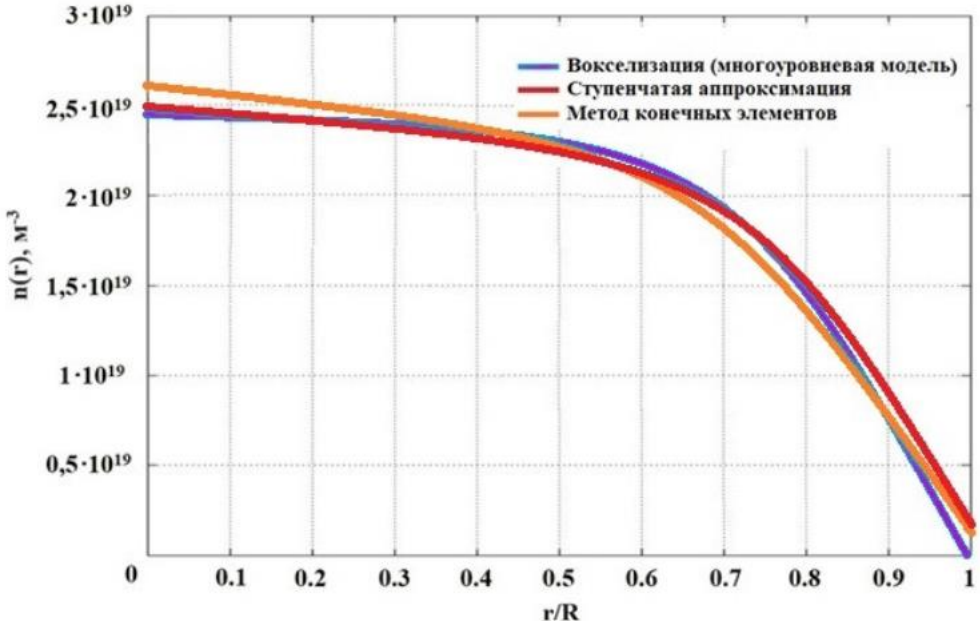


Рис. 4. Сравнение решений

Благодарности

Научный руководитель - Будак Владимир Павлович, НИУ «МЭИ»

Список литературы

1. Chandrasekhar S. Radiative Transfer. Dover Publications, New York, 1960. 416 p.
2. Michelson M., Mihalas D. Monte Carlo Methods for Radiative Transfer in Three-Dimensional Media. Journal of Quantitative Spectroscopy and Radiative Transfer, 1989. P. 88–96
3. Lux I., Koblinger L. Monte Carlo Particle Transport Methods: Neutron and Photon Calculations. CRC Press, 1991. 530 p.
4. Howell J. R. The Monte Carlo Method in Radiative Transfer. Journal of Heat Transfer, 1998.
5. Dimitrov D., Ivanov M. Adaptive Mesh Refinement in Radiative Transfer Calculations. Journal of Computational Physics, 2003.
6. Modest M. F. Radiative Heat Transfer. Academic Press, 2013.
7. Fang Z., Shu C.-W. High Order Methods for Computational Physics. Springer, 2011.
8. Астрелин В. Т., Солдаткина Е. И. Численное моделирование источника электронного пучка установки ГДЛ // Физика плазмы. 2022. Т. 48. № 1. С. 3-11.
9. Будак В.П., Железнов И.И. Применение научной визуализации в моделировании низкотемпературной газоразрядной плазмы // Научная визуализация. 2023. Т15. №4. С. 41 – 55.
10. Климонтович Ю. Л. Кинетическая теория электромагнитных процессов / М.: Наука, 1980. 374с.
11. Budak V. P., Zheleznov I. I., Grigoryev A. A. Features of solving the radiation transfer equation in low-temperature gas discharge plasma / / Light & Engineering. 2024. Vol. 32, No. 3. P. 4-10.
12. Amanatides J., Woo A.A. Fast Voxel Traversal Algorithm for RayTracing//Eurographics. 1987. Vol.87. P.3–10
13. Каланов В.П., Федоренко А.С. О сложной структуре спектральных линий атомов ртути // Тр. ВНИИИС. Вып.22. 1990. С.60-70.
14. Радциг А.А., Смирнов Б.М. Параметры атомов и атомных ионов, справочник // М.: Энергоатомиздат, 1986. 344 с.
15. Решенов С.П. Генерация и перенос излучения в плазме. М.: МЭИ, 1989. 118 с.
16. Свитнев С.А. Эффективный источник ультрафиолетового излучения на основе разряда низкого давления. Диссертация на соискание ученой степени кандидата технических наук. М.: МЭИ, 2012.
17. Zheleznov I.I., Popov O.A. Electrodeless source of UV radiation based on low pressure microwave mercury discharge // Light & Engineering. 2023. Vol. 31, No. 3. P. 36-42.
18. Zheleznov I., Konovalov S., Vanin I. Electrodeless microwave source of bactericidal ultraviolet radiation// Proceedings of the 13th Asia Lighting Conference (Beijing, China) P.789-791

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-105-122

Виртуальная лаборатория анализа музейного освещения
Virtual Laboratory for Analyzing Museum Lighting

В.П. Будак1, И.Р. Галлямов1, Д.А. Киселёв1
1Национальный Исследовательский Университет "МЭИ", Красноказарменная 14, Москва, 111250, Россия

V.P. Budak[1], I.R. Galliamov[1], D.A. Kiselev[1]
1 National Research University "MPEI", Krasnokazarmennaya 14, Moscow, 111250, Russia

Аннотация. Освещение играет ключевую роль в восприятии произведений искусства. Это действительно важный вопрос, которым долгое время не интересовались, так как существовали строгие нормативы о том, как нужно освещать, сформулированные еще в СССР. Но время летит вперёд и рекомендации, а также требования должны не отставать и идти в ногу с современными тенденциями и технологическими возможностями. Для эффективного восприятия произведений искусства важно правильно подобрать освещение. Оно должно подчеркивать форму, текстуру и цвет объекта, а также создавать естественные тени и атмосферу, чтобы помочь зрителю лучше понять и оценить произведение. Однако традиционные методы освещения могут привести к появлению неестественных теней и бликов, что негативно сказывается на восприятии произведения. Современные технологии позволяют создавать виртуальные дисплеи, доступные широкой аудитории, но они требуют реалистичной визуализации синтетических изображений скульптур. Модуляция освещения в программах глобального расчета освещенности и использование моделей цветового восприятия позволят анализировать и проверять разные подходы к освещению, учитывая множество параметров, и, проведя ряд экспериментов, с достаточной точностью выявлять положительные и отрицательные стороны решений и рекомендовать необходимые в разных ситуациях и обстоятельствах.

***Ключевые слова*:** CIECAM02, музейное освещение, произведения живописи, освещение скульптур, виртуальная лаборатория.

Аннотация. Lighting plays a key role in the perception of works of art. This is a really important issue, which for a long time was not interested in, because there were strict regulations on how to light, formulated back in the USSR, but time flies forward and recommendations and requirements should keep up and keep up with modern trends and technological capabilities. For effective perception of works of art it is important to choose the right lighting. It should accentuate the shape, texture and color of the object, as well as create natural shadows and atmosphere to help the viewer better understand and appreciate the artwork. However, traditional lighting methods can result in unnatural shadows and glare, which negatively affects the perception of the artwork. Current technologies allow for virtual displays that are accessible to a wide audience, but they require realistic visualization of synthetic images of the sculptures. Modulation of lighting in global calculation programs and the use of color perception models will make it possible to analyze and test different approaches to lighting, taking into account a variety of parameters, and, having conducted a number of experiments, to identify with sufficient accuracy the positive and negative sides of solutions and recommend the necessary ones in different situations and circumstances.

***Ключевые слова*:** CIECAM02, museum lighting, painting works, sculpture lighting, virtual laboratory.

I. Введение

Освещение играет ключевую роль в восприятии произведений искусства. Оно способно подчеркнуть особенности картины, добавить ей глубины и выразительности.[1] С появлением синтетических изображений важной задачей стало найти способы реалистичного отображения освещения на экране компьютера, при которой наблюдатель получит тот же визуальный опыт, как при просмотре вживую. [2].

Модуляция освещения в программах глобального расчета и использование моделей цветового восприятия позволят анализировать и проверять разные подходы к освещению, учитывая множество параметров. Проведя ряд экспериментов, с достаточной точностью выявлять положительные и отрицательные стороны решений и рекомендовать необходимые в разных ситуациях и обстоятельствах [3].

Влияние освещения на восприятие скульптуры - важный аспект, который необходимо учитывать при создании художественной композиции. Правильно подобранный свет может подчеркнуть форму и фактуру произведения искусства, добавить ему глубины и объема. Однако, если освещение выбрано неверно, это может привести к искажению формы и потере первоначального замысла автора. Тени, создаваемые светом, играют важную роль в визуальном образе скульптуры, помогая зрителю понять ее трехмерную природу и структуру.

Однако слишком яркий или слишком слабый свет, а также неправильный угол падения могут привести к появлению «неправильных» теней, искажающих форму скульптуры.

При всем вышесказанном, на дисплее монитора компьютера освещение может исказить восприятие музейные скульптуры. Таким образом, важно найти варианты реалистичного отображения освещения на дисплее монитора.

Данная тематика актуальна, так как остро стоит вопрос тенеобразования на скульптурах и реалистичного отображения изображения освещения скульптур на дисплее мониторов.

II. Постановка задачи

Цель данной работы — подтвердить применимость метода программой обработки синтетических изображений на основе CIECAM02 и с помощью нее сравнить различные приёмы освещения произведений искусства на экране компьютера, а также реализация путем образования «правильных» теней на скульптурах в музейном освещении. Для достижения этой цели будут рассмотрены различные методы освещения, их влияние на восприятие произведения искусства и возможности использования CIECAM02 для создания реалистичного отображения. Можно задаться вопросом: как рационально и грамотно подобрать прием освещения произведений искусства так, чтобы наблюдатель получил максимум положительных эмоций от просмотренной картины и его восприятие не исказилось за счёт освещения? Как оптимально проводить ряд экспериментов с подбором необходимых параметров освещения для получения наилучшего исхода, также прибегая к натурным методам?

III. Теория

Было проанализировано множество литературы, связанной с музейным освещением и подходам к его исследованию, продемонстрированные с списке литературы [8-23]. На основе предшествующих исследований в данной работе было проведено 2 эксперимента, касающихся освещения картин двух жанров и освещения скульптур различных простых форм. Для этих экспериментов применялось разное осветительное оборудования и разными светотехническими параметрами, такими как мощность, угол светораспределения и цветовая температура.

Для проведения экспериментов и исследования влияния освещения на полотна картин была создана сцена в программе DIALux Evo, воссоздающая помещение Главного здания Эрмитажа. За прототип помещения был выбран зал итальянской живописи среди многочисленных помещений (рис. 2.1). Первым этапом стало изучение плана здания и выбор одного из залов, которое подходило по всем критериям, включая возможность регулировать размещение источников освещения. Затем было проведено измерения всех параметров помещения, включая размеры, расположение окон и дверей, а также типы и количество светильников.

Рис. 1. Прототип выбранного помещения. Зал итальянской живописи

После этого созданию модели в программе Dialux Evo. Использованы все полученные данные для точного воссоздания интерьера. Учтены все особенности помещения, такие как: высота потолков, материалы отделки стен и пола, расположение картин. В результате создания пространственной 3D модели получили реалистично воссозданное в виртуальной среде помещение, которое позволяет визуализировать освещение и оценить его качество в условиях, приближенных к реальным. В работе использовались картины 2-х жанров: написанные полотна Васнецова и картины жанра Барокко. К сожалению, программы глобального расчета освещенности DiaLux Evo не даёт возможности учесть материалы, используемые в написании картин, а лишь возможность наложения текстурированного изображения полотна, которое по своим физическим параметрам имеет лишь коэффициент отражения. Поэтому создание реалистичного зала выполнено с допущением к свойствам полотна. Но даже такое приближение помогло лучше понять принципы работы освещения и его влияние на восприятие пространства (рис. 2).

Рис. 2. Сцена зала из программы Dialux Evo

Также для проработки приемов освещения были выбраны 3 типа светильников с разной оптикой. Световые приборы имеют угол раскрытия 20°, 40° и 60°. В дальнейших исследования планируется получить данные, используя осветительное оборудование с оптикой различных углов рассеяния, начиная от рассеянного мягкого света до узконаправленной. В данной работе для освещения картин применяется оборудование с цветовую температурой равной 4500К [5]. Нейтральный белый цвет свечения выбран как нулевое приближение в экспериментах, и корреляция в выборе цветовой температуры будет применено в дальнейших исследованиях. Такая оптическая составляющая осветительного оборудования наиболее распространена на рынке и широко используется в музейном освещении, а также легко реализуема для большой типологии светильников, которые можно грамотно интегрировать в различные интерьеры помещений музеев и выставок.

КСС выбранный светильников представлены на рис. 3.

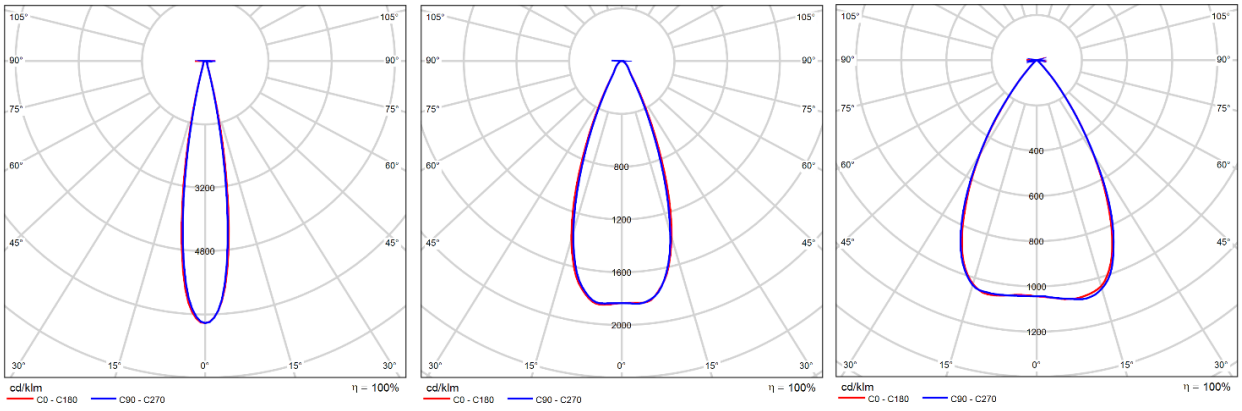


Рис. 3. Кривые силы света использованного в модели осветительного оборудования

В рамках данной работы были проведены эксперименты по изучению формирования "правильных" теней на музейных скульптурах. Для этого были выбраны простые геометрические фигуры: куб, цилиндр и шар.

Под «правильной» тенью понимается тень, которая помогает воспринимать фигуру такой, какая она есть на самом деле. Она позволяет видеть шар как шар, а цилиндр как цилиндр. Если тень сформирована неправильно, она может создать «театральный эффект», искажающий форму и вид фигуры.

Проводилось 2 эксперимента:

1. при помощи моделирования в лаборатории света «Light Point» компании «Точка Опоры»;
2. при помощи моделирования в программе DIALux EVO и программы для оценки восприятия студентов кафедры «Светотехники».

У компании "Точка Опоры" есть собственная световая лаборатория под названием «Light Point» (рис. 2.1) [4], где и были проведены необходимые исследования.

Для проведения эксперимента потребовалось согласование всех работ в лаборатории и использование светильников компании. Были определены ключевые характеристики светильников, такие как цветовая температура (Тцв) и угол светового пучка (табл. 1), которые были необходимы для проведения эксперимента. Когда все работы были согласованы, пришло время определить оптимальное время и даты проведения эксперимента.

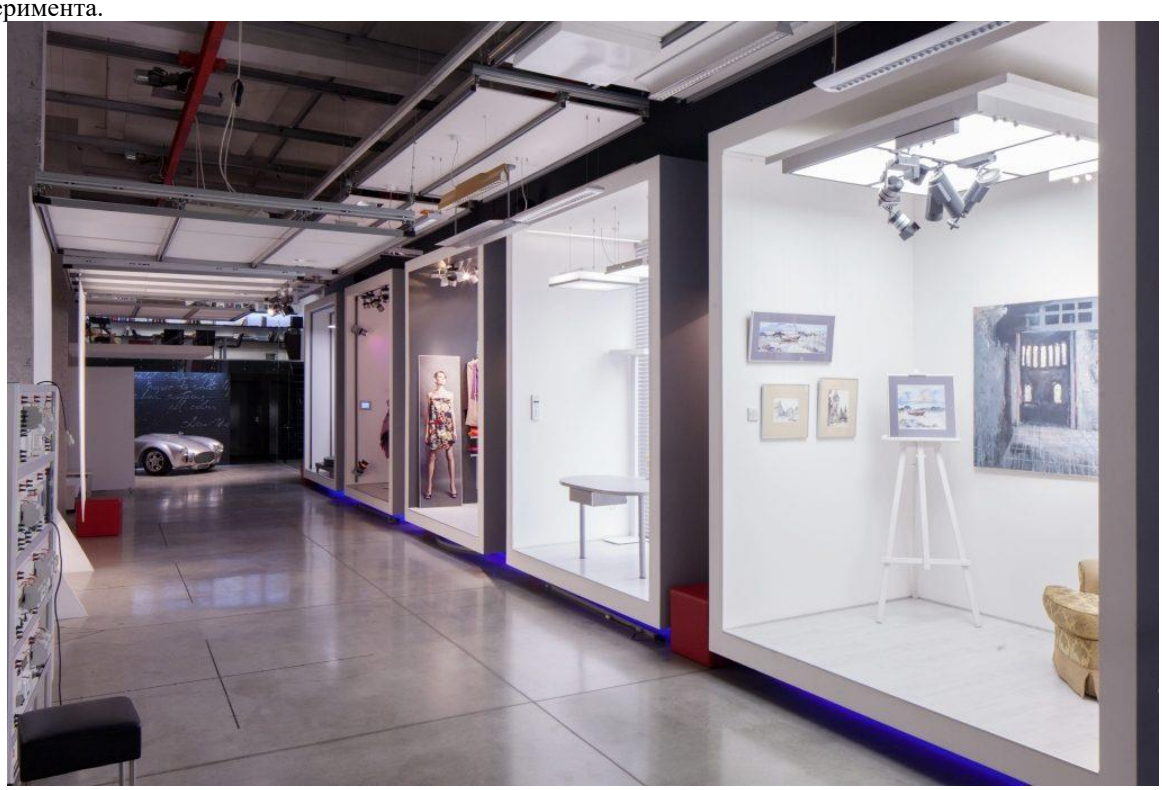


Рис. 4. Световая лаборатория «Light Point»

Таблица 1

Характеристики используемого светового прибора*

Световой прибор	Характеристики светового прибора	
	Цветовая температура, К	Угол раскрытия светового пучка, °
Трековый прожектор	3500	10

* - данные характеристики были согласованы с компанией «Точка Опоры», остальные характеристики являются конфиденциальной информацией.

В компании «Точка Опоры» работает большое количество специалистов в области светотехники, что позволило легко найти экспертов для участия в эксперименте.

Моделирование эксперимента проходило на фоне чистой белой стены, свободной от лишних элементов, чтобы обеспечить максимальную концентрацию внимания на фигурах. Материал фигур – гипс белого матового цвета с рассеивающими свойствами. Для моделирования использовался трековый светильник, который располагался в трех разных положениях на расстоянии 2,5 м от стены:

- на высоте 3 метра от пола под углом 0° относительно фигуры (ровно напротив фигуры);
- на высоте 3 метра от пола под углом 45° относительно фигуры;
- примерно на высоте глаз среднестатистического человека – 175 см от пола.

Далее на рис. 2.2-2.7 представлено, как выглядел данный натурный эксперимент:

Рис. 5. Моделирование с расположением светильника на высоте глаз среднестатистического человека

Рис. 6. Моделирование с расположением светильника на высоте 3 метра от пола под углом 45° относительно куба

Рис. 7. Моделирование с расположением светильника на высоте 3 метра от пола под углом 0° относительно куба

Рис. 8. Моделирование с расположением светильника на высоте 3 метра от пола под углом 0° относительно шара

Рис. 9. Моделирование с расположением светильника на высоте 3 метра от пола под углом 45° относительно шара

Рис. 10. Моделирование с расположением светильника на высоте глаз среднестатистического человека

Для проведения эксперимента была создана сцена в программе DIALux evo, воссоздающая сцену с натурного моделирования, упомянутого выше.

Для создания сцены в программе использовались базовые геометрические формы, что облегчило процесс проектирования. Материал, используемый в моделировании аналогичен натурному эксперименту. Сцена освещалась из нескольких выбранных точек, чтобы добиться оптимального эффекта. Было решено проанализировать освещение только с одной стороны, поскольку сравнивать левую и правую стороны не имело смысла, так как результат был идентичным.

Оптимальное расстояние от стены до точек освещения и их высота были определены на основе натурного моделирования. В качестве источника света был выбран прожектор, характеристики которого приведены в таблице 1.

Были выбраны следующие точки освещения:

1. На высоте 3 м от пола напротив фигуры.
2. На высоте 3 м от пола под углом 30° по отношению к фигуре.
3. На высоте 3 метров от пола под углом 45° по отношению к фигуре.
4. На высоте 0,5 метра от пола напротив фигуры.
5. На высоте 0,5 метра от пола под углом 30° по отношению к фигуре.
6. На высоте 0,5 метра от пола под углом 45° к фигуре.

Обработав получившиеся в ходе моделирования сцены зала с картинами через написанную и описанную программу по обработке синтетических изображений на основе CIECAM02 [6], получаем набор изображений для исследования и проведения экспериментов по выявлению наилучшего приема освещения произведений живопись с последующей обработкой этого исследования и создания шкалы ощущений. Результат обработанных синтетических изображений представлены на рис. 11 –15.

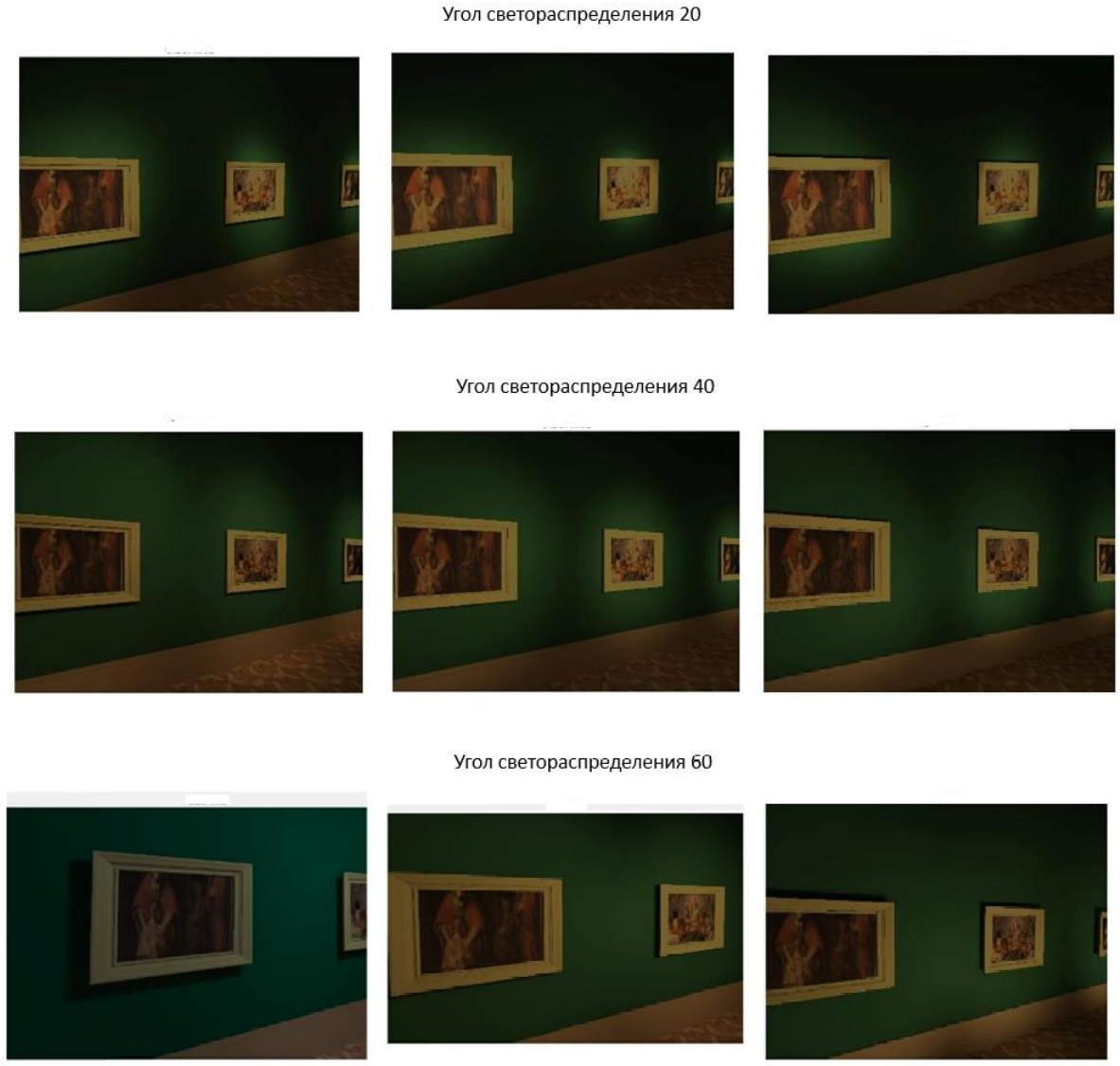


Рис. 11. Обработанные синтетические изображения картин в стиле Барокко, осветительное оборудования находится на разных высотах (3м, 1.75м и 0.5 м от пола)

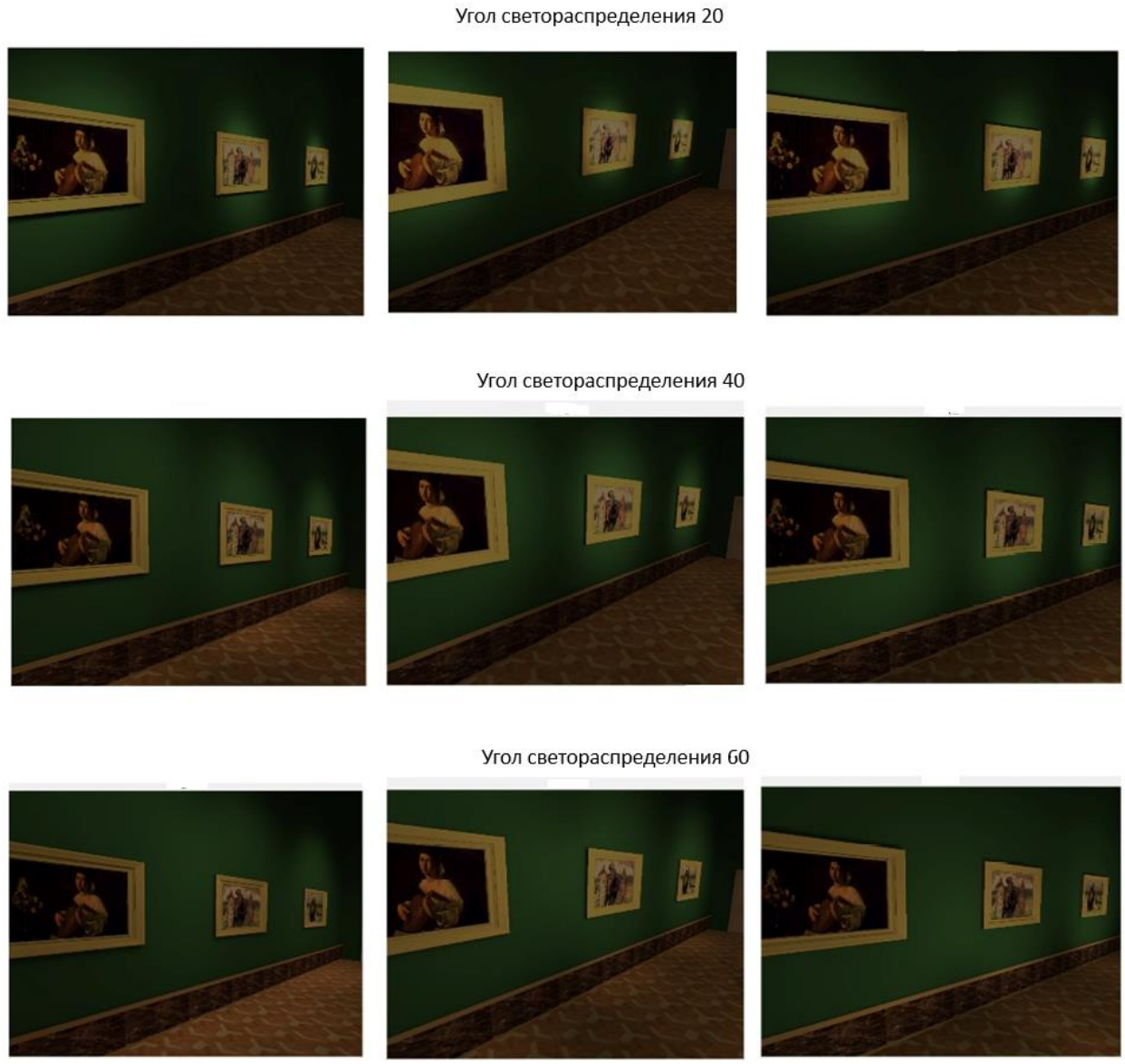


Рис. 12. Обработанные синтетические изображения картин Васнецова, осветительное оборудования находится на разных высотах (3м, 1.75м и 0.5 м от пола)

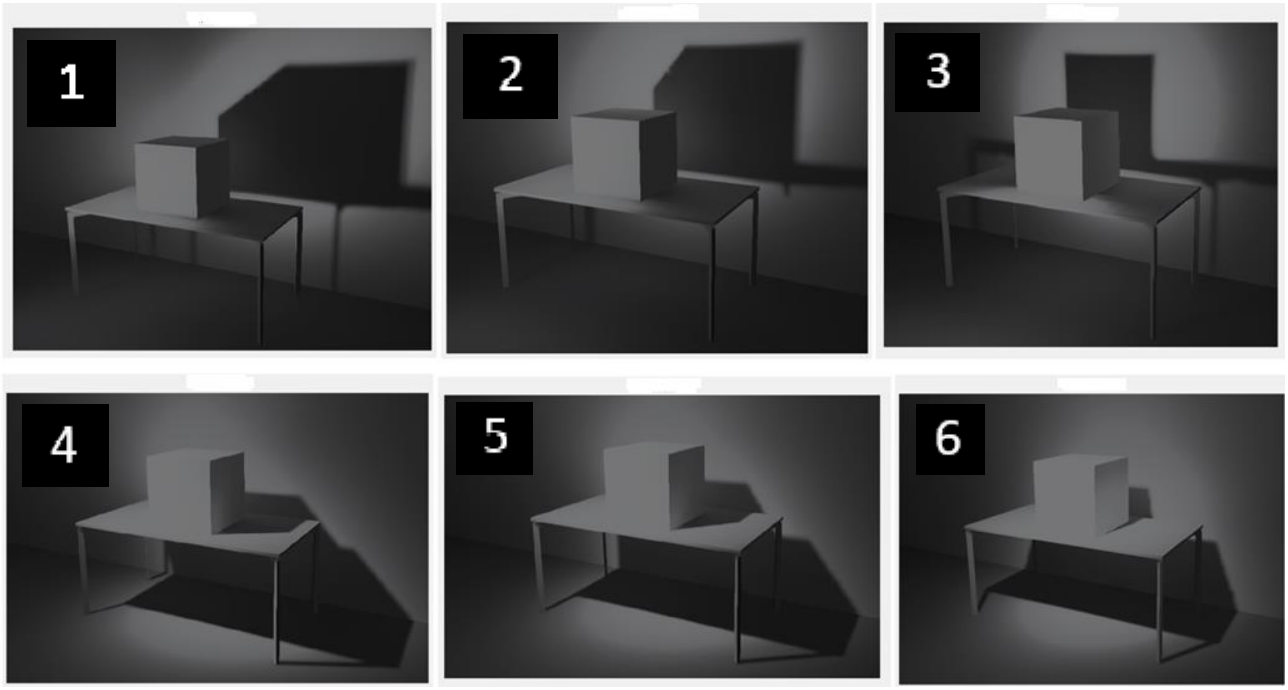


Рис. 13. Результат работы модели куб (1. На высоте 3 м от пола напротив фигуры.
2. На высоте 3 м от пола под углом 30° по отношению к фигуре.
3. На высоте 3 метров от пола под углом 45° по отношению к фигуре.
4. На высоте 0,5 метра от пола напротив фигуры.
5. На высоте 0,5 метра от пола под углом 30° по отношению к фигуре.
6. На высоте 0,5 метра от пола под углом 45° к фигуре.)

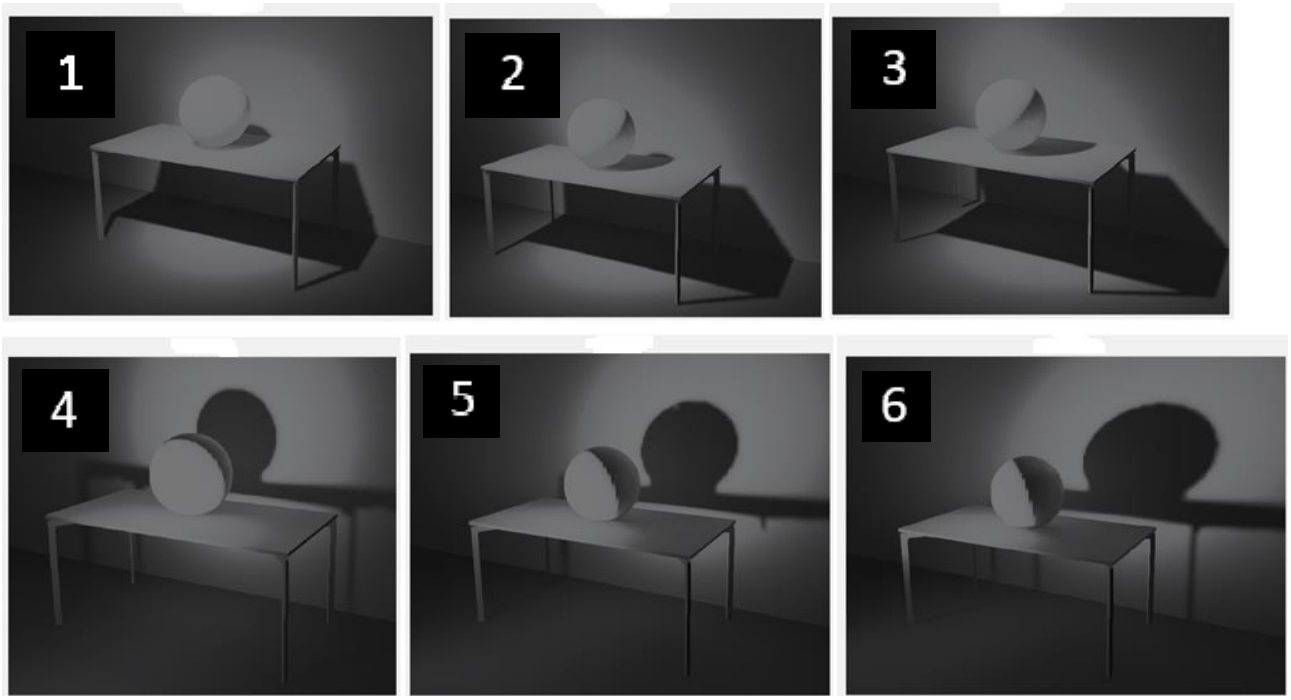


Рис. 14. Результат работы модели шар (1. На высоте 3 м от пола напротив фигуры.
2. На высоте 3 м от пола под углом 30° по отношению к фигуре.
3. На высоте 3 метров от пола под углом 45° по отношению к фигуре.
4. На высоте 0,5 метра от пола напротив фигуры.
5. На высоте 0,5 метра от пола под углом 30° по отношению к фигуре.
6. На высоте 0,5 метра от пола под углом 45° к фигуре.)

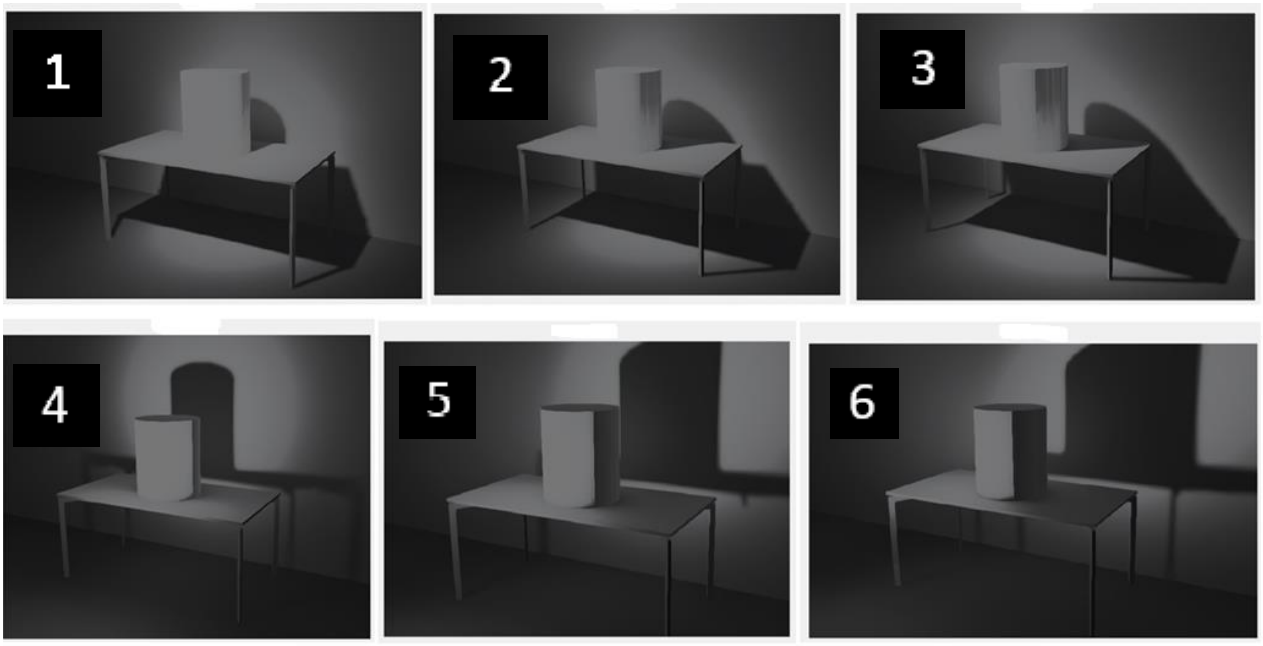


Рис. 15. Результат работы модели цилиндр (1. На высоте 3 м от пола напротив фигуры.
2. На высоте 3 м от пола под углом 30° по отношению к фигуре.
3. На высоте 3 метров от пола под углом 45° по отношению к фигуре.
4. На высоте 0,5 метра от пола напротив фигуры.
5. На высоте 0,5 метра от пола под углом 30° по отношению к фигуре.
6. На высоте 0,5 метра от пола под углом 45° к фигуре.)

Для удобной демонстрации получившихся синтетических изображений была написана программа, функционал которой завязан на демонстрации с выбором оценки. Программа должна иметь удобный интерфейс, понятный наблюдателю, работать исправно и быстро. Программа сравнения приемов освещения позволяет сравнивать различные методы и технологии освещения, поэтому ее можно использовать для анализа качества света и определения наиболее эффективных методов освещения. Основная цель программы - позволить наблюдателю выбрать лучший метод освещения для освещения произведений живописи. Для этого она предоставляет инструменты для сравнения различных параметров освещения. Кроме того, программа может автоматически генерировать отчет о результатах эксперимента, что делает её полезным инструментом. Приступив к написанию кода программы, важно разложить рабочий процесс на этапы, постепенное выполнение которых в конечном итоге приведет к желаемому результату. Итоговый интерфейс полученной в результате разработки программы представлен на рис. 16.

Критерием для отбора экспертов для проведения экспериментов, связанных с восприятием освещения произведений искусства послужил опыт данных лиц в работе с освещение. Студенты 1-го курса магистратуры кафедры «Светотехники» в своих многочисленных проектах в материалах учебных пособий и практических заданий за время обучения на кафедре не раз сталкивались с проектированием осветительных установок в различных сферах светотехники. Их опыт и намеренность положительно сказались на результатах исследовании и получении грамотной и правильной статистики.

Для подготовленных экспертов, количество которых составило 21 человек, был выбран кабинет в корпусе Е НИУ МЭИ, предоставленный кафедрой для проведения эксперимента. В кабинете сохранялся средний уровень освещенности с доступом естественного освещения, так как эксперимент проводился днем. Для наблюдателей был подготовлен ноутбук с правильными параметрами яркости дисплея. Таким образом, параметры, закладываемые в программу обработки синтетических изображений и условия просмотра получившихся изображений совпадали, что гарантирует грамотные итоги проведенного эксперимента. Для более точной понимания задачи, стоящей перед наблюдателями, был написан инструктаж, который точно и поэтапно описывает действия, которые необходимо выполнить наблюдателю.

Также для оценки изображений и последующем создании шкалы ощущений были составлены психометрические шкалы, которые пояснят значение каждой оценки и будет служить экспертам инструментом определения своих выводом по представленным изображениям.

Для изображений с полотнами живописи:

- *1 – не различимо;*
  Изображение плохо просматривается, не видно, что происходит на полотне. Граница шкалы снизу
- *2 – плохо различимо*
  Изображение слабо видно, мелкие детали не различимы, видно только доминанту;
- *3 – видно;*

Детали изображений различимы. Они могут быть легко распознаны.

- *4 – хорошо видно;*
  Человек легко может распознать мелкие детали на полотне. Чётко просматривается, что хотел донести автор.
- 5 – отлично видно
  Идеальные условия для просмотра. Момент, когда абсолютно все детали различимы и чётко просматриваются без искажений. Верхняя граница шкалы.

Для изображений с скульптурами:

- *1 – не различимо*
  Фигура плоская, объем и форма не просматриваются. Граница шкалы снизу.
- *2 – плохо различимо*
  Слабо видно тени и форму фигуры, фигура еще не объемная.
- *3 – видно*
  Тени образуются на фигуре, но слабо. Фигура не выглядит плоской.
- *4 – хорошо видно*
  Прослеживается форма фигуры. Тени на фигуре начинают образовывать объем, но они «ломают» восприятие фигуры.
- *5 – отлично видно*
  Идеальные условия для просмотра. Сценарий освещения, когда тени, образованные на фигуре, выделяют объем фигуры и не «ломают» форму. Верхняя граница шкалы.

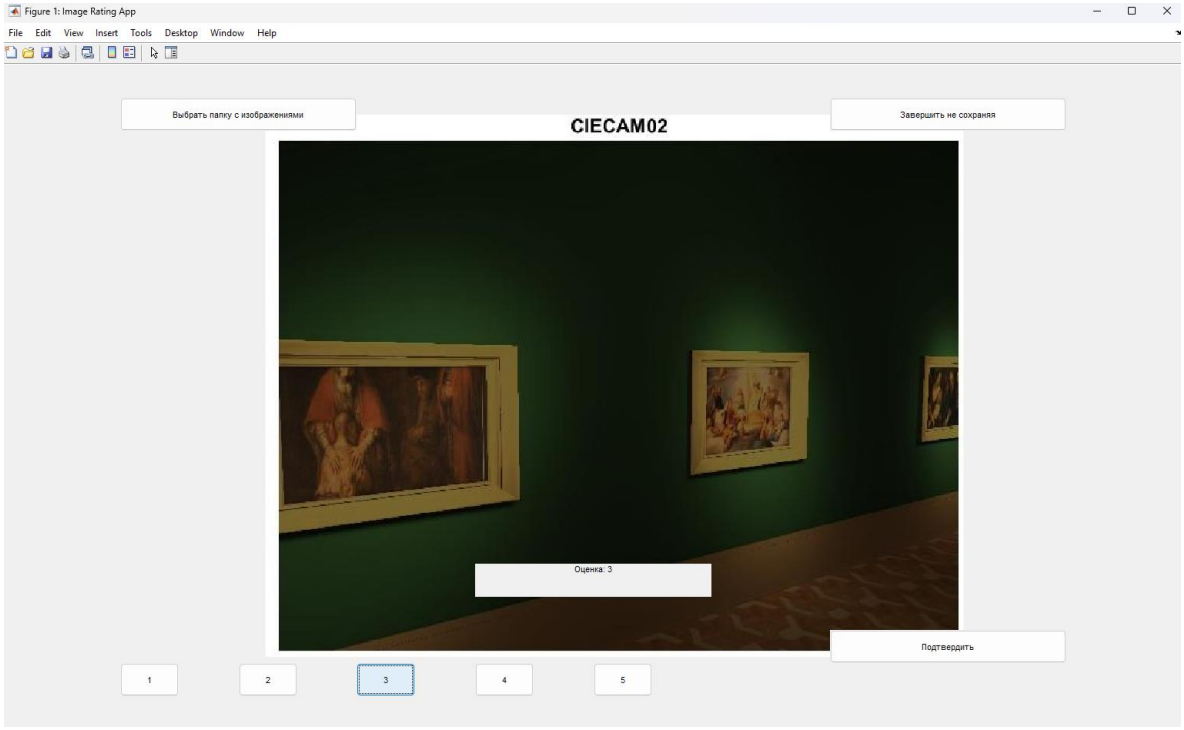


Рис. 16 Интерфейс программы получения оценок

IV. Результаты экспериментов

В исследовании принимали участие наблюдатели, которые давали свою оценку качеству полученных изображений и сценариям тенеобразования. Собранные данные были собраны в формате таблиц Excel для дальнейшего анализа. Для удобства интерпретации результатов была разработана программа для построения гистограммы нормального распределения оценок. Это позволило определить наиболее эффективный подход к созданию «правильной» тени. Однако для полного анализа данных требовалась статистическая обработка. Это позволило оценить точность и надежность полученных результатов. Важно отметить, что измерения всегда связаны с определенными ограничениями, обусловленными различными факторами, включая субъективные ошибки восприятия и внешние условия. Поэтому наблюдателей опрашивали несколько раз, чтобы уменьшить влияние этих факторов.

Для оценки точности измерений были рассчитаны средняя ошибка измерения и стандартное отклонение. Средняя ошибка отражает степень близости среднего значения измерения к истинному значению, а стандартное отклонение характеризует разброс данных вокруг среднего значения. После выполнения необходимых расчетов и визуализации результатов в виде графиков следующим шагом являлся анализ данных и формулирование

окончательных выводов. Для обработки и анализа собранной информации, представляющий статистику в виде гауссового распределения, учитывая математическое ожидание полученной величины.

Результаты написанной программы представлены на рисунках 17 – 23.

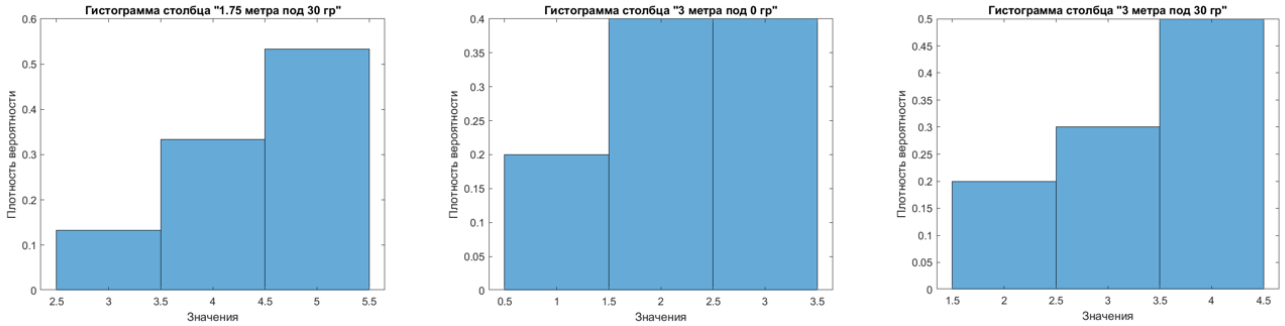


Рис. 17. Гистограммы распределения статистики натурного эксперимента (куб)

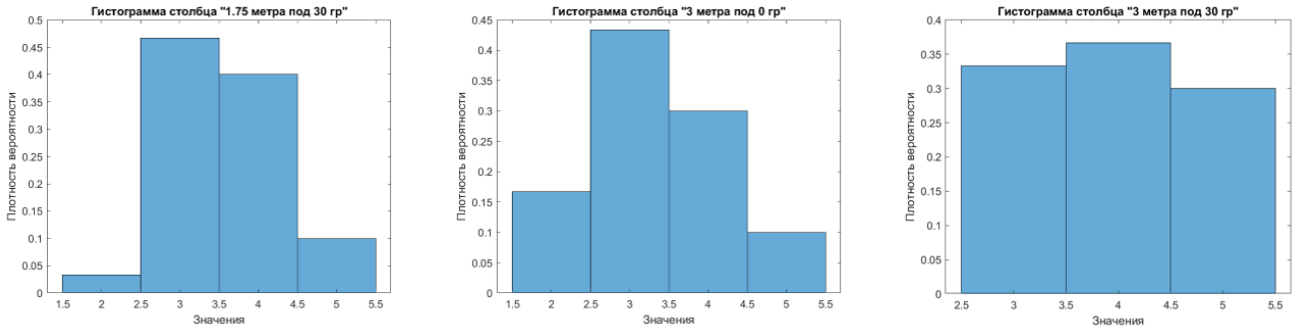


Рис. 18. Гистограммы распределения статистики натурного эксперимента (шар)

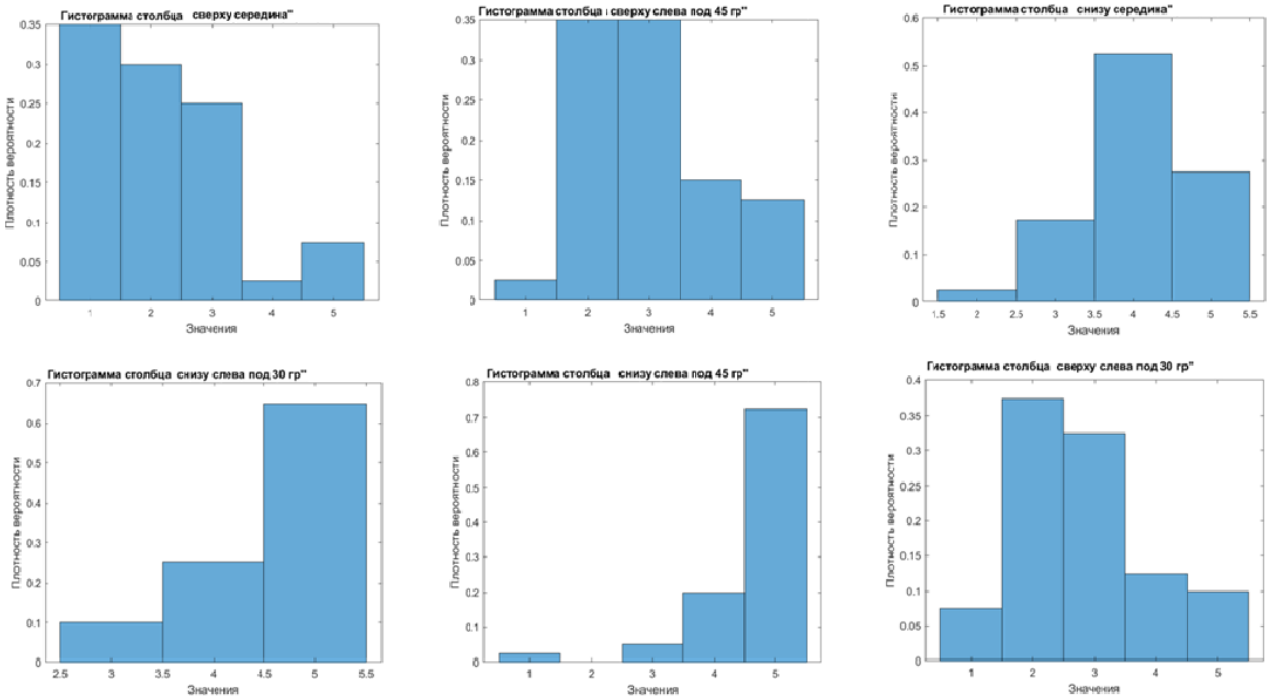


Рис. 19. Гистограммы распределения статистики программного эксперимента (куб)

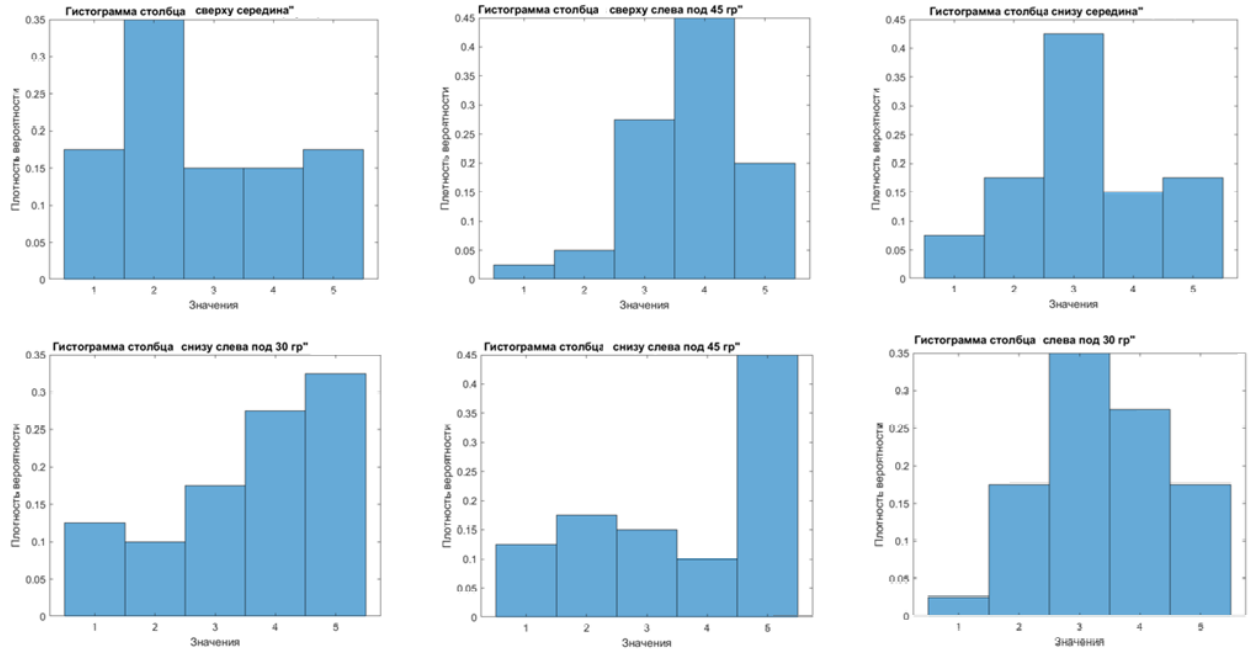


Рис. 20. Гистограммы распределения статистики программного эксперимента (шар)

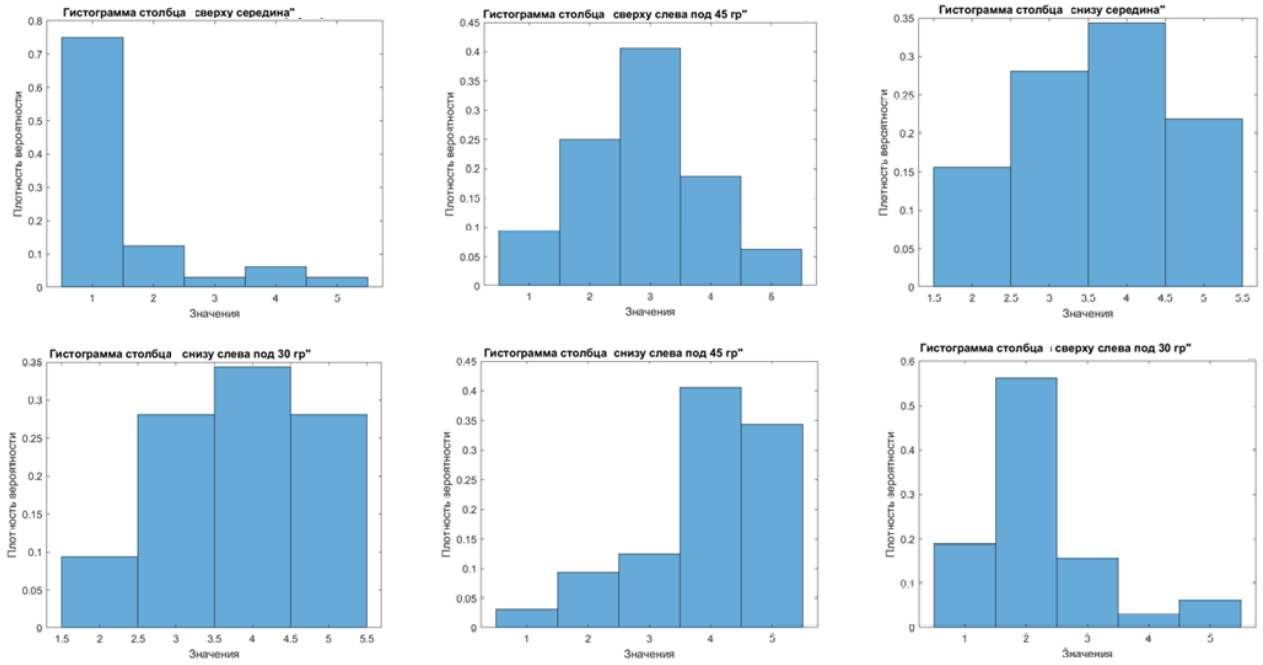


Рис. 21. Гистограммы распределения статистики программного эксперимента (цилиндр)

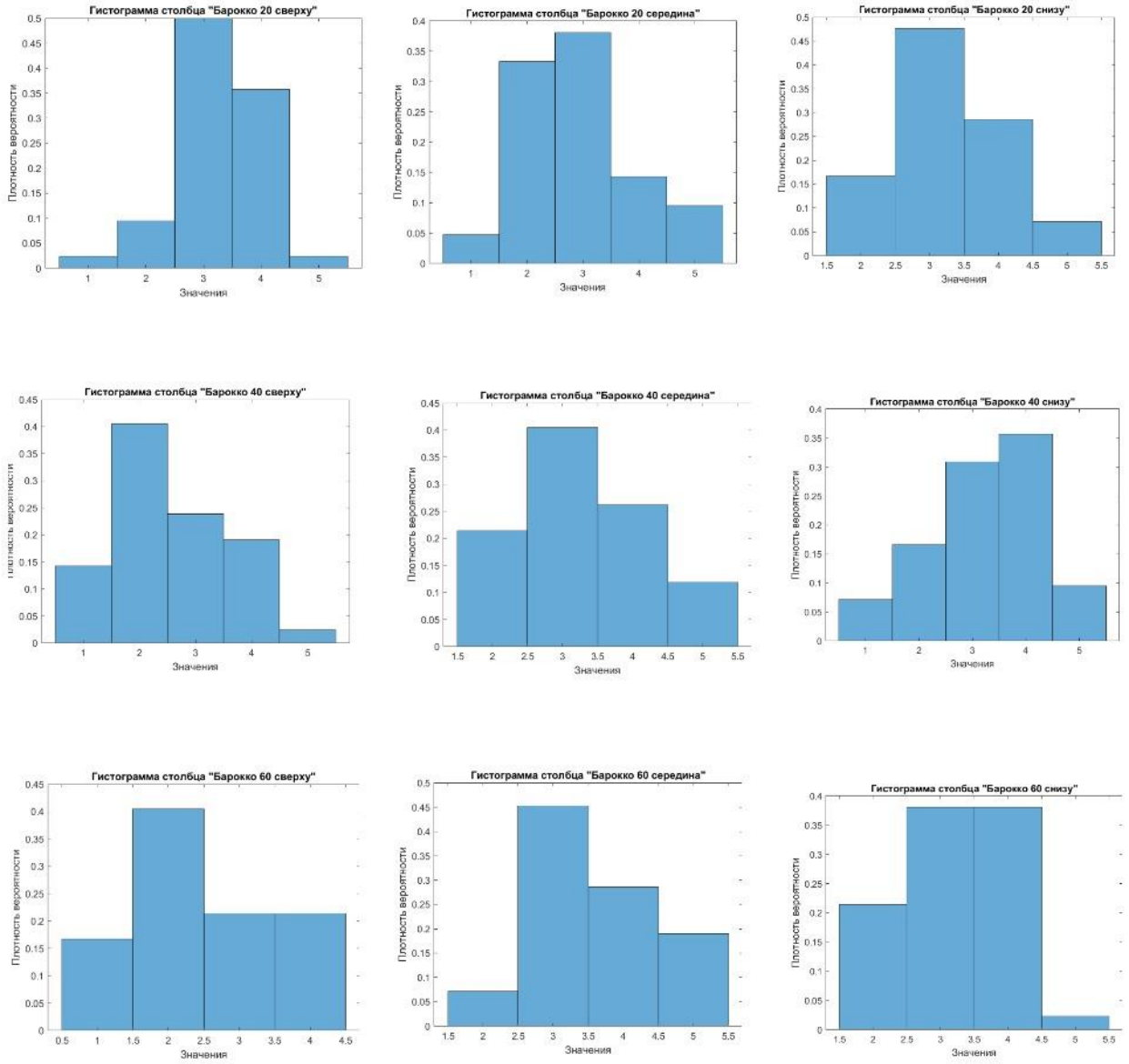


Рис. 22. Гистограммы распределения оценок ощущения картин Барокко.

Анализ освещения произведений искусства в музее с помощью цветовых моделей CIECAM02 позволяет оценить их цветовое качество и сравнить их цветовое восприятие в разных условиях, позволяет определить оптимальные параметры освещения, которые обеспечивают комфортное, привлекательное и достоверное цветовое восприятие экспонатов, позволяет выявить проблемы и недостатки существующего освещения и предложить рекомендации по их устранению или улучшению.

После проведения серии экспериментов и исследований, сделан вывод, что наиболее эффективным приемом освещения для картин из разных эпох и авторов оптимальной и наиболее приятной для экспертов является подсветка ниже глаз среднестатистического наблюдателя с использованием оптики в 60 градусов.

Этот метод позволяет создать мягкий, рассеянный свет, который подчеркивает текстуру и глубину произведения, делая его более объемным и выразительным. Кроме того, такой подход помогает избежать резких теней и бликов, которые могут отвлекать внимание зрителя от главного объекта и высветить фон картины – стену, которая также играет немаловажную роль в восприятии картины.

Оптика в 60 градусов обеспечивает равномерное распределение света по всей поверхности произведения, что делает его более гармоничным и приятным для глаз. Это особенно важно для работ, выполненных в темных тонах или имеющих сложную структуру.

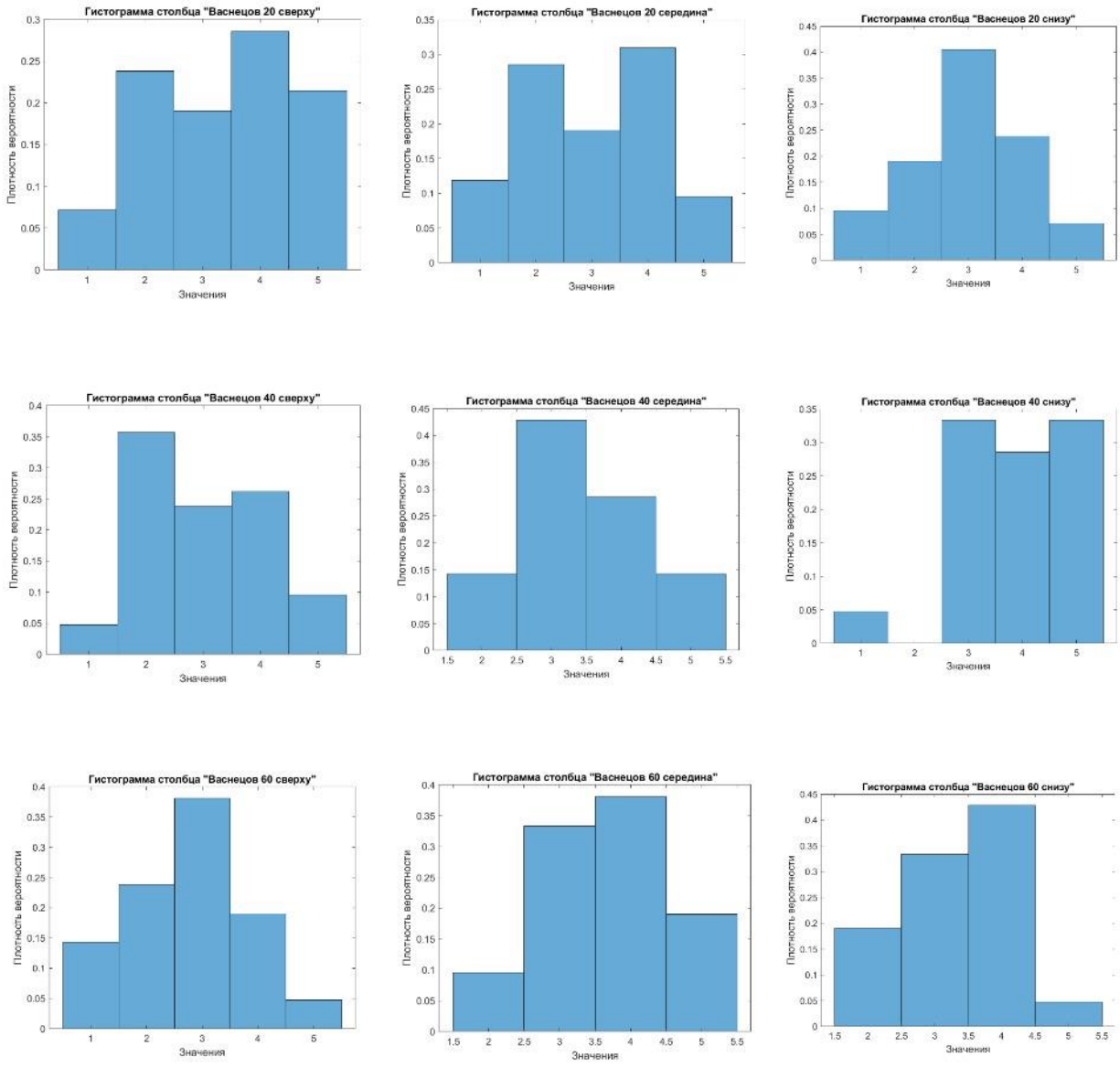


Рис. 23. Гистограммы распределения оценок ощущения картин Васнецова.

Таким образом, использование подсветки ниже глаз среднестатистического наблюдателя с оптикой в 60 градусов является оптимальным решением для эффективного освещения представленных в работе произведений искусства жанров Барокко и картин Васнецова. Реалистичный метод представления освещения картин позволяет подчеркнуть красоту и уникальность каждой работы, делая ее более привлекательной и запоминающейся для зрителей.

В работе удалось провести моделирование вживую в лаборатории света «Light Point» светотехнической компании «Точка Опоры» и в программе DIALux evo. По данным моделированиям были проведены эксперименты с экспертами компании и студентами кафедры «Светотехники» Национального исследовательского университета «Московского энергетического института». В ходе исследования был проведен экспертный опрос для оценки восприятия освещения для образования «правильной» тени на скульптурах среди сотрудников светотехнической компании, а также для оценки восприятия полученных изображений среди светотехников 4 курса бакалавриата, а также 1-2 курсов магистратуры НИУ «МЭИ». В исследованиях приняло участие 35 человек, из которых 15 работники «Точки Опоры» и 20 студенты. По результатам опросов выяснилось, что эксперты в натурном эксперименте отдают предпочтение освещению, которое подразумевает под собой расположение осветительного прибора на высоте 1,75 м от пола, нацеленный под углом 30° относительно куба и

на высоте 3 м от пола, нацеленный под углом 30° относительно шара. Что касается студентов, те при проведении эксперимента выбрали вариант освещения, при котором прожектор расположен на высоте 0,5 м под углом 45° относительно фигуры, то есть относительно каждой фигуры был выбран один и тот же вариант освещения.

V. Обсуждение результатов

Главной задачей построения шкалы ощущений - изучение закона соответствия между психическими и физическими переменными. В данном контексте основное внимание преимущественно уделялось стимульным переменным, поскольку предполагалось отсутствие влияния на ответы испытуемого в эксперименте так называемых несенсорных факторов: изменение функционального состояния человека, мотивации, влияние получения дополнительной информации об экспериментальной ситуации и др. Эти предположения, принимаемые по умолчанию, отражены в процедурных особенностях пороговых измерений и общем представлении о пороге как мере сенсорной чувствительности. [7] Отметим важную и очень характерную особенность пороговых методов: те статистические показатели, которые используются в этих методах в качестве пороговых мер, на самом деле являются, строго говоря, мерами исполнения сенсорной задачи, решаемой испытуемым в определенных условиях, так как определяются не только величиной сенсорной чувствительности испытуемого, но и рядом факторов несенсорной природы, от которых также зависит его ответ. Тем не менее, пороговые методы обладают рядом преимуществ, таких как простота, незначительные временные затраты на проведение измерений, удобство выражения пороговых показателей в единицах интенсивности стимула. Для построения шкалы ощущений обратимся к полной полученной статистике по результатам исследований на нахождение наилучших с точки зрения наблюдателей приемов освещения произведений искусства. Необходимость построения графика вызвана тем, что наблюдатель не всегда может точно поставить оценку увиденном изображению. В некоторых случаях он колеблется в итоговом решении и суть графика наглядно показать расхождение поставленной оценки и предполагаемой. Представив оценки в виде гистограммы нормального распределения, аппроксимируем ее подобранной Гауссовой функцией (рис. 24).

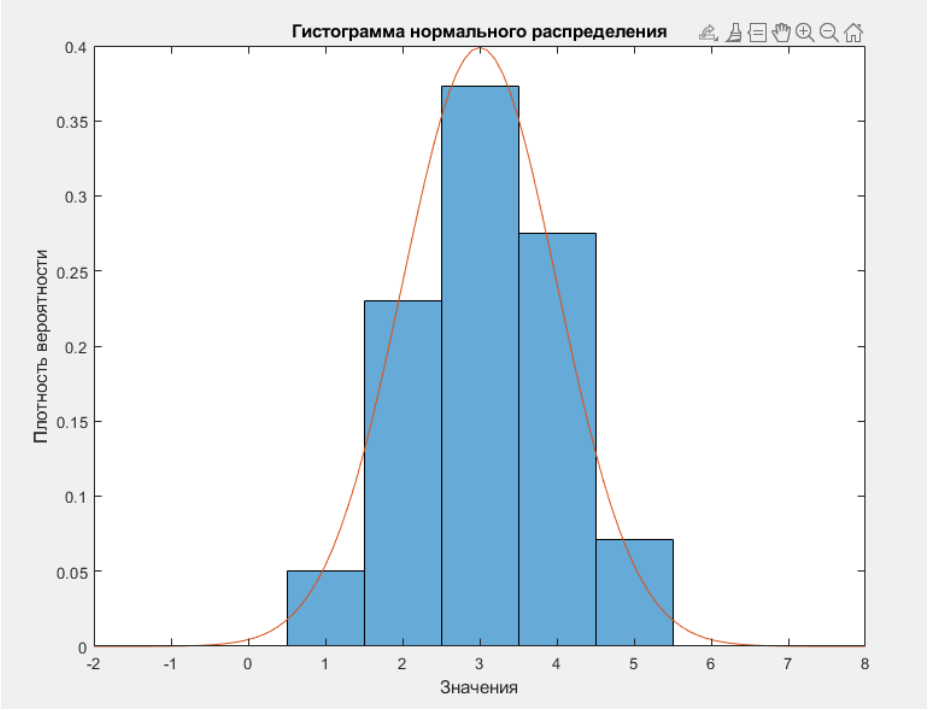


Рис. 24. Гауссово распределение всех полученных оценок.

Отчетливо заметно, что Гауссова функция практически идеально описывает получившуюся гистограмму.

На шкале ощущений задается единица измерения, соответствующая психометрической оценке, полученной от наблюдателя. Количество, а значит и вероятность получения оценки представлена на шкале в виде горизонтальных линий, сообщает о разбросе дисперсии получения этой оценки. Для получения наклонных

линий, символизирующих возможное отклонение оценки в большую или меньшую сторону необходимо решить систему интегральных уравнений. [23]

$$\Delta h_i = \int_{R_{i-1}}^{R_i} f(R)dR = \int_{R_{i-1}}^{\infty} f(R)dR - \int_{R_i}^{\infty} f(R)dR \quad (1)$$

Где

R_i- значение интервала оценки.

$f(R)$- функция нормального распределения этой оценки.

По итогу решение системы уравнений можно представить в виде прямых наклонных линий. Точкой пересечения синих вертикальных отрезков и наклонных линий на графике служит математическое ожидание величины оценки рис. 25.

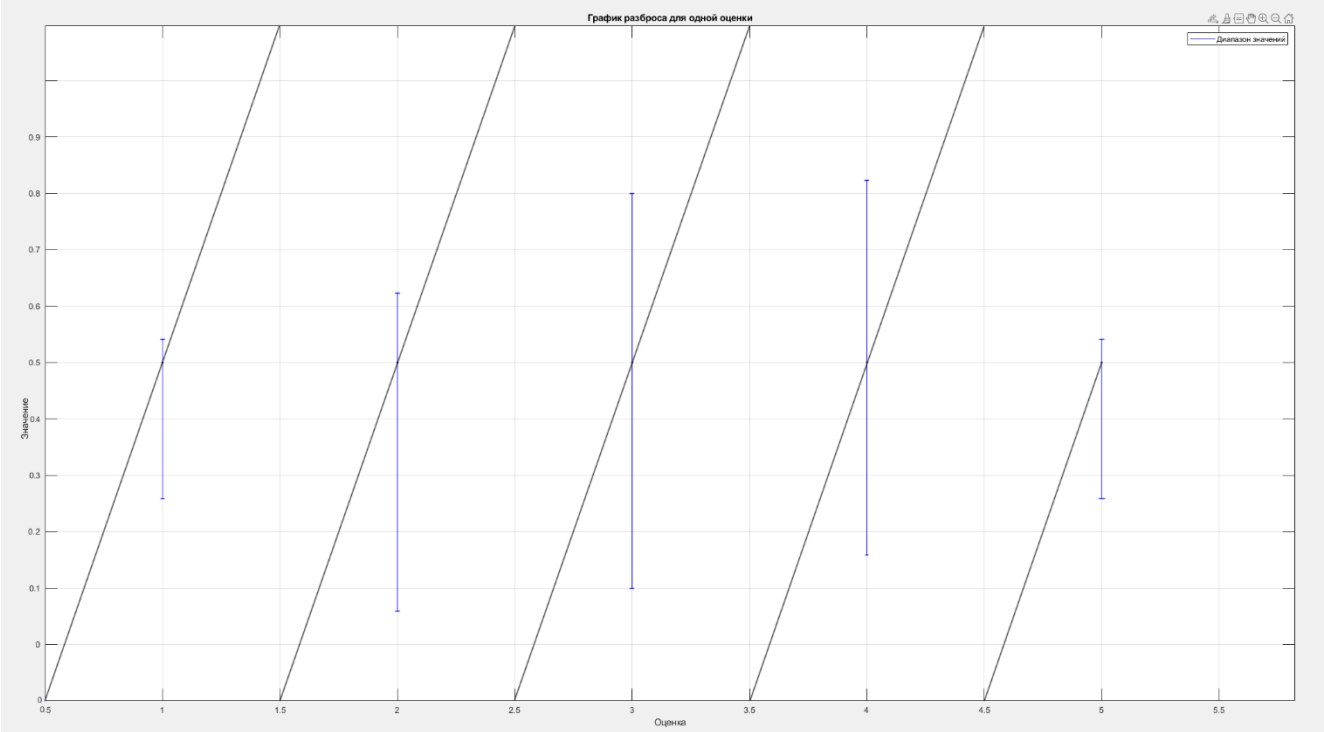


Рис. 25. Построенная шкала ощущений.

VI. Выводы и заключение

Исследования показали, что модель цветового восприятия CIECAM02, которая специально адаптирована для работы с синтетическими изображениями и позволяющая учесть различные факторы как при начальных условиях создания изображения, так и при просмотре на экране компьютера, является отличным инструментом для получения реалистичных результатов для последующих экспериментов.

По проведенному исследованию предложены оптимальные, с точки зрения экспертов, приемы освещения картин двух разных жанров, а также скульптур и результаты оказались такими, что оптика осветительного оборудования и место расположения совпадают и вызывают у наблюдателя желание поставить на специально подготовленной, сформулированной и описанной психометрической шкале оценку выше, по сравнению с остальными представленными в ходе эксперимента приёмами.

Так же визуализирована шкала ощущений на основе множества оценок, которая в свою очередь демонстрирует возможное отклонение физической выставленной оценки от предполагаемой экспертом.

Подводя итог работы, можно смело судить, что такой подход к анализу и усовершенствованию с точки зрения скорости и достоверности получаемых результатов может благотворно сказаться на тенденциях по изучению и модернизации проектирования музейного освещения и даёт стимул к дальнейшим исследованиям и движениям к созданию новых нормативных актов в этой области науки.

Список литературы

1. Боос Г.В. Освещение музеев как одно из приоритетных направлений светотехнической отрасли России.: «СВЕТОТЕХНИКА», 2022, № 6, с. 41-44.

2. V. Budak, I. Galliamov, D. Kiselev, Comparative analysis of color appearance models for realistic display of synthetic images, Light & Engineering 2023. №2. С.76-80.

3. Фершильд М.Д. Модели цветового восприятия: пер. с англ.: Издательство, 2004. 439с.

4. https://light-point.club/about/

5. https://www.svetcom.ru/poleznaya-informatsiya/4000-5000K/

6. Будак В.П., Галлямов И.Р. Киселев Д.А. Сравнительный анализ моделей цветового восприятия для реалистического отображения синтетических изображений, Светотехника. Москва, 2023. № 2. С. 76-80.

7. А. Н. Гусев, И. С. Уточкин Психологические измерения: Теория. Методы: Общепсихологический практикум / А. Н. Гусев, И. С. Уточкин. — М.: Аспект Пресс, 2011. — 317 с.

8. Боос Г.В. Освещение музеев как одно из приоритетных направлений светотехнической отрасли России.: «СВЕТОТЕХНИКА», 2022, № 6, с. 41-44.

9. Кузякин Б.Г. Особенности освещения экспозиций в Государственном Эрмитаже.: «СВЕТОТЕХНИКА», 2018, с. 10-13.

10. Куйдин Д.С., Фомин А.Г. Освещение музея современного искусства.: «ИННОВАЦИОННАЯ СВЕТОТЕХНИКА», 2023, № 1, с. 115-118

11. Волгина Л.Е. Свет в музее: опыт и проблемы.: «СВЕТОТЕХНИКА», 2018, с. 14-18.

12. Новаковский Л.Г. Освещение произведений живописи, графики, полиграфии, фотографии – проблемы и возможные решения.: «СВЕТОТЕХНИКА», 2018, с. 19-27.

13. Щепетков Н.И. Архитектурное освещение в музеях.: «СВЕТОТЕХНИКА», 2018, с. 28-35.

14. Черняк А.Ш., Кузнецова А.Б., Барцева А.А. Измерения параметров освещения залов и экспонатов Государственного Эрмитажа и Государственной Третьяковской галереи.: «СВЕТОТЕХНИКА», 2018, с. 45-49.

15. Белякова М.П. Работы по модернизации освещения в зале № 277 Государственного Эрмитажа.: «СВЕТОТЕХНИКА», 2018, с. 62-64.

16. Койнов С.В., Ходырев Д.М. Музейное освещение – подход, пример и направление движения.: «СВЕТОТЕХНИКА», 2018, с. 65-70.

17. Янкина А.С. Классификация круглой и рельефной скульптуры.: «Научный электронный журнал меридиан», 2020, с. 102 – 104

18. Искусство скульптуры [Электронный ресурс] – URL: https://infourok.ru/referat-po-izobrazitelnomu-iskusstvu-po-teme-iskusstvo-skulptury-6053091.html (дата обращения 24.11.2023)

19. Скульптуры: жанры и вид [Электронный ресурс] – URL: https://goodcoins.su/antic/sculpture/ganrsculpt.htm (дата обращения 12.10.2023)

20. Станковая скульптура [Электронный ресурс] – URL: https://www.derevorez.ru/skulpt/vidsk/37-stansk.html (дата обращения 25.12.2023)

21. Ли О.Д., Лебедкова С.М. Моделирование вариантов освещения скульптур в залах музея.: «ИННОВАЦИОННАЯ СВЕТОТЕХНИКА», 2022, С. 75 – 79

22. Атасян Д.А., Шлеюк С.Г. Приницпы Современной организации освещения музейного и выставочного пространства.: «Социально-гуманитарные инновации: стратегии фундаментальных и прикладных научных исследований», 2021, с.847 – 851

23. Котик Г.Г., Матвеев А.Б., Перейма В. В., Тохадзе И. Л. Категорийные оценки качества освещения и их взаимосвязь на психометрической шкале, Светотехника. Москва, 1975. № 7. С. 3-5.

УДК 004.946
DOI: 10.25206/978-5-8149-3873-2-2024-123-135

Проектирование и виртуальное прототипирование компактной системы дополненной реальности с двумерной дифракционной решеткой
Design and virtual prototyping of a compact two-grating augmented reality system

Д.Д. Жданов[1], А. М. Алексеев[1], А.Д. Жданов[1], В.Г. Соколов[1], И.С. Потемин[1], Е. М. Алексеев[1]
[1]Университет ИТМО, Санкт-Петербург, Россия

D.D. Zhdanov[1], A. M. Alekseev[1], A. D. Zhdanov [1], V.G. Sokolov[1], I.S. Potemin[1], E. M. Alekseev [1]
[1]ITMO University, St. Peterburg, Russia

Аннотация. В работе представлена программная модель оптически-прозрачной системы дополненной реальности, построенной по схеме с двумя дифракционными элементами. Отличительной чертой данного устройства является то, что второй дифракционный элемент представляет собой двумерную дифракционную решетку, которая совмещает функции расширения и вывода изображения из световода. Представлена технология изготовления двумерных дифракционных решеток. Предложены формулы для расчета направления лучей, дифрагированных на дифракционной решетке, включая двумерную решетку. Предложены два алгоритма трассировки лучей и формирования изображений, создаваемых оптическими системами дополненной реальности, содержащих дифракционные решетки, в том числе двумерные. На основе предложенных алгоритмов проведено виртуальное прототипирование ряда систем дополненной реальности. Выявлены особенности проектирования данных систем, их преимущества и недостатки. Даны рекомендации для проектирования и оптимизации систем дополненной реальности, построенных по схеме с двумя дифракционными элементами.

***Ключевые слова*:** световод, дифракционная решетка, дифракционный элемент, дополненная реальность, виртуальное прототипирование, стохастическая трассировка лучей.

Аннотация. The paper presents a software model of an optically transparent augmented reality system built according to a two-diffraction grating scheme. A distinctive feature of this device was that the second diffraction element was a two-dimensional diffraction grating that combined the functions of image expansion and output from the light guide. The technology of manufacturing two-dimensional diffraction gratings is presented. Formulas for calculating the direction of rays diffracted on a diffraction grating, including a two-dimensional grating, are proposed. Two algorithms for ray tracing and image formation generated by augmented reality optical systems containing diffraction gratings, including two-dimensional ones, are proposed. Based on the proposed algorithms, virtual prototyping of a number of augmented reality systems was carried out. The features of the design of these systems, their advantages and disadvantages are considered. Recommendations are given for the design and optimization of augmented reality systems built according to a scheme with two diffraction gratings.

***Ключевые слова*:** Lightguide, diffraction grating, diffraction element, augmented reality, virtual prototyping, stochastic ray tracing.

I. Введение

Оптически-прозрачные системы дополненной и смешанной реальностей, использующие световоды и дифракционные системы ввода-вывода изображения, являются одним из наиболее широко используемых компактных решений соответствующих устройств [1]. Ввод изображения в световод и последующий вывод в сторону глаза пользователя, как правило, осуществляются при помощи различного рода дифракционных оптических элементов [2]. Оптическая система проецирует виртуальное изображение с микродисплея, находящегося в ее фокусе, на бесконечность, что соответствует параллельным пучкам лучей, покидающим оптическую систему с каждой точки изображения на ее микродисплее. Далее, свет попадает в светопроводящую пластину, на поверхность которой нанесена дифракционная решетка. Период решетки и показатель преломления пластины выбираются таким образом, чтобы обеспечить полное внутреннее отражение всех световых лучей, создающих изображение и вошедших в светопроводящую пластину. Свет, претерпевая полное внутреннее отражение, распространяется в пластине и выводится дифракционной решеткой с периодом, равным входной решетки. Это позволяет вывести свет под тем же углом, под которым он был введен в дифракционную решетку, и создать изображение на сетчатке глаза наблюдателя, аккомодированного на бесконечность [3]. Если дифракционная эффективность решетки в порядке, который выводит изображение, невысока, а эффективность в

нулевом порядке высока, то глаз наблюдателя видит реальный мир сквозь выходную дифракционную решетку без каких-либо искажений. В результате на сетчатке глаза одновременно формируются изображения реального и виртуального миров [4]. Как и во всех устройствах виртуальной и дополненной реальностей данная схема не лишена проблемы конфликта вергенции-аккомодации человеческого зрения, поскольку виртуальные объекты находятся на бесконечности, а объекты реального мира могут находиться в любой точке пространства. Данная оптическая схема оптически-прозрачной системы дополненной реальности продемонстрирована на рис. 1.

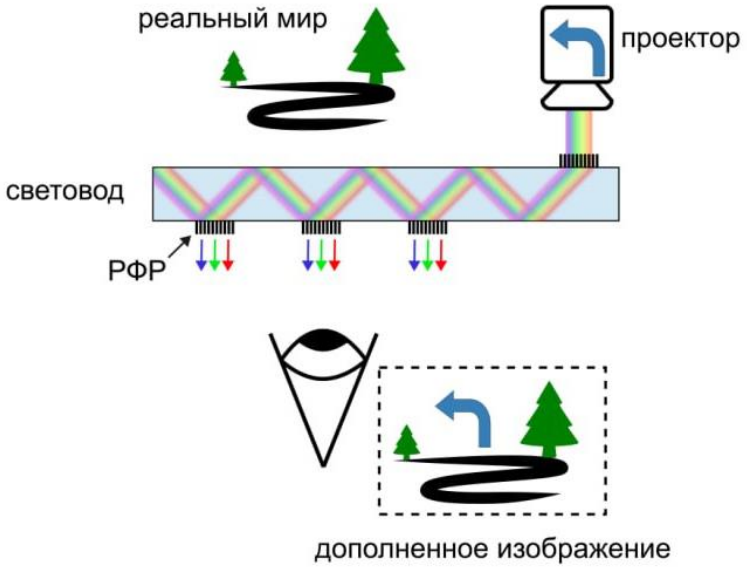


Рис. 1. Общая схема работы оптически-прозрачной системы дополненной реальности

Одна из основных проблем данной схемы – малое значение области, где можно наблюдать изображение (eye box). Эта область соответствует апертуре оптической системы, формирующей виртуальное изображение. Если выходная дифракционная решетка будет иметь больший размер, то это увеличит размер eye box только по направлению световых лучей, сохраняя практически тот же самый размер в поперечном направлении. Поэтому в большинстве случаев для увеличения размера eye box в двух направлениях используется дополнительная дифракционная решетка для расширения пучка лучей, распространяющегося в световоде.

Наиболее распространенная схема реализации подобных устройств содержит три одномерные дифракционные решетки, схематично представленные на рис. 2 (панель слева). Две дифракционные решетки используются для ввода и вывода пучков изображения, а одна – для их размножения, что позволяет увеличить область обзора изображения.

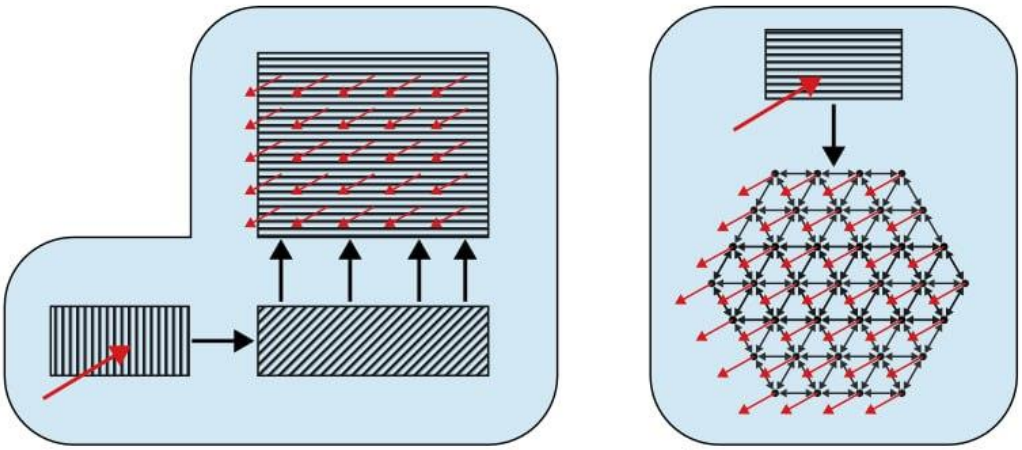


Рис. 2. Схема системы дополненной реальности с использованием трех дифракционных элементов (слева) и двух дифракционных элементов (справа)

Дифракционные решетки могут быть созданы либо в объеме самого световода (как правило, методом голографической записи), либо вытравлены или нанесены на его поверхность. Дифракционная эффективность используемых решеток должна иметь минимальную зависимость от длины волны и угла падения, а также балансировать вклад пучков изображения, приходящих с разных направлений. В противном случае виртуальное изображение будет иметь неравномерную яркость и хроматические аберрации. Это обеспечивается за счет варьирования дифракционной эффективности по поверхности выходной дифракционной решетки, что формирует равномерную яркость по всей поверхности eye box. Одно из возможных решений – это использование мета-решеток, состоящих из трехмерных нано-элементов, каждый из которых выполняет роль субволновой антенны. Данный подход также позволяет использовать двумерные решетки, объединяющие в себе функции вывода и размножения световых пучков, распространяющихся от входной дифракционной решетки.

Важным моментом является возможность производства таких дифракционных решеток. Изготовление дифракционных решето для устройств дополненной реальности представляет собой сложную технологическую задачу, требующую создания дифракционных элементов нанометровых размеров на площади в несколько квадратных сантиметров. Наиболее перспективными подходами к решению этой задачи являются использование наноимпринт литографии [5, 6] и контактного копирования голограмм [7, 8].

В основе технологии наноимпринт литографии лежит воспроизведение наноструктур с шаблона посредством механического изменения формы полимерного слоя, нанесенного на целевую подложку. Главным преимуществом данного подхода является то, что он позволяет разделить трудоемкий этап создания наноструктур и их копирование. Создание штампа для дифракционных элементов остается сложной задачей, как правило, осуществляемой при помощи электронной литографии и последующего химического или ионного травления подложки. Однако после создания штампа он может быть использован для быстрого и недорогого переноса наноструктур на поверхность целевых подложек методом термической или ультрафиолетовой наноимпринт литографии. Основные этапы технологического процесса ультрафиолетовой наноимпринт литографии показаны на рис. 3.

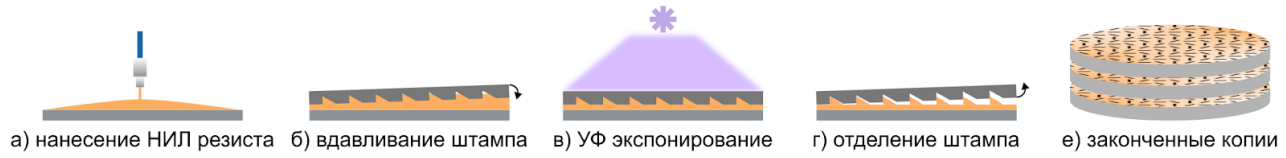


Рис. 3. Основные этапы технологического процесса ультрафиолетовой наноимпринт литографии

После изготовления штампа осуществляется его копирование. Прозрачный штамп погружается в слой наноимпринт резиста, нанесенного на подложку методом центрифугирования. После этого резист отверждается под воздействием ультрафиолета, а штамп может быть использован для создания следующей копии. Наноимпринт литография позволяет воспроизводить наноструктуры размером до 5–10 нм и среднестатистическим отклонением от заданных параметров меньше 1 нм. Так как структуры создаются непосредственно в самом полимерном слое, используемый резист должен иметь подходящий показатель преломления и соответствовать технологическим требованиям наноимпринт литографии [9, 10]. Использование полимера в качестве активного слоя также делает подобные устройства более подверженными механическим повреждениям, что требует нанесения дополнительного защитного слоя.

Таким образом, существуют технологии, позволяющие формировать дифракционные структуры практически любой сложности, однако создание штампа является достаточно дорогой процедурой, и это является серьезным ограничением для реального прототипирования такого рода устройств.

II. Постановка задачи

Важным моментом в создании и удешевлении процесса проектирования и производства устройств такого рода является виртуальное прототипирование, то есть построение виртуальной модели устройства дополненной реальности (оптическая система формирования изображения, входные (InCoupler) и выходные (OutCoupler) дифракционные решетки, световод) и формирование физически корректной визуализации результатов ее работы (модель наблюдателя). Данная модель представлена на рис. 4.

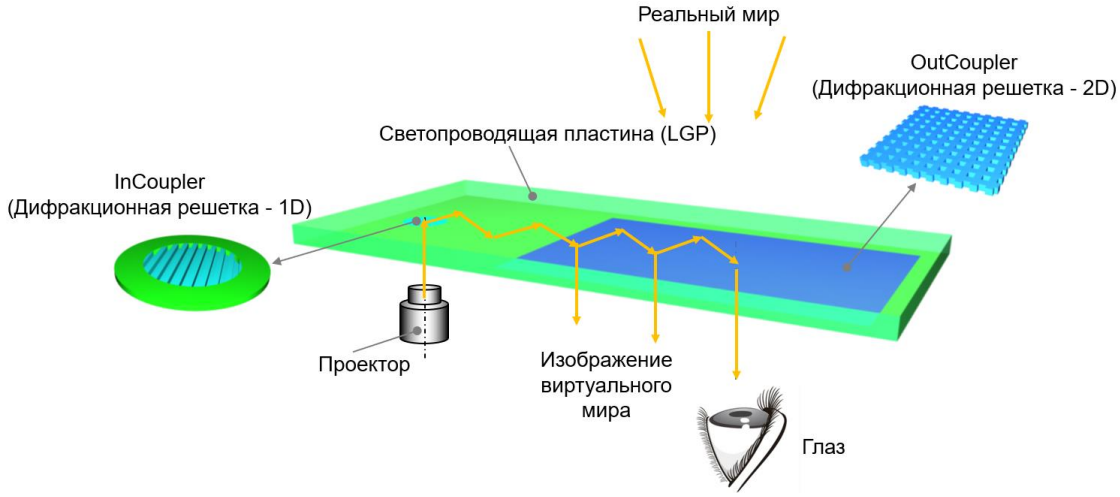


Рис. 4. Модель устройства дополненной реальности и системы физически корректной визуализации результатов ее работы

Как было отмечено выше, одно из возможных решений по уменьшению габаритов светопроводящей пластины и возможному повышению энергоэффективности устройства – это использование двумерных решеток, объединяющих в себе функции вывода и размножения световых пучков, формирующих изображение (см. рис. 2 (панель справа). Поэтому основной упор будет сделан на моделирование оптических устройств дополненной реальности, построенных с использованием двумерных дифракционных решеток.

Основная задача данного исследования – это создание виртуальной модели оптического устройства дополненной реальности, построенного на основе светопроводящей пластины с двумерной решеткой, определение основных параметров модели, влияющих на целевые показатели работы устройства дополненной реальности (распределение яркости по поверхности eye box), выполнение тестовых расчетов и анализ возможных недостатков работы данного устройства.

III. Теория

Главными элементами модели устройства дополненной реальности являются дифракционные решетки: одномерная решетка, вводящая изображение в оптическую систему, и двумерная решетка, расширяющая и выводящая изображение из светопроводящей пластины. Остальные элементы устройства дополненной реальности могут быть представлены традиционными моделями вычислительной оптики и реалистичного рендеринга. Поскольку информация об оптической системе, формирующей изображение микродисплея, отсутствует, данная модель была заменена на модель идеальной камеры, в фокусе которой находится микродисплей идеального контраста. При наличии параметров оптической системы и микродисплея может быть построена более точная модель, учитывающая аберрации объектива и параметры микродисплея, например, для моделирования эффекта обратной проекции, формирующей ложную засветку изображения. В качестве модели глаза также использовалась идеальная камера с фокусным расстоянием 22.5 мм и диаметром зрачка 3 мм, что соответствует средним параметрам человеческого глаза. Светопроводящая пластина – это стеклянный параллелепипед, боковые грани которого зачернены для устранения возможных бликов. Коэффициенты преломления и отражения света в светопроводящей пластине вычисляются по формулам Френеля, а направления лучей – формулами Снеллиуса. Преобразование луча на дифракционном элементе (одномерной дифракционной решетке) описывается следующей формулой [11, 12]:

$$n'\vec{s}' = n\vec{s} + m\lambda\vec{q} + \gamma\vec{N}, \quad (1)$$

где: $\vec{s}$ – направление падающего луча, n – показатель преломления среды падающего луча, $\vec{N}$ – локальная нормаль к поверхности в точке падения, $\vec{s}'$ – направление драгировавшего луча, n' – показатель преломления среды драгировавшего луча, m – порядок дифракции луча, λ – длина волны, на которой распространяется луч, $\vec{q}$ – вектор локальных пространственных частот, спроектированный на плоскость решетки ($|\vec{q}| = \frac{1}{T}$, где T – период дифракционной решетки в единицах длины волны), γ – нормировочная константа, определяемая из условия $\vec{N} = 1, \vec{s} = 1, \vec{s}' = 1$.

В случае двумерной дифракционной решетки формула (1) несколько меняет вид:

$$n'\vec{s}' = n\vec{s} + \lambda\left(m_x\vec{q}_x + m_y\vec{q}_y\right) + \gamma\vec{N}, \quad (2)$$

где: $\vec{q}_x$ и $\vec{q}_y$ – вектора локальных пространственных частот x и y решеток соответственно, спроектированные на плоскость решетки, m_x и m_y – порядки дифракции x и y решеток соответственно.

Исходя из условия $\left|\vec{N}\right| = 1$, $|\vec{s}| = 1$, $|\vec{s}'| = 1$, вычисляется γ. Делается замена для первого случая:

$$\vec{V} = \frac{n\vec{s}+m\lambda\vec{q}}{n'}. \quad (3)$$

И для второго случая:

$$\vec{V} = \frac{n\vec{s}+\lambda(m_x\vec{q}_x+m_y\vec{q}_y)}{n'}. \quad (4)$$

И вычисляются γ и $\vec{s}'$:

$$\gamma' = \frac{\gamma}{n'}$$
$$\gamma' = -(\vec{V}\cdot\vec{N}) \pm \sqrt{(\vec{V}\cdot\vec{N})^2 - \vec{V}^2 + 1} \quad . \quad (5)$$
$$\vec{s}' = \vec{V} - \left((\vec{V}\cdot\vec{N}) \pm \sqrt{(\vec{V}\cdot\vec{N})^2 - \vec{V}^2 + 1}\right)\vec{N}$$

Знак дискриминанта определяется из условия: $(\vec{s}\cdot\vec{N})(\vec{s}'\cdot\vec{N}) > 0$ – для преломления и $(\vec{s}\cdot\vec{N})(\vec{s}'\cdot\vec{N}) < 0$ – для отражения.

Дифракционная эффективность одномерной решетки с постоянным шагом вычисляется достаточно просто – это преобразование Фурье от набега фазы, создаваемой решеткой [13, 14]. Например, для треугольного профиля дифракционная эффективность по амплитуде равна:

$$D_{m\lambda} = (-1)^m e^{i\alpha_\lambda\pi} sinc(\pi(\alpha_\lambda - m)). \quad (6)$$

А для ступенчатого профиля:

$$D_{m\lambda} = (-1)^m e^{i\alpha_\lambda\pi} \frac{sin(\pi m/M) sin(\pi(\alpha_\lambda - m))}{\pi m \cdot sin(\pi(\alpha_\lambda - m)/M)}, \quad (7)$$

где: M – число ступеней профиля решетки, $\alpha_\lambda = \frac{(n_\lambda - 1)\lambda_0}{(n_{\lambda_0} - 1)\lambda}$, λ – длина волны, на которой работает решетка, λ_0 – длина волны, на которой производилась запись решетки, n_λ – показатель преломления среды решетки, на длине волны которой она работает, n_{λ_0} – показатель преломления среды решетки, на длине волны которой она была записана.

Для двумерных дифракционных решеток используются специальные программы, формирующие матрицы дифракционной эффективности по порядкам m_x и m_y. Для лучевых (энергетических) расчетов выполняется расчет таблиц квадратов дифракционной эффективности $\left\| D_{m_x m_y}(\psi, \sigma, \lambda) \right\|^2$ по всем длинам волн λ и углам падения ψ и σ.

Методы трассировки детерминистических лучей в оптических системах с голограммными элементами были разработаны достаточно давно [12]. Эти методы бои реализованы в компьютерных системах моделирования и проектирования оптических систем, например в компьютерной системе DEMOS [15-18], ориентированной на проектирование систем с голограммными элементами. В данном случае ситуация несколько другая. Распространение световых лучей в светопроводящей пластине носит случайный характер и для моделирования распространения лучей в светопроводящей системе с дифракционными элементами наиболее подходящим решением будет прямая и обратная трассировка лучей. Общий обзор методов, ориентированных на различные виды стохастической трассировки лучей, можно найти в работе [19]. Базовые алгоритмы методов прямой и обратной трассировки лучей для систем, содержащих голограммные элементы со светопроводящими элементами, представлены в Табл. 1.

ТАБЛИЦА 1
АЛГОРИТМЫ ТРАССИРОВКИ ЛУЧЕЙ В ОПТИЧЕСКОЙ СИСТЕМЕ ДОПОЛНЕННОЙ РЕАЛЬНОСТИ

Прямая трассировка лучей	**Обратная трассировка лучей**
Define probability of a given diffraction order $P_m(\psi,\sigma,\lambda) = \sum_m \|D_m(\psi,\sigma,\lambda)\|^2$ $P_{m_x m_y}(\psi,\sigma,\lambda) = \sum_{m_x,m_y} \left\|D_{m_x m_y}(\psi,\sigma,\lambda)\right\|^2$	Define probability of a given diffraction order $P_m(\psi,\sigma,\lambda) = \sum_m \|D_m(\psi,\sigma,\lambda)\|^2$ $P_{m_x m_y}(\psi,\sigma,\lambda) = \sum_{m_x,m_y} \left\|D_{m_x m_y}(\psi,\sigma,\lambda)\right\|^2$
do	do

{
Emit ray from light sources according to luminance probability function;
Transfer ray trough the lens (ideal or real with all effects light ghost and rare projection);
Transfer ray on InCoupler grating (randomly choose diffraction order m according to $P_m\ (\psi, \sigma, \lambda)$);
Trace ray in waveguide, if case of the diffraction event:
{
Transfer ray on OutCoupler grating (randomly choose diffraction order m_x, m_y according to $P_{m_x m_y}(\psi, \sigma, \lambda)$;
Accumulate ray on eye detector;
}
} while (Accuracy is not achieved)
	{
	Emit ray from eye detector;
	Transfer ray on OutCoupler grating (randomly choose diffraction order m_x, m_y according to $P_{m_x m_y}(\psi, \sigma, \lambda)$;
	Trace ray in waveguide, if case of the diffraction event:
	{
	Transfer ray on InCoupler grating (randomly choose diffraction order m according to $P_m\ (\psi, \sigma, \lambda)$);
	Transfer ray trough the lens (ideal or real with all effects light ghost and rare projection);
	Accept luminance of the light emitter to view point;
	}
	} while (Accuracy is not achieved)

На основе разработанных алгоритмов в системе физически корректного рендеринга Lumicept был реализован специальный объект (OPTOS) для моделирования и виртуального прототипирования оптических систем с дифракционными элементами методами прямой и обратной трассировки лучей.

IV. Результаты экспериментов

В рамках исследования был проведен ряд тестовых расчетов. Аналогичные расчеты в свое время были выполнены для светопроводящих систем с двумя одномерными дифракционными решетками [20]. В данном исследовании проводились расчеты светопроводящего устройства дополненной реальности с одномерной и двумерной дифракционными решетками. Программное обеспечение, используемое для расчетов, – Lumicept [21], компьютерная система физически корректного моделирования оптически сложных сцен. Базовые расчеты проводились методом прямой трассировки лучей. В рамках одного расчета производился расчет девяти изображений, соответствующих девяти положениям глаза в области eye box. Общая схема представлена на рис. 5. Все размеры заданы в мм, размер выходной решетки 35x25 мм, показатель преломления светопроводящей пластины 1.8, ее поверхности не просветлены.

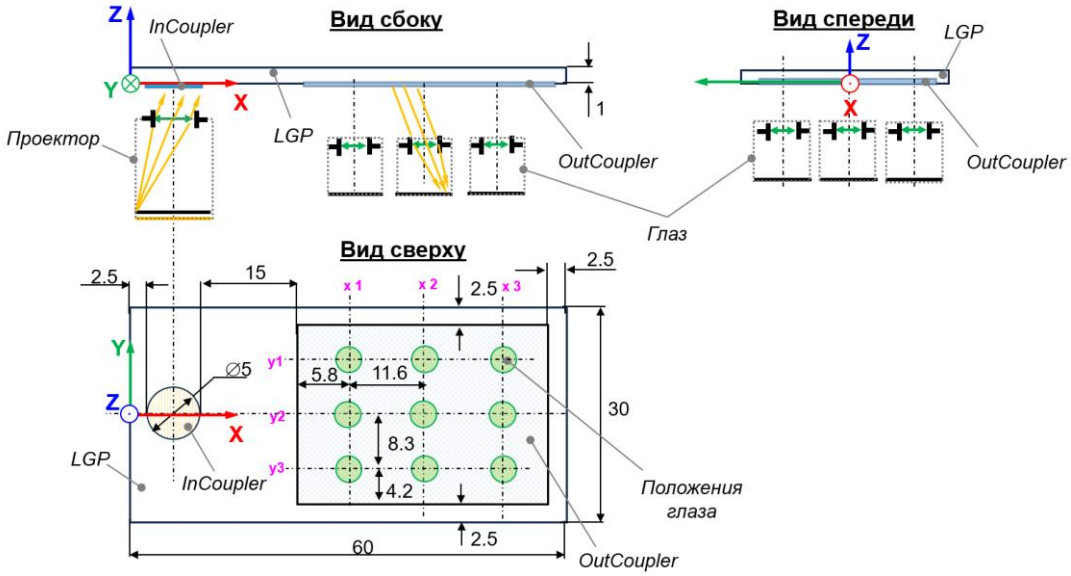


Рис. 5. Общая модель устройства дополненной реальности, используемая в виртуальном прототипировании

Параметры оптической системы формирования изображения представлены на рис. 6. Проекционная система была представлена идеальной линзовой системой с диаметром выходного зрачка 4 мм и фокальным отрезком равным 20 мм. В фокальной плоскости проекционного объектива располагался тестовый объект, представляющий собой модель LCD матрицы. Для повышения эффективности светового расчета (метод прямой трассировки лучей) LCD матрица задавалась прозрачным объектом, подсвечиваемым источником света с ламбертовской диаграммой излучения, ограниченной полем зрения проекционного объектива. При использовании метода обратной трассировки лучей LCD матрица представляла собой самосветящийся объект с ламбертовской диаграммой излучения. При моделировании использовалось два типа изображений на LCD

матрице: равнояркий объект и текстура, ориентированная по четырем сторонам изображения. Мощность излучения LCD матрицы выбиралась из условия 1 лм на входной дифракционной решетке. Расчет проводился в узком спектральном диапазоне в районе 540 нм.

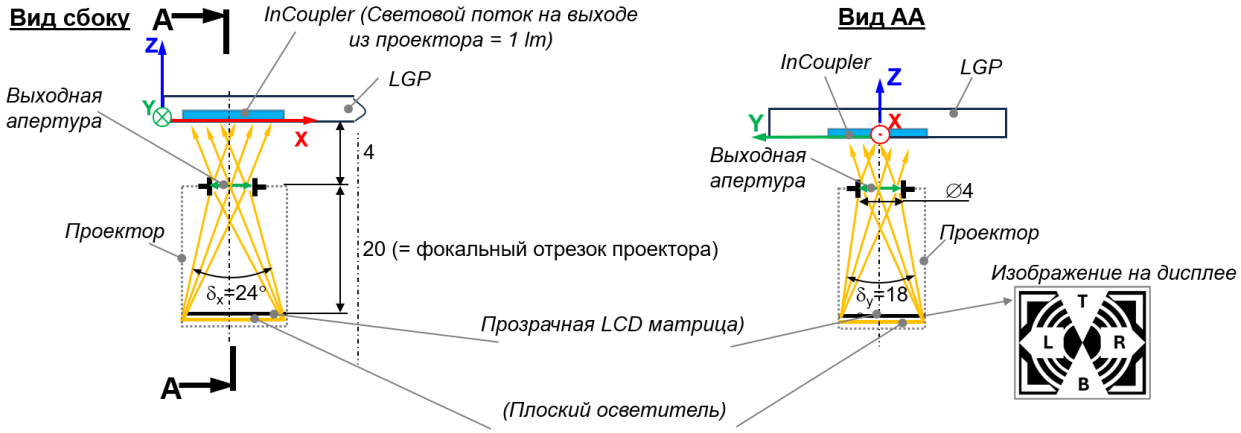


Рис. 6. Модель оптической системы формирования изображения

Параметры модели человеческого глаза (наблюдателя) представлены на рис. 7. Наблюдатель представлен моделью идеальной линзы с параметрами, близкими к параметрам человеческого глаза: фокальный отрезок 22.5 мм и диаметр входного зрачка 3 мм. При расчете использовалось девять позиций наблюдателя, показанных на рис. 5. В методе прямой трассировки лучей наблюдатель накапливал лучи на приемнике излучения (сетчатки глаза), выпущенные с LCD матрицы, а в методе обратной трассировки лучей наблюдатель испускал лучи с сетчатки глаза для считывания яркости с LCD матрицы.

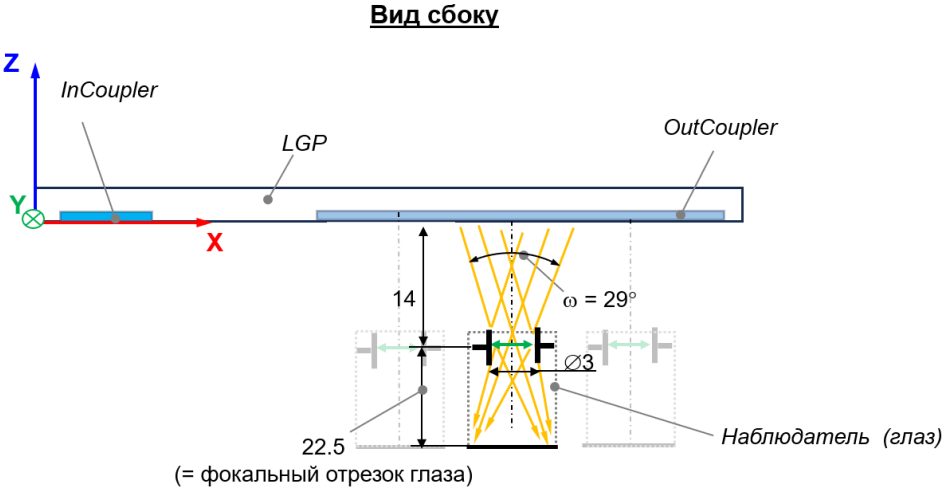


Рис. 7. Модель человеческого глаза (наблюдателя)

Базовые вычисления проводились для выходной двумерной дифракционной решетки с ортогонально ориентированными векторами. Вектора были направлены по осям X и Y. Расчет проводился для двух типов профиля входных решеток: бинарной и близкой по форме к эшелле. Выходная решетка имела бинарную структуру. Периоды входной и выходной решеток были равны и составляли 350 нм. Схема представлена на рис. 8.

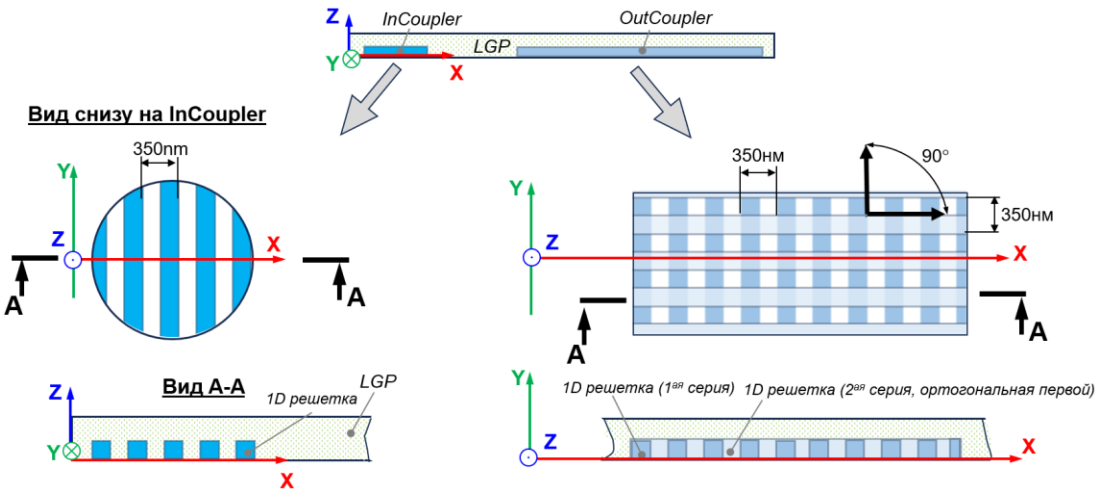


Рис. 8. Параметры дифракционных решеток

Визуализация трасс лучей в данной модели продемонстрирована на рис. 9. На левой части изображения показаны все лучи, распространяющиеся в данном устройстве, на правой части показаны только те лучи, которые попадают в глаз наблюдателя.

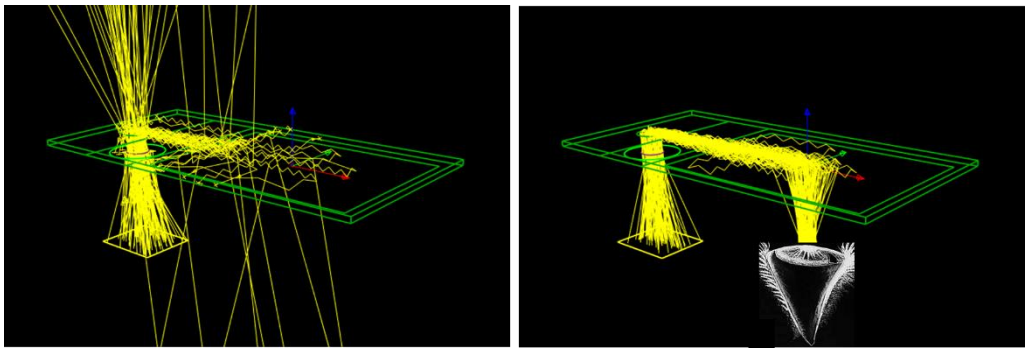


Рис. 9. Визуализация трасс лучей

Далее демонстрируются результаты моделирования для равнояркого источника света для двух типов дифракционных решеток: бинарная – левая часть рис. 10 и эшелле – правая часть рис. 10.

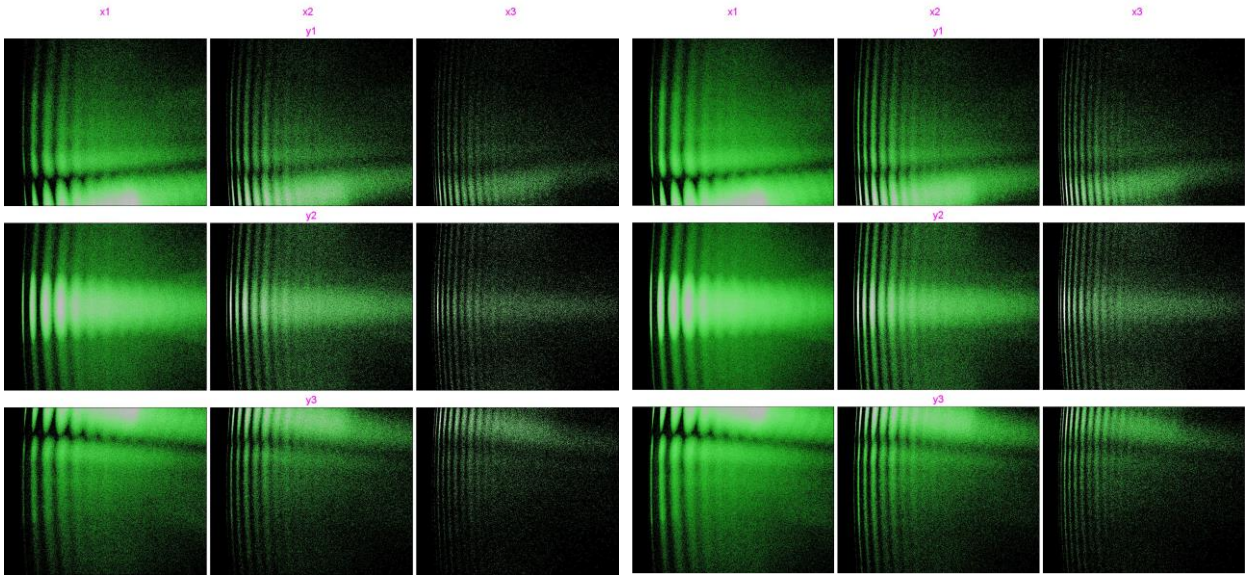


Рис. 10. Результаты расчета для равнояркого источника света. Максимальная яркость визуализации изображений – 5000 кд/м^2/ср.

Аналогичный расчет был проведен для текстурированного источника света на LCD матрице. Результат расчета для входной дифракционной решетки типа эшелле представлен на рис. 11.

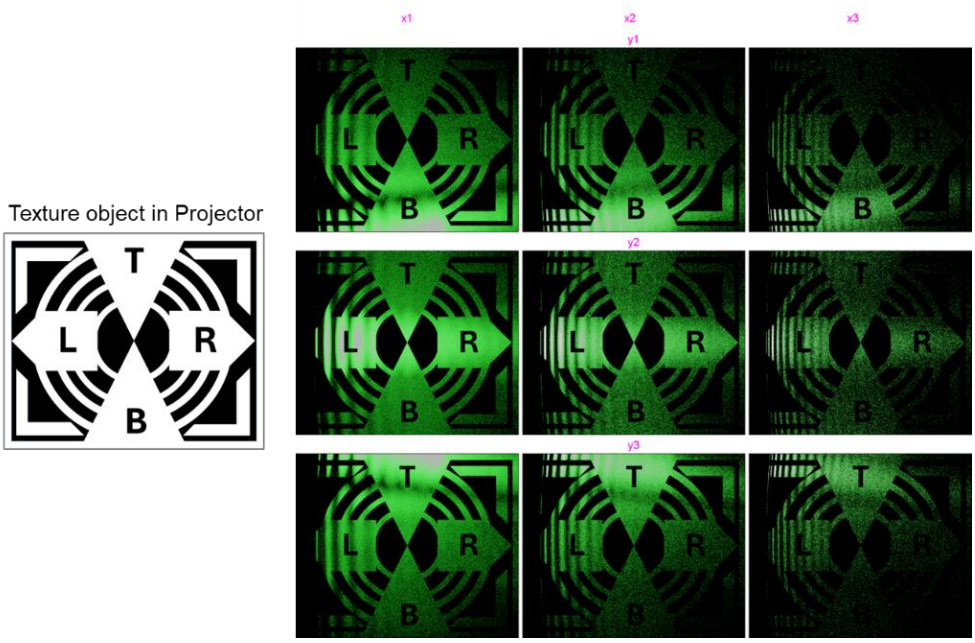


Рис. 11. Результат расчета для текстурированного источника света

Для следующего расчета структура выходной дифракционной решетки была заменена таким образом, что ориентация ее штрихов между векторами решетки составила угол 120°. Входная решетка осталась типа эшелле. Схема представлена на рис. 12.

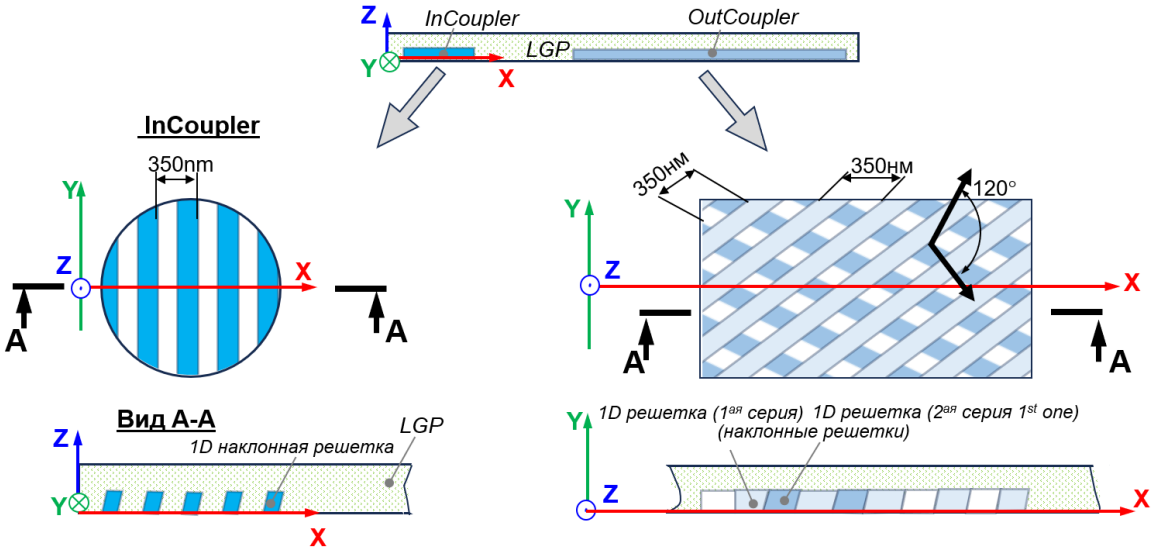


Рис. 12. Параметры дифракционных решеток

Для данной схемы был выполнен расчет с текстурированным источником света, результат представлен на рис. 13.

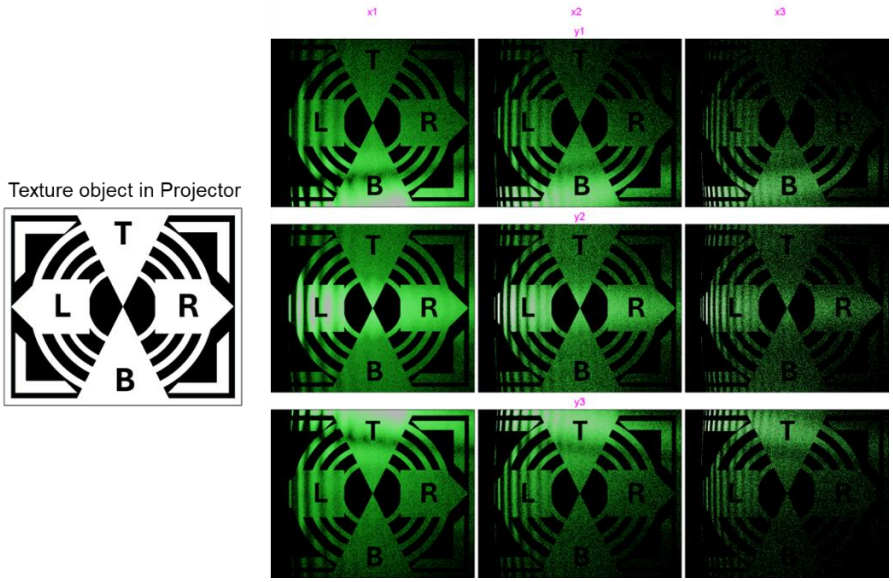


Рис. 13. Результаты расчета для текстурированного источника света. Максимальная яркость визуализации изображений – 5000 кд/м^2/ср.

Один из недостатков данной оптической схемы – это формирование ложного перевернутого изображения в направлении, противоположном наблюдателю. То есть, глядя в очки дополненной реальности с противоположной стороны, можно увидеть информацию, передаваемую с LCD матрицы. Ход лучей, демонстрирующий этот эффект показан на рис.14, а соответствующие изображения в обоих направлениях – на рис. 15, левое – основное изображение, правое – ложное. Изображения на рис. 15 были получены методом обратной трассировки лучей.

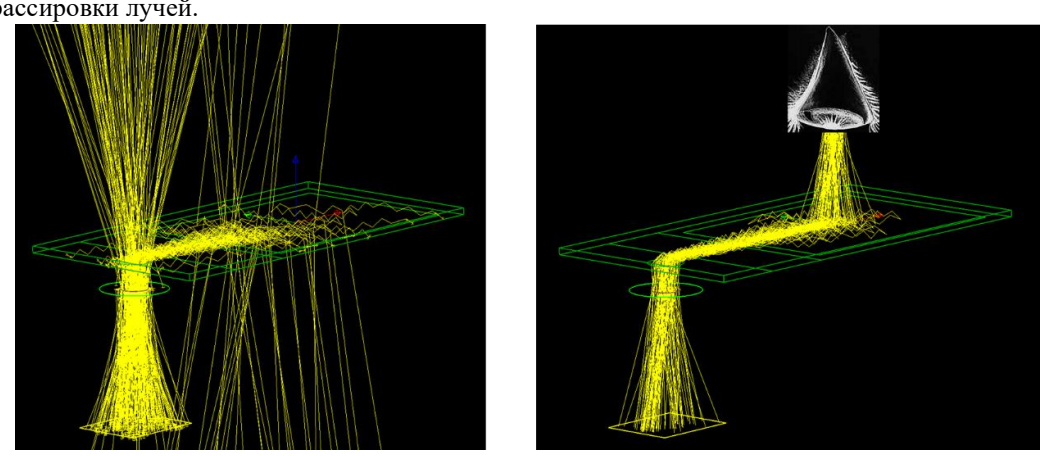


Рис. 14. Визуализация трасс лучей для ложного изображений

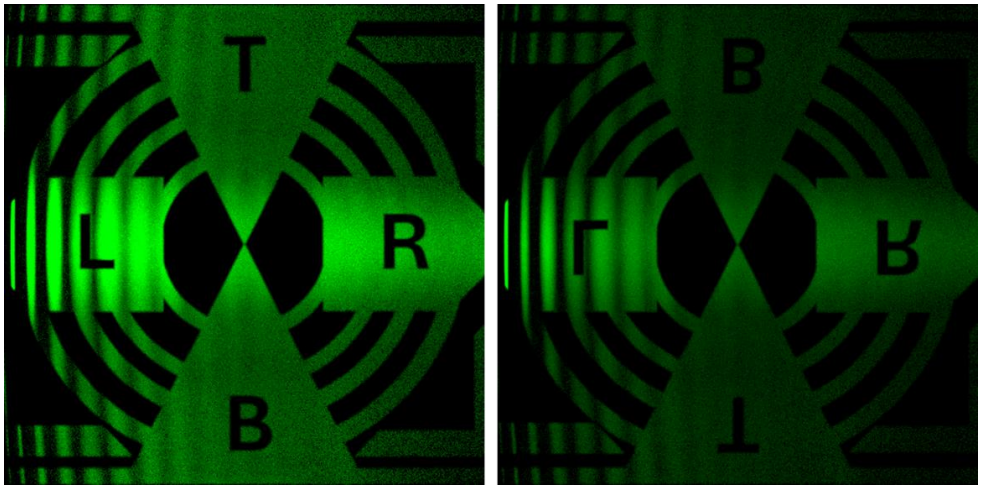


Рис. 15. Основное и ложное изображения

Кроме того, была исследована возможность устранения конфликта вергенции-аккомодации человеческого зрения. Для последней оптической схемы (рис.12) была произведена оптимизация параметров дифракционной структуры решеток на угол поля зрения 19° и выполнено 2 расчета: LCD матрица находилась в фокусе проекционного объектива (изображение на бесконечности), результат представлен на левой части рис. 16, и изображение находилось на конечном расстоянии, равном 941.5 мм (моделировался эффект конвергенции и аккомодации на данное расстояние), результат представлен на правой части рис. 16.

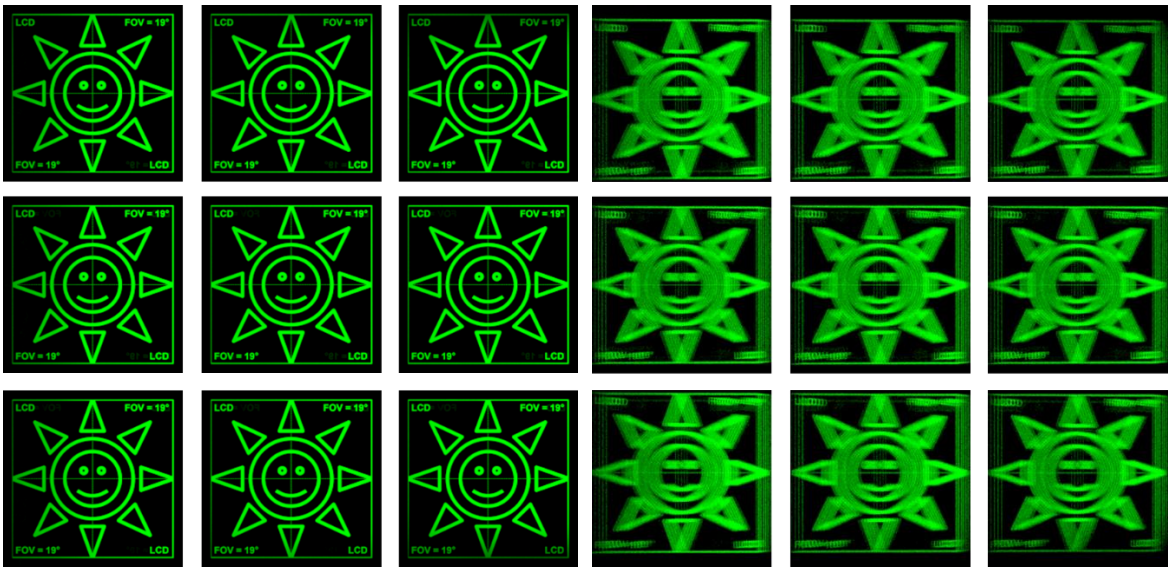


Рис. 16. Изображение, конвергированное на бесконечность и на 941.5 мм

Видно, что попытка устранить конфликт вергенции-аккомодации приводит к двоению изображения. Двоение изображения вызвано тем, что выходная решетка строит группу изображений, разнесенных вдоль ее поверхности, и во входной зрачок наблюдателя могут попадать все эти изображения с соответствующим смещением.

V. Обсуждение результатов

Проведенное исследование показало (например, рис. 10), что для достижения высокого качества изображения необходим правильный подбор всех параметров оптической схемы устройства дополненной реальности. Основными целевыми параметрами данного устройства являются:

- максимальное поле зрения,
- максимальная яркость изображения при заданном световом потоке на входной дифракционной решетке,
- максимальный eye box,
- минимальная вариация яркости видимого изображения,
- максимальное качество изображения, включая отсутствие эффектов двоения и обратной проекции.

Параметры оптимизации:

- показатель преломления светопроводящей пластины,
- толщина светопроводящей пластины,
- период решеток (периоды входной и выходной решеток должны быть одинаковы),
- ориентация и организация структуры выходной двумерной решетки,
- дифракционная эффективность решеток (определяется формой профиля решетки).

Очевидно, что «полосатость» структуры изображения на рис. 10, 11, 13, 15 вызвана тем, что входной зрачок глаза может не захватывать лучи, идущие по направлениям, близким к касательным к рабочей поверхности светопроводящей пластины, то есть шаг размножения пятна изображения выбран неверно. Для устранения данного эффекта можно уменьшить толщину пластины или период дифракционной решетки. Однако, данное решение может вызвать двоение изображения при наличии дефокусировки LCD матрицы.

Устранение ложного изображения, формируемого в обратном направлении, может быть выполнено с помощью линейного поляризационного фильтра, ориентированного в направлении, перпендикулярном поляризации виртуального изображения.

Для достижения максимальной яркости выходного изображения дифракционная эффективность входной решетки должна быть максимальна для минус первого порядка дифракции из воздуха в стекло (при условии падения лучей в пределах поля зрения системы дополненной реальности). При дифракции из стекла в стекло ее дифракционная эффективность в нулевом порядке должна быть максимальна.

Для достижения равномерной яркости по всей видимой поверхности изображения необходимо соответствующее проектирование двумерной дифракционной решетки. Проектирование может осуществляться двумя путями: итерационным прототипированием и подбором параметров дифракционной решетки до тех пор, пока равномерность яркости изображения не будет приемлема, или расчетом распределения дифракционной эффективности по поверхности выходной дифракционной решетки, обеспечивающей равномерность яркости изображения, с последующим решением обратной задачи восстановления параметров дифракционной решетки по вычисленной дифракционной эффективности.

VI. Выводы и заключение

В рамках проведенного исследования была разработана компьютерная модель оптически прозрачного устройства дополненной реальности, построенного по схеме с двумя дифракционными элементами. Особенностью данного устройства являлось то, что вторая второй дифракционный элемент представляла собой двумерную дифракционную решетку, которая совмещала функции расширения и вывода изображения из световода. Для программной реализации модели двумерной дифракционной решетки методами прямой и обратной трассировки лучей были предложены формулы расчета направления дифрагированных световых лучей для заданного направления падения и порядков дифракции (для расчета дифракционной эффективности использовалась внешняя программа). Построенная программная модель была интегрирована в систему физически корректного компьютерного моделирования и виртуального прототипирования, что позволило выполнить виртуальное прототипирование ряда моделей систем дополненной реальности и обозначить основные проблемы, возникающие при проектировании данного рода систем. Данная модель, интегрированная в систему физически корректного компьютерного моделирования, может служить основой для проектирования оптического канала систем дополненной реальности. В рамках дальнейшего исследования планируется интеграция метода расчета дифракционной эффективности двумерных дифракционных решеток в компьютерную модель дифракционного элемента.

Источник финансирования. Благодарности

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, грант № 22-11-00145. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации в рамках международного конкурса по привлечению высококвалифицированных специалистов ITMO Fellowship Program.

Список литературы

1. Xiong, J., Hsiang, E.L., He, Z., Zhan, T. and Wu, S.T., 2021. Augmented reality and virtual reality displays: emerging technologies and future perspectives. Light: Science & Applications, 10(1), pp.1-30.

2. US Patent: US 9,372,347 B1

3. B. C. Kress, Optical Architectures for Augmented-, Virtual-, and Mixed-Reality Headsets // SPIE Press Monograph Vol. PM316 (Society of Photo-Optical Instrumentation Engineers, Bellingham, Washington DC, 2020).

4. Igor S. Potemin, Dmitry D. Zhdanov, Andrey Zhdanov, Nikolay N. Bogdanov, "Virtual prototyping of imaging optics with HOE," Proc. SPIE Vol. 11548, Optical Design and Testing X, 115481S (10 October 2020); doi: 10.1117/12.2574080

5. Sreenivasan, S. V., Nanoimprint lithography steppers for volume fabrication of leading-edge semiconductor integrated circuits // Microsystems & nanoengineering 3, no. 1 (2017): 1-19.

6. Oh, Dong Kyo, Taejun Lee, Byoungsu Ko, Trevon Badloe, Jong G. Ok and Junsuk Rho, Nanoimprint lithography for high-throughput fabrication of metasurfaces // Frontiers of Optoelectronics 14 (2021): 229-251.

7. Chamberlin, G.R. and McCartney, D.J., 1989, March. Contact Copying Of Holographic Transmission Gratings. // Holography Techniques and Applications (Proc. SPIE Vol. 1026, pp. 64-68).

8. Waldern, Jonathan D., Ratson Morad, and Milan M. Popovich, Waveguide manufacturing for AR displays, past, present and future // Frontiers in Optics, pp. FW5A-1. Optica Publishing Group, 2018.

9. Guo, L.J., 2007. Nanoimprint lithography: methods and material requirements. Advanced materials, 19(4), pp.495-513.

10. Asif, Muhammad H., Mariusz Graczyk, Babak Heidari, and Ivan Maximov, Comparison of UV-curable materials for high-resolution polymer nanoimprint stamps // Micro and Nano Engineering 14 (2022): 100118.

11. Тударовский А.И. Формулы для расчета лучей, отраженных дифракционной решеткой / Труды ГОИ, 1958, т.26, вып.152, с.3-12.

12. Парицкая Г.Г., Медведев В.Е. Расчет хода лучей через оптические системы, содержащие голографические дифракционные решетки // Оптико-механическая промышл. 1975. - №3. - С.25-27.

13. Frank Wyrowski, Efficiency of quantized diffractive phase elements // Optics Communications 92 (1992), p,119-126.

14. Ган М.А. О точности определения ЧКХ // ЖНиПФиК, 1975, т. 4, с.269-272.

15. Mikhail A. Gan, Dmitriy D. Zhdanov, Vadim V. Novoselskiy, Sergey I. Ustinov, Alexander O. Fedorov, Igor S. Potyemin, and Sergey N. Bezdidko "Design of optical systems with HOE by DEMOS program" // Proc. SPIE 1574, Intl Colloquium on Diffractive Optical Elements, (1 October 1991); https://doi.org/10.1117/12.50131.

16. Mikhail A. Gan, Dmitriy D. Zhdanov, Vadim V. Novoselskiy, Sergey I. Ustinov, Alexander O. Fedorov, and Igor S. Potyemin DEMOS: state-of-the-art application software for design, evaluation, and modeling of optical systems" // Optical Engineering 31(4), (1 April 1992). https://doi.org/10.1117/12.56130.

17. Mikhail A. Gan, Dmitriy D. Zhdanov, Vadim V. Novoselskiy, Sergey I. Ustinov, Alexander O. Fedorov, and Igor S. Potyemin "DEMOS: new possibility for design and modeling of complex optical systems" // Proc. SPIE 1780, Lens and Optical Systems Design, 17802F (15 April 1993); https://doi.org/10.1117/12.142804.

18. Mikhail A. Gan, Dmitriy D. Zhdanov, Vadim V. Novoselskiy, Sergey I. Ustinov, Alexander O. Fedorov, and Igor S. Potyemin "DEMOS: tool for design and modeling of IR optical systems" // Proc. SPIE 2106, CIS Selected Papers: Iconics and Thermovision Systems, (11 November 1993); https://doi.org/10.1117/12.163677.

19. Фролов В.А., Волобой А.Г., Ершов С.В., Галактионов В.А. Современное состояние методов расчёта глобальной освещенности в задачах реалистичной компьютерной графики. Фролов В.А., and. // Труды Института системного программирования РАН – 2021. vol. 33. no. 2, P. 7-48.

20. Igor S. Potemin, Dmitry D. Zhdanov, Andrey Zhdanov, and Nikolay N. Bogdanov "Virtual prototyping of imaging optics with HOE", Proc. SPIE 11548, Optical Design and Testing X, 115481S (10 October 2020); https://doi.org/10.1117/12.2574080.

21. Hybrid Light Simulation Software Lumicept // Integera.jp. URL: https://integra.jp/en/products/lumicept (date of call 20.07.2024)

УДК 004.925.3
DOI: 10.25206/978-5-8149-3873-2-2024-136-146

Моделирование лесных массивов в системах виртуального окружения с помощью тиражирования облаков точек
Modeling of woodlands in virtual environment systems using point cloud instancing

П. Ю. Тимохин, М. В. Михайлюк
ФГУ «ФНЦ Научно-исследовательский институт системных исследований РАН», Москва, Россия

P. Y. Timokhin, M. V. Mikhaylyuk
Federal State Institution «Scientific Research Institute for System Analysis of the Russian Academy of Sciences», Moscow, Russia

Аннотация. На примере лесных массивов в статье исследуется задача моделирования уникальных виртуальных объектов из квазиповторяющихся трехмерных скалярных полей (облаков точек). Предлагается новый подход к решению этой задачи в реальном времени, основанный на тиражировании видоизмененных экземпляров облака точек (дерева-образца) с помощью аппаратно-ускоренной трассировки лучей. Разработан метод задания псевдослучайной трансформации (положения, ориентации, масштаба и цвета) экземпляров дерева-образца, обеспечивающий эффект уникальности лесного массива, а также метод выполнения такой трансформации на конвейере трассировки лучей графического процессора. Предложенные методы и подход реализованы в виде программного комплекса с использованием языков C++, GLSL и API Vulkan. Апробация комплекса показала, что созданное решение позволяет генерировать в реальном времени уникальные виртуальные модели лесных массивов, содержащие порядка миллиона деревьев. Полученные результаты подтвердили высокую эффективность применения аппаратно-ускоренной трассировки лучей для решения поставленной задачи в системах виртуального окружения, видеосимуляторах, геоинформационных системах и других приложениях.

***Ключевые слова*:** виртуальное окружение, визуализация, лесной массив, трассировка лучей, облако точек, тиражирование, реальное время, GPU

Аннотация. Exemplified on woodlands, the paper studies the task of modeling unique virtual objects from quasi-repeating 3D scalar fields (point clouds). A new approach to solve this task in real-time is proposed, based on instancing point cloud (tree-sample) by means of hardware-accelerated ray tracing. A method to set a pseudorandom transformation (of position, orientation, scale and color) of tree-sample instances, providing woodland uniqueness effect, as well as a method to perform such transformation on the ray tracing pipeline of the GPU, were developed. The proposed methods and approach are implemented in the form of a software complex using the C++, GLSL languages and the Vulkan API. The approbation of the complex showed that the created solution allows unique virtual models of woodlands, comprising about a million trees, to be generated in real-time. The obtained results confirmed high efficiency of using hardware-accelerated ray tracing to solve the task under research in virtual environment systems, video simulators, geoinformation systems and other applications.

***Ключевые слова*:** virtual environment, visualization, woodland, ray tracing, point cloud, instancing, real-time, GPU.

I. Введение

Задача виртуального моделирования лесных массивов актуальна во многих важных сферах человеческой деятельности (машиностроение, сельское хозяйство, ликвидация чрезвычайных происшествий и др.) [1, 2]. Типовые сценарии работы в системах виртуального окружения предполагают визуализацию обстановки как «от лица пешехода» («pedestrian view»), так и с высоты «птичьего полета» («bird's-eye view»). Ввиду этого требуются трехмерные виртуальные модели лесных массивов, которые, с одной стороны, реалистично выглядят вблизи, а, с другой стороны, эффективно масштабируются для покрытия больших площадей.

Один из подходов состоит в построении лесного массива из предварительно созданных полигональных моделей деревьев. Узким местом данного подхода является его геометрическая сложность, ограниченная числом связанных текстурированных треугольников, которое видеокарта способна разместить в своей памяти и визуализировать в реальном времени (со скоростью не менее 25 кадров в секунду). Ввиду того, что качественные высокополигональные («high-poly») модели деревьев могут содержать сотни тысяч треугольников [3], напрямую данный способ подходит для моделирования небольших лесных участков или отдельно стоящих деревьев. Чаще в компьютерной графике реального времени применяются полупроцедурные модели деревьев, у которых остов (ствол и ветки) создается заранее, а листья добавляются процедурно, в процессе визуализации. Для покрытия

больших площадей полупроцедурные деревья дополняются различными сложными техниками управления уровнем детализации [4], сжатия геометрии [5], распределенного тиражирования [6] и др.

Другой подход, исследуемый в настоящей работе, состоит в моделировании лесных массивов на основе трехмерных скалярных полей (облаков точек), полученных путем 3D-сканирования реальных деревьев [7, 8]. По сравнению с полигональным представлением данный подход имеет минимальную избыточность геометрии: 3D-описание дерева содержит только позиции и цвета точек. Чтобы такое 3D-описание можно было использовать в системе виртуального окружения, необходимо преобразовать его в объемную виртуальную модель, например, заменив точки процедурными сферами [9]. Так как качественные 3D-описания деревьев состоят из миллионов точек [10, 11], то, в контексте моделирования лесного массива, это приводит к задаче обработки в реальном времени огромного числа процедурных поверхностей.

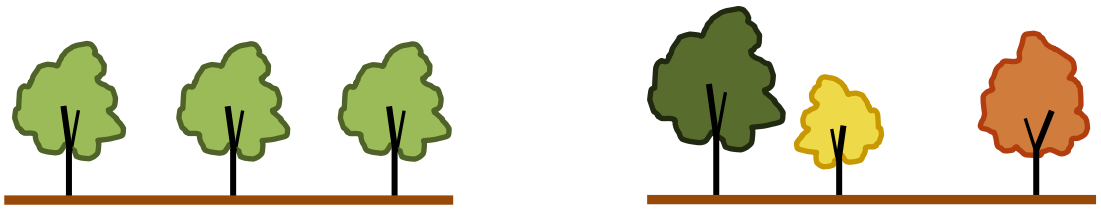
Эффективным путем является решение этой задачи непосредственно на графическом процессоре (GPU) с помощью аппаратно-ускоренной трассировки лучей [12]. Одной из ключевых опций данной технологии является реализованная на уровне архитектуры возможность тиражирования виртуальных объектов «на лету», т.е. в процессе формирования кадра визуализации [13]. Преимущество состоит в том, что вместо хранения множества копий объекта используется лишь один объект (образец) и набор отличительных признаков его экземпляров (в базовом исполнении это массив модельных матриц). Такой подход позволяет существенно сократить расход видеопамяти при тиражировании большого числа экземпляров, а также гибко управлять внешним видом экземпляров, изменяя их отличительные признаки с помощью шейдерных программ.

В контексте задачи моделирования лесного массива мы модифицировали базовый подход [13] для работы с квазиповторяющимися[1] облаками точек, а также добавили в него возможность коррекции цветов экземпляров тиражируемого облака точек.

II. Постановка задачи

Пусть имеется детализированное облако точек, описывающее некоторое реальное дерево (*дерево-образец*). Для простоты будем считать, что в виртуальной среде точки дерева-образца моделируется сферами одинакового радиуса[2]. Рассмотрим задачу моделирования на GPU виртуального лесного массива из $m \times n$ экземпляров дерева-образца, стоящих на плоском участке земной поверхности.

В базовой постановке данная задача может быть сведена к расположению экземпляров дерева-образца в виртуальной сцене с некоторым фиксированным шагом друг относительно друга (см. рис. 1а). Однако, в этом случае смоделированный лес будет иметь выраженную регулярную структуру, что отрицательно сказывается на реалистичности виртуальной сцены и не подходит для современных систем виртуального окружения. В данной работе предлагается расширенная постановка задачи, при которой экземпляры дерева-образца имеют отличные друг от друга шаг расстановки, ориентацию, масштаб и цвет (см. рис. 1б).



(а) (б)

Рис. 1. Базовая и расширенная постановки задачи

Параметризуем рассматриваемую задачу. Для этого вначале определим локальную систему координат (Object Coordinate System, OCS), в которой задаются координаты точек дерева-образца. Начало системы OCS дерева-образца расположено в центре основания ствола, а одна из осей (например, Y) направлена вдоль вертикали роста дерева (оси X и Z лежат в перпендикулярной плоскости некоторым произвольным образом). Для удобства будем считать, что у $(0,0)$-го экземпляра дерева-образца начало и направление осей системы OCS совпадает с мировой системой координат (World Coordinate System, WCS), в которой осуществляется моделирование лесного массива.

Рассмотрим (i, j)-ый экземпляр дерева-образца. Обозначим через $d = (d_x, 0, d_z)$ вектор сдвига данного экземпляра в плоскости земной поверхности, через φ - его угол поворота вокруг оси Y системы OCS, а через k

[1] Повторяющимися с некоторыми вариациями.
[2] Радиус сфер выбирается так, чтобы обеспечивалась визуальная неразрывность сплошных элементов модели дерева-образца.

- его коэффициент масштабирования по трем осям системы OCS. Данным параметрам будет соответствовать некоторая модельная матрица M_{mod}, преобразующая координаты точек экземпляра дерева-образца из системы OCS в систему WCS [14]:

$$P_{wcs} = M_{mod} P_{ocs}, \tag{1}$$

где P_{ocs} и P_{wcs} - координаты точки экземпляра дерева-образца в системах OCS и WCS.

Кроме трансформации геометрии каждый (i, j)-ый экземпляр дерева-образца претерпевает преобразование цвета. Обозначим через $q = (q_r, q_g, q_b)$ вектор коэффициентов коррекции (далее *корректор*) rgb-цветов точек экземпляра дерева-образца:

$$C_{inst} = qC_{src}, \tag{2}$$

где C_{src} и C_{inst} - исходный и скорректированный rgb-цвета точки экземпляра дерева-образца. В данной работе полагается, что остов дерева является условно черным и на него корректор q цвета не оказывает заметного влияния.

Таким образом, чтобы смоделировать лесной массив, необходимо: (1) задать трансформацию (модельную матрицу M_{mod} и корректор q цвета) для каждого (i, j)-го экземпляра дерева-образца и (2) выполнить заданные трансформации на GPU. Далее рассмотрим предлагаемые методы решения поставленной задачи.

III. Теория

8. *Метод задания трансформации экземпляра дерева-образца*

Согласно теории аффинных преобразований [14] представим модельную матрицу M_{mod} в виде комбинации следующих матриц

$$M_{mod} = TR_y S, \tag{3}$$

где T - матрица сдвига экземпляра дерева-образца (пространства) на вектор d (см. раздел II), R_y - матрица поворота пространства на угол φ вокруг оси Y системы OCS, а S - матрица растяжения/сжатия пространства в k раз вдоль осей X, Y, Z системы OCS, которые имеют вид

$$T = \begin{pmatrix} 1 & 0 & 0 & d_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & d_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \; R_y = \begin{pmatrix} \cos\varphi & 0 & \sin\varphi & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\varphi & 0 & \cos\varphi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \; S = \begin{pmatrix} k & 0 & 0 & 0 \\ 0 & k & 0 & 0 \\ 0 & 0 & k & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4}$$

Чтобы трансформация экземпляров дерева-образца выглядела более естественно и разнообразно, координаты вектора d, угол φ, коэффициент k, а также компоненты корректора q в данной работе задаются на основе псевдослучайной выборки из нормального распределения. Рассмотрим это на примере коэффициента k.

Вначале зададим функцию K нормального распределения коэффициента k с помощью шаблонного класса *std::normal_distribution* из стандартной библиотеки шаблонов C++ (STL) [15]:

$$K(x) = \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}, \tag{5}$$

где μ_k - математическое ожидание (среднее значение) коэффициента k, а σ_k - его среднеквадратическое отклонение (задаются один раз перед началом моделирования). Для вычисления аргумента x функции K создадим стандартный генератор псевдослучайных чисел (далее генератор G), основанный на алгоритме «Mersenne Twister», и инициализируем его значением системного генератора случайных чисел (см. шаблонные классы *std::mersenne_twister_engine* и *std::random_device* в [15]). В результате получим следующее выражение коэффициента k для (i, j)-го экземпляра дерева-образца

$$k = K(g_p), \tag{6}$$

где g_p - псевдослучайное значение, сформированное генератором G после его p-го вызова. Отметим, что в выражении (6) функция K нормального распределения (как и генератор G) используется одна и та же для всех экземпляров дерева-образца.

Остальные параметры трансформации (φ , q_r , q_g , q_b) экземпляра дерева-образца задаются аналогично (5) и (6), за исключением координат вектора d . Для координат d_x и d_z также задаются функции D_x и D_z нормального распределения вида (5), но с двумя отличиями. Во-первых, математические ожидания μ_{dx} и μ_{dz} задаются не напрямую, а изменяются в зависимости от номеров i , j экземпляра дерева-образца:

$$\mu_{dx} = jh_x \,,\ \mu_{dz} = ih_z \,,$$

где h_x и h_z - значения фиксированных шагов расположения экземпляров дерева-образца вдоль осей X и Z системы WCS, которые задаются перед началом моделирования (см. базовую постановку задачи в разделе II). Такой подход позволяет избежать при моделировании эффекта скученности и напирания деревьев друг на друга. Второе отличие следует из первого и состоит в том, что функции D_x и D_z задаются индивидуально для каждого (i, j) -го экземпляра дерева-образца, а не для всех экземпляров сразу.

Используя соотношения (3) - (6), запишем искомые выражения модельной матрицы M_{mod} и корректора q цвета, задающие трансформацию (i, j) -го экземпляра дерева-образца:

$$M_{mod} = \begin{pmatrix} K(g_p)\cos\Phi\left(g_{p+1}\right) & 0 & K(g_p)\sin\Phi\left(g_{p+1}\right) & D_x(g_{p+2}) \\ 0 & K(g_p) & 0 & 0 \\ -K(g_p)\sin\Phi\left(g_{p+1}\right) & 0 & K(g_p)\cos\Phi\left(g_{p+1}\right) & D_z(g_{p+3}) \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad q = (Q_r(g_{p+4}), Q_g(g_{p+5}), Q_b(g_{p+6})) \,, \quad (7)$$

где Φ , Q_r , Q_g , Q_b - функции нормального распределения величин φ , q_r , q_g , q_b , аналогичные (5).

9. Метод трансформации на GPU экземпляров дерева-образца

Как упоминалось выше, в данной работе моделирование лесного массива осуществляется на GPU с помощью аппаратно-ускоренной трассировки лучей. Ядром данной технологии является специальный программируемый конвейер (далее *RT-конвейер*), в котором процесс трассировки лучей разбит на ряд типовых стадий (генерация луча, обход ускоряющих структур данных, пересечение луча с процедурным объектом и др.) [12]. В предлагаемом методе трансформация экземпляров дерева-образца выполняется на RT-конвейере в два этапа. На **первом этапе** выполняется построение эффективного описания дерева-образца и его экземпляров в формате, «понятном» RT-конвейеру. На **втором этапе** на основе построенного описания на RT-конвейере реализуется непосредственно трансформация геометрии и цвета экземпляров дерева-образца. Рассмотрим эти этапы.

a. Построение описания дерева-образца и его экземпляров

Чтобы получить доступ к функционалу RT-конвейера, геометрию дерева-образца и их экземпляров необходимо представить в видеопамяти в виде иерархии ограничивающих объемов (Bounding Volume Hierarchy, BVH). Данная иерархия является древовидной и формируется на основе ускоряющих структур нижнего уровня (Bottom Level Acceleration Structure, BLAS) и верхнего уровня (Top Level Acceleration Structure, TLAS) [12]. В общем случае, в BLAS-структуру помещаются данные о геометрии (координаты вершин) объекта-образца, а в TLAS-структуру - ссылка на BLAS-структуру объекта-образца и массив модельных матриц его экземпляров.

В рассматриваемой задаче геометрия дерева-образца задана неявно (каждая точка облака моделируется процедурной сферой), ввиду чего мы помещаем в BLAS-структуру не сами точки облака, а параллелепипеды (Axis-Aligned Bounding Boxes, AABBs), ограничивающие процедурные модели этих точек (см. рис. 2). Для построения эффективного описания дерева-образца с помощью AABBs в данной работе применяется предложенный ранее *метод октантных групп точек* (ОГТ) [15]. Его суть состоит в упорядочивании точек облака по группам (одна группа - один AABB), соответствующим листьям разреженного воксельного октодерева, которое хорошо вписывается в идеологию BVH. Число точек в каждой группе не превышает некоторое пороговое значение K_{max}, которое устанавливается перед началом моделирования, исходя из объема исходных данных и аппаратных возможностей видеокарты. По сравнению с типовым подходом (одна сфера - один AABB) [9], метод ОГТ использует для описания облака точек в несколько раз меньше AABBs, которые при этом практически не пересекаются друг с другом. Это позволяет, с одной стороны, разгрузить «бутылочное горлышко» RT-конвейера (аппаратный блок обхода BVH-иерархии), что дает значительный прирост производительности, а с другой стороны - снизить накладные расходы видеопамяти на построение ускоряющих структур.

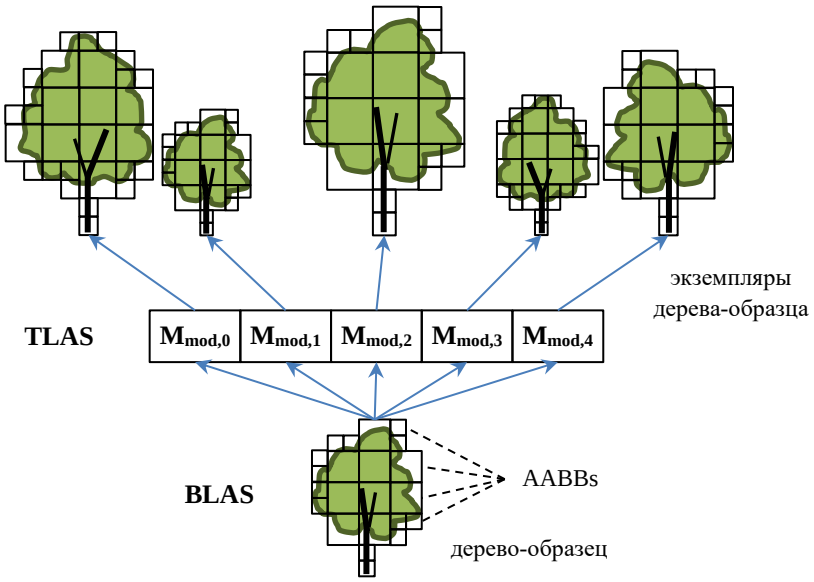


Рис. 2. Взаимосвязь BLAS- и TLAS-структуры в рассматриваемой задаче

В результате выполнения этапа построения описания дерева-образца и его экземпляров в видеопамяти формируется следующий набор данных:

- массив A_g позиций и массив C_g rgb-цветов точек дерева-образца, упорядоченных по группам (каждая группа включает от 1 до K_{max} точек);
- массив H_g дескрипторов групп точек дерева-образца (каждый дескриптор содержит индекс начала и длину соответствующей ему группы точек в массиве A_g);
- массив B AABBs, описывающих дерево-образец (каждый AABB ограничивает соответствующую ему группу точек из массива H_g и задается парой диагональных вершин);
- массив M_{inst} модельных матриц и массив Q_{inst} корректоров цветов экземпляров дерева-образца, вычисленные согласно (7).

Отметим, что через связку BLAS-TLAS в видеопамять передаются только массивы B и M_{inst} (согласно общей идеологии построения BVH [12]). Остальные массивы мы передаем отдельно с помощью механизма буферов хранения (Storage Buffers) [17].

b. Трансформация экземпляров дерева-образца на RT-конвейере

Согласно концепции аппаратно-ускоренной трассировки лучей [12], трансформация экземпляров дерева-образца на RT-конвейере включает в себя аппаратную и программную части. Аппаратная часть отвечает за поиск AABBs экземпляров дерева-образца, которые лежат на пути трассируемого луча[3], а программная часть - за трансформацию геометрии и цвета процедурных сфер, находящихся внутри найденных AABBs. Рассмотрим эти составляющие.

Аппаратная часть. Поиск AABBs, лежащих на пути луча, RT-конвейер выполняет на основе BVH-иерархии виртуальной сцены. Данная иерархия создается автоматически видеодрайвером до запуска трассировки лучей, исходя из заполненных BLAS- и TLAS-структуры. По сути, драйвер создает дерево ограничивающих объемов сцены, к листьям которого привязывает AABBs экземпляров дерева-образца (построение и обработка BVH-иерархии является закрытой частью технологии и зависит от конкретной ее реализации). В процессе трассировки луча специальный аппаратный блок RT-конвейера выполняет обход BVH-иерархии, при котором:

- ищет листья BVH-иерархии, которые пересекаются лучом;
- при обнаружении такого листа выполняет обработку ассоциированных с ним AABBs (далее *AABB-кандидатов*);
- при обработке AABB-кандидата выполняет трансформацию его вершин с помощью матрицы из массива M_{inst} и определяет пересечение луча с трансформированным AABB-кандидатом.

[3] В данной работе рассматриваются *первичные лучи* - лучи, испущенные из положения виртуального наблюдателя (по лучу на пиксел области вывода) в пространство сцены до первого пересечения с каким-либо объектом.

Как можно видеть, в масштабе задачи моделирования лесного массива на аппаратный блок RT-конвейера возлагается большой объем работы по отбору AABB-кандидатов, пересекаемых лучом, ввиду чего важно иметь эффективное описание дерева-образца с помощью AABBs.

Программная часть включает в себя трансформацию геометрии и цвета процедурных сфер внутри AABB-кандидатов, отобранных аппаратным блоком. Трансформация является поточечной и реализуется на двух программируемых стадиях RT-конвейера: в шейдере потенциального пересечения (Intersection Shader, I-шейдер) и шейдере ближайшего подтвержденного пересечения (Closest Hit Shader, CH-шейдер) [12].

На ***стадии I-шейдера*** выполняется вычисление ближайшей к наблюдателю точки пересечения луча с трансформированной процедурной сферой AABB-кандидата (см. рис. 3).

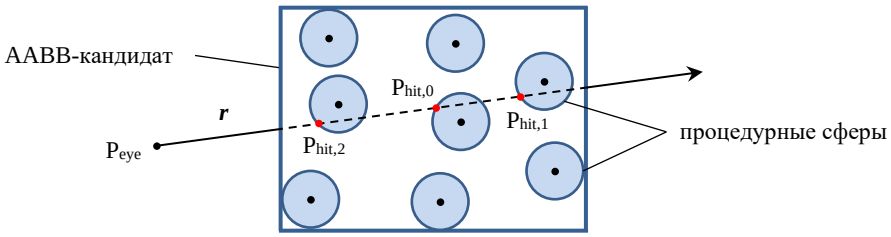


Рис. 3. Пересечение луча с процедурными сферами внутри AABB-кандидата

Чтобы найти такую точку, мы проверяем пересечение луча r со всеми трансформированными сферами внутри AABB-кандидата и из найденных точек (на рис. 3 это точки $P_{hit,0}$, $P_{hit,1}$ и $P_{hit,2}$) выбираем ближайшую к наблюдателю. Это реализует следующий алгоритм, адаптированный под облака точек, описанные с помощью октантных групп (см. раздел 2.1):

1. Получим данные из аппаратного блока RT-конвейера: координаты начала P_{eye} и вектора r направления луча в системе WCS; модельную матрицу M_{mod} экземпляра дерева-образца, которому принадлежит AABB-кандидат; индекс i_{AABB} обрабатываемого AABB-кандидата в массиве B AABBs, описывающих дерево-образец.
2. По индексу i_{AABB} получим данные из массива H_g дескрипторов групп точек: индекс i_{group} начала и длину n_{group} группы точек, соответствующей AABB-кандидату.
3. Вычислим радиус R_{tf} трансформированных процедурных сфер: $R_{tf} = kR$, где $k = M_{mod}[1][1]$ - коэффициент масштабирования экземпляра дерева-образца, а R - заданный исходно радиус сферы.
4. Инициализируем данные о ближайшей точке пересечения: индекс $p_{chit} = 0$ точки в массиве A_g (позиций точек дерева-образца) и параметр $t_{chit} = t_{big}$, где $t_{big} = 10^7$.
5. Цикл по индексу p в массиве A_g от i_{group} до $(i_{group} + n_{group} - 1)$
   Получим координаты $P_{sph,ocs}$ центра сферы (точки облака) в системе OCS: $P_{sph,ocs} = A_g[p]$.
   Вычислим координаты $P_{sph,wcs}$ центра сферы в системе WCS: $P_{sph,wcs} = M_{mod}\, P_{sph,ocs}$.
   Найдем параметр t_{hit} точки пересечения луча $\{P_{eye}, r\}$ со сферой $\{P_{sph,wcs}, R_{tf}\}$ с помощью алгоритма [18].
   Если $-0.5 < t_{hit} < t_{chit}$, то: $t_{chit} = t_{hit}$, $p_{chit} = p$.
   Конец цикла.
6. Если $t_{chit} < t_{big}$, то сообщим RT-конвейеру о наличии подтвержденного пересечения со сферой с помощью встроенного оператора *reportIntersectionEXT*.

Конец алгоритма.

Отметим, что, применяя описанный алгоритм к каждому следующему отобранному AABB-кандидату, RT-конвейер автоматически обновляет наименьшее значение параметра t_{chit} и связанного с ним индекса p_{chit}. После обработки всех AABB-кандидатов обновленная пара $\{t_{chit}, p_{chit}\}$ передается на стадию CH-шейдера.

На ***стадии CH-шейдера*** выполняется вычисление скорректированного цвета C_{cor} ближайшей к наблюдателю точки $\{t_{chit}, p_{chit}\}$ согласно выражению (2):

$$C_{cor} = Q_{inst}[u]C_g[p_{chit}],$$

где u - индекс обрабатываемого экземпляра дерева-образца, полученный из аппаратного блока RT-конвейера, Q_{inst} - массив корректоров цветов экземпляров дерева-образца, а C_g - массив rgb-цветов точек дерева-образца (см. раздел 2.1).

Таким образом, трансформация экземпляров дерева-образца выполняется путем описанной обработки на RT-конвейере каждого луча, испущенного из положения наблюдателя (по лучу на пиксел области вывода). После обработки всех лучей в области вывода формируется изображение смоделированного лесного массива.

IV. Результаты экспериментов

Для проведения экспериментальной оценки предложенные методы и подход были реализованы в созданном ранее программном комплексе визуализации облаков точек [16]. Архитектура комплекса включает в себя CPU-часть (оболочка, загрузка и предварительная обработка исходных данных) и GPU-часть (аппаратно-ускоренная трассировка лучей). CPU-часть разработана на языке C++ с использованием библиотеки STL, а GPU-часть реализована на языке GLSL программирования шейдеров с помощью API Vulkan (версия 1.3.290). В рамках настоящей работы была проведена модификация комплекса: в CPU-часть добавлена возможность тиражирования облаков точек с псевдослучайной трансформацией, а в GPU-часть - возможность визуализации тиражированных экземпляров облаков точек. Также для получения разнообразных вариантов тестовых сцен в комплекс была добавлена возможность задания параметров псевдослучайной трансформации через конфигурационный файл.

С помощью модифицированного комплекса был проведен ряд экспериментов по моделированию лесных массивов для различных времен года. В качестве деревьев-образцов были использованы цифровые модели листопадного дерева «Beautiful Autumn» [10] (около 1.6 млн. точек) и вечнозеленого дерева «Coastal redwood» [11] (около 5 млн. точек)[4]. На основе данных деревьев-образцов были разработаны «осенняя» и «вечнозеленая» конфигурации виртуальных лесных массивов (см. Таблицу 1).

ТАБЛИЦА 1
КОНФИГУРАЦИИ ВИРТУАЛЬНЫХ ЛЕСНЫХ МАССИВОВ

Тип	**Параметры псевдослучайной трансформации**						
	$\{h_x, \sigma_{dx}\}$	$\{h_z, \sigma_{dz}\}$	$\{\mu_\varphi, \sigma_\varphi\}$	$\{\mu_k, \sigma_k\}$	$\{\mu_r, \sigma_r\}$	$\{\mu_g, \sigma_g\}$	$\{\mu_b, \sigma_b\}$
«Осенняя»	(a) {0.7, 0.15}; (b) {1.4, 0.3}; (c) {2.8, 0.6}; (d) {5.6, 1.2}; (e) {11.2, 2.4}		{0, 90}	{1, 0.2}	{0.6, 0.15}	{0.6, 0.1}	{0.05, 0.01}
«Вечнозеленая»				{1, 0.1}	{0.9, 0.03}	{0.9, 0.03}	{0.9, 0.03}

В Таблице 1 значения параметров расстановки экземпляров дерева-образца (2-ой столбец) заданы в условных единицах (одна единица соответствует исходной высоте дерева-образца[5]), значения параметров вращения (3-й столбец) заданы в градусах, а значения параметров масштаба и цвета (4-7 столбцы) являются безразмерными.

Каждая из приведенных в Таблице 1 конфигураций была протестирована на 5-ти вариантах расстановок (см. пункты (a)-(e) в столбце 2), которые соответствуют расположению 1428x1428, 714x714, 357x357, 178x178 и 89x89 экземпляров дерева-образца на участке размером около 1000x1000 условных единиц. Результаты тестирования отражены на рисунке 4 в виде графиков зависимости среднего времени синтеза кадра (в миллисекундах) от высоты положения виртуального наблюдателя (в условных единицах).

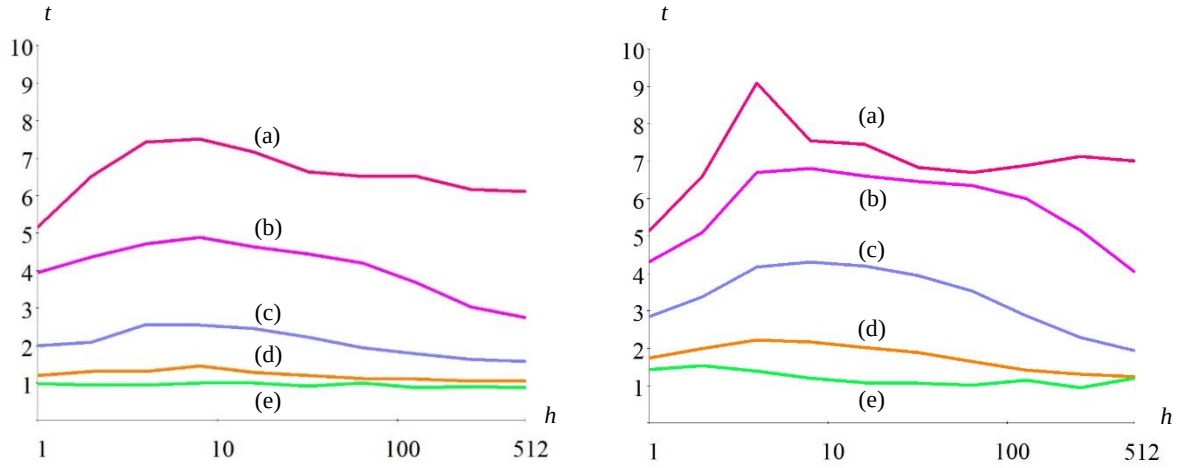


Рис. 4. Влияние высоты h положения наблюдателя на среднее время t синтеза кадра при вариантах (a)-(e) расстановок деревьев из Таблицы 1. Слева приведены графики для лесных массивов, построенных на основе листопадного дерева-образца [10], справа - на основе вечнозеленого [11]. Шкала оси h - логарифмическая.

[4] Данные цифровые модели распространяются по лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0), http://creativecommons.org/licenses/by/4.0/.
[5] При загрузке цифровой модели дерева-образца мы нормализуем координаты его точек и удваиваем исходный масштаб.

На рисунках 5 и 6 показаны примеры кадров визуализации лесных массивов «от лица пешехода» и с высоты «птичьего полета». В левом верхнем углу каждого изображения приведены значения средней скорости визуализации (в кадрах в секунду) и времени синтеза кадра.

Тестирование конфигураций лесных массивов проводилось на персональном компьютере (Intel Core i7-6800K 6x3.40 GHz, 16 GB RAM), оборудованном видеокартой NVidia RTX 2080 (драйвер DCH 536.40), при разрешении области вывода 1920x1080. Для построения AABB-описаний деревьев-образцов использовался пороговый размер K_{max} = 8 точек и радиус сферы $R = 8 \cdot 10^{-4}$ условных единиц.

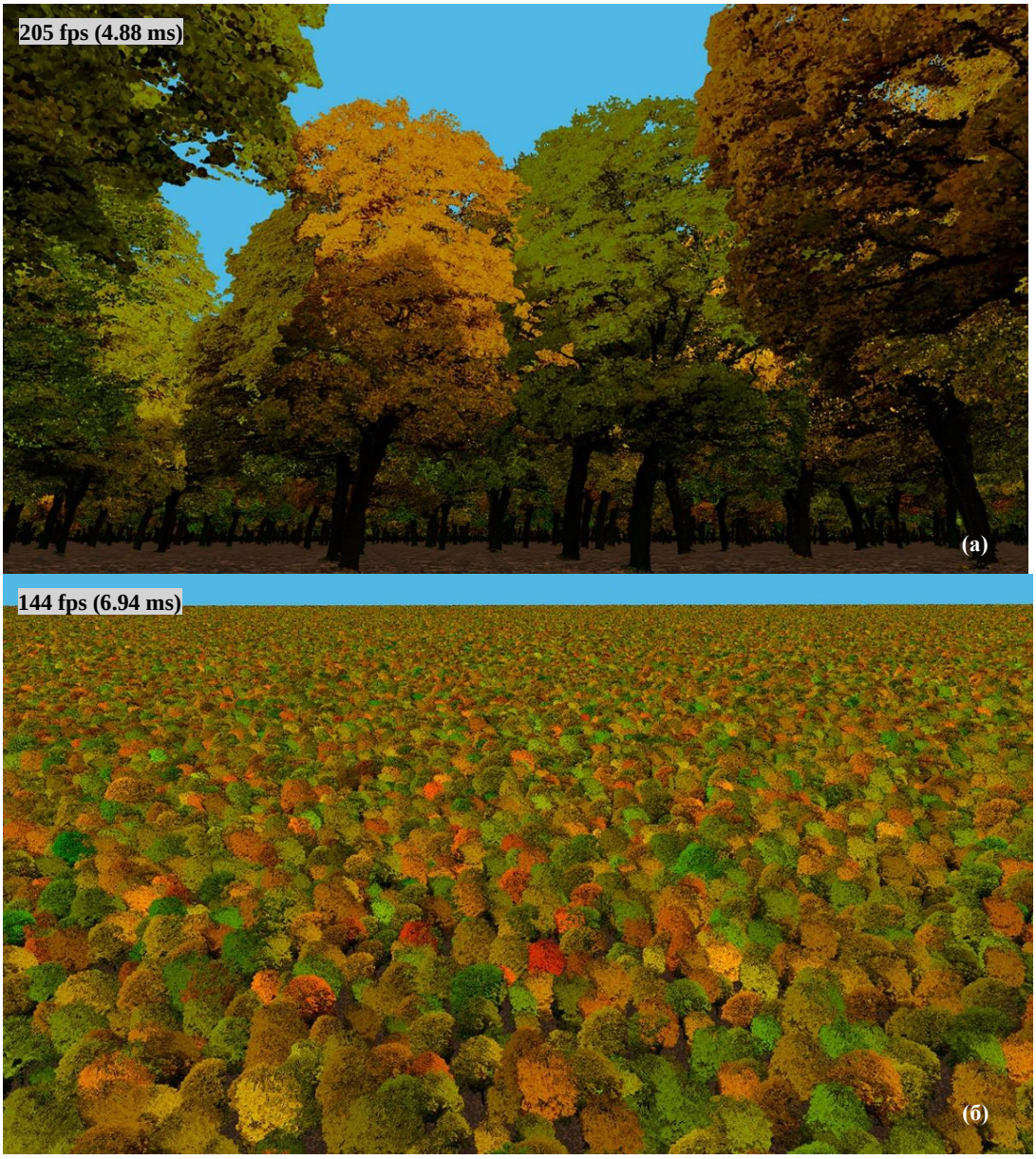


Рис. 5. Кадры визуализации «осеннего» лесного массива с помощью предложенного решения: (а) вид «от лица пешехода»; (б) вид с высоты «птичьего полета» (размеры массива - 1428x1428 деревьев)

V. Обсуждение результатов

Из рисунка 4 видно, что предложенное решение обеспечивает визуализацию всех исследуемых вариантов лесных массивов в реальном времени с запасом производительности 75-80% (в системах виртуального окружения время синтеза кадра не должно превышать 40 миллисекунд при визуализации в моно режиме). Также приведенные количественные результаты демонстрируют хорошую масштабируемость предложенного решения: при превышении высоты наблюдения более 10 условных единиц (когда в поле зрения попадают сотни тысяч деревьев), время синтеза кадра перестает расти и приобретает тенденцию к снижению. Важно отметить, что результаты получены на видеокарте с архитектурой аппаратно-ускоренной трассировки лучей первого поколения

(Turing), которая содержит всего 46 ядер трассировки лучей (RT-ядер). Согласно спецификации [19] в архитектуре Ada последнего (третьего) поколения значительно повышены количество и эффективность RT-ядер (в частности, увеличена в 10 раз скорость построения BVH-иерархии и в 2 раза скорость ее обхода с учетом полупрозрачности). Это дает основания полагать, что предложенное решение имеет намного больший потенциал, который может быть раскрыт с помощью новых графических карт.

Наши исследования качества визуализации показали, что увеличение разброса псевдослучайных параметров коррекции цвета, с одной стороны, повышает эффект уникальности лесного массива (см. рисунок 5), а с другой стороны, может приводить к образованию излишне «пестрых» цветов. По этой причине, например, достаточно сложно получить из «осеннего» дерева-образца убедительный виртуальный лесной массив, соответствующий летнему периоду. В этой связи предложенный метод коррекции цветов будет давать лучшие результаты, когда необходим небольшой разброс цветов (см. рисунок 6). Преодоление описанных ограничений, а также построение лесных массивов смешанного типа, представляется направлением для последующей работы.

Рис. 6. Кадры визуализации «вечнозеленого» лесного массива с помощью предложенного решения: (а) вид «от лица пешехода»; (б) вид с высоты «птичьего полета» (размеры массива - 1428x1428 деревьев)

VI. Выводы и заключение

В данной работе на примере задачи моделирования лесного массива была продемонстрирована возможность создания уникальных виртуальных объектов из квазиповторяющихся облаков точек. Уникальность лесного

массива обеспечивается с помощью разработанного метода псевдослучайной трансформации, который позволяет одновременно варьировать шаг расстановки, ориентацию, масштаб и цвет деревьев. Тиражирование деревьев выполняется с помощью подхода, основанного на аппаратно-ускоренной трассировке лучей, который был модифицирован для псевдослучайной трансформации облаков точек. Были предложены метод и алгоритм выполнения такой трансформации на конвейере трассировки лучей, адаптированные под эффективное описание облака точек с помощью ограничивающих параллелепипедов. Программная реализация предложенных методов и подхода позволила провести оценку качественных и количественных показателей на примере моделирования «осеннего» и «вечнозеленого» лесных массивов. Серия экспериментов показала, что предложенное решение позволяет генерировать в реальном времени уникальные виртуальные лесные массивы «от лица пешехода» и с высоты «птичьего полета» с существенным (75-80%) запасом по производительности. Также предложенное решение продемонстрировало хорошую масштабируемость при работе с виртуальной сценой, содержащей около 2-х миллионов деревьев. Полученные результаты подтверждают возможность использования предложенного решения в системах виртуального окружения и многих других приложениях компьютерной графики реального времени (видеосимуляторах, научной визуализации, геоинформационных приложениях и др.). В качестве дальнейшей работы планируется моделирование лесного массива смешанного типа на виртуальном рельефе местности.

Источник финансирования. Благодарности

Публикация выполнена в рамках государственного задания ФГУ ФНЦ НИИСИ РАН по теме № FNEF-2024-0002 «Математическое моделирование многомасштабных динамических процессов и системы виртуального окружения».

Список литературы

1. Thuvander L., Somanath S., Hollberg A. Procedural Digital Twin Generation for Co-Creating in VR Focusing on Vegetation // The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2022. Vol. XLVIII-4/W5-2022. P. 189-196. DOI: 10.5194/isprs-archives-XLVIII-4-W5-2022-189-2022.

2. You J., Huai Y., Nie X., Chen Y. Real-Time 3d Visualization of Forest Fire Spread Based on Tree Morphology and Finite State Machine // Computers & Graphics. 2022. Vol. 103. P. 109-120. DOI: 10.1016/j.cag.2022.01.009.

3. «Realistic High Poly Tree», a 3D Model by Chenuka Wijesundara // Sketchfab. https://skfb.ly/ooQUt.

4. Bao G., Li H., Zhang X., Dong W. Large-scale forest rendering: Real-time, realistic, and progressive // Computers & Graphics. 2012. Vol. 36. No. 3. P. 140-151. DOI: 10.1016/j.cag.2012.01.005.

5. Strnad D., Kohek Š., Nerat A., Žalik B. Efficient Representation of Geometric Tree Models with Level-of-Detail Using Compressed 3D Chain Code // IEEE Transactions on Visualization and Computer Graphics. 2020. Vol. 26. No. 11. P. 3177-3188. DOI: 10.1109/TVCG.2019.2924430.

6. Zellmann S., Morrical N., Wald I., Pascucci V. Finding Efficient Spatial Distributions for Massively Instanced 3-d Models // Eurographics Symposium on Parallel Graphics and Visualization. 2020. P. 1-11. DOI: 10.2312/pgv.20201070.

7. Weiser H., Schäfer J., Winiwarter L., Krašovec N., Fassnacht F. E., Höfle B. Individual tree point clouds and tree measurements from multi-platform laser scanning in German forests // Earth System Science Data. 2022. Vol. 14. No. 7. P. 2989–3012. DOI: 10.5194/essd-14-2989-2022.

8. Ткачева А. А, Фаворская М. Н. Моделирование трехмерных сцен лесных участков по данным лазерного сканирования и аэрофотоснимкам // Информационно-управляющие системы. 2015. № 6. С. 40-49. DOI: 10.15217/issn1684-8853.2015.6.40.

9. Lefrançois M.-K. Ray tracing intersection // NVIDIA DesignWorks. Vulkan Ray Tracing Tutorials. 2020-2023. https://github.com/nvpro-samples/vk_raytracing_tutorial_KHR/tree/master/ray_tracing_intersection.

10. «Beautiful Autumn» by Epic_Tree_Store // Sketchfab. https://skfb.ly/VrYw.

11. «Coastal redwood from Grove of Old Trees» by kungphil // Sketchfab. https://skfb.ly/oxL6p.

12. Rusch M., Bickford N., Subtil N. Introduction to vulkan ray tracing // Ray Tracing Gems II. NVIDIA. 2021. P. 213–255. DOI: https://doi.org/10.1007/978-1-4842-7185-8_16.

13. Lefrançois M.-K. Ray tracing instances // NVIDIA DesignWorks. Vulkan Ray Tracing Tutorials. 2020-2024. https://github.com/nvpro-samples/vk_raytracing_tutorial_KHR/tree/master/ray_tracing_instances.

14. Lengyel E. Mathematics for 3D Game Programming and Computer Graphics (Third Edition). Boston, MA: Course Technology PTR, 2012. 624 p.

15. Pseudo-random number generation: std::mersenne_twister_engine, std::random_device, std::normal_distribution. C++ reference. Numerics library. 2024. https://en.cppreference.com/w/cpp/numeric/random.

16. Timokhin P. Y., Mikhaylyuk M. V. A Method to Order Point Clouds for Visualization on the Ray Tracing Pipeline // Programming and Computer Software. 2024. Vol. 50. No 3. P. 264–272. DOI: 10.1134/S0361768824700075.

17. Resource Creation. Buffers // Vulkan 1.3.290 - A Specification (with all ratified extensions). The Khronos Vulkan Working Group. 2024. https://registry.khronos.org/vulkan/specs/1.3-khr-extensions/pdf/vkspec.pdf.

18. Raytracing - Ray Sphere Intersection. 2018. https://viclw17.github.io/2018/07/16/raytracing-ray-sphere-intersection.

19. NVIDIA Ada GPU Architecture (v.2.02). 2022. https://images.nvidia.com/aem-dam/Solutions/geforce/ada/nvidia-ada-gpu-architecture.pdf.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-147-163

Анализ производительности методов обхода двухуровневых BVH деревьев в трассировке лучей на графических процессорах
Performance analysis of traversal two-level BVH trees for GPU ray tracing

Л. М. Смирнов[1], В. А. Фролов[1,2], А.Г. Волобой[2],
[1] МГУ им. М.В. Ломоносова, факультет ВМК, ГСП-1, Ленинские горы, Москва, 119991, Россия
[2] ИПМ им. М. В. Келдыша РАН, Миусская пл. 4, Москва, 125047, Россия

L.M. Smirnov[1], V.A. Frolov[1,2], A.G. Voloboy [2],
[1]Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia
[2] Keldysh Institute of Applied Mathematics RAS, Miusskaya pl. 4, Moscow, 125047, Russia

Аннотация. В работе представлен сравнительный анализ производительности нескольких методов обхода BVH деревьев на стационарных (десктопных) и мобильных графических процессорах. Мы исследовали BVH деревья с различной глубиной и количеством дочерних узлов, реализовали несколько алгоритмов обхода со стэком и два различных алгоритма без-стэкового обхода; предложили свой вариант без-стэкового обхода, более производительный чем существующие в ряде случаев. Предложили свой вариант сжатия BVH дерева с 2 узлами, теряющий не более 15% производительности. Мы выявили некоторую общую проблему, встречающуюся почти во всех алгоритмах при их реализации на графических процессорах. Мы полагаем, что наш анализ поможет разработчикам аппаратных ускорителей трассировки лучей создавать более экономное аппаратное решение, не ограничиваемое при этом только лишь трассировкой лучей. Если говорить более конкретно, результаты наших экспериментов говоря о том, что можно получить ускорение до 5 раз при помощи изменения механизма работы L2 кэша, причем на десктопных GPU с аппаратным ускорением трассировки лучей это, по-видимому, уже сделано не только в рамках непосредственно механизма аппаратного ускорения трассировки лучей, но и в более общем случае.

***Ключевые слова*:** трассировка лучей, графические процессоры, BVH деревья

Аннотация. This paper presents a comparative analysis of the performance of several BVH traversal methods on PC(desktop) and mobile GPUs. We examined BVH with different depths and the number of child nodes, implemented several stack traversal algorithms and two different stackless traversal algorithms. We proposed our own stackless traversal option, which is more productive than existing ones in some cases, and our own version of BVH compression of a tree with 2 nodes, which loses no more than 15% of performance. As a result of the performance analysis, we have identified some common problems that occur in almost all algorithms, when they are implemented on GPUs. We believe that our analysis will help developers of ray tracing hardware accelerators to create a more economical hardware solution that is not limited to ray tracing alone. More specifically, the results of our experiments indicate that it is possible to get acceleration up to 5 times by changing the mechanism of the L2 cache, and on desktop GPUs with hardware acceleration of ray tracing, this seems to have already been done not only within the framework of the hardware acceleration of ray tracing itself, but also more generally the case.

***Ключевые слова*:** ray tracing, GPU, Mobile graphic application

I. Введение

Одна из целей нашей работы заключалась в том, чтобы определить основные узкие места в алгоритмах обхода BVH деревьях на практике, и предложить в программное или аппаратное решение (в будущем) для устранения этих узких мест. Несмотря на широко доступные в мире графические процессоры с аппаратным ускорением трассировки лучей, в силу закрытости этих решений на сегодняшний день до конца не понятно что же именно нужно ускорять аппаратно. Например, специалисты компании Imagination выделяют 6 уровней аппаратного ускорения [IMRT]:

1. Исторические решения
2. Программная реализация на GPU
3. Ускорение пересечения с некоторыми примитивами (например, прямоугольным параллелепипедом)
4. Аппаратный обход BVH дерева, реализованный на отдельных блоках
5. Когерентная обработка BVH дерева, сочетающая в себе сортировку лучей или узлов BVH дерева для более эффективного доступа в память

6. Предыдущий уровень + аппаратное ускорение построения дерева

Забегая немного вперёд, наши эксперименты показали, что имеет смысл рассматривать только 5-ый уровень и далее, а остальные варианты не дадут почти никакого ускорения даже на достаточно слабых мобильных графических процессорах, которые обделены вычислительной мощностью по сравнению со стационарными решениями. Следующий не очевидный вопрос: какой алгоритм обхода и какое именно BVH дерево выбрать, как его хранить в памяти. Некоторые виды деревьев в сочетании с определённым способом хранения могут быть менее эффективны с алгоритмической точки зрения, но более эффективны с точки зрения работы подсистемы памяти.

II. Анализ существующих подходов

Поскольку одна из наших целей заключалась в проверке существующих алгоритмов на практике, мы старались реализовывать наиболее перспективные варианты структур данных, алгоритмов обхода и измерять их производительность. С другой стороны, в некоторых случаях мы полагались на данные литературы, поскольку алгоритмы либо заведомо не подходили для наших целей, либо выигрыш от них, известный в литературе, был недостаточно большим. За отправную точку нашего исследования мы берём две работы: [SurveyBVH] и [MB22]. Работа [SurveyBVH] содержит всесторонний обзор алгоритмов трассировки лучей для различных видов BVH деревьев и способов их хранения в памяти, реализаций трассировки лучей на различных типах устройств. Работа [MB22] (hippie) содержит сравнение различных алгоритмов построения BVH на графических процессорах, а также современную реализацию трассировки лучей на CUDA/HIP. Эта статья основана на программном окружении, разработанным Timo Aila в работе [AL09], которая изначально была реализована на CUDA, и в основном была сосредоточена на изучении эффективности самого обхода. Особенность [MB22] в том, что эта работа исследует связь алгоритма построения и характеристик дерева с эффективностью алгоритмом обхода на GPU. Кроме того, исходные коды данной работы находятся в открытом доступе, что позволяет нам сравнить разработанное программное решение с hippie.

1. Основные направления

Анализируя [SurveyBVH], можно выделить несколько направлений, которые могут существенно влиять на производительность.

(1) Качественное дерево. Обычно это достигается с помощью алгоритма построения дерева, выбирающего способ разбиения на основе минимизации некоторой функции стоимости, такой как Surface Area Heuristic (SAH) [AKL13]. Однако на практике этого недостаточно для надежной работы с треугольными мешами, поскольку многие треугольники плохо аппроксимируются прямоугольными параллелепипедами. Эта проблема приводит подходам под-разбиения треугольников: (1.1) предварительное разбиение треугольников [EG07] — быстрый и надежный способ с точки зрения потребления памяти; (1.2) пространственное разбиение (spatial split) для BVH [SFD09; PGDS09], которое обеспечивает наилучшее качество по данным [SurveyBVH; SFD09; PGDS09] и далее упоминается как SBVH в работе [MB22].

Основываясь на анализе [SurveyBVH] и [MB22] мы выбрали вариант алгоритма SBVH, реализованный в библиотеке Embree [embree] 3-ей версии в качестве метода построения дерева на основе доступного в Embree примера построения BVH2 дерева. BVH4 дерево мы получали из BVH2 путём удаления внутренних узлов. Мы сравнили долгий, но качественный билдер SBVH против быстрого и часто-используемого решения LBVH на 2 известных сценах – Crytek Sponza и Bistro. Для данного сравнения вся геометрия сцены была объединена в 1 меш. Сравнение проводилось по количеству арифметических операций на 1 луч в среднем, которое мы оценивали из числа пересеченных лучом ограничивающих боксов и треугольников (рис. 1) по формуле AC = 25*NC + 60*TC, где NC – число пересеченных узлов, а TC – число пересеченных треугольников.

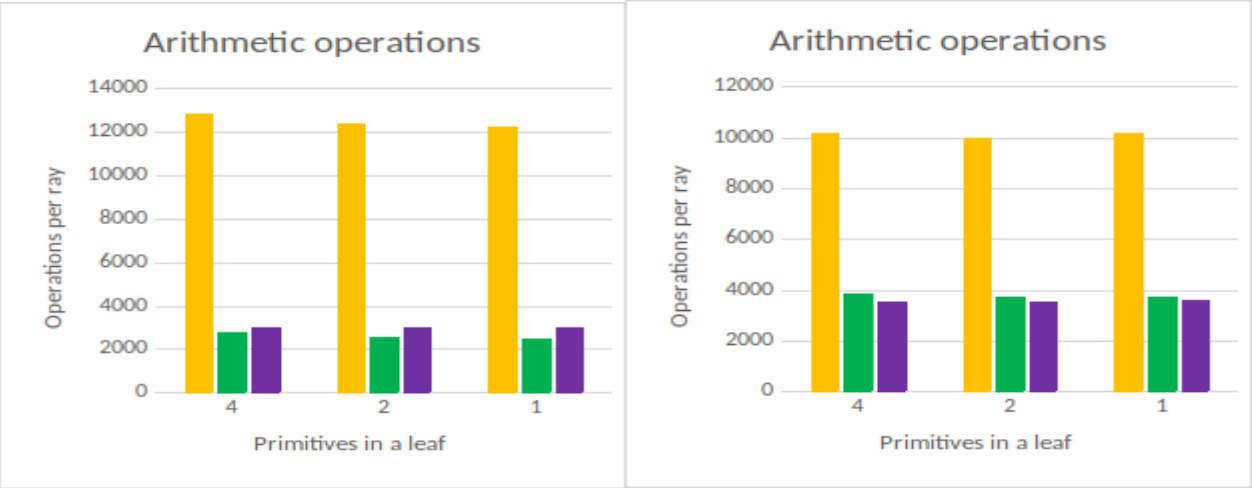


Рис. 1 Сравнение количества арифметических операций на луч для сцены Cry Sponza. Желтый (первый столбец) — LBVH билдер, реализованный в библиотеке Embree, зеленый и фиолетовый (второй и третий) — варианты SBVH. Зелёный – наш собственный построитель, фиолетовый – реализация из библиотеки embree. Левое изображение – данные для сцены Cry Sponza. Правок – для сцены bistro.

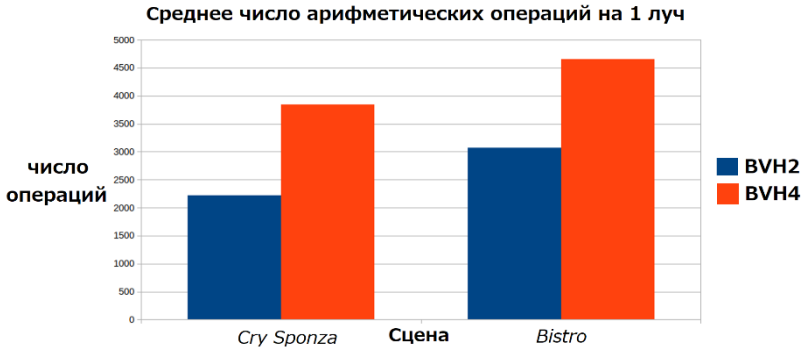
Таким образом, на сложных геометрических сценах качественно построенное дерево – один из необходимых шагов, дающих ускорение до 3 раз. При этом в наших экспериментах оптимальным числом треугольников в листьях можно считать число в интервале от 2 до 4. Более глубокое или мелкое подразбиение ухудшало производительность (рис. 1).

(2) Широкие BVH деревья (BVH4 или BVH8) [YKL17]. Широкие деревья в целом более компактны, поскольку содержат меньше внутренних узлов, чем двоичные деревья; кроме того, количество шагов в них обычно меньше, а соседние узлы, как правило, стараются упаковать так чтобы все N узлов (4 или 8) попадали в одну кэш-линию. Основываясь на анализе работы [YKL17], мы решили остановиться на BVH с 4 узлами (BVH4) по нескольким причинам:

- В идеале мы бы хотели найти комбинацию алгоритмов и структур данных, которая бы подходила под разные вычислительные устройства. Мобильные GPU обычно имеет кэш-линию, равную 64 байтам, что идеально подходит для 4 узлов в формате half-float (реализация «BVH4_16»), а центральные процессоры и десктопные GPU имеют кэш линию размером 128 байт, что также хорошо подходит для 4 узлов с 32 битной плавающей точкой (реализация «BVH4_16»). Кроме того BVH4 хорошо ложится на SIMD инструкции для бОльшего числа центральных процессоров чем BVH8.
- Подход BVH8 из [YKL17] использует нестрогий порядок обхода дочерних узлов, который предполагает, что узлы расположены в стиле окто-дерева. Обычно это справедливо для алгоритма построения LBVH, но не для общего случая и может привести к снижению производительности.
- С алгоритмической точки зрения BVH8 в 1,5 раза менее эффективен, чем BVH2 и BVH4, поскольку делает больше пересечений луча с ограничивающими параллелепипедами. Алгоритмическая эффективность BVH2 и BVH4 практически эквивалентна (за исключением пустых узлов в BVH4).
- Значительное улучшение BVH8 в [YKL17] было достигнуто только для некогерентных вторичных лучей, что не является основной целью нашего исследования, поскольку некогерентные лучи – это отдельная сложная проблема, для которой подходят, например метод из [LSS18]. Основной недостаток этого метода в том, что он активно использует взаимодействие между разными потоками и выделяет на 1 луч 8 потоков (Обход широких BVH в стиле SIMD-CPU). Такой алгоритм требует реализации трассировки лучей в виде отдельного вычислительного ядра, что не всегда применимо. Кроме того реализация в работе [LSS18] интенсивно использует разделяемую память GPU, что не всегда хорошо работает на мобильных графических процессорах.

Несмотря на многочисленные преимущества широкие BVH деревья имеют и недостатки, поэтому ими нельзя ограничиваться. Основной недостаток – усложненный алгоритм обхода со значительно большим регистровым давлением. Это вызвано необходимостью сортировки узлов BVH дерева, пересекаемых лучом во время обхода. Кроме того, наличие пустых узлов всё же делает BVH4 деревья менее эффективными с алгоритмической точки зрения, чем BVH2 деревья. Мы численно исследовали этот вопрос, оценив количество арифметических операций на тех же двух сценах для BVH2 и BVH4, и обнаружили неожиданный эффект. Несмотря на очевидные эмпирические соображения и данные литературы, оцененное на реальном эксперименте число выполненных операций пересечения с узлами дерева было значительно больше для BVH4 деревьев, чем для BVH2 (Рис. 2).

Рис. 2. Сравнение среднего числа арифметических операций на луч для сцен геометрически-сложных Cry Sponza



и Bistro. Для сцен с более простой геометрией такой разницы не наблюдается.

Причём, эта картина воспроизводилась только на сложных сценах с неравномерным размером треугольников. На более простых сценах с хорошо протесселированной геометрией (например, dragon) разницы, ожидаемо, почти не было. Мы объясняем это двумя факторами: (1) для рассмотренных нами сцены деревья получаются в высокой степени неравномерными, что приводит к большому числу пустых узлов в BVH4 и, как следствие, росту количества выполняемых арифметических операций; (2) В сценах со сложной геометрией для BVH4 дерева подъём становится в 2 раза более дорогим, чем для BVH2, поскольку при чтении номера узла из стэка алгоритм сразу же тестирует всех его детей. Результаты этого эксперимента натолкнули нас на мысль не останавливаться только лишь на широких BVH деревьях, а исследовать также возможность более повышения эффективности подсистемы памяти для BVH2 деревьев используя трилеты (см. далее подраздел 4).

(3) Обход широких BVH в стиле SIMD-CPU на графическом процессоре таким образом, чтобы каждый поток вычислял пересечение только с одним ограничивающим примитивом [LSS18; MB22]. В этом подходе на каждый луч выделяется более чем 1 поток (4 или 8 в зависимости от ширины дерева), и в процессе трассировки лучей потоки взаимодействуют между собой. Мы не реализовывали этот подход, поскольку он имеет ряд ограничений. Во-первых, его GPU реализация интенсивно использует разделяемую память, что не на всех GPU работает хорошо, особенно на мобильных. Во-вторых, этот подход выделяет на 1 луч 4 или 8 потоков (по ширине дерева), что вынуждает реализовывать трассировку лучей в виде отдельного вычислительного ядра и делает невозможным реализацию рендера в одном убер-ядре. Это не всегда применимо т.к. зачастую нужно иметь функцию трассировки луча, которую можно просто вызвать в шейдере или C++ коде.

(4) Трилеты. Трилетами называют небольшие участки дерева, помещаемые в памяти последовательно в некоторый блок определённого размера. Этот подход использует предположение о том, что прочитать некоторый последовательный блок памяти более эффективно, чем узлы из случайно-разбросанных областей памяти. Его можно разделить на Cache Oblivious BVH (COLBVH) и Cache-Aware BVH (CALBVH):

- **COLBVH.** Расположение узлов BVH в виде трилетов без явного учета параметров кэша [YM06], который пытается оптимизировать доступ к памяти для различных размеров кэша.
- **CALBVH**. Расположение узлов BVH в виде трилетов с явным учетом параметров кэша (CALBVH) [AK10]. В этом подходе узлы BVH дерева раскладываются строго по блокам, равным размеру кэш-линии вычислительного устройства. Обычно это 128 байт (для большинства CPU и десктопных GPU) и 64 байт для мобильных GPU.

Мы реализовали оба подхода и получили незначительный, но стабильный (+9%) на всех сценах прирост в производительности для CALBVH (рис. 3 CALBVH/COLBVH, слева) и почти полное отсутствие влияния типа и размера трилета на производительность в COLBVH (рис. 3 CALBVH/COLBVH, справа).

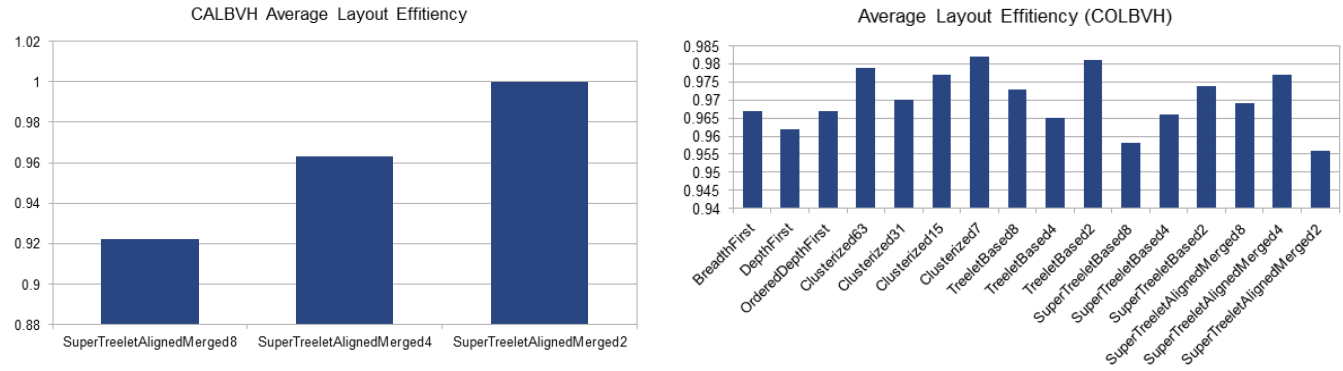


Рис. 3 CALBVH/COLBVH. Средняя скорость трассировки лучей для различных вариантов трилетов в % от самого быстрого в зависимости от различного типа и схемы расположения трилов в памяти. Поскольку измерения сделаны на мобильном GPU, наиболее эффективным оказался подход, в котором трилет целиком помещается в 64 байта (SuperTreeletAlignedMerged2).

Эти результаты вполне объяснимы. Для современной аппаратуры (неважно CPU или GPU) важно сокращение количества чтений кэш-линий, которая в нашем эксперименте для мобильных GPU была равна 64 байтам. Если трилет не помещается целиком в 1 кэш-линию, это ожидаемо не даёт ускорения. Однако на данном этапе нам не было понятно, почему CALBVH дает столь незначительное ускорение. Ответ на этот вопрос мы получили позже после анализа загрузки блоков GPU (см. раздел «Обсуждение результатов»). Забегая немного вперёд: трилеты снижают количество кэш-линий, которые должны быть загружены в кэш, однако они никак не помогают снизить количество запросов из мультипроцессора в кэш даже если многие из этих запросов идут к одним и тем же кэш-линиям.

(5) Сжатие данных узлов (или пониженная точность) для компактного представления деревьев BVH [MW06; YKL17;LV16]. Известно, что сжатый широкий BVH уменьшает трафик памяти и повышает производительность некогерентных лучей как для центральных [WBB08], так и для графических процессоров [YKL17]. В работах [WBB08, YKL17] утверждается, что квантование координат плоскостей для широких BVH деревьев уменьшает трафик памяти и повышает производительность для некогерентных лучей, причем и для CPU и для GPU. Однако для BVH2 подобный подход не был ранее исследован. Мы предложили свой вариант кодирования, основанный на небольших трилетах, помещающихся в 64-килобайтную (глубиной 2 уровня) или 128-килобайтную (глубиной 3 уровня) кэш-линию.

(6) Использование меньшего количества ограничивающих плоскостей, чтобы избежать избыточных тестов на пересечение плоскостей, общих для соседних в пространстве узлов [WK06; EWM08; LVY*20]. Эти подходы уменьшают трудоемкость арифметических операций, и поэтому их следует учитывать в случае, когда узким местом в алгоритме обхода становится математика. Также сообщается, что Dual-Split деревья уменьшают общий размер памяти для хранения дерева [LVY*20].

Мы отказались от этого направления на раннем этапе исследования по одной простой причине. Для BVH деревьев и даже для мобильных GPU средство анализа производительности (и некоторые другие наши эксперименты, о которых пойдет речь позже) показали, что мультипроцессор GPU загружен вычислениями лишь на 20%. Следовательно, сокращать число арифметических операций не имеет большого смысла.

(7) Сжатие геометрии, подобное квантованию данных в вершинах мешей [SE10; YM06]. Мы не стали исследовать это направление, поскольку анализ количества пересечений с узлами BVH дерева и с треугольниками в листьях показал, что в среднем узлов BVH дерева обычно пересекается на 1 порядок больше, чем треугольников.

(8) Реализация двухуровневого обхода. В большинстве существующих современных работ по трассировке лучей на графическом процессоре рассматривается производительность обхода одноуровневого дерева, содержащего геометрию в виде единой треугольной сетки [AL09;YKL17;PT22;MB22]. Двухуровневые реализации существуют, но они либо специфичны [GBPG11; ZQW23], либо используют устаревшие структуры данных [KBS11]. Поэтому объединение различных алгоритмов и структур данных для прохождения TLAS (Top Level Acceleration Structure) и BLAS (Bottom Level Acceleration Structure) открывает нам простор для исследований.

(9) Mega-kernel vs separate-kernel [Megakernel]. Этот пункт не столько относится к трассировке лучей как таковой, сколько к факторам, влияющим на производительность приложения в целом. Несмотря на ряд работ по

этой теме, до сих пор нет прямого сравнения, где бы одна и та же реализация трассировки лучей была бы выполнена в виде двух описанных способов. Мы выполнили такую реализацию благодаря возможностям используемой нами технологии программирования kernel_slicer, в которой из одной реализации на C++ поддерживается генерация кода как в режиме убер-ядра, так и в режиме разделённых ядер. При этом в отличие от DrJIT, в kernel_slicer программист сам определяет границы разбиения кода на вычислительные ядра, что позволяет произвести разделить несвязанные между собой участки кода более тщательно и, таким образом, провести более аккуратный эксперимент.

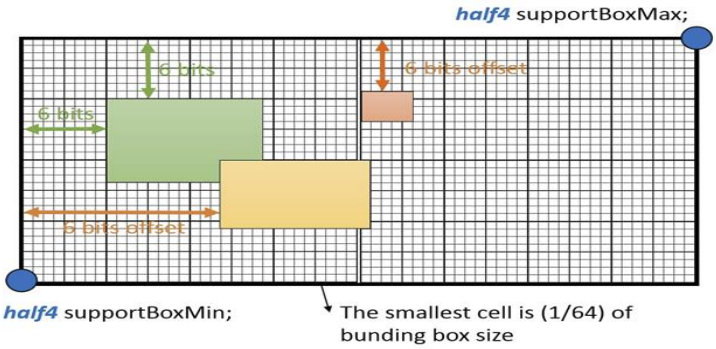
III. Предложенные методы

1. *Сжатие узлов*

При разработке метода, повышающего эффективность работы подсистемы памяти нужно принимать во внимание тот факт, что аппаратура (как CPU так и GPU) работает на уровне кэш-линий, блоков размером 64, 128 и т.п. байт. Поэтому, если только не принимать во внимание out-of-core подходы (где данные BVH дерева могут подгружаться из долговременной памяти), нет большого смысла в методах, которые сохраняют относительно-крупные блоки данных вроде COLBVH [YM06]. Трилет целиком должен помещаться в одну кэш-линию, в противном случае аппаратура так или иначе будет генерировать несколько запросов обращений в память.

Поэтому мы реализуем сжатие узлов дерева таким образом, чтобы один трилет помещался ровно в одну кэш-линию. Как было предложено в [LV16], мы использовали 6 бит на плоскость ограничивающего прямоугольника и сохранили «поддерживающий» ограничивающий прямоугольник с плавающей половинной точностью в «голове» трилета (рис. 4). В качестве ***базового*** метода мы рассматриваем FatBVH с двумя дочерними узлами *без использования трилетов для 64-байтной* кэш-линии и с использованием трилетов и *добавлением одного пустого узла* для выравнивания для большого размера кэш-линии. В базовом методе используются 16-битная плавающая точка и размер одного узла равен 32 байтам (size = 6*2*sizeof(half) + sizeof(int)*2 = 6*2*2 + 4*2 = 32). Нетрудно заметить, что для 64-байтной кэш-линии практически любой шаг в таком дереве, во-первых, приводит к чтению новой кэш-линии, а во-вторых, не использует половину прочитанных узлов (проблема чрезмерной выборки данных, overfetch). Мы называем наш подход «**BVH2Fat06**».

Рис. 4. Предлагаемая 6-битная схема кодирования координат плоскостей. Регулярная сетка внутри



опорной рамки показывает возможные положения плоскостей внутренних рамок. Синие точки отмечают минимум и максимум, сохраняемые с половинной точностью.

В нём один узел помещается в «int4» и занимает ровно 16 байт. Координаты опорного бокса хранятся во вспомогательном элементе «half8» в начале блока. Следовательно, размер нашего блока равен sizeof(half8) + (4+3)*sizeof(int4) = 128 байт для ячеек дерева с 4 конечными узлами; sizeof(half8) + (2+1)*sizeof(int4) = 64 байта для ячеек дерева с 2 конечными узлами; для ячеек трилета с 8 конечными узлами размер нашего блока равен sizeof(half8) + (8+7)*sizeof(int4) = 256 байтам и т.д.

Существенным недостатком низко-битных методов является необходимость в дополнительных арифметическая операциях и увеличенном регистровом давлении, которые появляются из-за распаковки данных. Сравнив производительность BVH2Fat16 (базового алгоритма обхода) и BVH2Fat06, мы можем оценить стоимость этой дополнительной арифметики.

2. Безстекэвый обход BVH4 дерева

В качестве базового алгоритма мы взяли реализацию без-стэкового обхода из работы [BK16]. Предложенный алгоритм (**BVH4_SHRT**)** расширяет этот метод на широкие BVH деревья, с 4 узлами. Структура BVH вершины в 4-узловом дереве повторяет структуру вершины бинарного дерева из работы [BK16]. Для нумерации вершин используется следующее правило: если номер текущей вершины равен k, то номера дочерних узлов слева направо равны: 4k, 4k + 1, 4k + 2, 4k + 3. Таким образом, в битовом представлении номеров при переходе с одного уровня на нижележащий происходит сдвиг ключа на 2 бита: k << 2. Это позволяет быстро вернуться к отложенному номеру вершины во время обратного прохода: нужно посчитать количество уровней N , на которые нужно подняться, сдвинуть текущий номер вершины на количество бит, равное 2 ∗ N , и определить, с какой из четырех вершин необходимо продолжить обход.

Так как широкое дерево имеет четыре дочерних узла, необходимо сохранять порядок пересечения вершин на текущем уровне. Для этого дополнительно используются 3 целых числа, представляющие собой битовые маски: firstOffset, secondOffset, и thirdOffset Алгоритм обхода и сохранения информации для обратного прохода на псевдокоде:

```
nodes F- новые вершины
oldPositions F- [1, 2, 3, 4]
сортировка(nodes)
сортировка(oldPositions)
если (пересечено 2 или более вершин)
{
  thirdOffset = thirdOffset F< 3
  secondOffset = secondOffset F< 3
  firstOffset = firstOffset F< 3
  numLevels = numLevels F< 1 | 1
}
если (пересечено 4 или более вершин) { thirdOffset = thirdOffset | oldPositions[3] }
если (пересечено 3 или более вершин) { secondOffset = secondOffset | oldPositions[2] }
если (пересечено 2 или более вершин) { firstOffset = firstOffset | oldPositions[1] }
если (пересечено 1 или более вершин)
{
  nodeKey = (nodeKey F< 2) + oldPositions[0] - 1
  если (пересечена ровно 1 вершина) { numLevels = numLevels F< 1 }
}
```

В firstOffset сохраняется номер первой вершины, которую нужно обработать при обратном проходе. Аналогично, в secondOffset и thirdOffset хранятся номера второй и третьей отложенных вершин. Во время обратного прохода необходимо определить адресы вершины, с которой нужно продолжить обход. Уровень считается обработанным, когда 3 младших бита каждой битовой строки равны 0. Алгоритм возврата к вершинам на псевдокоде:

```
numLev = КоличествоНулейВМладшихБитах(numLevels)
numLevels = numLevels F> numLev
pos = 0
19
first = firstOffset & 0x7
second = secondOffset & 0x7
third = thirdOffset & 0x7
isFirstZero = first F= 0
isSecondZero = second F= 0
isThirdZero = third F= 0
если (!isFirstZero) {
  pos = first
  firstOffset = firstOffset ^ first
  isFirstZero = true
}
иначе если (!isSecondZero) {
  pos = second
  secondOffset = secondOffset ^ second
  isSecondZero = true
}
иначе если (!isThirdZero) {
  pos = third
  thirdOffset = thirdOffset ^ third
  isThirdZero = true
}
если (isFirstZero F& isSecondZero F& isThirdZero) { F/ не осталось вершин на
```

```
  уровне↩→
  firstOffset = firstOffset ^ 3
  secondOffset = secondOffset ^ 3
  thirdOffset = thirdOffset ^ 3
  numLevels = numLevels ^ 1
}
nodeKey = ((nodeKey F> (2 * numLev)) & zeroTwoBits) + pos -1
address = hashTable[(nodeKey + displacementTable[nodeKey & dSize]) % hSize]
```

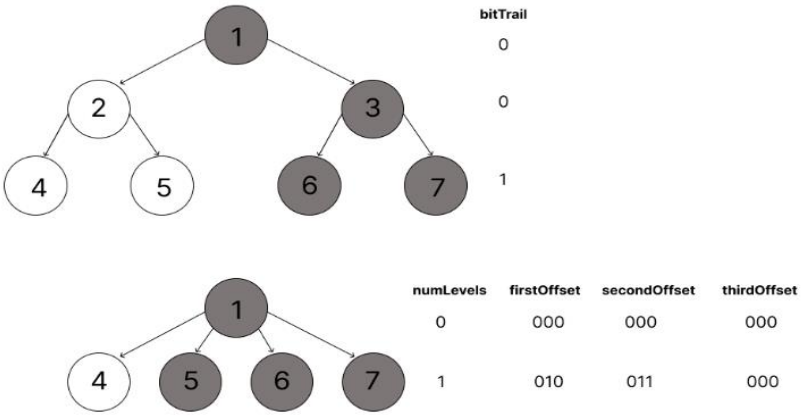


Рис. 5. *Схематичное представление алгоритма сохранения вершин. Сверху — базовый алгоритм, снизу — модифицированный.*

На рисунке (рис. 5) представлены схемы указанных выше алгоритмов. Серым цветом отмечены вершины, с которыми было найдено пересечение, белыми — вершины без пересечений. В примере для базового алгоритма на одном уровне пересечены обе вершины, поэтому младший бит равен 1. Пусть первой была пересечена вершина 7, тогда при обратном проходе достаточно посчитать количество младших нулей в bitTrail (равно 0), при обходе nodeKey = 7, следовательно, новый nodeKey = (nodeKey >> 0) XOR 1 = 6. В случае с 4 вершинами помимо количества уровней для подъема по дереву необходимо знать порядок обхода вершин. Порядок пересечения вершин на схеме: 7, 5, 6. Таким образом, в трех младших битах firstOffset хранится относительный порядок вершины с nodeKey = 5 (2, в битовом представлении 010), а в secondOffset — порядок вершины с nodeKey = 6 (3, в битовом представлении 011). Третья вершина не пересечена, поэтому младшие биты thirdOffset равны 0. При возврате младший бит numLevels обнуляется тогда, когда все три младших бита firstOffset, secondOffset и thirdOffset равны 0, то есть на уровне не осталось необработанных вершин.

3. Исследование алгоритмов обхода двух-уровневого дерева

При реализации двухуровневого обхода имеются некоторые неочевидные моменты. Во-первых, появляется значительно больше вариантов, чем while-while или if-if, рассмотренные в работе [AL09]. Во-вторых, в какое-то место обхода нужно поместить переход между уровнями дерева. Этот переход заключается в умножении луча на матрицу инстанса и на самом деле является достаточно тяжелым. Это является проблемой, поскольку в наиболее очевидной реализации (BVH2_LOFT) этот переход будет осуществляться внутри основного цикла обхода в том месте, в самой “горячей” его части. Мы рассмотрели несколько способов, позволяющих вынести этот переход за основной цикл и, таким образом, разгрузить его.

Таблица 1. Различные использованные нами форматы узла дерева деревьев.

Number	

F#1	`struct BVHNode {`
	`  float3 boxMin;`
	`  uint  leftOffset;  //!< left offset or leaf`
	`  float3 boxMax;`
	`  uint  escapeIndex; //!< escapeIdex (i.e. 'rope') [TR09]`
	`};`
	`rightOffset == leftOffset+1;`
	Благодаря escapeIndex, этот узел можно использовать как для перемещения по стеку, так и без него с помощью предложенной связи из работы [TR09]. Один узел занимает 32 байта, а два соседних узла дерева 64 байта, причем они всегда занимают одну строку кэша, если leftOffset кратно 2(что легко проверяется). Компоновка (т.е. расположение узлов относительно друг друга) может быть практически любой, однако на практике это зависит от глубины или ширины.

	Компоновка по ширине подходит для быстрого и параллельного переоборудования.

F#2	`struct BVHNodeFat {`
	`  float4 lmin_xyz_rmax_x;`
	`  float4 lmax_xyz_rmax_y;`
	`  float4 rmin_xyz_rmax_z;`
	`  uint32_t offs_left;`
	`  uint32_t offs_right;`
	`  uint32_t dummy1;  // добавлено для выравнивания`
	`  uint32_t dummy2;  // добавлено для выравнивания`
	`};`
	32-разрядная версия. Построение данных может быть любым, поддерживаются данные типа COLBVH.
	Занимает 64 байта для одного узла, сохраняет 2 дочерних ограничивающих полей.

F#3	`struct BVHNodeFat16 {`
	`  half4 lmin_xyz_rmax_x;`
	`  half4 lmax_xyz_rmax_y;`
	`  half4 rmin_xyz_rmax_z;`
	`  uint32_t offs_left;`
	`  uint32_t offs_right;`
	`};`
	16-разрядная версия. Построение данных может быть любым, поддерживаются данные типа COLBVH.
	Занимает 32 байта для одного узла, сохраняет 2 дочерних ограничивающих полей.

F#4	`struct BVHNode16 { // фактически, распакованный из одного uint4 в шейдерах`
	`  half2  minXY;`
	`  half2  minZmaxY;`
	`  half2  maxYZ;`
	`  uint32_t leftOffset;`
	`};`
	16 битная версия F#1. Не сохраняет escapeIndex, поэтому не может использоваться при обходе в безстековых алгоритмах с помощью верёвок из работы [TR09]. Для одного узла требуется 16 байт, что позволяет разместить 4 узла в одной 64-байтовой строке кэша для обхода BVH4.

F#5	`struct Treelet<N> {`
	`  half4 boxMin;`
	`  half4 boxMax;`
	`  int4  nodes[N*2-1];`
	`};`
	Предложенный формат CALBVH. В нем хранится 16 байт для поддержки ограничивающего объема и 16 байт для каждого узла с двумя сжатыми 6-битными ограничивающими рамками. Для 64-байтовой строки кэша N должно быть равно 2, что соответствует 2-уровневым ячейкам и содержит в общей сложности 6 ограничивающих объемов. Для 128-байтовой строки кэша N должно быть равно 3-уровневым ячейкам и содержать в общей сложности 12 ограничивающих рамок.

Таблица 2. Различные, исследованные нами алгоритмы двух-уровневого обхода.

Название	Детали реализации

BVH2_STACKLESS_II
BVH2_STSS_II
	Формат используемого узла: F#1
	Двухуровневый без стековый обход как TLAS и BLAS без вложенных циклов с использованием связи из работы [TR09], а также с реализацией if-if в соответствии с [AL09].
	`while(nodeId < ESCAPE_ROOT and TLAS):`
	`  node := nodes[nodeId]`
	`  nodeId := node.nextId();`
	`  if node.isLeaf() and TLAS:`
	`  nodeId := nextBLAS(node).root`
	`  if node.isLeaf() and BLAS:`
	`  intersectTris(node)`
	`  if nodeId >= ESCAPE_ROOT and BLAS:`
	`  nodeId := continueTLAS`

BVH2_STACKLESS_WW
BVH2_STSS_WW
	Формат используемого узла: F#1
	Двухуровневый без стековый обход как TLAS и BLAS с циклами двойной вложенности с использованием связи из работы [TR09], а также с реализацией while-while в соответствии с [AL09].
	`while(nodeId < ESCAPE_ROOT and TLAS):`
	`  while (node is not leaf):`
	`  node := next(node)`
	`  if TLAS:`
	`  nodeId := nextBLAS(node).root`
	`  if BLAS:`
	`  intersectTris(node)`

	`if nodeId >= ESCAPE_ROOT and BLAS:` `  nodeId := continueTLAS`

BVH2_STAS
 	
Stackless
Top
Acceleration
Structure
	Формат используемого узла: F#1
	Обрабатывает TLAS, используя связи [TR09] и BLAS с помощью алгоритма на основе стека, который в общей сложности приводит к трем вложенным циклам:
	`while(nodeId < ESCAPE_ROOT):`
	`  nodeT := nodes[nodeId]`
	`  nodeId:= nodeT.nextId()`
	`  if nodeT is leaf:`
	`  node := nextBLAS(nodeT).root`
	`  while(stack.notEmpty):`
	`  while(node is not leaf):`
	`  node := next(node)`
	`  intersectTris(node)`

BVH2_LOFT
 	
Left
Offset
Format of
Tree
	Формат используемого узла: F#1
	Обход, основанный на стеке как для ATLAS, так и для BLAS в двух вложенных циклах:
	`while(stack.notEmpty):`
	`  while(node is not leaf):`
	`  node := next(node)`
	`  if TLAS:`
	`  node := nextBLAS(node).root`
	`  else if BLAS:`
	`  intersectTris(node)`

BVH2Fat32	Формат используемого узла: F#2 Псевдокод обхода аналогичен обходу BVH2_STAS
BVH2Fat16	Формат используемого узла: F#3 Псевдокод обхода аналогичен обходу BVH2_STAS
BVH2Fat06	Формат используемого узла: F#5 Предложенный CALBVH обход. Считывает ограничивающий объем только в том случае, если изменен элемент дерева. Псевдокод обхода такой же, как и для BVH2_STAS.

BVH4_32	Формат используемого узла: F#1
	
	Использует без стековый обход BVH2 со связями, описанными в [TR09] для TLAS и обход на основе стека BVH4 для BLAS в цикле if-if в соответствии с [AL09].
	`while(nodeId < ESCAPE_ROOT):`
	`  nodeT := nodes[nodeId]`
	`  nodeId:= nodeT.nextId()`
	`  if nodeT is leaf:`
	`  node := nextBLAS(nodeT).root`
	`  while(stack.notEmpty):`
	`  if(node is not leaf):`
	`  node = next(node)`
	`  else:`
	`  intersectTris(node)`

BVH4_16	Формат используемого узла: F#4 Аналогичен BVH4_32

BVH2_SHRT32
 	
Stackless
Hierarchical
Ray
Tracing
	Формат используемого узла: F#1
	Безстековый алгоритм. Использует связи как в работе[TR09] для TLAS и иерархический обход без использования стека [BK16] для BLAS. Используется 32-битовая строка.
	`while(nodeId < ESCAPE_ROOT):`
	`  nodeT := nodes[nodeId]`
	`  nodeId:= nodeT.nextId()`
	`  if nodeT is leaf:`
	`  nodeAddr := nextBLAS(nodeT).rootAddr`
	`  while(nodeAddr < ESCAPE_ROOT):`
	`  while(node is not leaf):`
	`  node := next(node)`
	`  intersectTris(node)`
	`  nodeAddr := backtracking(node)`

BVH2_SHRT64	Формат используемого узла: F#1 Безстековая реализация. Использует связи как в работе[TR09] для TLAS и модифицированный иерархический обхода без стека [BK16] для BLAS. Используется 64-битная строка. Псевдокод аналогичен BVH2_SHRT32
BVH4_SHRT64	Формат используемого узла: F#1 Реализация без использования стека. Использует связи как в работе[TR09] для TLASи модифицированный иерархический обхода без стека [BK16] с дочерними 4 узлами для BLAS.

	Используется 64-битная строка. Псевдокод обхода такой же, как и предыдущий.
BVH4_SHRT64_HALF	Формат используемого узла: F#4 Аналогично предыдущему(BVH4_SHRT), за исключением сохранения ограничивающих объемов при помощи чисел с плавающей точкой половинной точности, 4 ограничивающих объема в одной строке кэша. Псевдокод обхода такой же, как и в предыдущем.

Тестовые сцены и устройства

Для наших тестовых запусков использовались известные сцены с различными характеристиками, чтобы охватить различные сценарии использования BVH. Использовались сцены с множеством различный объектов (instanced_objects), с одним единым мешем, состоящим из множества полигонов (cry_sponza_c), так и меши, состоящие из множества тонких\мелких частей (рис. 6).

Для тестирования BVH на мобильных платформах было решено использовать устройства разной производительности. Для тестирования слабой производительности был выбран Realme 10 (модель RMX3630) с GPU Mali-G57 MC2. Для среднего сегмента был выбран имеющийся в наличии Huawei MatePad Air с GPU Adreno 660. Разные конфигурации GPU выбраны, чтобы проверить одинакова ли соотношение производительности при разных характеристиках устройств. Так же для сравнения с десктопным GPU в эксперименте №4 использовался телефон с Adreno 730 как представитель высокопроизводительного GPU. Десктопными аналогами являлись RTX 2070 и RTX4090.

Рис. 6. *Изображения тестовых сцен. Первое число (до '/') для каждой сцены отображает количество уникальных треугольников, хранящихся в памяти, второе (после '/') – количество видимых (инстанцированных) треугольников. Число после запятой – число инстансов. Далее характеристики сцен:*
(1) instanced_objects: 170K/62M m., 601 инст; (2) dragon: 871K/871K m., 1 инст. (3) bonsai: 650K/650K m./ 1инст; (4) sponza_c: 66K/66K m., 1 инст. (5) cry_sponza_c: 244K /262K m., 17 инст; (6) cry_sponza: 244K/262K m., 281 инст. (7) hero: 28K/28K m., 4 инст; (8) human_sophi : 16.5K/16.5K m., 4 инст.

IV. Эксперименты

Во всех экспериментах замеры проводятся на 10000 кадров подряд. Полученные результаты усредняются. Это необходимо для нивелирования влияния перегрева и вмешательства планировщика задач на производительность мобильного устройства.

Эксперимент №1

В этом эксперименте мы хотели оценить влияние степени расхождения (некогерентности) лучей на производительность. Для того чтобы создать воспроизводимый эксперимент мы сделали следующим образом:

1. Зафиксировали параметры BVH дерева, алгоритма обхода и т.д. для всех сцен.
2. Использовали только трассировку только для первичной видимости
3. Постепенно увеличивали разрешение синтезируемого изображения и измеряли производительность трассировки в количестве миллионов лучей в секунду. Разрешение менялось по следующему правилу: 512x512, 1024x1024, 2048x2048, 4096x4096

Таким образом, в каждом следующем замере мы повышали степень когерентности лучей, всё чаще и чаще вызывая ситуацию, когда все лучи SIMD группы warp обращаются по одному и тому же адресу в процессе спуска по BVH дереву. Результат представлен в рис. 7.

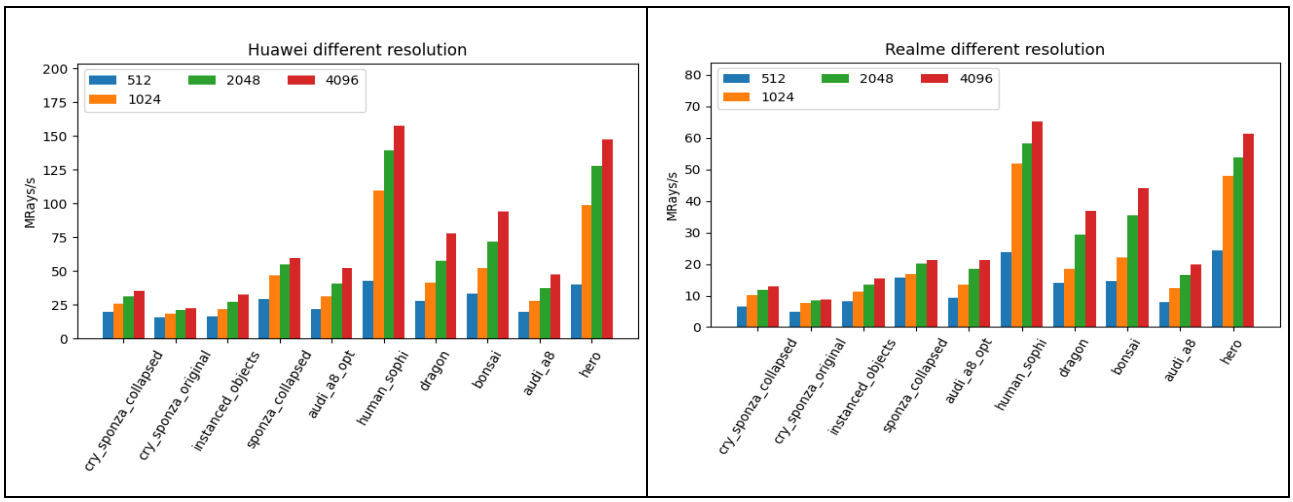


Рис. 7. Количество миллионов лучей в секунду на Adreno 660 (слева) и Mali-G57 (справа) при различных разрешениях на разных сценах.

Эксперимент №2 (рис. 8)

В этом эксперименте мы хотели протестировать различные способы реализации двухуровневого обхода и хотели измерить для них, насколько различается производительность обхода двухуровневого и одноуровневого дерева для одной и той же сцены. Для этого мы взяли сцену cry_sponza где TLAS состоит из примерно 300 объектов (инстансов), значительно перекрывающихся друг с другом и объединили все инстансы в одну сплошную геометрию (cry_sponza_collapsed).

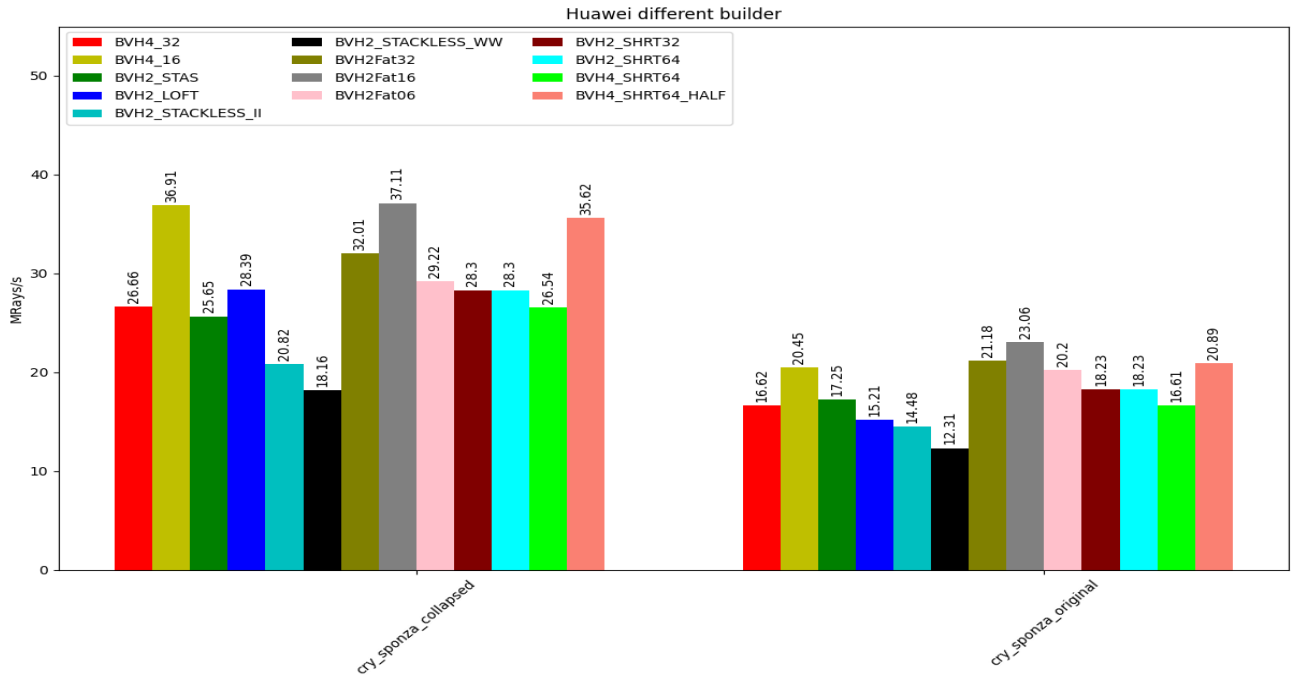


Рис. 8.1 Количество миллионов лучей в секунду на Adreno 660 при различных алгоритмах обхода на двух сценах.

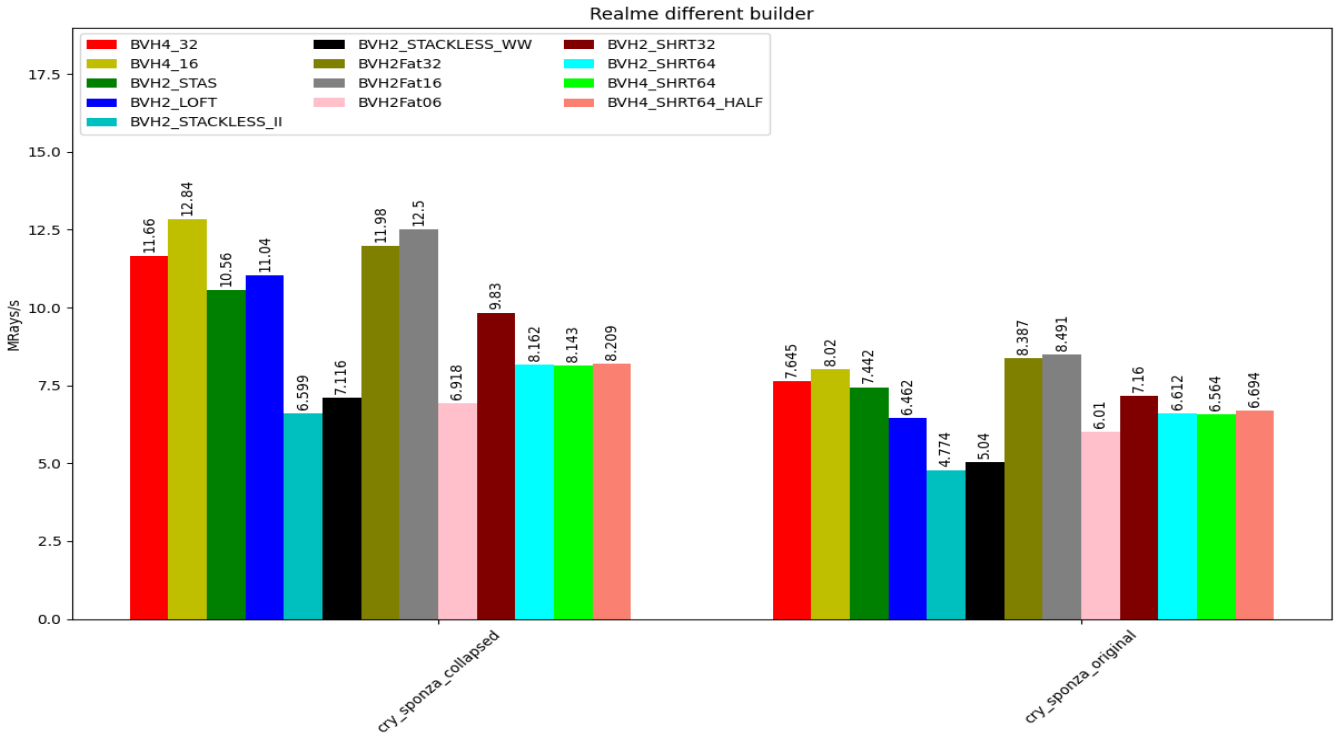


Рис. 8.2 Количество миллионов лучей в секунду на Mali-G57 при различных алгоритмах обхода на двух сценах.

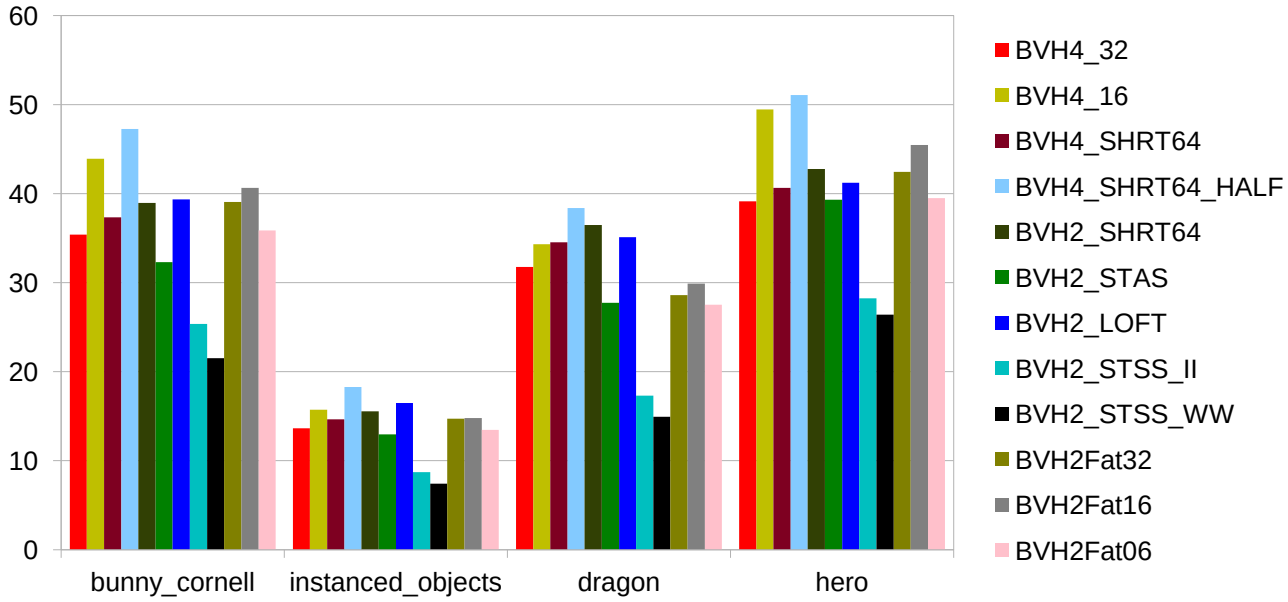


Рис. 8.3 Количество миллионов лучей в секунду на Adreno 660 при различных алгоритмах обхода на остальных сценах.

Эксперимент №3 (рис. 9)

В этом эксперименте мы хотели проверить гипотезу о низкой эффективности убер-ядер на современных GPU и оценить, насколько сильно регистовое давление может изменить критерий выбора алгоритма обхода. Для этого мы реализовали т.н. Whitted трассировку лучей и при помощи кодо-генератора, умеющего это делать из коробки [kernel_slicer], и сгенерировали по две реализации для каждого исследуемого нами алгоритма.

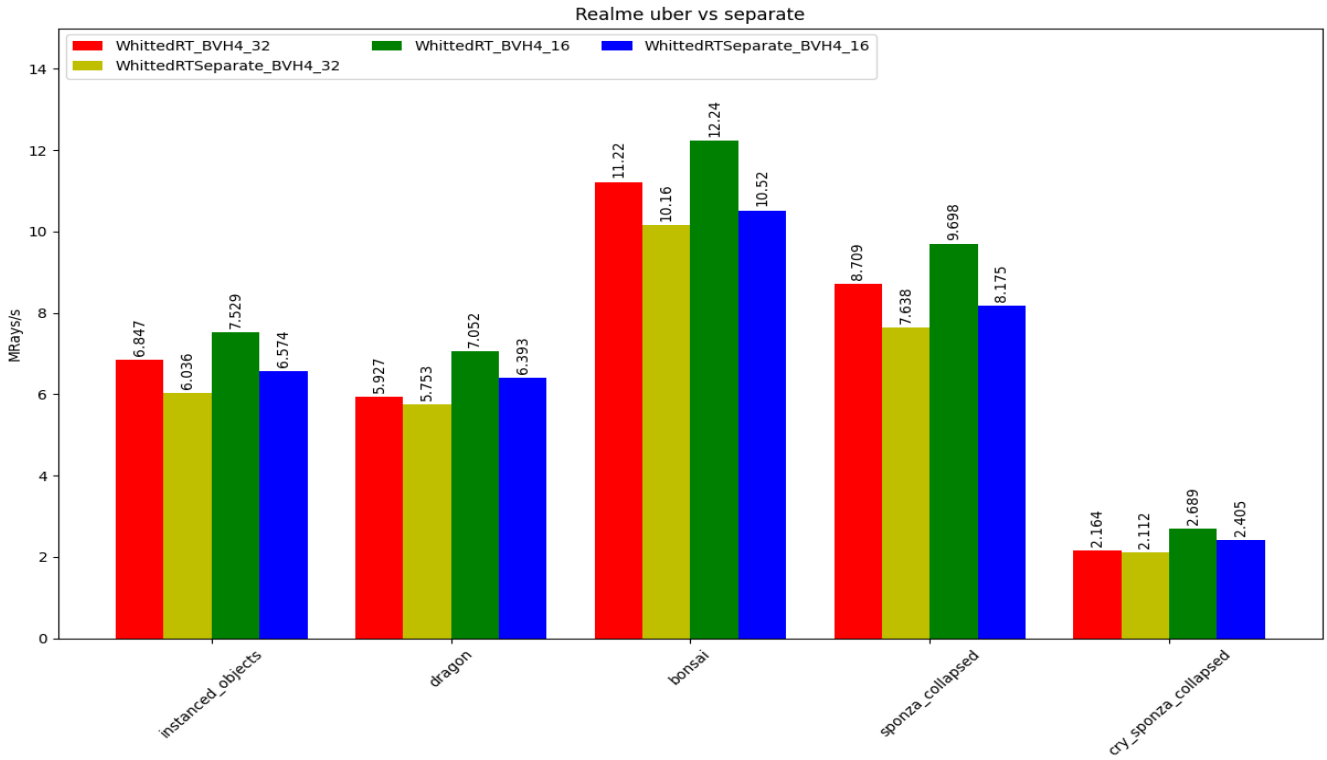


Рис. 9.1 Количество миллионов лучей в секунду на Mali-G57 при использовании одного большого ядра (WhittedRT) или раздельных ядер(WhittedRTSeparate) в двух различных реализациях обхода BVH4 дерева.

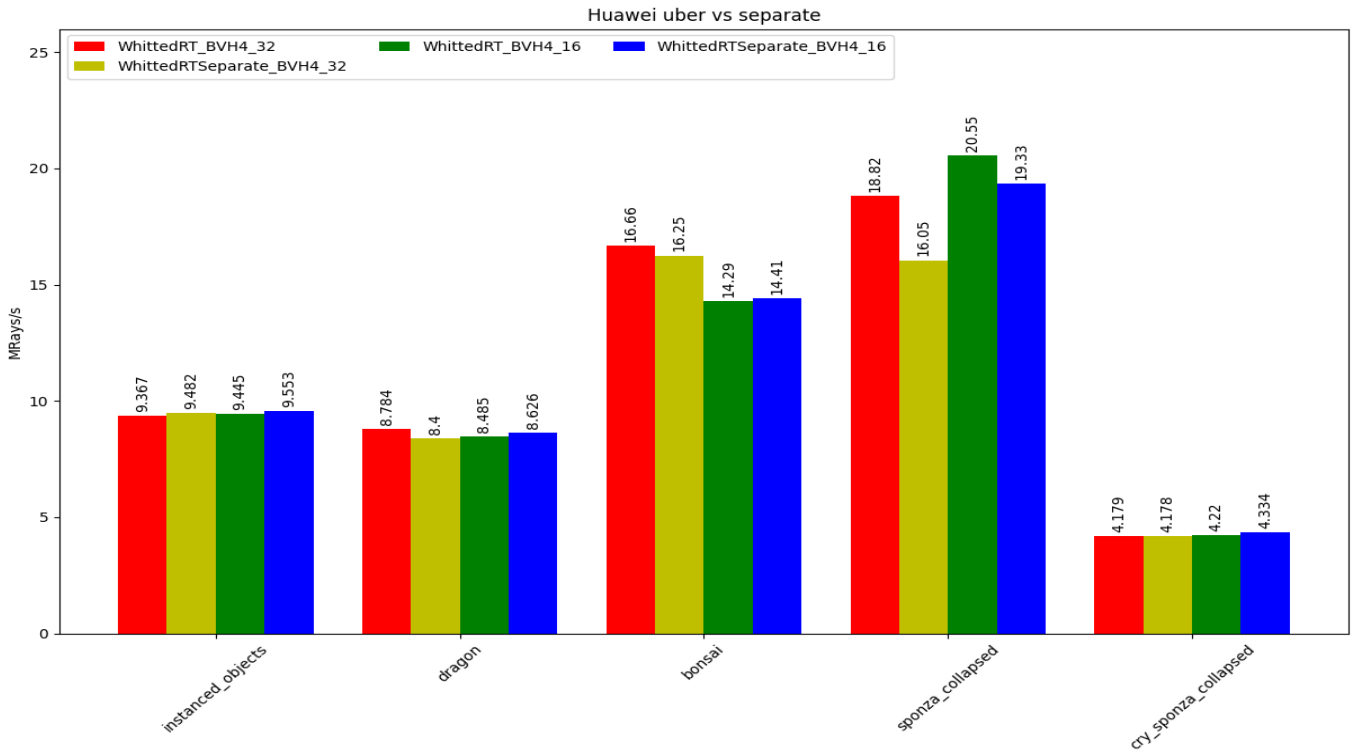


Рис. 9.2 Количество миллионов лучей в секунду на Adreno 660 при использовании одного большого ядра(WhittedRT) или раздельных ядер(WhittedRTSeparate) в двух различных реализациях обхода BVH4 дерева.

V. Сравнения с существующими решениями

Мы сравнили нашу скорость обхода BVH деревьев лучей с фреймворком Hippie [MB22] (рис. 10). Небольшая разница в производительности объясняется 4 основными факторами:

1. Другая реализация построения дерева. Например, в hippie SBVH строится крайне медленно (~ в 100 раз медленнее, чем у нас)
2. Реализация обхода в hippe одноуровневая, а у нас — двухуровневая.
3. Вершины в hippie сортируются и хранятся в непрерывном массиве («список треугольников»), в то время как наша реализация хранит индексированную геометрию в памяти. Дополнительный уровень косвенной адресации, таким образом, увеличит время пересечения луча и треугольника.

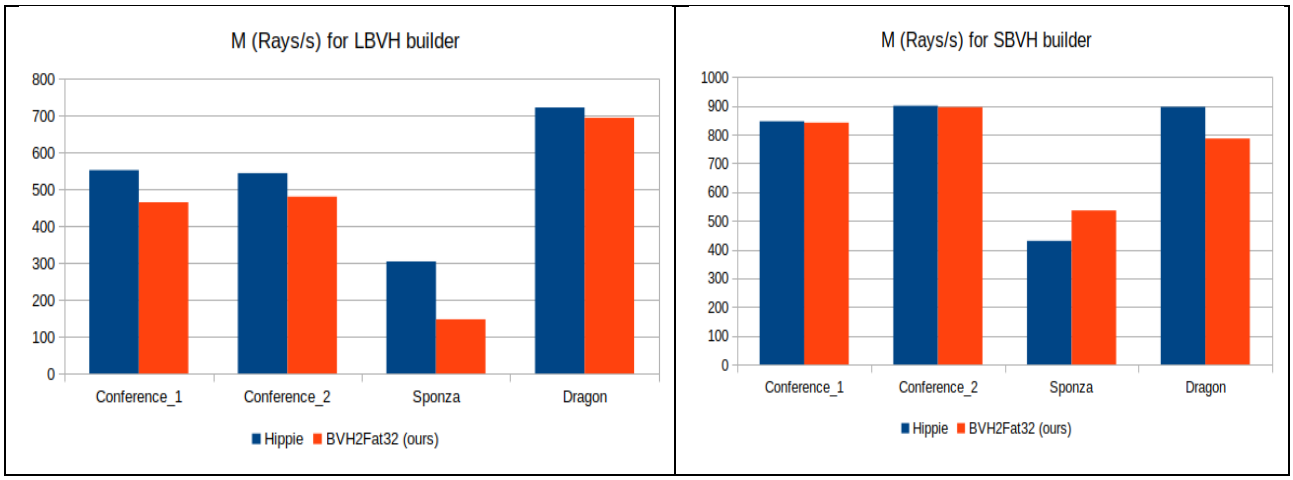


Рис. 10. *Сравнение разработанного решения (BVH2Fat32, красный столбец) и [MB22] (Hippie, синий столбец) на видеокарте Nvidia 3070M.*

V. Обсуждение результатов

Эксперимент №1 показывает, что при увеличении разрешения изображения (и соответственно степени когерентности лучей) на простых сценах производительность растёт быстро и нелинейно. Напротив, в сложных сценах производительность растет медленно и линейно. Это является следствием того, что на простых сценах при увеличении когерентности лучей всё чаще встречается ситуация, когда вся SIMD-группа потоков warp обращается по одному и тому же адресу при чтении данных из массива узлов BVH дерева. На сложных сценах с большой глубиной дерева этого практически не происходит, т.к. даже в большом разрешении лучи в соседних пикселах будут большую часть времени обходить разные участки BVH дерева.

Эксперимента №2 показывает несколько моментов:

(2.1) предложенная схема реализации двухуровневого обхода (все алгоритмы кроме BVH2_LOFT и **_STACKLESS), когда обход TLAS осуществляется в отдельном внешнем цикле (и умножение луча на матрицу находится именно в этом цикле), лучше, чем традиционный вариант, осуществляющий умножение луча на матрицу во внутреннем цикле (BVH2_LOFT).

(2.2) Наиболее производительными оказались варианты, сокращающие именно число обращений в память к отдельным кэш линиям: BVH4_16, BVH2Fat16 и предложенный нами алгоритм BVH2Fat06, который незначительно отстаёт от BVH2Fat16, но при этом является в 2 раза более компактным по памяти.

(2.3) Предложенные нами алгоритмы безстэкового обхода широких BVH деревьев BVH4_SHRT* побеждает базовый метод (BVH2_SHRT64 [BK16]) и другие методы обхода дерева (рис. 8.3) на сценах, в которых лучи преимущественно спускаются по дереву, но редко поднимаются. То есть почти всегда находят пересечение в первом просмотренном листе. Это сцены с преимущественно открытыми пространствами либо простой геометрией. На более сложной сцене Cry Sponza, с другой, стороны, дорогая процедура бэк-трекинга, заменяющего обращения в стек, в этих алгоритмах становится ограничивающим фактором.

Эксперимент №3 показывает, что для простой геометрии убер-ядро работает быстрее, поскольку затраты на передачу данных больше между вычислительными ядрами, чем время трассировки лучей. Для тяжелой геометрии реализация через разделённые ядра работает быстрее на мобильном GPU (но не на десктопном).

Итог. Исследовав большое число подходов, мы делаем неутешительный вывод: по большому счету не столь важно, какой именно обход и какие именно структуры данных используются. Различные опробованные нами подходы давали лишь небольшое ускорение, в лучшем случае пределах 25%-50% относительно более-менее прямолинейных и очевидных реализаций. Во всех этих алгоритмах узкое место – неэффективная операция сбора данных (gather), когда разные лучи читают узлы по разным адресам в памяти. Это не приводит к большому числу кэш-промахов, как можно было бы ожидать. Низкая производительность операции gather проявляется даже на тех сценах, которые целиком помещаются в L2 кэш. В сочетании с данными GPU профилировщиков мы заключаем, что узкое место находится между L2 кэшем и собственно мультипроцессором GPU. Мультипроцессор отправляет в кэш запросы. Почти все запросы попадают в кэш. Многие из этих запросов могут даже попадать в одну кэш-линию. Однако, если запросы идут по разным адресам, это ломает механизм распределения данных (broadcast), в результате чего разбивает одно обращение в кэш на несколько (скорее всего в пределах 8). Наши синтетические эксперименты показали, что, если большую часть времени все потоки будут выбирать данные по одному адресу, производительность вырастает до 5 раз.

VI. Заключение

Таким образом, мы провели анализ производительности множества подходов и предложили три улучшения:

1. Мы предложили низко-битное представление для BVH деревьев с двумя дочерними узлами в виде трилетов, которое несущественно снижает производительность трассировки лучей на GPU с производительными в плане вычислений, экономя при этом память почти в 2 раза для BVH2 дерева.
2. Мы предложили без-стековый обход для широких BVH деревьев (с 4 дочерними узлами). Этот подход выигрывает в скорости у существующих подходов на сценах, где лучи большую часть времени спускаются по дереву, но редко поднимаются.
3. Мы предложили более эффективный алгоритм обхода двухуровневого дерева (все наши против BVH2_LOFT, который является известным ранее) на GPU, чем ранее известные методы: исследуя различные способы реализации обхода двухуровневого дерева мы обнаружили, что верхний уровень (TLAS) лучше обходить без-стековым алгоритмом в отдельном внешнем цикле, поскольку в этом случае удается оптимизировать внутренний цикл для обхода нижнего дерева BLAS, который обычно выполняется большее число итераций.

Мы разработали программное решение, не уступающее по производительности передовым решениям трассировки лучей на GPU (рис. 10). Наконец, мы оценили влияние способа реализации трассировки лучей через убер и разделённые ядра, и пришли к выводу что на большинстве современных GPU убер-ядро работает быстрее, т.к. накладные расходы на передачу данных между вычислительными ядрами слишком большие. На слабых GPU и сложных сценах разделённые ядра всё ещё могут иметь некоторые преимущества в производительности, но не настолько значительные, чтобы это оправдывало более сложную в реализации схему рендера через разделённые ядра. Мы показали, что почти все существующие подходы имеют одно и то же узкое место (медленную реализацию gather): разделение данных кэша между разными потоками мультипроцессора, когда разные потоки читают данные из L2 кэша по различным адресам.

Список литературы

[IMRT] Kristof Beets, The six levels of ray tracing acceleration. Imagination white paper. URL = https://forums.macrumors.com/attachments/imagination-raytracing-primer-sept2020-pdf.1973926/

[MB22] Daniel Meister and Jiří Bittner, Performance Comparison of Bounding Volume Hierarchies for GPU Ray Tracing, *Journal of Computer Graphics Techniques (JCGT)*, vol. 11, no. 4, 1-19, 2022
Available online http://jcgt.org/published/0011/04/01/

[SurveyBVH] Daniel Meister, Shinji Ogaki, Carsten Benthin, Michael J. Doyle, Michael Guthe, Jiří Bittner. A Survey on Bounding Volume Hierarchies for Ray Tracing.

[AL09] Aila T., Laine S.: Understanding the Efficiency of Ray Traversal on GPUs. In Proceedings of HPG (2009), pp. 145–149. 16, 17, 23.

[AK10] AILA T., KARRAS T.: Architecture Considerations for Tracing Incoherent Rays. In Proceedings of PG (2010), p. 113–122. 18, 19

[AKL13] AILA T., KARRAS T., LAINE S.: On Quality Metrics of Bounding Volume Hierarchies. HPG(2013), pp. 101–108. 4, 5, 10

[LSS18] LIER A., STAMMINGER M., SELGRAD K. CPU-style SIMD ray traversal on GPUs. In Proceedings of the Conference on High-Performance Graphics, Association for Computing Machinery, 7:1–7:4. 2018. URL: https://doi.org/10.1145/3231578. 3231583. 7, 8, 11, 15

[EG07] ERNST M., GREINER G.: Early Split Clipping for Bounding Volume Hierarchies. In Proceedings of Symposium on Interactive Ray Tracing (2007), pp. 73–78. 9, 10

[SFD09] STICH M., FRIEDRICH H., DIETRICH A.: Spatial Splits in Bounding Volume Hierarchies. In Proceedings of the High-Performance Graphics (2009), pp. 7–13. 10, 22, 23
[SE10] S EGOVIA B., E RNST M.: Memory Efficient Ray Tracing with Hierarchical Mesh Quantization. In Proceedings of Graphics Interface (2010), pp. 153–160. 12, 13
[MW06] M AHOVSKY J., W YVILL B.: Memory-Conserving Bounding Volume Hierarchies with Coherent Raytracing. Computer Graphics Forum 25, 2 (2006), 173–182. 12
[LV16] LIKTOR G., VAIDYANATHAN K.: Bandwidth-efficient BVH Layout for Incremental Hardware Traversal. In Proceedings of High-Performance Graphics (2016), pp. 51–61. 8, 18, 19
[KBS11] Kalojanov J., Billeter M., Slusallek P. Two-level grids for ray tracing on GPUs //Computer Graphics Forum. – Oxford, UK : Blackwell Publishing Ltd, 2011. – Т. 30. – №. 2. – С. 307-314.
[PT22] Pieterjan Bartels and Takahiro Harada. 2022. Combining GPU Tracing Methods within a Single Ray Query. In SIGGRAPH Asia 2022 Technical Communications (SA '22). Association for Computing Machinery, New York, NY, USA, Article 17, 1–4. https://doi.org/10.1145/3550340.3564231
[ZQW23] Zellmann, S., Wu, Q., Ma, K. L., & Wald, I. (2023). Memory-Efficient GPU Volume Path Tracing of AMR Data Using the Dual Mesh.
[GBPG11] Garanzha, K., Bely, A., Premoze, S., & Galaktionov, V. (2011). Out-of-core GPU ray tracing of complex scenes. In *ACM SIGGRAPH 2011 Talks* (pp. 1-1).
[TR09] Torres, Roberto et al. “Ray casting using a roped BVH with CUDA.” Spring conference on Computer graphics (2009).
[BK16] BINDER N., KELLER A.: Efficient Stackless Hierarchy Traversal on GPUs with Backtracking in Constant Time. In Proceedings of High-Performance Graphics (2016), pp. 41–50. 17
[Megakernel] Megakernels Considered Harmful: Wavefront Path Tracing on GPUs. Samuli Laine, Tero Karras, and Timo Aila. High-Performance Graphics 2013.
[embree] Wald, I., Woop, S., Benthin, C., Johnson, G. S., & Ernst, M. (2014). Embree: a kernel framework for efficient CPU ray tracing. *ACM Transactions on Graphics (TOG)*, *33*(4), 1-8.
[WBB08] WALD I., BENTHIN C., BOULOS S.: Getting Rid of Packets - Efficient SIMD Single-Ray Traversal using Multi-Branching BVHs. In Symposium on Interactive Ray Tracing (2008), pp. 49–57. 10, 14, 22, 23
[PGDS09] POPOV S., GEORGIEV I., DIMOV R., SLUSALLEK P.: Object Partitioning Considered Harmful: Space Subdivision for BVHs. In Proceedings of High-Performance Graphics (2009), pp. 15–22. 4, 5, 10, 22
[YKL17] YLITIE H., KARRAS T., LAINE S.: Efficient Incoherent Ray Traversal on GPUs Through Compressed Wide BVHs. In Proceedings of High-Performance Graphics (2017), pp. 4:1–4:13. 10, 12, 16, 17, 22, 23
[YM06] YOON S.-E., M ANOCHA D.: Cache-Efficient Layouts of Bounding Volume Hierarchies. Computer Graphics Forum (2006). 8
[WK06] WÄCHTER C., K ELLER A.: Instant Ray Tracing: The Bounding Interval Hierarchy. In Proceedings Eurographics Symposium on Rendering (2006), pp. 139–149. 13
[EWM08] EISEMANN M., WOIZISCHKE C., MAGNOR M.: Ray Tracing with the Single Slab Hierarchy. In Proceedings of Vision, Modeling, and Visualization (2008), pp. 373–381. 13
[LVY*20] LIND., VASIOU E., YUKSEL C., KOPTA D., BRUNVAND E.: Hardware-Accelerated Dual-Split Trees. Proceedings of the ACM on Computer Graphics and Interactive Techniques 3, 2 (2020). 13
[kernel_slicer] URL=https://github.com/Ray-Tracing-Systems/kernel_slicer?tab=readme-ov-file

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-164-178

Восстановление параметров объёмного рассеяния на основе измерения ДФР
Reconstruction of volume scattering parameters on the base of BSDF measurements

В. Г. Соколов1, С. Куприянов, Д. Д. Жданов, И. С. Потемин, А. Д. Жданов
1Университет ИТМО, Санкт-Петербург, Российская Федерация

V. G. Sokolov, S. Kyprijanov, D. D. Zhdanov, I. P. Potemin, A. D. Zhdanov
ITMO University, St. Peterburg, Russian Federation

Аннотация. В работе исследуется восстановление параметров объёмного рассеяния с использованием двунаправленных функций рассеяния, измеренных для одного или нескольких образцов со светорассеивающей средой. Элементы с объёмным рассеянием широко распространены в современных оптических устройствах, таких как световоды, системы освещения, устройства подсветки, светорассеивающие плёнки и светильники. Корректное определение параметров объёмного рассеяния (размер, оптические свойства, концентрация светорассеивающих частиц) имеет решающее значение для корректного моделирования систем, имеющих элементы с объёмным рассеянием. Рассеяние света в таких элементах часто аппроксимируют с помощью двунаправленных функций рассеяния, но такое приближение возможно только в том случае, если толщиной объектов с объёмным рассеянием можно пренебречь. В статье рассматривается сложная проблема восстановления параметров объёмного рассеяния с акцентом на среды, в которых, светорассеяние можно определять на основе теории Ми. Основная проблема заключается в неоднозначности задачи, поскольку разные комбинации параметров объёмного рассеяния могут привести к одинаковому угловому рассеянию. В статье предложен метод восстановления параметров объёмного рассеяния с использованием процедуры двухуровневой локальной и глобальной оптимизации. Применение этого метода подтверждается успешным восстановлением нескольких образцов с различными рассеивающими свойствами, от сильно диффузных до практически прозрачных.

***Ключевые слова*:** Объёмное рассеяние, Теория Ми, Двунаправленная Функция Рассеяния, ДФР, Измерения ДФР, Уравнение переноса светового излучения

Аннотация. The work explores reconstructing volume scattering (VS) parameters using bi-directional scattering distribution functions (BSDF) measured for one or several samples with scattering medium. Elements with VS are prevalent in modern optical devices, such as light pipes, light guides, illuminating systems, backlight devices, diffuse sheets, and luminaires. Accurately defining VS parameters (size, optical properties, concentration of scattering agents) is crucial for simulating light scattering. Light scattering in such elements is often approximated with BSDF, but this is only feasible if the thickness of VS objects can be ignored. The article addresses the complex problem of reconstructing VS parameters, focusing on media described by Mie theory. The main challenge is ambiguity, as different combinations of VS parameters can result in similar angular scattering. The paper proposes a method to reconstruct VS parameters using a two-level optimization procedure based on local and global optimization algorithms. This method is validated by successfully reconstructing several samples with varying scattering properties, from highly diffuse to very specular.

***Ключевые слова*:** Volume Scattering, Mie theory, Bi-directional Scattering Distribution Function, BSDF, BSDF measurements, Light Transfer Equation, LTE.

I. Введение

Материалы с объёмным рассеянием (ОР) широко используются в современных оптических системах, таких как световоды, измерительные приборы, светопроводящие пластины для ЖК-дисплеев. Определение параметров сред с объёмным рассеянием (размер, оптические свойства, концентрация частиц) имеет решающее значение для моделирования и проектирования, но прямое точное измерение этих параметров является очень сложной задачей, требующей изготовления специальных образцов, сложного и дорогостоящего оборудования, а во многих случаях невозможно вообще. Вследствие описанных проблем для восстановления параметров объёмного рассеяния часто используются измерения углового рассеяния (отражения и/или пропускания) ОР образца, а другими словами, двунаправленная функция рассеяния (ДФР).

Задачи восстановления параметров объёмного рассеяния можно разделить на две группы (см. рис. 1). Первый тип — ДФР (см. рис. 1А), применим, когда толщиной (t) элемента с ОР можно пренебречь. Образец с ОР освещают с разных сторон и измеряют угловое распределение рассеянного света, в результате чего получают

ДФО образца [1–4]. Эта ДФР используется для восстановления рассеивающих свойств тонкой плёнки с ОР, смоделированного как однослойная поверхность нулевой толщины. Подробно метод восстановления ОР свойств описан в [5].

Когда требуется обеспечить точное распространение света в среде с ОР, например, в световодах, необходимо учитывать как угловые, так и пространственные преобразования света. Различные модели ОР требуют определения набора параметров (рис. 1Б). ДФР образца ОР измеряется и используется для восстановления параметров ОР, применимых к любой форме элемента ОР в оптической системе. Определение параметров ОР затруднено из-за высокой чувствительности параметров ОР к рассеянию на выходе, особенно в средах с малым рассеянием. Эксперименты показывают, что разные наборы параметров ОР могут приводить к одинаковым ДФР. В статье также исследуется, как эта неоднозначность влияет на рассеяние объектов ОР, имеющих разные размеры/форму.

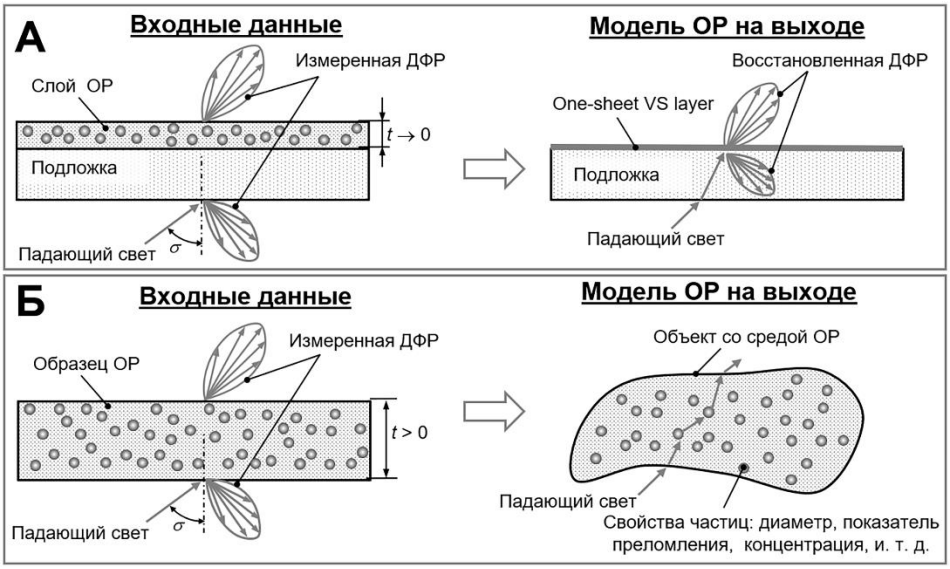


Рис. 1. Виды моделирования объёмного рассеяния: А – моделирование среды ОР с помощью ДФР, Б – моделирование среды ОР с помощью параметрической модели.

II. Модель объёмного рассеяния

Схематически распространение света в неоднородной среде показано на рис. 2. Обычно среда с объёмным рассеянием состоит из двух и более диэлектриков с разными показателями преломления: связующего (наполнителя) и рассеивающих агентов – микрочастиц. Свет распространяется внутри наполнителя и рассеивается на частицах.

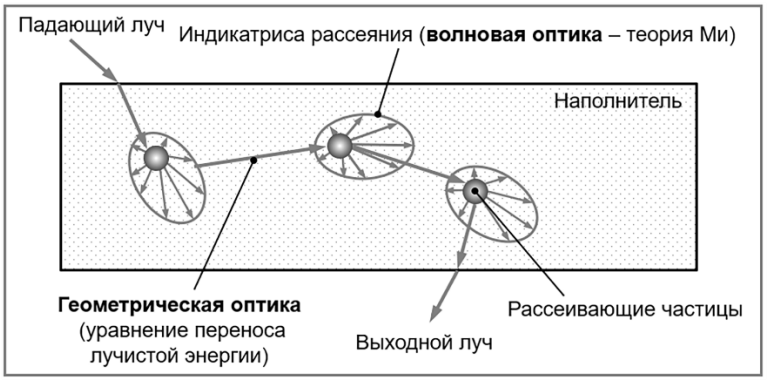


Рис. 2. Распространение света в неоднородной среде

Свет, попадающий в наполнитель, сталкивается с микрочастицами, что приводит к его частичному поглощению и изменению направления распространения. Это явление широко изучалось, формулы представлены во многих публикациях, например в [6–9]. При высоких концентрациях микрочастиц для точного моделирования рассеяния света требуется волновое решение, которое практически не используется из-за

сложности. Однако для моделирования ОР с малым светорассеянием (концентрацией частиц не более 3–5%) применяют гибридные приближенные подходы.

В гибридных подходах распространение света между частицами определяется путём решения уравнения переноса лучистой энергии согласно законам геометрической оптики, а рассеяние света на микрочастицах определяется с помощью индикатрисы рассеяния. Уравнение переноса может быть решено различными методами, как аналитическими [10-12], так и численными [13]. В данном исследовании выбран численный стохастический подход (Монте-Карло), который, хотя и менее эффективен, чем аналитические методы, но более гибок и доступен в широко используемом оптическом программном обеспечении, таком как Lumicept [14].

Разработано два основных подхода к расчёту индикатрисы рассеяния микрочастиц: теория Ми и модель Хеньи-Гринштейна. Теория Ми, основанная на волновой оптике и приближении формы частиц к сферической, является наиболее физически точной моделью и поддерживается большинством световых симуляторов. В этом исследовании используется теория Ми [15-18]. На рис. 3 представлены параметры, описывающие среду ОР с микрочастицами на основе теории Ми.

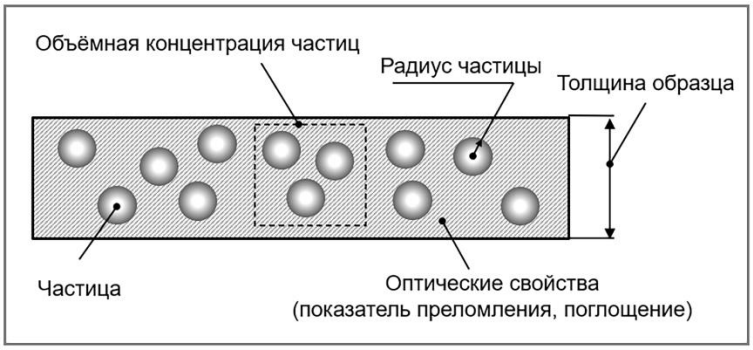


Рис. 3. Параметры среды с объёмным рассеянием

В таблице 1 представлен общий набор основных параметров объёмного рассеяния наполнителя и микрочастиц, описывающих светорассеяние образца ОР в соответствии с теорией Ми.

ТАБЛИЦА 1
ПАРАМЕТРЫ СРЕДЫ ОБЪЁМНОГО РАССЕЯНИЯ

1.	**№**	**2.**	**Элемент ОР**	**3.**	**Parameter**
4.	1	**5.**	**Образец**	6.	Толщина
7.	2	**8.**	**Наполнитель**	9.	Показатель преломления
10.	3			11.	Поглощение
12.	4	**13.**	**Рассеивающие частицы**	14.	Объёмная концентрация частиц (PVC)
15.	5			16.	Радиус
17.	6			18.	Показатель преломления
19.	7			20.	Поглощение (мнимая часть показателя преломления)

Следует отметить, что в таблице 1 приведено довольно большое количество параметров. Использование всего набора параметров значительно усложняет процесс восстановления, поэтому предпочтительно исключить из процесса восстановления как можно большее число параметров. Некоторые из этих параметров, такие как толщина образца и показатель преломления наполнителя, должны быть известны, а поглощение может быть близким к нулю и обычно не учитываться. В этом исследовании при восстановлении параметров ОР используются три основных параметра: объёмная концентрация частиц – PVC (Particle Volume Concentration), радиус частиц (R) и их показатель преломления (n).

III. Метод восстановления параметров объёмного рассеяния

Выбранный в работе алгоритм восстановления параметров ОР аналогичен подходу, предложенному в [5], где восстанавливается ДФО тонких плёнок с ОР. То есть, измеренные ДФР образца с ОР используется в качестве целевой функции. Затем аналог этой функции рассчитывается на основе некоторой исходной модели ОР. Параметры ОР модифицируются с помощью внешнего оптимизатора, и процедура циклически повторяется до достижения минимальной разницы между измеренным (целевым) ДФР и ДФР, рассчитанным на основе искусственной модели образца с ОР.

Однако, как уже упоминалось, задача восстановления параметров ОР существенно сложнее по сравнению с восстановлением ДФР для тонкой плёнки ОР. Таким образом, процедура оптимизации параметров ОР включает в себя комплекс усовершенствований, которые будут подробно рассмотрены ниже. Блок-схема метода реконструкции представлена на рис. 4. Отметим, что данный подход восстановления параметров ОР реализован с помощью скрипта Python и интегрированного в Lumicept [14].

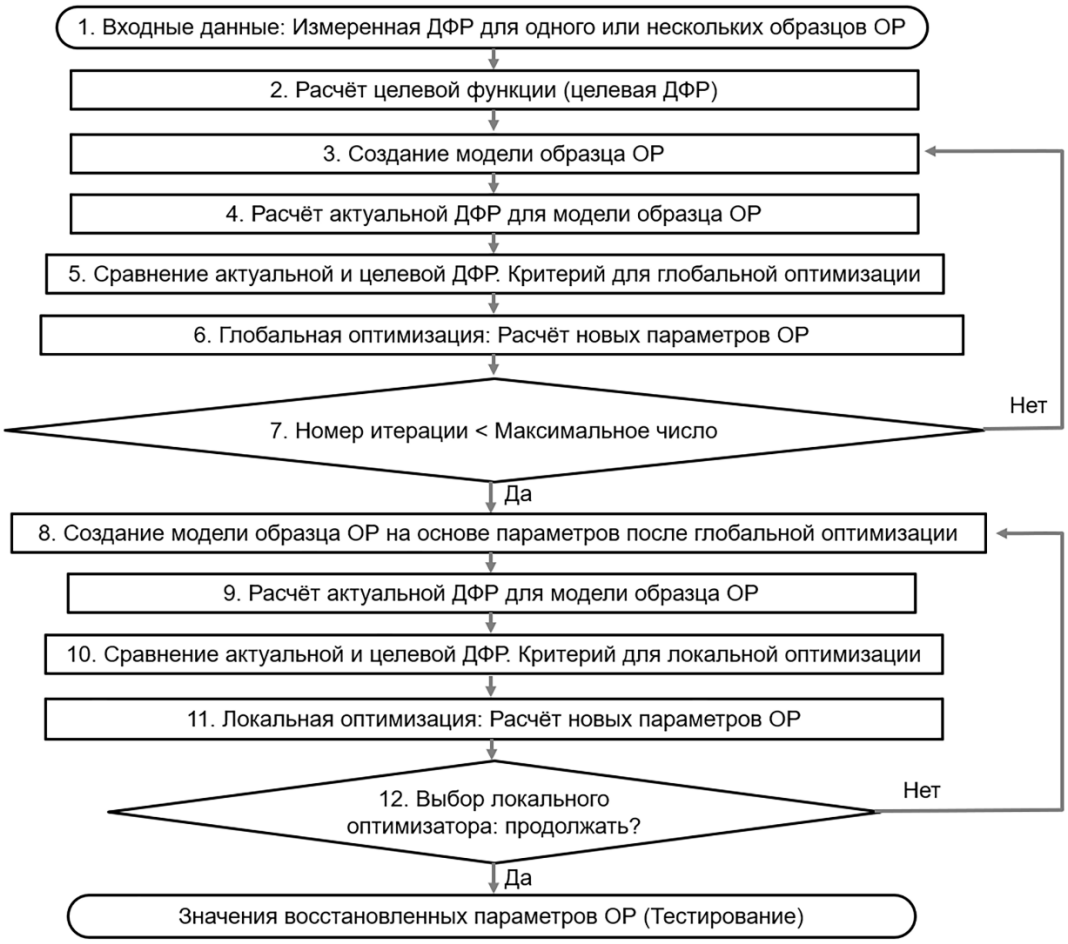


Рис. 4. Схема восстановления параметров объёмного рассеяния

Основной проблемой при восстановлении параметров ОС является неоднозначность, поскольку различные комбинации значений параметров ОР могут давать ДФР близкие к измеренным. Это указывает на то, что целевая функция оптимизации немонотонна и может иметь много локальных минимумов. Кроме того, критерий оптимизации очень чувствителен к значениям параметров ОР.

Для решения этих проблем подход к восстановлению параметров ОР включает анализ нескольких ДФР, измеренных для образцов ОР различной толщины, и разделение процедуры оптимизации на глобальный и локальный этапы, что повышает значительно качество восстановления параметров ОР. Процесс восстановления можно описать следующими шагами (см. также рис. 4):

1. Ввод исходных данных: измеренные ДФР для одного или нескольких образцов, данные для построения модели образца ОР (толщина образца, исходные параметры среды ОР) и технические параметры для процесса восстановления (список параметров ОР, тип оптимизации и другие параметры, необходимые для старта процесса восстановления);

2. Расчёт целевой функции: Измеренная ДФР, как правило, является сложной функцией с детальным разрешением, что делает ее неудобной для прямого использования в качестве целевой функции. Вместо этого упрощённая целевая функция рассчитывается путём назначения измеренной ДФР к однослойной поверхности, освещения ее параллельным источником света и расчёта углового распределения прошедшего/отражённого света с использованием приёмников света, поддерживаемого программным обеспечением, в котором осуществляется процесс восстановления (Lumicept [14]).

3. Построение модели образца ОР: Модель образца ОР строится с использованием начальных параметров рассеивающей объёмной среды, используемых в теории Ми;

4. Расчёт углового пропускания/отражения. Этот шаг включает расчёт углового пропускания/отражения для модели образца ОР при тех же условиях освещения и наблюдения, которые использовались для расчёта целевой функции;

5. Угловое распределение света, рассчитанное для модели образца ОР сравнивается с целевой функцией, и вычисляется критерий оптимизации. Используются один из критериев: абсолютное среднеквадратичное отклонение *RMSD*, см. формулу (1) или относительное $RMSD_{\text{relative}}$, см. формулу (2):

$$RMSD = \sqrt{\frac{\sum_i^n (I_{ei} - I_{si})^2}{n}} \tag{1}$$

$$RMSD_{\text{relative}} = \sqrt{\frac{\sum_i^n \left(\frac{I_{ei} - I_{si}}{0.5 \cdot (I_{ei} + I_{si})}\right)^2}{n}} \cdot 100\% \tag{2}$$

где: I_{ei} — ожидаемое значение силы света для *i* направления наблюдения и длины волны (берётся из целевой функции); I_{si} — рассчитанное значение силы света для *i* направления наблюдения и длины волны, рассчитанное на основе модели образца; *n* — общее число направлений наблюдения и длин волн.

Выбор абсолютного или относительного *RMSD* в качестве критерия оптимизации важен в случае, если целевая функция очень неравномерна. Абсолютный критерий *RMSD* предпочтительнее использовать в случае, если важно сделать акцент на восстановлении углового рассеяния в пиках, в зонах, близких к зеркальным направлениям рассеянного света, в то время как $RMSD_{\text{relative}}$ более предпочтителен в случае, если важны также зоны диффузного рассеяния. Однако в случаях сред ОР с малым рассеянием даже использование $RMSD_{\text{relative}}$ недостаточно. Для таких образцов с очень малым рассеянием мы предлагаем использовать данные в логарифмическом масштабе: $\ln(I_{ei})$ и $\ln(I_{si})$ вместо I_{ei} и I_{si} в формулах (1) и (2). В крайних случаях, когда образцы ОР близки к чисто прозрачным, предпочтительнее исключить все направления наблюдения, близкие к зеркальным, для определения критерия, см. рис. 5.

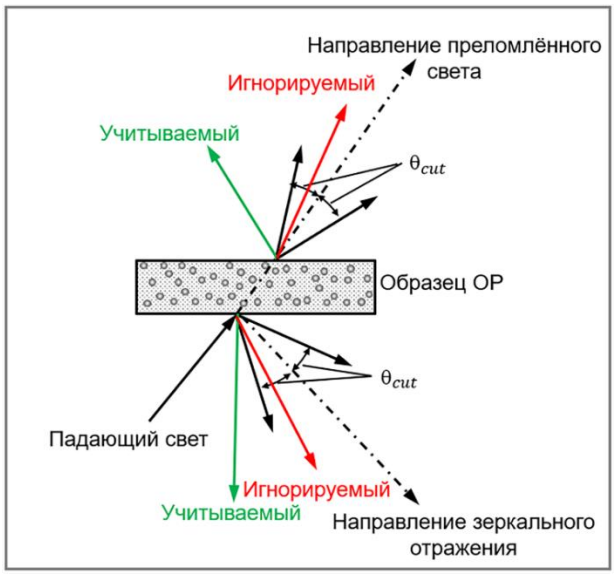


Рис. 5. Исключение зеркальных направлений при определении критерия оптимизации

6. Глобальная оптимизация запускается на следующем шаге с использованием подхода «brute force» [19] (в исследовании используется функция scipy.optimize.brute из SciPy). На этом шаге вычисляется значение критерия оптимизации в каждой точке многомерной сетки всех параметров оптимизации с целью найти глобальный минимум. Каждый интервал возможных значений параметров подразделяется на сетку с шагом =

(*max* - *min*) / *N*s, где *min* и *max* — минимально и максимально возможные значения параметров, а *N*s — количество точек сетки, одинаковое для каждого параметра ОР. Этот подход, использующий перебор дискретных значений параметров достаточно прост, но неэффективен из-за большого количества итераций Ns^{Pn}, где *P*n — количество оптимизируемых параметров. Шаги 4–6 алгоритма (см. рис. 4) повторяются до тех пор, пока не будут рассчитаны все комбинации параметров (см. 7 на рис. 4), и оптимальный набор параметров не будет передан в качестве исходных для локальной оптимизации.

7. После завершения всех итераций глобальной оптимизации определяются параметры объёмной среды, для которых актуальная ДФР (модели образца ОР) наиболее близка к целевой (измеренной) ДФР.

8. Далее строится новая модель образца на основе параметров объёмного рассеяния, полученных после глобальной оптимизации и. все последующие шаги подобны шагам, выполняемым при глобальной оптимизации:

9. Производиться расчёт ДФР для новой модели образца ОР.

10. Далее определяется значение критерия оптимизации на основе актуальной (рассчитанной) и целевой ДФР

11. Текущие значения оптимизируемых параметров и критерий оптимизации (RMSD) передаются в функцию локального оптимизатора. В качестве локального оптимизатора используется «симплекс» метод [20] (функция scipy.optimize.fmin из SciPy).

12. Локальный оптимизатор делает выбор о завершении или продолжении процесса оптимизации. В случае продолжения оптимизации производиться расчёт новых параметров ОР для следующей итерации и происходит повторение шагов 8–12. В случае завершения оптимизации, при выполнении максимального количества итераций или достижения цели оптимизации (минимального порога RMSD) определяются окончательные значения оптимизируемых параметров ОР.

На финальном шаге процесса восстановления выполняется тест: рассчитывается угловое светорассеяние для наилучшего набора параметров ОР, полученного в результате двух этапов оптимизации.

IV. Описание образцов с объёмным рассеянием

Для надёжной проверки подхода к восстановлению параметров ОР, описанного в предыдущих главах, нам нужны образцы с точно заданными параметрами ОР и широким спектром светорассеивающих свойств. Для этой цели в исследовании использовались три образца с различными рассеивающими свойствами. Их визуализация представлена на рис. 6.

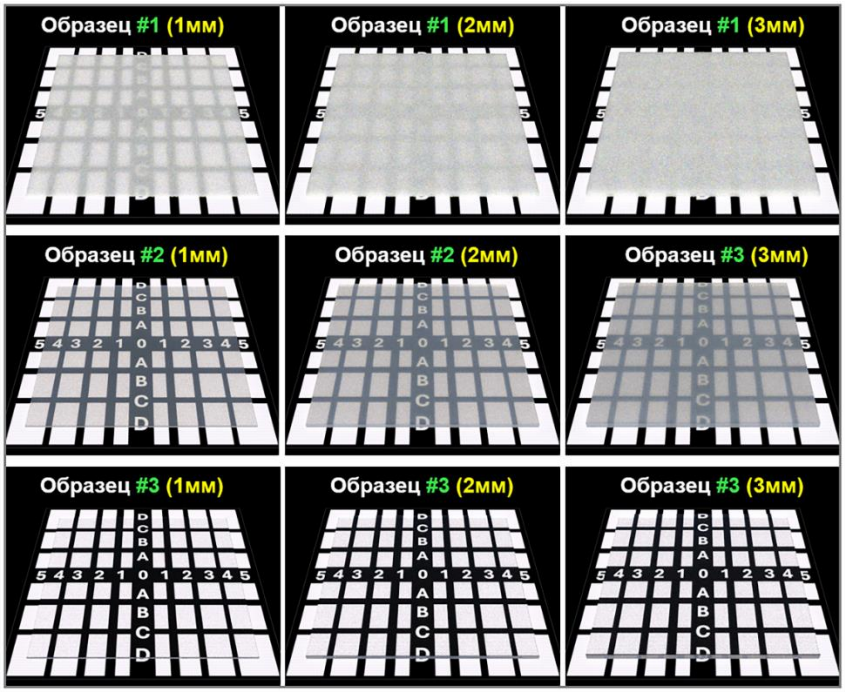


Рис. 6. Визуализация образцов с объёмным рассеянием выбранных для исследования

Изображения образцов ОР (см. рис 6.) были синтезированы с помощью коммерческого программного обеспечения Lumicept [14], обладающего передовым и физически корректным рендерингом «Path Tracing» [21]. Сцена для синтеза изображений состояла из пластины со средой ОР, помещённой над текстурированной плоскостью и освещённой параллельным светом. Каждый образец моделировался для трёх толщин: 1, 2 и 3 мм (отмечены жёлтым на рис. 6). Среда ОР состоит из наполнителя и сферических рассеивающих агентов. Оптические свойства наполнителя одинаковы для всех трех образцов, с показателем преломления 1.5 для всех длин волн. Это упрощение используется для всех параметров объёмного рассеяния. В принципе, можно также учитывать спектральную зависимость, как описано в [5], но в данном исследовании она не рассматривалась. Параметры рассеивающих агентов, которые должны быть определены во время процедуры реконструкции, представлены в таблице 2.

ТАБЛИЦА 2
ПАРАМЕТРЫ РАССЕИВАЮЩИХ ЧАСТИЦ ДЛЯ ОБРАЗЦОВ, ИСПОЛЬЗУЕМЫХ В АПРОБАЦИИ

21. Параметр	**22. Образец #1**	**23. Образец #2**	**24. Образец #3**
25. Объёмная концентрация частиц (PVC)	26. 1%	27. 0.1%	28. 0.01%
29. Радиус (R)	30. 100нм	31. 150 нм	32. 200 нм
33. Показатель преломления (n)	34. 1.7	35. 1.65	36. 1.55

Доля диффузного рассеивания света уменьшается от образца #1 к образцу #3. Это достигается за счёт снижения концентрации рассеивающих частиц (объёмной концентрации частиц - PVC) и разницы в показателях преломления между наполнителем и частицами. Согласно рис. 6, образец #1 сильно диффузный, напоминающий «молочное стекло», тогда как образец #3 практически прозрачен. Поглощение частицами в модели игнорируется, но наполнитель имеет небольшое значение поглощения около 2% на 10 см, но этот параметр не является параметром оптимизации.

V. Результаты восстановления параметров ОР

Для всех трех образцов, описанных в главе 4, были рассчитаны ДФР. Для повышения реалистичности восстановления использовалась точная модель реального измерительного прибора GCSM-4 [22]. ДФР были рассчитаны для всех трех толщин образцов и для обоих световых компонент: отражённой (ДФО) и преломлённой (ДФП). Таблица 3 представляет ограничения параметров, используемые в процессе реконструкции.

ТАБЛИЦА 3
ОГРАНИЧЕНИЯ ДЛЯ ПАРАМЕТРОВ ОР

37. Параметр	**38. Образец #1**		**39. Образец #2**		**40. Образец #3**	
	41. Min	**42. Max**	**43. min**	**44. max**	**45. min**	**46. max**
47. Объёмная концентр. (PVC)	48. 0.01 %	49. 5%	50. 0.01 %	51. 5%	52. 0.001 %	53. 0.1%
54. Радиус (R)	55. 10нм	56. 500нм	57. 10нм	58. 500нм	59. 10нм	60. 500нм
61. Показатель преломления (n)	62. 1.4	63. 2.2	64. 1.4	65. 2.2	66. 1.4	67. 2.2

Предполагается, что информация о свойствах рассеивающих агентов отсутствует, поэтому в качестве входных ограничений для первого уровня реконструкции — глобальной оптимизации — задаются очень широкие интервалы значений параметров ОР.

Результаты восстановления параметров ОР для образца #1

Первоначально было рассмотрено восстановление параметров ОР на основе одного образца #1 для трех вариантов толщины: 1, 2 и 3 мм. Свойства наполнителя исключены из процесса оптимизации (показатель преломления = 1.5). Граничные ограничения параметров ОР были определены в соответствии с таблицей 3. Для глобальной оптимизации использовался метод «brute force» (функция SciPy "brute") с разбивкой Ns = 6, а для локальной оптимизации применялся «симплекс» метод (функция SciPy "fmin") количество итераций = 30. Таким образом, общее количество шагов в процессе реконструкции составляет приблизительно 250 (см. главу III). Основные условия освещения были ограничены σ = 0° (см. σ на рис. 1). Это ограничение обусловлено несколькими причинами. Во-первых, из опыта измерения ДФР, чем меньше угол наклона освещения, тем более точны измерения. Кроме того, освещение под нормальным направлением приводит к осевой симметрии рассеянного света, что позволяет использовать более грубые параметры углового разрешения в процессе восстановления параметров ОР, тем самым ускоряя весь процесс. Исследования показывают, что форма рассеяния света для различных направлений освещения для образцов со средами ОР со сферическими частицами взаимно связана. Другими словами, лучшее согласование ожидаемого углового рассеяния света с достигнутым в процессе восстановления при одном условии освещении, приводит к лучшему согласию и для освещения при других углах. Это будет продемонстрировано ниже. Основные результаты реконструкции для образца #1 представлены в таблице 4.

ТАБЛИЦА 4
РЕЗУЛЬТАТЫ ВОССТАНОВЛЕНИЯ ПАРАМЕТРОВ ОР ДЛЯ ОБРАЗЦА #1

68. №	**69. Параметры для восстановления**	**70. Восстановленные параметры ОР**			**71. Отклонение ДФР** $RMSD_{relative}$ (%)		
		72. PVC (%)	**73. R (нм)**	**74. n**	**75. 1мм**	**76. 2мм**	**77. 3мм**
78. 0	79. Ожидаемые параметры	80. 1.000	81. 100.0	82. 1.700	83. -	84. -	85. -
86. 1	87. ДФП, t = 1мм, σ = 0°	88. 4.226	89. 96.5	90. 1.400	91. 0.971	92. 1.27	93. 1.35
94. 2	95. ДФП, t = 2мм, σ = 0°	96. 4.136	97. 103.5	98. 1.402	99. 1.44	100. 1.00	101. 1.12
102. 3	103. ДФП, t = 3мм, σ = 0°	104. 3.936	105. 111.6	106. 1.400	107. 1.89	108. 1.65	109. 1.26
110. 4	111. ДФП, t = 1мм, σ = 0, 30, 60°	112. 2.963	113. 102.1	114. 1.386	115. 1.10	116. 1.71	117. 1.37
118. 5	119. ДФП + ДФО, t = 1мм, σ = 0°	120. 3.005	121. 101.3	122. 1.384	123. 1.58	124. 1.51	125. 1.25
126. 6	127. ДФП, t = 1, 2мм, σ = 0°	128. 0.530	129. 97.1	130. 1.76	131. 1.00	132. 1.22	133. 1.21
134. 7	135. ДФП, t = 1, 3мм, σ = 0°	136. 1.865	137. 94.9	138. 1.651	139. 1.08	140. 0.95	141. 0.91
142. 8	143. ДФП, t = 2, 3мм, σ = 0°	144. 0.539	145. 105.2	146. 1.769	147. 1.21	148. 1.20	149. 0.73
150. 9	151. ДФП, t = 1, 2, 3мм, σ = 0°	152. 4.464	153. 98.5	154. 1.403	155. 1.24	156. 0.97	157. 1.01

Примечания к таблице 4: Строка с нулевым номером представляет ожидаемые (истинные) значения параметров объёмного рассеяния образца #1 (взяты из таблицы 2) Во втором столбце «Параметры для восстановления» перечислены ключевые условия восстановления параметров ОР, указывающие целевую функцию оптимизации: ДФП, используемую для описания углового пропускания и ДФО - для углового

отражения. «t» указывает толщину образца, а σ — направление освещения. В следующих трех столбцах представлены восстановленные параметры ОР, в последних трёх колонках даны отклонения между рассчитанными и ожидаемыми ДФР для всех толщин образца в формате $RMSD_{\text{relative}}$, см. формулу (2). Строки 1–3 иллюстрируют реконструкцию с использованием в качестве целевой функции углового пропускания рассеянного света при нормальном освещении (σ = 0°) для образца #1 с различной толщиной. Строка 4 представляет вариант восстановления, где в целевую функцию добавлены наклонные направления освещения под углами σ = 30° и 60°. Строка 5 представляет результаты восстановления с целевой функцией, использующей как угловое пропускание (ДФП), так и отражение (ДФО). Последующие строки используют много конфигурационный подход, где в процессе восстановления используются ДФП от нескольких толщин образца #1. Это все возможные комбинации двух толщин (1мм, 2мм); (1мм, 3мм); (2мм, 3мм) и в случае строки 9 все три толщины образца используются. Разница между ожидаемым и фактическим угловым рассеянием после реконструкции минимальна, с относительным среднеквадратичным отклонением 1–2%. Следовательно, графики в Таблице 4 аналогичны, поэтому представлены только выборочные примеры. На рис. 7 показаны результаты реконструкции для образца #1 толщиной 2 мм, соответствующие второй строке в таблице 4.

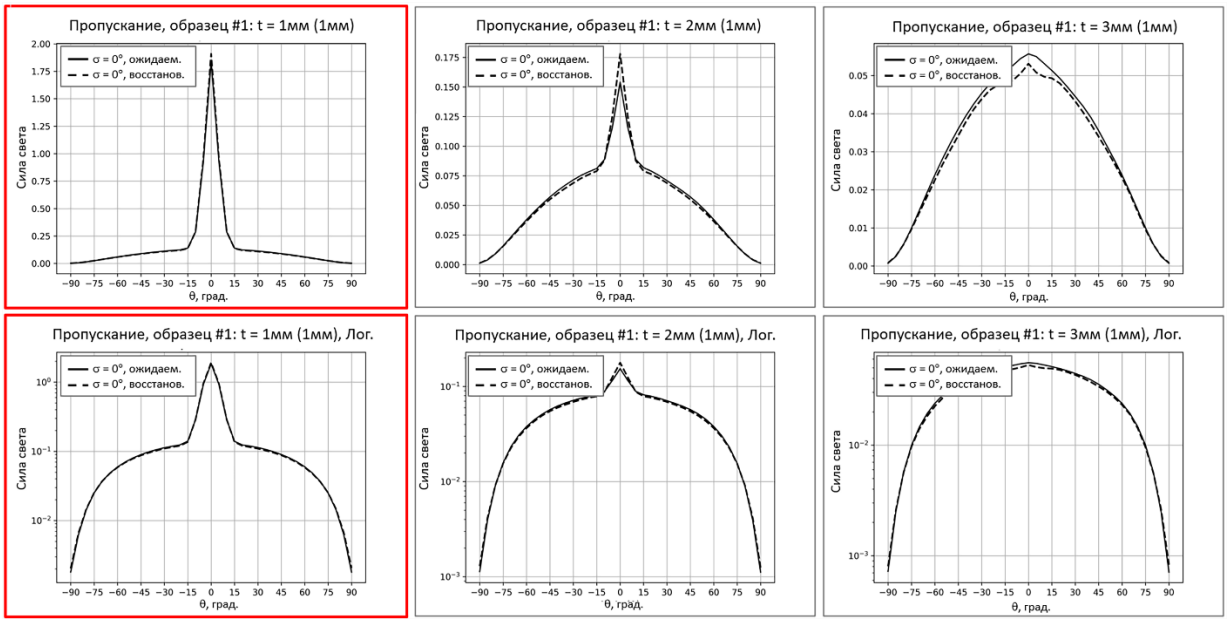


Рис. 7. Результаты восстановления параметров ОР для образца #1 с толщиной = 2 мм

Рис. 7: содержит шесть графиков. Каждый столбец представляет определённую толщину образца: 1, 2 и 3 мм (слева направо). Графики в первой строке показывают угловое распределение силы света, усреднённое для всех длин волн в обычной линейной шкале. Во второй строке показаны аналогичные данные, но в логарифмической шкале. Графики, отмеченные сплошными линиями, представляют собой ожидаемое угловое распределение пропущенного света («целевые функции» в оптимизации). Графики, отмеченные пунктирными линиями, представляют собой достигнутое угловое пропускание. В заголовках графиков толщина образца, используемая в процессе восстановления параметров ОР, указана в скобках, а сами графики для этой толщины отмечены красным контуром. На рис. 7 реконструкция была выполнена для образца #1 толщиной 2 мм. Близость ожидаемого и восстановленного распределения для 1 и 3 мм показывает превосходную масштабируемость параметров ОР.

На рис. 8 представлен ещё один вариант восстановления, где используются три направления освещения, см. строку 4 в таблице 4. Здесь реконструкция параметров ОР выполнена для толщины 1 мм.

На левых графиках рис.8 представлено угловое пропускание в линейном масштабе, а на правых — в логарифмическом. Из этих графиков и других исследований очевидно, что усложнение критерия оптимизации дополнительными направлениями освещения не улучшает результат реконструкции.

На рис. 9 представлен другой вариант исследования, соответствующий строке 5 в таблице 4. Здесь к критерию оптимизации добавлен компонент отражения.

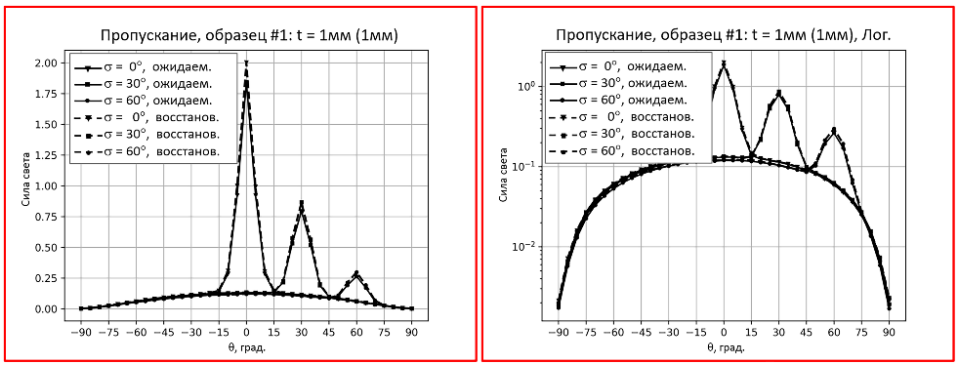


Рис. 8. Результаты восстановления параметров ОР для образца #1 с толщиной = 1 мм. Целевая функция расширена углами освещения σ = 0, 30 и 60°

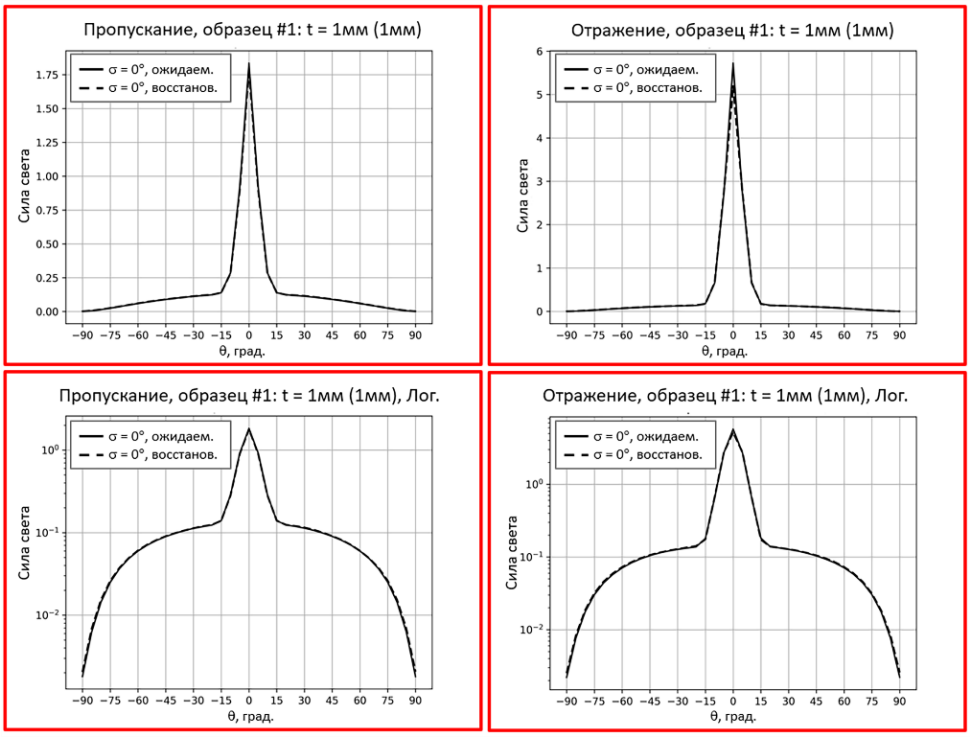


Рис. 9. Результаты восстановления параметров ОР для образца #1 толщиной 1 мм. Целевая функция расширена отражательной компонентой

Левый столбец на рис. 9 представляет угловое пропускание, а правый столбец представляет отражение. Первая строка показывает данные в линейном масштабе, а вторая строка - в логарифмическом масштабе. Исследования показывают, что пропускание и отражение в средах ОР тесно связаны. Чем меньше разница между желаемым и достигнутым пропусканием, тем меньше разница в отражении. Таким образом, добавление отражения (ДФО) к целевой функции не приводит к заметным улучшениям в реконструкции, только к замедлению.

Другие варианты реконструкции параметров ОР, перечисленные в таблице 4, представляющие различные комбинации толщины образца, дают аналогичные результаты, полученные для более простого восстановления параметров ОР на основе одной толщины образца. Сравнение ожидаемых параметров ОР, представленных в таблице 2 с достигнутыми параметрами (см. таблицу 4) выявляет существенные различия, несмотря на превосходные результаты в достигнутом угловом рассеянии. Однако согласие между угловым распределением света образцов с разной толщиной очень хорошее, что указывает на хорошую масштабируемость восстановленных параметров ОР, даже если они отличаются от своих истинных значений.

Результаты восстановления параметров ОР для образца #2

Результаты восстановления, достигнутые для образца #2 очень похожи на результаты образца #1. Поэтому мы ограничили презентацию двумя вариантами, как показано в Таблице 5. Первый вариант показывает

реконструкцию на основе одного образца толщиной 2 мм, тогда как второй вариант использует все три толщины образцов: 1 мм, 2 мм и 3 мм.

ТАБЛИЦА 5
РЕЗУЛЬТАТЫ ВОССТАНОВЛЕНИЯ ПАРАМЕТРОВ ОР ДЛЯ ОБРАЗЦА #2

158. №	**159. Параметры для восстановления**	**160. Восстановленные параметры ОР**			**161. Отклонение ДФР** $RMSD_{relative}$ (%)		
		162. PVC (%)	**163. R (нм)**	**164. n**	**165. 1мм**	**166. 2мм**	**167. 3мм**
168. 0	169. Ожидаемые параметры	170. 0.100	171. 150.0	172. 1.600	173. -	174. -	175. -
176. 1	177. ДФП, t = 2мм, σ = 0°	178. 0.008	179. 141.4	180. 1.967	181. 1.34	182. 1.17	183. 1.39.
184. 2	185. ДФП, t = 1, 2, 3мм, σ = 0°	186. 0.010	187. 131.5	188. 1.937	189. 1.72	190. 1.50	191. 1.66

Графики с результатами восстановления представлены только для одного варианта (см. рис. 10). Он аналогичен рис. 7, но данные представлены только в логарифмическом масштабе. Образец #2 имеет существенно меньшую диффузность по сравнению с образцом #1, что приводит к большим зеркальными пикам в угловом распределении, что делает их неинформативными в обычном линейном масштабе.

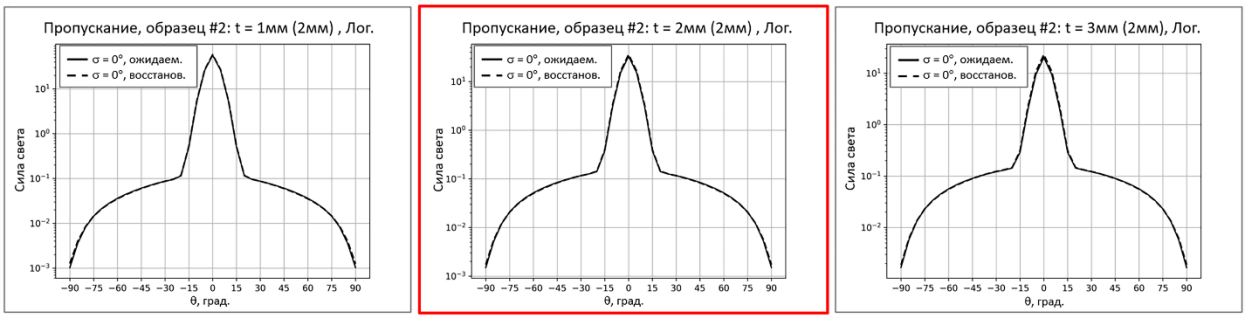


Рис. 10. Результаты восстановления параметров ОР для образца #2. Восстановление выполнено для толщины образца = 2 мм

На рис. 10 график с толщиной образца, используемой в процессе восстановления (в данном случае 2 мм), отмечен красным контуром. Анализируя данные в таблице 5 и на рис. 10, можно сделать вывод, аналогичный выводу для образца #1: в процессе восстановления достигается очень хорошее согласование в ДФП для всех толщин образца, что доказывает хорошую масштабируемость восстановленных параметров ОР. Однако достигнутые значения параметров ОР далеки от истинных.

Результаты восстановления параметров ОР для образца #3

Образец #3 очень прозрачный. Это усложняет процесс реконструкции параметров ОР из-за огромных пиков в ДФП, используемых в качестве целевых функций. Чтобы избежать проблемы с паразитным пиком пропускания, направления вблизи зеркальных углов исключаются из расчёта критерия оптимизации (в данном исследовании $\theta_{cut} = 10°$). Несмотря на высокую прозрачность образца #3, общий подход к восстановлению параметров ОР очень похож на подходы, применяемые для образцов #1, #2 с большей диффузностью. Таким образом, результаты восстановления для образца #3 представлены в форме, аналогичной образцу #2, см. таблицу 6.

ТАБЛИЦА 6
РЕЗУЛЬТАТЫ ВОССТАНОВЛЕНИЯ ПАРАМЕТРОВ ОР ДЛЯ ОБРАЗЦА #3

192. №	**193. Параметры для восстановления**	**194. Восстановленные параметры ОР**			**195. Отклонение ДФР** $RMSD_{relative}$ (%)		
		196. PVC (%)	**197. R (нм)**	**198. n**	**199. 1мм**	**200. 2мм**	**201. 3мм**

202. 0	203. Ожидаемые параметры	204. 0.01	205. 200.0	206. 1.55	207. -	208. -	209. -
210. 1	211. ДФП, t = 2мм, σ = 0°	212. 0.0008	213. 192.2	214. 1.652	215. 2.23	216. 2.41	217. 2.08
218. 2	219. ДФП, t = 1, 2, 3мм, σ = 0°	220. 0.0011	221. 188.9	222. 1.632	223. 1.77	224. 2.17	225. 2.17

Графики на рис. 11 сравнивают ожидаемое и достигнутое угловое пропускание для всех трех толщин образцов (1 мм, 2 мм и 3 мм). Как и в случае с образцом #2, толщина образца, использованная в процессе восстановления составляла 2 мм.

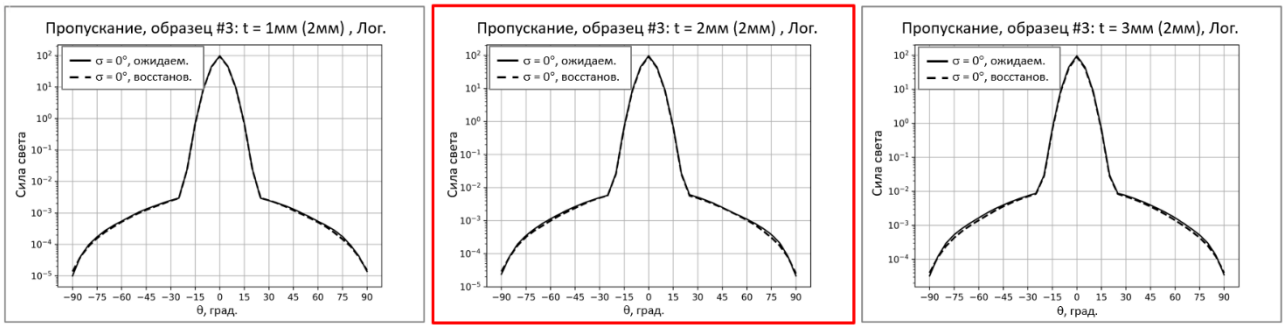


Рис. 11. Результаты восстановления параметров ОР для образца #3. Восстановление выполнено для толщины образца = 2 мм

После реконструкции образца #3 можно сделать выводы, аналогичные сделанным для более диффузных образцов #1 и #2. Наблюдается положительный результат в достижении ожидаемой формы углового рассеяния света (отличное соответствие целевой функции), но значения параметров существенно отличаются от реальных, за исключением радиуса рассеивающих частиц.

VI. Проверка результатов восстановления образцов ОР на основе визуализации образцов

Восстановление параметров ОР для всех трех образцов показала, что практически невозможно восстановить истинные параметры ОР на основе данных ДФР, используемых в качестве целевой функции. Однако главной целью этой реконструкции являются не сами параметры ОР, а возможность использования этих параметров в моделировании. Это касается масштабируемости этих параметров — их влияния на результаты оптического моделирования и визуализации сцен, содержащих элементы со средами ОР и имеющими различные формы и размеры. Для решения этой проблемы было проведено несколько тестов визуализации образцов.

Первый тест представляет визуализацию диффузной пластины с использованием модели, представленной на рис. 6. Пластина с ОР помещается над текстурированной плоскостью и освещается параллельным светом. Сравнительные изображения представлены на рис. 12. Эти изображения, как и на рис. 6, созданы с помощью рендерера «Path Tracing» программного пакета Lumicept [14]. Для лучшей визуализации вид камеры изменён на вид «сверху» — направление наблюдения проходит вдоль нормали к верхней/нижней граням пластины ОР. Для всех образцов была использована одна и таже толщина 2 мм. Учитывая плоскостную симметрию модели, изображения пластин ОР разделены на две части. Левая часть изображения подготовлена для истинных параметров ОР, взятых из таблицы 2. Правая часть пластины создана для восстановленных параметров ОР, взятых из таблиц 4 (вариант 3) для образца #1, таблицы 5 (вариант 1) для образца #2 и таблицы 6 (вариант 1) для образца #3. Обратите внимание, что изображения, полученные для истинных параметров ОР и восстановленных параметров ОР, очень близки, что вполне разумно. Этот результат ожидается на основе хорошего соответствия ДФР рассчитанного для образца толщиной 2 мм, см. графики представленные на рис. 7, 10 и 11. Изображения для других исследованных толщин образцов очень похожи и поэтому не включены в документ.

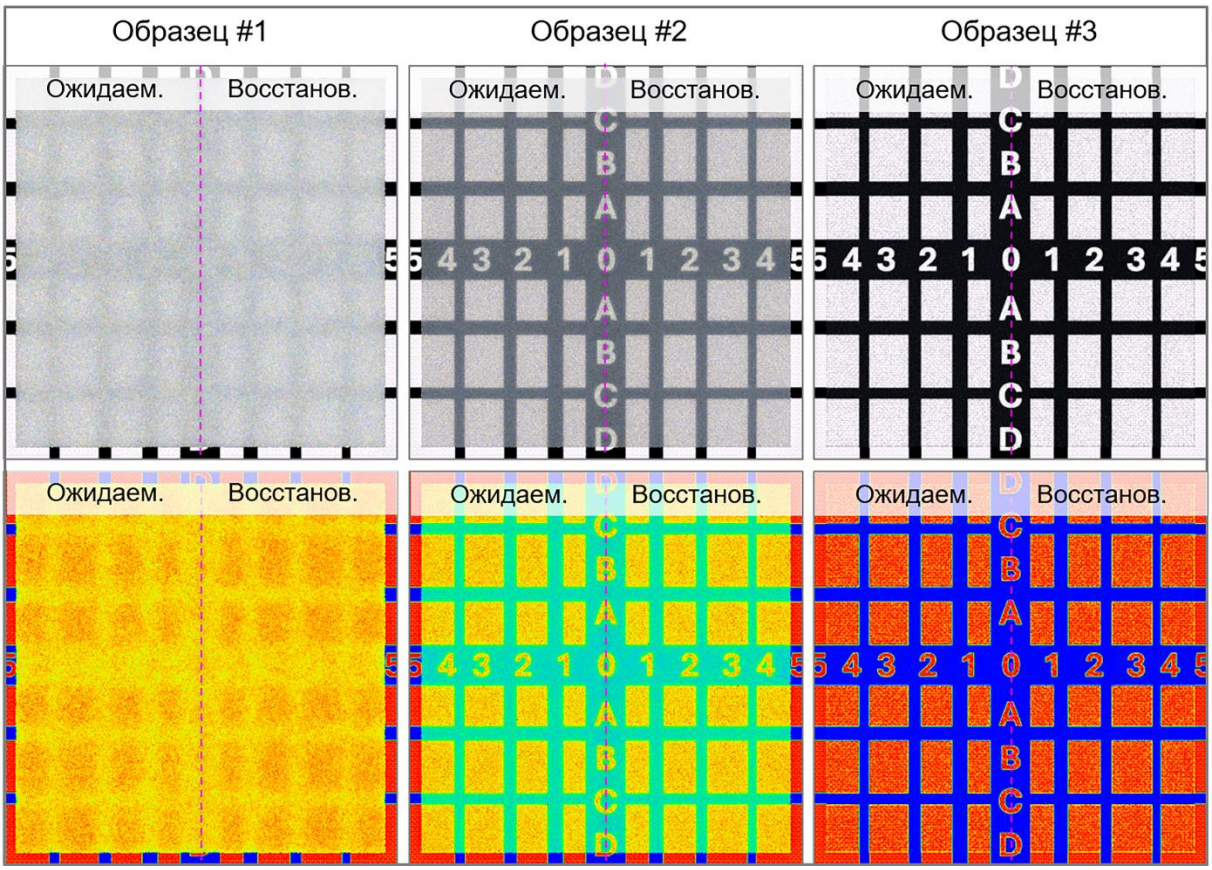


Рис. 12. Визуализация пластин с ОР для образцов #1, #2, #3. Толщина пластин = 2 мм.

На рис. 12 левые и правые части изображений представляют визуализацию пластин с истинными и восстановленными параметрами ОР. Они разделены пунктирной пурпурной линией для лучшей визуализации. Первая строка представляет изображения в формате RGB, а вторая строка - в искусственных цветах.

Очевидно, что изображения пластины ОР, подготовленные для истинных и восстановленных параметров ОР, должны быть достаточно близки, что подтверждается близостью их ДФР. Относительное среднеквадратичное отклонение ($RMSD_{\text{relative}}$) между изображениями (распределениями яркости), синтезированными для истинных параметров ОР и восстановленных не превышает 1%. Однако более интересно проверить масштабируемость параметров ОР на элементах с большими размерами, например, в случае применения ОР в световодах. Для этой цели была подготовлена ещё одна вторая модель. Схема этой модели представлена на рис. 13.

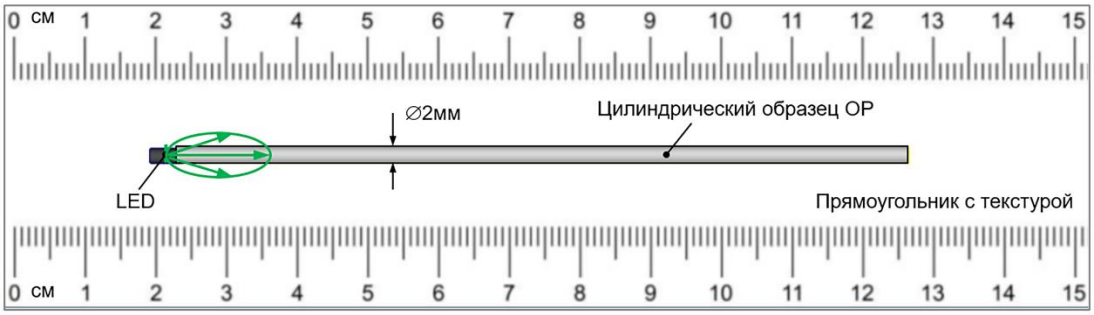


Рис. 13. Схема визуализации «световода» с параметрами ОР образца #3

Схема второй модели состоит из трубки, сделанной из материала с ОР, диаметром 2 мм и длиной 100 мм. Трубка лежит на текстурированной плоскости и освещается светодиодом (LED), размещённым на левом торце. Светодиод эмулируется круглым источником света, имеющим реалистичную гониометрическую диаграмму. Сгенерированные изображения показаны на рис. 14. Как и в случае с изображением пластины, они подразделяются на две половины: левая половина представляет собой синтез с истинными параметрами ОР, а правая половина представляет собой изображение, синтезированное для восстановленных параметров ОР.

Используются те же параметры ОР, что и в примере пластины, представленном на рис. 12 (правая колонка, образец #3).

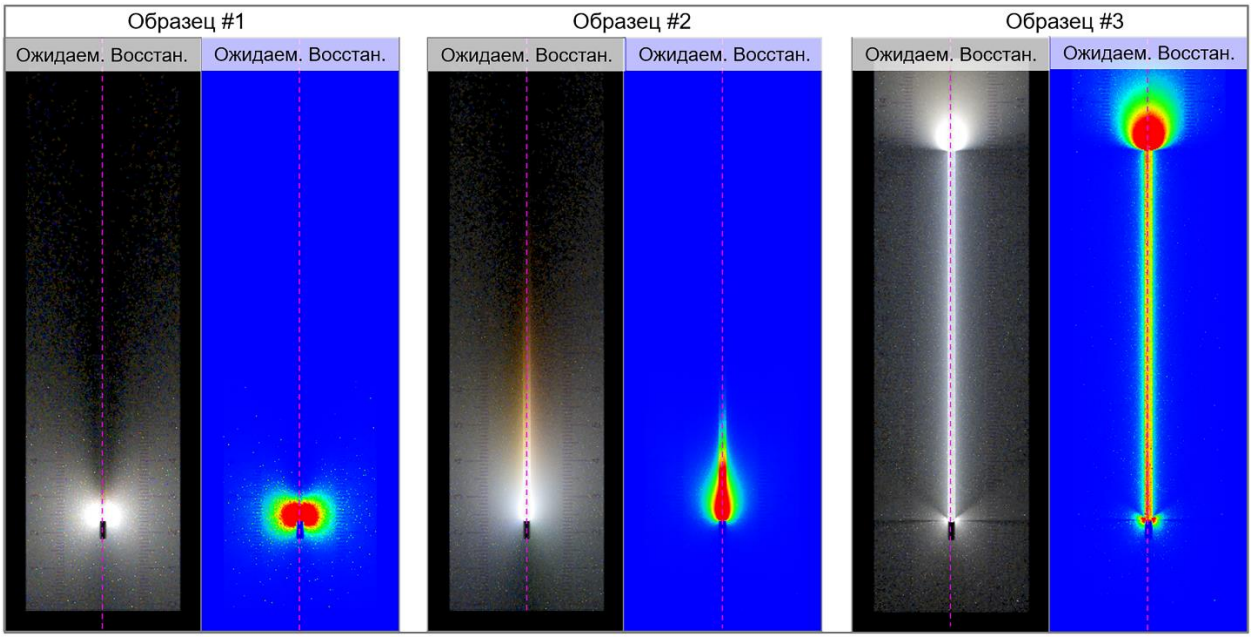


Рис. 14. Визуализация образцов ОР на модели световода

Изображения рис. 14 подобны изображениям, представленным на рис. 12 и отличаются от них только компоновкой. Каждый образец представлен изображением в RGB (левая часть) и в искусственных цветах (правая часть). Изображения, синтезированные для истинных параметров ОР отделены от восстановленных пунктирной пурпурной линией.

Сравнивая изображения на рис. 14, мы наблюдаем хорошее согласование между изображениями, синтезированными для истинных параметров ОР и их восстановленными аналогами, несмотря на заметные различия в значениях параметров ОР. Особенно интересен правый вариант, показывающий рассеяние света средой ОР, соответствующей образцу #3, рассеивающие свойства которой ближе к тем, которые используются в аналогичных моделях. Небольшую разницу можно заметить при внимательном рассмотрении, но она незначительна и в численном выражении $RMSD_{\text{relative}}$ не превышает 5%.

VII. Выводы и заключение

Исследования, отражённые в статье, представляют эффективный метод восстановления параметров ОР на основе измеренных ДФР. Использование двухуровневой оптимизации позволяет получить хорошее согласование между ожидаемой и восстановленной целевыми функциями (ДФР). Для восстановления параметров ОР достаточно использовать одно направление освещения, одну компоненту ДФР (измеренную для пропускания или отражения) и одного образца ОР (с одной толщиной). Усложнение критерия оптимизации с несколькими образцами разной толщины, использование нескольких компонентов ДФР не приводит к заметным улучшениям выходных данных восстановления, а лишь замедляют этот процесс. Восстановленные параметры ОР отличаются от истинных, за исключением размера рассеивающих агентов, однако масштабируемость этих параметров хороша, поэтому их можно использовать для корректного моделирования различных объектов, имеющих среду ОР. Все эти выводы верны только для сред ОР, которые можно аппроксимировать сферической формой рассеивающих диэлектрических частиц, ограниченных гладкими поверхностями, поскольку исследования в статье основаны только на теории Ми. Будущие улучшения возможны с использованием аналитических подходов к решению уравнений переноса света вместо стохастического подхода, использованного в данной работе. Вероятно, это может удалить паразитное влияние стохастического шума. Применение других целевых функций может быть рассмотрено также в качестве примера распределения яркости вдоль световодов, пластин, что может быть полезно для восстановления параметров ОР с очень малым диффузным рассеянием.

Источник финансирования. Благодарности

Работа выполнена при поддержке Российского научного фонда (Разработка методов, алгоритмов и вычислительной архитектуры реконструкции геометрии реального окружения для построения системы смешанной реальности, свободной от дискомфорта зрительного восприятия, №. 22-11-00145)

Список литературы

(исключаются учебники, учебно-методические пособия, рукописи (диссертации, авторефераты))

1. F O Bartell, E. L. Dereniak, and W. L Wolfe "The Theory And Measurement Of Bidirectional Reflectance Distribution Function (Brdf) And Bidirectional Transmittance Distribution Function (BTDF)", Proc. SPIE 0257, Radiation Scattering in Optical Systems, (3 March 1981); https://doi.org/10.1117/12.959611

2. Terms, International Commission on Illumination (CIE), URL: https://cie.co.at/e-ilv, (17 august, 2023)

3. BRDF, International Commission on Illumination (CIE), URL: https://cie.co.at/eilvterm/17-25-089, (17 august, 2023

4. Radiance coefficient, Luminance Factor, International Commission on Illumination (CIE), URL: https://cie.co.at/eilvterm/17-24-078, (17 august, 2023).

5. Sokolov V.G., Potemin I.S., Zhdanov D.D., Barladian B. BSDF Reconstruction of Inks and Paints for Light Guiding Plates // CEUR Workshop Proceedings - 2021, Vol. 3027, pp. 135-149

6. Mishchenko, M., 2015: Measurement and modeling of electromagnetic scattering by particles and particle groups. In Polarimetry of Stars and Planetary Systems (L. Kolokolova, J. Hough, and A.-C. Levasseur-Regourd, Eds.), Cambridge University Press, Cambridge, pp. 13-34.

7. D. Toublanc, "Henyey–Greenstein and Mie phase functions in Monte Carlo radiative transfer computations," Appl. Opt. 35, 3270-3274 (1996).

8. Mishchenko, M. I., 2018: "Independent" and "dependent" scattering by particles in a multi-particle group. OSA Continuum 1, 243-260.

9. Mishchenko, M. I., and L. D. Travis, 2003: Electromagnetic scattering by nonspherical particles. In Exploring the Atmosphere by Remote Sensing Techniques (R. Guzzi, Ed.), Springer-Verlag, Berlin, pp. 77-127

10. Budak V.P., Klyuykov D. K, Korkin S. V., “Convergence acceleration of radiative transfer equation solution at strongly anisotropic scattering”, Light Scattering Reviews 5”, July 2010, DOI: 10.1007/978-3-642-10336-0_5

11. Basov A. Y., Budak V. P., Mathematical Model of a Surface Radiance Factor, GraphiCon-2019 Bryansk, CEUR Workshop Proceedings. 2019. С. 43-47, DOI: 10.30987/graphicon-2019-2-43-47

12. Budak V., Efremenko D., Methods Calculating the Slab Radiance Factor, In: 30th International Conference on Computer Graphics and Machine Vision, GraphiCon 2020 Saint Petersburg, pp. 1-13, DOI: 10.51130/graphicon-2020-2-3-16, ISSN 1613-0073

13. Kokhanovsky A.A., Rozanov V.V., Budak V.P., Klyukov D.A., Korkin S.V., Cornet C., C-Labonnote L., Duan M., Emde C., Mayer B., Katsev I.L., Prikhach A.S., Zege E.P., Min Q., Nakajima T., Ota Y., Yokota T., Benchmark results in vector atmospheric radiative transfer, Journal of Quantitative Spectroscopy and Radiative Transfer. 2010. Т. 111. № 12-13. С. 1931-1946, DOI: 10.1016/j.jqsrt.2010.03.005.

14. Lumicept – A Hybrid Light Simulation Software. Available online: URL https://integra.jp/en/products/lumicept (accessed on 21 Dec 2022).

15. Bohren, C. F.; Huffmann, D. R. (2010). Absorption and scattering of light by small particles. New York: Wiley-Interscience. ISBN 978-3-527-40664-7.

16. Survikov ST (2011). "Mie Scattering". A-to-Z Guide to Thermodynamics, Heat and Mass Transfer, and Fluids Engineering. Thermopedia. Begel House. doi:10.1615/AtoZ.m.mie_scattering. ISBN 978-0-8493-9356-3. Retrieved 28 Jan 2019

17. Tzarouchis, D.; Sihvola, A. "Light Scattering by a Dielectric Sphere: Perspectives on the Mie Resonances". Appl. Sci. 2018, 8, 184.

18. Vasil'ev A.V., Ivlev L.S. Universal algorithm for calculating optical characteristics of two-layer spherical particles with homogeneous core and cover. Atmospheric and Oceanic Optics (1996), 9, No. 12, 982—988

19. "Brute Force Algorithms Explained", freeCodeCamp.org. 2020-01-06, Retrieved 2021-04-11

20. Nelder, John A.; R. Mead (1965). "A simplex method for function minimization". Computer Journal. 7 (4): 308–313. doi:10.1093/comjnl/7.4.308

21. A. G. Voloboi, V. A. Galaktionov, K. A. Dmitriev, and E. A. Kopylov. 2004. Bidirectional Ray Tracing for the Integration of Illumination by the Quasi-Monte Carlo Method. Program. Comput. Softw. 30, 5 (September-October 2004), 258–265. https://doi.org/10.1023/B:PACS.0000043051.47471.aa

22. Zdanov, D.D, et. al. Simulation of the BSDF measurement capabilities for various materials with GCSM-4 gonio-spectrophotometer // SPIE/COS Photonic Asia (2016)

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-179-193

Восстановление оптических свойств объектов сцены методом дифференцируемого рендеринга с применением оптимизации выбора наиболее важных точек

И.Е. Кинёв[1], С.И. Куприянов[1]

[1]ИТМО Университет, Кронверкский проспект 49, Санкт-Петербург, 197101 Россия.

I. E. Kinev[1], S. I. Kyprijanov[1]

[1]*ITMO University, Kronverksky Pr. 49, St. Petersburg, 197101, Russia*

Аннотация. В настоящее время для большинства VR/AR-приложений необходимо физически корректно реконструировать реальный мир и оптические свойства объектов. В данной работе для решения этой задачи предлагается использовать параметрическую модель оптических свойств, которая имеет небольшое количество параметров и корректно моделирует распространение света в сцене. Кроме того, на основе параметров модели выполняется дифференциальный рендеринг изображения, в результате которого полученные дифференциалы передаются в оптимизатор. Оптимизатор, в свою очередь, модифицирует параметры модели для улучшения качества рендеринга и тем самым приближает предлагаемую модель оптических свойств к реальным оптическим свойствам объектов. Для ускорения вычислений предлагается выбирать ограниченное количество значимых точек, которые будут анализироваться.

Ключевые слова: Рендеринг, дифференцируемый рендеринг, трассировка лучей, рассеяние света

Аннотация. In this paper, the authors propose a method to address the challenge of accurately reconstructing the real world and optical properties of objects in VR/AR applications. The main idea is to use a parametric model of optical properties, which has a small number of parameters and can effectively model the propagation of light in the scene. To achieve this, the authors introduce a differential rendering approach based on the model parameters. This means that the image is rendered using the current parameter values, and the resulting differentials are then transferred to an optimizer. The optimizer's role is to modify the model parameters in order to improve the rendering quality and bring the proposed model of optical properties closer to the real optical properties of objects. To speed up calculations, it is proposed to select a limited number of significant points to be analyzed.

***Ключевые слова*:** Rendering, differentiable Rendering, ray tracing, light scattering

I.Введение

В современном мире методы рендеринга реалистичных изображений широко используются для решения различных задач. Например, создание реалистичных кинематографических или анимационных эффектов, моделирование распространения света в различных оптических системах для повышения их надежности к появлению каких-либо помех (шумов, выбросов, ряби и т.п.) при формировании итогового изображения на приемнике, синтез реалистичных изображений для создания и расширения обучающей выборки нейронных сетей и т.д. В данной работе методы рендеринга реалистичных изображений будут рассмотрены применительно к восстановлению оптических свойств объекта по набору изображений с заранее уже известной геометрии сцены.

Для достижения более реалистичного представления сцен в компьютерной графике предложено множество моделей, имитирующих оптические свойства объектов, чтобы приблизить создаваемые изображения к реальным фотографическим образцам. Среди наиболее популярных моделей можно назвать модель Фонга [1] , Диснеевскую модель[2] , модель Кука-Торренса [3] и т. д. Задавая и изменяя различные параметры модели для того или иного объекта сцены, можно подобрать наиболее подходящие значения параметров, позволяющие получить максимально реалистичное изображение, а затем выполнить рендеринг с учетом свойств создаваемых объектов.

Рендеринг может быть выполнен различными способами, с использованием растеризации или бросания лучей (ray casting). Однако в данной работе предлагается использовать методы рендеринга, основанные на трассировке лучей, которые позволяют физически корректно моделировать распространение света в сцене, при этом рассматривая такие физические явления, как: преломление света, рассеяние, отражение, дисперсия, интерференция, каустика и т. д. В общем, данные методы вычисляют интеграл яркости для наблюдаемой точки поверхности, который рекурсивно собирает освещенность по всей сцене, учитывая свойства поверхности.

Свойства задаются через функцию двунаправленного рассеяния (BSDF) [3, 4], после чего решается уравнение рендеринга [5].

Кроме того, для реконструкции реальной сцены необходимо знать геометрию объектов, расположенных в ней. Для этого можно использовать различные камеры, которые могут строить изображение с картами глубины, например, камеры «Time of Flight», камеры «Structured Light» или камеры с датчиками LiDAR. Затем, используя изображения RGBD, можно сгенерировать облако точек сцены и использовать его для построения геометрии сцены.

Таким образом, данные методы могут быть использованы в VR/AR-приложениях для физически-корректной реконструкции реального мира. Однако, несмотря на это, на данный момент все еще существуют проблемы с несоответствием геометрии и оптических свойств реальных объектов с объектами в 3D-сцене, что негативно сказывается на самочувствии человека при длительном использовании VR/AR-устройств из-за возникновения визуальных конфликтов.

Для частичного решения этой проблемы мы предлагаем реконструировать 3D-модели сцен по облаку точек или картам глубины с использованием дифференцируемого рендеринга, учитывая следующие характеристики: геометрию, оптические свойства, источники света. В данной работе в качестве базовой оптической модели для упрощения вычислений используется параметрическая модель, состоящая из коэффициентов диффузного рассеивания, отражения и преломления (KD, KS, KTS). Она имеет небольшое количество параметров и в то же время корректно моделирует распространение света в сцене.

II.Обзор существующих работ

Как уже упоминалось выше, для реконструкции материальных свойств реальных объектов используются различные приближенные модели оптических свойств. Одной из самых простых моделей является Фонг [1]. Модель основана на трех базовых компонентах освещения: диффузная, зеркальная, амбиентная. Для каждой точки поверхности, имеющей свои координаты и нормаль к поверхности, освещенность складывается на основе этих трех компонент, при этом свойства материала поверхности отвечают за способность воспринимать отдельный тип освещения, а свойства источников света — за мощность излучения для каждого отдельно взятого компонента. Преимущества этого метода в том, что он прост в реализации, обратим и широко используется. Однако он не совсем физически корректен и не подойдет для реконструкции сложных оптических свойств. В этом контексте стоит упомянуть модель освещенности освещенности Кука-Торренса, которая используется для расчета отраженного света. Эта модель основана на трех компонентах: коэффициенте Френеля, геометрической компоненте, которая учитывает самозатенение, и компоненте, учитывающей шероховатость поверхности. Шероховатость моделируется набором блестящих микрограней (маленькое зеркало), ориентированных по-разному, и те микрограни, которые ориентированы правильно, вносят наибольший вклад в отраженный свет. Эта модель хорошо подходит для создания различных стеклянных и металлических поверхностей. Также стоит отметить модель оптических свойств Disney [2], которая содержит 10 параметров в диапазоне [0; 1], что дает пользователю гибкость в настройке, но она ориентирована на художественное искусство, а не физически правильную визуализацию.

Как уже было упомянуто выше, в данной работе, с целью упрощения вычислений, в качестве базовой оптической модели используется параметрическая модель, состоящая из коэффициентов диффузного рассеивания (KD), отражения (KS) и преломления (KTS). Эта модель обладает небольшим количеством параметров, что обеспечивает удобство использования, и при этом корректно моделирует распространение света в сцене, что позволяет достичь высокого уровня реалистичности визуализации. В дополнение к простым коэффициентам, используемым в данной работе, можно применить смесь различных моделей оптических свойств, каждая из которых будет вносить свой вклад в реконструкцию материальных свойств объектов. Для этого используются веса, которые определяют значимость каждой модели в зависимости от конкретных условий и задач. Такой подход позволяет объединить преимущества различных моделей и улучшить точность реконструкции оптических свойств реальных объектов.

Рассматривая методы рендера, основанные на трассировке лучей, можно выделить две категории. По способу интегрирования: с использованием обычного интегрирования Монте-Карло и интегрирования цепей Маркова. Простейшими методами, основанными на обычном интегрировании Монте-Карло, являются прямая, обратная и двунаправленная трассировка лучей. Данные методы позволяют достаточно точно рассчитать вторичное освещение, но плохо подходят для расчета прямого освещения, например, когда в сцене есть большой источник света или глянцевый материал. Это связано с тем, что при трассировке вероятность случайного попадания в камеру для прямой трассировки и в источник света для обратной трассировки крайне мала. Для решения этой проблемы предложено несколько модификаций обратной трассировки лучей [6, 7]. Методы, основанные на цепях Маркова (методы Метрополиса) [8, 9, 10], хорошо подходят для расчета вторичного освещения, а также

позволяют эффективно рассчитывать сцены, содержащие большое количество зеркальных или преломляющих элементов. Кроме того, предложены методы фотонных карт [11], которые достаточно хороши для моделирования каустического освещения. Также были предложены методы вычисления дифференциальных изображений: дифференцируемая трассировка лучей Монте-Карло с использованием выборки краев [12], дифференцируемая визуализация в пространстве путей [13] и сопряженный метод вычисления градиентов перемещения света [14]. Каждый метод основан на идее, что сопряженное уравнение для градиента почти идентично сопряженному уравнению для значимости в теории переноса, решенной с использованием трассировки пути, с единственным отличием в функции выборки значимости. Первые два метода учитывают такие параметры, как геометрические, оптические, местоположение камеры, местоположение источника света и являются общими.

III.Методы

Предлагаемый алгоритм реконструкции оптических свойств показан на Рис.1, он состоит из восьми основных шагов:

1. получение изображений с яркостью в каждом пикселе и геометрией реальной сцены;
2. Анализ изображения и геометрии для отбора точек и получения информации о видимости точек с разных камер;
3. выбор точек на разных объектах сцены, которые будут использоваться для расчета дифференциалов;
4. рендеринг методом трассировки пути по выбранным точкам;
5. сохранение пути траектории луча для каждой итерации;
6. изменение оптических свойств объектов на основе вычисленных дифференциалов для одного луча;
7. рендеринг методом трассировки пути на основе полученных коэффициентов параметрической оптической модели на шаге 6;
8. сравнение полученных изображений из пункта 7 с истинными значениями из оригинального изображения.

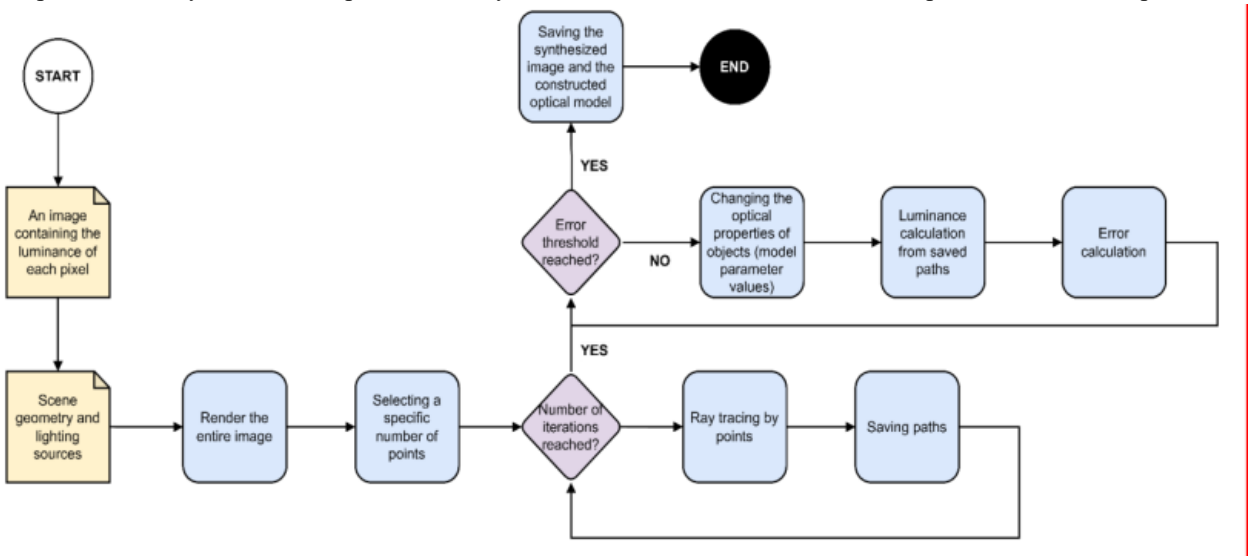


Рис. 1 Алгоритм восстановления оптических свойств

В данной работе предполагается, что первый шаг уже выполнен, т. е. мы уже знаем геометрию сцены, положение камеры в момент съемки под определенным углом и источники света. Предполагается, что на вход подается готовое изображение, содержащее значения яркости или RGB для каждого пикселя и описание геометрии и источников света в виде текстового файла, описывающего координаты вершин, треугольников и т. д. Для отбора точек необходимо выполнить анализ изображения и использовать геометрию сцены. При анализе выявляются потенциальные точки, содержащие большой вклад прямого освещения. Каждая точка проверяется на видимость с разных ракурсов камеры и дополняется информацией о яркости с этих камер. В итоге получается набор точек кандидатов с информацией.

В процессе реконструкции материальных свойств объектов, мы знаем положение камеры, геометрию сцены и источники света в сцене. Наложив изображение на геометрию сцены, мы можем сравнивать яркости исходного изображения с яркостями, полученными при рендеринге. В зависимости от разницы значений яркости, мы изменяем значения параметров оптических свойств объектов, таких как коэффициенты диффузного рассеивания,

отражения и преломления (KD, KS, KTS). Таким образом, мы можем приблизительно восстановить реальные свойства объектов.

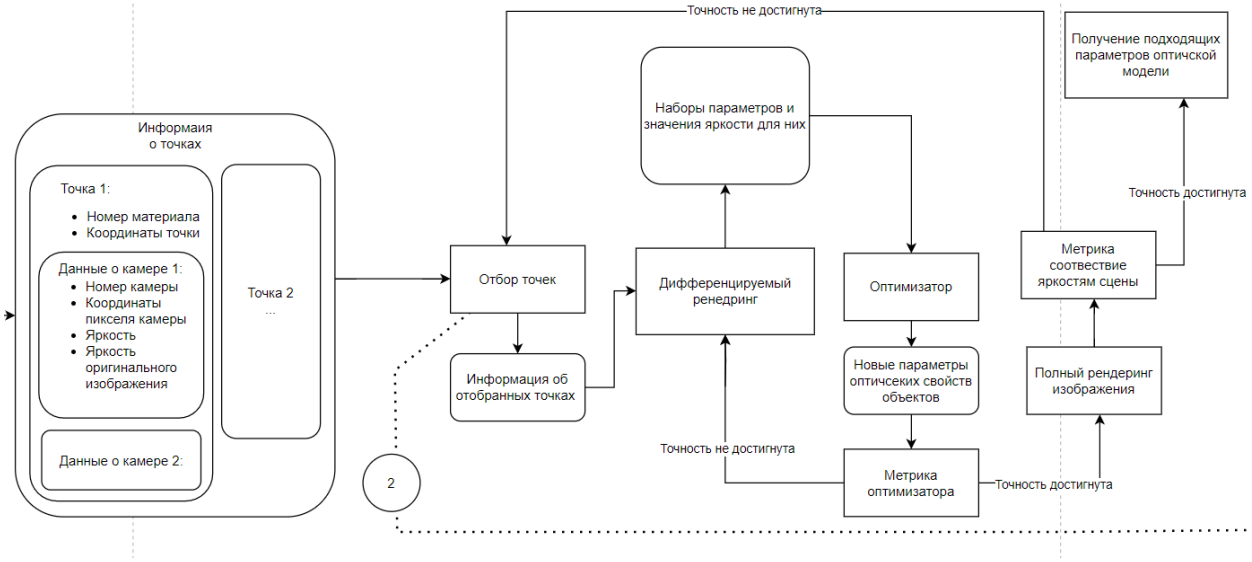


Рис. 2. Общий конвейер восстановления оптических свойств

Для рендеринга изображения, ввиду простоты его реализации, было решено использовать метод трассировки лучей, основанный на обычном интегрировании Монте-Карло. В этом методе трассировка траектории луча начинается с камеры, затем луч случайным образом перемещается по сцене, пока не достигнет источника света, не покинет сцену или не превысит максимальную глубину рекурсии.

Поскольку расчет распространения световых лучей в сцене требует сравнительно большого количества вычислительных ресурсов и времени, предлагается выбирать ограниченное количество точек на каждом объекте и анализировать только их при расчете разницы яркостей исходного изображения и синтезированного изображения. Проверяется, что точка не находится на границе геометрии объекта, и имеет прямое освещение. Точки сортируются по значению яркости на оригинальном изображении при запуске восстановления оптических свойств, а на последующих этапах вычисляют примерное процентное соотношение первичного освещения из оригинальной яркости, так как первичное освещение наиболее сильно показывает оптические свойства объекта.

Затем выполняется сравнительно большое количество итераций (10000-100000) расчетов этих точек для повышения точности расчетов значений яркости. В начале каждой итерации из камеры испускаются лучи в направлении выбранных точек на объектах, после чего весь путь, пройденный лучом при трассировке, сохраняется и все начинается заново. Поскольку точек немного, расчет этих трасс займет около нескольких секунд. Затем мы изменяем оптические свойства объектов и заново запускаем трассировку лучей. Поскольку вычисляется диффузная составляющая, лучи в принципе могут следовать по тем же путям, но они будут накапливать другую яркость, так как оптические свойства объектов изменились. Чтобы учитывать тот факт, что лучи могут пойти в зависимости от модифицированных коэффициентов по другому пути, мы производим нормировку оптических свойств в зависимости от весов свойств текущего объекта и предыдущего.

IV.Выбор точек

При восстановлении оптических свойств крайне важно оптимизировать процесс выбора точек. Если дифференцируемый рендеринг будет осуществляться по всему изображению, то процесс будет происходить существенно медленнее. При таком подходе для каждого пикселя будут рассчитываться дифференциалы и итоговый расчёт будет зависеть от разрешения изображения и шага дифференцируемого рендеринга. Для того что бы ускорить процесс предлагается воспользоваться знанием о геометрии сцены и яркостях исходных изображений.

В контексте задачи точки можно выбирать и случайно, однако для оптимизации процесса восстановления оптических свойств можно искать точки, которые хорошо описывают данный объект. Такие точки помогут достичь более быстрой сходимости оптимизирующей функции. Хорошо описывающая объект точка должна удовлетворять некоторым условиям: она сильно освещена прямым освещением, а также ее видно с разных ракурсов камеры, то есть под разными углами падения луча. В контексте дифференцируемого рендеринга процесс отбора точек должен включать в себя: анализ исходного изображения, анализ геометрии сцены и анализ значений яркостей для точек кандидатов.

При восстановлении оптических свойств объекта предполагается, наличие некоторого количества изображений (фотографии объекта с разных ракурсов камеры) и попиксельно информация о яркости. Для этого на вход программы подается изображения и оттуда формируется матрица с размерами равным разрешению изображения со значениями яркости. Однако для установления физически корректного значения яркости предлагается воспользоваться механизмами программного обеспечения Lumicept. Данная программа позволяет использовать свои интерфейсы взаимодействия, а также скрипты python для формирования матрицы яркостей как в фотометрических величинах, так и в спектральном и rgb представлении [15]. Она будет использовать физически корректный рендеринг для получения точных значений яркости используя информацию о геометрии. В данной работе использовалось api Lumicept для написания скрип файла python который получает на вход физически корректно отрендеренный файл с расширением .nit (выходной файл при рендеринге средствами Lumicept) и с помощью функции PostProcessor записывает попиксельно значения RGB в .csv файл.

Когда матрица сформирована происходит анализ геометрии сцены. Предполагается, что геометрия виртуальной сцены для фотографии уже сформирована и определена верно. Данные о геометрии сформированы в виде набора полигонов(треугольников) для объектов сцены и информация о их расположении в трехмерном пространстве сцены. Таким образом на вход программы подаются матрицы яркостей для каждого ракурса камеры и значения положении геометрии в трехмерном пространстве, вместе с данными о камерах и источниках света.

Когда все основные данные подготовлены происходит формирование точек кандидатов. Для этого происходит запуск виртуальных лучей попиксельно из заданных расположений камеры. Это поможет определить в какую именно геометрию попадут лучи при рендеринге изображения из данного пикселя. Знание о материале точки необходимо для изменения параметров модели оптических свойств при использовании дифференцируемого рендеринга. При условии, что луч из камеры через пиксель попал в объект, происходит проверка: на наличие компоненты первичного освещения (вектор из точки в источники освещения не перекрывают другие полигоны) и видимость данной точки с других ракурсов камеры. Для этого из точки пускается луч во все камеры и происходит проверка на разные условия соответствия. К таким условиям относиться: точка находится перед камерой и плоскостью экрана; нормаль треугольника направлена в сторону камеры; между точкой и камерой нет других треугольников, перекрывающих ее видимость; точка пересечения вектора из геометрии в новую камеру и плоскостью экрана этой камеры находиться в границах экрана. При удовлетворении всех условий данные о видимости точки с разных камер записываются в информацию о точки, а также вычисляются координаты пересечения вектора (из геометрии в новую камеру) с плоскостью экрана для того, чтобы найти информацию о яркости изображения по матрице яркостей оригинального изображения.

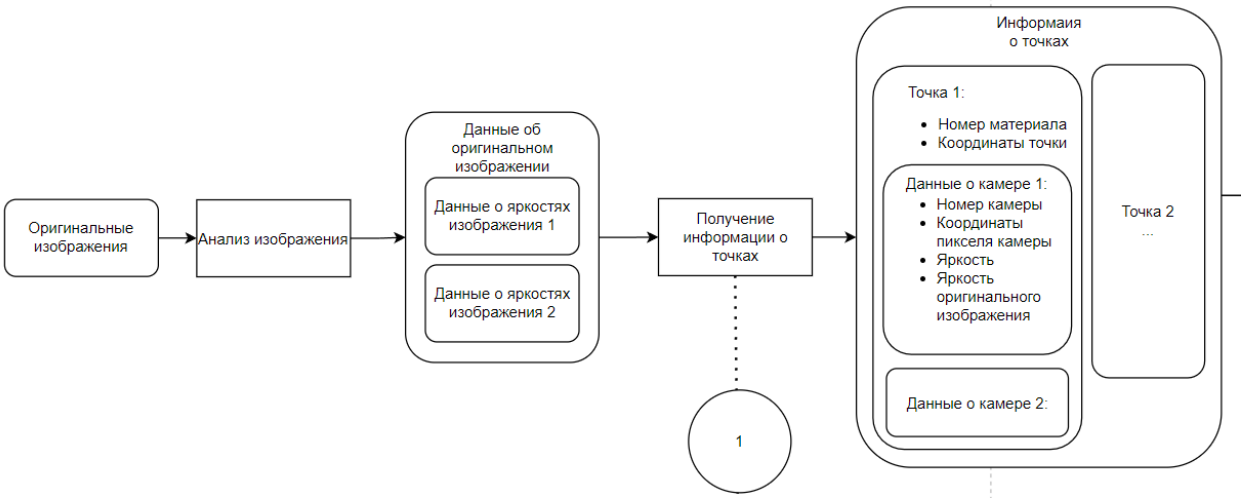


Рис. 2. Анализ изображения

На рис. 2 схематично представлен анализ изображения. На вход подаются оригинальные изображения и с помощью скриптов получаются попиксельно данные о яркостях. А после эта информация передается в функцию получения данных о точках.

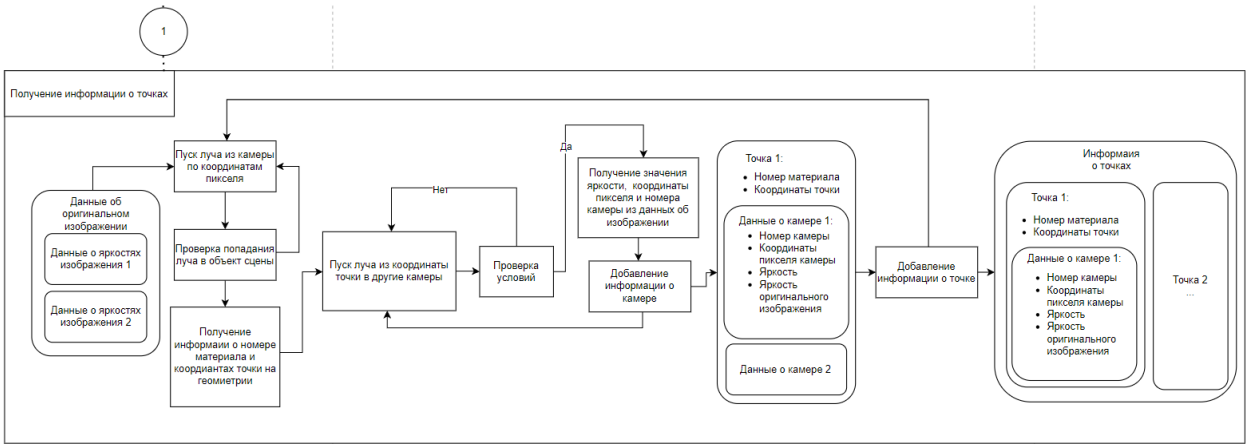


Рис. 3. Получения информации о точках

На рисунке 3 показан процесс получения информации о точках и проверки на видимость разных камер.

Когда все точки обработаны происходит предварительный отбор точек. Для этого в каждой точке суммируется яркость со всех экранов на которых она видна и происходит сортировка по этим значениям. Так как яркость оригинального изображения задана в формате RGB для сортировки, значения яркостей конвертируются в фотометрические значения[16]. Для корректной конвертации используются данные о белой точке и координаты цветов из программного обеспечения Lumicept. Далее некоторое количество точек отбирается для проведения начального дифференцируемого рендеринга и получения первых результатов его работы. Когда процессы дифференцируемого рендеринга завершили свою работу и сформировались примерные значения параметров модели оптических свойств происходит полный рендеринг изображения по всем камерам сцены и на основе этих данных формируется метрика ошибки по совпадению значений яркости с оригинальным изображением. При сравнении используется конвертация яркостей из длин волн, на которых работает рендеринг в фотометрические значения. Для этого спектрограмма сворачивается с Y кривой спектра видимости. При оценках ошибки используется среднеквадратичное отклонение яркости пикселя оригинального изображения и яркости, полученной при рендеринге. Если ошибка не приблизилась к заданным минимальным значениям, то происходит переопределение точек. Однако при сортировке по яркостям после нескольких итераций дифференцируемого рендеринга учитывается вклад первичного освещения. Для этого при дифференцируемом рендеринге определяется вклад первичного освещения по отношению ко вторичному, а после процентное соотношение вклада применяется к яркости оригинального изображения и получается значение оригинальной яркости, созданной первичным освещением. Это поможет лучше отобрать точки так как именно вклад первичного освещения дает хорошую оценку оптической модели для данной геометрии.

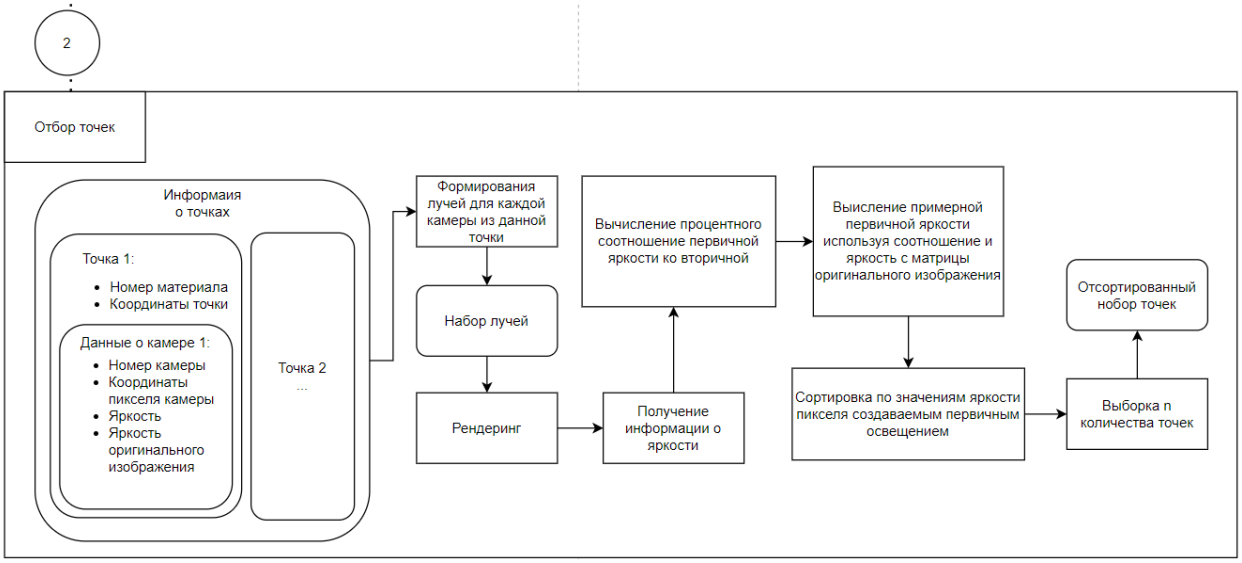


Рис. 4. Отбор точек

На рисунке 3 показана общая схема отбора точек для дифференцируемого рендеринга. После нескольких итераций дифференцируемого рендеринга происходит проверка на соответствие ошибки по яркостям оригинального изображения и полученного при рендеринге с полученными коэффициентами оптических

свойств. Если точность не достигнута, то происходит отбор новых точек с учетом пересчета примерной яркости, созданной первичным освещением. Однако при первом запуску дифференцируемого рендеринга коэффициенты неизвестны и тогда отбор точек осуществляется только по оригинальным яркостям изображения.

V.Расчет дифференциалов

Расчёт дифференциалов осуществляется следующим образом: на вход алгоритму подаются исходные изображения, полученные с различных ракурсов камер. Затем генерируются изображения, соответствующие положениям этих камер (Рис. 5). Для каждого объекта, в который попал луч во время трассировки, последовательно и с определённым шагом изменяется значение каждого коэффициента в оптической модели объекта. В соответствии с проведенными изменениями вычисляется яркость, формирующая изображение с модифицированными параметрами оптической модели. Полученные изображения с разных камер поступают в оптимизатор, который выбирает наилучший вариант изменения параметров. Этот процесс повторяется до достижения заданного порога точности.

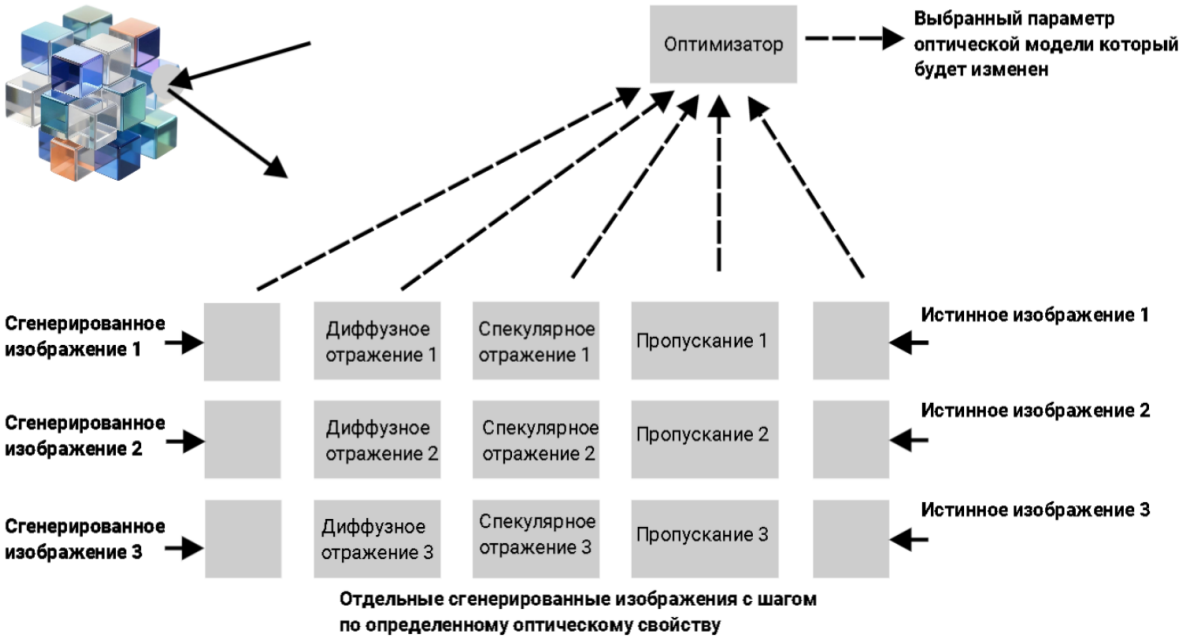


Рис. 5. Схема обновления параметров оптической модели

Расчёт яркостей и изменение параметров осуществляются в ходе трассировки луча для конкретного положения камеры (Рис. 6). Сначала луч взаимодействует с первым объектом, что приводит к формированию первичного освещения. Затем последовательно изменяются коэффициенты материала объекта, на который попал луч, с шагом 0.001, и на основе этих изменений рассчитываются яркости для текущего объекта. Далее, используя метод русской рулетки, определяется событие, которое произойдёт с лучом. В зависимости от выбранного события, луч меняет свою траекторию. Затем луч взаимодействует с другим объектом, где уже рассчитывается вторичное освещение. Этот процесс продолжается до тех пор, пока луч не покинет пределы сцены или не достигнет пороговой глубины распространения.

Далее вычисленные дифференциальные яркости для каждого объекта, полученные при трассировке луча, сохраняются в виде таблиц (Рис. 7). Данные таблицы также содержат информацию о глубине освещения, что учитывается при анализе и вычислении ошибки между оригинальным изображением и изображением с дифференцируемыми параметрами (Рис. 9). Чтобы правильно учесть значимость первичного, вторичного и последующего освещения, вычисляются коэффициенты значимости в зависимости от распределения значений яркости (Рис. 8). Например, если яркость вторичного освещения больше, то коэффициент значимости вторичного освещения будет преобладать над коэффициентом значимости первичного освещения и т.д.

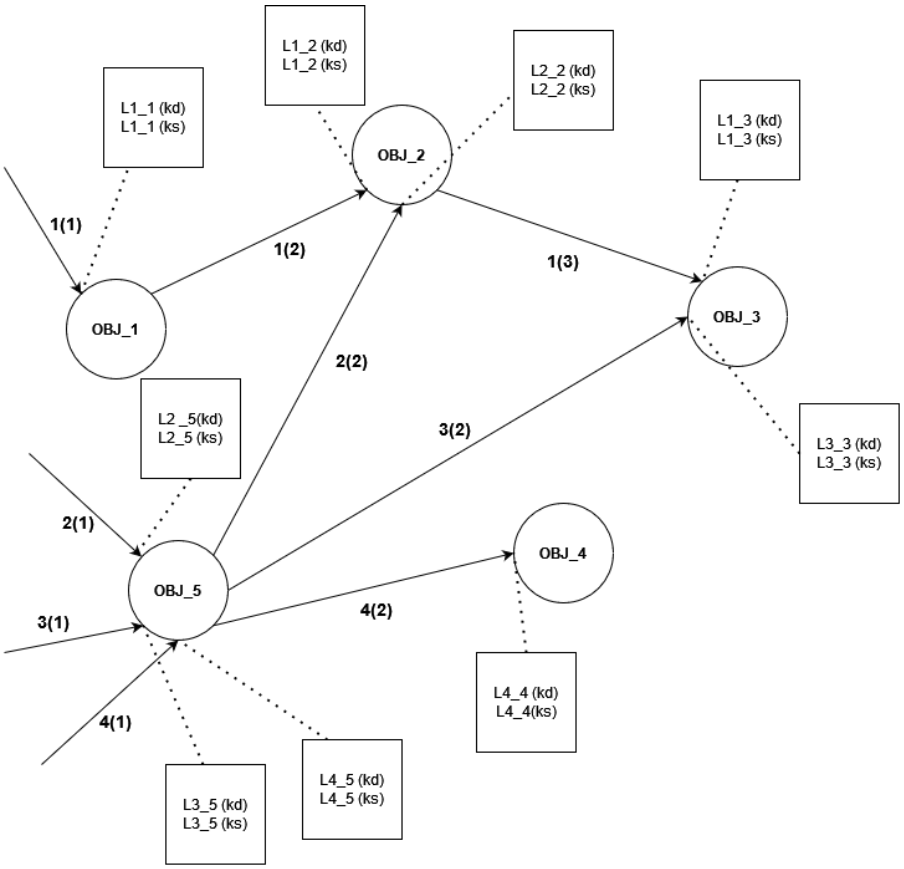


Рис. 6. Схема распространения лучей в сцене и вычисленные дифференциалы с определенным шагом

OBJ_1	
KD	L1_1
KS	L1_1
DEPTH	1

OBJ_2		
KD	L1_2	L2_2
KS	L1_2	L2_2
DEPTH	2	2

OBJ_3		
KD	L1_3	L3_3
KS	L1_3	L3_3
DEPTH	3	2

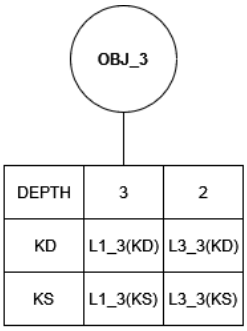
OBJ_4	
KD	L4_4
KS	L4_4
DEPTH	2

OBJ_5			
KD	L2_5	L3_5	L4_5
KS	L2_5	L3_5	L4_5
DEPTH	1	1	1

Рис. 7. Таблицы яркостей для каждого объекта

DEPTH	1	2	3	4	5
L	50	120	30	20	10
COEF	0,2174	0,5217	0,1304	0,870	0,435

Рис. 8. Коэффициенты значимости



OBJ_3_KD_LOSS = (coef2 * L_original + coef3 * L_original) - (coef2 * L1_3_KD * coef3 * L3_3_KD)

OBJ_3_KS_LOSS = (coef2 * L_original + coef3 * L_original) - (coef2 * L1_3_KS * coef3 * L3_3_KS)

Рис. 9. Вычисление ошибки

Псевдокод алгоритма восстановления оптических свойств:

```
# Инициализация переменных
original_image = original_image_init()
step_size = 0.001
learning_rate = 0.001
epochs = 15
coefs = init_coefs()
loss_bound = init_loss()

# Процесс восстановления
for epoch in range(epochs):
    rays = get_important_rays()

    # Рендеринг изображения
    rendered_image = render_image(coefs)

    # Вычисление общей потери
    image_loss = calculate_loss(original_image, rendered_image)  # MSE

    # Проверка потери
    if image_loss < loss_bound:
        return

    # Обновление коэффициентов
    for ray in rays:
        # Вычисление таблиц световых значений
        luminance_tables = calculate_diffs(ray, coefs, step_size)

        # Вычисление локальных потерь
        local_losses = calculate_local_losses(luminance_tables, original_image, coefs)

        # Обновление коэффициентов
        coefs = update_coefs(local_losses, learning_rate)
```

В начале инициализируются переменные: изображение с исходными материальными свойствами, шаг обучения, скорость обучения, количество эпох и начальные коэффициенты оптических свойств. Затем запускается процесс восстановления, который происходит в течение заданного количества эпох.

В каждой эпохе сначала вычисляются важные лучи, которые будут использоваться для обучения коэффициентов оптических свойств. Затем с использованием текущих коэффициентов происходит рендеринг изображения. После этого вычисляется общая потеря между исходным и рендеренным изображениями с использованием среднеквадратичного отклонения (MSE). Если потеря оказывается меньше заданной границы, обучение прекращается и возвращаются текущие коэффициенты.

В противном случае для каждого важного луча вычисляются таблицы световых значений, которые отражают влияние коэффициентов оптических свойств на распространение света в сцене. Затем для каждого пикселя в таблицах световых значений вычисляется локальная потеря, которая отражает разницу между исходным и

рендеренным значениями яркости для определенного пикселя. Наконец, коэффициенты оптических свойств обновляются с использованием градиентного спуска и текущих значений локальных потерь, учитывая скорость обучения. Таким образом, алгоритм постепенно приближается к реальным материальным свойствам объектов, используя методы рендеринга, основанные на трассировке лучей.

VI. Результаты

Для тестирования алгоритма была выбрана сцена Cornell Box (Рис. 10), а именно ракурс камеры, который направлен от задней стенки на маленький куб, вследствие чего куб имеет как прямое освещение на верхней части куба (прямоугольного источника света) и вторичное освещение на задней стенке. Стоит сделать уточнение по поводу цвета объектов сцены, так как рендер производился в 4 длинах волн (400, 500, 600, 700), а потом яркость переводилась в скалярное значение (cd/m^2), из-за малого диапазона волн цвет имеет не оригинальное значение.

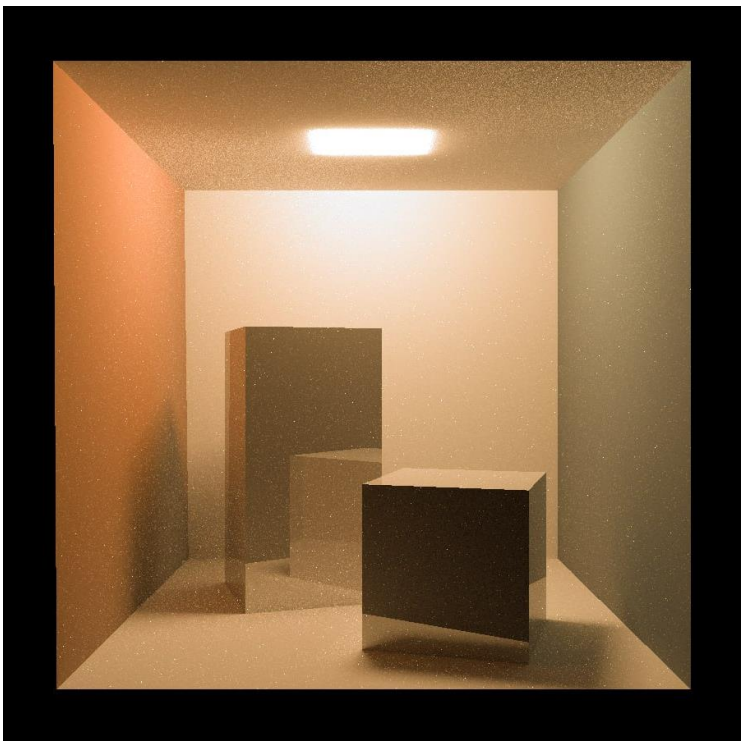


Рис. 10. Сцена Cornell Box (250 фаз рендера)

В ходе текущих экспериментов мы используем только коэффициент KD для упрощения расчетов. В дальнейшем мы планируем использовать смесь различных моделей оптических свойств с различными весами для улучшения точности реконструкции материальных свойств объектов.

На Рис. 11 представлены графики восстановления оптических свойств с оригинальными коэффициентами BRS_0=0.2, 0.3, 0.5 (задняя стенка) и BRS_1=0.2, 0.3, 0.5(куб) с заданными коэффициентами параметрической оптической модели BRS_0=BRS_1=0.01. В данном случае BRS_2 является материалом источника света, который не изменяется, так как мы знаем информацию об источнике света. Аналогичные тесты представлены на Рис. 12, только восстановление происходит со следующими заданными коэффициентами BRS_0=BRS_1=1.0. Количество отобранных лучей для анализа было равно 200.

Как можно заметить алгоритму удалось с небольшой погрешностью восстановить заданные оригинальные оптические свойства для куба (BRS_1), однако не совсем корректно были восстановлены свойства задней стенки (BRS_0). Это связано с тем, что камера в сцене направлена от стенки на куб и влияние BRS_0 не так велико на свойства куба, который виден практически на всем изображении.

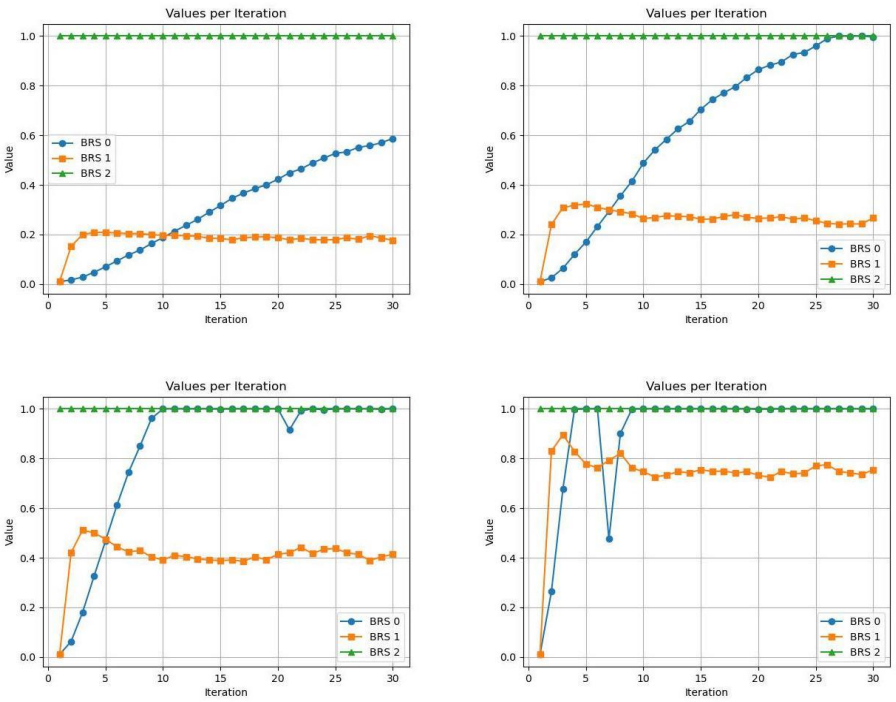


Рис. 11. Пример восстановления коэффициента BRS_1 (куб) если изначально коэффициенты BRS_1 (Маленький куб) и BRS_0 (задняя стена сцены Корнелл Бокс) заданы (0.01), количество лучей: 200. а) Пример восстановления для оригинального изображения, где коэффициенты (brs_0=0.2,brs_1=0.2) б) оригинальные коэффициенты (brs_0=0.3,brs_1=0.3) в) оригинальные коэффициенты (brs_0=0.5,brs_1=0.5) г) оригинальные коэффициенты (brs_0=1.0,brs_1=1.0)

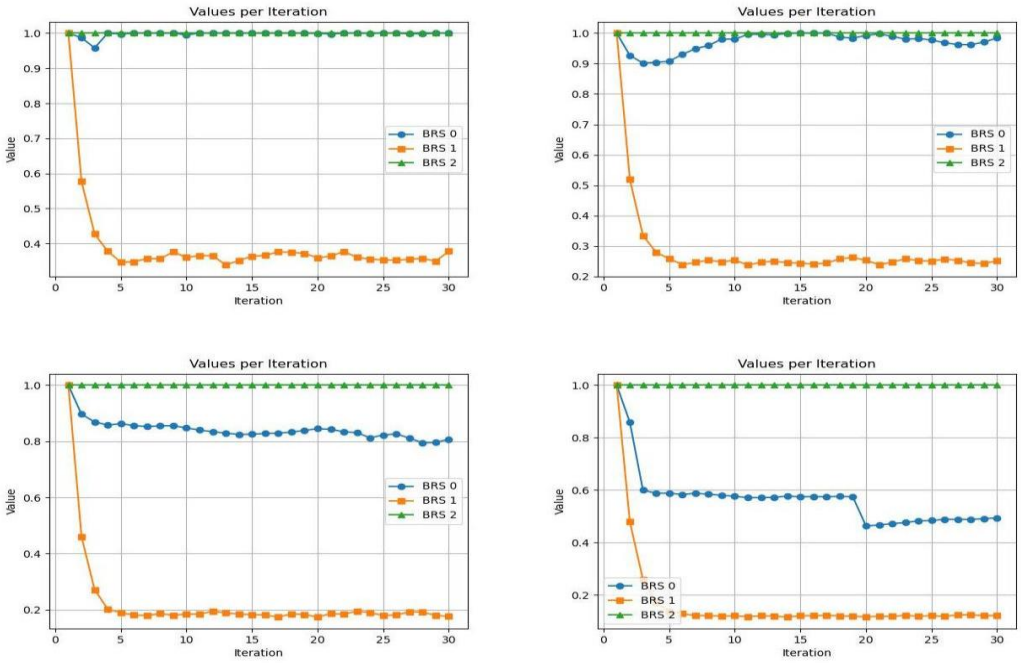


Рис. 12. Пример восстановления коэффициента BRS_1 (куб) если изначально коэффициенты BRS_1 (Маленький куб) и BRS_0 (задняя стена сцены Корнелл Бокс) заданы (1.0), количество лучей: 200. а) Пример восстановления для оригинального изображения, где коэффициенты (brs_0=0.5,brs_1=0.5) б) оригинальные коэффициенты (brs_0=0.3,brs_1=0.3) в) оригинальные коэффициенты (brs_0=0.2,brs_1=0.2) г) оригинальные коэффициенты (brs_0=0.1,brs_1=0.1)

На Рис. 13 представлен график восстановления оптических свойств для объекта BRS_1 с использованием 1000 лучей для анализа, чтобы увеличить точность, что дало положительный результат и восстановленный коэффициент стал практически равен оригинальному. Кроме того, для проверки, данные изображения были визуализированы, а также сравнены яркости наиболее важных областей (Рис. 14-15).

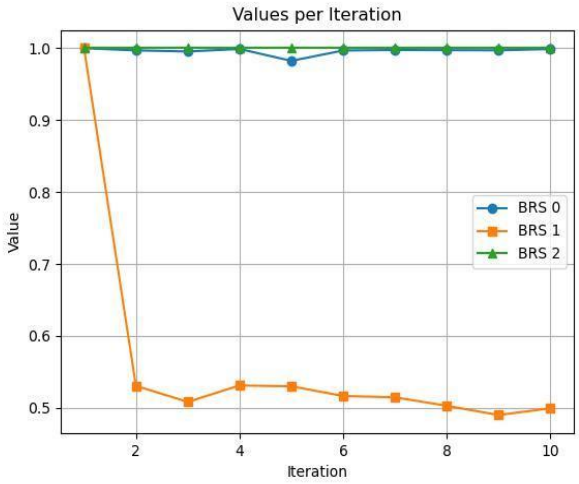


Рис. 13. Пример восстановления коэффициента BRS_1 (куб) если изначально коэффициенты BRS_1 (Маленький куб) и BRS_0 (задняя стена сцены Корнелл Бокс) заданы (1.0), оригинальные коэффициенты (brs_0=0.2,brs_1=0.2), количество лучей: 1000.

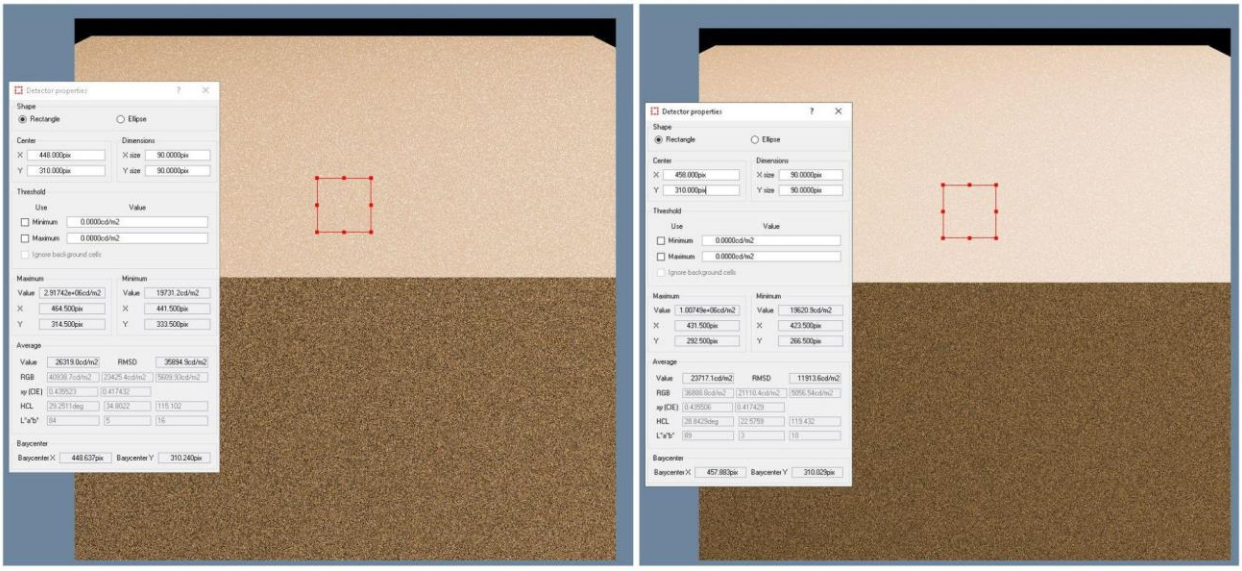


Рис. 14. Сравнение яркости на верхней части куба. Слева оригинальное изображение, справа восстановленное.

Как можно заметить верхняя часть куба, где преобладает первичное освещение практически не отличается по яркости от оригинального изображения, однако задняя стенка куба, где преобладает вторичное освещение, отличается практически в 2 раза. Как уже было сказано это связано с тем, что на данный момент мы используем только одну камеру и в сцене отсутствует задняя стенка сцены Cornell Box (BRS_0). Из-за этого плохо восстановился коэффициента для BRS_0, вследствие чего есть такое различие во вторичной яркости между оригинальным и восстановленным изображениями. Кроме того на рисунках 16-17 представлены изображения шагов восстановления оптических свойств с изначальными параметрами равными 0.01 к оригинальным значениям 0.5. На рисунке 16 представлены результаты рендера, а на рисунке 17 представлена тепловая карта по полученной яркости, в зависимости от шкалы на рисунке 18.

В ближайшем будущем мы планируем использовать несколько камер чтобы более точно восстанавливать оптические свойства всех объектов в сцене, а также использовать смесь различных моделей оптических свойств с различными весами для улучшения точности.

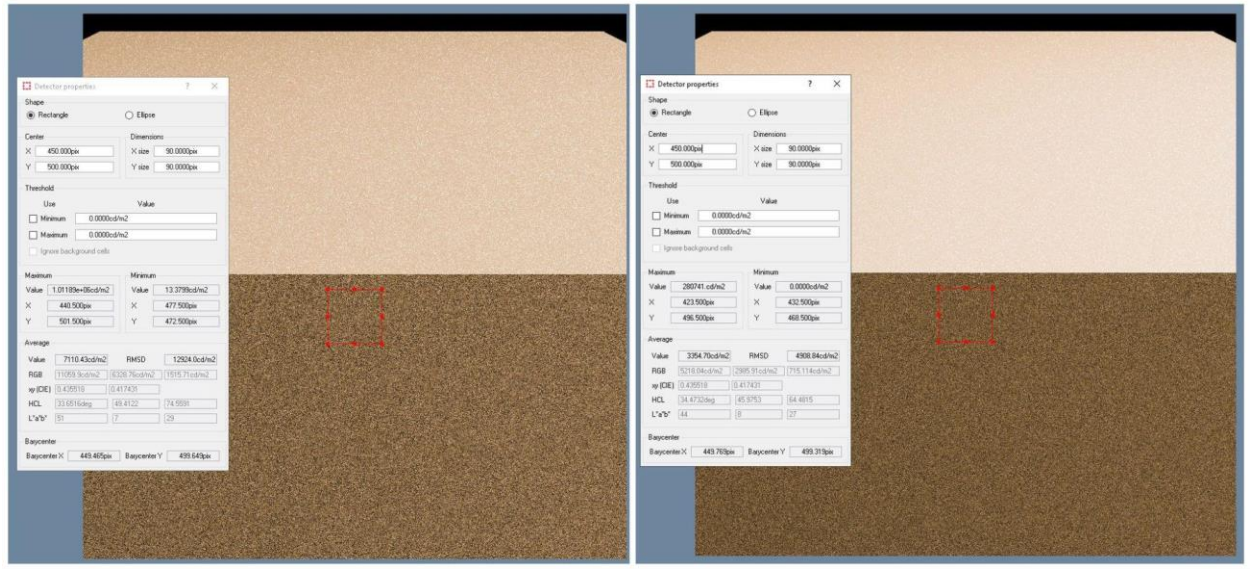


Рис. 15. Сравнение яркости на задней стенке куба. Слева оригинальное изображение, справа восстановленное.

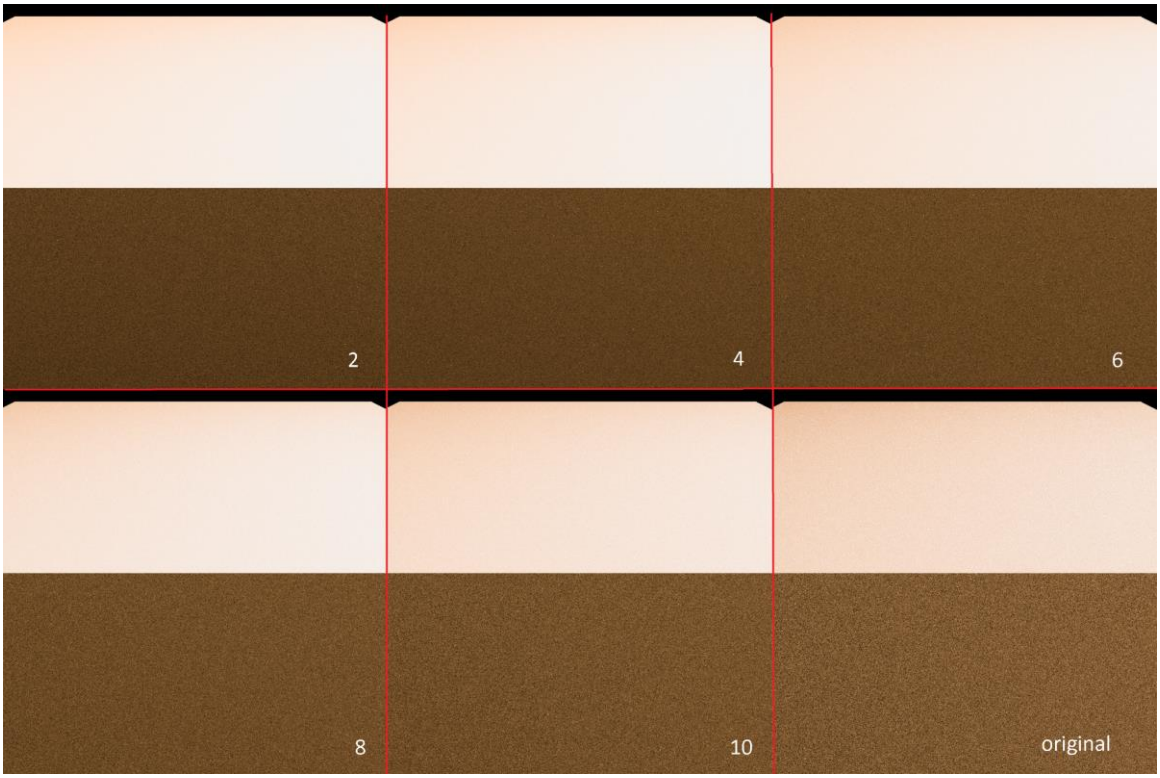


Рис. 16. Результаты рендера по итерациям восстановления изначальных коэффициентов 0.01 к оригинальным значениям 0.5.

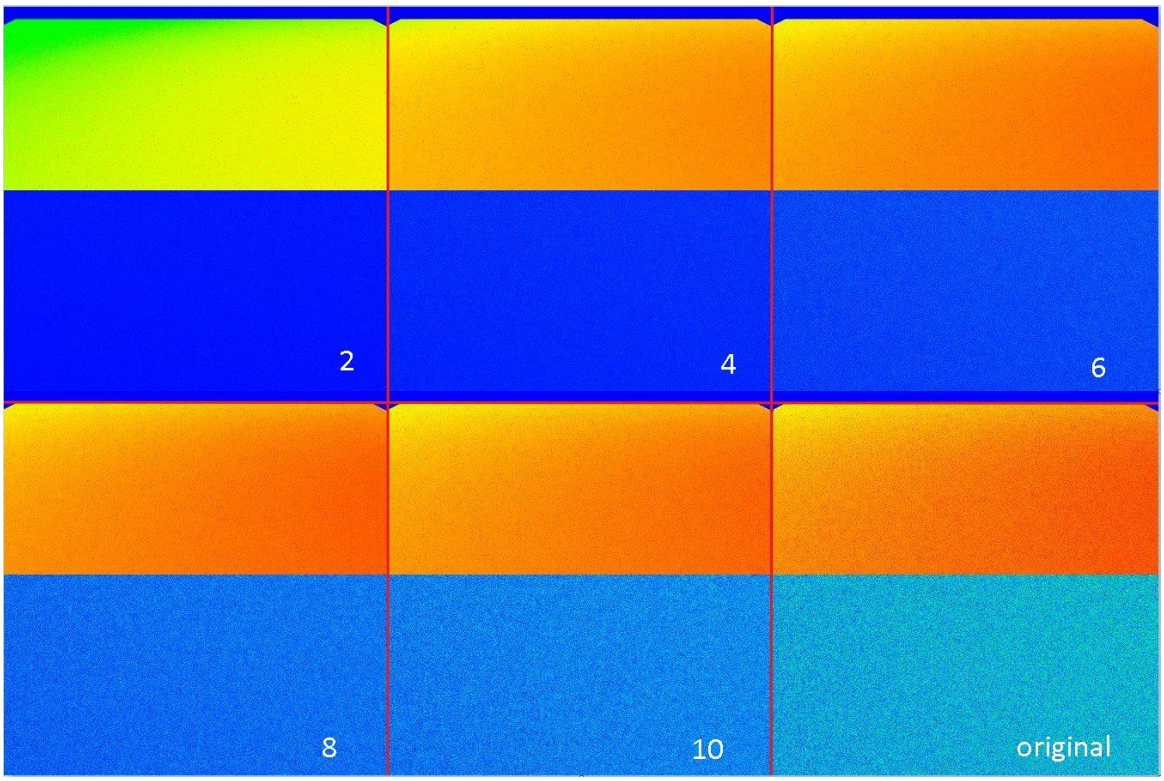


Рис. 17. Тепловые карты по итерациям восстановления изначальных коэффициентов 0.01 к оригинальным значениям 0.5.

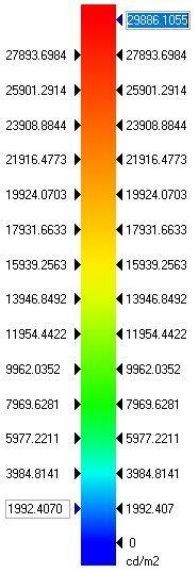


Рис. 18. Шкала значений яркостей

VII.Выводы

В результате данного исследования был предложен и реализован алгоритм расчета дифференциалов, основанный на методе трассировки лучей, который позволяет определять оптические свойства объектов в сцене. В основе предлагаемого алгоритма лежит метод дифференцируемого рендера. Алгоритм последовательно изменяет коэффициенты оптической модели объекта и вычисляет яркость для каждого объекта, взаимодействующего с лучом. Полученные изображения с разных ракурсов камер обрабатываются оптимизатором, который выбирает наилучший вариант изменения параметров. Кроме того, был предложен метод оптимизации расчета дифференциального рендеринга, суть которого заключается в выборе определенного количества точек на каждом объекте, проведении набора итераций трассировки лучей по этим точкам с

сохранением пути, получении значений яркости, а затем использовании их для оценки погрешности рендеринга с измененными параметрами, которые будут изменены оптимизатором.

Тестирование алгоритма на сцене Cornell Box показало, что алгоритм с небольшой погрешностью восстанавливает заданные оригинальные оптические свойства для объекта куба, однако не совсем корректно восстанавливает свойства задней стенки. Это связано с тем, что камера в сцене направлена от стенки на куб, и ее влияние на свойства куба является незначительным. В ближайшем будущем планируется использование нескольких камер для более точного восстановления оптических свойств всех объектов в сцене, а также смесь различных моделей оптических свойств с различными весами для улучшения точности.

Таким образом, разработанный алгоритм является перспективным инструментом для определения оптических свойств объектов в сцене и может быть применён в различных областях компьютерной графики и визуализации.

Список литературы

[1] Phong, B. T., "Illumination for computer generated pictures," Communications of ACM 18. Papers 6, 311-317 (1975).
[2] Burley, B., "Physically Based Shading at Disney," ACM Transactions on Graphics (ACM SIGGRAPH). Papers 2012, 1-7 (2012).
[3] Cook, R. L., Torrance, K. E., "A Reflectance Model for Computer Graphics," ACM Transactions on Graphics. Papers 1(3), 301-316 (1981).
[4] Hughes, J.F., "Computer Graphics: Principles and Practice, Third Edition, "Addison-Wesley Professional, 1262 (2013).
[5] Kajiya, J.T., "The rendering equation, " ACM SIGGRAPH Computer Graphics. Papers 20(4), 143-150 (1986).
[6] Veach, E., "Robust monte carlo methods for light transport simulation," Ph.D. Dissertation, Stanford University, http://graphics.stanford.edu/papers/veach_thesis/, 406 (1998).
[7] Bogolepov, D. K., Ulyanov, D., "GPU-Optimized Bidirectional Path Tracing," In Proc. of the 21th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, 1-15 (2013).
[8] Veach, E., Guibas, L.J., "Metropolis Light Transport, " In Proc. of the of the 24th Annual Conference on Computer Graphics and Interactive Techniques, 65-76 (1997).
[9] Bitterli, B., Jakob, W., Novak, J., and Jarosz W., "Reversible Jump Metropolis Light Transport Using Inverse Mappings," ACM Transactions on Graphics. Papers 37(1), 1-12 (2017).
[10]Gruson, A., West, R., Hachisuka, T., "Stratified Markov Chain Monte Carlo Light Transport," Computer Graphics Forum. Papers 39(2), 351-362 (2020).
[11] Jensen, H.W., "Global illumination using photon maps," Eurographics workshop on Rendering techniques. – Springer, Vienna, 21-30 (1996).
[12] Li, T., Aittala, M., Durand, F. and Lehtinen J., "Differentiable Monte Carlo Ray Tracing through Edge Sampling, " ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia). Papers 37(6), 222:1-222:11 (2018).
[13] Zhang, C., Miller, B., Yan, K., Gkioulekas, I. and Zhao, S., "Path-Space Differentiable Rendering, " ACM Transactions on Graphics (SIGGRAPH 2020). Papers 39(4), 143:1-143:19 (2020).
[14] Stam, J., "Computing Light Transport Gradients using the Adjoint Method, " arXiv preprint arXiv:2006.15059, https://doi.org/10.48550/arXiv.2006.15059, 23 (2020).
[15] Копылов М.С., Бирюков Е.Д., Барладян Б.Х., “Особенности использования сценариев в автоматическом тестировании систем оптического моделирования”, GraphiCon(2018), https://www.graphicon.ru/html/2018/papers/484-487.pdf
[16] И.Л. Ковалева “ПОЛУЧЕНИЕ И ОБРАБОТКА ИЗОБРАЖЕНИЙ”, Минск БНТУ(2008), https://rep.bntu.by/bitstream/handle/data/1411/Poluchenie_i_obrabotka_izobrazhenij.pdf?sequence=1&isAllowed=y

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-194-205

Сравнительный анализ методов трассировки лучей функций расстояний со знаком
Comparative study of ray tracing for signed distance function

А.Р. Гарифуллин (2), В.А. Фролов (1)
[1] Институт перспективных исследований проблем искусственного интеллекта и интеллектуальных систем МГУ имени М.В. Ломоносова, 119191 Москва, Ломоносовский пр-т, д.27, к.1.
[2] ИПМ им. М.В. Келдыша РАН, Миусская пл., д.4, Москва, 125047, Россия

A. R. Garifullin(2), V. A. Frolov(1)
[1] IAI Moscow State University, GSP-1, Lomonosovsky Prospect, 1, 27, Moscow, 119191, Russia
[2]Keldysh Institute of Applied Mathematics, Miusskaya sq., 4, Moscow, 125047, Russia

Аннотация. В работе проведён многокритериальный сравнительный анализ методов трассировки лучей для функций расстояний со знаком. В качестве основных критериев были выбраны: время работы алгоритма рендеринга; объём занимаемой памяти; точность представления поверхности, оцениваемая по рендеру при помощи метрики PSNR. Проанализировано 6 различных представлений и 4 алгоритма поиска пересечений. В качестве ускоряющей структуры во всех случаях использовалась иерархия ограничивающих объёмов (BVH деревья). Проведённое сравнение выявило перспективные представления и алгоритмы и показало, что функции расстояний в ряде случаев практически не уступают полигональным моделям по скорости, хотя при этом могут выигрывать по объёму потребляемой памяти и представлять поверхность с хорошим уровнем точности.

***Ключевые слова*:** компьютерная графика, трассировка лучей, функции дистанции со знаком

Аннотация. The paper provides a multi-criteria comparative analysis of ray tracing methods for signed distance functions. The following criteria were selected: rendering speed; the amount of memory used; the accuracy of the surface representation, measured by the PSNR metric. 6 different representations and 4 intersection search algorithms were analyzed. In all cases, the hierarchy of bounding volumes (BVH trees) was used as an accelerating structure. The comparison revealed promising representations and algorithms and showed that distance functions in some cases are practically equal to polygonal models in terms of speed, although they can outperform them in terms of memory consumption and represent a surface with a good level of accuracy.

***Ключевые слова*:** computer graphics, ray tracing, signed distance functions

I. Введение

Трассировка лучей в настоящее время становится одним из основных методов визуализации 3D моделей, особенно с учётом появления её аппаратного ускорения в современных графических процессорах. Одно из ключевых преимуществ трассировки лучей по сравнению с традиционно-используемым методом растеризации — универсальность, благодаря которой поверхность может быть представлена различными способами: аналитически-заданные поверхности, CSG-иерархия, полигональные сетки, поверхности второго порядка, сплайны и поверхности Безье, нейронные представления, функции расстояний и др. Однако рендеринг трассировкой лучей многих из этих представлений, безусловно обладающих рядом иных достоинств, оказывается вычислительно затратным. Сюда можно отнести, например, NURBS-поверхности (от англ. Non-Uniform-Beta-Splines [1]) или различные нейронные представления [2] [3], примерно на порядок отстающие от полигональных сеток по скорости трассировки. Полигональные сетки (особенно полученные в результате сканирования), с другой стороны, зачастую потребляют чрезмерно много памяти и не всегда подходят для задачи. Например, реконструкция/оптимизация поверхности состоящей из полигональных сеток в обратном рендеринге требует применения довольно сложных подходов регуляризации [4] [5], а исходное представление поверхности в 3D редакторе может быть задано в виде NURBS.

Функция расстояния со знаком (SDF, от англ. Signed Distance Function) – это такая функция, которая позволят вычислить расстояние от произвольной точки в 3D пространстве до поверхности: $f : R^3 \rightarrow R$. Нас интересуют функции расстояний по двум причинам: во-первых они могут быть использованы в качестве универсальных

аппроксимации других способов задания поверхностей, и при этом достигать хорошей точности. Во-вторых, функции расстояний по своей природе являются дифференцируемым представлением. Это позволяет использовать их в различных задачах как представление первого класса, т.е. непосредственно используемое и не подвергаемое каким-либо преобразованиям, например: для рендеринга; для оптимизации поверхности по изображениям [6] или облакам точек [7]; для решения задачи обнаружения столкновений. Кроме того, особенность функций расстояний в том, что в определённых случаях работа с ними может быть ограничена практически только лишь производительностью арифметических операций на текущем вычислительном устройстве с минимальной нагрузкой на подсистему памяти (например сюда можно отнести фрактальные поверхности и некоторые нейронные представления).

II. Обзор существующих работ

1. *Представления SDF-функций и другие представления поверхностей*

Рендеринг объектов, представленных функцией расстояния со знаком (Signed Distance Functions, SDF) впервые был предложен Джоном Хартом (John C. Hart) в 1996 году в работе [8]. Предложенный в работе метод трассировки лучей, называемый Sphere Tracing (он же Distance Aided Ray Marching), в силу своей простоты и относительной эффективности широко используется и в наши дни. В качестве основного способа задания таких функций долгое время использовался аналитическо-процедурный способ, представляющий функцию расстояния в виде набора аналитически-заданных примитивов и операций над ними [9]. Хотя таким способом можно задавать достаточно сложные, в том числе фрактальные поверхности [10] [11], его затруднительно использовать для представления произвольных сложных объектов, что ограничивает сферу применения. В нашей работе мы исследуем универсальные способы задани поверхностей в виде SDF-функций, которые позволяют, пусть и с некоторой долей апроксимации, задавать произвольную поверхность.

Далее по тексту для всех представлений, которые мы реализовывали/использовали указано сокращённое название на английском языке в круглых скобках следующим образом: **Представление (название)**. Рис. 1 наглядно демонстрирует 5 различных представлений SDF (hp_octree по своей структуре совпадает с frame_octree, за исключением того, что в каждом узле сохраняются не значения дистанций, а коэффициенты полинома).

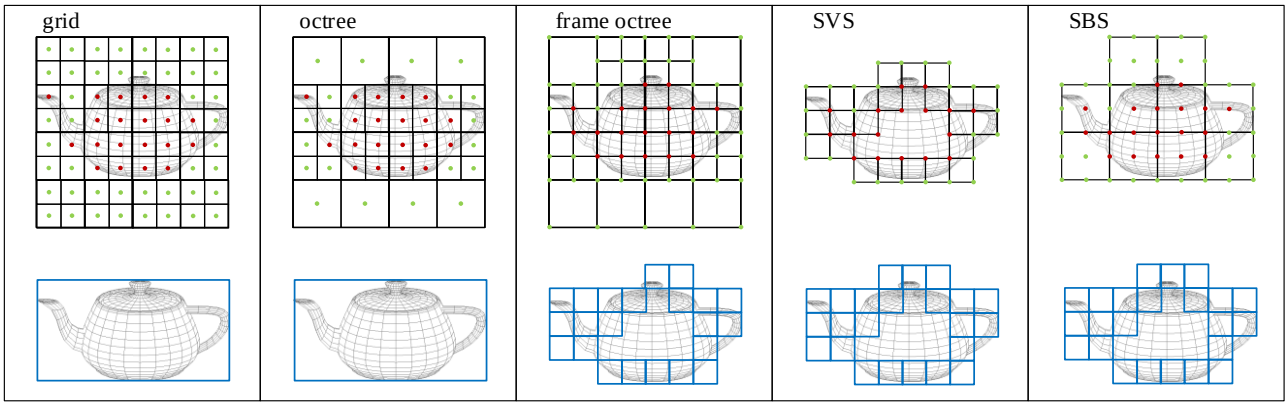


Рис. 1. Схематическое изображение рассматриваемых в работе представлений SDF. Слева направо: регулярная сетка (grid), октодерево (octree), октодерево с сохранением значений в углах (frame_octree), sparse voxel set (SVS), sparse brick set (SBS). Верхние схемы демонстрируют разбиение пространства для каждого из представлений. На нижних схемах показаны листовые узлы BVH для каждого из представлений. Зеленые и красные точки – позиции, SDF из которых сохраняется в данном представлении. Зелёные — снаружи объекта, красные — внутри. На нижних схемах показаны непустые (синие) листовые узлы BVH для каждого из представлений.

Регулярная сетка (sdf_grid). Самое простое что можно сделать – запомнить расстояния до поверхности в узлах трёхмерной регулярной сетки, образующей трёхмерный массив вокселов. Несмотря на очевидный недостаток такого подхода (низкая точность и высокий расход памяти), он иногда используется из-за простоты и низкой вычислительной стоимости. Например, в относительно недавней работе [6] для реконструкции 3D объектов по изображениям.

Октодерево (sdf_octree). Очевидным развитием регулярной сетки является окто-дерево, которое адаптивно разбивает пространство с большей детализацией вблизи поверхности, и сохряет расстояние в узлах [12]. Этот метод, как и его дальнейшие развития, исходят из предположения, что для многих задач (в том числе и для

рендеринга) не требуется точного вычисления расстояния до поверхности в каждой точки 3D пространства; достаточно сохранить лишь значение, которое будет не больше истинного расстояния и иметь правильный знак, в методе sphere tracing и его альтернативых, рассматриваемых в работе, это гарантирует, что луч не промахнётся (либо столкновения с целевым объектом не произойдёт).

Октодерево с сохранением в значений в углах (sdf_frame_octree). Сохраняет 8 дистанции в углах воксела или объема, образованного несколькими узлами окто-дерева. Такой подход позволяет более точно вычислять расстояния и значительно повышает локальность досткпа к памяти, однако требует значительного объема памяти, поскольку значения в углах дублируются между всеми вокселями, разделяющими этот угол.

Октодерево с полиномами (sdf_hp_octree) [14]. Этот подход сохраняет в узлах окто-дерева не сами расстояния, а коэффициенты полинома Лежандра, что позволяеть повысить точность и сократить потребление памяти. Поскольку реализация данной работы доступна на github в виде исходных кодов, мы использовали её оригинальный исходный код, добавив его в наше программное окружение, и перенесли его на GPU тем же способом, которым мы переносили остальные алгоритмы (с.м. раздел «реализация на графических процессорах»).

Sparse Voxel Set (sdf_SVS). Этот подход можно рассматривать как развитие frame_octree, адаптированное именно под задачу рендеринга. В отличие от ситуаций, когда нам необходимо узнавать значение функции дистанции в произвольной точке, все, что интересует нас в процессе рендеринга, это нахождение пересечения луча с поверхностью, заданной SDF, что может произойти только в тех узлах октодерева, которые содержат границу объекта. В SVS каждый воксел по-прежнему сохраняет 8 дистанций в углах, но помимо того, над множеством вокселей строится BVH дерево, что позволяет эффективно, без обхода самого октодерева обходить вокселы во время трассировки лучей, в том числе с использованием аппаратного ускорения [15]. Поскольку для каждого воксела хранится ограничивающий прямоугольный паралелипипед, значения внутри вокселей могут задавать расстояние относительно размера воксела, что позволяет использовать более компактное представление (8-битное в работе [15]).

Sparse Brick Set (sdf_SBS). Это аналогичный SVS подход, отличающийся тем, что в вокселах сохраняется небольшая регулярная сетка (2x2x2 или 4x4x4), что позволяет несколько уменьшить дублирование значений дистанции в сравнении с SVS, но требует большего количества вычислений при поиске пересечения. В данной работе рассматривались SBS с разным размером воксела и разным количеством байт на значения дистанции. Напрмер, название sdf_SBS-2-1 обозначает SBS с сеткой 2x2x2, где на каждое из значений выделяется 2 байта.

Нейронные представления. Эти представления следует разделить на два класса. Первый класс методов – это нейронные SDF, использующие нейро-сетевое представление для вычисления функций расстояния. Сюда можно отнести работы [2] и [3]. Второй класс методов – это недавно появившиеся нейросетевые методы трассировки лучей: [17], [18]. В этих подходах нейросеть непосредственно учат отвечать за запрос трассировки лучей, заменяя целые срезы BVH дерева нейронными представлениями. Мы не использовали эти подходы в нашем сравнении по причине их низкой производительности.

Нейронные SDF требуют значительного объёма вычислений. По сравнению с трассировкой лучей для полигонального представления этот объём даёт замедление примерно в 1000 раз. Что касается второго класса методов, здесь замедление уже всего в 5-6 раз по сравнению с трассировкой полигональных моделей. Однако в работе [17] это замедление измерено без учёта возможного аппаратного ускорения для обхода BVH дерева, что даст полигональным моделям дополнительный выигрыш ещё примерно в 5-10 раз с итоговым куммулятивным эффектом в 25-50 раз. В работе [18] при этом нейронная трассировка использовалась только для вторичных лучей из-за низкой точности получаемого результата пересечения луча и поверхности. Ускорение в 25%, полученное при этом благодаря предсказуемому и детерминированному паттерну обращения в память, на наш взгляд не окупает потерю в точности в данном подходе.

Близкой к нейронным представлениям группой методов являются работы, основанные на сжатии или декомпозиции 3D воксельной модели, например хеширование вокселей [25], вейвлет-преобразование [26] или Tensor Train декомпозицию [27]. Подобные методы способны существенно сократить объем памяти, занимаемый моделью, однако как и нейросетевые методы, плохо подходят для визуализации в реальном времени.

2. Алгоритмы поиска пересечений луча и SDF функции

Sphere Tracing (sphere_tracing) – итеративный алгоритм для поиска пересечения луча с поверхностью, заданной функцией дистанции. Разработанный Джоном Хартом в 1996 году [8], он является очень простым, но широко распространенным алгоритмом для решения данной задачи. Находясь в произвольной точке 3D

пространства он вычисляет дистинцию до поверхности, после чего шагает на эту дистанцию в направлении луча (именно поэтому его ещё называют Distance-Aided Ray Marching). Такой подход плохо работает в ситуации, когда луч проходит близко к поверхности, с которой не пересекается, поскольку он делает множество маленьких шагов. Более оптимальные алгоритмы поиска пересечения требуют некоторого априорного знания об устройстве SDF, а более точно – о проекции функции дистанции на луч. Это возможно, когда SDF задана аналитически (например, комбинацией примитивов конструктивной сплошной геометрии) или некоторыми вариантами октодеревьев с трилинейной или трикубической интерполяцией значений.

Interval Tracing (interval_tracing). Этот метод был предложен для примитивов конструктивной сплошной геометрии в 2020 году в работе [19]. Метод основан на оценке константы Липшица на некотором интервале вдоль луча. Константа Липшица — это числовая характеристика т. н. липшицевой непрерывности функции. Функция $f(x)$ называется липшицевой с константой Липшица K, если для всех x и y из её области определения выполняется неравенство: $|f(x)-f(y)| \leq K|x-y|$. Это условие означает, что функция не изменяется слишком быстро: разница значений функции $f(x)$ и $f(y)$ ограничена сверху константой K, умноженной на расстояние между x и y. Липшицевость — это более строгое условие, чем просто непрерывность, но менее строгое, чем условие дифференцируемости. По сравнению со Sphere Tracing, в работе [19] сообщают об ускорении до 10 раз на некоторых сценах за счет безопасного увеличения длины шага в направлении луча в сравнении со sphere tracing.

Аналитический метод (analytic). Метод, предложенный в работе [15] и предполагающий, что у нас есть аналитическое представление для функции $f(t)$, полученной проекцией SDF на луч и мы можем явно найти корни уравнения $f(t) = 0$. Этот метод вообще не требует итеративного приближения к поверхности и возволяет сразу найти точку пересечения, но накладывает очень жесткие требования на устройство исходной функции дистанции. В нашей работе, как и в [15], функция дистанции рассчитывается как линейная комбинация значений дистанции в 8 углах вокселя, и было показано, что в таком случае искомая функция будет многочленом степени не выше 3, что допускает аналитическое решение.

Метод Ньютона (newton). Как и следует из названия, этот метод предполагает численное нахождение корней уравнения $f(t) = 0$ методом Ньютона. Этот метод был впервые использован для трассировки лучей в компьютерной графике в работе [20] в 2004 году. Он также рассматривается в работе [15], где называется методом Ньютона-Рапсона. Этот метод также требует аналитического представления для функции $f(t)$, но применим к более широкому классу функций, чем аналитический (например, к полиномам степени 5 и выше, для которых не гарантируется наличие аналитического решения). Данный метод может также использоваться совместно с иным, более грубым приближением поверхности (например, грубой полигональной сеткой [24]) для ускорения работы со сложными аналитически заданными поверхностями.

3. Мотивация нашей работы

Наиболее близким аналогом нашей работы является статья сотрудников компании Nvidia [15] 2022 года. В этой работе помимо анализа способов представления функциий расстояния и алгоритмов пересечения предложен оптимизированный алгоритм вычисления пересечения между лучом и поверхностью, реализуемый через трилинейную интерполяцию функции в восьми углах вокселя --- аналитический метод. Обход BVH дерева осуществлялся аппаратно с использованием RT ядер видеокарты NVIDIA RTX 3090. Эта работа является отправной точкой для нашего исследования. Тем не менее есть ряд вопросов, на которые она не даёт ответа, а именно:

1. Какова точность представления поверхности у выбранных представлений для SDF по сравнению с некоторым эталоном? Ни численные оценки, ни детальные сравнения с эталоном не представлены в работе [15].
2. Каков баланс скорость/качество для SDF функций по сравнению с полигональными представлениями?
3. Насколько компактным может быть представление для определённого уровня точности? В [15] подходы сравниваются между собой по потреблению памяти для фиксированной глубины дерева, что в целом верно, но оставляет открытым вопрос о том, что можно всегда сделать дерево менее глубоким.
4. Какая взаимосвязь между этими критерими? Очевидно, что разная глубина дерева / разный размер вокселей напрямую влияет на все три критерия – точность, скорость, потребление памяти.
5. Не совсем понятно чем обусловлен выбор тестовых моделей в работе [15]. Они не являются ни общепринятыми, ни обще-доступным, а их SDF функции не известны. Это делает результаты работы [15] менее воспроизводимыми, и в большей степени эмпирическими.
6. Можно ли повысить точность, если хранить в вокселах не просто расстояния, а что-то более сложное, наподобие работы [14].

7. Сравнение по скорости трассировки лучей произведено только в комплексе, для алгоритма трассировки путей внутри дростаточно большой и сложной рендер-системсы Falcor. Такое сравнение не показательно, поскольку на производительность трассировщика путей, тем более с аппаратным ускорением и использованием т.н. конвейера трассировки лучей в современных GPU влияет множество факторов. В нашей работе помимо трассировки путей нам хотелось сравнить производительность в более чистом эксперименте.

Наконец, исходный код работы [15] не был ни опубликован в открытом виде, ни доступен в виде какой-либо закрытой библиотеки. По этой причине актуальность в реализации этих метов по прежнему высока.

III. Описания экспериментов

1. Выбор 3D моделей

Рис. 2. Некоторые из 3D моделей, использованные в экспериментах.

Мы провели сравнение на небольшом наборе из 7 различных тестовых моделей. В работе демонстрируются графики с результатами сравнений на 4 из 3D моделей, представляющих различные категории объектов, для которых потенциально полезно использование представлений в виде SDF. Результаты сравнения на всех рассмотренных моделях далее используются для получения усредненных показателей качества и производительности различных методов, что позволяет сравнить их между собой.

Модель №1. Чайник из Юты (teapot). Эта 3D модель хороша прежде всего тем, что она имеет точное аналитическое представление в существующих 3D редакторах, то есть имеется эталонное предстаавление поверхности. При этом есть возможность получить полигональное представление с хорошей сеткой и различным уровнем детализации.

Модель №2. Стэнфордский дракон (dragon). Эта 3D модель получена сканированием, и является открым и хорошо известным примером для подобного сценария. Сама полигональная модель является высоко-детализированной, поэтому её можно брать за эталон.

Модель №3. Деталь (clutch). Цель этой тестовой модели – тестирование SDF для задач CAD приложений и схожих применений.

Модель №4. Архитектура (house). Цель этой тестовой модели – проверка SDF на достаточно сложной модели, содержащей как относительно крупные плоские поверхности, так и большое количество мелких деталей. На рис. 2 изображены названные выше, а также некоторые другие использованные в эксперименте модели.

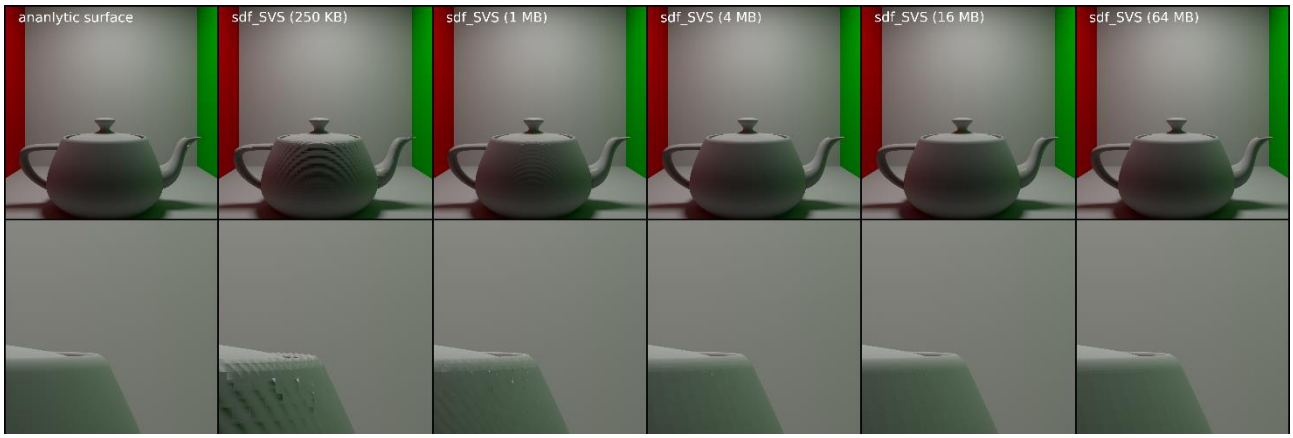


Рис. 3. Заданная сплайнами модель teapot и построенные по ней Sparse Voxel Set (sdf_SVS) разного размера.

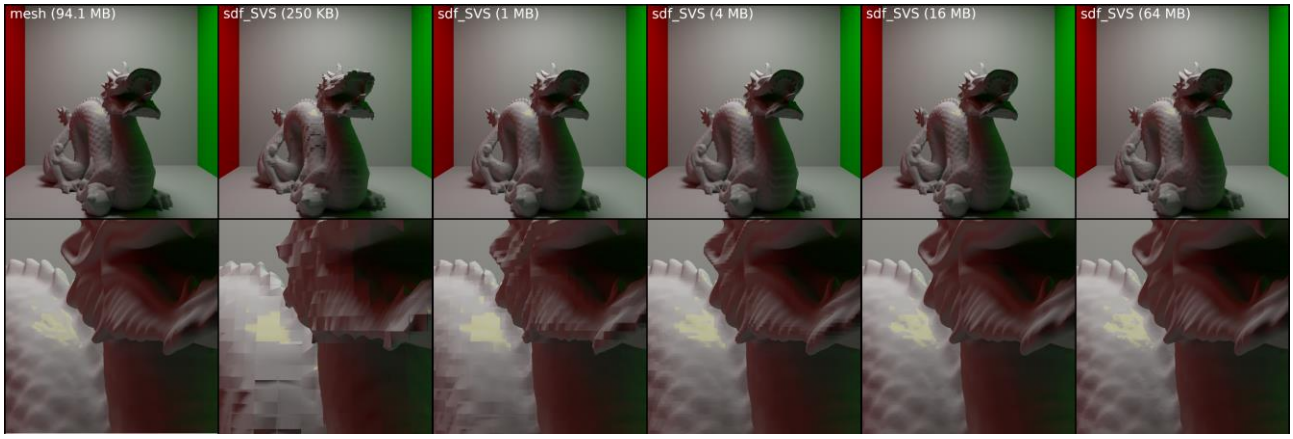


Рис. 4. Оригинальная 3D модель dragon (слева) и построенные по ней Sparse Voxel Set (sdf_SVS) разного размера.

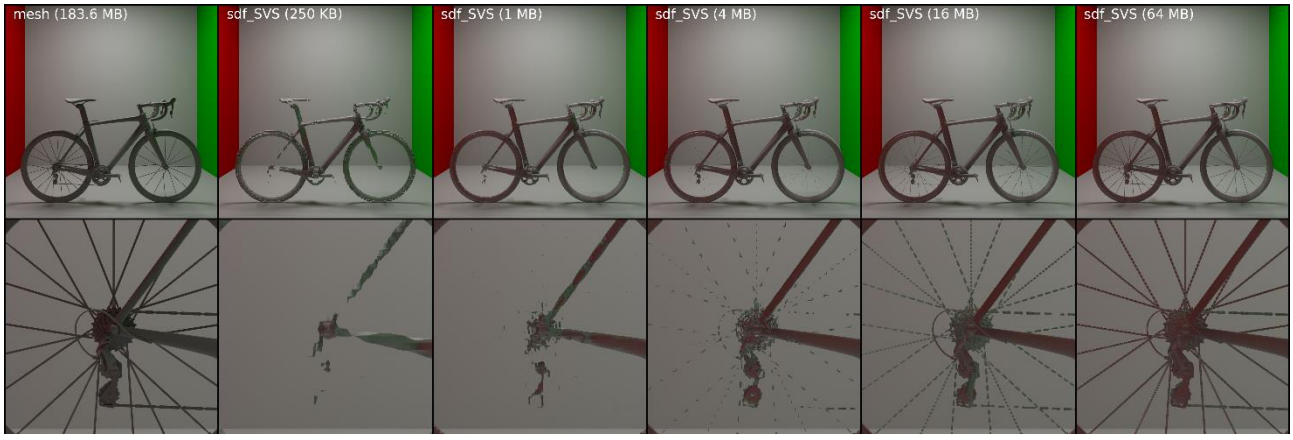


Рис. 5. 3D модель roadbike (слева) и построенные по ней Sparse Voxel Set (sdf_SVS) разного размера. Тонкие детали, такие как спицы на колесе, теряются или существенно искажаются при небольшом размере SVS.

2. Методика сравнения

Для каждой 3D модели и каждого представления мы фиксировали объём памяти, в который эту 3D модель можно было поместить в текущем представлении. Мы брали несколько порогов, от 250 Кб и до 64 Мб. Далее мы делали два эксперимента по оценке качества: (1) выполняли трассировку первичных лучей и измеряли

PSNR между двумя изображениями – рендером SDF для текущего порога по памяти и рендером эталона, исходной полигональной сеткой; (2) выполняли трассировку путей, помещая модель в известную сцену Cornell Box и тем же способом измеряли PSNR. На рис. 3-5 изображены примеры получаемых в ходе эксперимента изображений для исходных моделей и представления SVS. Для обоих случаев мы измеряли время выполнения алгоритма. Кроме того, для первичных лучей мы отдельно оценивали время выполнения трассировки внутри листовых узлов BVH дерева путём вычитания из полного времени трассировки время спуска лучей по BVH дереву до первого листа. Это позволяет нам более точно сравнить различные алгоритмы и представления друг с другом без учёта накладных расходов на обход BVH дерева. Дополнительно этот приём даёт нам оценку производительности для случая, когда обход BVH дерева ускорен на графических процессорах аппаратно. В нашей реализации обход BVH дерева выполнялся программно классическим алгоритмом со стеком. Было использовано BVH дерево с 2 листами; для построения дерева был использован вариант SBVH [21] из библиотеки embree3 [22].

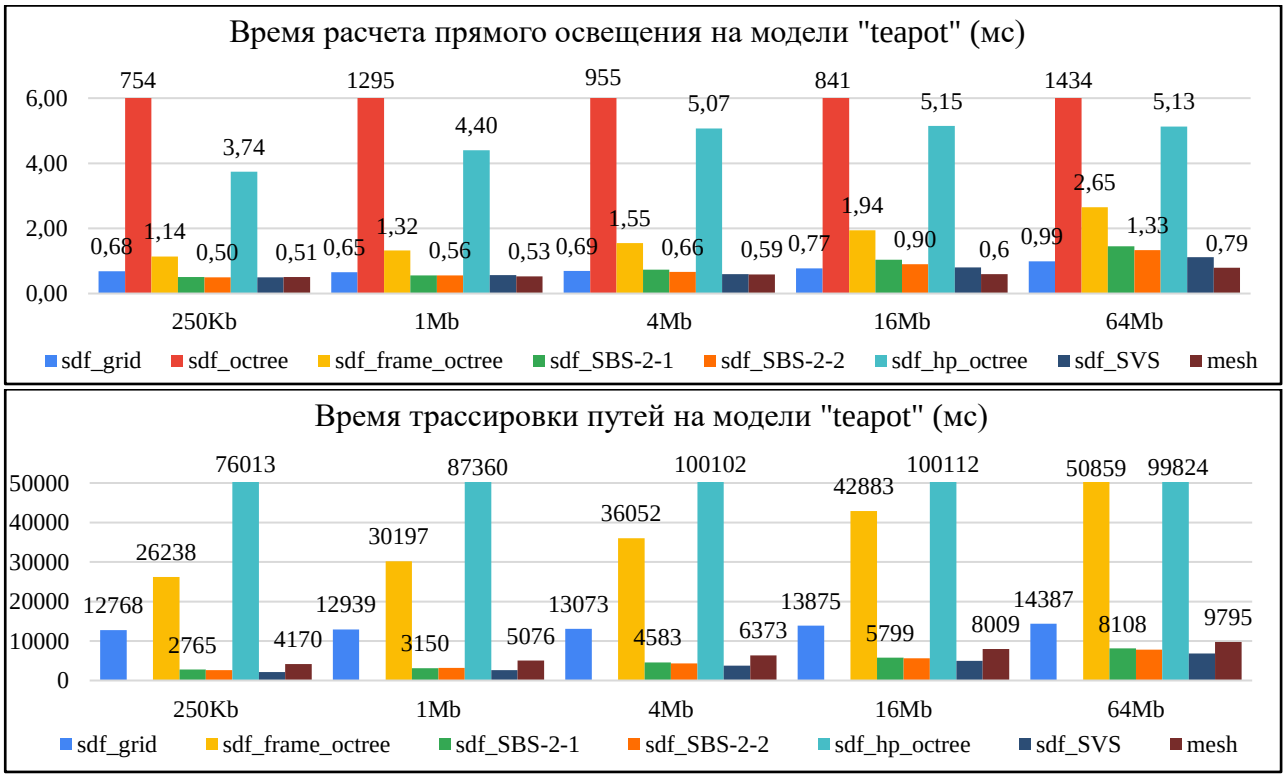
Для каждой 3D модели мы рендерили 10 ракурсов, поворачивая модель в центре Cornell Box вокруг своей оси. Значение метрики PSNR мы усредняли по всем 10 ракурсам. Для повышения точности измерения времени мы выполняли 100 запусков первичной трассировки лучей для каждого ракурса и усредняли их результаты. Для трассировки путей мы измеряли время 1 раз и усредняли его для всех 10 ракурсов.

3. *Перенос на графические процессоры*

Мы реализовывали все алгоритмы на C++ и параллельную реализацию на CPU при помощи OpenMP. Для переноса алгоритмов на GPU мы использовали технологию программирования kernel_slicer [23], осуществляющую автоматический перенос исходного кода с C++ на Vulkan [13]. Она же позволила нам сравнить производительность нашей программной реализации со скоростью аппаратной трассирвки лучей для полигональных моделей на современных GPU.

Полученная реализация, таким образом, является кросс-платформенной и может быть запущена на различных GPU. Однако на текущий момент мы протестировали наши алгоритмы лишь на одной машине со следующей аппаратной конфигурацией: AMD Ryzen 7 3700X 3.6 GHz, 32 Gb RAM, GPU Nvidia RTX 3070 8 Gb.

IV. Результаты экспериментов



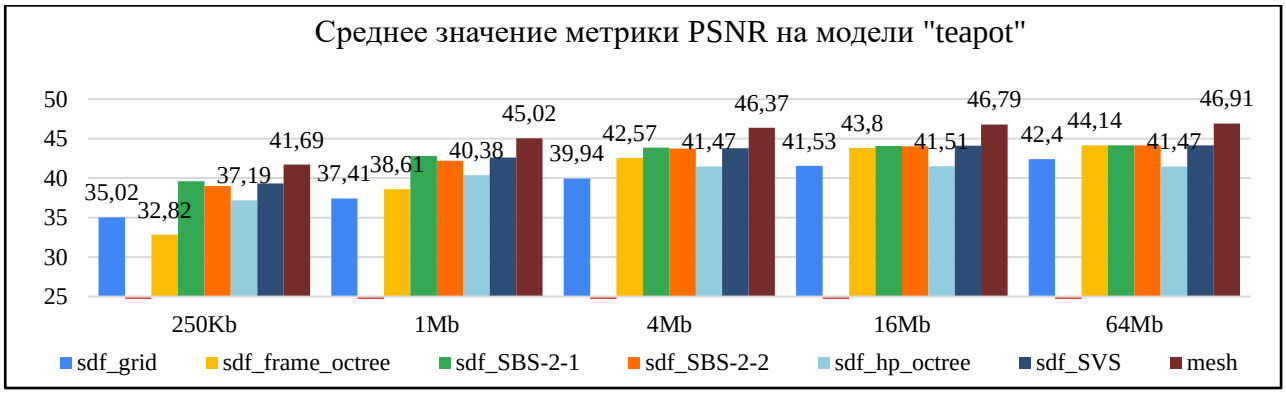


Рис. 6. Результаты экспериментов с различными представлениями модели "teapot". Помимо различных представлений SDF, в сравнение добавлены полигональные модели различной детализации, полученные из исходной параметрической поверхности.

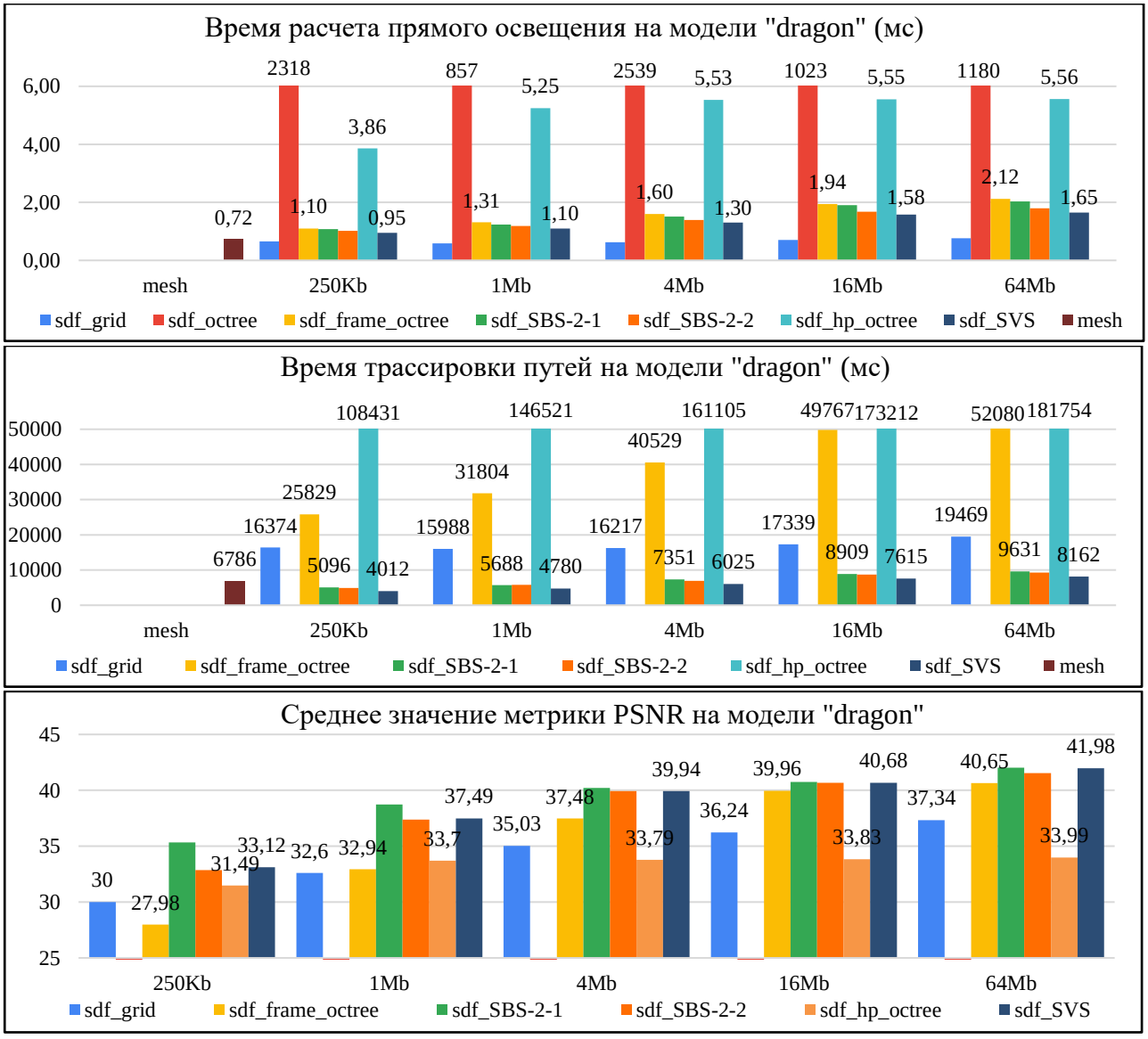


Рис. 7. Результаты экспериментов с различными представлениями модели " dragon". Размер исходной полигональной модели – 94.1 Мб.

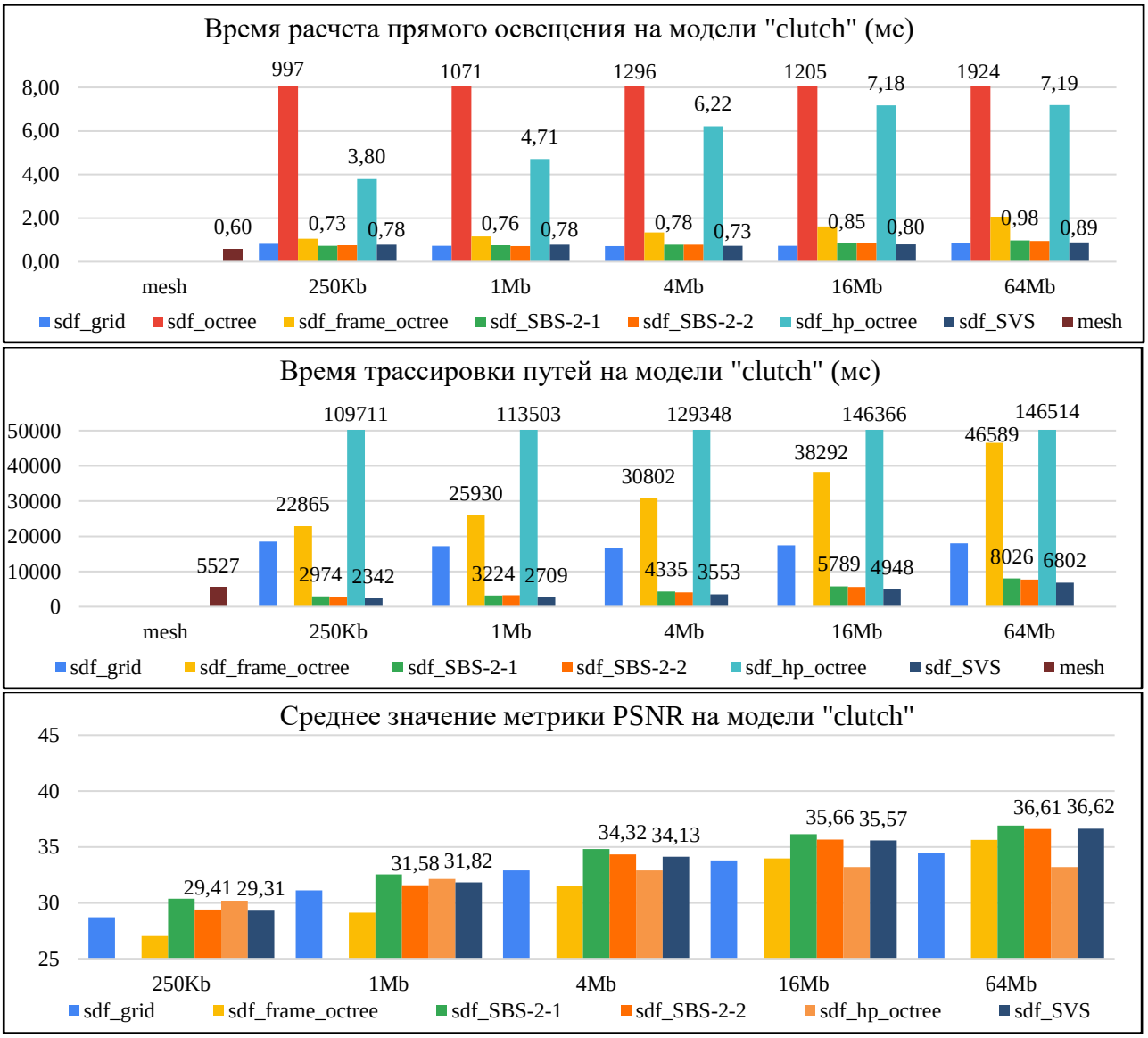
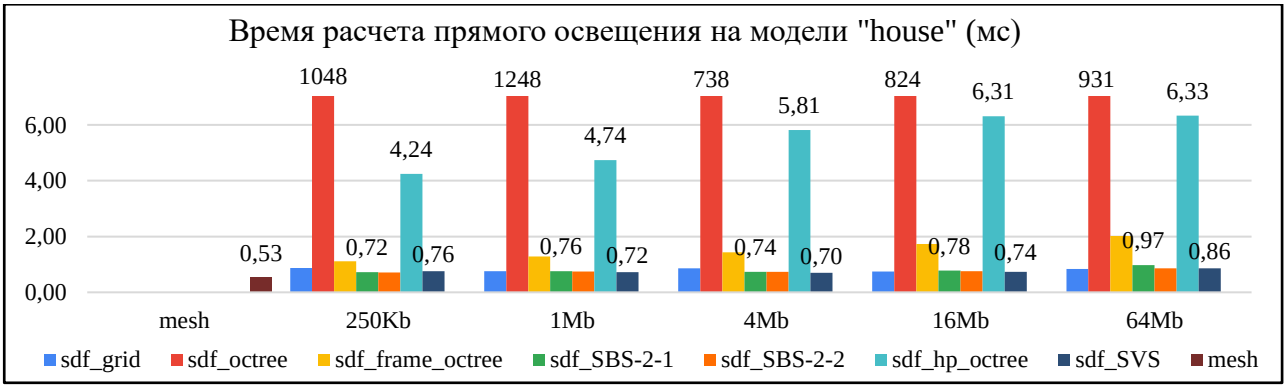


Рис. 8. Результаты экспериментов с различными представлениями модели "clutch". Размер исходной полигональной модели – 19 Мб.



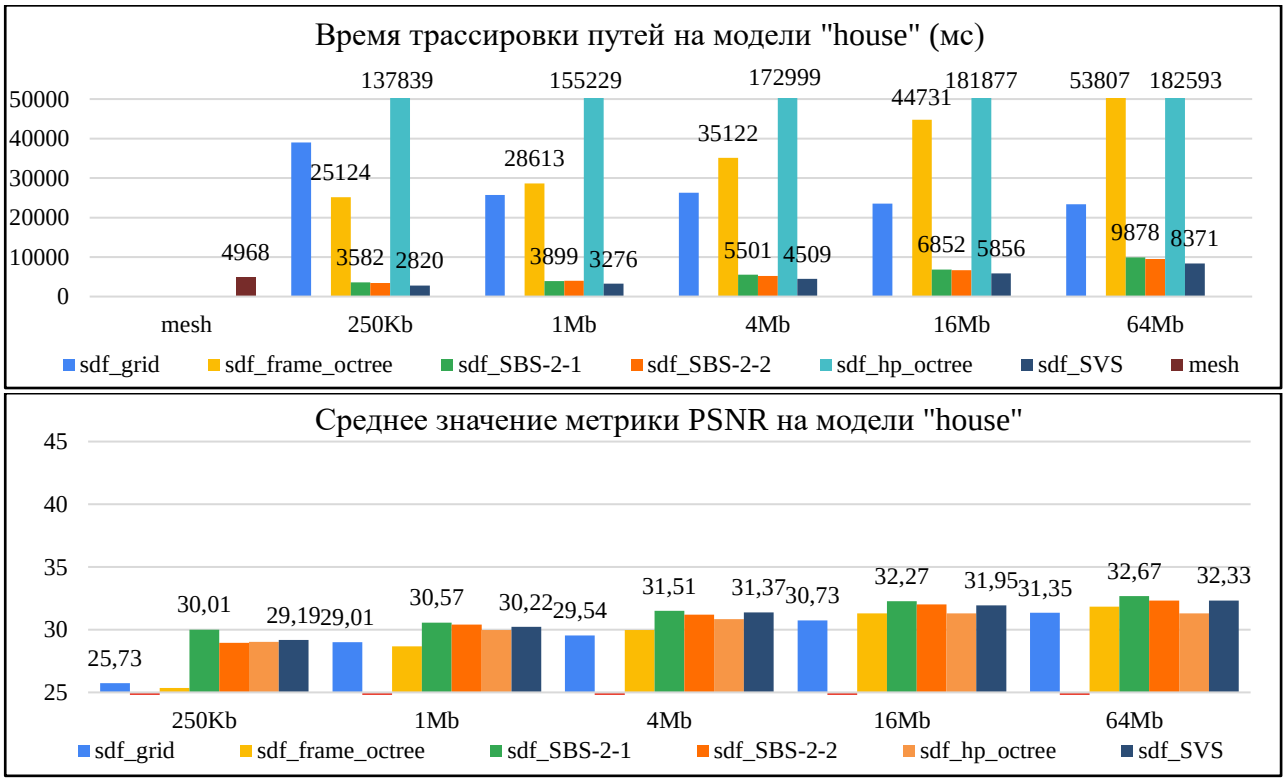


Рис. 9. Результаты экспериментов с различными представлениями модели "house". Размер исходной полигональной модели – 20.7 Мб.

На рис. 6-9 показаны результаты экспериментов (время расчета прямого освещения, трассировки путей и значение PSNR) на моделях “teapot”, “dragon”, “clutch” и “house” соответственно. Эксперименты с расчетом прямого освещения показали, что SDF octree значительно (до 1000 раз) уступает другим методам в скорости работы и поэтому он не рассматривался в экспериментах с трассировкой путей. Метод hp-octree также показал себя хуже остальных во всех экспериментах в скорости работы и по метрике PSNR. При этом для всех методов значение PSNR существенно зависит от исходной полигональной модели. Лучше всего методы справляются с гладкими, «органическими» моделями с малым числом острых углов («teapot», «dragon») и хуже – с моделями искусственных объектов, с большим числом плоских поверхностей и острых углов. В среднем на всех моделях (см. Таблицу 1) лучшие значения метрики PSNR показывает вариант sdf_SBS-2-1 (Sparse Brick Set размера 2x2x2 с точностью в 1 байт на значение дистанции), однако sdf_SVS (Sparse Voxel Set) совсем незначительно отстает от него на больших моделях (16 и 64 Мб). Это существенно, поскольку именно sdf_SVS обеспечивает наименьшее время трассировки и значительно превосходит SBS в этом показателе (Таблица 2). Учитывая достаточно хорошие значения PSNR для SVS, мы видим именно его предпочтительным представлением функции дистанции для произвольных моделей.

Таблица 3 показывает результаты экспериментов с разными методамами поиска пересечения луча и поверхности внутри вокселя. Выбор метода не влияет на качество рендера (разница в PSNR во всех наших экспериментах не превысила 0.03), поэтому главным критерием выбора является производительность. Для этого вычислялось время, потраченное непосредственно на поиск пересечения (время трассировки за вычетом времени, необходимого на обход BVH-дерева). Это время для каждого метода делилось на время, затраченное на поиск пересечения с использованием sphere tracing. Видно, что полученные значения для всех методов и размеров модели меньше 1, т.е. все альтернативные методы пересечения дают ускорение относительно базового, а лучший результат для моделей всех размеров показал аналитический метод (analytic), что подтверждает выводы работы [15]. Метод interval tracing не был представлен в [15] и, наколько нам известно, не был ранее использован для иерархических представлений SDF, подобных SVS, однако экперименты показали, что он уступает другим рассмотренным методам поиска пересечения.

ТАБЛИЦА 1

Среднее значение метрики PSNR для различных представлений SDF при трассировке путей

Представление SDF	250 Кб	1 Мб	4 Мб	16 Мб	64 Мб
sdf_grid	29.6	31.8	33.9	35.2	35.9
sdf_frame_octree	28.4	32.6	35.7	37.4	38.3
sdf_SBS-2-1	**34.2**	**36.5**	**37.8**	**38.4**	**39.0**
sdf_SBS-2-2	32.8	35.6	37.4	38.2	38.8
sdf_hp_octree	31.6	33.2	33.6	33.7	33.7
sdf_SVS	33.0	35.7	37.4	**38.3**	**38.9**

ТАБЛИЦА 2

Среднее время трассировки путей для различных представлений SDF (относительно меша)

Представление SDF	250 Кб	1 Мб	4 Мб	16 Мб	64 Мб
sdf_grid	2.22	2.67	2.71	2.74	2.84
sdf_frame_octree	3.67	4.31	5.33	6.52	7.62
sdf_SBS-2-1	0.53	0.63	0.83	1.02	1.49
sdf_SBS-2-2	0.51	0.60	0.81	0.99	1.47
sdf_hp_octree	17.13	20.19	21.63	22.67	22.77
sdf_SVS	**0.44**	**0.52**	**0.70**	**0.93**	**1.21**

ТАБЛИЦА 3

Среднее время поиска пересечения луча с поверхностью (относительно sphere tracing) без учета времени обхода BVH для представления sdf_SVS.

Алгорим пересечения	250 Кб	1 Мб	4 Мб	16 Мб	64 Мб
sphere_tracing	1	1	1	1	1
analytic	**0,515**	**0,581**	**0,673**	**0,733**	**0,788**
newton	0,550	0,610	0,699	0,756	0,810
interval_tracing	0,627	0,680	0,756	0,806	0,859

V. Выводы и заключение

Анализ данных, полученных в ходе экспериментов, позволяет сделать несколько важных выводов относительно использования функций дистанции со знаком как представления объектов в рендере трассировкой лучей:

1. SDF могут выступать компактным и эффективным представлением для сложных сцен как альтернатива полигональным мешам.
2. Хорошие результаты показывают те представления, которые были разработы специально для рендера и в полной мере используют BVH. Методы octree и hp_octree, успешно используемые для представления функции дистанции в задачах, не связанных с рендером (физические симуляции, детекция столкновений), оказались значительно менее эффективными.
3. Подтверждается вывод [15] о том, что аналитический метода и метод Ньютона для поиска пересечения с поверхностью ускоряют рендер в сравнении со sphere tracing.
4. Использование interval tracing ускоряет рендер в сравнении со sphere tracing (но ускорение значительно меньше, чем в [19], где рассматривались аналитически заданные функции дистанции геометрических примитивов), но его результаты хуже, чем у аналитического метода.
5. Наилучшей по совокупности рассматриваемых факторов, является комбинация SVS - analytic. Она превосходит остальные по скорости рендера на всех размерах моделей и очень незначительно отстает от SBS в качестве представления.

Список литературы

1. Rogers D. F. An introduction to NURBS: with historical perspective. – Elsevier, 2000.

2. Sitzmann V. et al. Implicit neural representations with periodic activation functions //Advances in neural information processing systems. – 2020. – Т. 33. – С. 7462-7473.

3. Takikawa T. et al. Neural geometric level of detail: Real-time rendering with implicit 3d shapes //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2021. – С. 11358-11367.

4. Luan F. et al. Unified shape and svbrdf recovery using differentiable monte carlo rendering //Computer Graphics Forum. – 2021. – Т. 40. – №. 4. – С. 101-113.

5. Nicolet B., Jacobson A., Jakob W. Large steps in inverse rendering of geometry //ACM Transactions on Graphics (TOG). – 2021. – Т. 40. – №. 6. – С. 1-13.

6. Vicini D., Speierer S., Jakob W. Differentiable signed distance function rendering //ACM Transactions on Graphics (TOG). – 2022. – Т. 41. – №. 4. – С. 1-18.

7. Cao Y., Li H. DiffSDF: Learning Implicit Surface from Noisy Point Clouds //2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA). – IEEE, 2023. – С. 197-204.

8. Hart J. C. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces //The Visual Computer. – 1996. – Т. 12. – №. 10. – С. 527-545.

9. Inigo Quilez. Distance Functions. WEB publication. URL = https://iquilezles.org/articles/distfunctions/ (accesed 12.06.2024)

10. Ingo Quilez. Distance Functions. WEB publication. URL = https://iquilezles.org/articles/terrainmarching/ (accesed 12.06.2024)

11.Darren Polkowski. San Base: Авангард компьютерной графики. WEB publication. URL = https://www.youtube.com/@lashenko (accesed 12.06.2024)

12. Frisken S. F. et al. Adaptively sampled distance fields: A general representation of shape for computer graphics //Proceedings of the 27th annual conference on Computer graphics and interactive techniques. – 2000. – С. 249-254.

13. Sellers G., Kessenich J. Vulkan programming guide: The official guide to learning vulkan. – Addison-Wesley Professional, 2016.

14. Koschier D. et al. An hp-adaptive discretization algorithm for signed distance field generation //IEEE transactions on visualization and computer graphics. – 2017. – Т. 23. – №. 10. – С. 2208-2221.

15. Söderlund H. H., Evans A., Akenine-Möller T. Ray Tracing of Signed Distance Function Grids //Journal of Computer Graphics Techniques Vol. – 2022. – Т. 11. – №. 3.

16. Fujimoto A., Iwata K. Accelerated ray tracing //Computer Graphics: Visual Technology and Art Proceedings of Computer Graphics Tokyo′ 85. – Tokyo : Springer Japan, 1985. – С. 41-65.

17. Weier P. et al. N-BVH: Neural ray queries with bounding volume hierarchies //arXiv preprint arXiv:2405.16237. – 2024.

18. Fujieda S., Kao C. C., Harada T. Neural Intersection Function //arXiv preprint arXiv:2306.07191. – 2023.

19. Galin E. et al. Segment tracing using local lipschitz bounds //Computer Graphics Forum. – 2020. – Т. 39. – №. 2. – С. 545-554.

20. Marmitt G. et al. Fast and Accurate Ray-Voxel Intersection Techniques for Iso-Surface Ray Tracing //VMV. – 2004. – Т. 4. – С. 429-435.

21. Friedrich H., Dietrich A., Stich M. Spatial Splits in Bounding Volume Hierarchies. 2009.

22. Wald I. et al. Embree: a kernel framework for efficient CPU ray tracing //ACM Transactions on Graphics (TOG). – 2014. – Т. 33. – №. 4. – С. 1-8.

23. Frolov V., Sanzharov V., Galaktionov V. kernel_slicer: high-level approach on top of GPU programming API //2022 Ivannikov Ispras Open Conference (ISPRAS). – IEEE, 2022. – С. 11-17.

24. Жданов Д. Д., Потемин И. С., Жданов А. Д. Использование технологий Embree для трассировки лучей в оптических системах с поверхностями свободной формы //Графикон-конференции по компьютерной графике и зрению. – 2023. – Т. 33. – С. 97-107.

25. M. W. Jones, “Distance field compression,” Journal of WSCG, vol. 12, no. 2, pp. 199–204, 2004

26. M. Niesner, M. Zollhofer, S. Izadi, and M. Stamminger, “Real-time 3D reconstruction at scale using voxel hashing,” ACM Transactions on Graphics, vol. 32, no. 6, 2013.

27. Boyko, Alexey I., Mikhail P. Matrosov, Ivan V. Oseledets, Dzmitry Tsetserukou, and Gonzalo Ferrer. "Tt-tsdf: Memory-efficient tsdf with low-rank tensor train decomposition." In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10116-10121. IEEE, 2020.

УДК 004.92
DOI: 10.25206/978-5-8149-3873-2-2024-206-212

Тестирование шейдеров для графической компоненты ОС, используемой на гражданских авиалайнерах

Testing of shaders for the graphics component of the OS used on civil airliners

Б.Х. Барладян, А.Г. Волобой, Л.З. Шапиро, Е.Ю. Денисов, В.А. Галактионов
Институт прикладной математики им. М.В. Келдыша РАН, Москва, Россия

B.Kh. Barladian, A.G. Voloboy, L.Z. Shapiro, E.Yu. Denisov, V.A. Galaktionov
[2]Keldysh Institute of Applied Mathematics RAS, Moscow, Russia

Аннотация. Комплексы бортового оборудования современных гражданских авиалайнеров относятся к системам, критическим с точки зрения безопасности (safety critical). Обязательным требованием для такого оборудования является его сертификация. Графическая компонента операционной системы должна использовать библиотеку OpenGL SC (SC – safety critical). Графической компоненте необходимо использовать шейдеры, но компилятор шейдеров не может быть сертифицирован. Поэтому мы предложили и реализовали подход, при котором компилятор не используется в бортовом программном обеспечении. Он создает шейдеры заранее, и они загружаются в процессе работы в виде двоичного программного объекта. При этом возникает проблема проверки правильности работы компилятора шейдеров и корректности выполнения шейдеров. В данной работе предлагается технология верификации шейдеров, используемых при визуализации различных приложений на дисплее пилота гражданского воздушного судна. Был разработан набор тестов достаточно простой для эффективного тестирования, но при этом позволяющий проверить корректность всех операций набора шейдеров, применяемых в приложениях гражданской авиации.

***Ключевые слова*:** OpenGL SC, компилятор шейдеров, тестирование шейдеров, вершинный шейдер, фрагментный шейдер

Abstract. The on-board equipment systems of modern civil airliners are classified as safety critical systems. A mandatory requirement for such equipment is its certification. The graphical component of the operating system must use OpenGL SC (SC – safety critical). The graphics component needs shaders, but the shader compiler cannot be certified. Therefore, we proposed and implemented an approach in which the shader compiler is not used in the on-board software. It creates shaders in advance and they are loaded at runtime as a binary software object. This raises the problem of checking the correct operation of the shader compiler and the correct execution of shaders. This paper proposes a technology for verifying shaders used in visualizing various applications on the display of a civil aircraft pilot. A set of tests was developed that is simple enough for effective testing, but at the same time allows to check the correctness of all operations of a set of shaders used in civil aviation applications.

***Keywords*:** OpenGL SC, shader compiler, shader testing, GPU, vertex shader, fragment shader

I. Введение

Комплексы бортового оборудования современных гражданских авиалайнеров относятся к системам, критическим с точки зрения безопасности (safety critical). Поэтому каждое разрабатываемое гражданское воздушное судно и его комплектующие должны проходить сертификацию. Программный комплекс современного гражданского самолета работает под управлением операционной системы реального времени (ОСРВ). Для сертификации программного обеспечения бортового оборудования воздушного судна существует стандарт DO-178C (в русскоязычной редакции – КТ-178С) [1], определяющий, в частности, требования к процессу разработки ПО. Эта специфика не позволяет применять готовые, известные решения для той или другой функциональности.

Одним из важных требований к графической компоненте ОСРВ является использование специальной версии графической библиотеки OpenGL SC (SC – safety critical), предназначенной для систем, критических с точки зрения безопасности. В настоящее время библиотека OpenGL SC представлена двумя версиями стандарта: 1.0.1 и 2.0. Стандарт OpenGL SC 1.0.1 является подмножеством стандарта OpenGL 1.3. Стандарты OpenGL SC 2.0 и OpenGL SC 1.0.1 значительно отличаются. Некоторые функции версии 1.0.1 не существуют в версии 2.0 и могут

быть заменены шейдерными программами. Можно сказать, что в OpenGL SC 2.0 упор делается на программируемый конвейер трехмерной графики, а в OpenGL SC 1.0.1 – на фиксированную функциональность.

Стандарт OpenGL SC 2.0 основан на стандарте OpenGL ES 2.0, но не является его подмножеством. Он разработан для графического оборудования, работающего на встроенных платформах. В нем удалены те аспекты OpenGL ES 2.0, которые не соответствуют детерминированной безопасности для критических по безопасности программных приложений, в частности, удален встроенный компилятор шейдеров. Предполагается, что вершинный (vertex) и фрагментный (fragment) шейдеры отдельно компилируются и связываются с двоичным объектом программы, который генерирует исполняемый код из скомпилированных шейдеров. При этом в стандарте не оговорен способ создания и интерфейс двоичного объекта программы. Стандартизован только интерфейс функции glProgramBinary(), которая позволяет загрузить предварительно скомпилированный двоичный объект.

Необходимость разработки сертификационного пакета для графической компоненты в соответствии с требованиями КТ-178С требует доступа ко всем ее кодам. В большинстве случаев драйверы, предоставляемые производителями GPU, не имеют открытого программного кода, поэтому они не могут быть сертифицированы для использования в авиации. Также драйвер необходимо адаптировать для работы под управлением операционной системы реального времени. Поэтому наша разработка [2] базировалась на пакете Mesa 3D [3] с открытым исходным кодом. Следует отметить, что в пакете Mesa для внутренней компиляции шейдеров существенно используется язык C++, который практически невозможно сертифицировать по авиационным нормам. Исключить этот код для стандарта 1.0.1 невозможно, но это можно относительно просто сделать для стандарта 2.0, поскольку компилятор шейдеров выносится в отдельную внешнюю программу. Компилятор шейдеров используется только на этапе подготовки приложений для создания двоичного программного объекта, который будет работать на программируемых конвейерах рендеринга.

При этом возникает проблема квалификации компилятора шейдеров. Он должен быть квалифицирован, как и любой другой инструмент, используемый при подготовке сертифицируемого программного продукта. В случае, когда прямая сертификация программы невозможна, общепринятый подход состоит в использовании специальных тестов для доказательства корректности работы программы.

Задача тестирования компилятора шейдеров является актуальной, поскольку он в значительной мере определяет эффективность и надежность программ, в которых используется. Это направление активно развивается, существует ряд работ, в которых исследуются различные подходы к тестированию компиляторов шейдеров [4-10]. Основной целью подходов к тестированию, представленных в этих работах, является подготовка таких тестов, которые будут максимально покрывать возможные применения компиляторов. Важное место здесь занимает генерация тестов с одновременным получением корректных результатов их работы.

Однако задача квалификации компилятора шейдеров для применения в авионике существенно отличается от задач, рассматриваемых в этих работах. Это скорее задача не квалификации компилятора, а сертификации результатов его работы в конкретном случае. На практике в комплексе бортового оборудования используется всего несколько наборов шейдеров. В большинстве случаев вообще один, который используется в коде, генерируемом пакетом SCADE [11]. Таким образом, наша задача сводится к доказательству правильности выполнения операций, заданных в двоичном программном объекте. Объект создается с помощью компилятора из описания шейдеров на специальном языке (The OpenGL ES Shading Language) [12]. С одной стороны, задача упрощается, так как нам необходимо проверить правильность работы всего нескольких подготовленных компилятором двоичных программных объектов, но с другой стороны эти проверки должны быть более строгими, поскольку код предполагается использовать в системах, критических с точки зрения безопасности (safety critical). Мы должны проверить все возможные случаи использования полученного программного объекта.

II. Постановка задачи

Спецификации OpenGL ES 2.0 и SC 2.0 реализуют программируемый графический конвейер и являются производными от спецификации OpenGL 2.0. На рисунке 1 показан графический конвейер OpenGL ES 2.0 (SC 2.0) [13]. Затененные прямоугольники на рисунке 1 указывают программируемые этапы конвейера в OpenGL ES 2.0, наличие которых является принципиальной особенностью стандарта ES 2.0 и разработанного на его базе стандарта OpenGL SC 2.0.

Следует отметить, что ряд операций в OpenGL ES выполняется вне шейдеров. Для нашей задачи здесь наиболее существенны такие операции, как Scissor Test, Stencil Test, Depth Test, Blending и Dithering. Поскольку эти операции производятся вне шейдеров, то их следует исключить из разрабатываемых тестовых приложений. Отдельно следует упомянуть этап растеризации, который также выполняется вне шейдеров. Мы не можем исключить его в тестах, но должны учитывать, что «теоретическое» значение координат пикселя может отличаться на одну единицу от значения, сгенерированного GPU.

Исходя из схемы использования шейдеров в графическом конвейере OpenGL ES 2.0 (SC 2.0) мы можем сформулировать следующие общие требования к тестовым приложениям:

1. Тестовые приложения должны быть максимально простыми и понятными так, чтобы результаты работы шейдеров могли быть надежно вычислены в независимой программе.
2. Приложения должны позволять получать результаты в виде изображений для всех допустимых вариантов использования шейдеров (всех допустимых операций, выполняемых шейдерами) таким образом, чтобы правильность полученных изображений доказывала корректность работы компилятора.
3. Тестовые приложения не должны содержать операции, выполняемые вне шейдеров, поскольку это только усложнит анализ полученных изображений. К этим операциям относятся тест трафарета, тест глубины, тест ножниц и использование прозрачности при смешивании текущего цвета с цветом буфера цвета.

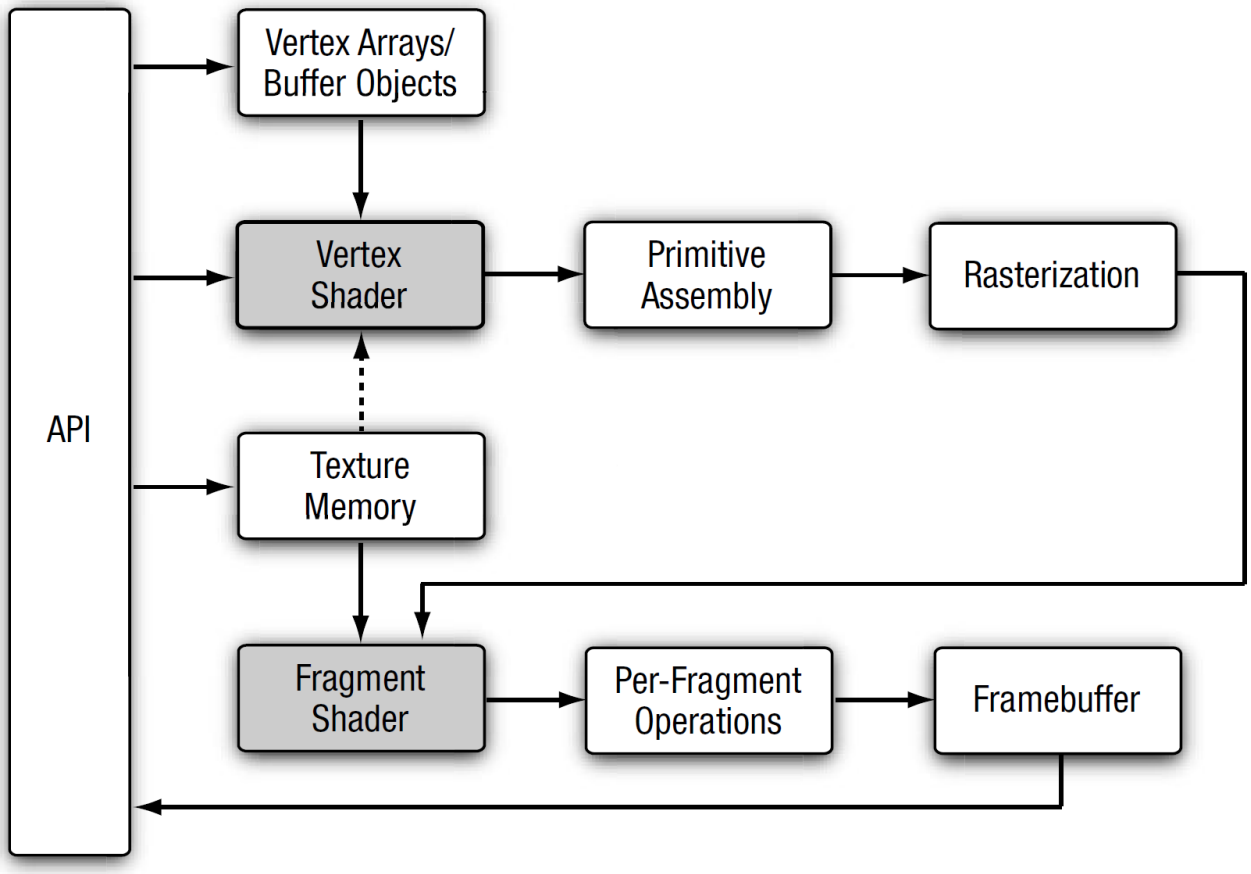


Рис. 1. Графический конвейер OpenGL ES 2.0 (SC 2.0).

Более детальные требования должны задаваться для конкретного набора шейдеров, используемого в бортовом оборудовании. В нашей работе мы исследовали три набора шейдеров, используемых в практических приложениях:

1. Простой набор шейдеров, используемый в приложении AMM (Airport Moving Map, рисунок 2);

В приложении AMM используется простейший набор шейдеров. Вершинный (vertex) шейдер используется для умножения координат на матрицу преобразования, а фрагментный (fragment) шейдер присваивает фрагменту заданный цвет.

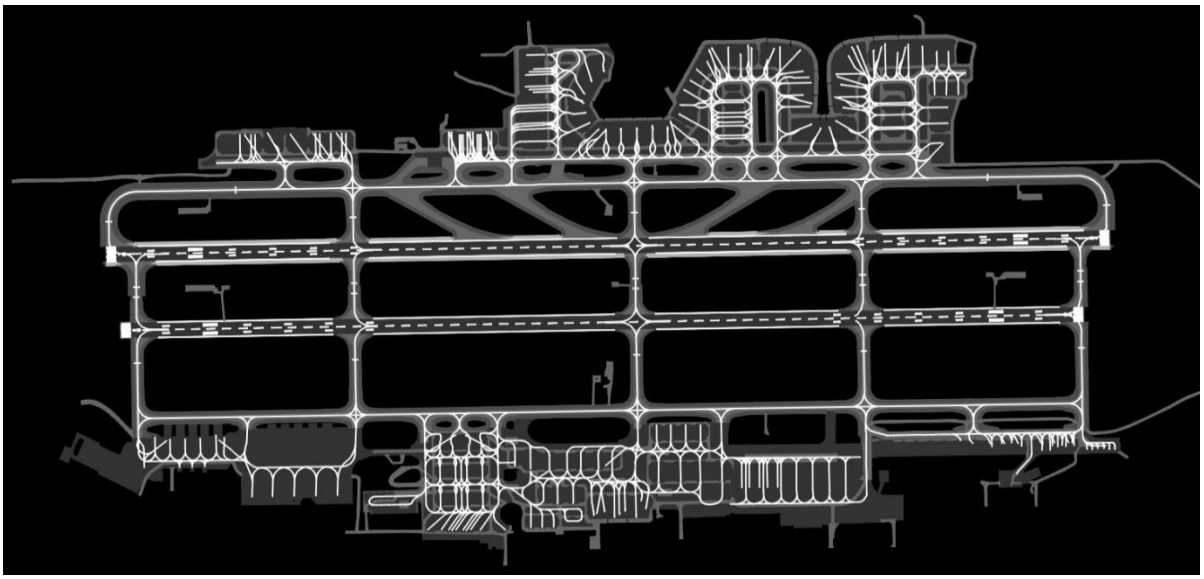


Рис. 2. Приложение Airport Moving Map, использующее вершинный и фрагментный шейдеры.

2. Набор шейдеров, используемый в пакете SCADE R16;

Шейдеры, используемые в пакете SCADE R16, дополнительно обеспечивают возможность использования текстуры для задания цвета фрагмента и вершины.

3. Набор шейдеров, используемый в пакете SCADE R19;

Шейдеры, используемые в пакете SCADE R19, также обеспечивают функциональность плоскости отсечения (clipping plane), которая важна для разработчиков, но отсутствует в обоих стандартах OpenGL SC 1.0.1 и SC 2.0.

Поскольку требования к тестированию шейдеров из пакета SCADE R19 покрывают требования к двум другим наборам шейдерам, то далее будем рассматривать только их.

Таким образом, разрабатываемые тестовые приложения должны обеспечить проверку следующих операций, выполняемых шейдерами:

- Умножение координат на матрицу преобразования (vertex шейдер);
- Использование плоскостей отсечения (vertex шейдер + fragment шейдер);
- Использование 4-х компонентных и черно-белых текстур с различными вариантами смешивания их цветов с цветом геометрического примитива.

III. Тестовое приложение

Исходя из сформулированных требований, тестовое приложение было построено как обычное OpenGL приложение, написанное по стандарту OpenGL SC 2.0 для работы под управлением операционной системы реального времени. Для упрощения и оптимизации процесса тестирования оно генерирует несколько вариантов изображений для последующего анализа их независимой программой, которая должна проверить корректность операций, выполненных с помощью шейдеров.

Нами был разработан набор геометрических примитивов, с одной стороны, достаточно простой для эффективного тестирования, но, с другой стороны, позволяющий проверить корректность всех операций исследуемого набора шейдеров. Этот набор состоит из:

1. Квадрат, расположенный в плоскости XY с координатами вершин

   -5.0 -5.0
   5.0 -5.0
   5.0 5.0
   -5.0 5.0

   Квадрат присутствует в изображении всегда.
2. Набор точек с размером 3 с разными цветами (рис.2), расположенных на этом квадрате по равномерной двумерной сетке размером 16x16. Набор точек присутствует в изображении только по требованию. Он используется для проверки правильности умножения координат на матрицу преобразования

3. В качестве матрицы преобразования выбрана матрица с коэффициентами

   0.070711 0.070711 0.000000 0.000000
   -0.070711 0.070711 0.000000 0.000000
   0.000000 0.000000 1.000000 0.000000
   0.000000 0.000000 0.000000 1.000000

   Это преобразование масштабирует геометрический примитив по осям X и Y и поворачивает его на 45 градусов вокруг оси Z. Преобразование является постоянным для приложения.
4. В приложении используются 2 плоскости отсечения с уравнениями

   1.0 0.0 0.0 0.0
   -0.707106 -0.707106 0.0 0.506066,

   которые вращаются вместе с квадратом геометрии. Их уравнения являются постоянными для приложения. Использование плоскости отсечения может включаться и выключаться для различных тестов.
5. В качестве текстур в приложении используются две текстуры: цветная 4-х компонентная и однокомпонентная черно-белая. Изображение каждой из текстур представляет собой checker (шахматная доска) размером 128x128 пикселей. Размер одной клетки составляет 8x8 пикселей. 4-х компонентная текстура имеет следующие цвета клеток в формате 0xAABBGGRR: 0xFF0000FF, 0xFF00FF00, 0xFFFF0000, 0xFF000000, 0xFF00FFFF, 0xFFFF00FF, 0xFFFFFF00, 0xFFFFFFFF. Для 4-х компонентной текстуры цвета клеток в формате 0xAABBGGRR – 0xAA00FF00 (зеленый) и 0x55FF0000 (синий). Они упакованы в 4 байта. Цвета клеток для черно-белой текстуры упакованы в 1 байт и составляют 0xAA и 0x00. Эти две текстуры генерируются (создаются) непосредственно в приложении.

   Текстуры используются только с квадратом. Для сетки точек они не используются. В качестве текстурных координат для вершин квадрата в приложении устанавливаются значения:

   0.0 0.0
   1.0 0.0
   1.0 1.0
   0.0 1.0

С помощью этих объектов можно проверить корректность всех операций, реализуемых в шейдерах, используемых в библиотеке OGLX пакета SCADE R19 KCG 6.7 [11].

IV. Тестирование шейдеров

Следуя традициям авиационной промышленности, тестовое приложение выполняется на некоторой целевой платформе, в то время как независимая программа реализуется под управлением операционной системы Linux. Для генерации изображений используются параметры, определяющие, какие операции будут проверяться, и вид построенного изображения. Приложение генерирует изображение для каждого набора тестовых параметров и сохраняет его в файле. Файл содержит значения тестовых параметров, использованных при построении изображения, и само изображение в виде массива пикселей. Разработанная технология использовалась для тестирования всех трех наборов шейдеров, описанных в разделе II. Для наиболее полного набора тестов для шейдеров, используемых в библиотеке OGLX пакета SCADE R19, генерировались десять изображений (примеры тестовых изображений представлены на рисунках 3 и 4).

На следующем этапе происходит анализ сгенерированных изображений и проверка их корректности. Для этого тестовая программа последовательно загружает все сгенерированные изображения и параметры, использованные для их генерации, вычисляет теоретические значения цветов в тестовых точках графического примитива и сравнивает его с цветами изображения. Для каждого изображения выполняются следующие действия:

1. Для каждой тестовой точки вычисляются ее координаты в пространстве камеры и ее пиксельные координаты на изображении. По пиксельным координатам определяется цвет в данной точке на изображении.
2. Определяется возможность отсечения точки какой-либо из плоскостей отсечения. Если точка отсечена, то ее цвет равен цвету очистки буфера цвета.
3. Определяется цвет геометрического примитива. Если в тесте отображены точки, то цвет примитива для точки должен быть равен цвету точки, иначе он равен цвету квадрата.
4. Если в тесте не используется текстура, то цвет точки равен цвету геометрического примитива.
5. Если используется текстура, то для точки вычисляются текстурные координаты по ее исходным координатам, и определяется цвет используемой текстуры. Итоговый цвет является результатом

смешивания цвета примитива и цвета текстуры, в соответствии с используемым в данном тесте типом смешивания.

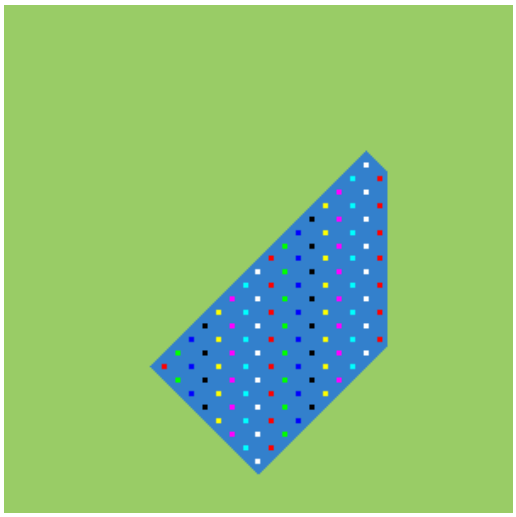
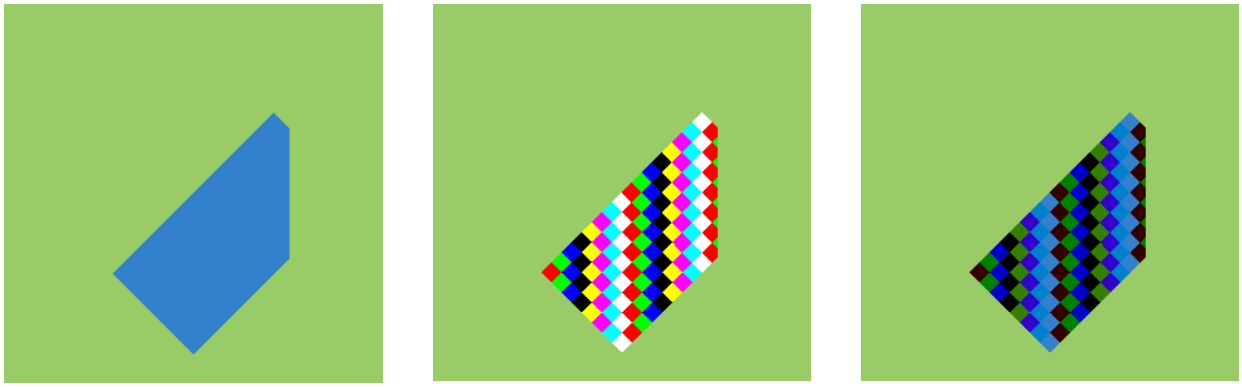


Рис. 3. Изображение теста для проверки правильности работы плоскостей отсечения.



A b c

Рис. 4. Тест для проверки модулирования цвета примитива цветом текстуры:
a - цвет примитива, b – цвет текстуры, c – результат модулирования.

Цвет в пикселе на изображении и теоретически вычисленный цвет сравниваются. Их совпадение для всех точек всех изображений означает корректную работу шейдеров. Если же цвет не совпадает в какой-либо точке, то выдается сообщение об ошибке с указанием исходных координат точки, ее пиксельных координат и цветов теоретического и пиксельного в формате 0xAABBGGRR.

Для тестирования шейдера выполняется десять тестов в соответствии с возможными вариантами использования шейдера.

Тест 1 создает изображение (рис. 3) для проверки матричного преобразования координат вершин геометрических примитивов (квадрата и точек). В примитивы включены точки сетки с различными цветами, в которых проверяется совпадение теоретически вычисленного цвета и цвета изображения.

Тест 2 создает изображение (рис. 3) для проверки правильности работы плоскостей отсечения. Используются две плоскости отсечения, определенные в разделе III. В примитивы включены точки сетки с различными цветами, в которых проверяется совпадение теоретически вычисленного цвета и цвета изображения.

Остальные тесты создают изображения для проверки правильности смешивания цвета примитива и цвета текстуры. Программа шейдера обеспечивает четыре способа такого смешивания:

1. в качестве цвета выбирается цвет примитива, цвет текстуры игнорируется (рис. 4a).
2. в качестве цвета выбирается цвет текстуры, цвет примитива игнорируется (рис. 4b).
3. в качестве компонент цвета берется цвет примитива, а альфа-компонента результата берется из альфа-компоненты текстуры (рис. 4a, т.к. значение альфа-компоненты не видно визуально).
4. цвет примитива модулируется (умножается по-компонентно) цветом текстуры (рис. 4c).

Тесты 3, 4, 5, 6 создают изображения для проверки правильности смешивания цветов этими способами для 4-х компонентной текстуры. Их результаты изображены на рисунках 4a, 4b и 4c.

Тесты 7, 8, 9, 10 создают изображения для проверки правильности смешивания цветов этими способами для черно-белой текстуры. Отметим, что черно-белая текстура перед использованием в шейдере переводится в 4-х компонентную текстуру установкой нулей в компоненты цвета и установкой альфа-компоненты из текстуры. Установленная из черно-белой текстуры альфа-компонента используется (как правило) затем для смешивания цвета примитива с цветом, хранящимся в буфере цвета, с использованием прозрачности (Blending). Эти тесты не имеют наглядного визуального отображения и проверяются только численно. В качестве тестируемых точек выбираются середины клеток текстуры.

Работа шейдера считается корректной, когда цвета на изображениях и теоретически рассчитанные цвета совпали для всех тестируемых точек во всех тестах.

V. Результаты

Разработанная технология тестирования использовалась для тестирования программных объектов, полученных с помощью компилятора шейдеров для трех наборов шейдеров: набор шейдеров, написанных специально для приложения AMM, и два набора шейдеров используемых в пакетах SCADE версий R16 и R19 соответственно. Во всех трех случаях было доказано, что скомпилированные программные объекты выполняют в точности те операции, которые описаны на языке шейдеров The OpenGL ES Shading Language. Следует отметить, что поскольку операции, выполняемые в шейдерах для пакета SCADE R19, покрывают необходимые тесты для двух других наборов, то код, необходимый для тестирования этих наборов, можно получить простым урезанием кода тестирования для пакета SCADE R19.

Список литературы

[1] DO-178C Software Considerations in Airborne Systems and Equipment Certification, http://www.rtca.org/store_product.asp?prodid=803.
Квалификационные требования часть 178С, АР МАК, 2014

[2] Barladian B.Kh , Deryabin N.B., Voloboy A.G., Galaktionov V.A., Shapiro L.Z., Valiev I.V. and Solodelov Yu.A., Specifics of the Development of an On-Board Visualization System for Civil Aircrafts // Programming and Computer Software, 2024, Vol.50, No.3, pp.215-223. DOI: 10.1134/S0361768824700014

[3] The Mesa 3D Graphics Library, http://www.mesa3d.org

[4] Donaldson A, Evrard H., Lascu A., Thomson P., Automated testing of graphics shader compilers. Proceedings of the ACM on Programming Languages, Volume 1, Issue OOPSLA. (Article No.: 93, Pages 1 – 29, 2017). https://doi.org/10.1145/3133917

[5] Donaldson A, Evrard H., Lascu A., Thomson P., Putting Randomized Compiler Testing into Production (Experience Report), 34th European Conference on Object-Oriented Programming (ECOOP 2020). Leibniz International Proceedings in Informatics (LIPIcs), Volume 166, pp. 22:1-22:29, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2020) , https://doi.org/10.4230/LIPIcs.ECOOP.2020.22

[6] Crawford L. and O'Boyle M., A Cross-platform Evaluation of Graphics Shader Compiler Optimization. 2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Belfast, UK, 2018, pp. 219-228, doi: 10.1109/ISPASS.2018.00035

[7] Xiao D. , Liu Z. and Wang S. , Metamorphic Shader Fusion for Testing Graphics Shader Compilers, 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), Melbourne, Australia, 2023, pp. 2400-2412, doi: 10.1109/ICSE48619.2023.00201

[8] Ádám István Szűcs, Improving graphics programming with shader tests. Pollack Periodica Volume 14: Issue 1.

[9] Donaldson A, Thomson P., Teliman V., Milizia S., Maselco A. P., Karpiński A., Test-case reduction and deduplication almost for free with transformation-based compiler testing. PLDI 2021: Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, June 2021, Pages 1017 - 1032, https://doi.org/10.1145/3453483.3454092

[10] Kuo L. -W. , Yang C. -C. , Lee J. -K. and Tseng S. -Y. , The design of LLVM-based shader compiler for embedded architecture, 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), Hsinchu, Taiwan, 2014, pp. 961-968, doi: 10.1109/PADSW.2014.7097916

[11] https://cae-expert.ru/product/scade-suite

[12] www.khronos.org/files/opengles_shading_language.pdf

[13] Aaftab Munshi Dan Ginsburg Dave Shreiner, OpenGL ES 2.0 Programming Guide, Boston: Addison.Wesley, 2008.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-213-224

Концепция федеративного рендерига
Federative rendering concept

А. Д. Жданов 1, Д. Д. Жданов 1
1 Университет ИТМО, Санкт-Петербург, Россия

A. D. Zhdanov 1, D. D. Zhdanov 1
[1]ITMO University, Saint Petersburg, Russia

Аннотация. В последнее время все чаще возникают ситуации, когда заказчику требуется провести оптическое моделирование или рендеринг сцен, представляющих собой коммерческую тайну. Поэтому заказчик не готов делиться данными сцены с владельцами программного обеспечения (ПО) систем реалистичного рендеринга, обеспечивающего соответствующий сервис, при этом он также не готов приобрести дорогостоящее ПО. Таким образом возникает проблема, когда заказчик услуги рендеринга опасается утечек данных, а разработчик ПО боится пиратского использования дорогостоящей системы реалистичного рендеринга, и требуется найти решение, позволяющее выполнить рендеринг сцен без передачи заказчику программного обеспечения и без предоставления данных сцены владельцам ПО. В данной работе рассматривается решение данной проблемы средствами федеративного рендеринга, который позволяет оставить все данные на стороне заказчика рендеринга, а управление всем процессом рендеринга выполнить на стороне владельца системы реалистичного рендеринга. В рамках исследования был реализован федеративный рендеринг методами прямой и обратной трассировки лучей, представлены результаты рендеринга и сравнение их с результатами, полученными средствами традиционной системы реалистичного рендеринга.

***Ключевые слова*:** рендеринг, трассировка лучей, федеративные методы, распределенное программирование

Аннотация. Recently, situations have increasingly arisen when the customer needs to conduct optical modeling or rendering of scenes that represent a trade secret. Therefore, the customer is not ready to share scene data with the owners of software for realistic rendering systems, and he is also not ready to purchase expensive software. Thus, a problem arises when the customer of the rendering service is afraid of data leaks, and the software developer is worried about the pirated use of an expensive realistic rendering system. As a result, a task arises when it is required to render scenes without transferring the software to the customer and without providing the scene to the software owners. In this paper, we consider the solution to this dilemma through the use of federated rendering methods, when all the data remains completely on the side of the rendering customer, and the entire rendering process is managed by the owners of the realistic rendering system. As part of the study, federated rendering using forward and reverse ray tracing methods was implemented, and the results of rendering and their comparison with the results of using a traditional realistic rendering system were presented.

***Ключевые слова*:** rendering, ray tracing, federative methods, distributed computing.

I. Введение

Физически корректный рендеринг пользуется большим спросом не только в архитектурных проектах, дизайне интерьеров, игровых приложениях и при создании обучающих наборов данных для систем искусственного интеллекта, но и для решения задач виртуального прототипирования, где важно понять, как сложное оптическое устройство будет работать в условиях реального мира. При решении таких задач интеллектуальной собственностью являются не только результаты рендеринга, но и данные сцены (например, оптическое устройство), а также программное обеспечение системы реалистичного рендеринга. В ряде случаев заказчик не готов приобрести систему рендеринга, но готов купить сервис моделирования и визуализации. Однако, при этом может возникнуть конфликт интересов, связанный с тем, что заказчик рендеринга не хочет передавать данные для визуализации разработчику системы рендеринга, опасаясь утечек данных и, как следствие, финансовых потерь, а разработчик системы рендеринга не готов предоставлять заказчику свое программное обеспечение, опасаясь возможности последующего его нелицензионного использования и распространения. Это создает проблему: программное обеспечение не может быть установлено на компьютер клиента, а клиент не желает передавать свои данные на сервер для расчетов.

Решением данной проблемы должно стать отделение системы рендеринга от пользовательских данных. Изоляция системы рендеринга от данных сцены возможна за счет объектно-ориентированной организации данных сцены. Базовые методы программного интерфейса сцены обеспечивают взаимодействие светового излучения с объектами сцены и тем самым определяют физическую сущность системы рендеринга. Эти методы реализуются в виде ряда сервисов, выполняемых на сервере заказчика. Сервисы, в соответствии с базовым программным интерфейсом объектов сцены, выполняют основные операции с данными сцены, не передавая их системе рендеринга. Они передают только информацию о выполнении операции и в ряде случаев дополнительные данные, необходимые для оптимизации алгоритма рендеринга. Для организации такого взаимодействия предлагается специальная трехуровневая организация сцены, которая позволяет преобразовать уникальные данные пользователя в унифицированное представление данных рендеринга, представленных в виде абстрактных программных интерфейсов. На основе этой объектно-ориентированной организации и базовых алгоритмов рендеринга строится система федеративного рендеринга, разделяющая данные сцены от системы рендеринга.

II. Постановка задачи

Существует большое число алгоритмических и программных решений физически корректного реалистичного рендеринга. Большинство из существующих решений можно найти в обзоре [1]. Как правило, эти решения локализованы на одном компьютере. Даже если предлагается распределенное или облачное решение, то каждый узел сети содержит полноценную систему рендеринга, то есть ядро системы (вычислительную часть и сцену). Такое решение обеспечивает высокую вычислительную эффективность при минимизации затрат на передачу данных в системе распределенных или облачных вычислений. Существующие решения обеспечивают два основных сценария работы программной системы: либо пользователь приобретает компьютерную систему рендеринга и самостоятельно выполняет необходимые расчеты, либо пользователь арендует вычислительные ресурсы, на которые он передает свои данные для выполнения необходимых расчетов. Оба решения имеют свои преимущества и недостатки. В первом случае пользователь получает возможность выполнять необходимые вычисления, сохраняя конфиденциальность своих данных, однако, он должен приобрести соответствующее программное обеспечение за достаточно высокую стоимость, а правообладатель программного обеспечения при этом не защищен от несанкционированного копирования и распространения его программного обеспечения. Во втором случае пользователь получает достаточно простое и дешевое решение (тем более, если расчеты разовые), но конфиденциальность данных теряется.

Предлагается решение федеративного рендеринга, которое позволяет физически разделить систему рендеринга и данных сцены. При этом данные пользователя могут находиться на одном физическом устройстве (или устройствах, если допускаются распределенные вычисления на стороне пользователя), а система рендеринга (правообладателя программного обеспечения) – на другом физическом устройстве. При этом система рендеринга ничего не знает о сцене (кроме общей статистики по результатам вычислений) и не получает доступа к ее данным, а пользователь не имеет доступа к алгоритмам и программным решениям системы рендеринга. Физическая корректность и реалистичность результатов рендеринга сохраняется, но допускается некоторое снижение производительности.

III. Реалистичный рендеринг

Реалистичный рендеринг решает задачу физически корректного расчета яркости (освещенности, интенсивности, светового потока) на приемнике излучения оптически сложной сцены, которая может содержать оптические системы освещения и наблюдения. Расчет физически корректного значения видимой яркости точки сцены выполняется с использованием уравнения рендеринга, изначально представленного в работах Джеймса Каджия [2]:

$$L_o(x,\Theta_o) = L_e(x,\Theta_o) + \int_{\Omega_i} L_i(x,\Theta_i) f_r(x,\Theta_i,\Theta_o)|\Theta_i \cdot \mathrm{N}_x| d\omega_i,$$

где x – точка поверхности сцены, для которой вычисляется яркость; Θ_o – направление наблюдения данной точки; $L_e(x,\Theta_o)$ – собственная яркость точки x при наблюдении с направления Θ_o; $L_i(x,\Theta_i)$ – яркость, наблюдаемая из точки x в направлении Θ_i; $f_r(x,\Theta_i,\Theta_o)$ – значение двунаправленной функции рассеивания для точки x и углов наблюдения Θ_o и освещения Θ_i.

Для решения уравнения рендеринга используется большое число подходов, таких как обратная трассировка лучей [3], трассировка путей [4], прямая трассировка лучей [5], двунаправленная трассировка лучей [6], карты освещенности [7], фотонные карты [8], прогрессивные прямые [9] и обратные [10] фотонные карты, методы Метрополиса [11] и другие. Рассмотрим, например, алгоритм рендеринга, основанный на методе обратной трассировки лучей, схематично представленный на рис. 1, и алгоритм, основанный на методе прямой

трассировки лучей, схематично представленный на рис. 2, использующие выборку по значимости [12] и расчет яркости прямого виденья на поверхностях сцены.

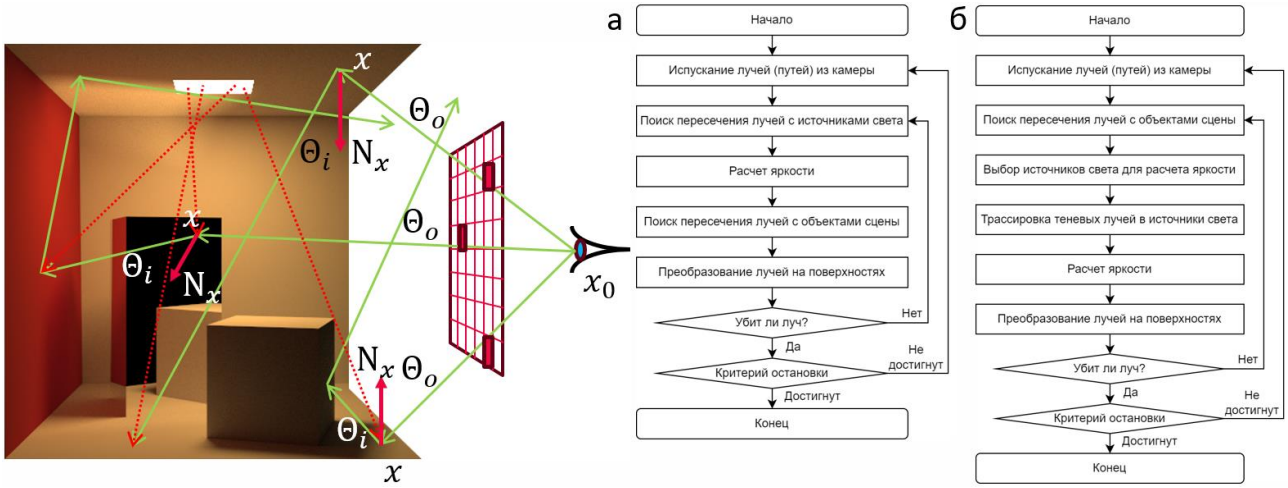


Рис. 1. Алгоритм рендеринга, основанный на методе обратной трассировки лучей, использующий выборку по значимости на поверхностях сцены (а) и выборку по значимости и расчет яркости прямого виденья на поверхностях сцены (б)

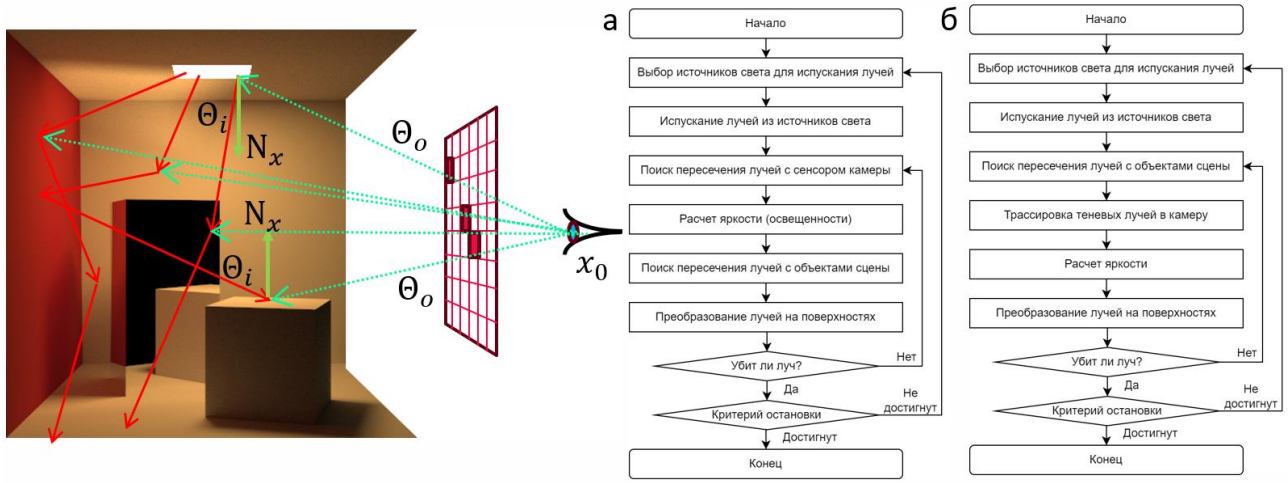


Рис. 2. Алгоритм рендеринга, основанный на методе прямой трассировки лучей, использующий выборку по значимости на поверхностях сцены (а) и выборку по значимости и расчет яркости прямого виденья на поверхностях сцены (б)

Из представленных выше примеров можно увидеть, что алгоритмы рендеринга выполняют однотипные действия с данными сцены, как то испускание лучей объектами сцены, поиск пересечения лучей с объектами сцены, трансформация лучей на поверхностях и прочее. Таким образом, разработав универсальные абстрактные интерфейсы для данных сцены и отделив систему рендеринга от непосредственно данных, можно реализовать модель федеративного рендеринга.

В результате, задача, которую необходимо решить в рамках данного исследования, сводится к размежеванию модели данных сцены и системы фотореалистичной визуализации для реализации различных алгоритмов рендеринга, построенных на использовании абстрактных интерфейсах сцены. Это позволяет разделить систему фотореалистичного рендеринга на два независимых набора сервисов, первый из которых должен отвечать за хранение данных сцены и выполнение операций, связанных с взаимодействием светового излучения с ее объектами, и второй – непосредственно за реализацию алгоритмов рендеринга, построенных на абстрактных интерфейсах взаимодействия светового излучения с объектами сцены. При этом если коммуникация данных сервисов заключается исключительно в передаче команд и получении простых численных ответов, то это не позволит разработчикам системы рендеринга получить доступ к конфиденциальным данным клиента. Решение данной задачи позволит полностью разделить данные сцены и систему рендеринга и исключить необходимость передачи данных между владельцами данных и владельцами системы рендеринга, а также защитить конфиденциальную информацию с обоих сторон.

IV. Модель представления данных сцены

Для отделения данных сцены от системы рендеринга необходимо, чтобы все данные, включая итоговые результаты рендеринга, хранились исключительно на стороне сервисов клиента, а также, чтобы сервисы клиента могли производить все операции, необходимые для работы алгоритмов рендеринга. В результате необходимо разработать максимально абстрактную модель представления данных, которая будет поддерживать набор интерфейсов, позволяющих реализовать на их основе различные алгоритмы рендеринга, как показано на рис 3. Таким образом, данные будут находиться исключительно на стороне клиента, а алгоритмы рендеринга, выполняющиеся на стороне системы рендеринга, будут реализовывать различные методы рендеринга, используя их абстрактные интерфейсы. Результаты выполнения будут оставаться на стороне клиента, тогда как на сторону рендеринга будут передаваться статистические данные (например, количество трассированных и убитых лучей) и, при необходимости, уникальные идентификаторы результатов, сохраненных на стороне клиента, которые могут быть использованы как аргументы последующих вызовов.

Рис. 3. Использование абстрактных интерфейсов данных для взаимодействия между данными сцены и алгоритмами рендеринга в системе фотореалистичной визуализации

Рассмотрим, что содержат в себе данные сцены. Во-первых, это геометрические объекты, описывающие геометрические данные сцены, как, например, поверхности линз оптической системы объектива камеры, для которых необходимо задать оптические свойства поверхностей и сред. Во-вторых, для работы алгоритмов рендеринга требуется задать источники света и виртуальную камеру. Несмотря на кажущуюся простоту перечисленных выше объектов, возникает проблема представления данных. Например, если рассматривать геометрию сцены, то для большинства алгоритмов рендеринга наиболее удобным представлением является треугольная сетка, для которой разработано множество высокопроизводительных алгоритмов поиска точки пересечения лучей с поверхностью. Для пользователя системы трехмерного моделирования наиболее удобным может быть задание геометрии с использованием аналитических форм поверхности или различного рода сплайнов. Таких представлений пользовательской геометрии может быть бесконечно много.

Очевидным решением данной проблемы будет преобразование данных из их внешнего (пользовательского) представления в системе трехмерного моделирования в представление, наиболее подходящее для алгоритмов рендеринга. Однако, данное решение содержит проблему, которая заключается в том, что различные алгоритмы рендеринга могут требовать различного представления данных. Таким образом, при добавлении нового формата пользовательских данных возникает необходимость поддерживать их преобразование во все имеющиеся представления данных рендеринга, а при добавлении нового алгоритма рендеринга, который будет требовать особое представление данных, будет необходимо поддержать все имеющиеся типы входных данных. На начальном этапе разработки системы федеративного рендеринга данное решение допустимо. Однако, в процессе развития системы или расширении списка поддерживаемых алгоритмов рендеринга дальнейшее развитие системы станет крайне трудозатратным и может приводить к сбоям в ее работе.

В рамках данной работы для решения вышеописанной проблемы предлагается использовать трехуровневую модель представления данных сцены с дополнительным промежуточным (или каноническим) представлением данных. Этот канонический уровень данных имеет ограниченное число представлений, как правило, имеющих вид, близкий к представлениям, используемым алгоритмами рендеринга. В результате данные модели сцены, определенные в пользовательском представлении в системе трехмерного моделирования, преобразуются в промежуточное представление, а затем уже в представление, требуемое для конкретного алгоритма рендеринга, как показано на рис. 4. Если появляется новый тип представления данных пользователя, то отпадает необходимость поддержки всех алгоритмов рендеринга, а достаточно поддержать преобразование данных пользователя в промежуточное представление. При добавления нового представления данных для нового

алгоритма рендеринга, необходимо только добавить методы преобразования из промежуточного представления данных в представление данных, требуемое алгоритмом рендеринга.

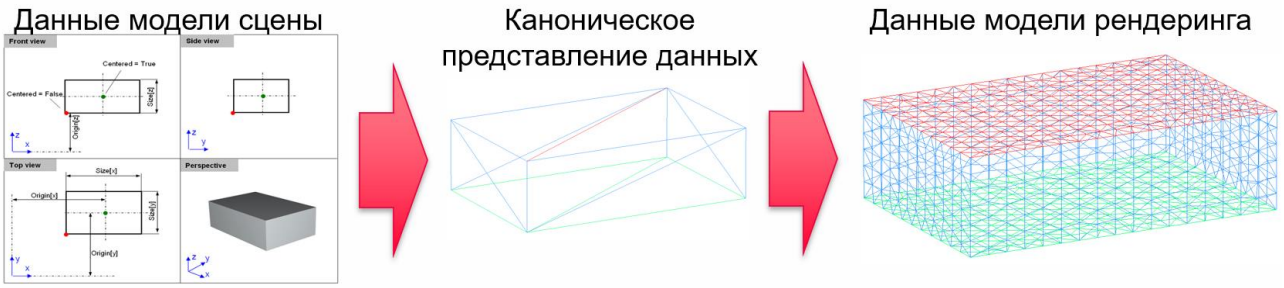


Рис. 4. Преобразование данных из параметрического задания в системе трехмерного моделирования на уровне пользователя (слева) в треугольную сетку (по центру) и в треугольную сетку с разбиением треугольников для расчета распределения освещенности сцены (справа)

Данные на уровне пользователя

Первым уровнем представления данных сцены является представление данных в виде, определенном пользователем. На данном уровне представление данных соответствует представлению данных в системе трехмерного моделирования, используемого пользователем для создания данных для симуляции. Это могут быть, например, данные в системе Blender [13] или Dassault Systems CATIA [14]. Программные интерфейсы данного уровня содержат методы, позволяющие пользователю формировать объекты сцены в удобном для него представлении, а также преобразовывать их в промежуточное представление.

Данные на промежуточном уровне

Вторым уровнем представления данных сцены является промежуточное представление данных в каноническом представлении с ограниченным числом возможных типов объектов. Основной задачей данного уровня является ограничение потенциально бесконечного количества типов объектов данных на уровне пользователя до минимального достаточного, обеспечивающего реализацию физически корректного рендеринга. Программные интерфейсы данного уровня содержат методы, позволяющие преобразовать данные пользователя в представление данных, совместимое с системой рендеринга.

Данные на уровне рендеринга

Третьим уровнем представления данных сцены является представление данных на уровне системы рендеринга, то есть данные и программные интерфейсы объектов, используемые непосредственно в процессе расчета. Программные интерфейсы данного уровня содержат методы, необходимые для реализации различных алгоритмов рендеринга, которые будут подробно рассмотрены в следующей главе.

Интеграция трехуровневой модели представления данных сцены в систему рендеринга

На рис. 5 показана блок-схема интеграции трехуровневой модели представления данных сцены в систему реалистичного рендеринга.

Рис. 5. Блок-схема интеграции трехуровневой модели представления данных сцены в систему рендеринга

Пунктиром на схеме обозначены блоки, которые необходимо разработать и реализовать в случае появления новых типов данных пользователя, новых способах их формирования или построения нового алгоритма рендеринга. Как видно из представленной блок-схемы, использование трехуровневой модели представления данных сцены существенно снижает трудозатраты для расширения системы рендеринга и снижает вероятность внесения ошибок, исключая необходимость поддержки всех видов входных данных всеми алгоритмами рендеринга.

V. Интерфейсы объектов сцены на уровне рендеринга

Для реализации различных алгоритмов рендеринга объекты данных сцены в представлении модели рендеринга должны иметь интерфейсы, достаточные для осуществления соответствующего взаимодействия светового излучения с объектами сцены. Рассмотрим основные действия, которые обеспечивают взаимодействие светового излучения с объектами сцены для корректной работы алгоритмов рендеринга, основанных на методах прямой или обратной трассировки лучей. Эти действия включают в себя испускание световых лучей, теневых лучей или путей из камеры, поиск пересечения лучей с объектами сцены, преобразование лучей на поверхностях сцены или в среде между поверхностями, а также методы, необходимые для расчета освещенности и яркости на поверхностях геометрических объектов сцены. Все эти действия схематично представлены на рис. 6 и должны быть поддержаны абстрактными интерфейсами модели данных сцены для разделения данных сцены и системы рендеринга. Для повышения эффективности все рассматриваемые интерфейсы спроектированы для работы с пучками лучей, а не с индивидуальными лучами.

Такое разделение позволит хранить данные исключительно на стороне клиента, тогда как управление ими в процессе рендеринга будет выполняться со стороны системы рендеринга, исключая его доступ к конфиденциальным данным сцены.

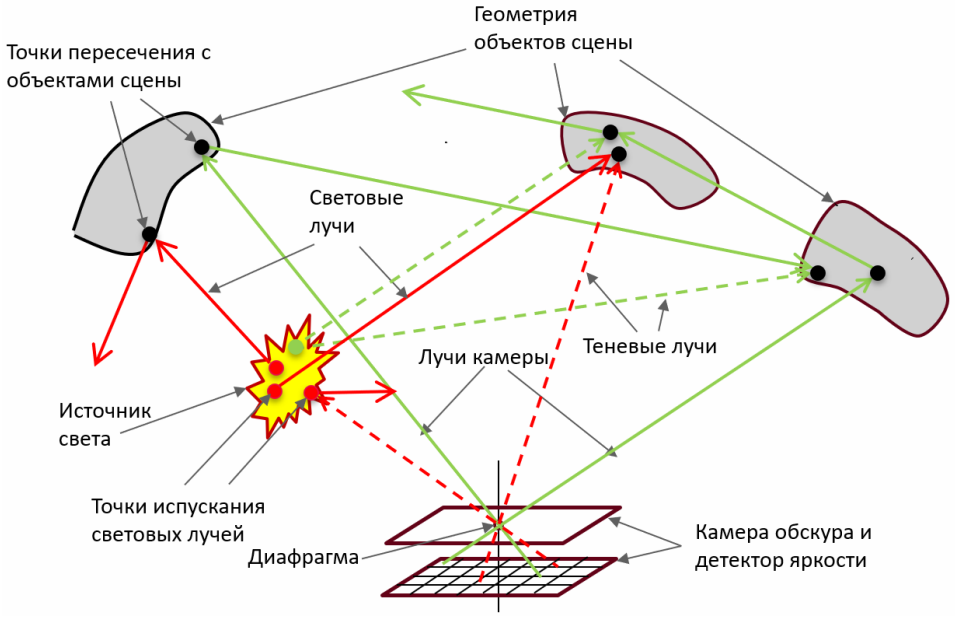


Рис. 6. Объекты сцены и их взаимодействие

Рассмотрим более подробно интерфейсы, которые должны быть поддержаны объектами данных сцены в представлении модели рендеринга. Для ускорения работы и возможности векторизации вычислений все интерфейсы объектов сцены должны поддерживать работу как с одним элементом (например, лучом), так и с вектором элементов (например, пучком лучей).

1. *Геометрический объект поверхности сцены*

Основной задачей интерфейса геометрической поверхности сцены является поиск точки пересечения лучей с ее поверхностью.

Методы интерфейса геометрического объекта поверхности сцены включают в себя:

- Поиск точки встречи пучка лучей с геометрическими объектами сцены (прямых, обратных, теневых).
- Формирование внешней границы геометрического объекта (для построения ускоряющей структуры, обеспечивающей реализацию методов быстрого поиска точек встречи пучка лучей с геометрическими объектами сцены).

Оптические свойства поверхности объекта сцены

Основными задачами интерфейса оптических свойств поверхности объекта сцены являются методы преобразование луча на поверхности сцены, а также вычисление видимой яркости точки на ее поверхности.

Методы интерфейса оптических свойств поверхности объекта сцены включают в себя:

- Преобразование пучка лучей на поверхности (как правило, один луч формирует не более одного луча) и информирование о результате данного преобразования, включая событие поглощения.
- Расчет видимой яркости поверхности исходя из ее локальной освещенности, ориентации, направлений освещения и наблюдения.

Оптические свойства среды сцены

Основной задачей интерфейса оптических свойств среды сцены является преобразование луча в процессе его распространения. Стоит отметить, что распространение света в среде не обязательно является прямолинейным (например, в случае градиентного изменения показателя преломления света в среде) и может формировать собственную яркость, однако, в рамках выбранных моделей рендеринга эти явления не рассматриваются.

Методы интерфейса оптических свойств среды сцены включают в себя:

- Перенос пучка лучей на заданную дистанцию с возможностью поглощения отдельных лучей (в общем случае допускается непрямолинейное распространение света).
- Вычисление показателя преломления для рядя точек среды (для расчета направления преломленных лучей и коэффициентов отражения/пропускания).

Источник света

Основными задачами интерфейса источника света являются испускание световых лучей с ее поверхности, а также методы, позволяющие оценить яркости прямого виденья и прямого освещения.

Методы интерфейса источника света включают в себя:

- Расчет видимой яркости источника света для пучка лучей зрения.
- Испускание пучка лучей с источника света в соответствии с заданной функцией плотности распределения вероятности.
- Выбор точек на поверхности источника света (в соответствии с заданной функцией плотности распределения вероятности) для расчета яркости прямого освещения.
- Расчет локальной освещенности на поверхности сцены от выбранного набора точек.
- Расчет светового потока источника света.

Камера

Основной задачей интерфейса камеры является испускание путей из камеры, формирование теневых лучей в камеру и методы для накопления яркости на детекторе камеры.

Методы интерфейса камеры включают в себя:

- Испускание пучка лучей из камеры в направлении сцены с заданной функцией плотности распределения вероятности выбора точек излучения.
- Формирование пучка теневых лучей от точек на объектах сцены в камеру.
- Проверка попадания пучка лучей на детектор камеры (результат преобразуется в соответствующие световые единицы детектора).

Сцена

Для повышения производительности сцена содержит интерфейсы, объединяющие группы однотипных объектов с общими интерфейсами, например, позволяющими испускать лучи из всех источников сцены одновременно.

Методы интерфейса сцены включают в себя:

- Поиск пересечения пучка лучей со всеми геометрическими объектами сцены.
- Испускание пучка лучей со всех источников света в сцене (в соответствии с заданной функцией плотности распределения вероятности).
- Формирование точек на всех источниках света сцены для расчета яркости прямого освещения.
- Обработка результатов расчета для безопасной записи и выполнения арифметических операций с данными при работе в параллельных потоках.

VI. Система федеративного рендеринга

Система федеративного рендеринга строится на следующих базовых положениях:

- Трехуровневая модель представления сцены, которая позволяет отделить модель представления данных пользователя от их представления в системе рендеринга с возможностью гибкого расширения как форматов ввода данных, так и поддерживаемых интерфейсов для алгоритмов рендеринга.
- Абстрактные интерфейсы объектов представления данных в системе рендеринга, которые позволяют реализовать различные алгоритмы рендеринга без обращения непосредственно к конфиденциальным данным пользователя.
- Разделение данных пользователя и алгоритмов рендеринга на два независимых набора сервисов, которые взаимодействуют друг с другом через сетевые протоколы, что позволяет ограничить доступ к конфиденциальным данным обеих сторон в процессе работы алгоритмов рендеринга.

1. Архитектура системы федеративного рендеринга

Архитектура системы федеративного рендеринга включает в себя два ключевых компонента: сервер системы рендеринга и сервер заказчика:

- Сервер системы рендеринга отвечает за работу сервисов алгоритмов рендеринга, формирует команды заказчику на выполнение элементарных действий со сценой и управляет процессом рендеринга на компьютере заказчика. Он не имеет полного доступа к данным сцены, что обеспечивает конфиденциальность и безопасность информации заказчика, а оперирует исключительно статистическими данными, на основании которых формирует команды для выполнения на сервере заказчика.
- Сервер заказчика отвечает за работу сервисов заказчика и выполнения команд, поступающих от сервисов системы рендеринга в соответствии с интерфейсами, описанными в предыдущих главах. Он обладает полными данными о сцене, может произвести преобразование из представления данных на уровне пользователя в промежуточное представление, а далее в представление, требуемое для конкретного алгоритма рендеринга. Все результаты выполнения команд от сервера системы рендеринга сохраняются исключительно локально, что гарантирует конфиденциальность и безопасность информации заказчика, при этом алгоритм рендеринга работает на стороне сервера системы рендеринга, что гарантирует конфиденциальность и безопасность системы рендеринга.

Перед началом работы с системой федеративного рендеринга заказчик устанавливает на своем сервере компоненту ПО системы рендеринга, позволяющую загрузить данные сцены и через промежуточное представление сформировать объекты, реализующие необходимые интерфейсы рендеринга и подключение к серверу системы рендеринга. Данная компонента может быть реализована различными способами, например, как простейший загрузчик файлов из командной строки, плагин к системе трехмерного моделирования, полноценная система моделирования или библиотека программных интерфейсов для интеграции в собственную инфраструктуру заказчика.

Взаимодействие сервисов в процессе рендеринга

Рассмотрим взаимодействие компонент системы федеративного рендеринга при реализации алгоритма рендеринга, основанного на методе обратной трассировки лучей, использующем выборку по значимости и расчет яркости прямого виденья на поверхностях сцены. Процесс рендеринга представляет собой серию запросов и ответов между серверами заказчика и системы рендеринга. Блок-схема взаимодействия серверов системы рендеринга и заказчика для данного алгоритма представлена на рис. 7.

В процессе федеративного рендеринга сервера системы рендеринга и заказчика выполняют следующие действия:

1. На этапе инициализации сервер заказчика инициализирует соединение с сервером системы рендеринга и передает ему минимально необходимую информацию о параметрах рендеринга и сцене, включающую в себя требуемую точность, параметры изображения, метод рендеринга и т. д.
2. После установки соединения между серверами заказчика и системы рендеринга сервисы системы рендеринга посылают сервисам заказчика команды на выполнение элементарных действий, таких как сформировать представление модели данных для требуемого алгоритма рендеринга на основе входных данных пользователя, выпустить пучок лучей с камеры, найти пересечения пучка лучей с поверхностями сцены, выбрать источники света и рассчитать освещенность от этих источников света, проверить затенение лучей пучка, вычислить яркость по рассчитанной освещенности, преобразовать лучи на поверхностях сцены и т. п.

3. Сервер заказчика посредством локальных сервисов выполняет запросы, поступаемые от сервисов системы рендеринга. За выполнение команды, как правило, отвечает конкретный интерфейс объекта, к которому была адресована команда. Интерфейс реализуется в виде сервиса, который имеет доступ не только к данным данного объекта, но и ко всей сцене и внутреннему кэшу, в котором хранятся результаты всех расчетов. Для многоядерных компьютеров предполагается параллельное исполнение команд в синхронном режиме с общей памятью.
4. При завершении запроса сервер заказчика информирует систему рендеринга о результатах работы своих сервисов. Как правило, информация имеет статистический характер, например, число лучей, которые нашли пересечение с объектами сцены, число лучей, которые не поглотились после взаимодействия с поверхностями сцены, оценки точности вычислений и т. п.
5. Сервисы сервера системы рендеринга анализируют результаты выполнения запросов, получаемых от заказчика, и принимают определенные решения о следующих шагах выполнения алгоритмов рендеринга, например, сформировать новую порцию лучей из камеры, окончить расчет и т. п.

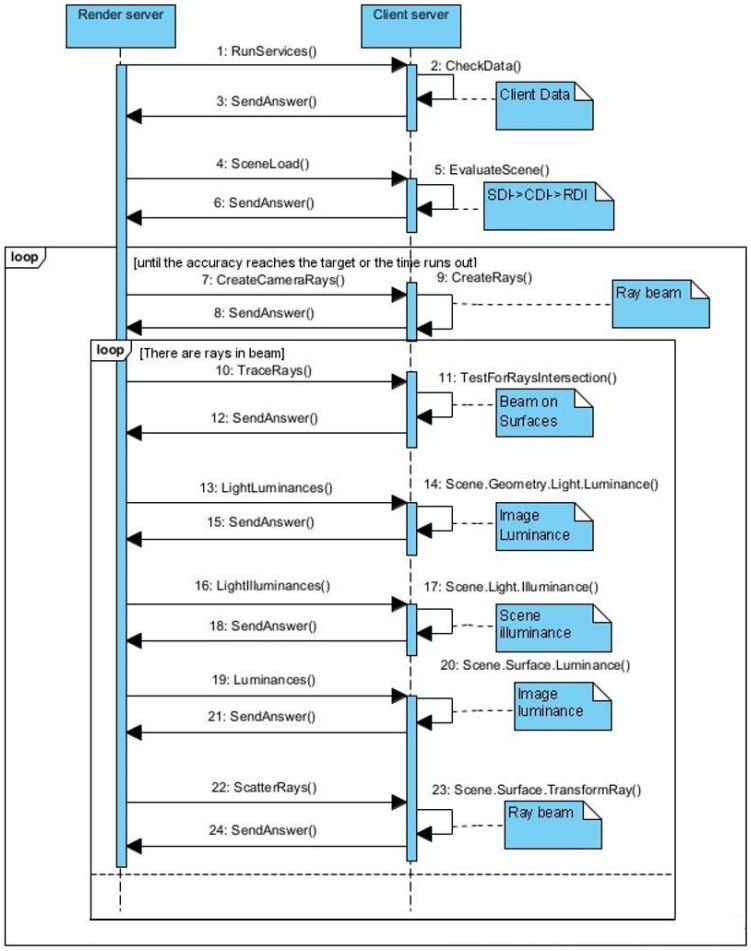


Рис. 7. Взаимодействие сервисов клиента и системы рендеринга при работе алгоритма, основанного на методе обратной трассировки лучей, использующем выборку по значимости и расчет яркости прямого виденья на поверхностях сцены

В процессе работы данной системы рендеринга все результаты, относящиеся к сцене, хранятся исключительно на сервере заказчика. Кроме того, результаты расчета (лучи, локальные значения освещенности и яркости, локальные нормали, барицентрические координаты и т. п.) также хранятся на сервере заказчика, что позволяет избежать передачи по сети больших объемов данных, снижающих общую производительность вычислений.

Стоит отметить, что описанная в данной работе модель федеративного рендеринга может быть расширена для поддержки распределенной модели вычислений с несколькими серверами заказчика и одним сервером системы рендеринга. Для этого потребуется расширить интерфейсы сервисов на стороне заказчика, обеспечив возможность обмена промежуточными данными о результатах рендеринга и слияния этих результатов.

VII. Результаты рендеринга

Концепция федеративного рендеринга была реализована и протестирована на ряде сцен, включая стандартные тестовые сцены, например, Cornell box [15] и сцены, формирующие изображения в сложных оптических системах. Все расчеты выполнялись на многопроцессорной рабочей станции с двумя шестнадцатиядерными процессорами AMD EPIC 7281 2.1GHz. Сервисы системы рендеринга и системы заказчика были запущены на одной системе. Сервисы работали на независимых процессах, без использования общей памяти и на разных физических процессорах рабочей станции. Обмен данными между сервисами был реализован с использованием протокола TCP/IP с целью более корректного моделирования поведения системы в условиях реальной эксплуатации.

На рис. 8 показан результат рендеринга стандартной тестовой сцены Cornell Box с использованием федеративного рендеринга и алгоритма, основанного на методе обратной трассировки лучей с выборкой по значимости и расчетом яркости прямого виденья на поверхностях сцены.

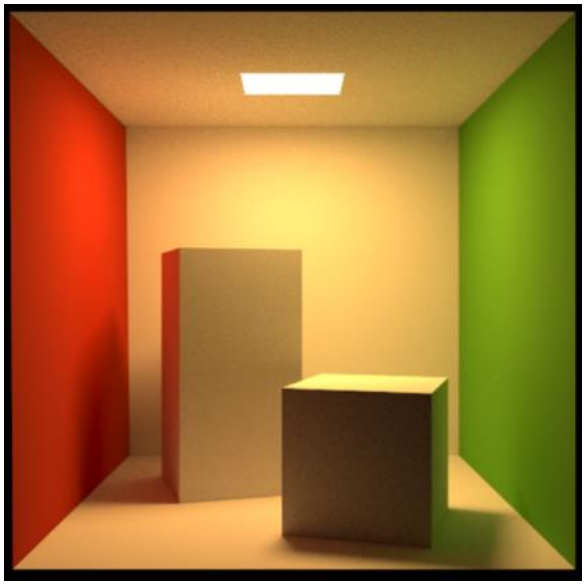


Рис. 8. Результат применения федеративного рендеринга к сцене Cornell box

Для проверки работоспособности метода федеративного рендеринга для решения задач виртуального прототипирования было проведено моделирование функции рассеивания точки сложной оптической системы [16], изображенной на рис. 9(а), состоящей из десяти линз с асферическими поверхностями, заданными полиномами Якоби 34-й степени. Моделирование проводилось методом прямой трассировки лучей. Результаты моделирования для внеполевого бесконечно удаленного источника света представлены на рис. 9(б) и 9(в). Поскольку источник света находился вне поля зрения объектива, то функция рассеивания точки есть изображение блика, полученного в результате переотражений от поверхностей линз объектива. Результаты расчета функции рассеивания точки для полевых источников света представлены на рис. 10.

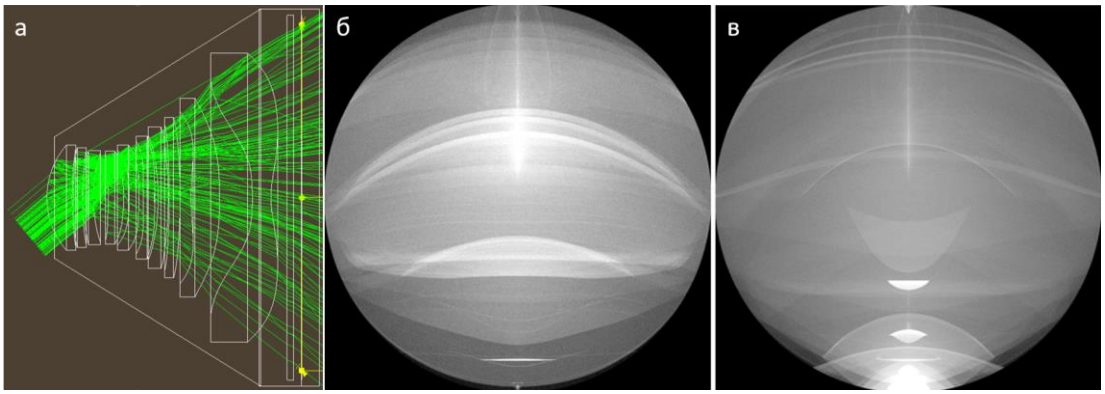


Рис. 9. Схема десятилинзового объектива (а) и результаты расчета функции рассеивания точки данного объектива методом федеративного рендеринга для внеполевых бесконечно удаленных источников света, находящихся под углами 45º (б) и 40º (в)

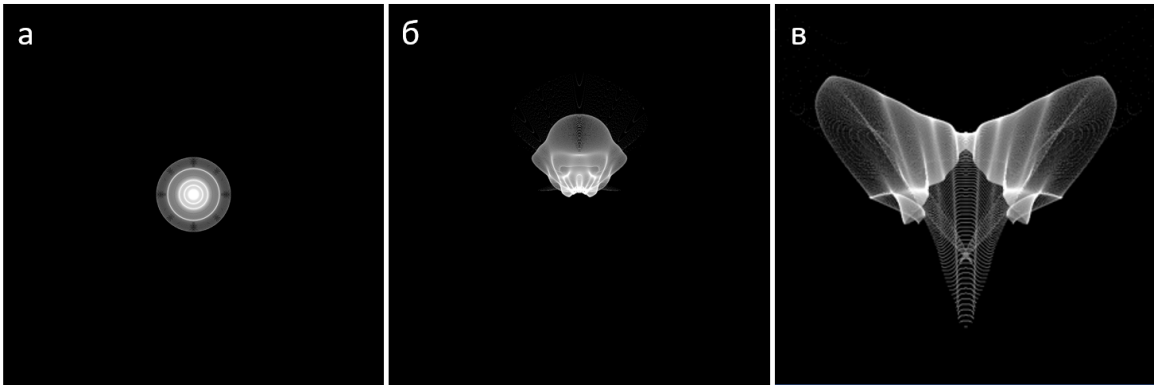


Рис. 10. Результаты расчета функции рассеивания точки десятилинзового объектива методом федеративного рендеринга для полевых бесконечно удаленных источников света, находящихся под углами 0º (а),28º (б) и 38º (в)

Во всех рассмотренных случаях результаты сравнивались с эталонными, полученными с использованием традиционных систем рендеринга. Полученные изображения совпадают с эталонными в рамках погрешности вычислений. Время работы системы федеративного рендеринга незначительно больше, чем в традиционной системе рендеринга, что связано с особенностями организации вычислений и задержками взаимодействия по сети. В случае увеличения задержек сети скорость рендеринга упадет, но незначительно, так как передаваемые объемы данных минимальны и скорость работы всей системы в целом останется приблизительно прежней. Кроме того, можно увеличивать размеры задач для выполнения на сервере заказчика, что позволит нивелировать задержки сети, особенно в случае длительных расчетов.

VIII. Выводы и заключение

В рамках данной работы была предложена, реализована и протестирована концепция системы федеративного рендеринга, которая разделяет систему рендеринга на две составляющие: сервисы системы рендеринга и сервисы заказчика, что позволяет защитить обе стороны: и заказчика, и разработчиков системы рендеринга. Конфиденциальность данных заказчика гарантируется тем, что данные заказчика никогда не покидают его сервер, на котором производятся все вычисления. Разработчики системы рендеринга в свою очередь защищены от нелицензионного копирования своего программного обеспечения тем, что алгоритмы рендеринга выполняются на их серверах, удаленно управляя данными заказчика в виде запросов на вычисление и получение ответов в виде минимальной статистической информации.

Система федеративного рендеринга строится на абстрактных интерфейсах, которые позволяют реализовать различные алгоритмы рендеринга, и при необходимости может быть расширена дополнительными методами. За счет малого объема данных, передаваемых между серверами заказчика и рендеринга, система федеративного рендеринга практически не уступает по скорости традиционной системе рендеринга, что было проверено на реализованном в рамках работы программном прототипе.

Источник финансирования. Благодарности

Работа выполнена при финансовой поддержке Российского научного фонда, грант № 22-11-00145.

Список литературы

1. Фролов В.А. Современное состояние методов расчёта глобальной освещенности в задачах реалистичной компьютерной графики. / В.А. Фролов, А.Г. Волобой, С.В. Ершов, В.А. Галактионов // Труды Института системного программирования РАН. 2021. Т. 33. №. 2. С. 7–48.

2. Kajiya J.T. The rendering equation // Proceedings of the 13th annual conference on Computer graphics and interactive techniques. 1986. P. 143–150.

3. Kay T.L., Kajiya J.T. Ray tracing complex scenes // ACM SIGGRAPH computer graphics. 1986. Vol. 20. No. 4. P. 269-278.

4. Hughes J.F. Computer Graphics: Principles and Practice. Third Edition / J.F. Hughes, A. Van Dam, M. McGuire, D.F. Sklar, J.D. Foley, S.K. Feiner, K. Akeley. USA, Addison-Wesley Professional, 2013. 1262 p.

5. Chattopadhyay S., Fujimoto A. Bi-directional ray tracing // Computer Graphics 1987. Springer, Tokyo, 1987. P. 335-343.

6. Lafortune E.P., Willems Y. Bi-directional path tracing // Compugraphics' 93. 1993. P. 145-153.

7. Miller G. S. Illumination and reflection maps / G. S. Miller, C. R. Hoffman // ACM SIGGRAPH. 1984. Vol. 4.

8. Jensen H.W. Global illumination using photon maps // Eurographics workshop on Rendering techniques. Springer, Vienna, 1996. P. 21-30.

9. Jensen H.W. Importance driven path tracing using the photon map // Eurographics Workshop on Rendering Techniques. Springer, Vienna, 1995. P. 326-335.

10. Zhdanov A.D., Zhdanov D.D. Progressive backward photon mapping // Programming and Computer Software. 2021. Vol. 47. No. 3. P. 185-193.

11. Veach E., Guibas L.J. Metropolis light transport // Proceedings of the 24th annual conference on Computer graphics and interactive techniques. 1997. P. 65-76.

12. Veach E., Guibas L.J. Optimally combining sampling techniques for Monte Carlo rendering // Proceedings of the 22nd annual conference on Computer graphics and interactive techniques. 1995. Pp. 419-428.

13. blender.org - Home of the Blender project - Free and Open 3D Creation Software [Электронный ресурс] / Blender Foundation. – Режим доступа: https://www.blender.org, свободный. (Дата обращения: 20.07.2024 г.).

14. 3D CAD Software: Shape the World We Live In | CATIA - Dassault Systèmes [Электронный ресурс] / Dassault Systèmes. – Режим доступа: https://www.3ds.com/products/catia, свободный. (Дата обращения: 20.07.2024 г.).

15. Niedenthal S. Learning from the Cornell box // Leonardo. 2002. Vol. 35. No. 3. P. 249-254.

16. Жданов Д.Д., Потемин И.С., Жданов А.Д. Использование технологий Embree для трассировки лучей в оптических системах с поверхностями свободной формы // Труды международной конференции по компьютерной графики и зрению "ГрафиКон" [International Conference on Computer Graphics and Vision, GraphiCon] - 2023. № 33. С. 97-107.

УДК 004.92
DOI: 10.25206/978-5-8149-3873-2-2024-225-237

Простой метод вычисления цвета краски металлик
A Simplified Method for Calculating Metallic Paint Color

С.В. Ершов, А.Г. Волобой
Институт прикладной математики им. М.В. Келдыша РАН, Москва, Россия
S.V. Ershov, A.G. Voloboy
Keldysh Institute of Applied Mathematics RAS, Moscow, Russia

Аннотация. Современные автомобильные краски имеют сложную структуру и состоят из множества пигментов различной природы. Моделирование красок «металлик», т.е. содержащих как окрашивающие пигменты, так и отражающие частицы, – сложная задача. В основном, в этих красках алюминиевые частицы отвечают за текстуру, а красящие диффузные пигменты за цвет. Это позволяет разделить задачу подбора краски на два этапа: отдельно подобрать вид алюминиевых частиц, а затем смоделировать изменение цвета с помощью красящих пигментов. В этой статье мы фокусируемся на задаче второго этапа и предлагаем упрощенный метод вычисления цвета окрашенного металлика. Сравнения с базовым точным методом, применяющим стохастическую трассировку лучей света в структуре краски, и с измеренными данными показали приемлемую точность предложенного метода. При этом наш метод значительно быстрее. Это позволяет использовать его в задаче оптимизации для подгонки цвета, где вычисление цвета производится многократно.

***Ключевые слова*:** автомобильная краска металлик, цвет краски, структура краски, диффузный пигмент, подгонка цвета, измерение краски

Abstract. Modern automotive paints have a complex structure and consist of many pigments of different nature. Modeling metallic paints which contain both coloring pigments and reflective particles, is a complex task. In these paints, aluminum particles are mainly responsible for the texture, and coloring diffuse pigments are responsible for the color. This allows us to divide the paint matching task into two stages: separately find the type of aluminum particles, and then model the color change using coloring pigments. In this paper, we focus on the second stage and propose a simplified method for calculating the color of metallic paint. Comparisons with the baseline exact method which uses stochastic ray tracing in the paint structure, and with measured data demonstrated acceptable accuracy of the proposed method. At the same time, our method is significantly faster. This allows us to use it in the optimization problem of color matching, where the color calculation is performed multiple times.

***Keywords*:** car paint metallic, paint color, paint structure, diffusion pigment, color matching, paint measurement

I. Введение

Современные автомобильные краски имеют сложную структуру и состоят из множества пигментов различной природы, таких как интерференционные пластинки, зеркальные чешуйки и обычные (мелкие частицы размером меньше длины волны) пигменты, помещенных в прозрачный лак. Внешней вид краски характеризуется ее цветом, яркостью, глянцевитостью поверхности, пространственной неоднородностью (текстурой) и т.д. В общем случае визуальные характеристики окрашенной краски поверхности выражаются двунаправленной функцией отражения (ДФО или BRDF, т.е. Bidirectional Reflection Distribution Function). Однако, эта функция не может передать неоднородность (текстуру) краски, часто свойственную металликам. Проблеме правильного отображения внешнего вида красок посвящено множество работ. Например, авторы статей [1–3] предлагают модели визуализации, основанные на анализе фотоизображений или результатах измерений реальных образцов краски.

Другой подход заключается в моделировании распространения света в слоях краски со сложной микроструктурой с целью воспроизведения их внешнего вида. Такое моделирование является достаточно сложной задачей и не всегда обеспечивает требуемую точность. Модель краски, состоящая из плоскопараллельных однородных слоев, была предложена в [4]. Она основана на статистическом подходе и

точно описывает взаимодействие света внутри краски, включая отражения на алюминиевых чешуйках и перламутровые эффекты на интерференционных пластинах. Модель эта оказалась вполне реалистичной и впоследствии использовалась и развивалась в [5–7].

Моделирование красок «металлик», т.е. содержащих (сравнительно) крупные отражающие пластинки, значительно сложнее, чем обычных диффузных красок, композиция которых состоит только из окрашивающих диффузных пигментов. Это связано как с использованием различных вычислительных методов, так и со сложностью задания корректных входных данных. При моделировании диффузных красок хорошо себя зарекомендовали методы расчета непрерывной среды, в то время как для красок металлик точность этих методов может быть недостаточной. Реальные алюминиевые частицы, используемые в красках металлик, могут быть не плоскими, иметь сложную форму многослойных чешуек. Поверхность их может быть полированной и тогда отражать луч зеркально, но может быть шероховатой со сложной функцией отражения. Все эти особенности часто неизвестны, и поэтому не все из них могут быть учтены при моделировании.

Одной из наиболее сложных и практически необходимых задач является получение состава краски (т.е. ее композиции — набора чешуек, пигментов и их концентраций), обеспечивающего заданный внешний вид, а именно цвет и текстуру. Чаще всего эта задача необходима при кузовных ремонтных работах, поскольку исходный состав краски может быть неизвестен. Более того, даже если состав краски известен, внешний вид окрашенной поверхности может меняться со временем под воздействием атмосферных явлений, реагентов, используемых на дорогах и т.д. В результате внешний вид поверхности может отличаться от первоначального. Несколько коллективов пытались найти автоматическое решение этой весьма сложной задачи [8–11]. Нами также были предложены общие подходы к решению этой проблемы [12, 13], а также проработаны различные аспекты ее решения [14-16].

II. Постановка задачи

В основном, в красках металлик алюминиевые частицы отвечают за яркость и текстуру, а красящие диффузные пигменты за цвет. Как уже говорилось, текстура не может быть передана одной функцией ДФО. Но из практики известно, что существует весьма небольшое число (порядка двух десятков) разных видов алюминиевых чешуек, используемых в автомобильных красках. Это позволяет разделить задачу подбора краски на два этапа: отдельно подобрать вид алюминиевых частиц (например, методами ИИ, обученными на наборе фотографий поверхности красок металлик), а затем смоделировать изменение цвета с помощью красящих пигментов. В такой постановке для второго этапа итоговая ДФО является критерием достижения результата, так как вопрос правильности текстуры решается на первом этапе.

В итоге, постановка задачи для второго этапа сводится к следующему. Существует серебристая краска металлик, в составе которой нет окрашивающих пигментов. Надо найти, какой диффузный пигмент необходимо добавить для получения требуемого цвета. На первом шаге необходимо научиться корректно моделировать цвет такого окрашенного металлика, т.е. вычислять ДФО поверхности краски. Следует заметить, что диффузный элемент добавляется в небольшой концентрации (обычно несколько процентов), частички металлика в краске преобладают.

В автомобильной индустрии для быстрого изменения цвета и текстуры краски часто используется портативный гониорефлектометр XRite T12 [17]. Поэтому в нашей задаче мы будем рассматривать ДФО поверхности для углов падения и наблюдения, которые могут быть получены с помощью этого прибора. Это позволит нам сравнить результаты нашего моделирования с измеренными данными реальных образцов краски, специально приготовленных по известным составам. И при изготовлении образцов, и при моделировании мы предполагаем, что краска нанесена толстым (укрывистым) слоем, и подложка не видна.

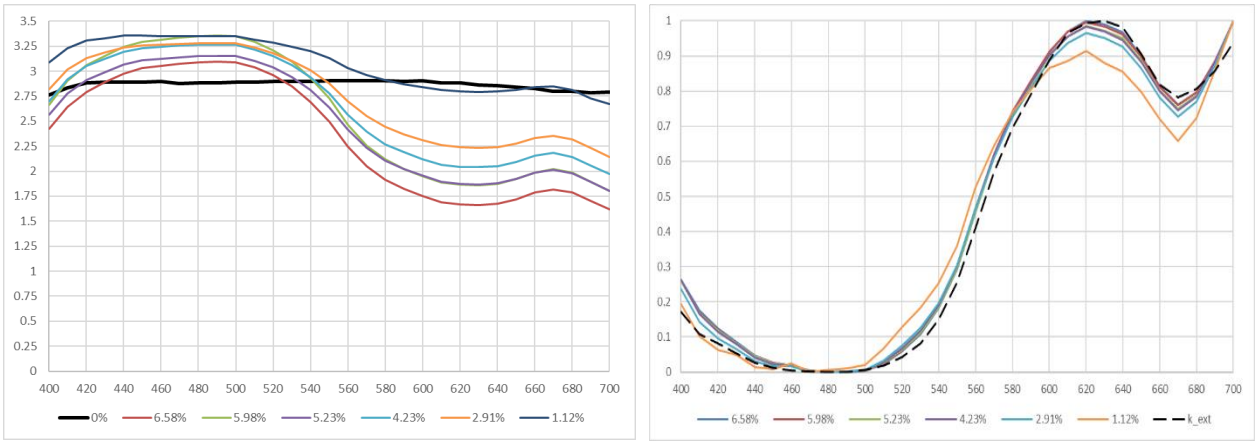
Также в данной работе делается допущение, что окрашивающий пигмент имеет слабое рассеяние и, в основном, является поглощающим.

III. Модель взаимодействия света с краской

Для экспериментов были выбраны две компоненты краски металлик: алюминиевые частицы TYM24 и диффузный красящий пигмент TJM51 из серии пигментов MARIPOSA [18]. Были изготовлены и измерены несколько образцов с разными соотношениями концентраций алюминиевых чешуек и красящего пигмента.

Как показал анализ измерений, спектр чистого металлика без красящего пигмента почти плоский (т.е. коэффициент яркости почти не зависит от длины волны), а добавление красителя приводит к заметной неоднородности этого спектра (рис.1а). Чтобы посмотреть, как краситель меняет цвет, возьмем отношение

$g(\lambda) \equiv \frac{f_0(\lambda)}{f(\lambda)} - 1$, где $f(\lambda)$ — спектр окрашенного металлика, а $f_0(\lambda)$ — спектр чистого металлика. При малой концентрации красителя эта величина близка к 0, а при более высоких концентрациях получается сильная вариация. Отнормируем это выражение: $\tilde{g}(\lambda) \equiv \frac{g(\lambda) - \min g(\lambda)}{\max g(\lambda) - \min g(\lambda)}$. Если теперь наложить «растянутые» кривые $\tilde{g}(\lambda)$ для разных концентраций пигмента, то они почти совпадают. Также они близки к спектру экстинкции красителя, аналогичным образом нормированному (рис. 1b). Совпадение не идеально точное, но весьма хорошее; слишком хорошее, чтобы быть случайным.



(a) (b)

Рис. 1. (a) Спектр коэффициента яркости слоя чистого TYM24 и его смесей с TJM51 в разных концентрациях. По вертикали — коэффициент яркости, по горизонтали — длина волны (нм). Различные цвета графиков соответствуют разным концентрациям пигмента. (b) Преобразование спектра в растянутых координатах (то есть $\tilde{g}(\lambda)$) для разных концентрация красителя TJM51. Штриховая линия — аналогичным образом растянутый спектр экстинкции пигмента TJM51

Надо также заметить, что экстинкцию пигмента мы знаем приблизительно. При измерениях пробных образцов для сильно поглощающих пигментов она оказывает меньшее влияние, а потому определяется с худшей точностью.

Коль скоро форма спектра почти фиксирована, то его можно однозначно задавать величиной $\max \frac{f_0(\lambda)}{f(\lambda)} - \min \frac{f_0(\lambda)}{f(\lambda)}$. Назовем ее «степень неоднородности». Она увеличивается с ростом концентрации красителя. Как видно из Рисунка 2, «размах» изменения спектра с высокой точностью линейно растет с концентрацией красителя. Форма же спектра примерно совпадает с (нормированным) спектром экстинкции.

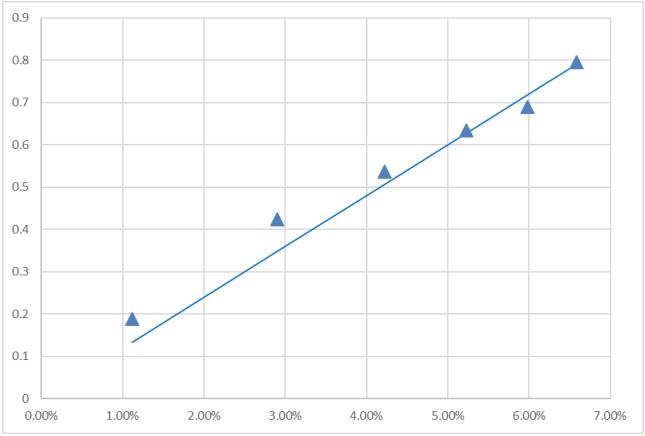


Рис. 2. Степень неоднородности преобразования спектра как функция концентрации TJM51. По вертикали — $\max \frac{f_0(\lambda)}{f(\lambda)} - \min \frac{f_0(\lambda)}{f(\lambda)}$, по горизонтали — концентрация TJM51. Прямая линия — наилучшее линейное приближение.

Поэтому можно ожидать, что спектр будет выражаться формулой

$$\frac{f_0(\lambda)}{f(\lambda)} \approx 1 + const \times k_{ext}(\lambda), \quad (0)$$

где $k_{ext}(\lambda)$ — спектр экстинкции при данной концентрации красителя (пропорционален этой самой концентрации), а множитель $const$ зависит от самого металлика, его концентрации, а также, возможно, углов падения и наблюдения. Но не от длины волны и не от свойств красителя.

Наблюдаемое поведение (постоянство нормированного спектра + линейная зависимость его «размаха» от концентрации) характерны для одностолкновительной модели. Для смеси с сильно преобладающей долей алюминиевых чешуек и чисто поглощающим пигментом красителя эта модель показана на рис. 3.

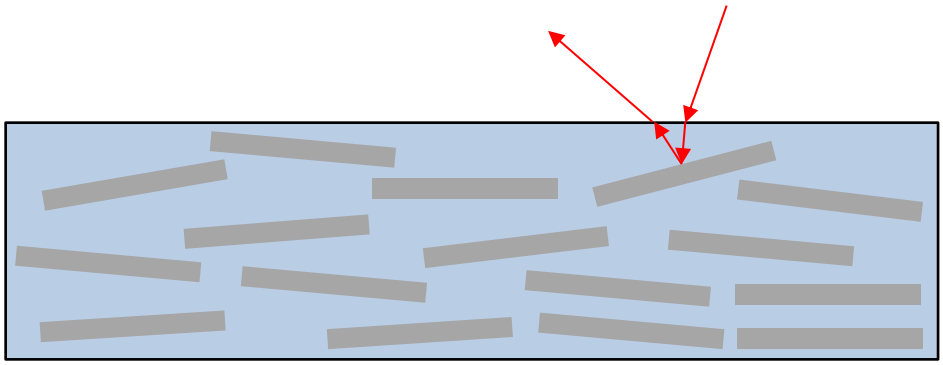


Рис. 3. Ход лучей света в одностолкновительной модели. Показано сечение слоя краски; серые прямоугольники — алюминиевые чешуйки, голубой фон — смола с введенным (голубоватым) красителем. Красные стрелки — путь света от источника к детектору.

Алюминиевые частицы могут быть и не плоскими (и даже в пределе вообще шариками). Поверхность их может быть полированной, и тогда отражать луч зеркально, или шероховатой. Однако самый простой случай — это тонкие плоские частицы с зеркально отражающей поверхностью. Для многих металликов такое предположение соответствует действительности.

В этой модели луч света меняет свой спектр, пока распространяется до алюминиевой частицы за счет поглощения красителем. Затем он меняет свой спектр при отражении от слегка не идеально белого алюминия (на Рис. 1a видно, что спектр отражения чистого металлика слегка спадает к краям диапазона — это спектр отражения и есть). И, наконец, он меняет спектр, пока идет от алюминиевой частицы к поверхности красящего слоя.

IV. Простой метод вычисления цвета металлика

1. Вывод метода

Итак, измеренные данные показывают, что поведение света в краске наиболее корректно описывается одностолкновительной моделью. Это не что иное, как модель ДФО (BRDF) 1-го порядка

Для слоя непрозрачных чешуек его ДФО 1-го порядка без красителя определяется уравнением (1) из [14]:

$$f_1(v,u) = t_F(v)t_F(u)\int_0^H a\,(v',u',z)\mathfrak{f}_1(v',u';z)dz, \quad (1)$$

где v',u' – преломленные лучи, $a(v',u',z)$ — это среднее затенение продукта вдоль обоих сегментов луча и $\mathfrak{f}_1(v',u';z)dz$ является ДФО 1-го порядка для слоя толщиной dz на глубине z, $t_F(v)$ — Френелевское пропускание границы воздух-краска при падении луча из воздуха в направлении v. Добавление экспоненциального поглощения света красителем вносит дополнительный множитель

$$f_1(v,u) = t_F(v)t_F(u)\int_0^H a\,(v',u',z)e^{-k_{ext}(\lambda)\ell}\mathfrak{f}_1(v',u';z)dz, \quad (2)$$

где $a(v',u',z)$ – это то же самое затухание из-за затенения чешуйками (т.е. та же функция, что и в уравнении (1)), член $e^{-k_{ext}(\lambda)\ell}$ учитывает поглощение красителем, а ℓ является суммой длин падающего и отраженного лучей.

В [14] уравнение (1) представляет собой аппроксимацию с коэффициентом $a(v',u',z)$, полученную для идеально выровненных чешуек нулевой толщины. Затем ДФО вырождается в дельта-функцию зеркального отражения. В действительности чешуйки имеют случайный наклон, и в результате ДФО представляет собой лопасть конечной ширины. Проведенные в [14] испытания показали, что результаты модели достаточно точны для чешуек со случайным изменением наклона, вплоть до значений, типичных для реальных красок.

Что касается толщины чешуйки, то она не существенна для чисто геометрического затенения $a(v',u',z)$, полученного в [14]. Но при наличии красителя она уже играет определенную роль. Для идеально выровненной чешуйки конечной толщины h сумма длин падающего и отраженного лучей, естественно, равна

$$\ell = \frac{z-\frac{h}{2}}{\cos\sigma'} + \frac{z-\frac{h}{2}}{\cos\gamma'} = \mu z - \mu\frac{h}{2}$$
$$\mu \equiv \frac{1}{\cos\sigma'} + \frac{1}{\cos\gamma'},$$

где z — глубина центра чешуйки, σ' — угол между направлением освещения и вертикалью после преломления на границе краска-воздух, γ' — угол между направлением наблюдения и вертикалью после преломления на границе краска-воздух, поэтому $\mu \equiv \frac{1}{\cos\sigma'} + \frac{1}{\cos\gamma'}$ является функцией направления освещения и наблюдения. Для измерительных точек прибора T12 оно довольно близко к 2. $k_{ext}(\lambda)$ — коэффициент экстинкции среды между чешуйками, очевидно

$$k_{ext}(\lambda) = \frac{\mathrm{PVC_{col}}}{1-\mathrm{PVC_{flake}}} k_{ext}^{(0)}(\lambda)$$

Рассмотрим случай непрозрачных чешуек алюминия. Для простоты будем считать, что это тонкие одинаковые диски. Случаи, когда они имеют некоторую прозрачность и изменение площади, были рассмотрены в [15], но эффект не очень важен, и здесь мы рассматриваем более простой случай.

Таким образом, формула (2) становится

$$f_1(v,u) = t_F(v)t_F(u)e^{k_{ext}(\lambda)\mu\frac{h}{2}}\int_0^H a\,(v',u',z)e^{-k_{ext}(\lambda)\mu z}\mathfrak{f}_1(v',u';z)dz$$

Как и в [14], это верно для идеально выровненных чешуек, тогда как для чешуек с некоторым уклоном эта формула — некоторое хорошее приближение, поэтому результаты, полученные на ее основе, следует использовать с определенной осторожностью. Наши эксперименты доказывают, что для наклонов частиц в реальных красках они остаются вполне справедливыми.

В случае, когда слой краски однороден по вертикали, $\mathfrak{f}_1(v',u';z)$ не зависит от глубины z и поэтому

$$f_1(v,u) = t_F(v)t_F(u)\mathfrak{f}_1(v',u')e^{k_{ext}(\lambda)\mu\frac{h}{2}}\int_0^H a\,(v',u',z)e^{-k_{ext}(\lambda)\mu z}dz \qquad (3)$$

Функция $a(v',u',z)$ определяется уравнением (6) из [14]. Оценим значение $2R/c$ (используемое в критерии этой формулы) для типичных чешуек. Предполагая наблюдение и освещение в плоскости падения

$$c \equiv \sqrt{\left(\frac{v_x}{v_z}-\frac{v_{o,x}}{v_{o,z}}\right)^2+\left(\frac{v_y}{v_z}-\frac{v_{o,y}}{v_{o,z}}\right)^2} = \frac{v_x}{v_z}-\frac{v_{o,x}}{v_{o,z}} = \tan\sigma'-\tan\gamma'$$

что для угла падения, близкого к нормальному, дает

$$c \approx \sigma'-\gamma' \approx \frac{\sigma-\gamma}{\eta}$$

Для таких измерительных точек T12, как «15as15», это дает $c \approx 1/6$, а для «15as–15» ≈ 0.5. Полагая радиус чешуек R=10 микрон, мы видим критерий $z \overset{?}{\lesseqgtr} 2R/c$ — это $z \overset{?}{\lesseqgtr} 120\ \mu m$. Обычно толщина слоя краски составляет менее 30 микрон, поэтому по всему слою $z \le 2R/c$, и, следовательно,

$$a(v',u',z) = e^{\frac{8b}{3\pi}}e^{-DSz}e^{-\frac{4b}{3\pi}\left(3\frac{cz}{2R}\arcsin\frac{cz}{2R}+\left(2+\left(\frac{cz}{2R}\right)^2\right)\sqrt{1-\left(\frac{cz}{2R}\right)^2}\right)} = e^{\frac{8b}{3\pi}}e^{-2b\left(\frac{cz}{2R}+\frac{2}{\pi}\frac{cz}{2R}\arcsin\frac{cz}{2R}+\frac{2}{3\pi}\left(2+\left(\frac{cz}{2R}\right)^2\right)\sqrt{1-\left(\frac{cz}{2R}\right)^2}\right)}$$

Используя аппроксимацию (7) из [14], можно воспользоваться более простой формулой

$$a(v',u',z) \approx e^{-2b\left(\frac{cz}{2R}+0.58\left(\frac{cz}{2R}\right)^2\right)}$$

Как мы видели выше, обычно $\frac{cz}{2R} \ll 1$ и $0.58\frac{cz}{2R} \le 0.3$. Поэтому

$$e^{-2b\left(\frac{cz}{2R}+0.58\left(\frac{cz}{2R}\right)^2\right)} = e^{-DSz}e^{-0.29DS\frac{c}{R}z^2} \approx e^{-DSz}\left(1-0.29DS\frac{c}{R}z^2\right)$$

Подставив его в наш интеграл (3) и предположив, что слой краски укрывистый (т.е. достаточно толстый), получим

$$f_1(v,u) \approx t_F(v)t_F(u)\mathfrak{f}_1(v',u')\frac{e^{k_{ext}(\lambda)\mu\frac{h}{2}}}{DS+k_{ext}(\lambda)\mu}\left(1-0.58\frac{cDS}{R}\frac{1}{(DS+k_{ext}(\lambda)\mu)^2}\right)$$

В отсутствии красителя формула 1-го порядка ДФО

$$f_1^{(0)}(v,u) \approx t_F(v)t_F(u)\mathfrak{f}_1(v',u')\left(\frac{1}{DS}-0.58\frac{cDS}{R}\frac{1}{(DS)^3}\right),$$

поэтому

$$f_1(v,u) \approx f_1^{(0)}(v,u)e^{k_{ext}(\lambda)\mu\frac{h}{2}}\frac{1}{1+\frac{k_{ext}(\lambda)\mu}{DS}}\frac{1-\frac{0.58\frac{c}{DSR}}{\left(1+\frac{k_{ext}(\lambda)\mu}{DS}\right)^2}}{1-0.58\frac{c}{DSR}} \tag{4}$$

Наше приближение (простой метод вычисления цвета металлика) состоит в том, чтобы применить то же соотношение к полной ДФО со всем порядкам рассеяния, следуя тому же закону:

$$f(v,u) \approx f^{(0)}(v,u)e^{k_{ext}(\lambda)\mu\frac{h}{2}}\frac{1}{1+\frac{k_{ext}(\lambda)\mu}{DS}}\frac{1-\frac{0.58\frac{c}{DSR}}{\left(1+\frac{k_{ext}(\lambda)\mu}{DS}\right)^2}}{1-0.58\frac{c}{DSR}} \tag{5}$$

Следует отметить, что

- μ и c зависят от направлений освещения и наблюдения, поэтому совершенно не зависят от состава краски;
- *DS* и *DSR* зависят только от металлических чешуек, поэтому они одинаковы для всех красителей и длин волн;
- зависимость от единственного красителя (через k_{ext}) очень проста, и поэтому может использоваться для расчета целевой функции в процессе оптимизации (подгонки, проектирования).

Обратите внимание, что основная часть масштабирования $\frac{f^{(0)}(v,u)}{f(v,u)} \approx 1+\frac{k_{ext}(\lambda)\mu}{DS}$ имеет именно вид (0), выведенного нами из поведения экспериментальных данных.

2. Преимущества расчета через масштабирование

Формула (5) получена на основе полной модели взаимодействия световых лучей со слоем чешуек, распределенных в цветном связующем. Таким образом, формально можно предсказать ДФО этой краски по ее составу. Почему мы выразили результаты в виде «масштабированной неокрашенной ДФО»?

Это связано с тем, что на практике прогнозирование изменения цвета (спектра) точнее, чем расчет полной ДФО. Самая простая причина: важнейшим параметром, который сильно влияет на ДФО красочного слоя, является распределение ориентации чешуек. Оно почти всегда неизвестно, поэтому формально мы вообще не можем вычислить ДФО. Но можно сделать некоторые правдоподобные предположения об этом распределении (например, что оно гауссово), изменить его параметры в разумном диапазоне и посмотреть, какие свойства полученной ДФО остаются стабильными по отношению к этому изменению. Именно эти характеристики мы можем затем использовать в ситуации неопределенности этого распределения.

Формула (5) в каком-то смысле делает именно это. Масштабирующий фактор с точностью до $f^{(0)}(v,u)$ совершенно не зависит от распределения ориентации чешуек. Это очевидно, так как он был рассчитан для идеально выровненных чешуек. Затем мы используем этот результат как приближение для реальных чешуек. Как

показывают тесты, это хорошее приближение. В результате мы прогнозируем цвет краски не с нуля, а как масштабирование ДФО для чистого металлика. Другими словами, это гибридная модель, в которой некоторые члены рассчитываются, а некоторые измеряются.

Кроме того, масштабирование (5) представляет собой простую алгебраическую функцию, ее можно легко вычислить. А для расчета полной ДФО требуется гораздо более затратная по времени трассировка лучей.

Резюмируя, мы предлагаем прогнозировать спектры окрашенного металлика, используя расчетное масштабирование измеренной ДФО чистого металлика, так как

1. точность значительно выше (по сравнению с расчетом с нуля);
2. для расчета масштабного коэффициента требуется меньше параметров (многие из тех, которые необходимы для расчета ДФО с нуля, не используются);
3. расчет масштабного коэффициента выполняется очень быстро и просто.

Более того, в практической работе можно сделать следующий шаг. Масштабный фактор (5) требует таких параметров чешуек, как их радиус R и произведение DS, а также толщина чешуйки h, которые трудно измерить напрямую. В этой ситуации можно просто подобрать эти три параметра, чтобы добиться наилучшего соответствия какому-то одному красителю. После этого мы сможем предсказать цвет краски, содержащей произвольный набор красителей в произвольных концентрациях.

V. Проверка метода

Как было сказано ранее, уравнение (1) и, следовательно, все, что из него выведено, формально верно только для идеально выровненных чешуек нулевой толщины. Для не выровненных чешуек конечной толщины это некоторое приближение, и необходимо проверить, насколько оно точно или неточно для реалистичных параметров. Кроме того, формально мы выводим все отношения к ДФО 1-го порядка, но применяем полученные отношения к полной ДФО. Далее, (5) предполагает, что чешуйки представляют собой диски, хотя на самом деле они имеют более сложную форму, и что рассеяние красителя незначительно по сравнению с его поглощением. Все эти допущения требуют проверки, насколько это работает в реальном случае (с рассеянием).

1. Вычислительная верификация: точная модель вычислений

Если краситель подчиняется закону Ламберта-Бера (экспоненциальное затухание) и волновые эффекты незначительны, мы можем вычислить ДФО композиции «металлик+краситель», используя очень подробную и надежную модель, которая работает с отдельными чешуйками, распределенными в смеси связующего и красителя. Взаимодействие с веществом между чешуйками использует модель непрерывной среды [14], тогда как взаимодействие света с чешуйками рассматривает их как реальные объекты с некоторой формой, размером, ориентацией и оптическими свойствами поверхности. В принципе можно принять во внимание, что настоящие чешуйки имеют форму «кукурузных хлопьев» (неправильную), не идеально плоские, их поверхность имеет некоторую шероховатость и так далее. Проблема в том, что все эти подробные характеристики обычно недоступны, и можно только постулировать некоторые правдоподобные значения, чтобы увидеть, оказывают ли они существенное влияние на итоговую ДФО. Если не оказывают, моделирование может использовать некоторые правдоподобные характеристики и при этом обеспечивать хорошую точность. Если, наоборот, неизвестные параметры (например, форма чешуек) имеют сильное влияние, то точный расчет ДФО невозможен без знания их фактических значений.

Но в любом случае, эта детальная модель позволяет проверить нашу упрощенную модель. Наиболее важным является то, что эта модель выводит коэффициент окраски из сравнения только ДФО 1-го порядка. Затем эти ДФО 1-го порядка рассчитываются с использованием модели «коррелированных световых событий» из [14], которая сама по себе является приближением. И, наконец, важно, что мы рассматривали только поглощение красителя и пренебрегали его рассеянием, поскольку оно переходит в более высокие порядки ДФО.

Поэтому имеет смысл проверить, насколько эти допущения изменяют результирующую ДФО, т.е. достаточна ли точность упрощенной модели. Или, по крайней мере, убедиться, что неточность окупается простотой расчетов. Здесь мы имеем в виду следующее. Мало того, что расчеты в точной модели настолько медленные, что ее трудно включить в процесс оптимизации при проектировании или подгонки краски. Но, кроме того, точная модель использует много параметров пигментов и их распределения, тогда как упрощенная модель оперирует скорее агрегированными значениями, которых гораздо меньше. Таким образом, для упрощенной модели более реалистично получить набор данных путем измерений.

2. Определение вероятных характеристик пигментов

Формально, поскольку эти тесты являются вычислительными, мы можем использовать любые предполагаемые характеристики пигментов (чешуек и окрашивающих пигментов), при условии, что они находятся в разумных пределах, т.е. того же порядка величины, что и в реальных красках. Но поскольку в дальнейшем мы будем проводить сравнение нескольких измеренных образцов краски, мы решили использовать набор параметров именно для этих составов.

Есть характеристики, которые известны явно, например, концентрации пигментов. Некоторые, например, экстинкцию красителя (т.е. ослабление света красителем) можно измерить, используя подход, описанный в [14]. Но многие параметры, особенно распределение ориентации чешуек, практически невозможно измерить напрямую. Для этого распределения мы попытались найти наиболее подходящие значения, которые обеспечивали бы совпадение измеренной и рассчитанной ДФО краски, состоящей только из металлических чешуек. Для других параметров, таких как толщина чешуек, слабо влияющих на неокрашенную ДФО, мы приняли некоторые разумные предположения, согласующиеся с микрофотографиями, данными производителя, известными измерениями и т.д.

Для получения параметров алюминиевых частиц мы использовали несколько образцов с чешуйками TYM24 в высокой концентрации, предоставленных ООО «Два стахановца». Они были измерены на нашем гониоспектрофотометре [19] и на приборе T12 [17]. Мы предполагали, что чешуйки представляют собой идеально плоские тонкие алюминиевые диски, равномерно распределенные в слое краски, а их ориентация случайна и определяется плотностью вероятности угла β между нормалью чешуйки и вертикальным направлением (т.е. распределение вращательно-симметрично). Мы имеем следующие параметры алюминиевых чешуек:

- концентрация;
- радиус;
- толщина;
- отражательная способность поверхности чешуйки $r_{fl}(\lambda)$;
- распределение ориентации чешуек $P(\beta)$.

Точно они неизвестны, но мы знаем некоторые реалистичные границы по данным производителей, по микрофотографиям, по измерениям отражательной способности чистого полированного алюминия. Затем мы варьировали указанные выше параметры в реалистичных пределах и наблюдали, как изменяется расчетная ДФО, приближается ли она к измеренным значениям или отклоняется от них. Мы не выполняли настоящую оптимизацию методом спуска, просто нашли некоторый вариант, выглядящий лучше остальных.

TYM24 представляет собой пигментную пасту, а не порошок из чешуек чистого алюминия. В пасте чешуйки распределены в связующем веществе с добавлением растворителей и других летучих компонентов. Сами по себе чешуйки не являются основным компонентом. Мы провели несколько экспериментов по удалению связующего вещества и получили значение объемной концентрации (PVC, Pigment Volume Concentration) около 13% по отношению к сухому TYM24, которую мы и использовали в дальнейшем. Для толщины частиц было взято более-менее реалистичное значение 0.5 микрона. Радиус чешуек можно было оценить по микрофотографии слоя сильно растворенного TYM24. В вычислительных экспериментах мы использовали три реалистичных значения: 5, 10 и 20 микрон. Отражательная способность чешуйки и распределение их ориентации, выбранные нами, представлены на рис. 4. Позже мы использовали эти значения радиусов, толщины, отражательную способность и распределение ориентации чешуек при расчете роли красителя.

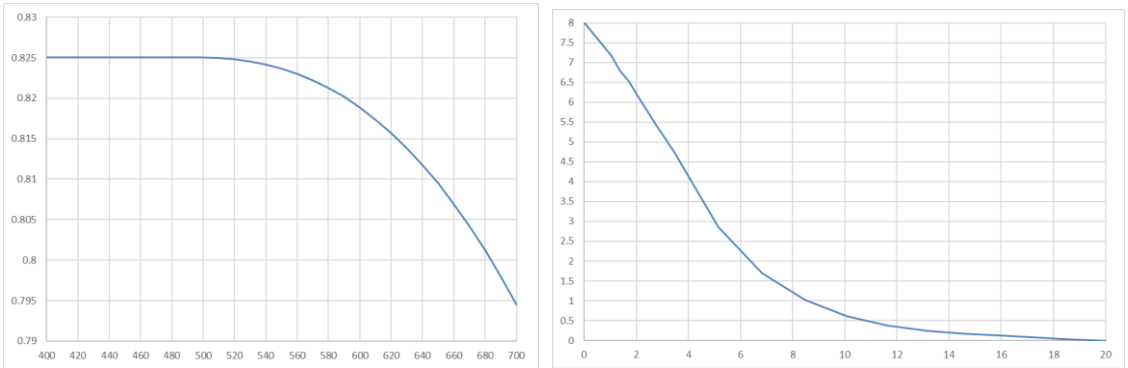


Рис. 4. Слева — спектр отражательной способности поверхности чешуек TMY24 (по вертикали — коэффициент отражения, по горизонтали — длина волны в нм), справа — распределение отклонения нормали

чешуек от вертикального направления (по вертикали — доля частиц с заданным наклоном, по горизонтали — угол наклона нормали частицы к вертикали, в градусах).

Для проверки выбранных параметров мы рассчитали ДФО только чешуек TYM24 без добавления других пигментов или смолы. Как видно на рис. 5 результат вычислений близок к измеренному образцу краски, содержащей только частицы TYM24.

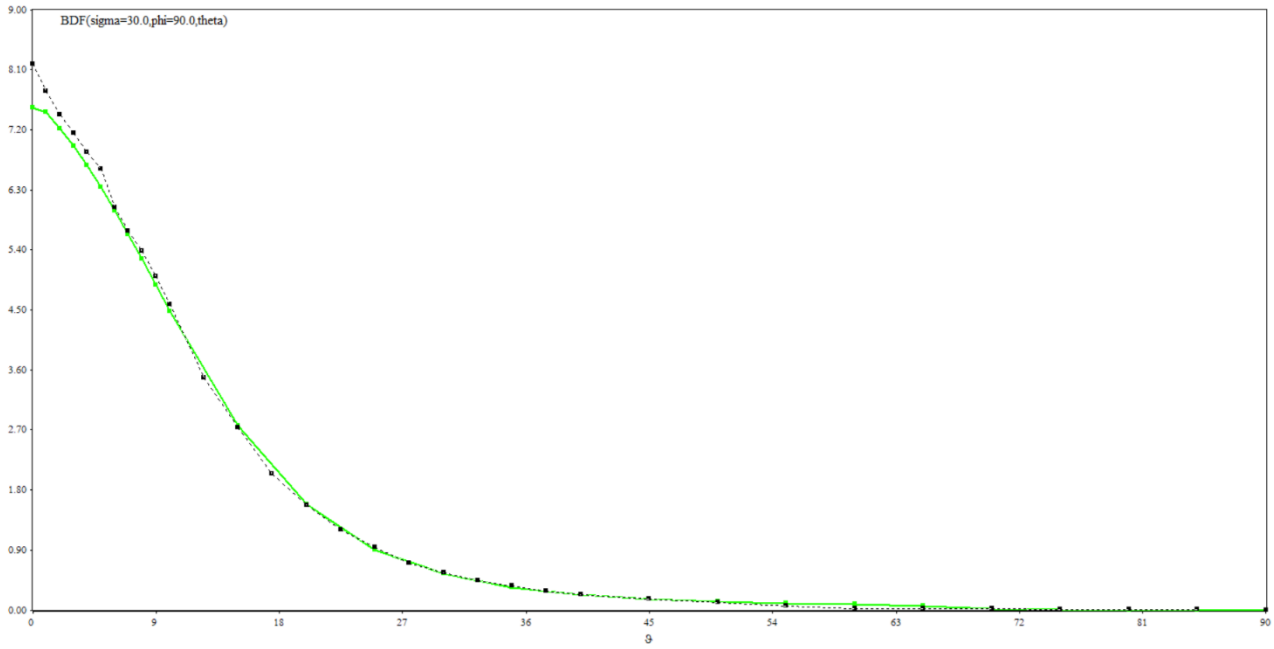


Рис. 5. Измеренная (яркая сплошная линия) и сосчитанная (пунктир) ДФО для образца с чистым TYM24. По вертикали — коэффициент яркости, по горизонтали — угол между направлением наблюдения и зеркальным лучом, в градусах, для случая угла падения 30 градусов.

Мы предположили, что красящий диффузный пигмент подчиняется модели непрерывной среды, и его рассеяние, если таковое имеется, может быть аппроксимировано фазовой функцией Хеньи-Гринштейна [20]. Таким образом, пигмент характеризуется экстинкцией, рассеянием и параметром асимметрии g. Для их измерения мы разработали и опробовали простой метод [16], давший весьма хорошие результаты. Для экспериментов мы использовали краситель TJM51, для которого у нас были приведенные выше данные измерений (графики экстинкции и рассеяния приведены на рис. 6), а также образец цветного металлика, состоящего из чешуек TYM24 и пигмента TJM51.

Были изготовлены и измерены 6 образцов с разными соотношениями концентраций алюминиевых чешуек и красящего пигмента (таблица 1). Концентрацию (PVC, Pigment Volume Concentration) пигментов в сухой краске рассчитывали на основе значений исходного состава в пасте (PWC, Pigment Weight Concentration, первые два столбца таблицы 1), как описано в [16]. Концентрация чешуек представляет собой просто концентрацию TYM24, умноженная на 0,13 (предполагаемая доля объема чешуек в сухом TYM24).

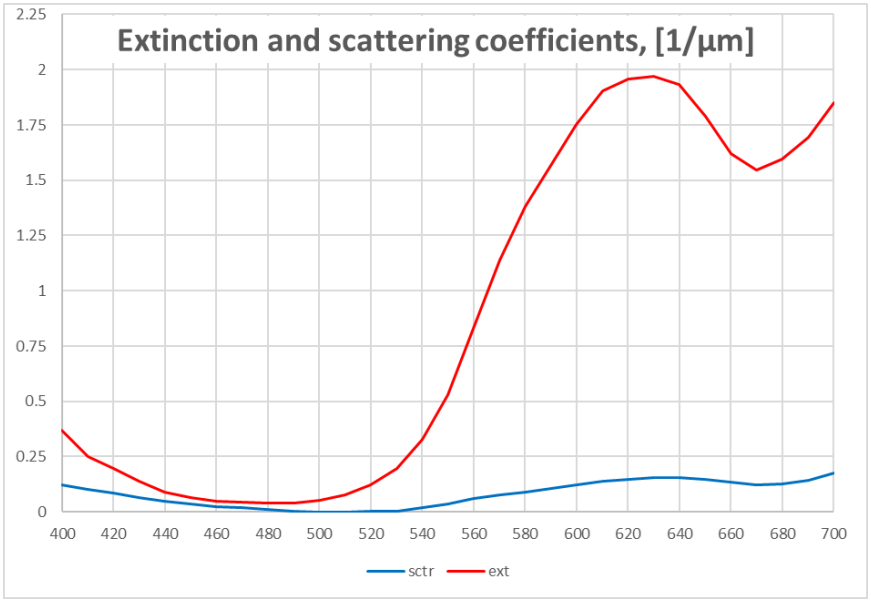


Рис. 6. Коэффициенты экстинкции и рассеяния чистого TJM51, по горизонтали — длина волны (нм).

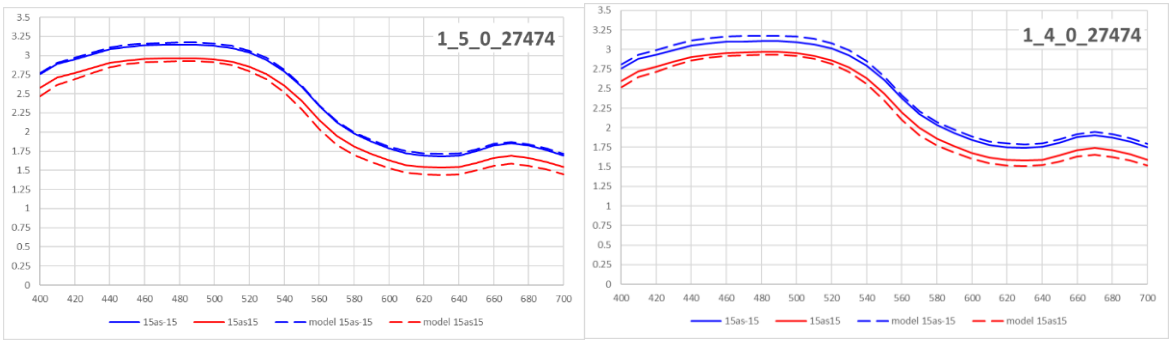
ТАБЛИЦА 1

Концентрации алюминиевых чешуек TYM24 и красящего пигмента TJM51 в тестовых образцах.

Образец	PWC_{wet}		PVC_{dry}		
	TYM24	TJM51	TYM24	TJM51	flakes
1_0_0_27474	99.0099%	0.9901%	98.88%	1.12%	12.1%
1_1_0_27474	97.4200%	2.5800%	97.09%	2.91%	12.2%
1_2_0_27474	96.2406%	3.7594%	95.77%	4.23%	12.3%
1_3_0_27474	95.3445%	4.6555%	94.77%	5.23%	12.5%
1_4_0_27474	94.6695%	5.3305%	94.02%	5.98%	12.6%
1_5_0_27474	94.1315%	5.8685%	93.42%	6.58%	12.9%

3. Сравнение с точной моделью вычислений

Сравнение результатов вычислений цвета окрашенного металлика точным и нашим упрощенным методами для всех образцов приведено на рис. 7. Разные цвета графиков соответствуют двум разным углам наблюдения.



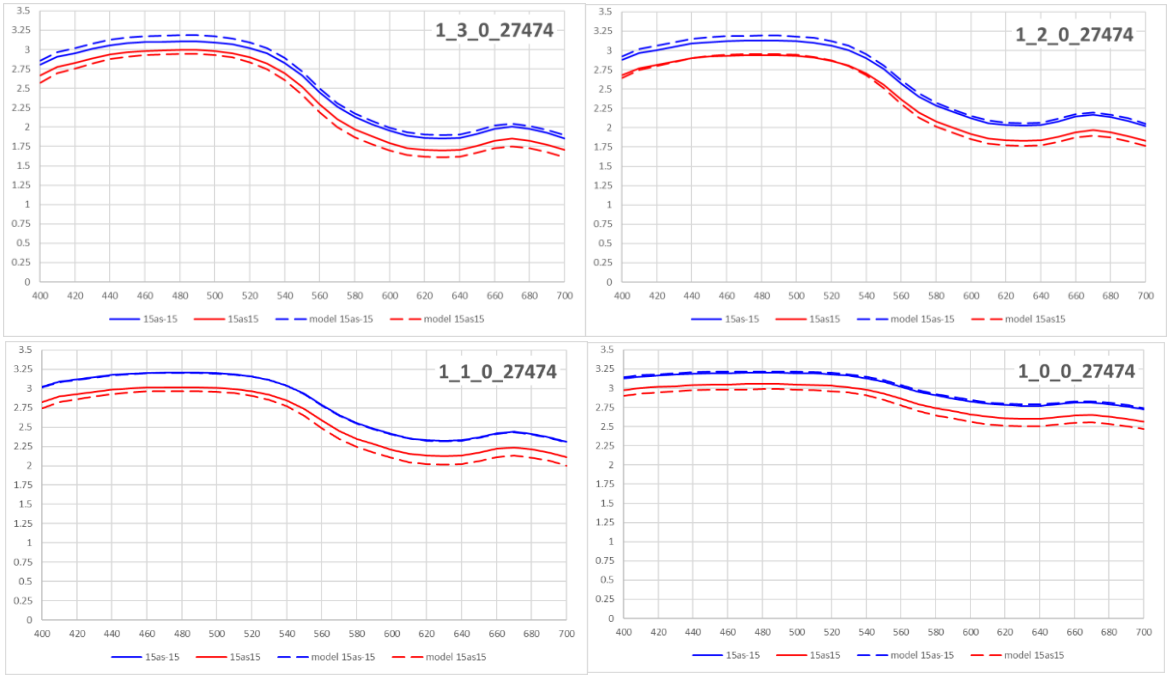


Рис. 7. Сравнение спектров, рассчитанных по точной модели (сплошные линии) с моделью, представленной уравнением (5) (пунктирные линии). По вертикали – коэффициент яркости, по горизонтали – длина волны (нм).

Из графиков, приведенных на рис. 7, мы можем заключить, что упрощенная модель может быть использована в качестве замены дорогостоящей полной модели. Хотелось бы заметить, это не доказывает, что она учитывает все физические эффекты и делает это точно, это просто доказывает, что те эффекты, которые мы учитывали, обрабатываются примерно так же точно, как и в полной модели.

4. Сравнение с измеренными данными

Сравнение результатов вычислений цвета окрашенного металлика нашим упрощенным методами с измеренными данными для всех образцов приведено на рис. 8. Сплошной линией показаны измеренные спектры, пунктирной — результат нашей упрощенной модели. Разные цвета графиков соответствуют двум разным углам наблюдения.

На рис. 8 можно увидеть довольно хорошее совпадение вычисленных и измеренных спектрограмм.

VI. Заключение

Нами был предложен и испытан упрощенный метод вычисления цвета краски металлик на основе ее композиции. Используя предложенный метод, мы прогнозируем цвет краски не с нуля, а как масштабирование ДФО для чистого металлика. Таким образом, это гибридная модель, в которой некоторые члены рассчитываются, а некоторые измеряются. Вычисление цвета краски требует существенно меньше вычислительных ресурсов, или другими словами, значительно быстрее (доли секунды вместо нескольких минут для точного расчета). Это позволяет использовать его в задаче оптимизации для подгонки цвета, где вычисление цвета производится многократно.

Как показало сравнение с измененными данными, предложенный нами метод оказался также точнее, чем базовый метод на основе, стохастической трассировки лучей. Из сравнения рисунков 7 и 8 видно, что результаты, полученные нашим упрощенным методом, ближе к измеренным данным. И это несмотря на то, что в базовом методе параметры алюминиевых чешуек, такие как распределение ориентации, толщина, диаметр и т.д., были подобраны так, чтобы рассчитанная ДФО чистого металлика была близка к измеренной. Другими словами, несмотря на подбор оптимальных параметров, базовый метод имеет ограниченную точность. При этом, если применить расчетное масштабирование (5) к измеренной базовой ДФО, то результирующие спектры предсказываются с вполне допустимой точностью (рис. 8).

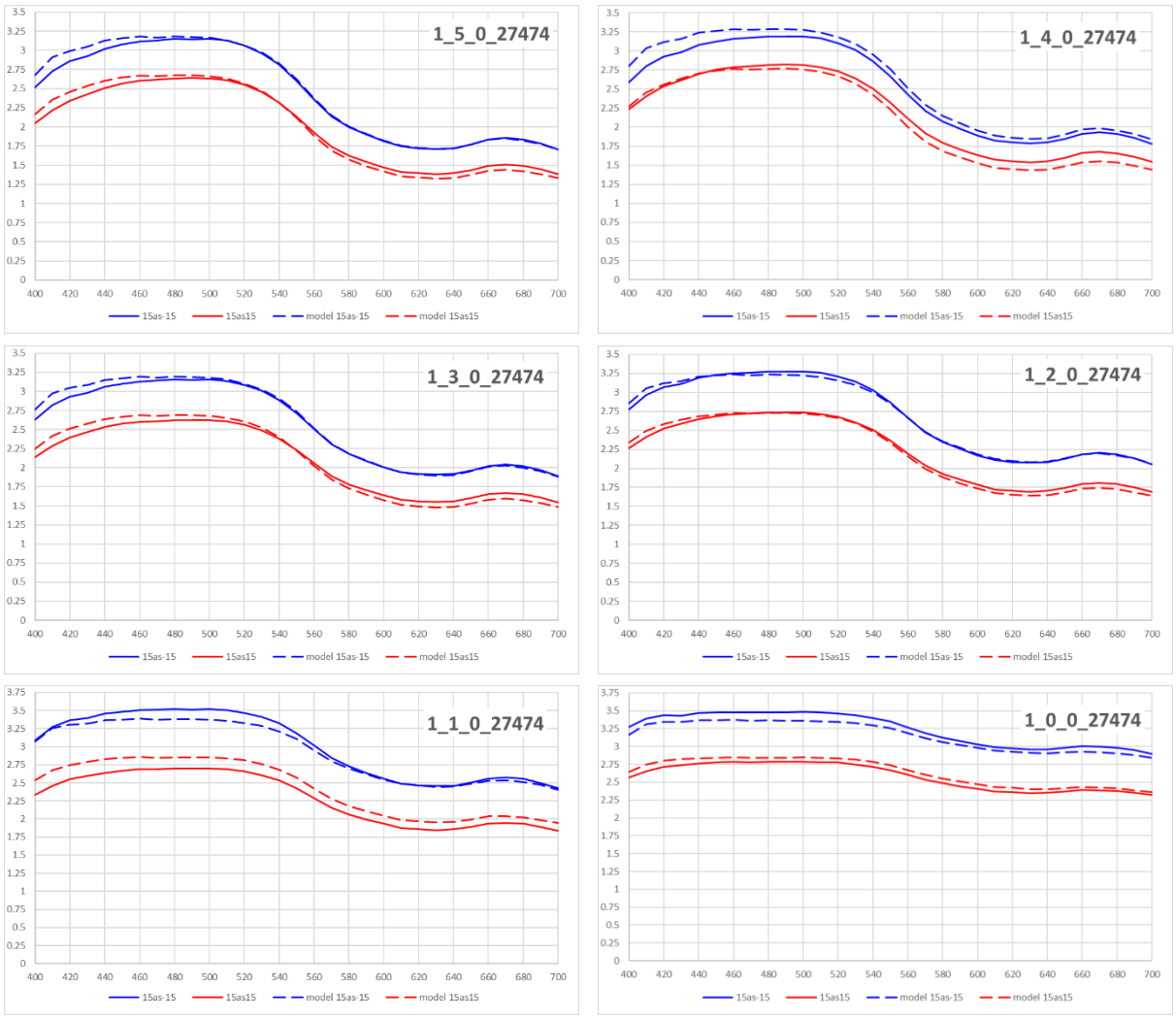


Рис. 8. Сравнение измеренных спектров (сплошные линии) со спектрами, вычисленными в соответствии с моделью, представленной уравнением (5) (пунктирные линии). По вертикали — коэффициент яркости, по горизонтали — длина волны (нм).

Благодарности

Авторы благодарят ООО «Два стахановца» (г. Челябинск) и лично Александра Кузнецова и Владислава Леушкина за подготовку и измерения образцов автомобильных красок, без которых это исследование было бы невозможно.

Список литературы

[1] M. Rump, G. Müller, R. Sarlette, D. Koch, R. Klein, Photo-realistic Rendering of Metallic Car Paint from Image-Based Measurements, Computer Graphics Forum 27(2) (2008): 527–536.

[2] T. Kneiphof, R. Klein, Real-time image-based lighting of metallic and pearlescent car paints, Computers & Graphics 105 (2022): 36–45. doi:10.1016/j.cag.2022.04.009

[3] D.B. Kim, M.K. Seo, K.Y. Kim, K.H. Lee, Acquisition and representation of pearlescent paints using an image-based goniospectrophotometer, Optical Engineering 49(4) (2010): 043604. doi:10.1117/1.3407431

[4] S. Ershov, K. Kolchin, K. Myszkowski, Rendering pearlescent appearance based on paint-composition modeling, Computer Graphics Forum 20(3) (2001): 227–238.

[5] W. Jakob, E. d'Eon, O. Jakob, S. Marschner, A comprehensive framework for rendering layered materials, ACM Transactions on Graphics 33(4) (2014): Article No. 118, 1–14. doi:10.1145/2601097.2601139

[6] L. Belcour, Efficient Rendering of Layered Materials using an Atomic Decomposition with Statistical Operators, ACM Transactions on Graphics 37(4) (2018): 1. doi:10.1145/3197517.3201289

[7] J. Guo, Y.J. Chen, Y.W. Guo, J.G. Pan, A physically-based appearance model for special effect pigments, Computer Graphics Forum 37(4) (2018): 67–76.

[8] S. Ergun, S. Onel, A. Ozturk, A general micro-flake model for predicting the appearance of car paint, EGSR'16: Proceedings of the Eurographics Symposium on Rendering: Experimental Ideas & Implementations (2016): 65–71.

[9] T. Golla, R. Klein, Interactive interpolation of metallic effect car paints, EG VMV'18: Proceedings of the Conference on Vision, Modeling, and Visualization (2018): 11–20. doi:10.2312/vmv.20181248

[10] H. Wu, J. Dorsey, H. Rushmeier, Physically-based interactive bi-scale material design, ACM Transactions on Graphics 30(6) (2011): 1–10. doi:10.1145/2070781.2024179

[11] M. Bati, P. Barla, R. Pacanowski, An inverse method for the exploration of layered material appearance, ACM Transactions on Graphics 40(4) (2021): Article No. 176, 1–15. doi:10.1145/3450626.3459857

[12] S. Ershov, K. Kolchin, K. Myszkowski, Reverse engineering approach to appearance-based design of metallic and pearlescent paints, The Visual Computer 20 (2004): 586–600.

[13] А.Г. Волобой, С.В. Ершов, С.Г. Поздняков, Интерактивное моделирование автомобильных красок, Труды 22-й Международной Конференции по Компьютерной Графике и Зрению – Графикон-2012 (2012): 242–247.

[14] Sergey Ershov, Alexey Voloboy, Boris Barladian, Vadim Sokolov, Igor Potemin, Dmitry Zhdanov. Light scattering in automotive paints: continuous medium approach vs correlations between particles // Proc. SPIE 11874, Illumination Optics VI, 1187408 (12 September 2021); DOI: 10.1117/12.2597166

[15] Ershov S., Voloboy A., Galaktionov V. Simulation of Light Scattering in Automotive Paints: Role of Particle Size // *Mathematics* **2023**, 11(11), 2429. https://doi.org/10.3390/math11112429

[16] Sergey Ershov, Alexey Voloboy and Sergey Pozdnyakov. Extraction of characteristics of solid pigments from paint samples // Proceedings of the 33 International Conference on Computer Graphics and Machine Vision, 2023, p. 146-156. https://doi.org/10.20948/graphicon-2023-146-156

[17] MA T12 Multi-angle Spectrophotometer. URL: https://www.xrite.com/-/media/xrite/files/literature/l10/l10-500_l10-599/l10-565_mat12_sellsheet/l10-565-ma-t12-sell-sheet-en.pdf

[18] Пигменты MARIPOSA, серии TYM и TJM. URL: https://www.paintbc.ru/TDS%20MARIPOSA/TJM%20TYM_%D0%A1%D0%B5%D1%80%D0%B8%D1%8F.pdf

[19] Волобой А.Г., Галактионов В.А., Ершов С.В., Летунов А.А., Потемин И.С. Аппаратно-программный комплекс для измерения светорассеивающих свойств поверхностей. Информационные технологии и вычислительные системы. 2006, № 4, с. 24–39.

[20] The Henyey-Greenstein phase function. URL: https://www.astro.umd.edu/~jph/HG_note.pdf.

УДК 004.92, 004.94
DOI: 10.25206/978-5-8149-3873-2-2024-238-249

Использование визуальной оболочки для реконструкции геометрии по набору RGB изображений с помощью дифференцируемого рендеринга
Using a Visual Shell for Geometry Reconstruction from a Set of RGB Images via Differentiable Rendering

А. И. Лысых, Д. Д. Жданов, М. И. Сорокин
ИТМО, Кронверкский пр-кт, д. 49, лит. А, Санкт-Петербург, 197101, Россия

A. I. Lysykh, D. D. Zhdanov, M. I. Sorokin
ITMO University, Kronverkskiy Prospekt, 49, St Petersburg, 197101, Russia

Аннотация. Использование методов дифференцируемого рендеринга является актуальным решением задачи реконструкции геометрии реального мира. Основными проблемами подобных методов являются возникающие в ходе оптимизации искажения геометрии и высокая вычислительная сложность. В статье рассматривается возможность использования трёхмерных визуальных оболочек в качестве первичного представления для оптимизации геометрии методами дифференцируемого рендеринга. Предлагается алгоритм формирования первичной геометрии объекта для дальнейшей оптимизации этой геометрии методами дифференцируемого рендеринга. Введение дополнительного этапа формирования геометрии позволяет не только увеличить точность восстановления геометрии, но и создать гомеоморфную форму для возможности корректной реконструкции геометрии объекта путём непрерывных преобразований параметров модели, что повышает универсальность упомянутых методов.

***Ключевые слова*:** дифференцируемый рендеринг, восстановление геометрии, реконструкция геометрии, визуальная оболочка, вогнутая оболочка, сегментация изображений, детектирование объектов, поиск контуров, определение силуэтов

Аннотация. The use of differentiable rendering methods is a relevant solution to the problem of reconstructing real-world geometry. The main issues with such methods are the geometric distortions that occur during optimization and the high computational complexity. The article considers the possibility of using three-dimensional visual hulls as a primary representation for optimizing geometry using differentiable rendering methods. An algorithm for forming the primary geometry of an object for further optimization of this geometry by differentiable rendering methods is proposed. The introduction of an additional stage of geometry formation allows not only to increase the accuracy of geometry reconstruction but also to create a homeomorphic form for the possibility of correct reconstruction of the object's geometry through continuous transformations of the model parameters, which increases the versatility of the mentioned methods.

***Ключевые слова*:** differentiable rendering, geometry restoration, geometry reconstruction, visual hull, concave hull, image segmentation, object detection, contour finding, silhouette detection

I. Введение

Задача реконструкции геометрии реального мира возникает во многих областях человеческой жизни, таких как сохранение культурного наследия, создание наборов данных для методов машинного обучения, медицинская визуализация, виртуальная и смешанная реальности, робототехника и автономные системы, археология и проч. Существует множество методов решения задачи реконструкции геометрии, основными из которых являются:

- Стереоскопическая реконструкция — методы, состоящие в использовании камер для получения перекрывающих изображений с известными внутренними и внешними параметрами [1];
- Фотограмметрия — методы, основанные на использовании нескольких изображений и определении ключевых точек для вычисления их пространственного положения [2];
- Структурированный свет — методы, использующие проекцию известного паттерна на объект, геометрия которого подлежит реконструкции, и дальнейший анализ искажений этого паттерна [3].
- Лазерное 3D сканирование — методы, использующие лазерный луч для измерения расстояний до точек в области видимости сканера с помощью учёта времени возвращения луча [4];
- Методы компьютерного зрения и машинного обучения — методы, основанные на анализе большого набора данных и поиска закономерностей в обрабатываемых изображениях объектов [5, 6, 7];

К другому классу методов, показавших высокую эффективность в решении задачи восстановления геометрии, относятся методы, основанные на использовании дифференцируемого рендеринга. Эти методы также называют методами обратного рендеринга — получения трёхмерной сцены из изображения [8].

В основе реконструкции геометрии методами с использованием дифференцируемого рендеринга лежит алгоритм обратного распространения ошибки (backpropagation) [8, 9]. Обычно этот алгоритм применяется в контексте обучения многослойных нейронных сетей для оптимизации сети путём минимизации функции потерь. Его идея состоит в распространении сигналов ошибки от выходов нейронной сети к её входам. Применение этого алгоритма в контексте дифференцируемого рендеринга имеет некоторые особенности: вместо оптимизации весов нейронной сети, алгоритм предполагает оптимизацию параметров рендеринга и направлен на минимизацию разницы между исходным и конечным изображением процесса обратного рендеринга.

В общем виде реконструкция геометрии с использованием метода дифференцируемого рендеринга состоит из следующего набора шагов, шаги 2–5 из которых повторяются циклично до достижения необходимой точности оптимизации геометрии:

1. Инициализация сцены — создание базовой формы геометрии, без которой начать цикл оптимизации геометрии невозможно;
2. Рендеринг сцены — расчёт яркости сцены с текущими параметрами, который может быть осуществлён как с помощью растеризации, так и с помощью методов физически-корректного рендеринга;
3. Вычисление функции потерь — эта функция характеризует разницу между синтетическим и целевым изображениями и служит основанием для оптимизации параметров сцены;
4. Вычисление градиентов — расчёт градиентов относительно входных параметров, которыми могут являться, например, положения вершин треугольников или прозрачность вокселя;
5. Оптимизация параметров геометрии — изменение параметров относительно вычисленных градиентов, например, с помощью методов градиентного спуска.

Существует несколько типов данных, которыми может быть представлена геометрия трёхмерных объектов в методах реконструкции геометрии с использованием дифференцируемого рендеринга: полигональная сетка, воксельная сетка, облако точек и неявные представления геометрии, используемые в комбинации с нейронными сетями. Каждый тип представления геометрии требует особенного метода рендеринга, имеет собственный набор оптимизируемых параметров и ограничений [8]. В рамках данной работы мы будем рассматривать реконструкцию треугольной сетки трёхмерных объектов.

Актуальным на сегодняшний день методом, реализующим реконструкцию треугольной сетки геометрии объектов с помощью дифференцируемого рендеринга, является метод Large Steps, предложенный B. Nicolet et al. в 2021 году [10]. Этот метод решает основной набор проблем, затрудняющих оптимизацию геометрии, представленной в виде треугольных сеток. Проблемы такой реконструкции состоят в возникающих в процессе самопересечениях геометрии, выворачивании некоторых областей сетки наизнанку или неадекватном распределении числа вершин на поверхности объекта. Метод основан на использовании предварительно обусловленного градиентного спуска, смещающего шаги оптимизации в сторону гладких решений.

В следующей работе авторы предлагают обобщить процесс coarse-to-fine оптимизаций путём регулировки средних размеров треугольника сетки относительно шага оптимизации [11]. Адаптивное изменение сетки поверхности на каждом этапе оптимизации позволяет избежать возникновения большинства критических искажений. Треугольники, потенциально ведущие к созданию проблем схлопываются при перестроении сетки.

Авторы выделяют два типа градиентов, получаемых с помощью дифференцируемого рендеринга: градиенты нормалей и градиенты силуэтов. Авторы отмечают, что одной из основных причин возникновения критических искажений является непостоянность второго типа градиентов. На рисунке 1 изображены возникающие при оптимизации искажения сетки. Для растущих регионов градиенты силуэта постоянно выталкивают одни и те же редкие вершины из силуэта, сохраняя все остальные вершины неизменными. Напротив, для сужающихся областей градиенты силуэта притягивают вершины внутрь силуэта, порождая самопересекающиеся треугольники. Кроме того, градиенты, получаемые из разных изображений, могут быть непостоянны, что ведёт к дополнительным искажениям формы геометрии.

Таким образом можно сделать вывод, что точность реконструкции геометрии треугольной сетки с использованием дифференцируемого рендеринга напрямую связана с числом и качеством возникающих артефактов оптимизации, а также эффективностью поиска оптимальной формы объекта методами градиентного спуска.

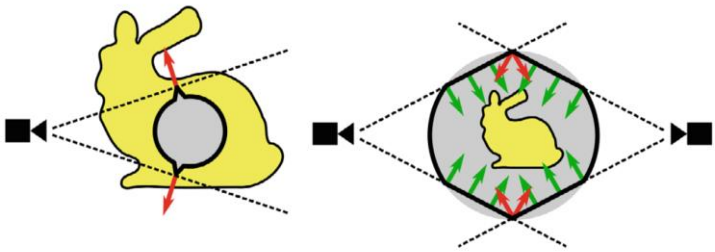


Рис. 1. Искажения, возникающие при оптимизации треугольной сетки. Два примера разреженных и непостоянных градиентов отмечены красными стрелками. Серая сфера оптимизируется по жёлтой целевой сетке. Слева: редкие градиенты силуэта выталкивают небольшое количество вершин, сохраняя при этом все остальные вершины неизменными. Справа: несовместимые градиенты силуэтов из двух изображений сминают сетку на пересечении конусов силуэтов [11].

II. Постановка задачи

Оптимизация геометрии в процессе реконструкции треугольной сетки представляет собой циклический процесс непрерывного преобразования трёхмерной формы. Из этого следует, что геометрия первичного объекта должна быть гомеоморфна геометрии эталонного объекта, то есть топологии этих объектов должны быть эквивалентны. В противном случае процесс оптимизации будет ограничен в своём стремлении достичь эталонной формы. Кроме того, соответствие первичной формы объекта искомой напрямую влияет на скорость оптимизации этой формы, что особенно актуально для объектов, имеющих геометрию, которая значительно отличается от геометрии примитивных объектов.

Из чего следует, что важным этапом, накладывающим определённые ограничения на возможность реконструкции геометрии описанным методом, является этап инициализации. Рассмотренные в первой главе работы, которые описывают реконструкцию геометрии с помощью дифференцируемого рендеринга, опираются на тот факт, что первичная форма реконструируемого объекта, представленная, как правило, в виде трёхмерного примитива эквивалентной топологии, заранее известна или выбрана вручную [8, 10, 11].

Как было указано в первой главе, основной причиной возникающих в ходе оптимизации искажений является неустойчивость градиентов силуэта. Такие искажения напрямую влияют на точность реконструкции геометрии методами дифференцируемого рендеринга, поскольку в некоторых случаях делают невозможной сходимость формы объекта за разумное время без введения регуляризации сетки, либо требуют значительного уменьшения скорости оптимизации [12].

В рамках данной работы рассматривается задача создания дополнительного этапа предобработки изображений для реконструкции геометрии из набора RGB изображений с помощью использования методов дифференцируемого рендеринга, который позволит не только сформировать первичную форму объекта, избежав ограничений топологии, но и увеличить точность реконструкции геометрии упомянутыми методами, путём минимизации возникающих артефактов и увеличения скорости сходимости оптимизируемой формы.

III. Теория

Для решения поставленной задачи предлагается сформировать первичное трёхмерное представление объекта с помощью формирования визуальной оболочки. Визуальная оболочка была представлена A. Laurentini et al. в работе 1994 года как концепция трёхмерной реконструкции с использованием метода Shape-From-Silhouette [13, 14]. Основной принцип метода заключается в создании трёхмерного представления объекта на основе его силуэтов, которые извлечены из нескольких изображений, полученных с разных точек наблюдения.

В оригинальной работе, формулирующей теоретической описание визуальной оболочки, как и во многих последующих работах, предлагающих алгоритм её формирования или использующих её [13-18], существует допущение о том, что силуэт объекта может быть выделен из эталонных изображений с высокой точностью. Задача выделения контура объекта на изображениях реального мира, является задачей сегментации изображения, и широко описана в различных работах [19-21]. В данной работе предлагается полный алгоритм формирования визуальной оболочки из набора изображений.

В актуальной работе, предлагающей метод реконструкции текстурированной треугольной геометрии с помощью дифференцируемого рендеринга [17], указывается возможность использования визуальной оболочки в качестве исходного приближения формы объекта, однако не указывается конкретный алгоритм её формирования, не делается ссылки на него, также не отражается степень влияния сформированной визуальной оболочки на скорость или точность оптимизации геометрии описанным методом. В данной работе предлагается оценка

влияния первичной формы объекта и её детализации на скорость и точность оптимизации треугольной сетки актуальными методами.

3. *Визуальная оболочка*

Визуальная оболочка представляет собой пересечение обобщённых конусов, сформированных с помощью поверхности силуэтов объекта из точек наблюдения камер. На рисунке 2 изображён пример построения визуальной оболочки бегущего человека на основе конусов силуэтов двух камер. При формировании конуса учитывается не только положение камер, но и их фокусное расстояние. Силуэт объекта проецируется из точки наблюдения и формирует основание конуса, точка наблюдения становится его вершиной. Пространство, описываемое каждым сформированным конусом, содержит в себе реконструируемый объект, таким образом пересечение этих пространств позволяет получить форму, ограничивающую полную форму объекта со всех видимых точек.

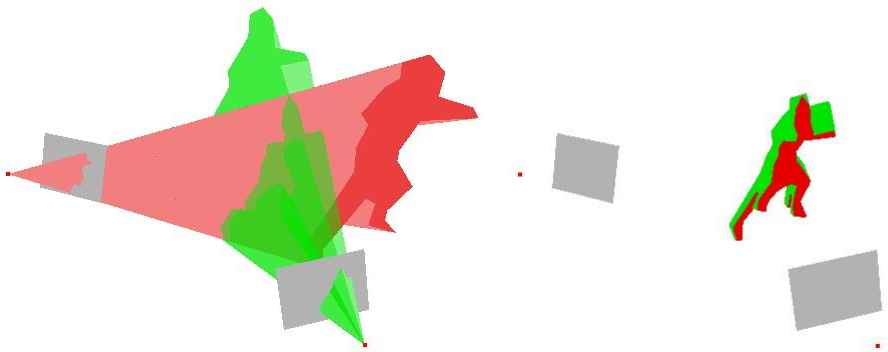


Рис. 2. Конусы силуэтов для двух виртуальных камер и визуальная оболочка бегущего человека, сформированная этими конусами

Кроме формирования описывающего объёма, приближенного к форме реконструируемого объекта, визуальная оболочка также формирует и необходимую топологию объекта. В случае, если реконструируемый объект содержит отверстия, хотя бы одно изображение каждого из отверстий формирует на визуальной оболочке объекта такие же отверстия. Таким образом визуальная оболочка является трёхмерным пространством, включающим в себя пространство формы реконструируемого объекта и гомеоморфным ему.

4. *Поиск силуэтов*

Для решения проблемы получения силуэтов из изображений реального мира предлагается использовать нейросетевые решения, специализируемые на сегментации изображений. Для корректной сегментации изображения, как правило, требуется сначала детектировать необходимый объект, то есть сформировать описывающий его прямоугольник. Таким образом для формирования силуэтов объектов потребуется использовать две модели нейронных сетей: детекции и сегментации.

Введение в процесс реконструкции нейросетевых алгоритмов может лишить методы дифференцируемого рендеринга одного из главных преимуществ: универсальности применения, которое заключается в том, что эти методы не ограничены классами реконструируемых объектов, в отличии от большинства нейросетевых методов. Для исключения этой проблемы предлагается использовать нейронные сети, созданные для решения задач класса "zero-shot learning" [22], то есть таких задач, решение которых не предполагает обучения на всех классах возможных объектов. Такие нейронные сети способны обобщать способы решения поставленной перед ними задачи, что позволит сохранить универсальность метода.

Как правило, модели сегментации изображений в качестве результата сегментации производят битовую маску. Для использования этой маски в качестве контура, формирующего основание конуса, требуется преобразовать эту маску в набор граней и вершин, то есть извлечь из полученного силуэта контуры.

5. *Формирование первичного представления геометрии*

Существует два основных подхода к построению визуальной оболочки: поверхностный и объёмный. Первый подход опирается на вычисление пересечения полигональных конусов зрения, второй — на регуляризацию пространства и объединение ячеек воксельной сетки, соответствующих силуэтам объекта [23]. За счёт отсутствия регуляризации первый подход оказывается более точным, однако требует больших вычислительных ресурсов, потому является предпочтительным в нашем случае.

Для формирования конусов из контуров используется экструзия и примитивные преобразования положения и размера сформированных плоскостей в пространстве относительно оси камеры. Контуры отверстий и формы должны быть обработаны отдельно, поскольку они являются независимыми друг от друга замкнутыми кривыми.

В простейшем случае формирование визуальной оболочки из конусов происходит с помощью булевых операций пересечения конусов, сформированных по всем изображениям. Более производительным методом, позволяющим снизить вычислительную сложность, является вычисление пересечений в двумерном, а не в трёхмерном пространстве. Описание эффективного алгоритма, который использует правила эпиполярной геометрии для упрощения задачи вычисления точек пересечения, можно найти в работе W. Matusik et al. [24].

Для сохранения топологии реконструируемого объекта важной деталью формирования визуальной оболочки является необходимость формирования отверстий. Упрощённый метод вычисления визуальной оболочки не позволяет произвести исключение контуров отверстий во время объединения конусов. Решением этой проблемы предлагается раздельная обработка конусов формы объекта и отверстий. Применив метод формирования визуальной оболочки к значащим конусам, возможно исключить конусы отверстий из полученной оболочки с помощью булевых операций.

Предлагается представить полный алгоритм формирования первичного представления объекта из набора изображений объекта, включающий метод формирования силуэтов, следующим образом:

1. Детектирование объекта — поиск описывающего прямоугольника, включающего объект на каждом изображении;
2. Выделение силуэтов — создание бинарного изображения, значащие пиксели которого содержат объект;
3. Поиск контуров — построение вершин, соединённых рёбрами обозначающих контуры формы или отверстий в силуэте объекта;
4. Формирование конусов — построение усечённых конусов относительно точек наблюдения камер из полученных контуров;
5. Пересечение конусов — этап, объединяющий результат формирования конусов видимости каждой из камер, осуществляющийся отдельно для конусов формы и конусов отверстий;
6. Исключение отверстий — этап, формирующий отверстия в полученной на предыдущем этапе визуальной оболочке с помощью исключения суммы их объёмов;
7. Реконструкция сетки — окончательный этап построения первичного представления геометрии объекта, необходимый для уменьшения искажений, возникающих в результате оптимизации, который позволит установить оптимальный уровень детализации первичной сетки.

Описанный алгоритм визуально отображён на рисунке 3. Представлены формируемые в ходе выполнения алгоритма объекты данных и операции над ними.

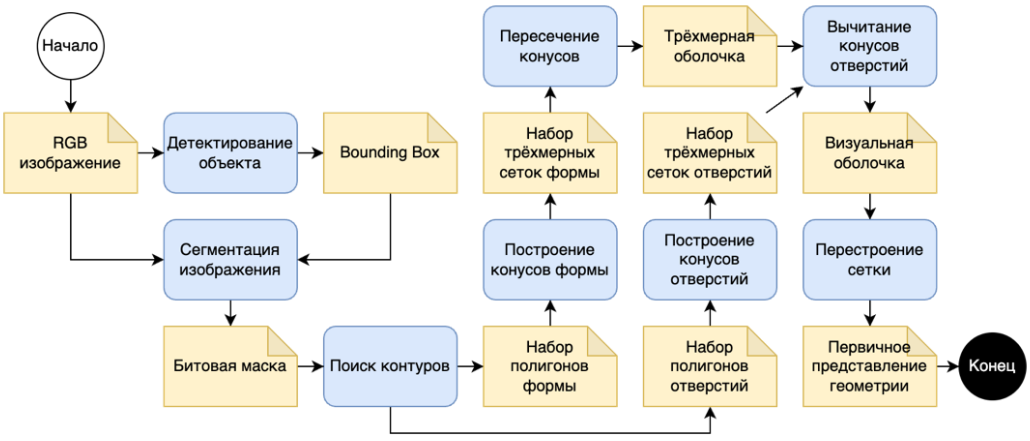


Рис. 3. Схема предложенного алгоритма формирования первичного представления геометрии для дальнейшей оптимизации с использованием метода дифференцируемого рендеринга. На схеме представлены формируемые в результате выполнения алгоритма объекты данных и операции над ними.

Первичное представление объекта, сформированное предложенным алгоритмом, используется в качестве исходной геометрии для преобразований её методами реконструкции с помощью дифференцируемого рендеринга. В качестве этих методов могут быть использованы любые методы, работающие с данными, представленными полигональной сеткой.

IV. Реализация

В качестве нейронной сети для детектирования объектов была использована модель Grounding DINO предложенная в 2024 году, которая позволяет создать описывающий прямоугольник искомого объекта по текстовому обозначению. Эта модель является моделью детектирования с открытым словарём (Open-Set Object Detection), и основана на комбинации детектора DINO [25] и Grounded Language-Image Pre-training [26], что обеспечивает мультимодальность сети и детекцию классов, не содержащихся в тренировочных данных.

Для выделения силуэтов путём сегментации изображений по найденному описывающему прямоугольнику использована предложенная в 2023 году модель Segment Anything in High Quality (SAM-HQ), которая является прямым последователем модели Segment Anything. Модель SAM-HQ обеспечивает zero-shot сегментацию изображений по набору точек или описывающему прямоугольнику. Результатом сегментирования такой модели является битовая маска, положительные значения которой обозначают принадлежность пикселя силуэту объекта.

Поиск контуров из битовой маски обеспечивается с помощью алгоритма, предложенного Satoshi Suzuki et al. в 1985 году [27]. Этот алгоритм преобразует бинарное изображение в набор контуров, обозначающих форму или отверстия силуэта. Формирование контуров осуществляется с помощью обхода пикселей на границах областей изображения, содержащих отличные друг от друга значения. Алгоритм реализован в библиотеке OpenCV [28].

Формирование конусов происходит в несколько шагов. В первую очередь осуществляется определение типа контуров: внешняя граница или отверстие. Далее происходит нормализация размера контура и положения контура в пространстве — контур располагается таким образом, чтобы его положение на изображении камеры соответствовало контуру реального объекта. После чего для каждого отдельного контура формируется экструзия из точки наблюдения, соответствующей ему камеры. Размеры сформированных экструзией плоскостей рассчитываются относительно их расстояния до точки наблюдения и фокусного расстояния камеры. К полученному усечённому контуру применяется матрица трансформации камеры для преобразования локальных координат камеры в координаты сцены. С помощью булевых операций конусы отверстий вычитаются из конусов формы.

Пересечение конусов осуществляется с помощью алгоритма, предложенного в работе W. Matusik et al. [24]. Исключение отверстий осуществляется с помощью булевых операций вычитания треугольных сеток полученных конусов отверстий из визуальной оболочки. Для осуществления булевых операций треугольные сетки представляются в виде топологического многообразия, что обеспечивается усечённостью конусов. Дополнительная операция перестроения сетки обеспечивает возможность регулировки разрешения треугольной сетки для её дальнейшей оптимизации.

В качестве метода для реконструкции геометрии с использованием дифференцируемого рендеринга применяется комбинация метода Large Steps in Inverse Rendering of Geometry [10] и Continuous remeshing for inverse rendering [11]. Первый метод обеспечивают сглаженный расчёт градиентов, исключая возникающие в результате оптимизаций артефакты, такие как пересекающиеся плоскости или переворот нормалей, второй метод обеспечивает высокую точность реконструируемых объектов благодаря постоянному перестроению треугольной сетки. Обратный рендеринг изображений с использованием этих методов осуществляется с помощью Mitsuba 3 [29] или Nvdiffrast [30].

V. Результаты экспериментов

Реализован предложенный алгоритм формирования первичного представления геометрии реконструируемого объекта из набора изображений, включающий полный алгоритм формирования визуальной оболочки. Сеть Grounded DINO предобучена на наборе данных Common Objects in Context (COCO), который включает более 330 тысяч изображений, разметку объектов и маски [31]. Сеть Segment Anything in High Quality предобучена 44 тысячах высококачественных масок.

Для формирования эталонных изображений созданы сцены со следующими тестовыми объектами на неоднородном фоне: suzanne, teapot, dragon, lucy, bunny, armadillo. Каждая сцена содержит восемь камер, наблюдающих тестовые объекты с различных ракурсов. Эталонные изображения получены путём прямого рендеринга с помощью Mitsuba 3. На рисунке 4 изображены эталонные изображения, полученные описанным методом для перечисленного набора сцен с одного ракурса.

Рис. 4. Эталонные изображение тестовых сцен с одного ракурса. Слева направо: bunny, teapot, Suzanne, lucy, dragon, armadillo.

Реализован предложенный алгоритм формирования первичного представления геометрии. Для каждой из сцен произведено формирование первичного представления геометрии с помощью реализованного алгоритма. На рисунках 5 и 6 изображены этапы формирования первичного представления геометрии для двух сцен: bunny и lucy.

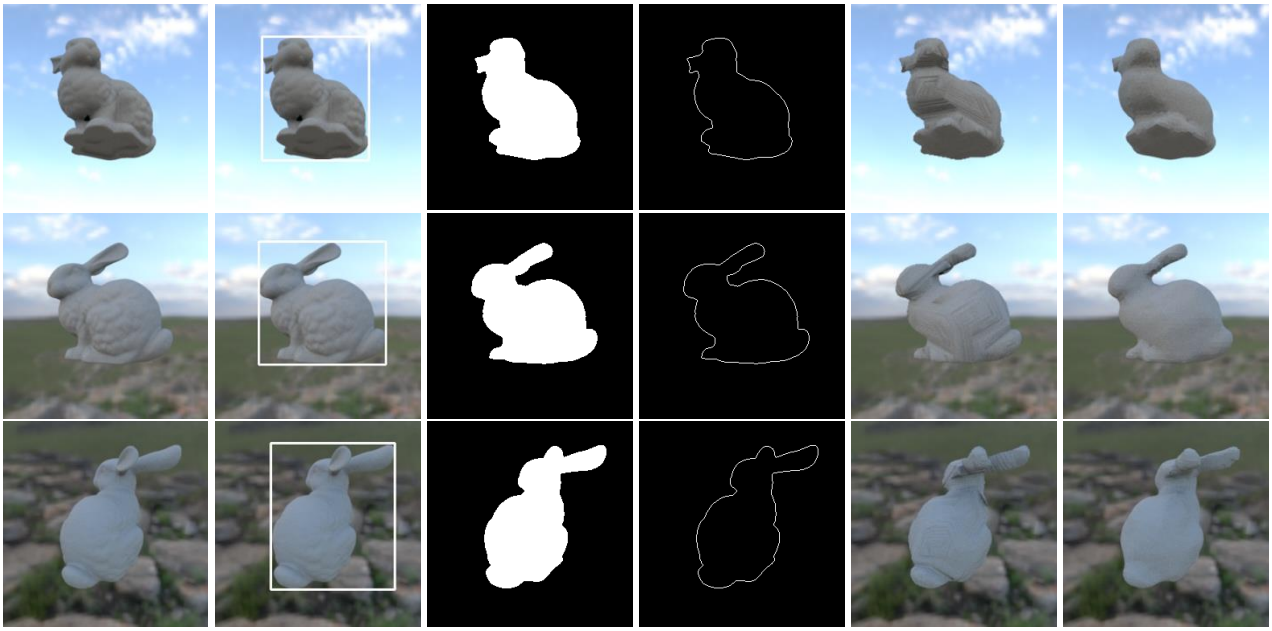


Рис. 5. Этапы формирования визуальной оболочки для тестовой сцены bunny показанные с нескольких ракурсов. Слева направо: исходное изображение, описывающий прямоугольник детектирования, сегментированный силуэт, контуры силуэта, визуальная оболочка, финальный результат предложенного алгоритма после перестроения сетки.

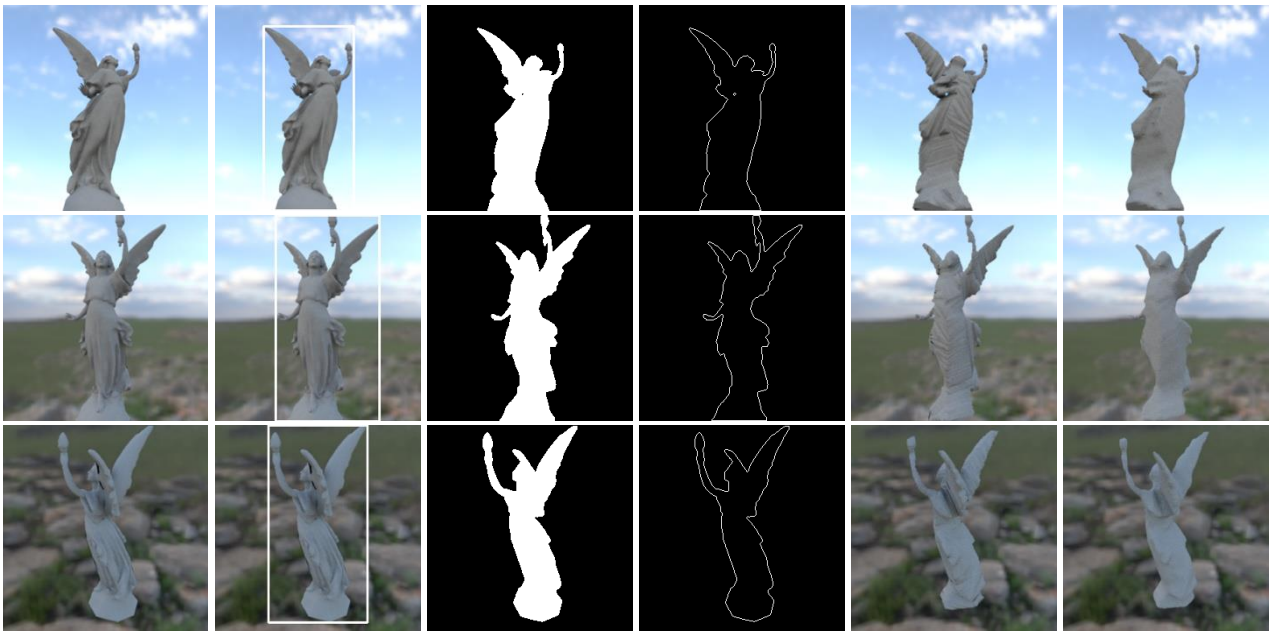


Рис. 6. Этапы формирования визуальной оболочки для тестовой сцены lucy показанные с нескольких ракурсов. Слева направо: исходное изображение, описывающий прямоугольник детектирования, сегментированный силуэт, контуры силуэта, визуальная оболочка, финальный результат предложенного алгоритма после перестроения сетки.

Для каждого набора изображений произведено формирование предложенного первичного представления и последующая реконструкция геометрии эталонного объекта с использованием метода Large Steps in Inverse Rendering of Geometry, реализованного в Mitsuba 3. Реконструкция геометрии осуществлялась в 32 итерации. В качестве исходной модели геометрии для оптимизации в первом случае использован примитив, топология которого соответствует топологии эталонного объекта сцены, во втором случае использовано предложенное первичное представление геометрии. Для полученных сеток геометрии рассчитана метрика Хаусдорфа по отношению к эталонной сетке объекта. На рисунках 7–12 изображены исходные модели геометрии объектов, промежуточные результаты оптимизации и эталонные объекты для каждой из шести сцен в случайном ракурсе. На таблице 1 представлены результаты сравнения времени оптимизации, времени этапа предобработки, а также метрики Хаусдорфа для каждой сцены в двух случаях: с использованием первичного представления и без.

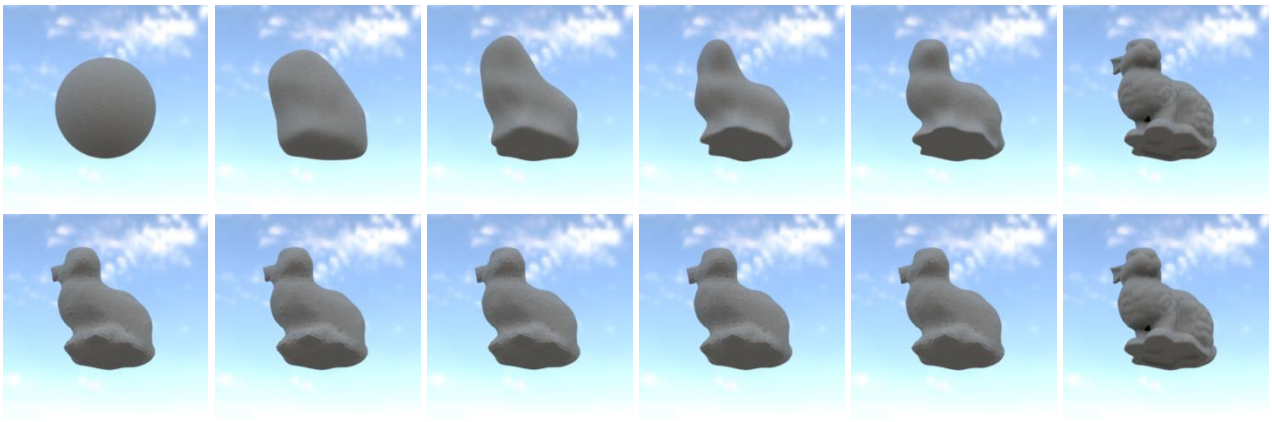
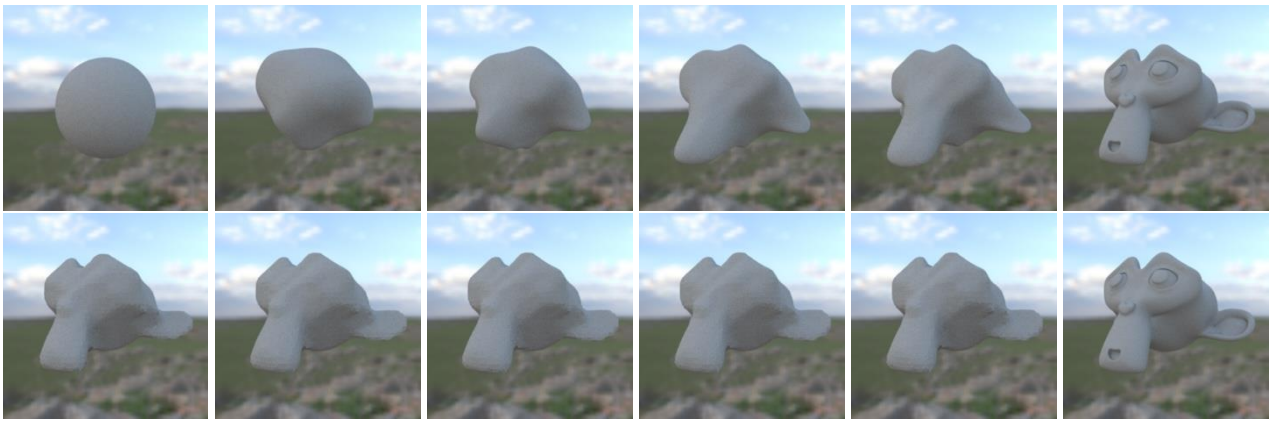


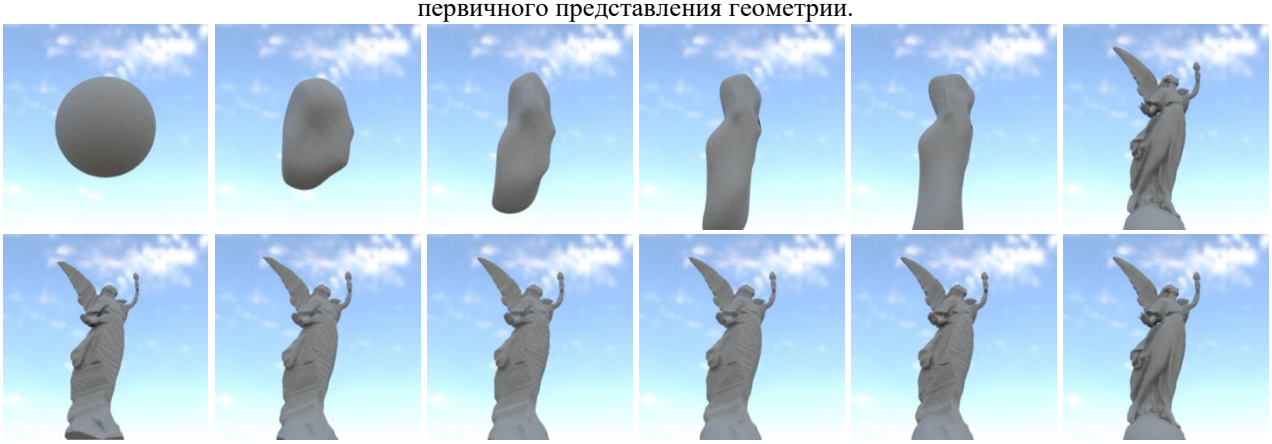
Рис. 7. Исходные модели геометрии, промежуточные этапы оптимизации геометрии и эталонное изображение для сцены bunny. Сверху оптимизация осуществлена для примитива (сферы), снизу для предложенного первичного представления геометрии.

Рис. 8. Исходные модели геометрии, промежуточные этапы оптимизации геометрии и эталонное изображение для teapot. Сверху оптимизация осуществлена для примитива (тороида), снизу для предложенного первичного представления геометрии.



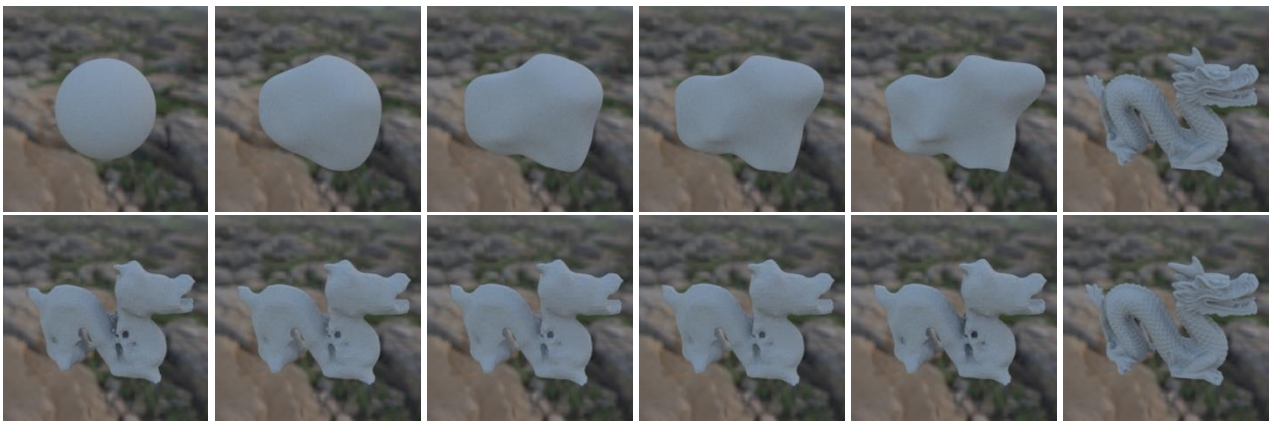
initial 8th iteration 16th iteration 24th iteration 32nd iteration reference

Рис. 9. Исходные модели геометрии, промежуточные этапы оптимизации геометрии и эталонное изображение для сцены suzanne. Сверху оптимизация осуществлена для примитива (сферы), снизу для предложенного первичного представления геометрии.



initial 8th iteration 16th iteration 24th iteration 32nd iteration reference

Рис. 10. Исходные модели геометрии, промежуточные этапы оптимизации геометрии и эталонное изображение для сцены lucy. Сверху оптимизация осуществлена для примитива (сферы), снизу для предложенного первичного представления геометрии.



initial 8th iteration 16th iteration 24th iteration 32nd iteration reference

Рис. 11. Исходные модели геометрии, промежуточные этапы оптимизации геометрии и эталонное изображение для сцены dragon. Сверху оптимизация осуществлена для примитива (сферы), снизу для предложенного первичного представления геометрии.

Рис. 12. Исходные модели геометрии, промежуточные этапы оптимизации геометрии и эталонное изображение для сцены armadillo. Сверху оптимизация осуществлена для примитива (сферы), снизу для предложенного первичного представления геометрии.

ТАБЛИЦА 1
СРАВНЕНИЕ РЕЗУЛЬТАТОВ РЕКОНСТРУКЦИИ ОБЪЕКТОВ.

Сцена	Исходная геометрия	Время предобработки	Время оптимизации	Метрика Хаусдорфа
bunny	Сфера	-	3м 56с	0.039
	Визуальная оболочка	5.3с	4м 02с	0.029
teapot	Тороид	-	4м 07с	0.120
	Визуальная оболочка	5.5с	4м 17с	0.013
suzanne	Сфера	-	4м 08с	0.040
	Визуальная оболочка	5.2с	4м 21с	0.022
lucy	Сфера	-	4м 47с	0.073
	Визуальная оболочка	6.0с	4м 59с	0.026
dragon	Сфера	-	4м 16с	0.123
	Визуальная оболочка	6.3с	4м 22с	0.047
armadillo	Сфера	-	4м 41с	0.133
	Визуальная оболочка	5.7с	4м 38с	0.021

Описанные эксперименты проведены на компьютере со следующими характеристиками:

- Процессор AMD Ryzen 7 5700G, 3.80ГГц;
- Оперативная память GoodRAM, 32 ГБ DDR4, 3200 МГц;
- Графический ускоритель GIGABYTE GeForce RTX™ 2080 GAMING OC 8G, 1800МГц.

VI. Выводы и заключение

Анализ результатов показал, что использование визуальной оболочки позволяет не только обеспечить гомеоморфность геометрии исходного и эталонного объектов, но и увеличить точность реконструкции методов, опирающихся на использование дифференцируемого рендеринга, до десяти раз при сохранении аналогичного времени реконструкции. На основании чего можно сделать вывод об эффективности использования визуальной оболочки в качестве исходной геометрии объектов, подлежащих реконструкции с помощью методов дифференцируемого рендеринга. Использованный алгоритм обратного рендеринга оказывается неэффективным в воссоздании мелких деталей формы объектов или уточнении построенного первичного представления, однако показывает высокую эффективность в формировании силуэта изображения. Дальнейшим направлением исследований может являться оптимизация метода построения визуальной оболочки для минимизации возникающих артефактов геометрии и ускорения методов реконструкции, а также разработка метода реконструкции геометрии с использованием дифференцируемого рендеринга, в котором будет уделено особое внимание операциям с градиентами нормалей, а не с градиентами силуэта.

Источник финансирования. Благодарности

Работа выполнена при финансовой поддержке Российского Научного Фонда, грант “Разработка методов, алгоритмов и вычислительной архитектуры реконструкции геометрии реального окружения для построения системы смешанной реальности, свободной от дискомфорта зрительного восприятия” №22-11-00145.

Список литературы

1. Cardenas-Garcia J. F., Yao H. G., Zheng S. 3D reconstruction of objects using stereo imaging //Optics and Lasers in Engineering. – 1995. – Т. 22. – №. 3. – С. 193-213.
2. Mikhail E. M., Bethel J. S., McGlone J. C. Introduction to modern photogrammetry. – John Wiley & Sons, 2001.
3. He C., Shen Y., Forbes A. Towards higher-dimensional structured light //Light: Science & Applications. – 2022. – Т. 11. – №. 1. – С. 205.
4. Collis R. T. H. Lidar //Applied optics. – 1970. – Т. 9. – №. 8. – С. 1782-1788.
5. Zhou Z. et al. Three-Dimensional Geometry Reconstruction Method from Multi-View ISAR Images Utilizing Deep Learning //Remote Sensing. – 2023. – Т. 15. – №. 7. – С. 1882.
6. Liu Z. N. et al. High-quality textured 3D shape reconstruction with cascaded fully convolutional networks //IEEE Transactions on Visualization and Computer Graphics. – 2019. – Т. 27. – №. 1. – С. 83-97.
7. Henderson P., Ferrari V. Learning to generate and reconstruct 3d meshes with only 2d supervision //arXiv preprint arXiv:1807.09259. – 2018.
8. Kato H. et al. Differentiable rendering: A survey //arXiv preprint arXiv:2006.12057. – 2020.
9. Periyasamy A. S., Behnke S. Towards 3D Scene Understanding Using Differentiable Rendering //SN Computer Science. – 2023. – Т. 4. – №. 3. – С. 245.
10. Nicolet B., Jacobson A., Jakob W. Large steps in inverse rendering of geometry //ACM Transactions on Graphics (TOG). – 2021. – Т. 40. – №. 6. – С. 1-13.
11. Palfinger W. Continuous remeshing for inverse rendering //Computer Animation and Virtual Worlds. – 2022. – Т. 33. – №. 5. – С. e2101.
12. O’Riordan A. et al. Stereo vision sensing: Review of existing systems //2018 12th International Conference on Sensing Technology (ICST). – IEEE, 2018. – С. 178-184.
13. A Laurentini A. The visual hull concept for silhouette-based image understanding //IEEE Transactions on pattern analysis and machine intelligence. – 1994. – Т. 16. – №. 2. – С. 150-162.
14. Schneider D. C. Visual hull //Computer Vision: A Reference Guide. – Cham: Springer International Publishing, 2021. – С. 1363-1366.
15. Oxholm G., Nishino K. Multiview shape and reflectance from natural illumination //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – С. 2155-2162.
16. Seitz S. M. et al. A comparison and evaluation of multi-view stereo reconstruction algorithms //2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). – IEEE, 2006. – Т. 1. – С. 519-528.
17. Lin L., Zhu J., Zhang Y. Multiview textured mesh recovery by differentiable rendering //IEEE Transactions on Circuits and Systems for Video Technology. – 2022. – Т. 33. – №. 4. – С. 1684-1696.
18. Marí R. Accelerating NeRF with the Visual Hull //Image Processing On Line. – 2024. – Т. 14. – С. 217-231.
19. Kass M., Witkin A., Terzopoulos D. Snakes: Active contour models //International journal of computer vision. – 1988. – Т. 1. – №. 4. – С. 321-331.
20. Boykov Y. Y., Jolly M. P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images //Proceedings eighth IEEE international conference on computer vision. ICCV 2001. – IEEE, 2001. – Т. 1. – С. 105-112.
21. Khan W. Image segmentation techniques: A survey //Journal of image and graphics. – 2013. – Т. 1. – №. 4. – С. 166-170.
22. Wang W. et al. A survey of zero-shot learning: Settings, methods, and applications //ACM Transactions on Intelligent Systems and Technology (TIST). – 2019. – Т. 10. – №. 2. – С. 1-37.
23. Siudak M., Rokita P. A survey of passive 3D reconstruction methods on the basis of more than one image //Machine Graphics & Vision. – 2014. – Т. 23. – №. 3/4.
24. Matusik W. et al. An efficient visual hull computation algorithm. – 2002.
25. Zhang H. et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection //arXiv preprint arXiv:2203.03605. – 2022.
26. Li L. H. et al. Grounded language-image pre-training //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2022. – С. 10965-10975.

27. Suzuki S. et al. Topological structural analysis of digitized binary images by border following //Computer vision, graphics, and image processing. – 1985. – Т. 30. – №. 1. – С. 32-46.
28. Bradski G. et al. OpenCV //Dr. Dobb's journal of software tools. – 2000. – Т. 3. – №. 2.
29. Jakob W. et al. Dr. jit: A just-in-time compiler for differentiable rendering //ACM Transactions on Graphics (TOG). – 2022. – Т. 41. – №. 4. – С. 1-19.
30. Laine S. et al. Modular primitives for high-performance differentiable rendering //ACM Transactions on Graphics (ToG). – 2020. – Т. 39. – №. 6. – С. 1-14.
31. Lin T. Y. et al. Microsoft coco: Common objects in context //Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. – Springer International Publishing, 2014. – С. 740-755.

Секция 2

НАУЧНАЯ ВИЗУАЛИЗАЦИЯ И ВИЗУАЛЬНАЯ АНАЛИТИКА

Scientific Visualization and Visual Analytics

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 004.514
DOI: 10.25206/978-5-8149-3873-2-2024-251-257

Факторы контекстуальной точности в UX-тестах и их комбинаторное влияние на эффективность взаимодействия с интерфейсом
Factors of contextual accuracy in UX tests and their combinatorial influence on the effectiveness of interaction with the interface

А. В. Якунин, С. С. Бодрунова
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Alexander V. Yakunin, Svetlana S. Bodrunova
School of Journalism and Mass Communications, St. Petersburg State University, 7-9 Universitetskaya nab., St. Petersburg 199004, Russia

Аннотация. В предлагаемом исследовании представлены результаты эксперимента, посвященного влиянию графического интерфейса на опыт пользователя при решении задач разной когнитивной интенсивности. В качестве эмпирической базы выступают четыре сайта с разным качеством дизайна на двух разных уровнях композиционно-графической модели (микро- и макро-), на которых тестируются задачи двух типов – в режиме цейтнота и без лимита времени. В качестве зависимых переменных использованы метрики психофизиологических состояний пользователя – эффективность когнитивных процессов («производительность») и уровень психоэмоционального напряжения (в показателях уровня стресса и утомления). Как показали результаты эксперимента, различия на макроуровне дизайна сильнее влияют на эффективность когнитивных процессов, при этом эмоциональные состояния пользователей проявляют относительную стабильность. В случае различий на микроуровне влияние на психоэмоциональное состояние пользователей оказывается более значительным, чем на когнитивные процессы.

***Ключевые слова*:** контекстуальная модель точности, юзабилити-тестирование, графический интерфейс, опыт пользователя

Аннотация. The proposed study presents the results of an experiment examining the influence of a graphical interface on the user's experience when solving problems of varying cognitive intensity. The empirical base is four sites with different design quality at two different levels of the compositional-graphic model (micro- and macro-), on which two types of tasks are tested - under time pressure and without a time limit. Metrics of the user's psychophysiological states were used as dependent variables – the effectiveness of cognitive processes ("performance") and the level of psycho-emotional stress (in terms of the level of stress and fatigue). As the results of the experiment showed, differences at the macro level of design have a stronger impact on the efficiency of cognitive processes, while the emotional states of users show relative stability. In the case of differences at the micro level, the impact on the psycho-emotional state of users is more significant than on cognitive processes.

***Ключевые слова*:** contextual fidelity model, usability testing, GUI, UX.

I. Введение

В последние годы в работах о человек-компьютерном взаимодействии активно обсуждаются комбинаторные эффекты юзабилити (UX-эффекты), являющиеся результатом взаимодействия групп факторов, влияющих на взаимодействие человека и интерфейса. Обобщенная модель анализа таких эффектов, включающая четыре группы факторов (свойства интерфейса, характер задачи, черты пользователя и условия эксперимента), была предложена в 2011 году как «модель контекстуальной точности» (contextual fidelity model, CFM). В рамках данной модели юзабилити изучается совокупное, комбинаторное и/или компенсаторное влияние указанных факторов на когнитивные и эмоциональные процессы в психике пользователя в ходе взаимодействя с интерфейсом [1-4]. Понимание психофизиологических эффектов, формирующихся в процессе такого взаимодействия, значительно способно существенно продвинуть вперед теорию и методологию эффективного проектирования интерфейсов информационных интерактивных систем и технологий визуализации данных. Но «модель контекстуальной точности» пока не применялась для оценки порталов крупных медиапроектов. Данная статья предлагает частично восполнить этот пробел, поскольку изучает комбинаторные эффекты взаимодействия двух факторов CFM для новостных агентств России и Китая.

II. Постановка задачи

В предлагаемом исследовании рассматривается взаимодействие и влияние на пользовательский опыт двух факторов CFM – дизайна продукта (веб-сайта) и свойств задачи. Применительно к организации исследования это означает измерение воздействия графического интерфейса сайтов с разным качеством дизайна на двух разных уровнях композиционно-графической модели (микро- и макро-) в условиях выполнения задач двух типов – на скорость («быстрой») и без лимита времени («медленной»). В качестве зависимых переменных предполагается использовать метрики психофизиологических состояний пользователя – эффективность когнитивных процессов («производительность») и уровень психоэмоционального напряжения (в показателях уровня стресса и утомления).

III. Теория

К настоящему моменту методы исследования эффективности интерфейсов опираются на четыре самостоятельных подхода, каждый из которых в качестве ведущего рассматривает один из факторов contextual fidelity model.

Подход с точки зрения контекста рассматривает эффекты различного качества, порожденные непосредственным окружением и условиями деятельности. В этом подходе исследуется зависимость качества взаимодействия от присутствия других людей [5] или формата взаимодействия [6].

Подход с точки зрения свойств задачи учитывает зависимость переживаний пользователя от структурной [7] и когнитивной [8] сложности исполняемых задач. Влияние демографических и психографических характеристик пользователя на результаты взаимодействия является объектом интереса третьего подхода – с точки зрения свойств пользователя [9-11].

Но, безусловно, наиболее интенсивно разрабатываемой областью является зависимость опыта пользователя от дизайна продукта. Объектом исследовательского внимания в этом направлении HCI становились, в частности, связи между эстетической оценкой и воспринимаемым юзабилити [12], между субъективной удовлетворенностью пользователя, его производительностью и функциональностью дизайна [13].

При этом очевидным недостатком большинства исследований является слишком общий, недифференцированный подход к оценке качества дизайна, основанный на субъективных показателях (таких как эстетическая оценка, например). Между тем, как показали некоторые исследования [14], различные уровни композиционно-графической модели по-разному влияют на процессы визуального восприятия и интеллектуальную активность реципиента.

В этой связи особое значение, как представляется, имеет выделение в композиционно-графической модели двух структурных уровней – микро- и макро- [15], каждый из которых связан с различными когнитивными режимами. Согласно данной концепции, формирование пользовательского опыта в процессе взаимодействия с графическим интерфейсом представляет собой многомерный процесс, включающий в себя множество дискретных элементов и принципов дизайна – таких, например, как макет страницы, соответствие графики и текста, стиль шрифта, межстрочный интервал и высота символов. Эти элементы способны напрямую влиять на психологическое состояние пользователей, которое, в свою очередь, определяет их общую оценку и удовлетворенность веб-сайтом. Располагаясь на различных уровнях композиционной архитектуры, тем не менее, данные элементы составляют локальные системные единства, связанные общим воздействием на определенные когнитивные процессы:

- Макроуровень композиции (F-паттерн, цветовое зонирование, модульная система макета, креолизация контента) объединяет критерии, организующие общую архитектонику страницы и управляющие эффективностью поискового когнитивного режима;
- Микроуровень композиции (адаптивность, размер и контрастность шрифтов, тип гарнитуры, интерлиньяж и длина строки) объединяет критерии, обеспечивающие детальное изучение контента и управляющие эффективностью фокального (сукцессивного) когнитивного режима.

Анализ воздействия каждого из уровней на опыт пользователя в различных сочетаниях CFM представляет несомненную ценность для понимания пластической, нелинейной динамики психофизиологических процессов, составляющих саму суть интерактивного взаимодействия человека с информационным объектом. Без детального изучения такой динамики дальнейшее совершенствование методов дизайн-проектирования и оценки качества интерфейсов не представляется возможным.

IV. Результаты экспериментов

1. Метод

В процессе эксперимента исследовалось воздействие интерфейса четырех сайтов на аудиторию русскоязычных студентов, при этом в оценке изменений пользовательского опыта мы ориентировались на два типа процессов – когнитивные («производительность») и эмоциональные («стресс» и «усталость»). Для формирования этих состояний мы разработали два типа задач – задачу ускоренного поиска контента («быстрая») и задачу без ограничения времени («медленная»). Для измерения качества пользовательского опыта мы выбрали следующие показатели:

1.1. Оценка производительности

Данный параметр измерялся с помощью двух психодиагностических методов оценки:

- Тест на подвижность нервных процессов «Расстановка чисел» [16];
- Тест на емкость оперативной памяти.

1.2. Оценка психоэмоционального состояния

Данный параметр измерялся с помощью двух методов:

- Теста Спилбергера-Ханина на ситуативную тревожность (URL: https://psytests.org/result?v=sphA3oc);
- Теста Fatigue Assessment Scale (FAS) на степень усталости (URL: https://www.waso 75).

Все психодиагностические методы применялись дважды - до и после эксперимента. Разница между показателями представляет наглядное подтверждение влияния определенных факторов дизайна интерфейса на состояние пользователей.

Эмпирической базой для разработки заданий выступили аналогичные друг другу медиапорталы крупных СМИ России (RT, РИА «Новости») и Китая (CGTN, агентство «Синьхуа») с интерактивными элементами и с различным качеством микро- и макроуровней композиции веб-страницы, определенным по методике расчета индекса юзабилити для каждого уровня (U-index) [14]. В соответствии с данной методикой были сопоставлены сайты с различиями в проектировании одного из уровней композиции и тождественными дизайнерскими решениями на другом (см. ТАБЛИЦА 1):

- <**Макроуровень различается; микроуровень тождественен**>: CGTN vs. RT;
- <**Макроуровень тождественен; микроуровень различается**>: Синьхуа vs. РИА «Новости».

ТАБЛИЦА 1
ПОКАЗАТЕЛИ U-INDEX ДЛЯ ЧЕТЫРЕХ ПОРТАЛОВ

Медиапроект	**Макроуровень**	**Микроуровень**
CGTN	U=6	U=5
RT	U=4	U=5
РИА «Новости»	U=5	U=6
Синьхуа	U=5	U=4

2. Дизайн эмпирического исследования

Тесты выполнялись в 8 группах (по 5 человек), 4 из которых работали с заданием на быстрый поиск, и две – без лимита времени. В соответствии с поставленными вопросами мы разработали структуру эксперимента из следующих этапов.

1. Стартовое измерение параметров производительности и психоэмоционального состояния перед основным заданием;

2. Выполнение заданий с определенной скоростью на сайтах CGTN и RT; РИА «Новости» и Синьхуа.

3. Финальное тестирование параметров производительности и психоэмоционального состояния для определения сдвигов в показателях пользовательского опыта.

V. Обсуждение результатов

Результаты эксперимента представлены далее в виде таблиц 2 и 3.

ТАБЛИЦА 2

Медиапроект		**CGTN**			**RT**		
Тип задачи		**«быстрая» задача**					
Время выполнения теста		до задачи	после задачи	Δ	до задачи	после задачи	Δ
Производительность	подвижность нервных процессов (баллы 1-10)	9	10	**1 слабый рост**	9	6	**-3 снижение**
	оперативная память (% верных ответов, 0-100)	60	65	**5 умеренный рост**	52	50	**-2 снижение**
Стресс (баллы, 10-50)		4	4	0	5	8	**3 умеренный рост**
Усталость (баллы, 0-10)		20	20	0	18	22	**4 умеренный рост**
Тип задачи		**«медленная» задача**					
Время выполнения теста		до задачи	после задачи	Δ	до задачи	после задачи	Δ
Производительность	подвижность нервных процессов (баллы 1-10)	6	10	**4 умеренный рост**	8	6	**-2 снижение**
	оперативная память (% верных ответов, 0-100)	72	83	**11 значительный рост**	49	47	**-2 снижение**
Стресс (баллы, 10-50)		5	5	0	4	5	**1 слабый рост**
Усталость (баллы, 0-10)		19	19	0	21	24	**3 слабый рост**

ТАБЛИЦА 3

Медиапроект		**Синьхуа**			**РИА «Новости»**		
Тип задачи		**«быстрая» задача**					
Время выполнения теста		до задачи	после задачи	Δ	до задачи	после задачи	Δ
Производительность	подвижность нервных процессов (баллы 1-10)	8	5	**-3 снижение**	8	8	0
	оперативная память (% верных ответов, 0-100)	54	52	**-2 снижение**	60	60	0
Стресс (баллы, 10-50)		4	7	**3 умеренный рост**	3	5	**2 слабый рост**

Усталость (баллы, 0-10)		19	19	0	20	20	0
Тип задачи		**«медленная» задача**					
Время выполнения теста		до задачи	после задачи	Δ	до задачи	после задачи	Δ
Производительность	подвижность нервных процессов (баллы 1-10)	8	7	**-1 снижение**	4	4	0
	оперативная память (% верных ответов, 0-100)	58	58	0	65	65	0
Стресс (баллы, 10-50)		3	3	0	4	4	0
Усталость (баллы, 0-10)		18	19	**1 слабый рост**	20	20	0

1. Для ситуации <Макроуровень различается; микроуровень тождествен> (CGTN vs. RT; ТАБЛИЦА 2) характерны следующие наблюдаемые эффекты:

Во-первых, в целом дизайн макроуровня значительно влияет на рост производительности и, следовательно, интенсивность когнитивной активности – в случае с большим значением индекса юзабилити для макроуровня (CGTN; U=6) и для быстрой, и для медленной задач характерен рост подвижности нервных процессов.

Во-вторых, разница в дизайне макроуровня веб-страницы оказывается значима для формирования негативных эмоциональных состояний пользователя. И в быстрой, и в медленной задаче наблюдаются одинаковые эффекты: в случае с большим значением индекса юзабилити для макроуровня (CGTN) уровень стресса и утомления остается неизменным, что характеризует способность пользователей при данном индексе юзабилити сохранять стабильное психическое состояние и эффективно справляться со стрессом при выполнении заданий с разной скоростью. В случае с меньшим индексом RT (U=4) наблюдается рост обоих негативных эмоциональных состояний. Этот результат говорит о том, что, хотя медленные задания дают пользователям больше времени на обдумывание и обработку информации, длительная когнитивная нагрузка все равно приводит к накоплению усталости. Возможно, это связано со стереотипностью когнитивных операций. Примечательно, что хотя при выполнении быстрых и медленных заданий уровень усталости повышался по-разному, в целом скорость выполнения задания не оказывала существенного влияния на уровень усталости. Этот вывод опровергает традиционное мнение о том, что чем быстрые, форсированные задания вызывают большую усталость. Более вероятно, что за влиянием задачи на уровень усталости стоят другие факторы, такие как сложность задачи, состояние самого пользователя и продолжительность выполнения задачи.

В-третьих, наблюдается отчетливая связь между дизайном макроуровня и скоростью задачи: в случае с более высоким индексом юзабилити (CGTN) при увеличении скорости выполнения задания тенденция роста подвижности нервных процессов замедляется. Возможно, это свидетельство в пользу прямой зависимости между периодом адаптации к когнитивной нагрузке и степенью подвижности нервных реакций – медленная задача предоставляет больше возможностей для ускорения когнитивных способностей. Аналогичная тенденция наблюдается и с объемом оперативной памяти: с ростом скорости работы на сайте с более качественным дизайном макроуровня этот объем сокращается. Медленная задача лучше активизирует ресурсы оперативной памяти, при этом быстрые задачи перераспределяют когнитивные ресурсы в пользу скорости обработки информации и в ущерб емкости оперативной памяти.

2. Для ситуации <Макроуровень тождествен; микроуровень различается> (Синьхуа vs. РИА «Новости», ТАБЛИЦА 3) различия на микроуровне дизайна также оказываются значимыми для формирования дисфункций и имеют следующую специфику.

Во-первых, в случае с производительностью дизайн с меньшим индексом юзабилити (Синьхуа, U=4) оказывается чувствителен к скорости выполнения задачи - чем выше скорость, тем сильнее снижается производительность.

Во-вторых, в случае с эмоциональным состоянием наблюдается сквозной эффект для обоих сайтов – линейная корреляция между скоростью выполнения задачи и индексом стресса. С увеличением скорости пропорционально

растет и уровень стресса – возможно, срочность заставляет пользователя острее ощущать сложность задачи. В свою очередь, в случае медленной задачи стресс отсутствует.

VI. Выводы и заключение

Таким образом, в экспериментальных данных проявляются два различных паттерна совместного влияния композиционной архитектуры сайта и содержания исполняемых задач.

Во-первых, при наличии макроуровневых различий в композиционно-графической модели сайта и отсутствии различий на микроуровне скорость навигации пользователей по сайту существенно влияет на их когнитивную эффективность. При этом эмоциональные состояния пользователей проявляют относительную стабильность, что позволяет предположить для данного типа различий более тесную связь колебаний когнитивной эффективности с содержанием исполняемой задачи и особенностями дизайна.

Во-вторых, в случае совпадения юзабилити сайтов на макроуровне и различиях на микроуровне, оказалось, что скорость навигации по сайту начинает оказывать влияние на негативное психоэмоциональное состояние пользователей. Согласно результатам исследования, индекс стресса пользователей увеличился как на сайте Синьхуа, так и на сайте РИА «Новости». Более того, для Синьхуа наблюдалось значительное снижение подвижности нервных процессов пользователей вне зависимости от скорости выполнения задания. Этот результат позволяет предположить, что различия на микроуровне дизайна могут вызывать психологический стресс у пользователей, что, в свою очередь, определяет скорость их психофизиологических реакций.

Источник финансирования.

Исследование выполнено в рамках проекта «Центр международных медиаисследований» Санкт-Петербургского государственного университета, #94033584.

Список литературы

1. Gilal N. G., Zhang J., Gilal F. G. The four-factor model of product design: scale development and validation. // J. Prod. Brand Manag. 2018, 27(6), 684–700.
2. Sauer J., Sonderegger A. The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. // Applied Ergonomics, 2009, 40, pp. 670-677.
3. Sonderegger A., Sauer J. The influence of design aesthetics in usability testing: effects on user performance and perceived usability. Applied Ergonomics, 2010, 41, pp. 403–410.
4. Yakunin A. V., Bodrunova S. S. Cumulative Distortions in Usability Testing: Combined Impact of Web Design, Experiment Conditions, and Type of Task and Upon User States During Internet Use. // Design, User Experience, and Usability: UX Research, Design, and Assessment - 11th International Conference, DUXU 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Proceedings. Springer Nature - pp. 526-535.
5. Als B. S., Jensen J. J., Skov M. B. Exploring Verbalization and Collaboration of Constructive Interaction with Children. // Proceedings of the IFIP 9th International Conference on Human-Computer Interaction - INTERACT 2005, pp. 443-456
6. Razak F. H. A., Hafit H., Sedi N., Zubaidi N.A., Haron H. Usability testing with children: Laboratory vs field studies. // User Science and Engineering (i-USEr), 2010 International Conference, vol., no., pp.104-109.
7. Sauer J., Sonderegger A. Visual aesthetics and user experience: A multiple-session experiment. // International Journal of Human-Computer Studies Volume 165, September 2022.
8. Bara I., Binney R. J., Ward R., Ramsey R. A generalised semantic cognition account of aesthetic experience. // Neuropsychologia Volume 173, 13 August 2022 https://www.sciencedirect.com/science/article/pii/S0028393222001476
9. Aledaily Arwa, Gannouni Sofien, Belwafi Kais, Aboalsamh Hatim A Framework for Usability Testing using EEG Signals with Emotion Recognition. // Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems.
10. Gannouni S., Aledaily A., Belwafi K., Aboalsamh H. Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. // Scientific Reports, 2021, 11(1), 1–17.
11. Gannouni S., Aledaily A., Belwafi K., Aboalsamh H. Adaptive emotion detection using the valence-arousal-dominance model and eeg brain rhythmic activity changes in relevant brain lobes. // IEEE Access 8, 2020, 67444–67455.

12. De Angeli A., Sutcliffe A.G., Hartmann J. Interaction, usability and aesthetics: What influences users' preferences? // Proceedings DIS-06, Designing Interactive Systems, New York: 2006, ACM Press.
13. Tuch A.N., Roth S.P., Hornbaek K., Opwis K., Bargas-Avila J.A. Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. // Comput. Hum. Behav. 2012, 28(5), 1596–1607.
14. Yakunin A.V., Bodrunova S.S. U-index: An eye-tracking-tested checklist on webpage aesthetics for university web spaces in Russia and the USA. // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, 2017, Vol. 10288, pp. 219-233.
15. Якунин А.В. Композиционно-графическая модель веб-сайта как фактор его эстетической привлекательности // Проблемы, перспективы и направления инновационного развития науки: Сборник статей международной научно-практической конференции (1 октября 2016, Уфа) / в 2 ч. Ч. 2. Уфа: АЭТЕРНА. С. 226-231.
16. Касьянов С. Психологические тесты - М.: Эксмо, 2006. - С. 552–553.

УДК 535.8
DOI: 10.25206/978-5-8149-3873-2-2024-258-263

Применение оптимизации Гаусса-Ньютона при гильберт-диагностике пламени Бунзена
Gauss-Newton optimization application in Hilbert diagnostics of Bunsen flame

Э. В. Арбузов1,3, Ю. Н. Дубнищев2,3, О. С. Золотухина2,3
1Институт математики им. С.Л. Соболева СО РАН, г. Новосибирск, Россия
2Институт теплофизики им. С.С. Кутателадзе СО РАН, г. Новосибирск, Россия
3Новосибирский государственный технический университет, г. Новосибирск, Россия

E. V. Arbuzov1,3, Yu. N. Dubnishchev2,3, O. S. Zolotukhina2,3
[1]Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia
[2]Kutateladze Institute of Thermophysics SB RAS, Novosibirsk, Russia
[3]Novosibirsk State Technical University, Novosibirsk, Russia

Аннотация. В работе на примере данных о температуре и основных соединениях ламинарного метано-воздушного пламени Бунзена рассматривается возможность применения алгоритма оптимизации Гаусса-Ньютона в гильберт-диагностике реагирующих сред (пламён). Предложенная ранее и проверенная на численных моделях методика реконструкции полей фазовой оптической плотности использована для решения обратной задачи гильберт-оптики: в рассматриваемом осесимметричном сечении пламени вычислено распределение гильбертограммы, из которого далее восстановлено исходное значение фазовой функции и показателя преломления среды. Развитием работы станет расширение сферы приложений полученных результатов к обработке экспериментальных данных.

***Ключевые слова*:** гильберт-оптика, алгоритм Гаусса-Ньютона, оптимизация, пламя Бунзена, диагностика пламени

Аннотация. The possibility of using the Gauss-Newton optimization algorithm in the Hilbert diagnostics of reacting media (flames) is considered in the work using the example of temperature data and the main connections of a laminar methane-air Bunsen flam. The previously proposed and tested on numerical models technique for reconstructing phase optical density fields was used to solve the inverse problem of Hilbert optics. The hilbertogram distribution was calculated in the considered axisymmetric flame section, from which the initial value of the phase function and the refractive index of the medium was then restored. The work development will be the scope expansion of obtained results applications to the experimental data processing.

***Ключевые слова*:** Hilbert optics, Gauss-Newton algorithm, optimization, Bunsen flame, flame diagnostics

I. Введение

Исследования процессов горения имеют важное практическое значение во многих областях промышленности и технологиях [1, 2]. Наиболее значимой является оптическая диагностика, позволяющая получать информацию о термодинамических и структурных параметрах пламён, основными достоинствами которой являются бесконтактность применения, возможность проведения измерений в высокотемпературных зонах. Методы оптической диагностики условно можно разделить на две группы. Первую составляют способы, основанные на явлениях рассеяния, поглощения и флуоресценции при прохождении светового излучения через исследуемую среду [3–5]. Ко второй относятся методы, заключающиеся в фиксации изменений коэффициента преломления (фазовой плотности) газовой смеси [6, 7].

II. Постановка задачи

Оптическая гильберт-диагностика позволяет с высокой чувствительностью визуализировать фазовые структуры исследуемых сред в пространстве, заданном зондирующим световым полем [8]. Ранее в [9] были представлены результаты полихроматической гильберт-диагностики водородно-воздушного диффузионного пламени с реконструкцией из полученной фазовой функции температурного поля и определением концентраций продуктов горения. Однако остаётся нерешённым вопрос автоматического эффективного расчёта фазовых полей по данным гильберт-измерений.

В [10] для оптимизации реконструкции фазовой функции при гильберт-диагностике на примере численных моделей было предложено применение итерационного алгоритма Гаусса-Ньютона [11, 12], который

используется для нахождения минимума целевой функции. В настоящей работе выполнена адаптация оптимизации Гаусса-Ньютона для решения обратной задачи гильберт-оптики с использованием экспериментальных данных о температуре и основных соединениях ламинарного метано-воздушного пламени Бунзена, описанных в [13].

III. Теория

Математическая модель гильберт-визуализации фазовых возмущений в некотором сечении *y* диагностируемого объекта может быть задана соотношением (ось *z* совпадает с направлением светового пучка) [10]:

$$H\left[\psi\right]\left(x^*, y\right) = \left\{\int_{-\infty}^{+\infty} \frac{\cos\left[\psi\left(x, y\right)\right]}{x^* - x} dx\right\}^2 + \left\{\int_{-\infty}^{+\infty} \frac{\sin\left[\psi\left(x, y\right)\right]}{x^* - x} dx\right\}^2. \quad (1)$$

Функция $\psi\left(x, y\right)$ в свою очередь зависит от геометрической длины пути и показателя преломления $n(x,y,z)$:

$$,\psi\left(x, y\right) = \frac{2\pi}{\lambda} \int_{z_1}^{z_2} \left[n(x,y,z) - n_\infty\right] dz \quad (2)$$

где λ – длина волны, n_∞ – показатель преломления среды, окружающей пламя, z_1 и z_2 – начальная и конечная точки распространения светового луча в объекте. В случае осевой симметрии выражение (2) можно представить через интеграл Абеля [9]:

$$\psi\left(x, y\right) = \frac{4\pi}{\lambda} \int_x^R \left[n(r,y) - n_\infty\right] \frac{r}{\sqrt{r^2 - x^2}} dr \; ; \; r^2 = x^2 + z^2 \; ; \quad (3)$$

R – радиус осесимметричного сечения пламени; $n(r,y)$ – зависимость показателя преломления от радиуса в пределах сечения.

Связь показателя преломления и параметров горящей смеси определяется соотношением:

$$n - 1 = \frac{p}{p_{n.c.}} \frac{T_{n.c.}}{T} \sum_k A_k \left(1 + \frac{B_k}{\lambda^2}\right) C_k \; , \quad (4)$$

где *p* – давление в пространственной структуре пламени; $p_{n.c.}$ – давление при нормальных условиях; *T* – температура пламени; $T_{n.c.}$ – температура при нормальных условиях; C_k – мольная доля *k*-го компонента горящей смеси; A_k и B_k – дисперсионные коэффициенты *k*-го компонента горящей смеси (справочные данные). Для нахождения параметров исследуемого пламени *T* и C_k необходимо реконструировать фазовую функцию $\psi\left(x, y\right)$ из гильбертограммы $H\left[\psi\right]$.

Для автоматического выполнения данной операции авторами на примере численных моделей был предложен метод [10], основанный на итерационном алгоритме Гаусса-Ньютона, который заключается в подборе профиля фазы, заданного суммой кривых Безье третей степени [14], последующем вычислении гильбертограммы и сопоставлении её с экспериментальными данными. Критерием достоверности реконструкции $\psi\left(x, y\right)$ служит совпадение структур экспериментальной и модельной гильбертограмм.

Если задать некоторое начальное приближение фазовой функции через совокупность участков, каждый из которых моделируется кривой Безье:

$$\psi_q^{\text{mod}}\left(x,y\right): \begin{cases} x\left(t, V_x^q\right) = (1-t)^3 V_{x,0}^q + 3(1-t)^2 t V_{x,1}^q + 3(1-t) t^2 V_{x,2}^q + t^3 V_{x,3}^q \\ y\left(t, V_y^q\right) = (1-t)^3 V_{y,0}^q + 3(1-t)^2 t V_{y,1}^q + 3(1-t) t^2 V_{y,2}^q + t^3 V_{y,3}^q \end{cases} ; \; t \in \left[0,1\right];$$

где V_x^q и V_y^q – компоненты векторов опорных вершин кривой Безье с номером ($q = 1,\ldots,Q$), и рассчитать от него $H^{\text{mod}}\left[\psi_q^{\text{mod}}\right]$, задача оптимизации сводится к определению V_x^q и V_y^q, при которых достигается минимум целевой функции:

$$f(\vec{V}) = \left\| F\left(\vec{V}\right) \right\|^2 = \sum_{m=0}^{M} \left[H(x_m) - H^{\text{mod}}(x_m) \right]^2 .$$

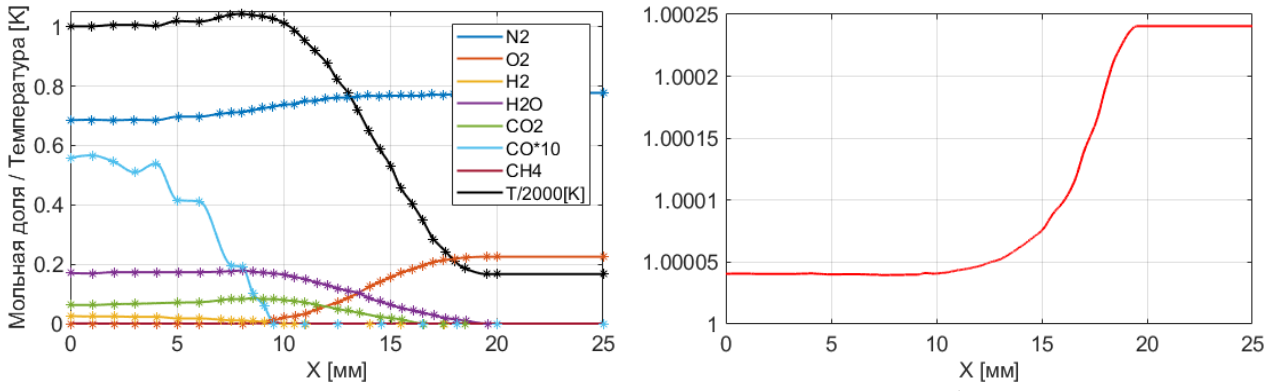
Для нахождения $\vec{V}$ осуществляется выполнение итерационных приближений:

$$\vec{V}_{d+1} = \vec{V}_d - \alpha \left[J^T\left(\vec{V}_d\right) J\left(\vec{V}_d\right) \right]^{-1} J^T\left(\vec{V}_d\right) F\left(\vec{V}_d\right), \quad (5)$$

где α – коэффициент регулировки «шага» итерации, J – матрица Якоби производных первого порядка функции $F\left(\vec{V}\right)$. В результате оптимизация заключается в задании вектора $\vec{V}_0$, определяющего начальное распределение $\psi\left(x, y\right)$, и дальнейшего применения алгоритма Гаусса-Ньютона (5) до тех пор, пока сумма квадратов расстояний между координатами экспериментальной и оптимизируемой гильбертограммами не станет меньше заданного значения.

IV. Результаты экспериментов

Обратимся к рис. 1.*а*, на котором представлено радиальное распределение температуры T и мольных концентраций продуктов горения топлива C_{N_2}, C_{O_2}, C_{H_2}, C_{H_2O}, C_{CO_2}, C_{CO} и C_{CH_4} в некотором сечении ламинарного метано-воздушного пламени Бунзена. Данные взяты из работы [13], в которой они были получены с использованием методов комбинационного рассеяния Рамана-Релея и лазерно-индуцированной флуоресценции.



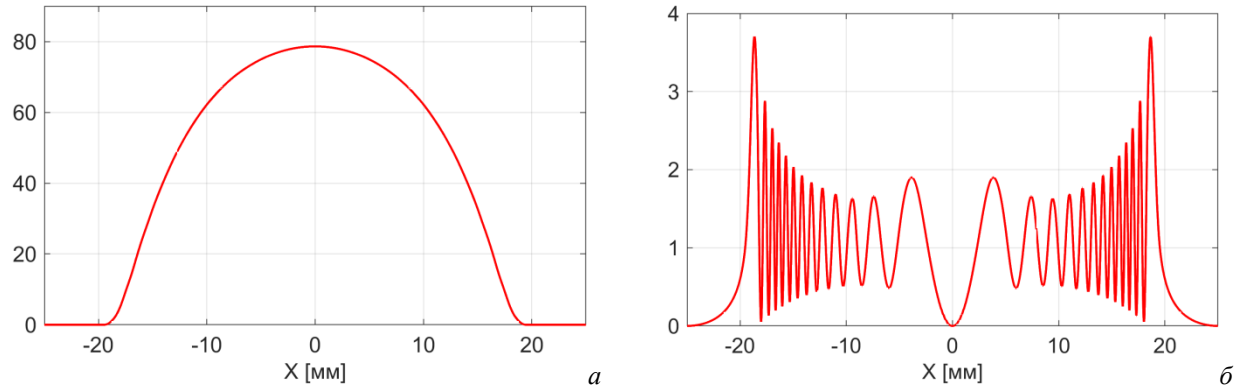
а *б*

Рис. 1. (а) Данные радиального распределения температуры и мольных концентраций продуктов горения топлива в сечении метано-воздушного пламени Бунзена; (б) радиальное распределение показателя преломления

Величина C_{CH_4} определена следующим образом:

$$C_{CH_4} = 1 - C_{N_2} - C_{O_2} - C_{H_2} - C_{H_2O} - C_{CO_2} - C_{CO} .$$

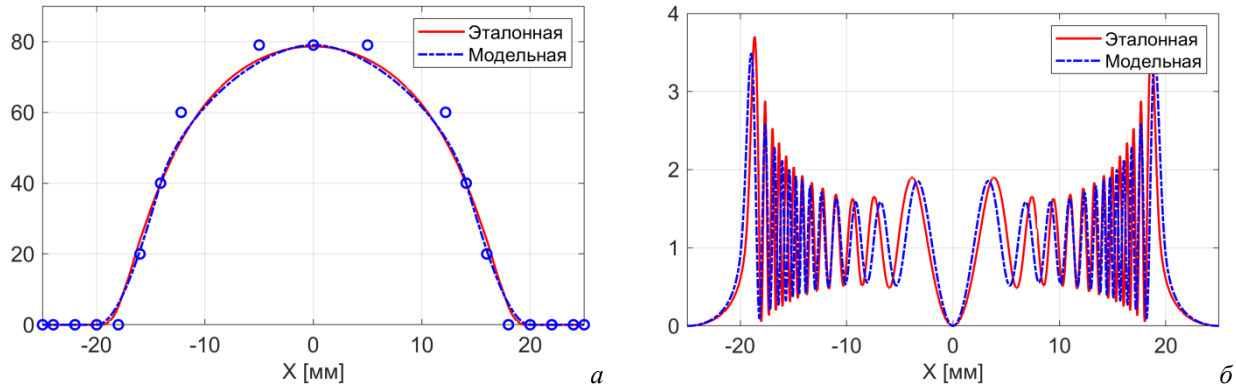
Используя полученные данные, с применением (4) при λ = 532 нм и $p = p_{n.c.}$ выполним расчёт показателя преломления в рассматриваемом сечении пламени (рис. 1.б). Далее по формулам (3) и (1) определим величину фазовой функции ψ (рис. 2.а) и соответствующее распределение гильбертограммы $H\left[\psi\right]$ (рис. 2.б).



а б

Рис. 2. Эталонные данные: (а) фазовая функция; (б) гильбертограмма

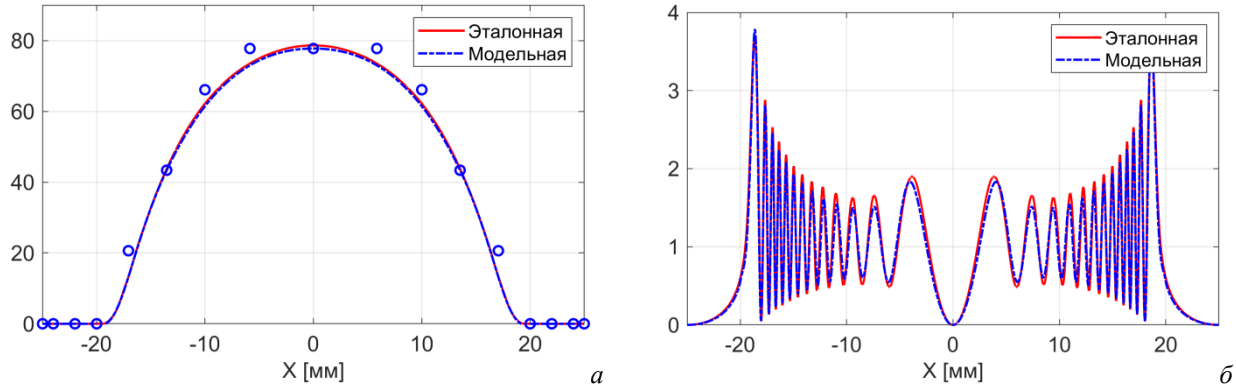
В экспериментальных измерениях обратная задача гильберт-диагностики заключается в реконструкции ψ из зарегистрированной гильбертограммы $H[\psi]$. Гильберт-преобразование позволяет визуализировать экстремумы и градиенты фазовой оптической плотности исследуемой среды, которые трансформируются в структуры гильберт-полос, причём в области экстремумов образуются «широкие» полосы. Применим эту информацию для задания некоторого начального приближения ψ^{mod} и $H^{\text{mod}}[\psi^{\text{mod}}]$, которое необходимо оптимизировать к «эталонным» величинам (рис. 3). Так как пламя является осесимметричным, потребовалось использовать три кривые Безье и отобразить их относительно центральной оси.



а б

Рис. 3. Начальное приближение: (а) фазовая функция, представленная суммой кривых Безье (с указанными опорными вершинами); (б) гильбертограмма

Зафиксируем известные координаты опорных вершин полиномов: значение фазы на краях исследуемой области и местоположение экстремума функции. Результат применения алгоритма оптимизации Гаусса-Ньютона (5) представлен на рис. 4.



а б

Рис. 4. Результат оптимизации: (а) фазовая функция; (б) гильбертограмма

Из реконструированного профиля фазовой функции выполним расчёт показателя преломления (рис. 5) и сравним с исходными данными, изображёнными на рис. 1.б.

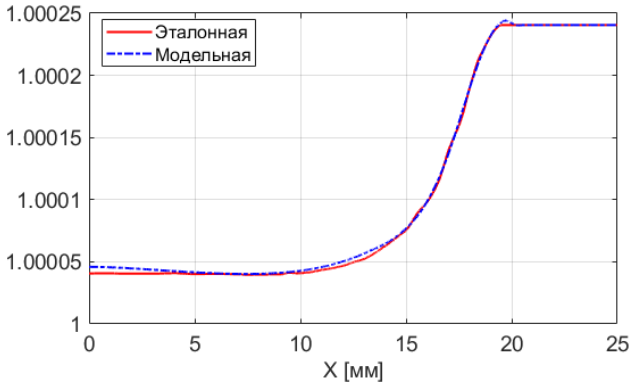


Рис. 5. Сравнение исходного распределения показателя преломления (рис. 1.б) в сечении пламени с реконструированным в результате решения обратной задачи гильберт-оптики

V. Обсуждение результатов

Алгоритм оптимизации Гаусса-Ньютона через 52 итерации позволил приблизить оптимизируемые данные H^{mod} к эталонным значениям H со среднеквадратической погрешностью

$$\sigma = \sqrt{\frac{\sum_{m=0}^{M}\left(H_m - H_m^{\mathrm{opt}}\right)^2}{M}} \approx 0.12;$$

при этом максимальное отклонение реконструированной фазовой функции от эталонной оказалось локализованным в центральной части сечения и составило

$$\delta_{\psi} = \max_{m=0,\ldots,M}\left|\psi_m - \psi_m^{\mathrm{opt}}\right| \approx 0.88;$$

что также наблюдается и в результирующем распределении показателя преломления:

$$\delta_{n} = \max_{m=0,\ldots,M}\left|n_m - n_m^{\mathrm{opt}}\right| \approx 2.5\cdot 10^{-6}.$$

Как видно на рис. 5, на краю области (при $x \approx 20$ мм) имеет место небольшой «выброс» в восстановленных данных, что говорит о недостаточно «гладком» соединении участков, моделируемых кривыми Безье. Отсюда можно сделать вывод, что применение алгоритма Гаусса-Ньютона в решении обратной задачи гильберт-оптики позволяет реконструировать значения фазовой функции с достаточной точностью, однако при этом требуется введение дополнительных ограничений для исключения «смешения» оптимизируемых опорных вершин.

VI. Выводы и заключение

В работе на примере данных о температуре и основных молярных концентрациях продуктов горения метано-воздушного пламени Бунзена апробирован метод автоматического восстановления фазовых структур по данным гильберт-диагностики с применением алгоритма оптимизации Гаусса-Ньютона. В осесимметричном сечении исследуемого пламени выполнен итерационный подбор профиля фазовой функции, адаптированного суммой кривых Безье третей степени, с последующим вычислением пространственной структуры показателя преломления. Установлено, что критерием достоверности реконструкции фазы является совпадение структур эталонной и модельной гильбертограмм. В дальнейшем авторами будет решаться проблема повышения точности оптимизации, а также вопрос расширения сферы приложений полученных результатов к обработке экспериментальных данных.

Источник финансирования. Благодарности

Работа выполнена в рамках государственного задания ИМ СО РАН № FWNF-2022-0009, государственного задания ИТ СО РАН № 121031800217-8.

Список литературы

1. Li J., Huang H., Bai Y., and etc. Combustion and heat release characteristics of hydrogen/air diffusion flame on a micro-jet array burner // Int. J. Hydrog. Energy. 2018. Vol. 43 (29). P. 13563–13574. doi.org/10.1016/j.ijhydene.2018.04.195.

2. Masri A. R. Challenges for turbulent combustion // Proc. Combust. Inst. 2021. Vol. 38 (1). P. 121–155. doi.org/10.1016/j.proci.2020.07.144.

3. Jiang N., Roy S., Hsu P. S., and etc. High-speed, two-dimensional, multi-species Raman imaging for combustion and flow diagnostics // AIAA 2018-2040. 2018. doi.org/10.2514/6.2018-2040.

4. Yang H. N., Yang B., Cai X. S., and etc. Three-dimensional (3-D) temperature measurement in a low pressure flame reactor using multiplexed tunable diode laser absorption spectroscopy (TDLAS) // Laser. Eng. 2015. Vol. 31. P. 285–297.

5. Gazeli K., Lombardi G., Aubert X., and etc. Progresses on the use of two-photon absorption laser induced fluorescence (TALIF) diagnostics for measuring absolute atomic densities in plasmas and flames // Plasma. 2021. Vol. 4 (1). P. 145–171. doi.org/10.3390/plasma4010009.

6. Островский Ю. И., Бутусов М. М., Островская Г. В. Голографическая интерферометрия / М. : Наука, 1977. 336 с.

7. Karaminejad S., Askari M. H., Ashjaee M. Temperature field investigation of hydrogen/air and syngas/air axisymmetric laminar flames using Mach-Zehnder interferometry // Appl. Opt. 2018. Vol. 57 (18). P. 5057–5067. doi.org/10.1364/AO.57.005057.

8. Арбузов В. А., Дубнищев Ю. Н. Методы гильберт-оптики в измерительных технологиях / Новосибирск: Изд-во НГТУ, 2007. 316 с.

9. Dubnishchev Yu. N., Arbuzov V. A., Arbuzov E. V., and etc. Optical diagnostics of hydrogen-air diffusion jet flame // J. Eng. Thermophys. 2022. Vol. 31 (3). P. 402–413. doi.org/10.1134/S1810232822030031.

10. Арбузов Э. В., Арбузов В. А., и др. Метод Гаусса-Ньютона в задаче оптимизации расчёта осесимметричной фазовой функции по данным гильберт-диагностики // Научная визуализация. 2023. Т. 15 (4). С. 56–67. doi.org/10.26583/sv.15.4.05.

11. Дэннис Дж., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений / Под ред. Евтушенко Ю.Г. М.: Мир, 1988. 440 с.

12. Lai W. H., Kek S. L., Tay K. G. Solving nonlinear least squares problem using Gauss-Newton method // Int. J. Innov. Sci. 2017. Vol. 4 (1). P. 258–262.

13. Nguyen Q. V., Dibble R. W., Carter C. D. Raman-LIF measurements of temperature, major species, OH, and NO in a methane-air Bunsen flame // Combust. Flame. 1996. Vol. 105 (4). P. 499–510.

14. Кривые Безье. 2022. URL: https://learn.javascript.ru/bezier.

УДК 535.8
DOI: 10.25206/978-5-8149-3873-2-2024-264-270

Визуализация поля течения, сопряженных параметров и источников тепла при управлении структурой течения
Visualization of the Flowfield, Adjoint Parameters and Heat Sources at Flow Structure Control

А.К. Алексеев[1] , А.Е.Бондарев[2]
[1]РКК Энергия им. С.П. Королева, Ленина 4а, Королев, 141700, Россия
[2]ИПМ им М.В. Келдыша РАН, Миусская пл., 4, 125047, Москва, Россия.

A.K. Alekseev[1], A.E. Bondarev[2]
[1] RSC Energia, Lenina 4a, Korolev, 141700, Russia
[2] Keldysh Institute of Applied mathematics of RAS, Miusskaya sq. 4, Moscow, 125047, Russia

Аннотация. Рассмотрено управление полем течения с помощью стационарных пространственно распределенных тепловых источников с целью снижения максимального давления на поверхности тела. В качестве примера использовано двумерное обтекание цилиндра сверхзвуковым потоком невязкого газа при взаимодействии скачков уплотнения IV типа по Edney. Представлены результаты визуализации распределения параметров течения, сопряженных параметров и тепловых источников.

Ключевые слова: взаимодействие скачков уплотнения, источники тепла, сопряженные уравнения, управление полем течения, вычислительная аэрогазодинамика

Abstract. The flowfield control using steady spatially distributed heat sources is addressed with the purpose to diminish the maximum value of the pressure at the surface of body. The two-dimensional inviscid flow around the cylinder is considered as an example for the Edney IV shocks interaction mode. The results for the visualization of the flowfield paramenters, adjoint parameters and heat sources are provided.

Keywords: shock wave interaction, heat sources, adjoint equations, flow field control, Computational Fluid Dynamics.

I. Введение

В настоящее время достаточно актуальны задачи активного управления течением с помощью различных способов воздействия, включая выделение энергии в поле течения [1-5]. Одной из областей применения, особо интересной с практической точки зрения, является управление течениями, в которых реализуется взаимодействие скачков IV типа по классификации Edney [6,7]. Как известно, при пересечении косых и прямых скачков уплотнения существует шесть различных типов взаимодействия [6]. Они сильно отличаются как по структуре течения, так и по величинам давления и тепловых потоков на поверхности тела. Наиболее интенсивное воздействие на поверхность тела реализуется при взаимодействии IV типа. При этом типе течения образуется узкая струйка, которая тормозится в нескольких последовательных косых скачках уплотнения и достигает поверхности тела малыми потерями полного давления. Это приводит к очень высоким величинам давления и теплового потока на поверхности тела, которые представляют большую проблему с точки зрения защиты конструкции аппарата. В связи с этим представляет интерес возможность активного управления течением с целью разрушения структуры течения типа IV и замены его менее опасной, с точки зрения нагрузок, структурой. В ряде работ экспериментально и численно исследовано управление взаимодействием скачков IV типа. В работах [1,2] экспериментально рассмотрено управление течением при взаимодействии скачков типа IV с помощью импульсного нагрева. В работе [3] в численных экспериментах продемонстрирована возможность снижения максимального давления и тепловых потоков на поверхности тела с помощью выделения тепловой энергии в окрестности

падающего скачка, однако, в работе указывается, что управляющее воздействие не является оптимальным по величине. Одной из причин, препятствующих практическому применению методов управления течением, является достаточно большая энергоемкость, поэтому вопрос поиска минимального по величине воздействия представляет определенный интерес. В этой связи, в данной работе рассмотрен численный поиск пространственно распределенных тепловых источников минимальной суммарной мощности, снижающих максимальную величину давления на поверхности тела. Расчеты проведены в рамках двумерной модели невязкого газа. В качестве основного инструмента использовано итерационное решение задачи минимизации функционала, аппроксимирующего максимум давления, градиентными методами. Поле градиента функционала рассчитывалось с помощью сопряженных уравнений.

II. Управление течением с помощью распределенных источников

Рассмотрим обтекание сверхзвуковым потоком плоского тела с цилиндрическим затуплением при падении на него наклонного скачка уплотнения. Используем нестационарные уравнения Эйлера:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho U_k)}{\partial x^k} = 0; \tag{1}$$

$$\frac{\partial (\rho U_i)}{\partial t} + \frac{\partial (\rho U_k U_i + P\delta_{ik})}{\partial x^k} = 0; \tag{2}$$

$$\frac{\partial (\rho E)}{\partial t} + \frac{\partial (\rho U_k h_0)}{\partial x^k} + q(x,y) = 0; \tag{3}$$

Здесь $U_1 = U, U_2 = V$, $\gamma = \frac{C_p}{C_v}$, $h = \frac{\gamma}{\gamma - 1}\frac{P}{\rho} = \gamma e$, $h_0 = (U^2 + V^2)/2 + h$, $e = \frac{RT}{\gamma - 1}$, $E = \left(e + \frac{1}{2}(U^2 + V^2) \right)$, $P = (\gamma - 1)\rho e$.

Область расчета $\Omega = (0 < x < X_{\max}, 0 < y < Y_{\max})$. Временной интервал $(0 < t < t_f)$ подбирался из условий выхода на стационарный режим. На входной границе принимаем условия, соответствующие наклонному скачку уплотнения в однородном потоке. На выходной границе и боковых границах области задаем условия вытекания. На границе тела Γ заданы условия непротекания. Граничные условия на теле реализованы методом погруженной границы [10,11].

Уравнения Эйлера решались с использованием реконструкции второго порядка точности [8] и метода [9] для решения задачи Римана.

На рис. 1 представлены изолинии плотности для конфигурации IV типа, соответствующие условиям работы [3] и полученные в расчете на сетке 400×475. Параметры течения: $M = 6$, угол отклонения потока за скачком $\Theta_1 = 15^o$, $\gamma = 1.4$.

При перемещении косого скачка вверх по координате Y наблюдался последовательный переход от схемы III типа через схему IV типа к схеме V типа. При этом течение по схеме IV имеет разновидности, существенно различающиеся по давлению на теле [13]. При малых Y (на границе со схемой III) давление максимально, затем, при увеличении Y наблюдается второй максимум, наконец, схема IV переходит в схему IVa, в которой высоконапорная струйка загибается вверх и не попадает на тело, что приводит к снижению максимального давления.

Рассмотрено стационарное течение с взаимодействием IV типа по Эдни, представленное на Рис. 1 (изолинии плотности) и Рис. 2 (распределение давления).

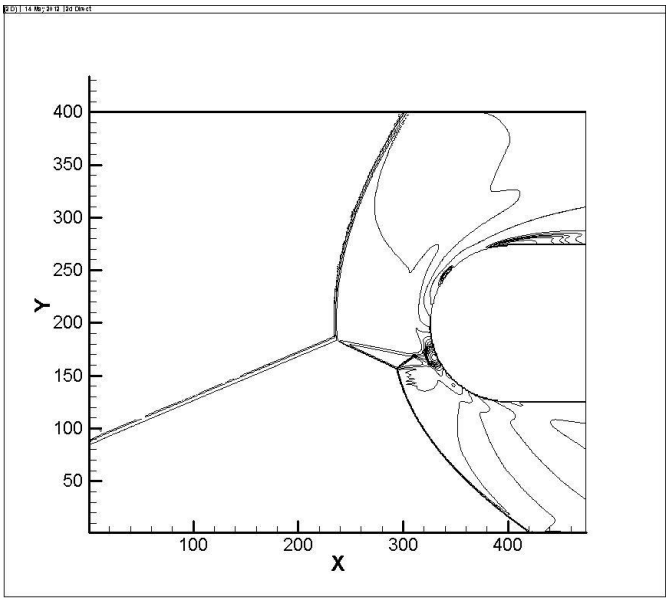


Рис. 1 Изолинии плотности при IV типе взаимодействия

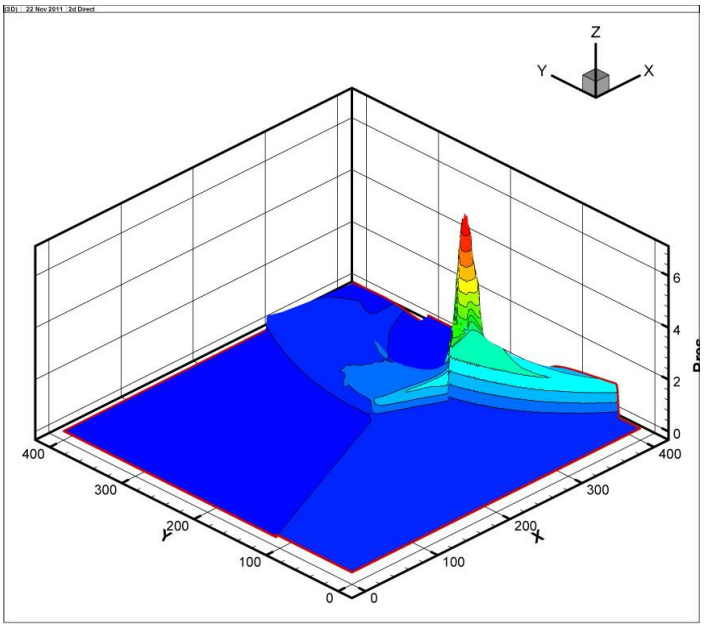


Рис. 2. Распределение давления при IV типе взаимодействия

Для того, чтобы снизить величину максимального статического давления на поверхности тела, мы будем искать некоторое стационарное распределение теплового потока $q(x, y)$. Для этой цели представляется естественным использование целевого функционала

$$\varepsilon(q) = \max_{\Gamma} p = \int_{\Gamma} p\delta(s - s_{\max})d\Gamma \tag{4}$$

($S_{\max}$ - координаты точки максимума давления на теле) минимизируемого итерационно. При этом интегральная мощность тепловых источников

$$\eta = \int_{\Omega} qd\Omega \tag{5}$$

также должна быть минимальна. Здесь, в отличие от работ [4,5], мы используем постоянные по времени источники тепла, а не возмущения начального поля температуры.

Однако в связи с возможными перемещениями зоны повышенного давления по поверхности тела при вариации источников, функционал (4) не обладает достаточной гладкостью, что препятствует применению градиентных методов. В связи с этим в работе использовались функционалы типа:

$$\varepsilon(q) = 1/2\int_{\Gamma} p^n d\Gamma + \alpha/2\int_{\Omega}\left(q(x,y)\right)^2 d\Omega,\ n = 2,4... \tag{6}$$

Здесь α-коэффициент регуляризации, обеспечивающий поиск минимального возмущения (5) и подбираемый в ходе расчетов. Увеличение показателя n в выражении $\varepsilon(q) = 1/2\int_{\Gamma} p^n d\Gamma, n = 1,2...$, делает функционал более чувствительным к изменению давления и, вообще говоря, позволяет более точно выявлять зоны максимального давления. Однако, для данного течения результаты расчетов при $n = 1$ и $n = 2$ практически не различаются по форме (с точностью до масштаба), что делает их эквивалентными при использовании градиентных методов оптимизации.

Поиск оптимального управления осуществляется итерационным методом с использованием метода наискорейшего спуска

$$q^{n+1}(x,y) = q^n(x,y) - \tau\nabla\varepsilon_q \tag{7}$$

Каждая итерация соответствует решению одной прямой и одной сопряженной задачи. Ключевым моментом здесь является расчет градиента целевого функционала. Для расчета последнего использованы сопряженные уравнения, имеющие форму [5]:

$$\frac{\partial\Psi_\rho}{\partial t} + \frac{\partial\Psi_i}{\partial t}U^i + \frac{\partial\Psi_e}{\partial t}E + U^k\frac{\partial\Psi_\rho}{\partial X^k} + U^kU^i\frac{\partial\Psi_i}{\partial X^k} + (\gamma-1)\frac{\partial\Psi_k}{\partial X^k}e + U^kh_0\frac{\partial\Psi_e}{\partial X^k} = 0 \tag{8}$$

$$\frac{\partial\Psi_i}{\partial t}\rho + \frac{\partial\Psi_e}{\partial t}\rho U^i + U^i\frac{\partial\Psi_k}{\partial X^i} + U^i\frac{\partial\Psi_i}{\partial X^k} + \frac{\partial\Psi_\rho}{\partial X^k} + h_0\frac{\partial\Psi_e}{\partial X^k} + U^iU^k\frac{\partial\Psi_e}{\partial X^i} = 0 \tag{9}$$

$$\frac{\partial\Psi_e}{\partial t} + \gamma U^k\frac{\partial\Psi_e}{\partial X^k} + (\gamma-1)\frac{\partial\Psi_k}{\partial X^k} = 0 \tag{10}$$

Начальные условия:

$$\Psi_{\rho,U,V,e}\big|_{\tau=0} = 0. \tag{11}$$

Граничные условия:

$$\Psi_{\rho,U,V,e}\big|_{\partial\Omega} = 0, \tag{12}$$
$$\Psi_\rho\big|_\Gamma = nP^{n-1}(\gamma-1)e,\ \Psi_e\big|_\Gamma = nP^{n-1}(\gamma-1)\rho.$$

Градиент целевого функционала можно определить, используя решение сопряженной задачи

$$\nabla \varepsilon_q = \Psi_e(t, x, y) + \alpha q(x, y). \quad (13)$$

Сопряженные уравнения рассчитывались с помощью метода [12].

III. Результаты расчетов

Распределение величины градиента в расчетном поле представлено на Рис. 3. Можно различить три зоны, нагрев в которых уменьшит давление на поверхности и одну, которая позволит увеличить давление. Первая (со стороны малых *Y*) зона отрицательного градиента находится в окрестности падающего скачка, что соответствует управлению течением за счет смещения падающего скачка. Вторая зона отрицательного градиента, практически смыкающаяся с первой, соответствует попаданию линии тока строго в область высоконапорной струйки, третья соответствует возмущению, распространяющемуся по противоположному, относительно падающего скачка, семейству характеристик.

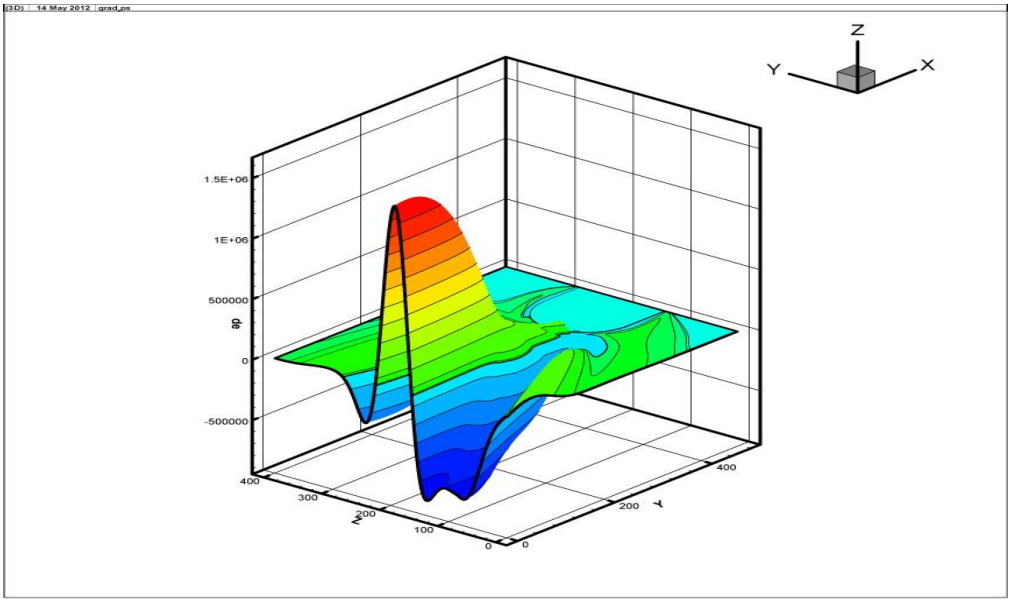


Рис. 3. Градиент целевого функционала при IV режиме взаимодействия

Градиент целевого функционала верно отражает изменения при малых величинах тепловых источников, однако при относительно интенсивных источниках, взаимодействие типа IV разрушается при любом их расположении. В связи с этим, выбор зоны воздействия с помощью градиента для интенсивных источников не целесообразен. Из Рис. 3 видно, что градиент в разных зонах имеет разные знаки, что при итерациях может привести к отрицательной плотности тепловых источников и, как следствие, нефизичности результата. Потому в данных расчетах во внимание принимались только тепловые источники с положительной плотностью.

При решении задачи (6) особую роль играет коэффициент регуляризации α. При малой величине коэффициента регуляризации α тепловые источники полностью искажают поле течения. Например, для значения $\alpha = 0$ мощность тепловых источников составила $\eta = 0.84$, что существенно больше, чем в работе [5], $P_{\max} = 0.82$. Поле течения подверглось существенной перестройке. Увеличение коэффициента регуляризации позволяет уменьшить интегральную интенсивность источников за счет снижения эффективности управления. Для $\alpha = 1$ распределение тепловых потоков представлено на Рис. 4, а соответствующее ему распределение давления на Рис. 5. При этом мощность тепловых источников $\eta = 0.0247$, максимальное давление $P_{\max} = 2.3$.

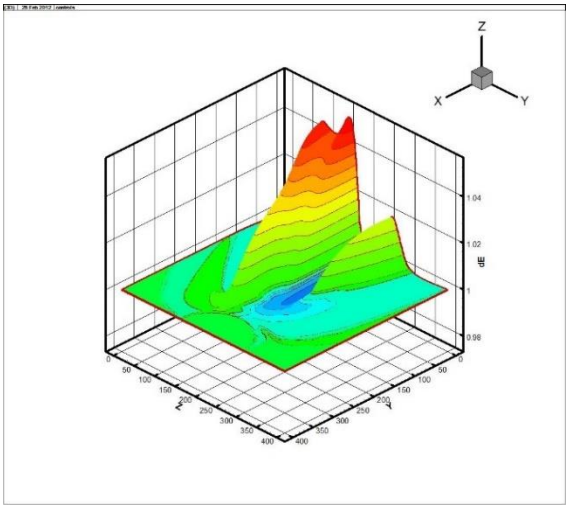


Рис. 4. Плотность теплового источника по результатам расчетов, $\eta = 0.0247$

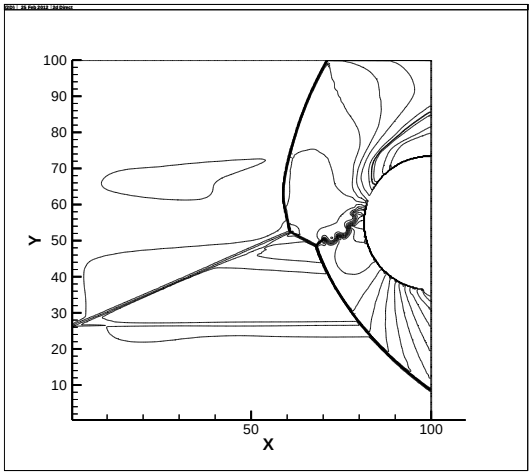


Рис. 5. Поле давления для теплового источника, $\eta = 0.0247$

Следует отметить, что в данном случае поле течения подверглось перестройке, соответствующей переходу к типу взаимодействия IVa.

Данный подход был сравнен с результатами работы Когана и Стародубцева [5], в которой использовался тепловой источник следующей формы:

$$q = q_s \exp(-\frac{r^2}{r_s^2}) \ , \ q_s = 2 , \ r_s = 0.1 \tag{14}$$

Для параметров течения $M = 6$, $\Theta_1 = 15^o$ (угол отклонения потока) результаты [5] воспроизводятся, при этом удается понизить максимальное значение давления с $P_{\max} = 6.587$ до $P_{\min} = 0.88$. Требуемая при этом мощность тепловых источников $\eta = 0.068$,

Расчеты демонстрируют возможность поиска управляющих воздействий с помощью решения обратной задачи а также существенное влияние коэффициента регуляризации. Результаты решения обратной задачи демонстрируют трансформацию течения не по сценарию “преломления скачка” с переходом от структуры IV к V [5], а переход от взаимодействия IV типа, к структуре типа IVa. Полученный в результате тепловой источник является достаточно гладким и занимает значительную часть расчетной области, то есть относится к другому классу воздействий в сравнении с [5].

IV. Обсуждение

Целью данной работы является поиск теплового источника, который переводил бы течение в состояние, соответствующее снижению максимального статического давления на поверхности. В отличие от работы [5], в которой рассмотрено управление течением в системе, имеющей два устойчивых состояния (переход между маховским и регулярным отражениями скачков) с помощью импульсного нагрева, в данной работе тепловые источники постоянны по времени, в случае их выключения течение вернется к начальному состоянию.

В численных экспериментах удается достаточно легко найти тепловые источники, способные разрушить низкоэнтропийную струйку, представляющую основную опасность с точки зрения высоких значений статического давления на поверхности. Это связано с относительно узким диапазоном параметров, в котором существует взаимодействие IV рода. Более сложным является поиск воздействия, минимального с точки зрения суммарной мощности тепловых источников.

В работе рассмотрено два различных варианта выбора управляющего воздействия. Первый использует “преломление скачка” в соответствии с результатами [5] и такой же суммарной мощностью тепловых источников. Второй подразумевает поиск управления с помощью решения обратной задачи.

“Преломление скачка ” дает существенное снижение максимального давления на теле, однако достаточно энергоемко.

Поиск минимального управления с помощью решения обратной задачи дает существенное снижение затрат энергии и соответствует трансформации течения из типа IV в тип IVa. При этом тепловые источники достаточно малы по амплитуде, но занимают большую область в пространстве, что относит их к другому классу воздействий, по сравнению с теми, которые формируют “ преломление скачка”.

V. Заключение

Рассмотрены следующие варианты управления течением на режиме взаимодействия IV типа с помощью источников тепла в поле течения:

- “Преломление скачка ” с помощью локального интенсивного источника тепла.
- Трансформация взаимодействия скачков типа IV в тип IVa с помощью распределенных источников тепла умеренной интенсивности.

Результаты расчетов указывают на то, что последний из приведенных методов требует меньших затрат энергии. Эти два метода соответствуют качественно разным распределениям тепловых источников.

Визуализация поля сопряженных параметров позволяет определить зоны, в которых тепловое воздействие наиболее продуктивно, а в которых, наоборот. Визуализация поля течения позволяет определить изменения в структуре течения, вызванные управляющими параметрами. В целом для успешного управления структурой течения требуется одновременная визуализация распределения параметров течения, сопряженных параметров и тепловых источников.

Список источников

[1] Adelgren R. G., et al, Control of Edney IV Interaction by Pulsed Laser Energy Deposition//AIAA J.- 2005.- V. 43, № 2.-P. 256-263.

[2] Yan H., Gaitonde D., Effect of Energy Pulse on 3-D Edney IV Interaction, AIAA J. 2008. V. 46. № 6. P. 1424-1431.

[3] Коган М.Н., Стародубцев М.А., Уменьшение пиковых тепловых потоков путем подвода тепла в набегающий поток, Изв. РАН, сер. Механика жидкости и газа, N 1, 2003, 134-146

[4] Георгиевский П.Ю., Левин В.А., Сверхзвуковое обтекание тела при наличии внешних источников тепловыделения// Письма в ЖТФ. 1988, Т. 14. С. 684-687.

[5] Алексеев А.К., К управлению переходом между регулярным и Маховским режимами взаимодействия ударных волн, ЖВМ и МФ. 2012. т. 52. № 6. С. 1134-1142.

[6] B. Edney, Effects of Shock Impingement on the Heat Transfer around Blunt Bodies. *AIAA J.,* 6(1) (1968) 15-21.

[7] Боровой В.Я., Течение газа и теплообмен в зонах взаимодействия ударных волн с пограничным слоем.- М.: Машиностроение, 1983.-128 с.

[8] van Leer B. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov’s method//J. Comput. Phys. 1979. V. 32. P. 101–136

[9] Sun M., Katayama K. An artifically upstream flux vector splitting for the Euler equations//JCP. 2003. V. 189. P. 305-329.

[10] Farooq M. A., Muller B., Accuracy assessment of the Cartesian grid method for compressible inviscid flows using a simplified ghost point treatment, J. of Structural Mechanics, V. 44, № 3, 2011, pp. 279-291.

[11] Gorsse Y., Iollo A., Telib H., Weynans L., A simple second order Cartesian scheme for compressible Euler flows. Inria, RR-7773. 2011. P. 24

[12] Toro E. and Siviglia A. PRICE: Primitive centred schemes for hyperbolic system of equations//Int. Journal for Numerical Methods in Fluids. 2003 V. 42 P. 1263–1291.

[13] Borovoy V. Ya. et al, Interference Between a Cylindrical Bow Shock and a Plane Oblique Shock//AIAA J. 1197. V. 35, №. 11. P. 1721-1728. (2013), pp. 1758–1789.

УДК 677.027
DOI: 10.25206/978-5-8149-3873-2-2024-271-282

Создание цифровой модели намоточного изделия
Development of a Digital Model of a Winding Product

О. В. Томилова, Н. В. Рокотов
1 ФГБОУ ВО «Санкт-Петербургский государственный университет промышленных технологий и дизайна», Санкт-Петербург, РФ
2 ФГБОУ ВО «Санкт-Петербургский государственный университет промышленных технологий и дизайна», Санкт-Петербург, РФ

O. V. Tomilova, N. V. Rokotov
[1]Saint Petersburg State University of Industrial Technologies and Design, Saint Petersburg, Russian Federation
[2]Saint Petersburg State University of Industrial Technologies and Design, Saint Petersburg, Russian Federation

Аннотация. В статье рассматривается проблема анализа и проектирования намоточных изделий с использованием компьютерного представления в виде 3D моделей. Авторы предлагают использовать аппарат цепных дробей для описания структуры намоточных изделий и оптимизации производственных процессов. Статья содержит информацию о процессе создания 3D модели намоточного изделия, а также о возможностях уменьшения вычислительных ресурсов для моделирования. Представлены результаты компьютерного моделирования различных типов намоток. В заключении статьи авторы делают вывод о том, что разработанная методика анализа и синтеза структуры прецизионной намотки с использованием аппарата цепных дробей позволяет исследовать различные типы намоток и создавать оптимальные 3D модели, повторяющие их структуру. Эти модели могут быть использованы для дальнейших исследований и в качестве цифровых моделей намоточных изделий.

***Ключевые слова*:** трёхмерная модель намотки, компьютерное моделирование, намоточное изделие, цифровая модель, типы намоток, виртуальный двойник, получение намоток, расчёт намотки, расчёт модели намотки, экспериментальный образец.

Аннотация. The article deals with the problem of analysis and design of winding products using computer representation in the form of 3D models. The authors propose to use the apparatus of chain fractions to describe the structure of winding products and optimize production processes. The article contains information about the process of creating a 3D model of a winding product, as well as about the possibilities of reducing computational resources for modeling. The results of computer simulation of various types of windings are presented. In conclusion, the authors conclude that the developed method of analysis and synthesis of the structure of precision winding using the apparatus of chain fractions allows us to study different types of windings and create optimal 3D models that repeat their structure. These models can be used for further research and as digital models of winding products.

***Ключевые слова*:** three-dimensional winding model, computer simulation, winding product, digital model, winding types, virtual twin, winding production, winding calculation, winding model calculation, experimental sample.

I. Введение

Намоточные изделия (НИ), представляющие из себя законченный продукт или заготовку, которые применяют для создания композиционных материалов, электрических катушек различных форм и назначений, фильтрующих элементов намоточного типа, в которых слой намотанных нитей различной природы от объёмных текстурированных нитей до проволоки образовывают фильтрующие перегородки для задержания механических частиц, жидкости и газа.

Вопросам применения и создания волокнистых полимерных композиционных материалов (ПКМ) в промышленности посвящено множество работ. Так, например, трубам в нефтехимической промышленности посвящена статья [1]. Обеспечению создания легких и прочных материалов на основе ПКМ в сфере авиационного и ракетного производства посвящены работы: [2, 3, 4, 5, 6] и другие. ПКМ широко применяются и в области судостроения, об этом свидетельствуют статьи [7, 8]. В химической промышленности востребованы трубы с определённым набором гидравлических и механических свойств, исследования по этому направлению описаны в статьях [9, 10]. В машиностроении также востребованы легкие и прочные материалы на основе ПКМ. Исследованиям в этой области посвящены работы: [11, 12].

В качестве фильтрующих аэраторов намоточные изделия представлены в работах [13, 14, 15].

II. Постановка задачи

Традиционные методы анализа и проектирования намоточных изделий ограничены своими возможностями. Сложная внутренняя структура намотки делает трудным анализ ее характеристик и оптимизацию производственных процессов. Отсутствие эффективных средств визуализации и анализа структуры намоточных изделий затрудняет проектирование, приводя к высоким издержкам и времени разработки. Одним из направлений изучения свойств намоточных изделий является разработка их компьютерного представления в виде 3D моделей. Применение данного подхода позволяет на этапе проектирования намоточных изделий анализировать внутреннее строение НИ, определять оптимальные параметры процесса намотки, решать задачи оптимизации структуры НИ с целью получения требуемых физико-механических характеристик (прочность, жёсткость, проницаемость изделия).

Созданные цифровые модели могут получить широкое применение при создании композитных материалов, задании свойств фильтрующим компонентам, определении качества крашения пряжи в паковках и в других сферах применения.

В статье [16] был рассмотрен процесс создания 3D модели намоточного изделия. Высота его составила 250 мм, учитывая ещё большое количество витков, более 100, модель получилась «тяжелой» и мало пригодной для последующего расчета.

Одной из возможностей уменьшения вычислительных ресурсов, требуемых для моделирования, является вычленение из тела НИ элемента, достоверно отражающего свойство структуры НИ.

III. Теория

В статье [17] показано, что для описания структур тел НИ может быть эффективно использован аппарат цепных дробей.

Исходными данными для синтеза поковки служат:

- диаметр d наматываемой нити круглого поперечного сечения;
- размеры наматываемой паковки (*B* – ширина паковки, мм; *D* – текущий диаметр паковки, мм);
- пределы, в которых должен находиться угол подъема витков нити a, град.

Структуру НИ можно описать последовательностью целых чисел $\alpha_0, \alpha_1, \ldots \alpha_n$. Число α_0 характеризует шаг спирали наматываемой нити H, а в совокупности со значениями *B* и *D* - угол подъема витков нити, причем $a_0 < \frac{2B}{H} < a_0 + 1$.

Для того чтобы обеспечить необходимую структуру намотки с заданными размерами шагов соответствующих порядков, можно использовать представление k в виде цепной дроби:

$$K = \frac{2*B}{H} = \frac{500}{124} = 4 + \frac{4}{124} = 4 + \frac{1}{31} \quad (1)$$

$a_0=4; a_1=31; a_2=a_3 \ldots =0$

Допустим мы хотим получить намотку со следующими параметрами:

Высота намотки = 250 мм, $H \sim 124$ мм, $H_1 \sim 1$ мм, тогда $a_0 = E\left(\frac{500}{124}\right) = 4$, а формула (1) при заданном виде намотки примет вид $k = 4 + \frac{1}{124} = \frac{497}{124}$

к представляет собой передаточное отношение. За 497 оборотов паковки нитеводитель сделает 124 двойных хода.

Найдем *H*

$$H = \frac{124}{497} * 500 = 124.7485;$$

Тогда H_1 можно найти по формулам:

$$H_1 = min \begin{cases} 2B - H * E\left(\frac{2*B}{H}\right) \\ H * \left[E\left(\frac{2*B}{H}\right) + 1\right] - 2 * B \end{cases} \quad (2)$$

$$H_1 = min \begin{cases} 2 * 250 - 124.7485 * E\left(\frac{2 * 250}{124.7485}\right) = 1.0060 \text{ мм} \\ 124.7485 * \left[E\left(\frac{2 * 250}{124.7485}\right) + 1\right] - 2 * 250 = 123.7425 \text{ мм} \end{cases}$$

На основе вычислений искомая намотка имеет следующие параметры:

Передаточное отношение $\frac{497}{124}$-; *B*-250 мм; *H*=124,7485 мм; H_1=1,0060 мм; H_2=0 мм (см. Таблица 1)

На основании формулы (1) мы видим, что структура поковки имеет повторяющиеся элементы. Для экономии резервов компьютера оставим 2 витка первого порядка и найдём новые параметры передаточных отношений и высоту намотки, сохранив при этом структуру намотки (H и H_1).

$$k = 2 + \frac{1}{124} = \frac{249}{124};$$

$$H = \frac{124}{249} * \big(2 * B - H * (4 - 2)\big) = \frac{124}{249} * (500 - 124.7485 * 2) \ = 124.7485 \text{ мм}$$

Выведем формулу для получения высоты намотки при двух витках первого порядка:

$$B' = (2B - H * (E - 2))/2 \quad (3)$$

Из полученного выражения получим B’

$$B' = \frac{500 - 124.7485 * 2}{2} = 125.2515 \text{ мм}$$

На основе полученных данных можно выполнить все вычисления для остальных намоточных изделий и построить компьютерные модели различных структур и подобрать необходимые параметры для реализации их на намоточных устройствах. (см. Таблица 1.) Для большей наглядности построены модели с двумя витками.

Эта последовательность представляет собой коэффициенты цепной дроби полученной в результате разложения рациональной дроби $k = \frac{2B}{H}$, например, с помощью алгоритма Евклида.

$$k = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}$$

Рассмотрим повторяющиеся витки нити на развертке (рис.1).

Треугольники ABC, MDE – равнобедренные. Высоты BH и DK делят основание треугольников на равные части. Следовательно, AH+LM=HC+CL,

А это означает, что HL= H/2, где H – шаг витка.

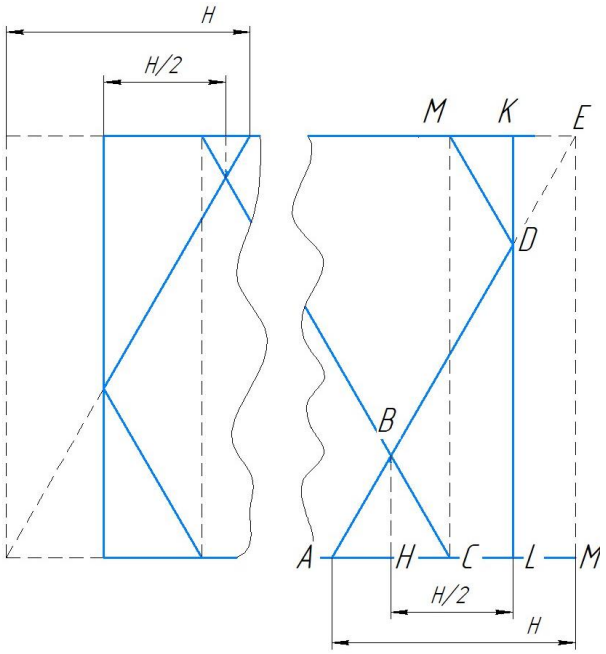


Рисунок 1 Развертка намотки нити

IV. Результаты экспериментов

ТАБЛИЦА 1
НАМОТОЧНЫЕ ИЗДЕЛИЯ

Сомкнутая структура
$u = \frac{491}{49} = 10 + \frac{1}{49} = 9 + \frac{99}{49}$;
B = 250 мм
D = 95 мм
H=49.8982 мм
H_1=1.0183 мм
H_2=0
α=0.1747

Та же структура, но при других значениях передаточного отношения
$u = \frac{99}{49} = 2 + \frac{1}{49}$;
B’ = 50.4073 мм
D = 95 мм
H=49.8982 мм
H_1=1.0183 мм
H_2=0
α=0.1747
Диаметр нити d = 0.9 мм

Рисунок 3 Паковка нитей с сомкнутой структурой намотки, передаточное отношение $u = \frac{491}{49}$

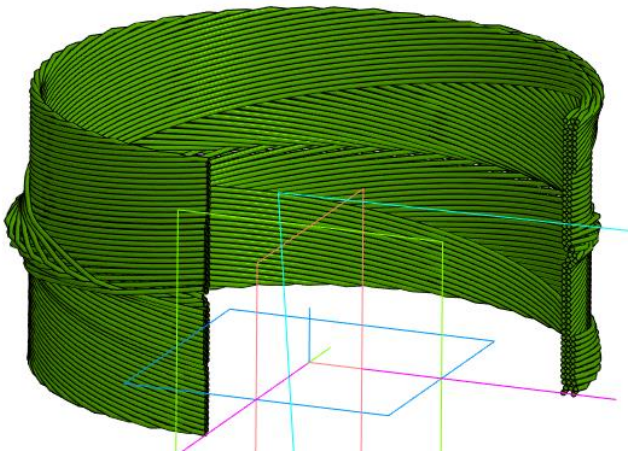


Рисунок 4 Компьютерная модель намотки нитей с сомкнутой структурой, передаточное отношение $u = \frac{99}{49}$

Сомкнутая структура
$u = \frac{497}{124} = 4 + \frac{1}{124}$;
B = 250 мм
D = 95 мм
H=124.7485 мм
H_1=1.0060 мм
H_2=0
α=0.4155
Диаметр нити d= 0.85 мм

Таже структура, но при других значениях передаточных отношений
$u = \frac{249}{124} = 2 + \frac{1}{124}$;
B’ = 125.2516 мм
D = 95 мм
H=124.7485 мм
H_1=1.0060 мм
H_2=0
α=0.4155
Диаметр нити d= 0.85 мм

Рисунок 5 Паковка нитей с сомкнутой структурой намотки, передаточное отношение $u = \frac{497}{124}$

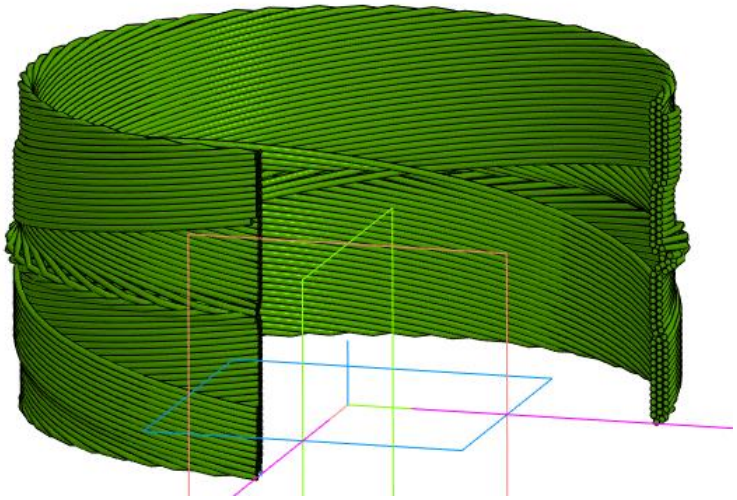


Рисунок 6 Компьютерная модель намотки нитей с сомкнутой структурой, передаточное отношение $u = \frac{249}{124}$

Сомкнутая структура
$u = \frac{494}{61} = 8 + \frac{1}{124}$;
B = 250 мм
D = 95 мм
H=61.7409 мм
H_1=6.0729 мм
H_2=1.0121
α=0.4155
Диаметр нити d= 1 мм

Таже сомкнутая структура, но при других значениях передаточных отношений
$u = \frac{128}{61} = 2 + \frac{6}{61} = 2 + \frac{1}{10 + \frac{1}{6}}$;
B’ = 64.7773 мм
D = 95 мм
H=61.7409 мм
H_1=6.0729 мм
H_2=1.0121
α=0.4155
Диаметр нити d= 1 мм

Рисунок 7 Паковка нитей с сомкнутой структурой намотки, передаточное отношение $u = \frac{494}{61}$

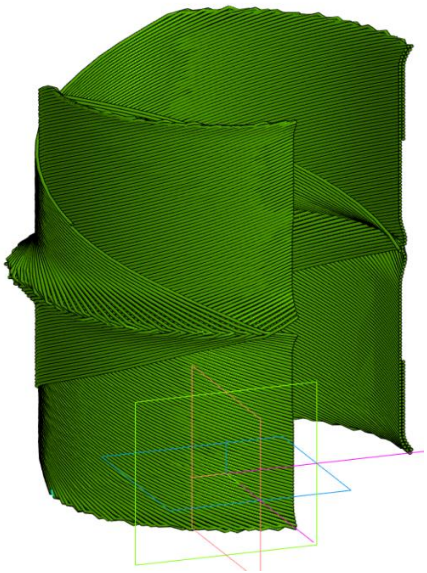


Рисунок 8 Компьютерная модель намотки нитей с сомкнутой структурой, передаточное отношение $u = \frac{128}{61}$

Сотовая структура
$u = \frac{46}{5} = 9 + \frac{1}{5};$
B = 250 мм
D = 95 мм
H=54.3478 мм
H_1=10.8696 мм
H_2=0
α=0.2124
Диаметр нити d= 1 мм

Таже сотовая структура, но при других значениях передаточных отношений
$u = \frac{11}{5} = 2 + \frac{1}{5};$
B’ = 59.3478 мм
D = 95 мм
H=54.3478 мм
H_1=10.8696 мм
H_2=0
α=0.2124
Диаметр нити d= 1 мм

Рисунок 9 Компьютерная модель намотки нитей с сотовой структурой, передаточное отношение $u = \frac{46}{5}$

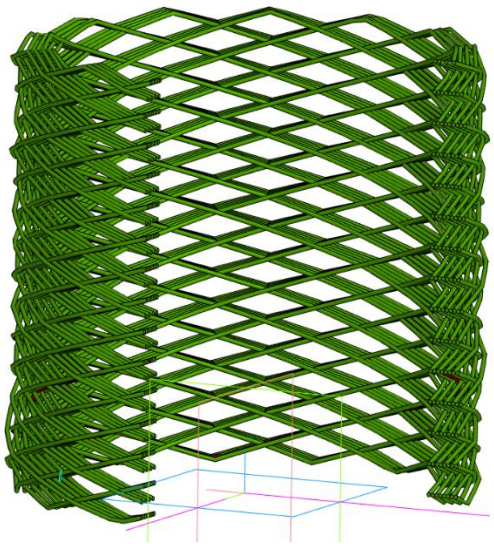


Рисунок 10 Компьютерная модель намотки нитей с сотовой структурой, передаточное отношение $u = \frac{11}{5}$

Сотовая структура

$$u = \frac{68}{13} = 5 + \frac{3}{13} = 5 + \frac{1}{4 + \frac{1}{3}};$$

B = 250 мм
D = 95 мм
H=95.5882 мм
H_1=22.0588 мм
H_2=7.3529 мм
α=0.3260
Диаметр нити d= 1 мм

Таже сотовая структура, но при других значениях передаточных отношений

$$u = \frac{29}{13} = 2 + \frac{29}{13} = 2 + \frac{1}{4 + \frac{1}{3}};$$

B = 106.61765 мм
D = 95 мм
H=95.5882 мм
H_1=22.0588 мм
H_2=7.3529 мм
α=0.3260
Диаметр нити d= 1 мм

Рисунок 11 Компьютерная модель намотки нитей с сотовой структурой, передаточное отношение $u = \frac{68}{13}$

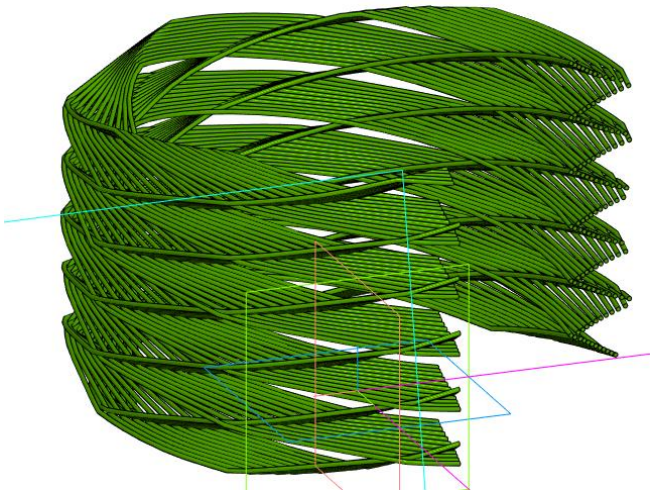


Рисунок 12 Компьютерная модель намотки нитей с сотовой структурой, передаточное отношение $u = \frac{29}{13}$

Спиралевидная структура

$$u = \frac{992}{121} = 8 + \frac{24}{121} = 8 + \frac{1}{5 + \frac{1}{24}};$$

B = 250 мм
D = 95 мм
H=60.9879 мм
H_1=12.0968 мм
H_2=0.5040 мм
α=0.2124
Диаметр нити d= 1 мм

Таже структура, но при других значениях передаточных отношений

$$u = \frac{266}{121} = 2 + \frac{24}{121} = 2 + \frac{1}{5 + \frac{1}{24}};$$

B’ = 67.0363 мм
D = 95 мм
H=60.9879 мм
H_1=12.0968 мм
H_2=0.5040 мм
α=0.2124
Диаметр нити d= 1 мм

Рисунок 13 Компьютерная модель намотки нитей со спиралевидной структурой, передаточное отношение $u = \frac{992}{121}$

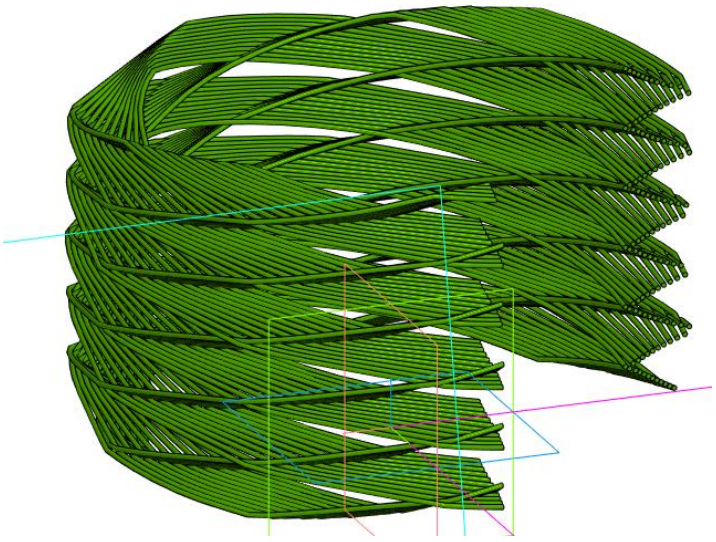


Рисунок 14 Компьютерная модель намотки нитей со спиралевидной структурой, передаточное отношение $u = \frac{266}{121}$

Спиралевидная опережающая

$$u = \frac{1468}{179} = 8 + \frac{36}{179} = 8 + \cfrac{1}{5 + \cfrac{1}{8 + \cfrac{1}{11 + \cfrac{1}{1 + \frac{1}{2}}}}}$$

B = 250 мм
H=60.9673 мм
H_1=12.2616
H_2=0.3406
α=0.2124

Таже спиралевидная структура, но при других значениях передаточных отношений

$u = \frac{394}{179} = 2 + \frac{36}{179}$;
B’ = 67.0981 мм
D = 95 мм
H=60.9673 мм
H_1=12.2616
H_2=0.3406
α=0.2124
Диаметр нити d= 1 мм

Рисунок 15 Компьютерная модель намотки нитей со спиралевидной опережающей структурой, передаточное отношение $u = \frac{1468}{179}$

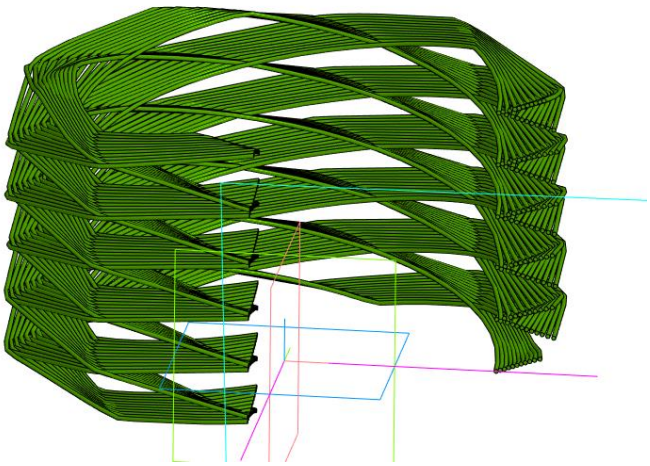


Рисунок 16 Компьютерная модель намотки нитей со спиралевидной опережающей структурой, передаточное отношение
$u = \frac{394}{179}$

Сведем все данные в таблицу.

ТАБЛИЦА 2
СТРУКТУРА ПАКОВКИ

		B, мм	H	H_1	H_2	α	a_0	a_1	a_2
1	**а**	250	124,7485	1,006	0	0,3960	4	124	0
	б	125,2515	124,7485	1,006	0	0,3960	4	124	0
2	**а**	250	49,8982	1,0183	0	0,1672	10	49	0
	б	50,4073	49,8982	1,0183	0	0,1672	2	49	0
3	**а**	250	61,7409	1,021	1,0121	0,2070	8	10	6
	б	64,7773	61,7409	1,021	1,0121	0,2070	2	10	6
4	**а**	250	54,3478	10,8696	0	0,1821	9	5	0
	б	59,7827	54,3478	10,8696	0	0,1821	2	5	0
5	**а**	250	95,5882	22,0588	7,3529	0,3101	5	4	3
	б	106,61765	95,5882	22,0588	7,3529	0,3101	2	4	3
6	**а**	250	60,9879	12,0968	0,5040	0,2045	8	5	24
	б	67,0363	60,9879	12,0968	0,5040	0,2045	2	5	24
7	**а**	250	60,9673	12,2616	0,3406	0,2044	8	4	5
	б	67,0981	60,9673	12,2616	0,3406	0,2044	2	4	5

V. Выводы и заключение

Рассмотрено формирование намотки в виде цепных дробей, выделены повторяющиеся элементы, которые представлены на развертке намотки нити и проанализированы с точки зрения геометрии. Представлены 3D модели разных типов намоток. Приведены расчёты паковки с заранее заданными параметрами. Выделены повторяющиеся элементы структуры намоточного изделия. На их основе выведена формула расчета паковки с двумя витками (аналогичная формула с одним витком). Таким образом, может быть создана оптимальная 3D модель повторяющая структуру любых типов намотки. Модель может быть применена для дальнейших исследований и использована в качестве цифровой модели намоточного изделия.

Литература

1. Котов Н. С. Нефть и газ. "Аттика"// Сфера. 2021. №3(82). С. 66-69.
2. Васильев В. В., Барынин В. А., Разин А. Ф. Анизогридные композитные сетчатые конструкции - разработка и приложение к космической технике // Композиты и наноструктуры. 2009. № 3(3). С. 38-50.
3. Серкейра Х., Фариа Х., Фанк, Р.. Изготовление композитных цилиндров с интегрированной решетчатой структурой с использованием намотки. // 16-я Европейская конференция по композитным материалам: ECCM 2014. С. 1-8.
4. Муриц А.П., Геллерт Э., Берчилль П., Чаллис К. Обзор перспективных композитных конструкций для кораблей и подводных лодок ВМФ. // Композитные конструкции. 2001. С. 21-42.
5. Прядицкий А. А. Обзор исследований по проектированию сетчатых конструкций из композитов для авиа- и ракетостроения // Вопросы проектирования и производства конструкций летательных аппаратов. 2014. №1. С. 15-21.
6. Самипур С. А. Проектирование и технология изготовления сетчатых конструкций летательных аппаратов с плетёной системой армирования: дис. кан-та техн. наук / Самипур Саджад Алиасгар. Казань: 2018. 134 с.
7. Судостроение. Юматекс, Госкорпорация РОСАТОМ. URL: https://umatex.com/applications/shipbuilding/ (дата обращения: 09.07.2024)
8. Гненной , Олег Композиты в Российском судостроении: каковы перспективы?. URL: https://www.korabel.ru/news/comments/kompozity_v_rossiyskom_sudostroenii_kakovy_perspektivy.html (дата обращения: 09.07.2024)
9. Радулови Ю. Композитные пластиковые трубы с намоткой нити: взаимосвязь между намоточными конструкциями и их гидравлическими и механическими свойствами. // Научно-техническое обозрение. 2011 Т. 61. № 3-4. С. 73 – 77.
10. Система Heisskraft-therm Руководство по проектированию и монтажу напорных трубопроводов, изготовленных из полипропилена рандомсополимера PP-R. // Heisskraft Plastic Piping Materials. 82 с.

11. Отмахов Д. В. Исследование полимерных композиционных материалов, полученных методом намотки. // Ученые заметки ТОГУ. 2014. Т. 5. №2. С. 97-102.

12. Захарычев С П, Иванов В А. Разработка установки радиальной намотки для изготовления армированных полимерных композиционных материалов // Вестник ТОГУ. 2008. №1(8). С. 101 – 110.

13. Панин А. И. Построение модели зависимости удельной плотности намотки от угла сдвига витков и толщины наматываемого слоя пряжи. //Вестник ДИТУД. 2001. №3.

14. Губейдуллина Х. Х., Шигапов И. И., Кадырова А. М. Аэрация сточных вод в животноводческих фермах // Вестник УГСХА 2012 № 4(20). С 114-117.

15. Николаева С. Д., Панин И. Н., Цимбалюк Е. П., Бояркина М. А. О структуре диспергирующего слоя пневматических аэраторов «Пантекс» // Известия высших учебных заведений .Технология текстильной промышленности. 2010. №2(323). С. 48-52.

16. Рокотов Н.В., Томилова О.В., Марковец А.В., Пономарь А.А. Разработка 3D-модели намоточного изделия // Вестник Санкт-Петербургского государственного университета технологии и дизайна. Серия 4: Промышленные технологии. 2022. № 3. С. 25-30.

17. Рокотов Н.В., Беспалова И.М., Марковец А.В. Анализ прецизионной намотки // Известия высших учебных заведений. Технология легкой промышленности. 2016. Т. 32. № 2. С. 22-26

УДК 004.032.26
DOI: 10.25206/978-5-8149-3873-2-2024-283-291

Роль плотности информационного поля объекта в графических нейросетях
The role of object information field density in graphical neural networks

Н.А. Бондарева [1]
[1] *ИПМ им. М.В. Келдыша РАН, Миусская пл., д.4, Москва, 125047, Россия*

N.A. Bondareva [1]
[1] *Keldysh Institute of Applied Mathematics, Miusskaya sq. 4, Moscow, 125047, Russia*

Аннотация. Работа предлагает краткий обзор генеративных нейросетей и рассматривает роль информации при обучении генеративных нейросетей. В цифровой среде каждый объект окружён обширным информационным полем, включающим в себя неупорядоченную информацию и совокупность упоминаний о нём. Плотность информационного поля объекта определяет способность таких технологий, как искусственный интеллект, воссоздавать его образ на основе собранных данных. Чем больше данных доступно, тем точнее и полнее можно воссоздать цифровой образ. В работе рассматривается ряд проблем, возникающих при использовании сетей text-to-image и возможные методы их решения. Статья рассматривает различные аспекты роли личных данных и возможных этических и социальных последствий в эпоху генеративных технологий, а также перспективы и риски дальнейшего развития генеративных нейросетей в специализированных областях, таких как медицина и производство. Стремительное развитие нейросетевых технологий может оказать значительное влияние на образование и социальные явления.

***Ключевые слова*:** Машинное обучение, компьютерное зрение и распознавание образов, нейронная сеть, компьютерная графика, плотность информационного поля, Text-to-image

Abstract. This paper offers a brief overview of generative neural networks and examines the role of information in learning generative neural networks. In a digital environment, every object is surrounded by a vast information field that includes unordered information and a collection of references about it. The density of an object's information field determines the ability of technologies such as artificial intelligence to recreate its image based on collected data. The more data available, the more accurate and complete the digital image can be recreated. The paper discusses a number of problems that arise in the use of text-to-image networks and possible methods of solving them. The paper considers various aspects of the role of personal data and possible ethical and social implications in the era of generative technologies, as well as the prospects and risks of further development of generative neural networks in specialized fields such as medicine and manufacturing. The rapid development of neural network technologies may have a significant impact on education and social phenomena.

Keywords: Machine learning, computer vision and pattern recognition, neural network, computer graphics, information field density,Text-to-image

I. Введение

В настоящее время область нейросетевых технологий развивается стремительно, с каждым днем обретая все больше умений и возможностей. Особенно популярными становятся генеративные диффузионные модели, в которых берут свою основу большинство текстовых нейросети, способных собирать, анализировать и генерировать текстовую информацию по запросу, и графических которые могут обрабатывать медиа-контент разными способами, от анимирования фотографий до автоматического создания изображений и видео по текстовому запросу. Результаты их работы применяются во многих отраслях, от медиа до медицины, но быстрый прогресс вызывает и социальные изменения. Раньше перспективной альтернативой диффузионным моделям считались генеративно-состязательные нейросети (GAN), но они оказались менее эффективными для генерации изображений по тексту и нестабильными в обучении [1].

Модели диффузии - это итеративные алгоритмы, превращающие случайный шум в изображение. Примером служит модель DDPM (Denoising Diffusion Probabilistic Model) [2], обучаемая на тысячах изображений, к которым последовательно добавляется шум. Модель учится удалять этот шум, улучшая качество изображения. Если применять обученную модель к случайному шуму, она сможет создать новое изображение, постепенно очищая

его от шума. На рисунке показан пример, где пользователь задает схематичный рисунок. Изображение зашумляется, а затем модель восстанавливает его с высокой точностью через процесс обратной диффузии.

Графические нейросети интерпретируют лингвистическую структуру запроса, обрабатывают его и формируют реалистичные визуальные образы. Они управляют несколькими объектами, их атрибутами и пространственными отношениями, устанавливая правильные связи между характеристиками объектов. Основой для таких задач служат диффузионные нейросетевые модели [2,3], которые появились в 2015 году, но получили популярность после работы [2]. Сегодня они достигают впечатляющих результатов в генерации и изменении изображений, таких как генерация изображений, музыки и видео по текстовому запросу (text-to-image), восстановление деталей (inpainting), удаление объектов и увеличение разрешения (super-resolution).

Модели text-to-image используют лингвистическую конструкцию (текстовый запрос), направляющую процесс обработки. Языковые модели, обученные на парах изображений и текстов, понимают содержание обоих типов данных. Например, модель CLIP (Contrastive Language — Image Pre-training) от OpenAI преобразует изображения и тексты в общее латентное векторное пространство, где вектора представляют собой набор значений. В таком пространстве можно находить ближайшие изображения к текстовому запросу, просто оперируя векторами. Модель Latent Diffusion [4], представленная в 2021 году, обучает генерации изображений из направленного шума, используя латентное пространство для текстов и изображений. Такие же принципы применяются в моделях Stable Diffusion, Imagen и других крупных нейросетях для преобразования текста в изображения. Основные принципы их работы описаны в [5-8].

II. Роль информации при обучении генеративных нейросетей

Генеративные нейросети, такие как Generative Adversarial Networks (GAN), Variational Autoencoders (VAE), и трансформеры (например, GPT), представляют собой сложные системы, способные создавать новые данные, которые невозможно отличить от реальных. Роль информации в этих моделях многогранна и критична для их успешного функционирования и обучения. Основываясь на принципах машинного обучения, данные становятся фундаментом, на котором строится обучающая модель, позволяя ей обобщать, интерполировать и экстраполировать закономерности.

Отдельной задачей помимо обучения нейронной сети становится формирование обучающих датасетов [9]. Датасет является основным источником информации, на котором нейросеть строит свою модель. Модель учится на датасете, анализируя его, выявляя паттерны и закономерности, которые она впоследствии использует для генерации. Его качество и разнообразие напрямую влияют на способности генерируемой модели: чем лучше подготовлены данные, тем меньше времени потребуется на отладку модели, ее тренировки, поиск и устранение неточностей распознавания.

Среди основных критериев качественного датасета можно выделить следующие:

1. Полнота данных. Это означает достаточность объема, глубины и широты наборов данных. Набор данных должен содержать достаточное количество параметров или признаков, чтобы не осталось неохваченных пограничных случаев. Неполнота приводит либо к невозможности анализа, либо к необходимости отталкиваться от некоторых предположений или допущений относительно пропущенной информации.
2. Точность. Данные должны быть максимально приближены к реальным условиям, в которых будет работать модель нейросети.
3. Правильность и корректность. Этот пункт подразумевает соответствие данных реальности и корректность их интерпретации, а также соответствия формата и аннотаций данных в датасете с теми, в которых работает фреймворк и архитектура модели нейросети.
4. Однородность (uniformity) — значения всех атрибутов должны быть сравнимыми для всех данных. Неравномерности или наличие выбросов в наборах данных отрицательно влияют на качество данных обучения.
5. Наличие разделенных между собой датасетов для обучения, валидации и тестирования.

Большие языковые модели (LLM) произвели революцию в обработке естественного языка в таких задачах, как понимание прочитанного, рассуждение и генерация языка. Они представляют собой мощные инструменты в области обработки естественного языка (NLP). Такие модели обучаются на основе огромных наборов текстовых данных, и в результате оказываются способны улавливать сложные паттерны и нюансы языка.

Основной принцип работы LLM заключается в использовании архитектуры трансформеров, которая была предложена в работе [10]. Трансформеры позволяют выявлять и анализировать зависимости между словами в предложении, независимо от их позиции. Это улучшает качество генерации текста и понимания контекста.

LLM активно используются в генеративных нейросетях многих типов, — в частности в таких, как «text-to-image», «text-to-video», и «text-to-3D» — а также при разнообразной обработке текста и текстовых запросов от автоматического перевода до генерации полноценного связного текста или кода. Модели обучаются на многих

тысячах гигабайт текстовых данных, в которые входят книги, статьи, веб-сайты и прочие различные текстовые ресурсы. Объем информации, содержащейся в модели LLM, можно охарактеризовать как чрезвычайно высокий. Например, модель GPT-3 от OpenAI имеет 175 миллиардов параметров, а её предшественница GPT-2 - всего 1,5 миллиарда параметров.

Модель графической нейросети Kandinsky [] обучалась на платформе SberCloud ML Space в течение двух месяцев, потратив 20 352 GPU-V100 дней. Для её обучения использовался датасет из 60 млн пар изображений и текстовых описаний, который впоследствии был сокращён до 28 млн. В обучении были задействованы такие известные датасеты, как ConceptualCaptions [11] (датасет, содержащий более 3 миллионов изображений с подписями на естественном языке, необработанные описания для которых собираются из Интернета) и YFCC100m [12] (крупнейшая общедоступная и свободно используемая коллекция мультимедиа, содержащая метаданные около 99,2 миллионов фотографий), переведённые на русский язык. Первый этап обучения составил 250 тыс. итераций.

Однако возможности нейросетей не ограничиваются обработкой естественного языка и генерацией изображений. Для специализированных задач — таких, как производство или медицина, возникают потребности в специальных нейросетях, обученных на профессионально ориентированных датасетах. Такие нейросети должны быть способны воспринимать узкопрофильные жаргоны и научно-технические термины. Это необходимо, чтобы предотвратить возможные двусмысленные трактовки. Учитывая огромное количество накопленных данных и наличие специализированных архивов для множества областей, создание графических нейросетей с конкретной направленностью — лишь вопрос времени. Потенциальное их применение открывает широкие возможности для анализа и сопоставления различных типов данных, а также для их визуализации в доступной и наглядной форме.

Такие нейросети могут оказаться полезными и в учебно-методических пособиях. Например, нейросеть способна отразить типичное состояние органа или тканей при наличии определенного набора симптомов, упомянутых в запросе. Если в тексте описания указана какая-либо патология, визуализация может помочь выделить её особенности, что способствует принятию правильного диагностического решения.

В области производства огромные базы данных технических чертежей и стандартных формализованных и подробно описанных 3D моделей дают потенциальную возможность для обучения специализированных нейросетей, ориентированных на строгую формализацию запроса и ограниченную тематику. Такие специализированные нейросети могут быть использованы для генерации CAD-файлов с помощью текстовых подсказок, как показано на рисунке 1, для восстановления 3D модели по чертежам и для оценки и сопоставления деревьев 3D моделей с целью выявления идентичных моделей, несмотря на различия в их построении. Генерируемые модели можно импортировать в выбранную программу CAD или создавать специализированные генераторы Text-to-CAD без построения и поддержания инфраструктуры.

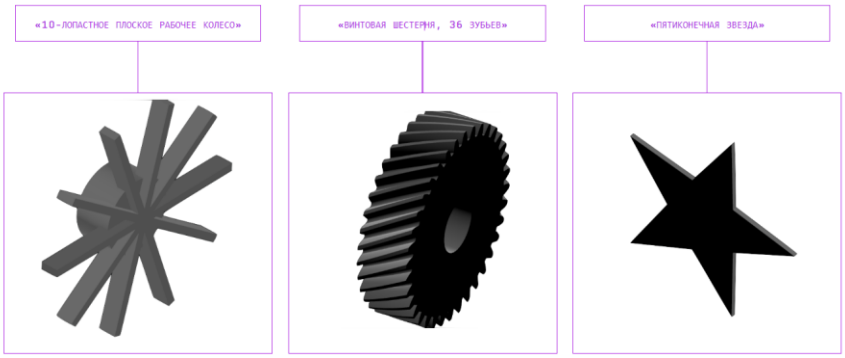


Рисунок 1 — Text-to-CAD, интерфейс для создания файлов CAD с помощью текстовых подсказок [13]

Кроме того, современные нейросети text-to-image, позволяющие создавать изображения или 3D модели по текстовому запросу, могут применяться в производственном процессе для получения предварительного визуального вида детали, который в дальнейшем может быть переработан в соответствии с видением конструктора. Пример такой нейросети на рисунке 2. Предварительный концепт, не требующий значительных затрат и предоставляемый в неограниченном количестве вариантов, может значительно сократить трудоемкость и расходы по созданию прототипов в процессе научно-исследовательских и опытно-конструкторских работ (НИР и НИОКР).

Рисунок 2 — Пример создания САПР по текстовым подсказкам

III. Информационное поле объекта

Каждый объект в цифровой среде обладает так называемым информационным полем. Информационное поле объекта определяется как весь объем неупорядоченной информации, связанной с искомым объектом и совокупность упоминаний в цифровой среде. Иными словами, это то количество открытой и публичной информации, которое окружает искомый объект и позволяет воссоздать его образ искусственно.

Информационное поле включает в себя все упоминания объекта в цифровой среде. Это могут быть:

1. Тексты: Статьи, отзывы, комментарии, посты в социальных сетях, блоги, научные работы.
2. Мультимедиа: Фотографии, видео, графики, аудиозаписи.
3. Структурированная информация: Базы данных, таблицы, анкеты, опросы.
4. Метаданные: время создания, авторство, местоположение и другие характеристики, которые помогают в идентификации и обработке информации об объекте.
5. Контекстуальные связи: связи между объектом и другими объектами, событиями, или влияющими факторами.

Плотность характеризует способность технологии, такой как искусственный интеллект, воссоздавать образ объекта на основе собранных данных. Чем больше данных доступно, тем точнее и полнее можно воссоздать цифровой образ объекта.

Плотность информационного поля должна напрямую коррелировать с такими факторами, как, к примеру, частота упоминаний объекта в различных медиа, разнообразие источников информации, глубина и детальность предоставленных данных.

Самый простой вид формулы, которой можно охарактеризовать плотность информационного поля представлен ниже:

$$П = \sum(x_1 + x_2 + \cdots + x_n),$$

где x_n — величина, характеризующая количество определенного типа данных. Из суммы этих величин складывается общее значение плотности.

Также каждое x_n может домножаться на определенные коэффициенты, характеризующие частоту упоминаний, разнообразие информации, её реальность и иные качественные характеристики представленных данных.

Таким образом, объекты, которые наиболее часто и разнообразно упоминаются в публичных источниках, будут иметь высокую плотность информационного поля, поскольку они часто оказываются в центре внимания СМИ и общественности. А вот объекты или люди, которые менее известны и реже упоминаются в публичных источниках, будут иметь плотность более низкую. Простейшей моделью визуализации, которая может в общих чертах отразить подобное представление, является облако тегов, представленное на рисунке 3. На нем отражено информационное распределение слов одного из разделов данной статьи: чем чаще слово упоминалось в тексте, тем больше оно по размеру.

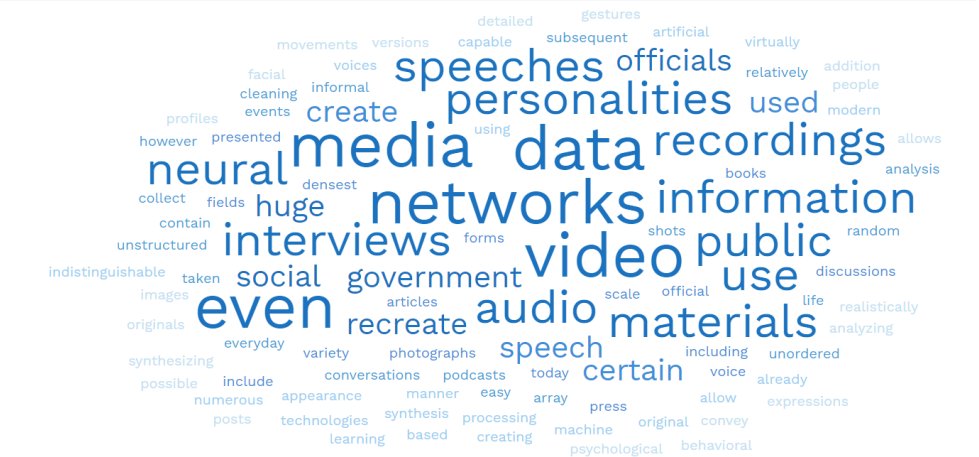


Рисунок 3 — Пример визуализации плотности информационного поля в самом простом воплощении (облако тегов, отражающее наиболее часто повторяющиеся в статье термины).

Самые плотные информационные поля на сегодняшний день имеют медиа-персоны и лица, представляющие государство. Информация о них представлена в огромных масштабах и самых разнообразных формах, что включает в себя видеоматериалы, записи голоса, фотографии, книги, статьи в прессе, обсуждения в социальных сетях и многое другое. Видеоматериалы включают как официальные выступления и интервью, так и случайные кадры, снятые на публичных мероприятиях или даже в повседневной жизни. Аудиозаписи могут содержать речи, интервью, подкасты и даже неформальные разговоры.

Это огромный массив неупорядоченной и неструктурированной информации, собрать которую, тем не менее, относительно легко. Последующие анализ, чистка данных и обработка позволяют использовать эти данные в целях воссоздания искусственного облика, манеры речи, видео и аудио материалов.

Современные технологии синтеза речи уже способны создавать практически неотличимые от оригинала голоса медиа-персон, основываясь на многочисленных аудиозаписях. С помощью методов машинного обучения и нейронных сетей можно также воссоздавать видеоверсию этих людей, синтезируя изображения, которые будут с максимальным приближением к реалистичности передавать мимику, жесты и движения оригиналов.

Кроме того, все эти данные позволяют создавать подробные психологические и поведенческие профили медиа-персон и государственных деятелей. Анализируя их публичные выступления, интервью и посты в социальных сетях, можно выявить их предпочтения, убеждения и мотивации. Например, даже использование определенных слов и фраз может дать представление о стиле общения человека, его эмоциональном состоянии и даже профессиональной компетенции. Эти данные могут использоваться не только для создания точных симуляций, но и для прогнозирования поведения этих личностей в тех или иных ситуациях.

Использование таких потенциальных возможностей нейронных сетей может быть разным. В статье [14] рассматриваются примеры известных случаев вредоносного использования нейросетей, от мошенничества до манипулирования общественным мнением.

Сейчас в большинстве популярных нейросетей, особенно тех, которые предоставляют свои услуги на платной основе, разработчики предпринимают широкий ряд специальных мер, направленных на уменьшение потенциального вредоносного влияния и обеспечение безопасности общества. Разработчики вводят всё больше ограничений, касающихся использования образов известных государственных и медийных лиц в изображениях, сгенерированных нейросетями. Однако поскольку остается возможность локального пользовательского обучения отдельных моделей нейросетей, то можно предположить, что проблема будет оставаться актуальной на протяжении продолжительного времени.

Вместе с тем набирает силу и популярность тенденция создания цифровых двойников реальных людей, как живых, так и давно умерших. Идея «цифрового пантеона» не нова, но теперь, имея возможность обучить нейросетевую модель на кадрах хроники, выдержках из личной переписки и собраниях сочинений известных исторических деятелей, возникает риск возникновения новых приемов манипуляции общественным мнением и засорения информационного пространства и образовательной системы ложной или искусственно сгенерированной информацией.

Подобное явление можно назвать созданием так называемой цифровой псевдоличности, которая сможет с определенной мерой достоверности воспроизводить речь и образ мышления определенного человека. В настоящее время уже существует услуга создания такой псевдоличности для публичных людей или ведущих бизнес [15]. Для этого потребуется загрузить в базу данных свой голос, цифровые образцы внешности (фото и

видео), образцы личной и деловой переписки, а также примеры текстов в различных стилях. Утверждается, что такая цифровая псевдоличность сможет имитировать стиль общения своего оригинала и вести переговоры от его имени (например, с клиентами).

В этой идее существует ряд рискованных моментов:

1) Необходима строгая конфиденциальность личных данных. Если разработчик позволит им «уплыть» в свободный интернет, то проконтролировать дальнейшее развитие своих потенциальных двойников пользователь уже не сможет.

2) Фейковые факты и высказывания. Пользователь может столкнуться с тем, что в его биографии всплывают ему самому неизвестные факты, которые на самом деле являются порождением генерации цифровой модели. Также ему могут приписать слова, которые он не говорил.

3) Манипуляции и мошенничество. Сбор и хранение биометрических и поведенческих данных в частных коммерческих организациях несут в себе риск, что данные окажутся в руках злоумышленников. В случае использования цифровой копии теми, кто не является оригиналом, легко распространить дезинформацию или совершать действия от имени другого человека.

4) Юридическая ответственность. Неясность в правовом регулировании создания и использования цифровых копий может создать правовые вакуумы, которые будут использоваться для незаконных целей, так как непонятно, кто будет нести ответственность за слова и действия цифровой копии реального человека.

5) Этические аспекты. В августе 2024 года впервые распространилось видео от якобы «мертвого» человека, который продолжил существовать в цифровом виде и вести блог. Видео было признано фейком, а оригинальный человек оказался жив и объявил произошедшее арт-перфомансом [16], тем не менее, в обществе это создало информационный прецедент использования лица и личных данных умершего человека для воспроизведения его цифровой псевдоличности и дальнейших манипуляций от его имени. Злоупотребление подобными действиями приведет к нарушениям прав на неприкосновенность частной жизни и личных данных.

При использовании личности известных исторических деятелей, обладающих плотным информационным полем, позволяющим создать цифровую псевдоличность, может возникнуть следующий ряд проблем:

1) Искажение исторической правды, неточности и фальсификации. Станет сложно отделить сгенерированные высказывания (особенно если они становятся крылатыми) от реальных, а также будет затруднительно установить подлинность того или иного высказывания, благодаря чему возрастет риск манипуляции общественным мнением в исторических и политических диспутах. Использование таких копий для интерпретации исторических событий может внести значительные изменения в восприятие и понимание истории, что не всегда будет соответствовать действительности.

2) Политическая и социальная манипуляция. Цифровые копии исторических деятелей могут быть использованы для пропаганды и политической манипуляции, а некорректное представление исторических личностей (особенно если они были спорными фигурами в своей исторической эпохе) может вызвать социальную напряженность.

3) Образовательные риски. У будущих поколений существует риск столкнуться с явлением мнимой достоверности, когда цифровая копия будет восприниматься как достоверное представление личности исторического деятеля, что в свою очередь приведет к поверхностному восприятию и массовым искажениям реальных фактов и культурного контекста той или иной эпохи.

Между тем, по мере наполнения интернета информацией, сгенерированной нейросетями, возникают новые проблемы и потенциальные риски от последствий такого обучения. Обучение на данных, созданных самим ИИ, может приводить к накоплению ошибок и избыточным повторениям. Это сопровождается риском появления артефактов, которые трудно заметить и исправить. Статья [17] рассматривает процесс того, как использование сгенерированного моделью контента в обучении вызывает необратимые дефекты в полученных моделях, где хвосты исходного распределения контента исчезают. Авторы называют этот эффект коллапсом модели и демонстрируют, что он может возникать в вариационных автоэнкодерах, моделях гауссовой смеси и LLM.

Специалисты всего мира высказывают разные мнения [18] на предмет будущего нейросетей и тех мер, которые можно предпринять для минимизации будущего коллапса, однако многие сходятся в том, что чистота и релевантность данных будут играть ключевое значение при обучении будущих поколений нейросетей.

IV. Роль личных данных в эпоху генеративных нейросетей

В эпоху расцвета генеративных нейронных сетей особую роль приобретает та личная информация о человеке, которую чаще всего он разглашает о себе сам. Локальные модели вполне реально обучить даже на том небольшом информационном поле, которое складывается из социальных сетей и иных источников.

К числу таких данных можно отнести:

- фотографии в социальных сетях;

- видео;
- аудиозапись голоса — из видео или телефонных разговоров
- личные записи и блоги
- дневники
- образцы подписи

В наше время люди постоянно становятся объектами съемки вне зависимости от собственного желания — от уличных видеокамер, до случайного попадания на чужие видео, выложенные в интернет. Персонализировано собирать такую информацию непросто, это требует очень большого количества ресурсов, но, тем не менее, подобные материалы могут стать обучающим материалом для новых моделей нейросетей. На рисунке 4 показана аннотированная фотография из обучающего набора данных Diversity in Faces от IBM, подготовленная для обучения процессу распознавания лиц.

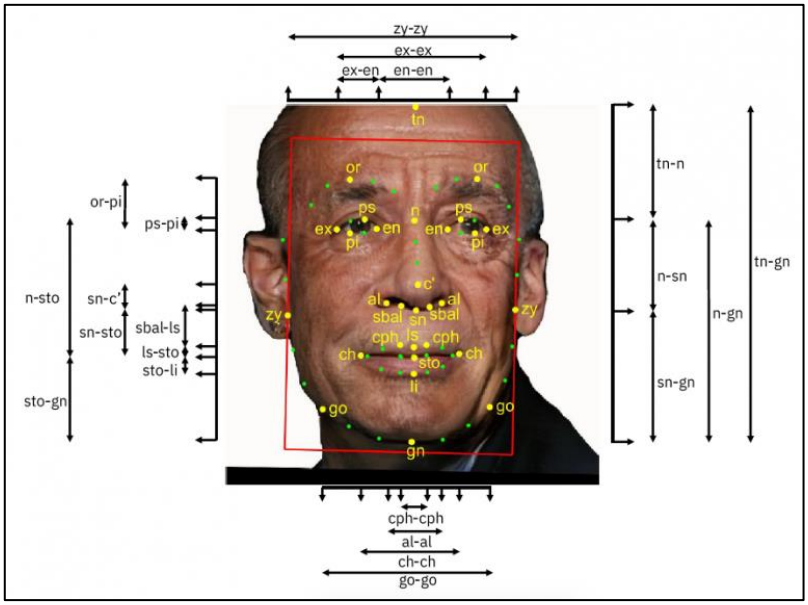


Рисунок 4 — Аннотированная фотография из набора данных Diversity in Faces от IBM

Поскольку алгоритмы распознавания лиц требуют огромного количества изображений, то методы ручного сбора фотографий никак не могут покрыть их потребности, так что исследователи начали массово собирать изображения с сайтов, таких как Flickr, Facebook, YouTube и других.

Статья [19] в подробностях раскрывает проблематику несанкционированного извлечения миллионов фотографий из Интернета для обучения соответствующих алгоритмов распознавания лиц. В январе 2019 г. IBM выпустила коллекцию почти миллиона фото с платформы Flickr, упомянув о намерении сократить объективные ошибки с помощью многообразного обучающего датасета, что в свою очередь вызвало серьезный резонанс в среде фотографов, чьи работы попали в эту коллекцию без уведомления как автора, так и модели. В частности были высказаны опасения, что собранные данные могут быть использованы для ограничения основных прав и конфиденциальности, а также репрессивной и дискриминационной политики.

Такие случайные фото также становятся элементами информационного поля, связанного с теми или иными объектами, к примеру, с людьми, чьи лица можно определить на фотографии и далее установить их личность более детально.

Учесть такие данные в личном контроле невозможно, в отличие от той информации, которую человек самостоятельно размещает в открытом доступе. В частности люди, которые плотно работают с публикой удаленно с участием камер и микрофонов (даже как любители) фактически предоставляют в открытый доступ записи своего голоса, своей мимики, используемого лексикона и многое другое, что может быть использовано в различных целях. К примеру, телефонными мошенниками для вредоносных махинаций, которые способны «украсть голос» даже при помощи телефонного разговора.

Пользователями уже разрабатываются идеи защиты от подобных краж, к примеру, разработка голосовых аватаров для защиты абонента при звонках от неизвестных ему лиц [19].

Что касается крупных коммерческих нейросетевых систем по обработке изображений, которые сотрудничают с государственными и силовыми структурами всё тоже не так однозначно. В новаторском исследовании 2018 года, оказавшем значительное влияние на исследования ИИ, Джой Буоламвини и Тимнит Гебру [20] первыми

выявили, что все популярные системы распознавания лиц с наибольшей точностью идентифицируют мужчин со светлой кожей (2,4% ошибок) и наиболее часто ошибаются при распознавании женщин с более темной кожей (61% ошибок). Вероятными причинами данного феномена являются следующие: меньшее количество фотографий женщин с темной кожей в базах данных, преимущественно белый мужской состав разработчиков таких систем, а также худшая работа датчиков камер при распознавании деталей в темных оттенках. Это усугубляется тем фактом, что некоторые коммерческие компании подходят к разработке нейросетевых алгоритмов с позиции «черного ящика», когда получают результаты, и сравнивают их с теми, которые хотели бы получить, не исследуя суть внутренних процессов.

Несмотря на выявляемые проблемы, данные системы продолжают активно применяться в различных областях, включая правоохранительные органы России и Китая. Исследования подтверждают, что представители расовых меньшинств в этих государствах подвергаются большему риску быть ошибочно идентифицированными как преступники. Такая тенденция обусловлена тем, что алгоритмы системы чаще устанавливают совпадение с лицами, чьи черты схожи с чертами лица подозреваемого. Известный пример подобного ложного совпадения произошел в 2023 году, когда ученый-гидролог был арестован по делу об убийствах 20-летней давности на основе данных искусственного интеллекта (ИИ). По данным новостных статей [21], программа ИИ определила, что фото задержанного на 55% совпадало с изображением подозреваемого в убийствах, совершенных в 2003 году. Дело было прекращено лишь спустя год.

В связи с этим, становится очевидной необходимость регуляций в области разработки искусственного интеллекта (ИИ). Регуляции должны заставлять разработчиков ИИ следовать общим стандартам, чтобы те не экономили на безопасности. Хотя регуляции сами по себе не создают технических решений, они всё же могут дать мощный стимул к их разработке и внедрению. Компании будут более интенсивно вырабатывать меры безопасности, если без них нельзя будет продавать свои продукты, особенно если другие компании подчинены тем же стандартам. Какие-то компании, может, и регулировали бы себя сами, но государственная регуляция помогает предотвратить то, что менее аккуратные конкуренты на безопасности сэкономят. Регуляции должны быть проактивными, а не реактивными. Часто говорят, что в авиации регуляции “написаны кровью” — но тут их надо разработать до катастрофы, а не после. Они должны быть устроены так, чтобы давать конкурентное преимущество компаниям с лучшими стандартами безопасности, а не компаниям с большими ресурсами и лучшими адвокатами. Регуляторов надо набирать независимо, не из одного источника экспертов (например, больших компаний), чтобы они могли сосредоточиться на своей миссии для общего блага без внешнего влияния [22, 23].

Для повышения прозрачности и подотчетности ИИ-систем необходимо требовать от компаний документации данных, объясняющей, какие источники данных используются при обучении и развёртывании своих моделей. Принятые компаниями решения использовать датасеты, в которых есть персональные данные или агрессивный контент, повышают и без того бешеный темп разработки ИИ и мешают подотчётности. Документация должна описывать мотивацию выбора, устройство, процесс сбора, назначение и поддержку каждого датасета. Общественный контроль за ИИ-системами общего назначения также становится необходимым, учитывая риски, которые частные компании никогда в должной мере не учтут. Чтобы удостовериться, что они адекватно принимаются во внимание, может потребоваться прямой общественный контроль за такими системами.

Идеальный сценарий предполагал бы, что ИИ разрабатываются, тестируются, а потом развёртываются только тогда, когда все их катастрофические риски пренебрежимо малы и находятся под контролем. Прежде чем начать работу над новым поколением ИИ-систем, должны проходить годы тестирования, мониторинга и внедрения в общество предыдущего поколения.

V. Заключение

Быстрое развитие диффузионных моделей позволило достичь больших результатов в различных областях, от медиа до медицины и производства. Способность этих моделей к генерации реалистичных изображений и текстов открывает новые возможности, но также порождает немало проблем. Одним из основных рассматриваемых моментов является критическая роль данных и их качества в процессе обучения моделей, что напрямую влияет на их работоспособность и точность. Введено и обосновано понятие «информационное поле». Отмечена необходимость уделять внимание вопросам конфиденциальности и безопасности данных для обучения, так как с увеличением влияния нейросетей, возрастают и риски, требующие решения проблем, связанных с возможным коллапсом моделей и манипуляцией общественным мнением через цифровые копии и псевдоличности. Важно продолжать исследовать и развивать нейросетевые технологии, внедряя при этом необходимые меры для их безопасного использования.

Благодарности

Вычислительная работа проведена с помощью гибридного суперкомпьютера К-100, установленного в Центре коллективного пользования ИПМ им. М.В. Келдыша РАН.

Список источников

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Networks, 2014, https://doi.org/10.48550/arXiv.1406.2661

[2] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, 2020, https://doi.org/10.48550/arXiv.2006.11239 (дата обращения 29.03.2023)

[3] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Zhu, S. Ermon, SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, 2022, https://doi.org/10.48550/arXiv.2108.01073

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, 2021, https://doi.org/10.48550/arXiv.2112.10752

[5] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., 2017. Image-toimage translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

[6] Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y., 2021. Text-toimage generation grounded by fine-grained user attention. In Proceedings of the IEEE/CVF Winter Con-ference on Applications of Computer Vision, pp. 237–246.

[7] Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M, Sutskever I., 2021. Zero-Shot Text-to-Image Generation, https://doi.org/10.48550/arXiv.2102.12092Д

[8] Radford A., Jong W.K., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020 [cs.CV]. https://doi.org/10.48550/arXiv.2103.

[9]

[10] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. 2017. Attention Is All You Need, arXiv preprint arXiv:1706.03762 https://doi.org/10.48550/arXiv.1706.03762

[11] Conceptual Captions Dataset. URL: https://github.com/google-research-datasets/conceptual-captions (дата обращения 27.08.2024)

[12] YFCC100M, URL: https://paperswithcode.com/dataset/yfcc100m (дата обращения 27.08.2024)

[13] Generate CAD from text prompts, URL: https://zoo.dev/text-to-cad (дата обращения 27.08.2024)

[14] Бондарева Н.А.. Графические нейронные сети и проблемы верификации изображений // Труды 33-ой Международной конференции по компьютерной графике и машинному зрению ГрафиКон 2023, Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия, 19-21 сентября 2023 г., с.317-327, DOI: 10.20948/graphicon-2023-317-327 https://www.graphicon.ru/html/2023/papers/paper_031.pdf

[15] Ломакина Я. «Первый в мире мертвый блогер» оказалась живой актрисой: что это вообще было, 2024, URL: https://journal.tinkoff.ru/dead-blogger/ (дата обращения 27.08.2024)

[16] Clone yourself, 2024, URL: https://www.delphi.ai/ (дата обращения 27.08.2024)

[17] Shumailov I., Shumaylov Z., Zhao Y., Gal Y, Papernot N., Anderson R., 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv preprint arXiv:2305.17493 https://doi.org/10.48550/arXiv.2305.17493

[18] Рожков Р. Градация деградации: ожидает ли генеративный искусственный интеллект вырождение, 2023, URL: https://www.forbes.ru/tekhnologii/491359-gradacia-degradacii-ozidaet-li-generativnyj-iskusstvennyj-intellekt-vyrozdenie (дата обращения 27.08.2024)

[19] Solon O. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent, 2019, URL: https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921 (дата обращения 27.08.2024)

[20] Сообщение из телеграм канала, 2023, https://t.me/sburyi/182 (дата обращения 26.08.2024)

[21] Buolamwini J., Gebru T., Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification // Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018

[22] ТАСС: Суд Москвы освободил ученого, арестованного с помощью ИИ по делу 20-летней давности, 2023, URL: https://tass.ru/proisshestviya/19508893 (дата обращения 27.08.2024)

[23] Харитонова Ю.С., Савина В.С., Паньини Ф. Предвзятость алгоритмов искусственного интеллекта: вопросы этики и права // Вестник Пермского университета. Юридические науки. 2021. №53. URL: https://cyberleninka.ru/article/n/predvzyatost-algoritmov-iskusstvennogo-intellekta-voprosy-etiki-i-prava (дата обращения: 05.09.2024).

[24] Hendrycks D., Mazeika M., Woodside T. An Overview of Catastrophic AI Risks arXiv preprint arXiv:2306.12001 https://doi.org/10.48550/arXiv.2306.12001

УДК 535.8
DOI: 10.25206/978-5-8149-3873-2-2024-292-299

Отображение точек многомерного информационного текстового массива на упругую карту для оценки взаиморасположения областей смысловой близости
Displaying points of a multidimensional information text array on an elastic map to assess the relative position of semantic proximity areas

А.Е.Бондарев[1]
[1]ИПМ им М.В. Келдыша РАН, Миусская пл., 4, 125047, Москва, Россия.

A.E. Bondarev[1]
[1] Keldysh Institute of Applied mathematics of RAS, Miusskaya sq. 4, Moscow, 125047, Russia

Аннотация. Представлены результаты вычислительных экспериментов по отображению точек исходного многомерного информационного массива на развертку упругой карты для оценки взаиморасположения областей смысловой близости с целью совершенствования обработки текстовой информации. Упругие карты рассматриваются в роли инструмента обеспечения аналитической работы с текстовой информацией. Как показывают предыдущие работы, для получения нужных расстояний, соответствующих кластерной картине изучаемого многомерного объема, необходимо использовать расстояния на упругой карте, которая и отражает кластерный портрет исследуемого многомерного объема данных. В работе представлены полученные таким способом кластерные структуры точек изучаемого многомерного объема на развертке упругой карты в плоскости двух первых главных компонент. Представлен анализ взаиморасположения кластеров различных конфигураций на разные моменты времени.

Ключевые слова: Многомерные текстовые данные, кластерная структура, упругие карты, анализ взаиморасположения кластеров

Abstract. The article presents the results of computational experiments on displaying the points of the original multidimensional information array on the elastic map scan to assess the relative positions of semantic proximity areas in order to improve the processing of text information. Elastic maps are considered as a tool for providing analytical work with text information. As previous works show, in order to obtain the required distances corresponding to the cluster picture of the studied multidimensional volume, it is necessary to use the distances on the elastic map, which reflects the cluster portrait of the studied multidimensional data volume. The paper presents the cluster structures of points of the studied multidimensional volume obtained in this way on the elastic map scan in the plane of the first two principal components. An analysis of the relative positions of clusters of different configurations at different points in time is presented.

Keywords: Multidimensional text data, cluster structure, elastic maps, cluster position analysis

Введение

Данная работа представляет результаты проведения вычислительных экспериментов по кластерному анализу для многомерных информационных массивов. Многомерные информационные массивы представляют собой текстовую информацию, являясь в цифровом виде частотами совместного употребления слов из различных частей речи (например, существительное + прилагательное). Для получения такого многомерного массива проводится анализ текстовых коллекций. В результате, для 300-мерного массива, построенного таким образом, мы рассматриваем 300 точек в пространстве из 300 измерений. Данная работа является продолжением работ [19-26] и во многом использует алгоритмы, результаты и наработки из предыдущих работ.

На современном этапе чрезвычайно актуальной задачей становится изучение и анализ многомерных объемов данных. Анализ многомерных данных является актуальной проблемой достаточно давно. Для подобного изучения разрабатывались методы анализа данных, методы визуальной аналитики. Применение этих методов дает возможность на практике узнать структуру исследуемого объема многомерных данных, увидеть его кластерную картину, определить области сгущения данных (кластеры), и т.д.

Изучение многомерных данных сформировало новую междисциплинарную область исследований, как визуальная аналитика (Visual Analytics). Визуальная аналитика – это комплекс методов и подходов, направленных на проведение анализа и визуального представления многомерных данных независимо от природы их происхождения. Основные концепции визуальной аналитики представлены в первых работах по данной дисциплине [1-4]. Подходы и методы визуальной аналитики и позволяют решать ряд практических задач исследования многомерных данных. К таким задачам можно отнести задачи классификации данных, обнаружения кластеров, выявление ключевых определяющих параметров, установление взаимосвязей между ключевыми параметрами и т.д.

Для отображения многомерного объема данных во вложенные в исходный объем многообразия меньшей размерности были разработаны подходы самоорганизующихся карт [5-7]. К подобным картам относятся и так называемые упругие карты (Elastic Maps). Теория упругих карт была разработана А. Горбанем и А. Зиновьевым вместе с коллегами и достаточно подробно описана в работах [8-14]. Упругие карты нашли довольно широкое применение в практическом анализе многомерных данных, причем в самых разных областях – анализ экономических данных, анализ полезных ископаемых, анализ медицинских данных и т.п. Примеры исследований представлены в работах [15-18]. Важнейшим свойством упругих карт является способность успешно использовать их для любых многомерных данных независимо от природы их происхождения. Особую эффективность в решении задач кластерного анализа многомерных данных упругие карты приобретают при совместном использовании с методом главных компонент (PCA). Отображение упругой карты и ее развертки в пространстве, образованном первыми тремя главными компонентами, в задачах кластеризации и классификации позволяет более точно и четко определять кластерную структуру изучаемых многомерных объемов данных. На рисунке 1 приведен пример развертки упругой карты на плоскость, расположенную в пространстве главных компонент с раскраской по плотности данных. Подобное представление позволяет определить кластерную структуру изучаемого объема многомерных данных без применения специальных алгоритмов кластеризации.

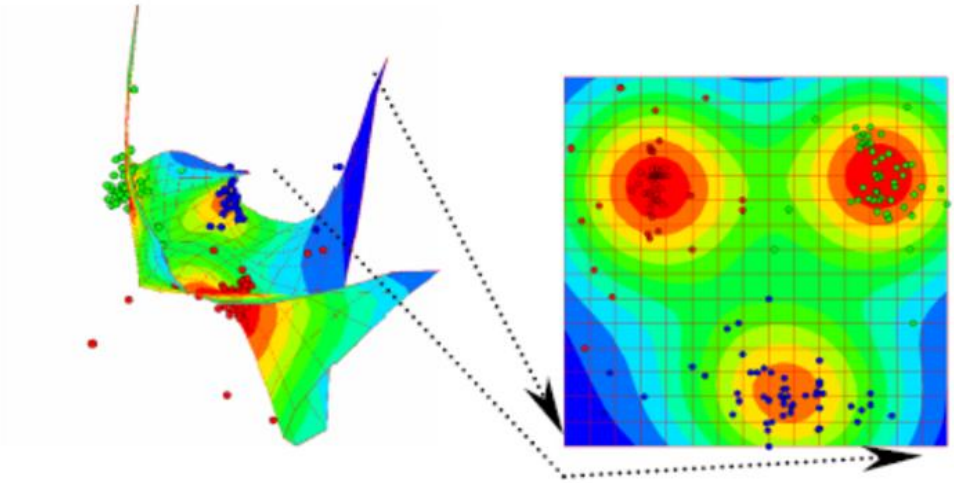


Рис.1. Пример упругой карты и ее развертки на плоскость с раскраской по плотности данных.

Важное значение имеет применение построения упругих карт к задачам анализа многомерных текстовых данных. Это направление появилось из необходимости решения практических задач кластерного анализа и становится все более актуальным. Применение построения упругих карт к задачам анализа многомерных текстовых данных впервые было реализовано на практике А.Е. Бондаревым и В.А. Галактионовым в ИПМ им. М.В. Келдыша РАН. В качестве изучаемых объемов многомерных данных использовались массивы частот совместного употребления различных частей речи, получаемых из текстовых коллекций. Результаты исследований подробно представлены в работах [19-26]. Однако в процессе решения задач кластерного анализа многомерных информационных объемов выяснилось одно крайне важное обстоятельство. Дело в том, что получаемая в численных экспериментах картина распределения кластеров в многомерном объеме текстовых данных существует лишь на развертке упругих карт в пространстве первых двух главных компонент. Определение конкретных внутрикластерных и межкластерных расстояний должно вестись на цифровых данных. Таковыми данными являются координаты точек многомерного объема в многомерном пространстве. Однако, как показывают работы [24-26], определение внутрикластерных и межкластерных расстояний на основе исходных координат точек многомерного пространства не дает надежных результатов. Для получения нужных расстояний

необходимо использовать расстояния на упругой карте, которая и отражает кластерный портрет исследуемого многомерного объема данных. Именно этому направлению исследований посвящена данная работа.

Упругие карты и применение метода главных компонент (PCA) в задачах визуального анализа многомерных текстовых информационных массивов

Упругие карты являются логическим развитием карт Кохонена. Идеология и алгоритмы реализации построения упругих карт подробно представлены в работах [8 - 14]. Подобная карта представляет собой систему упругих пружин, вложенную в многомерное пространство данных. Метод упругих карт формулируется как оптимизационная задача, предполагающая оптимизацию заданного функционала от взаимного расположения карты и данных.

Согласно [8 - 14] основой для построения упругой карты является двумерная прямоугольная сетка G, вложенная в многомерное пространство, которая аппроксимирует данные и обладает регулируемыми свойствами упругости по отношению к растяжению и изгибу. Расположение узлов сетки ищется в результате решения оптимизационной задачи на нахождение минимума функционала:

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{m} + \mu \frac{D_3}{m} \to min,$$

где $|X|$ - число точек в многомерном объеме данных X; m - число узлов сетки, λ, μ - коэффициенты упругости, отвечающие за растяжение и изогнутость сетки соответственно; D_1, D_2, D_3 - слагаемые, отвечающие за свойства сетки.

$$D_1 = \sum_{ij} \sum_{x \in K_{ij}} \left|\left|x - r^{ij}\right|\right|^2.$$

D_1 является мерой близости расположения узлов сетки к данным.

Здесь Kij - подмножества точек из X, для которых узел сетки r_{ij} является ближайшим:

$$x \overset{\text{П}}{\to} r^{ij}, \qquad \left|\left|x - r^{ij}\right|\right|^2 \to \min, \qquad K_{ij} = \left\{x \in X, x \overset{\text{П}}{\to} r^{ij}\right\},$$

Слагаемое D_2 представляет меру растянутости сетки:

$$D_2 = \sum_{ij} \left\|r^{ij} - r^{i,j+1}\right\|^2 + \sum_{ij} \left\|r^{ij} - r^{i+1,j}\right\|^2.$$

Слагаемое D_3 представляет меру изогнутости (кривизны) сетки:

$$D_3 = \sum_{ij} \left\|2r^{ij} - r^{i,j-1} - r^{i,j+1}\right\|^2 + + \sum_{ij} \left\|2r^{ij} - r^{i-1,j} - r^{i+1,j}\right\|^2.$$

Варьирование параметров упругости заключается в построении упругих карт с последовательным уменьшением коэффициентов упругости, в силу чего карта становится более мягкой и гибкой, наиболее оптимальным образом подстраиваясь к точкам исходного многомерного объема данных. После построения упругую карту можно развернуть в плоскость для наблюдения кластерной структуры в изучаемом объеме данных. На развернутой плоскости можно раскраской отобразить распределение плотности данных по упругой карте. В ряде случаев подобная раскраска может оказаться весьма полезной. Особо эффективными упругие карты являются при совместном использовании с методом главных компонент (PCA). Отображение упругой карты и ее развертки в пространстве, образованном первыми тремя главными компонентами, позволяет резко улучшить результаты, особенно в задачах кластеризации и классификации. Применение упругих карт позволяет более точно и четко определять кластерную структуру изучаемых многомерных объемов данных.

Описанный подход был многократно применен к многомерным текстовым массивам данных. Рассматривались многомерные данные, получаемые из текстовых коллекций, и представляющие частоты совместного употребления различных частей речи. Например, 300 существительных и 300 прилагательных. Прилагательные рассматривались, как измерения, а существительные – как точки в пространстве измерений. То есть, рассматривалось 300 точек в 300-мерном пространстве. На рисунке 2 представлен типичный вид упругой карты.

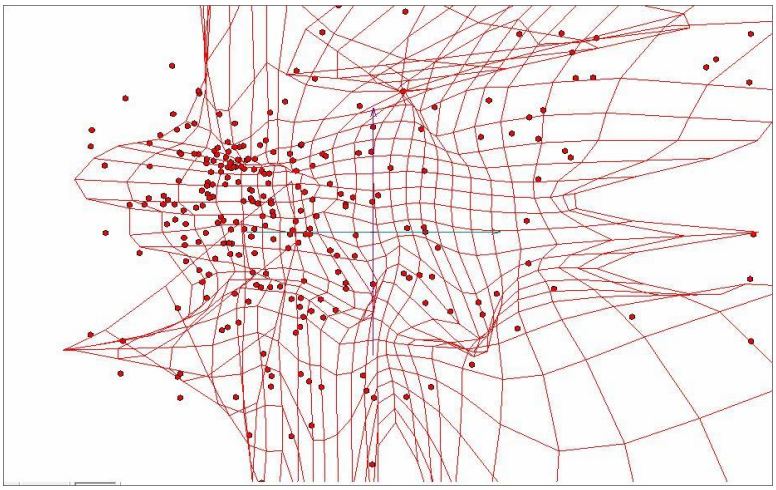


Рис.2. Типичный вид упругой карты частот совместного употребления.

Построение исходного набора данных и проведение тестовых расчетов взаиморасположения кластеров

Ранее в предыдущих работах [20-26] было показано, что подход оценки межкластерных расстояний с помощью построения гиперсфер в пространстве исходных данных не обеспечивает желаемого результата. Все численные исследования проводились с использованием двух типов метрик, задающих расстояние между объектами – манхэттенской метрики и косинусной метрики. Для оценки межкластерных расстояний необходимо построение других способов оценки. По результатам работ [20-26] был сделан вывод о том, что проблемы с изучением кластерных свойств связаны с тем, что исходная качественная информация о кластерах берется с развертки упругой карты, а для расчетов используются исходные координаты точек в многомерном пространстве. Было принято решение в дальнейшем использовать координаты аннотированных точек на развертке упругой карты в силу того, что они непосредственно отражают кластерную картину изучаемого объема данных.

Реальные данные были получены из текстовых корпусов новостной информации. Для получения информации из текстовых корпусов использованы процедуры, подробно описанные в [20]. Для получения необходимых многомерных массивов использовались группы сочетаний «существительное + прилагательное». Выбирались 300 существительных и 300 прилагательных. В качестве цифрового значения использовались частоты совместного употребления существительного и прилагательного в каждом изучаемом многомерном массиве. Таким образом, мы как и в прошлых экспериментах, получили возможность рассматривать прилагательные как координатные измерения в многомерном пространстве, а существительные – как точки в этом пространстве. Рассматривались 300 точек, находящихся в 300-мерном пространстве. Таким образом построено 5 многомерных массивов – за март 2005 года, за апрель 2005 года, за май 2005 года, за май 2006 года и за май 2007 года. Первые три массива позволяют прослеживать эволюцию кластерной структуры многомерного объема через месяц. Третий четвертый и пятый массивы позволяют прослеживать эволюцию кластерной структуры многомерного объема через год. Для всех массивов были построены упругие карты и их развертки в пространстве первых главных компонент.

Результаты вычислительных экспериментов

Вычислительные эксперименты включали в себя построение упругих карт и их разверток на различные моменты времени, выделение кластеров, нахождение их центров и радиусов. Напомним, что в данной задаче использовалась евклидова метрика для определения расстояний, центр каждого конкретного кластера определялся как среднее арифметическое входящих в данный кластер точек. В качестве характерного размера кластера выбиралось расстояние от центра кластера до наиболее удаленной точки кластера.

Начнем рассмотрение результатов вычислительных экспериментов с развертки упругой карты на момент времени 2005-4 (апрель 2005). Здесь в многомерном массиве данных на правом нижнем краю развертки упругой карты имеются два четко выраженных кластера, отображенных на рис.3, где представлен крупным планом фрагмент развертки, содержащий данные кластеры вместе с их вычисленными центрами.

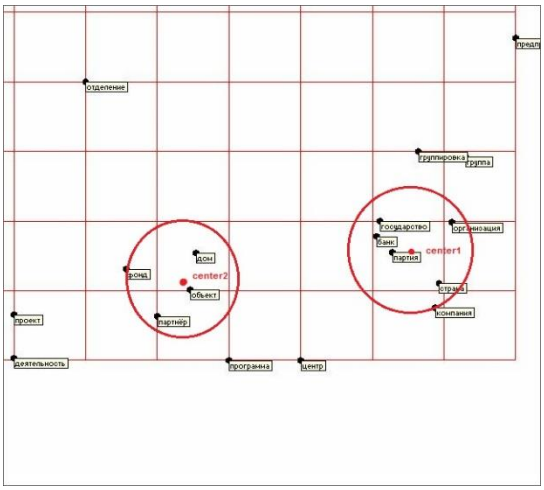


Рис. 3. Кластеры нижнего правого края для изучаемого многомерного массива на момент времени 2005-4 с центрами кластеров.

На рисунке 4 представлен фрагмент развертки упругой карты с аннотациями, соответствующая моменту времени 2005-5 (май 2005) с двумя кластерами по нижнему правому краю и центрами кластеров.

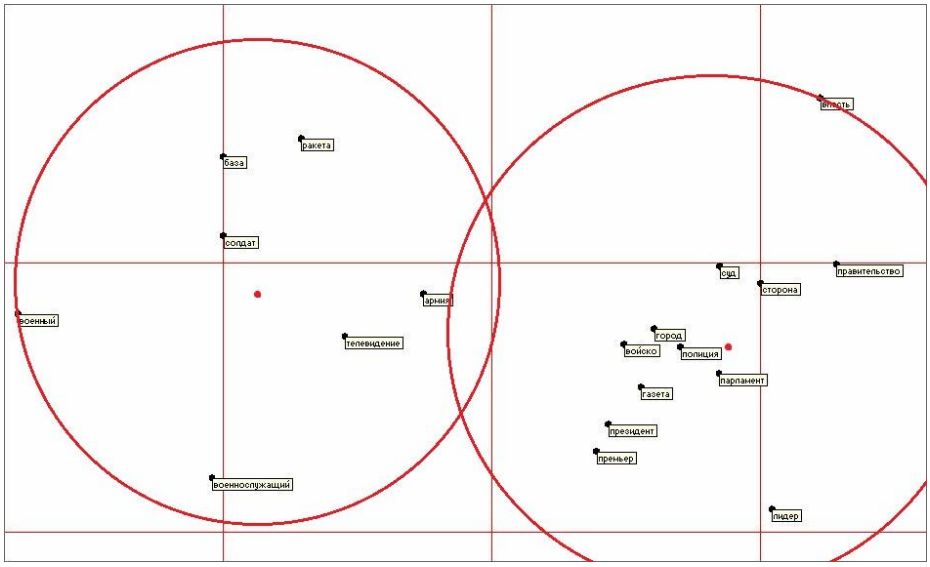


Рис. 4. Кластеры верхнего правого края для изучаемого многомерного массива на момент времени 2005-05 с центрами кластеров и радиусами.

Наиболее удаленными точками после вычисления расстояний от центра кластера являются точки «ВЛАСТЬ» и «ВОЕННЫЙ». Согласно полученным результатам в данном случае образуется маленькая область пересечения кластеров, но ни одна точка изучаемого массива многомерных данных в эту точку не попадает.

Теперь рассмотрим результаты для момента времени 2006-05. На рисунке 5 представлена развертка упругой карты с аннотациями, соответствующая моменту времени 2006-5 (май 2006).

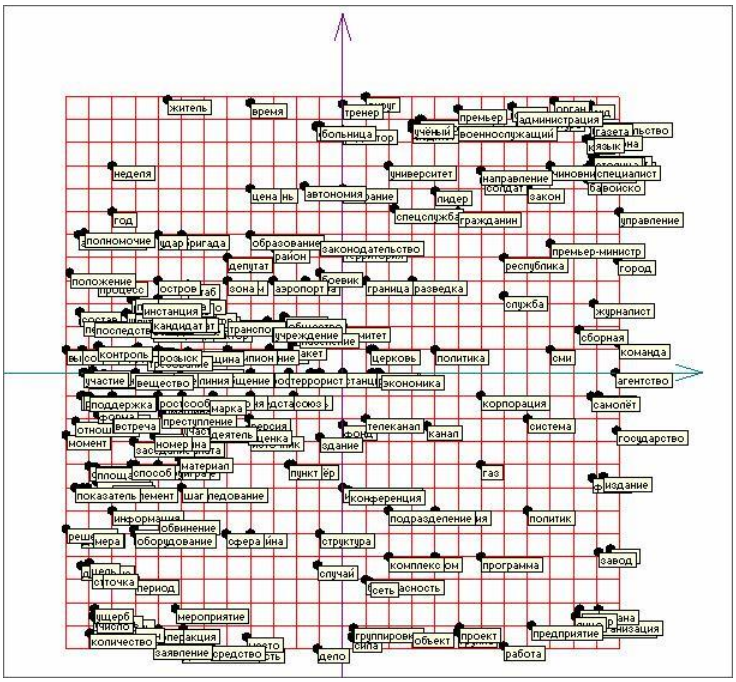


Рис.5. Аннотированная развертка упругой карты на момент времени 2006-05.

Здесь в многомерном массиве данных на правом краю в верхней и нижней частях развертки упругой карты имеются по два четко выраженных кластера. Кластеры верхней части представлены на рисунке 6.

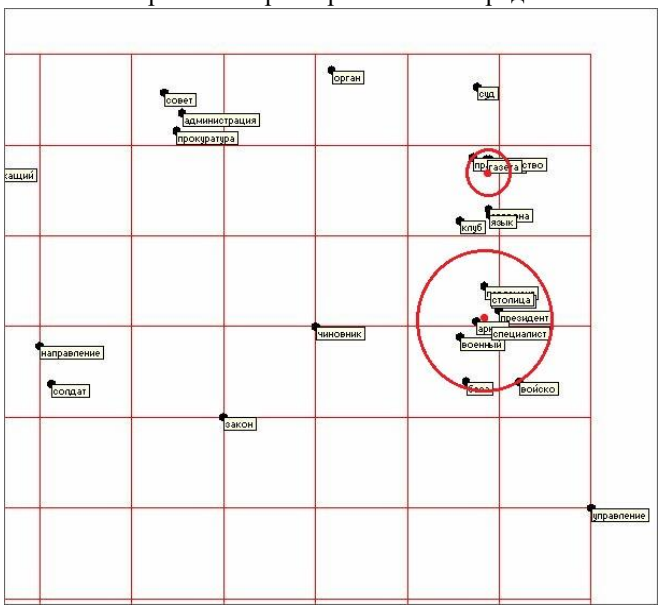


Рис. 6. Кластеры верхнего правого края для изучаемого многомерного массива на момент времени 2006-05 с центрами кластеров.

На рисунке 7 представлены кластеры нижнего правого края на момент времени 2006-5 (май 2006).

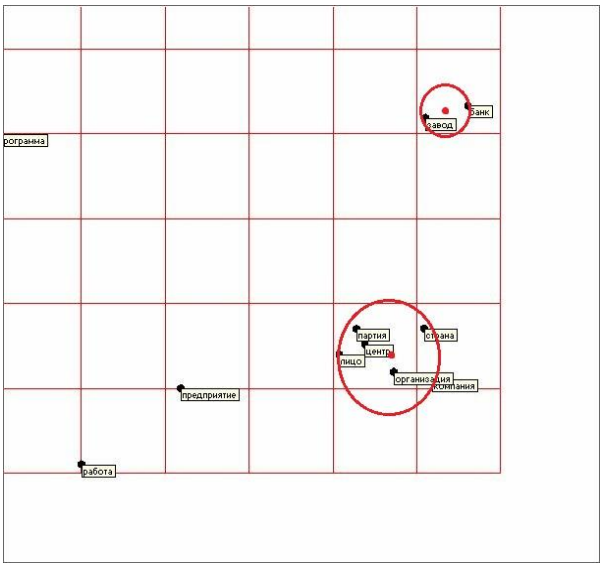


Рис. 7. Кластеры нижнего правого края для изучаемого многомерного массива на момент времени 2006-05 с центрами кластеров.

Приведенные результаты показывают, что в рассмотренных случаях пересечения кластеров и возникновения «спорных» точек нет.

Заключение

Упругие карты, будучи представленными в развертке на плоскость, образованную двумя первыми главными компонентами, позволяют увидеть кластерную картину изучаемого многомерного объема. В качестве многомерного объема используется массив частот совместного употребления из различных частей речи (например, «прилагательное + существительное»), получаемый из текстовых коллекций. Для анализа кластерной структуры используются координаты точек многомерного объема, получаемые с развертки упругой карты. Рассмотрены взаимные расположения кластеров, возникающие на различные моменты времени. Кластеры представлены в виде окружностей. Для построения радиусов использовался принцип выбора расстояния от центра кластера до наиболее удаленной его точки.

В результате проведенных вычислительных экспериментов можно утверждать, что:

- применение упругих карт и их разверток является эффективным инструментом для аналитика при изучении многомерной текстовой информации;
- процедуры построения кластеров и их взаиморасположения надо совершенствовать таким образом, чтобы они могли отражать «вытянутые» кластеры, проблемы пересечения кластеров, проблемы вложения кластеров;
- необходимо на основании этих процедур разработать четкую методологию применения упругих карт и их разверток к анализу кластерной структуры многомерных информационных данных.

Список источников

[1] Thomas J. and Cook K. 2005 Illuminating the Path: Research and Development Agenda for Visual Analytics (IEEE-Press)

[2] Wong P. C., Thomas J. Visual Analytics // IEEE Computer Graphics and Applications. 2004. V. 24, N. 5. — P. 20-21.

[3] Keim D., Kohlhammer J., Ellis G. and Mansmann F. (Eds.) Mastering the Information Age – Solving Problems with Visual Analytics, Eurographics Association, 2010.

[4] Kielman, J. and Thomas, J. (Guest Eds.) (2009). Special Issue: Foundations and Frontiers of Visual Analytics / Information Visualization, Volume 8, Number 4, p. 239-314.

[5] T. Kohonen, Self-Organizing Maps (Third Extended Edition), New York, 2001, 501 pages.

[6] Дебок Г., Кохонен Т. Анализ финансовых данных с помощью самоорганизующихся карт, Альпина Паблишер, 2001, 317 стр.

[7] Кохонен Т. Самоорганизующиеся карты. — М.: БИНОМ. Лаборатория знаний, 2008. — 655 с.

[8] Gorban A. et al. Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE 58, Springer, Berlin – Heidelberg – New York, 2007.

[9] A. N. Gorban, A. Y. Zinovyev, Principal Graphs and Manifolds, Из: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, Olivas E.S. et al Eds. Information Science Reference, IGI Global: Hershey, PA, USA, 2009. 28-59.

[10] Zinovyev A. Vizualizacija mnogomernyh dannyh [Visualization of multidimensional data]. Krasnoyarsk, publ. NGTU. 2000. 180 p. [In Russian}

[11] Zinovyev A. Data visualization in political and social sciences, In: SAGE «International Encyclopedia of Political Science», Badie, B., Berg-Schlosser, D., Morlino, L. A. (Eds.), 2011.

[12] Питенко А.А. Нейросетевой анализ в геоинформационных системах. Красноярск, Изд. КГТУ, 2000. 97 с.

[13] Россиев А.А. Итерационное моделирование неполных данных с помощью многообразий малой размерности. Красноярск, Изд. КГТУ, 2000. 83 с.

[14] ViDaExpert, http://bioinfo.curie.fr/projects/vidaexpert, last accessed (01 March 2020).

[15] Niedoba T., Multi-parameter data visualization by means of principal component analysis (PCA) in qualitative evaluation of various coal types, Physicochemical Problems of Mineral Processing, vol. 50, iss. 2, pp. 575-589, 2014.

[16] H. Shaban, S. Tavoularis, Identification of flow regime in vertical upward air–water pipe flow using differential pressure signals and elastic maps, International Journal of Multiphase Flow 61 (2014) 62-72.

[17] H. Shaban, S. Tavoularis, Measurement of gas and liquid flow rates in two-phase pipe flows by the application of machine learning techniques to differential pressure signals, International Journal of Multiphase Flow 67(2014), 106-117

[18] M. Resta, Computational Intelligence Paradigms in Economic and Financial Decision Making, Series Intelligent Systems Reference Library, Volume 99, Springer International Publishing, Switzerland 2016

[19] Bondarev A.E., Bondarenko A.V., Galaktionov V.A., Klyshinsky E.S. Visual analysis of clusters for a multidimensional textual dataset / Scientific Visualization. V.8, № 3, pp.1-24, 2016, URL: http://sv-journal.org/2016-3/index.php?lang=en

[20] A.E. Bondarev, A.V. Bondarenko, V.A. Galaktionov (2018) Visual analysis procedures for multidimensional data. Scientific Visualization 10.4: 109 - 122, DOI: 10.26583/sv.10.4.09 http://www.sv-journal.org/2018-4/09?lang=en

[21] Bondarev, A. E.: The procedures of visual analysis for multidimensional data volumes, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W12, 17-21, https://doi.org/10.5194/isprs-archives-XLII-2-W12-17-2019 , 2019.

[22] Bondarev A.E. Visual analysis and processing of clusters structures in multidimensional datasets // Proceedings of the 2nd International ISPRS Workshop on PSBB, 15–17 May 2017, Moscow, Russia, ISPRS Archives, Volume XLII-2/W4, 2017, pp.151-154. http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W4/151/2017/

[23] Бондарев А.Е., Галактионов В.А., Шапиро Л.З. Обработка и визуальный анализ многомерных данных / Научная визуализация, Т.9, № 5, 2017, с. 86-104. http://sv-journal.org/2017-5/08/index.php?lang=ru

[24] Alexander Bondarev, Alexander Bondarenko, Vladimir Galaktionov, Lev Shapiro. Visual Analysis of Textual Information on the Frequencies of Joint Use of Nouns and Adjectives // CEUR Workshop Proceedings, V. 2744, Proc. of the 30th International Conference on Computer Graphics and Machine Vision GraphiCon 2020, Saint Petersburg, Russia, September 22-25, 2020, p. paper20-1 — paper20-10, DOI: 10.51130/graphicon-2020-2-3-20

[25] A.E. Bondarev, A.V. Bondarenko, V.A. Galaktionov. Visual Analysis of Text Data Volume by Frequencies of Joint Use of Nouns and Adjectives (2020). Scientific Visualization 12.4: 9 - 22, DOI: 10.26583/sv.12.4.02

[26] Bondarev, A. E., Bondarenko, A. V., and Galaktionov, V. A.: Visual analysis of text data collections by frequencies of joint use of words, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIV-2/W1-2021, 21–26, 2021. https://doi.org/10.5194/isprs-archives-XLIV-2-W1-2021-21-2021

УДК 004.4'2
DOI: 10.25206/978-5-8149-3873-2-2024-300-314

Разработка инструментов визуализации данных на основе предметно-ориентированного моделирования
Development of Data Visualization Tools Based on Domain Specific Modeling

А. Д. Джейранян[1], И. Д. Ермаков[2], К. А. Проскуряков[1], Л. Н. Лядова[1]
[1]*Национальный исследовательский университет «Высшая школа экономики», Пермь, Россия*
[2]*Пермский государственный национальный исследовательский университет, Пермь, Россия*

A. D. Dzheiranian[1], I. D. Ermakov[2], K. A. Proskuryakov[1], L. N. Lyadova[1]
[1]*National Research University – Higher School of Economics (HSE University), Perm, Russia*
[2]*Perm State National Research University (PSU), Perm, Russia*

Аннотация. Описывается подход к разработке средств визуализации данных, обеспечивающий возможность настройки на потребности пользователей и специфику предметных областей, в которых они работают, основанный на предметно-ориентированном моделировании. Кратко представлены результаты анализа инструментов визуализации данных и возможности их настройки на предметные области исходя из потребностей пользователей и решаемых ими задач. Показано, что существующие инструменты требуют от пользователей навыков программирования для настройки формата визуализации данных или разработки новых моделей визуализации. Предлагается использовать инструменты предметно-ориентированного моделирования (языковой инструментарий) для создания предметно-ориентированных языков (DSL), предназначенных для разработки новых моделей визуализации данных, отражающих специфику решаемых пользователями задач. Использование разработанных инструментов не требует от пользователей профессиональных знаний языков программирования. Описывается архитектура программной системы, управляемой знаниями. Ядро системы – многоаспектная онтология, которая включает описания языков и предметных областей, а также правила генерации новых языков и трансформации построенных моделей. Языки предназначены для описания различных классов диаграмм. Система включает инструменты для автоматизации создания новых DSL через отображение онтологии предметной области на метамодель базового языка по заданным пользователем правилам, которые также сохраняются в онтологии. Классификация различных типов диаграмм составляет основу для создания онтологии языков визуализации данных. Описывается пример базового языка для создания диаграмм. Демонстрируется возможность настройки DSL и трансформации разработанных с его помощью моделей визуализации, генерации кода, реализующего модель.

***Ключевые слова*:** визуализация данных, предметно-ориентированное моделирование, предметно-ориентированный язык, метамоделирование, грамматика, многоаспектная онтология, трансформация моделей

Аннотация. An approach to the development of data visualization tools is described that provides the ability to customize to the needs of users and the specifics of the domains in which they work, based on domain-specific modeling. The results of the analysis of data visualization tools and the possibility of customizing them to subject area based on the needs of users and the tasks they solve are briefly presented. It is shown that existing tools require programming skills from users to customize the data visualization format or to develop new visualization models. It is proposed to use domain-specific modeling tools (language toolkits) to create domain-specific languages (DSL) designed to develop new data visualization models that reflect the specifics of the tasks solved by users. The use of the developed tools does not require users to have professional knowledge of programming languages. The architecture of a knowledge-driven software system is described. The core of the system is a multifaceted ontology, which includes descriptions of languages and domains, as well as rules for generating new languages and transforming constructed models. Languages are designed to describe different classes of diagrams. The system includes tools for automating the creation of new DSLs by mapping the domain ontology onto the metamodel of the base language according to user-specified rules, which are also stored in the ontology. The classification of different types of diagrams forms the basis for creating an ontology of data visualization languages. An example of a basic language for creating diagrams is described. The ability to customize the DSL and transform visualization models developed with its help and generate code that implements the model is demonstrated.

***Ключевые слова*:** data visualization, domain-specific modeling, domain-specific language, metamodeling, grammar, multifaceted ontology, model transformation.

I. Введение

Инструменты визуализации получили широкое распространение в различных отраслях, бизнес-функциях и ИТ-дисциплинах, как в частном, так и в государственном секторе. Они активно используются в таких областях, как энергетика, картография, медицина, финансы, социология и многие другие. В этом контексте визуализация данных служит методом анализа данных. Эффективность работы экспертов, аналитиков во многом определяется качеством визуализации данных, результатов их анализа.

Вопросам качества визуализации данных посвящены многочисленные исследования.

Нечитабельные, запутанные визуализации распространяют дезинформацию, вводят в заблуждение и снижают эффективность работы исследователей, качество управленческих решений и т. д. Авторы статьи [1] подчёркивают масштаб проблемы, ссылаясь на ресурсы, собирающие десятки тысяч пользователей, критикующих графики и диаграммы (reddit.com/r/dataisugly), выполненные с очевидными ошибками. Существующие методы визуализации данных дают возможность строить в основном простые и чрезмерно обобщенные визуализации, которым, к тому же, часто не хватает систематичности и универсальности. Отсутствие надлежащего инструментария для оценки качества визуализации данных еще больше усугубляет проблему. В статье предлагается VisQualdex – систематический набор рекомендаций по статической визуализации данных. Категоризация основана на теории грамматик графики. Предлагаются десятки критериев, предназначенных для выявления различных ошибок (ошибок разных категорий и масштаба). Предложенный набор рекомендаций рецензирован и протестирован экспертами в областях визуализации данных, науки о данных, графического дизайна, информационных технологий и информатики. Реализация рекомендаций доступна в виде веб-сервера, разработанного как одностраничное приложение на JavaScript с использованием принципов Vue.js и Material Design.

Особое внимание исследователями уделяется вопросам применения средств визуализации в образовании.

Авторы статьи [2] показывают, что инфографика важна для представления и передачи сложной информации, в частности, обсуждают влияние визуализации на результаты учебного процесса, организованного при взаимодействии с инфографикой, мультимедийными средствами. Результаты исследования показали, что использование инфографики, визуального обучения является эффективным подходом к организации учебного процесса. Цель исследования, результаты которого представлены в [3], – изучение теоретических и практических аспектов визуализации знаний в образовательном процессе. Авторы сравнивают различные модели представления знаний, рассматривают принципы визуализации и структуру компетенций визуализации, включающую аналитические, алгоритмические и пр. компоненты. Используя различные методы моделирования, визуализации и систематизации данных, авторы делают вывод, что визуализация знаний требует учёта особенностей образовательного процесса в контексте дидактических принципов и пр. Формирование навыков визуализации в ходе учебной деятельности – это сложная задача, зависящая от множества факторов. Эти навыки являются базовыми компонентами формирования визуализационной компетентности. Решение задачи визуализации знаний требует использования современных инновационных технологий, охватывающих базовые цифровые технологии, технологии обработки больших данных, методы интеллектуального анализа многомерных данных и т. д. В статье [4] рассматриваются программные средства, которые используются в образовательном процессе и влияют на формирование профессиональных компетенций у студентов. Цель проекта – разработка технологии создания образовательного VR-контента, который должен повысить эффективность обучения с помощью множества различных программных сред, используемых в процессе формирования конечного виртуального пространства.

В статье [5] описывается подход к пониманию идеи когнитивной ясности графовых моделей. Авторами описывается концептуальная схема структурирования понятий, связанных с когнитивной ясностью. В соответствии с этой схемой они выделяют факторы формирования когнитивной ясности. Когнитивная ясность определяется как набор внутренних характеристик визуального образа модели, эффекты присутствия которых проявляются при визуальном анализе модели. Авторы делают вывод, что наибольший интерес представляют именно факторы формирования когнитивной ясности (в силу их конструктивности). В статье приводится детализированная схема подхода к пониманию идеи когнитивной ясности и обсуждаются отдельные компоненты этой схемы. Предлагается обобщенный алгоритм подготовки и проведения эксперимента по решению аналитиком определенной задачи визуального анализа с участием факторов, о влиянии которых перед проведением эксперимента формируется гипотеза. В результате оценивается эффект изменения уровня когнитивной ясности и выявляется характер зависимости от заданных факторов или ее отсутствие, что позволяет принять, отвергнуть или уточнить исходную гипотезу.

Вопросы эффективности визуализации данных рассматриваются в статьях [6, 7]. Восприятие графической информации связано с определёнными зрительными когнитивными процессами, опираясь на исследование которых, авторы дают общие рекомендации по созданию эффективных научных визуализаций.

Потребности конечных пользователей (аналитиков данных) включают необходимость создания пользовательских типов диаграмм для конкретных задач и предметных областей, поскольку базовые типы диаграмм с базовой геометрией могут ограничивать передачу информации [1] и приводить к неэффективной визуализации, что, в свою очередь, может привести к ошибкам в принятии решений [8].

Пользователям требуется создание визуализаций, адаптированных к специфике решаемых ими задач и предметных областей. Таким образом, существует необходимость внедрения пользовательских спецификаций визуализации в используемые средства. Эти спецификации определяют, как пользователи могут указывать свои требования для создания визуализаций [9]. Статическая визуализация может оказаться неэффективной для решения задач анализа больших наборов данных, поэтому существует потребность в быстром и простом создании интерактивных визуализаций [10]. Эти методы должны быть способны работать с различными типами и источниками данных [9].

Исследователи [8, 11] отмечают ограниченную степень гибкости в манипулировании элементами диаграмм и отсутствие ориентированности на реальные потребности пользователя в существующих инструментах визуализации данных. Часто такая настройка требует использования языка программирования [10]. Отсутствие глубоких знаний программирования у пользователей приводит к необходимости создания no-code или low-code платформ.

В статье [12] рассматриваются различные техники и инструменты визуализации данных, приводится их классификация, формулируются проблемы, с которыми сталкиваются пользователи, и новые возможности, которые должны быть обеспечены перспективными средствами визуализации. В частности отмечается, что одной из ключевых задач является разработка методов автоматизации визуализации данных, обеспечения эффективной и интерактивной визуализации независимо от размера и сложности данных. Реализация этих методов облегчает исследования в областях с интенсивным использованием данных [13], позволяет непрерывно отслеживать и анализировать эти данные, визуализировать результаты анализа.

Автоматический визуальный анализ предполагает решение задач поиска, очистки, интеграции и визуализации данных. Для решения этих задач широко используются онтологии.

В работах [13, 14] было предложено использовать онтологии как часть архитектуры, как ядро аналитической платформы, управляемой знаниями. В данном случае используется многоаспектная онтология, которая позволяет реализовать семантический поиск и индексацию данных, избежать их дублирования, расширять функциональность информационных и аналитических систем, создавать предметно-ориентированные языки и модели и сценарии исследований, а также автоматически интерпретировать данные и результаты анализа данных для предоставления их разным группам пользователей согласно знакомой им терминологии.

Доступные системы инструментов визуализации данных можно разделить на следующие группы:

1) электронные таблицы (например, Excel, Google Sheets),

2) аналитические платформы (например, Microsoft Power BI, Tableau),

3) редакторы диаграмм (например, Miro, ChartBlocks).

Стандартные инструменты первых двух групп ограничены базовыми типами диаграмм и возможностями настройки визуальных эффектов. Инструменты третьей группы позволяют создавать собственные визуализации, редактировать расположение элементов, но не предоставляют возможность настройки на предметные области.

Использование языков программирования общего назначения (например, библиотек Python для визуализации данных: Matplotlib, Seaborn, Plotly и др.) способствует созданию выразительных визуализаций для решения конкретных задач, но требует глубоких знаний программирования от разработчиков диаграмм. Также созданное решение нельзя повторно использовать для других визуализаций, и оно представляет собой «черный ящик», где непонятно, как настроена визуализация [10].

Визуализация, отражающая специфику предметной области, может в значительной степени способствовать лучшему пониманию представленных данных, результатов их анализа, формальных моделей. Создание таких средств может быть основано на предметно-ориентированном моделировании (Domain Specific Modeling, DSM), предполагающем создание и использование предметно-ориентированных языков (Domain Specific Language, DSL) для разработки новых вариантов (моделей) визуализации данных.

Публикаций по созданию DSL для визуализации данных немного. Судя по обзору, большинство DSL фокусируются на небольшом наборе стандартных диаграмм (круговых диаграмм, гистограмм и т. д.) или визуализации конкретных типов данных (например, геопространственных и т. п.). Они различаются уровнями абстракции, контекстами использования и возможностями реализации.

В статье [15] описан процесс разработки DSL для построения и преобразования методов визуализации данных. Этот DSL встроен в язык программирования Haskell. Авторы предоставляют несколько уровней абстракции: на самом низком уровне пользователь может создать элемент, состоящий из определенной примитивной формы и набора визуальных параметров. Важно отметить, что основные конструкции языка

ограничены гистограммой и круговой диаграммой. Однако допускается располагать их элементы по-разному для создания более сложных вариантов визуализации.

В статье [16] представлена вариационная модель визуализации, реализованная с помощью DSL, встроенного в язык программирования PureScript. Этот DSL позволяет создавать вариационные визуализации и их комбинации, например наложение альтернативных гистограмм. В статье также обсуждаются методы представления вариаций и добавления вариаций к визуализациям через DSL. Разработанные средства обеспечивают создание, управление, навигацию и рендеринг различных визуализаций.

Исследователи в статье [17] представляют DSL, ориентированный на геовизуализацию данных. Они используют компилятор для облегчения автоматического создания визуализаций и предварительной обработки данных. Их система использует возможности многоядерного параллелизма для ускорения предварительной обработки данных.

В статье [18] предлагается подход, который помогает пользователю создавать визуализации моделей для конкретной предметной области с использованием CSP (Communicating Sequential Processes) – формального языка, который в основном используется для описания параллельных и распределенных систем. Авторами успешно созданы различные визуализации моделей CSP, демонстрирующие возможности предлагаемого подхода. Однако рассматриваемые средства не являются универсальными, возможности языка ограничивают его использование.

Языково-ориентированный подход (в данном случае подход, основанный на DSM) к реализации средств визуализации данных может стать основным при разработке системы визуализации данных с возможностью настройки на потребности пользователей. Создание новых вариантов визуализаций, настроенных на конкретные предметные области и задачи пользователей, разбивается на два этапа:

1. Разработка предметно-ориентированного языка (DSL), отражающего потребности пользователей.
2. Разработка предметно-ориентированных моделей визуализации с использованием созданного языка.

Рассмотренные инструменты предоставляют возможность разработки новых типов диаграмм. Но возможности настройки на различные предметные области в них отсутствуют или они ограничены. Кроме того, от пользователей требуются профессиональные навыки программирования, что также ограничивает возможности использования описанных средств.

Для создания DSL используются DSM-платформы (языковые инструментарии), которые автоматизируют разработку редакторов моделей и средств их трансформации, генераторов кода. Актуальной становится задача снижения трудоёмкости создания самих языков, автоматизации их разработки.

Таким образом, конечные пользователи сталкиваются с двумя вопросами при решении задач визуализации данных:

1. Как обеспечить создание или настройку визуализаций в соответствии с потребностями пользователей?
2. Как снизить уровень требований к компетенциям пользователя в программировании при использовании средств визуализации, обеспечивающих возможность настройки на потребности пользователей?

Для расширения возможностей инструментов визуализации предлагается языково-ориентированный подход, использование языкового инструментария, управляемого знаниями, позволяющего автоматизировать создание предметно-ориентированных языков на основе использования многоаспектной онтологии [19, 20, 21].

II. Постановка задачи

Цель проекта, результаты которого представлены в данной статье, – разработать подход к созданию средств эффективной визуализации данных с учётом рекомендаций, предложенных в [6, 7], обеспечивающих возможность создания пользовательских типов диаграмм для конкретных задач и предметных областей, а также возможность создания интерактивных визуализаций [10].

Специфика требований заключается в необходимости настройки визуализации на специфику предметных областей и решаемых пользователями задач или разработки новых типов диаграмм. Этого результата можно достичь за счет интерактивности, комбинирования разных типов диаграмм и т. д. Эти требования должны быть учтены при создании средств для разработки моделей визуализации (графиков, схем, диаграмм).

Предлагаемая в статье основа для реализации инструментов визуализации данных – DSM-платформа, управляемая знаниями, обеспечивающая возможность создания иерархий предметно-ориентированных языков [19, 20, 21], предназначенных для создания новых типов моделей (новых графических нотаций), отвечающих потребностям пользователей. Снижение трудоёмкости разработки новых DSL достигается за счёт использования базовых языков, их дополнения и комбинирования в соответствии с правилами, определяемыми пользователями. Построенные с использованием созданных DSL модели визуализации могут быть переведены в программный код, реализующий эффективную пользовательскую визуализацию данных, на основе заданных правил трансформации типа «Модель-Текст».

Основные функциональные требования к инструментам визуализации данных показаны в формате диаграммы прецедентов на рис. 1. Эксперт в предметной области должен разработать онтологию предметной области, в которой пользователи – разработчики визуализаций – решают свои задачи. DSM-эксперт отвечает за разработку метамоделей базовых языков (в данном случае – языков для разработки базовых типов диаграмм, которые могут послужить основой для создания новых DSL, реализующих потребности пользователей), а также за разработку правил генерации новых языков, отражающих специфику предметных областей, в которых работают пользователи, – правил отображения онтологии предметной области на метамодель выбранного базового языка. При разработке визуализации пользователь может использовать созданный для неё DSL, выполнив настройку параметров, или разработать новый DSL. Генерация кода на языке программирования по построенной модели визуализации позволяет создавать интерактивную визуализацию, управляемую событиями, комбинировать различные типы визуализаций при работе со сложными структурами данных, получаемых из различных источников. Правила трансформации (генерации кода) разрабатываются DSM-экспертом.

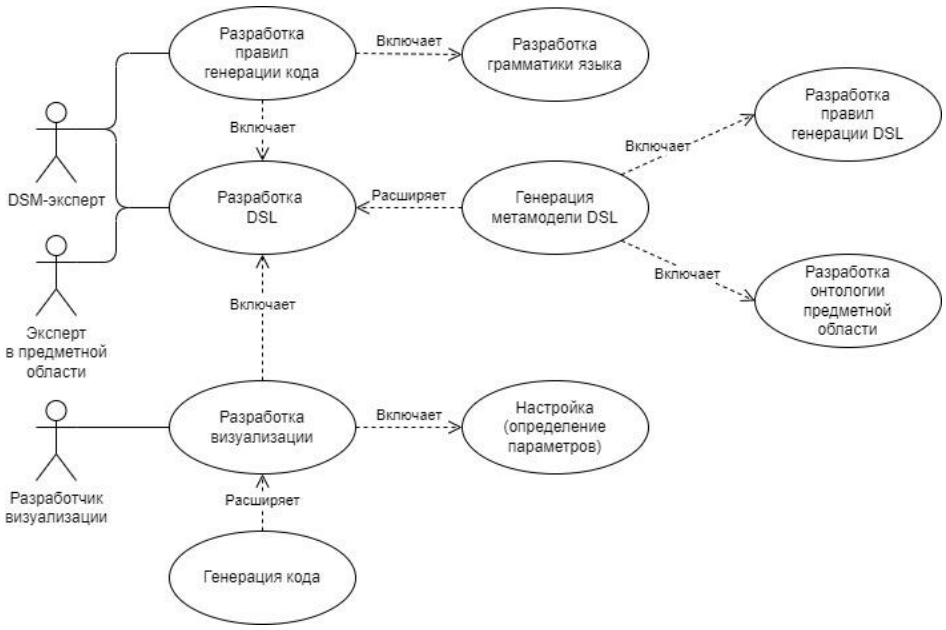


Рис. 1. Основные функции средств разработки моделей визуализации на основе DSM-подхода

Ядром языкового инструментария, с помощью которого должны быть реализованы описанные функции, является многоаспектная онтология [19, 20, 21], которая кроме описания предметной области включает и онтологию языков, где визуальные языки (языки создания схем, диаграмм и т. п.) описываются с помощью метамоделей, а текстовые – их грамматиками. Правила генерации DSL и кода также включаются в онтологию.

Разработка метамоделей DSL, правил генерации новых DSL и генерации кода не требует знания языков программирования – пользователи работают с редакторами и конструкторами в визуальной среде. Для создания онтологии предметной области предлагается использовать метод Mask [22]. Таким образом снижаются требования к квалификации пользователей, к навыкам программирования.

Для реализации предлагаемого подхода необходимо решить следующие задачи:

1. Проанализировать различные типы диаграмм и построить онтологию языков визуализации данных на основе классификации диаграмм.
2. Разработать метамодели базовых языков для различных типов диаграмм, используемых для визуализации данных, – основу для создания новых типов визуализации.
3. Разработать правила генерации новых DSL – новых языков визуализации данных, отражающих специфику предметных областей, и сгенерировать новые типы диаграмм на основе разработанных правил.
4. Разработать правила трансформации (генерации кода) для реализации нового типа визуализации данных.

Результаты необходимо проиллюстрировать примерами визуализации данных с использованием разработанных средств.

III. Теория

Визуальные языки основаны на абстрактных – графовых – моделях: любая модель, диаграмма представляется как граф, в котором представлены объекты некоторой предметной области и связи между ними; абстрактная графовая модель расширяется – дополняется атрибутами, характеристиками, которые могут быть приписаны

элементам графа, представляющим объекты конкретной предметной области; для этой конкретизированной модели строится её визуальное представление с использованием конкретного типа визуализации – графической нотации, которую пользователь выбирает, основываясь на своих потребностях. Таким образом, модель с одной и той же структурой может по-разному визуализироваться при решении пользователями различных задач.

6. *Графовая модель для реализации методов визуализации данных на основе DSM-подхода*

В основе реализации языкового инструментария – графовая модель, которая должна, с одной стороны, обладать достаточными выразительными возможности для представления структуры моделируемых объектов в различных предметных областях, а с другой – позволять разрабатывать эффективные алгоритмы для операций над построенными моделями.

Для реализации языкового инструментария предложена графовая модель – иерархические гиперграфы с полюсами (*HP*-графы) [19].

HP-граф – это ориентированный граф, который представляется как тройка $G = (P, V, W)$, где $P = \{\pi_1, \ldots, \pi_n\}$ – это множество внешних полюсов графа, $V = \{v_1, \ldots, v_m\}$ – это непустое множество вершин, а $W = \{w_1, \ldots, w_l\}$ – это множество гиперрёбер графа. Пусть *Pol* – абстрактное множество всех полюсов графа. Полюс – это базовое понятие выбранной графовой модели, через которое определяются все элементы графа:

- *Вершины* $v \in V$ – это непересекающиеся подмножества *полюсов* из множества *Pol*, при этом объединение всех этих подмножеств даёт всё множество полюсов *Pol*, то есть каждый полюс принадлежит ровно одной вершине. Полюсы – это «интерфейс» вершины, через который вершина связывается с другими вершинами. Каждая вершина $v \in V$ представленная множеством полюсов $P_v = \{p_{v_1}, \ldots, p_{v_n}\}$, имеет множество входных и выходных полюсов ($I(v)$ и $O(v)$), эти множества могут пересекаться, то есть один и тот же полюс может быть и входным, и выходным. В простейшем случае вершина может быть представлена единственным полюсом, тогда $I(v) = O(v)$.
- Множество *внешних полюсов* P графа – это также подмножество множества полюсов *Pol*. Это множество состоит из входных и выходных полюсов графа (эти полюсы являются «входами» и «выходами» графа): $P = I(G) \cup O(G)$, эти полюсы являются входными или выходными полюсами вершин.
- Каждое *гиперребро* графа $w \in W$ определяет связи между вершинами и представляется подмножеством множества полюсов этих вершин: $w = P_w = \{p_{w_1}, \ldots, p_{w_k}\}$, каждое такое множество должно содержать хотя бы один входной и один выходной полюс (эти полюсы могут принадлежать нескольким вершинам, между которыми устанавливаются связи, или одной вершине).

Как вершины, так и рёбра могут быть «раскрыты» – расшифрованы графом, который детализирует их структуру. Таким образом может быть построена иерархическая диаграмма (с декомпозицией её элементов, которая может быть выполнена в интерактивном режиме, по правилам, заданным пользователем).

Показано, что эта модель обладает всеми возможностями, которые обеспечивают другие широко используемые графовые модели (метаграфы, гиперграфы, графы с полюсами и пр.), а также позволяет оптимизировать алгоритмы – минимизировать их временную сложность, упростить объектную модель при реализации программных средств. Иерархическая модель строится с помощью операции расшифровки, определённой для этого типа графов (операции декомпозиции, которая может быть применена как к вершинам, так и к гиперрёбрам). Такие возможности графовой модели обеспечивают эффективную реализацию группировки элементов при построении диаграмм, декомпозиции при разработке интерактивных визуализаций.

7. *Метаязык для описания визуальных языков в языковом инструментарии*

При разработке визуальных языков в языковых инструментариях строятся их описания в виде графовых грамматик или метамоделей. Для создания этих описаний должны использоваться соответствующие метаязыки (по аналогии с формальными грамматиками, которые можно задать в форме БНФ (форма Бэкуса–Наура) или диаграмм Вирта). Метаязык должен обеспечивать возможность представления *HP*-графа, всех его элементов. Оценка метаязыков представлена в табл. 1. Наиболее подходящим метаязыком с точки зрения представления всех элементов, которые могут быть использованы в графической модели, является язык *GOPPRR* (Graph, Object, Port, Property, Relationship, Role) DSM-платформы MetaEdit+, где Graph представляет модель в целом, Object – элемент модели, представляемый вершиной, Port представляет точку, через которую вершина-объект связывается с другими вершинами, Property – это свойство вершины или связи, Relationship – связь между объектами, Role – роль вершины в связи.

Ни один из языков не позволяет описывать в модели «обобщённые» связи, в которых могут участвовать множества вершин.

Для реализации языкового инструментария предложен язык [21], который является расширением GOPPRR.

ТАБЛИЦА 1
СРАВНЕНИЕ МЕТАЯЗЫКОВ, ИСПОЛЬЗУЕМЫХ В ЯЗЫКОВЫХ ИНСТРУМЕНТАРИЯХ

Метаязык	**Модель**	**Объект модели**	**Связи между объектами**	**Роли**	**Полюсы**	**Свойства**
1. EMOF	2. –	3. Class	4. Property, Association	5. –	6. –	7. Property
8. Ecore	9. –	10. EClass	11. EReference	12. –	13. –	14. EAttribute
15. GOPPRR	16. Graph	17. Object	18. Relationship	19. Role	20. Port	21. Property
22. ArkM3	23. –	24. Class	25. Association, Composition	26. –	27. –	28. Property
29. WebGME MML	30. FCO	31. FCO	32. Pointer, Set, Connection, Containment, Inheritance, Mixin	33. Connection Role	34. –	35. Attribute

Разработанный метаязык (рис. 2) позволяет создавать с его помощью метамодели визуальных языков с использованием всех описанных выше элементов *HP*-графов.

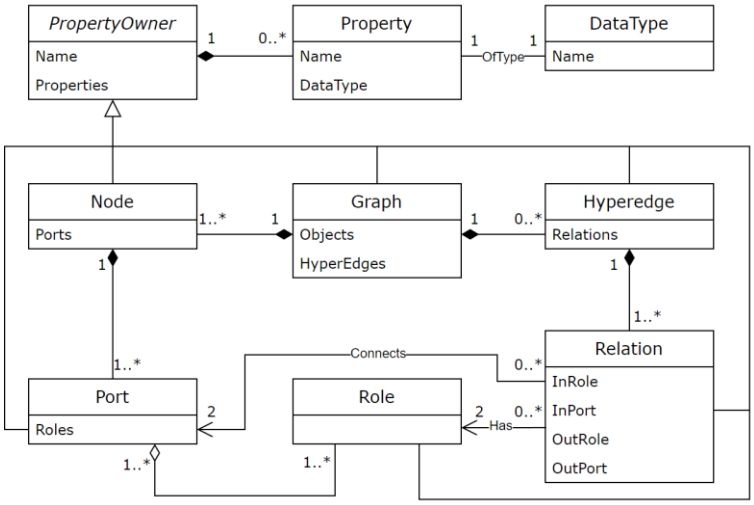


Рис. 2. Описание метаязыка *HPGPR* исследовательского прототипа языкового инструментария

Описанные выше модели использованы при разработке исследовательского прототипа языкового инструментария, позволяющего создавать новые предметно-ориентированные языки (визуальные DSL), настраиваемые на потребности пользователей, на основе базовых языков, метамодели которых представлены в системе. Разработанные языки также могут использоваться в качестве базовых для создания новых языков.

8. Обобщённая структура языкового инструментария

Обобщённая структура языкового инструментария, использованного для разработки прототипа средств визуализации данных, основанных на языково-ориентированном подходе, представлена на рис. 3. Показаны только те компоненты, которые должны быть реализованы для создания настраиваемых средств визуализации данных на основе DSM-подхода. В состав *многоаспектной онтологии* входят:

– *онтология визуальных языков* моделирования, включающая метамодели базовых языков, предназначенных для разработки визуальных моделей, описанные при помощи метаязыка платформы *HPGPR*, а также языков (DSL), созданных на их основе;

– *онтология текстовых языков*, содержащая описание грамматик текстовых языков;

– *онтология правил трансформаций*, хранящая двойки, первыми элементами которых являются элементы метамоделей визуальных языков, а вторыми – нетерминальные символы конструкций текстовых языков;

– *онтология предметных областей*, содержащая информацию о предметных областях, в которых пользователи решают свои задачи, для которых создаются DSL;

– *онтология правил генерации метамоделей* новых DSL, отражающих специфику предметных областей, включающая описание правил проецирования онтологий предметных областей на метамодели визуальных языков, выбранных в качестве базовых;

– *онтология моделей*, содержащая информацию о созданных моделях;

– *онтология пользовательских действий*, хранящая информацию о пользовательских действиях при работе с моделями (например, клики или наведения курсора мышки);

– *онтология взаимодействия*, представляющая информацию о возможных действиях пользователей при работе в интерактивном режиме в системе, например, построения новой модели, или приближения её части и пр.;

– *онтология правил интерпретации*, хранящая тройки, первыми элементами которых являются элементы метамоделей визуальных языков, вторыми элементами – пользовательские действия, а третьими – интерактивные действия, которые могут быть использованы при разработке интерактивных визуализаций.

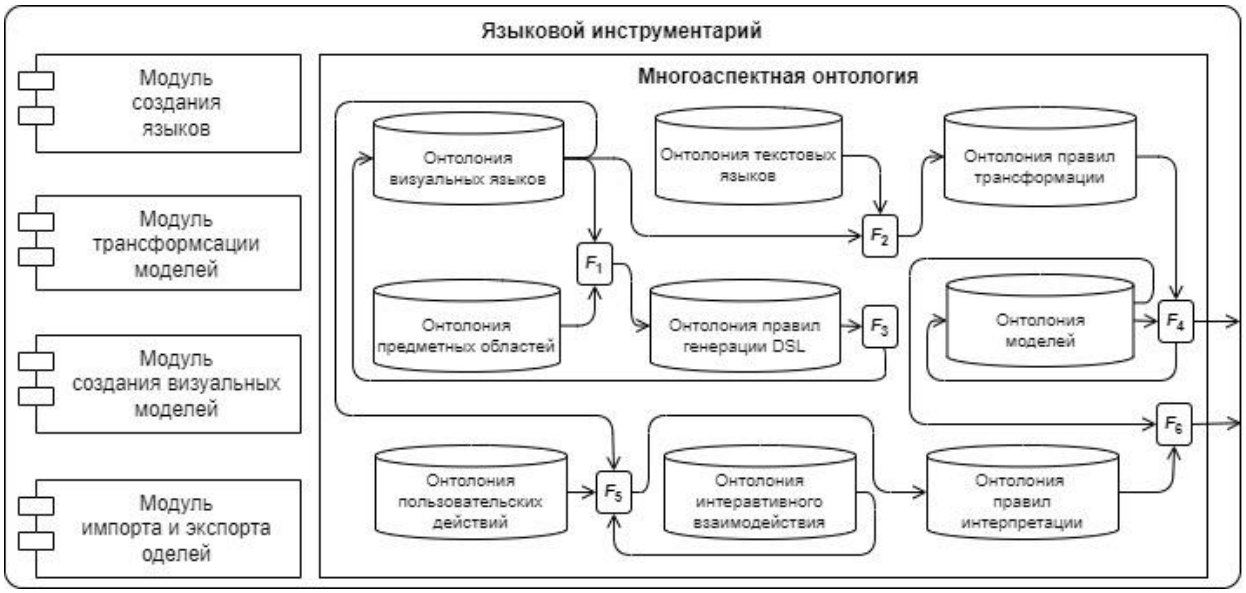


Рис. 3. Упрощённая структура языкового инструментария, управляемого знаниями

Многоаспектная онтология включает также функции, реализующие базовые алгоритмы, обеспечивающие автоматизацию генерации метамоделей языков, трансформацию моделей (в частности, генерацию кода – трансформацию моделей типа «Модель–Текст»), создание интерактивных моделей, их интерпретацию:

– F_1 – функция, обеспечивающая автоматизацию разработки новых языков путём отображения элементов онтологии предметной области на элементы метамодели выбранного базового языка, создания правил генерации DSL. Результат – правила генерации (отображения), определённые пользователем с помощью специального конструктора.

– F_2 – функция, с помощью которой определяется соответствие конструкций текстовых языков элементам метамоделей визуальных языков для создания правил трансформации моделей (генерации кода).

– F_3 – функция, реализующая генерацию метамодели нового DSL в соответствии с правилами, полученными с помощью F_1. Метамодель нового DSL сохраняется в онтологии визуальных языков. Новый DSL – подкласс базового языка, к которому применены правила отображения онтологии предметной области на его метамодель, заданные пользователем при выполнении F_1.

– F_4 – функция, реализующая трансформации моделей в соответствии с заданными правилами (для текстовых целевых языков выполняется генерация кода – трансформация типа «Модель-Текст»).

– F_5 – функция, принимающая элементы метамоделей визуальных языков, пользовательские сценарии и интерактивные действия для создания правил интерпретации для интерактивных моделей (визуализаций).

– F_6 – функция, которая обеспечивает применение заданных правил интерпретации к визуальным моделям для предоставления пользователям возможностей взаимодействия с моделями, создания интерактивных визуализаций.

Многоаспектная онтология включает также онтологию источников данных, онтологию функциональных модулей, которые также используются при создании моделей визуализации.

Далее описаны результаты разработки пользовательского прототипа средств визуализации данных, настраиваемых на потребности пользователей, основанные на использовании описанного выше подхода.

IV. Результаты экспериментов

Для апробации подхода (реализации прототипа средств визуализации) необходимо построить онтологию языков и метамодели базовых языков для описания диаграмм (визуализации данных традиционными способами), задать правила генерации новых DSL, с помощью которых можно выполнить визуализацию данных,

отражающую потребности пользователей, определить правила трансформации моделей для создания интерактивной визуализации. Далее приведено краткое описание результатов решения этих задач.

1. Классификация методов визуализации

Разнообразие способов визуализации достаточно велико и продолжает расширяться, что подтверждается постоянным появлением новых специализированных типов диаграмм (древовидных, хордовых, сетевых и т. д.).

Выбор наиболее подходящего типа визуализации в первую очередь определяется тем, какую задачу решает пользователь, какую идею он должен передать с помощью диаграммы [23, 24]. Чтобы определить, какие именно методы визуализации достаточны для реализации большинства идей визуализации, необходимо остановиться на определенной таксономии – выполнить классификацию методов визуализации данных.

Было решено сосредоточиться на пятикатегорийной структуре, предложенной Энди Кирком [24]:

1) сравнение категорий;
2) оценка иерархий и отношений части к целому;
3) представление изменений с течением времени;
4) определение типов связей и отношений;
5) картографирование геопространственных данных.

При разработке классификации учтены следующие характеристики:

1) степень популярности методов визуализации;

2) включение в дополнение к стандартным методам визуализации (линейные графики, гистограммы, круговые диаграммы, гистограммы, диаграммы рассеяния и пр.) методов визуализации абстрактных данных, характеризующихся многомерностью и отсутствием явных пространственных привязок;

3) возможность представлять любой тип визуализации в интерактивной форме.

В результате была разработана классификация, включающая шестнадцать методов визуализации, выделенных в зависимости от цели создания визуализации (рис. 4). Эта классификация является основой для разработки онтологии языков визуализации данных, создания метамоделей этих языков.

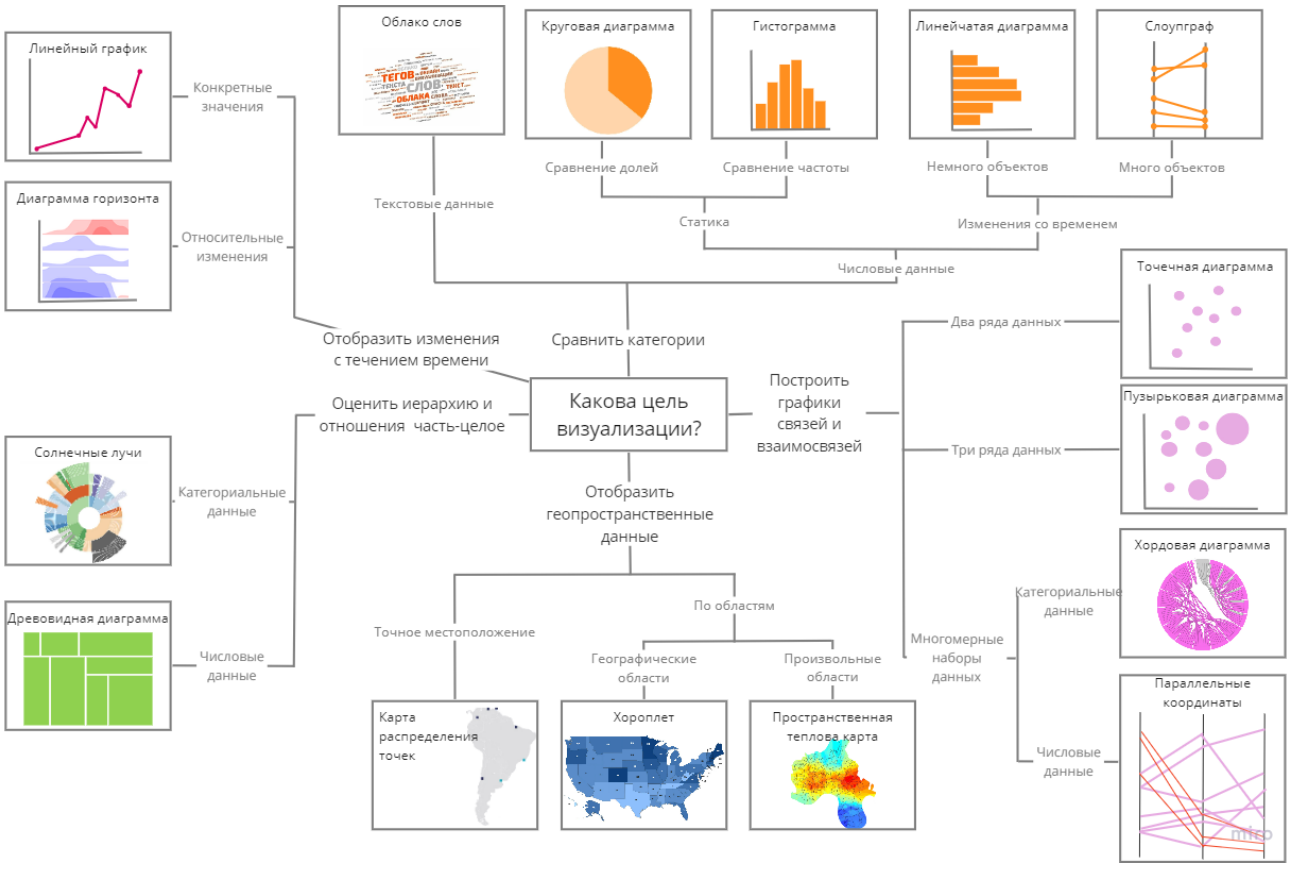


Рис. 4. Классификация методов визуализации по назначению

2. Разработка языков для создания моделей визуализации данных

Процесс построения онтологии языков визуализации данных включает в себя следующие этапы:

1. Формальное описание абстрактной диаграммы, которая будет включать свойства, общие для всех типов диаграмм (заголовок, легенда, ширина, высота и т. п.).
2. Выделение типов диаграмм в отдельные классы на основе разработанной классификации (рис. 4).
3. Добавление в описания классов диаграмм специфичных для каждого класса элементов диаграмм.
4. Определение отношений между классами.

Для каждого типа диаграмм формируется своя формализованная модель – представление диаграммы в онтологии, для которого создаётся DSL, позволяющий описывать этот тип визуализации.

На рис. 5 показан класс абстрактной диаграммы и его потомки в онтологии. Каждая из диаграмм состоит из отдельных элементов, имеющих свои свойства, которые также представляются в онтологии (рис. 6).

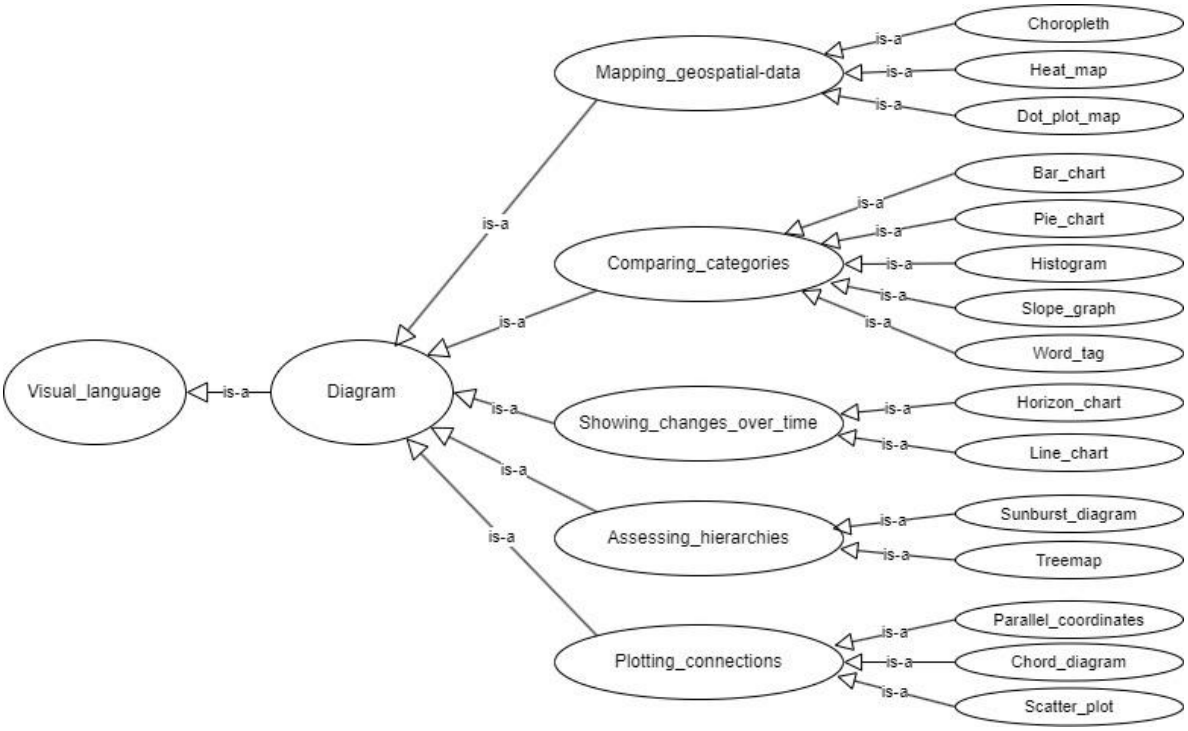


Рис. 5. Фрагмент многоаспектной онтологии: иерархия классов диаграмм в онтологии

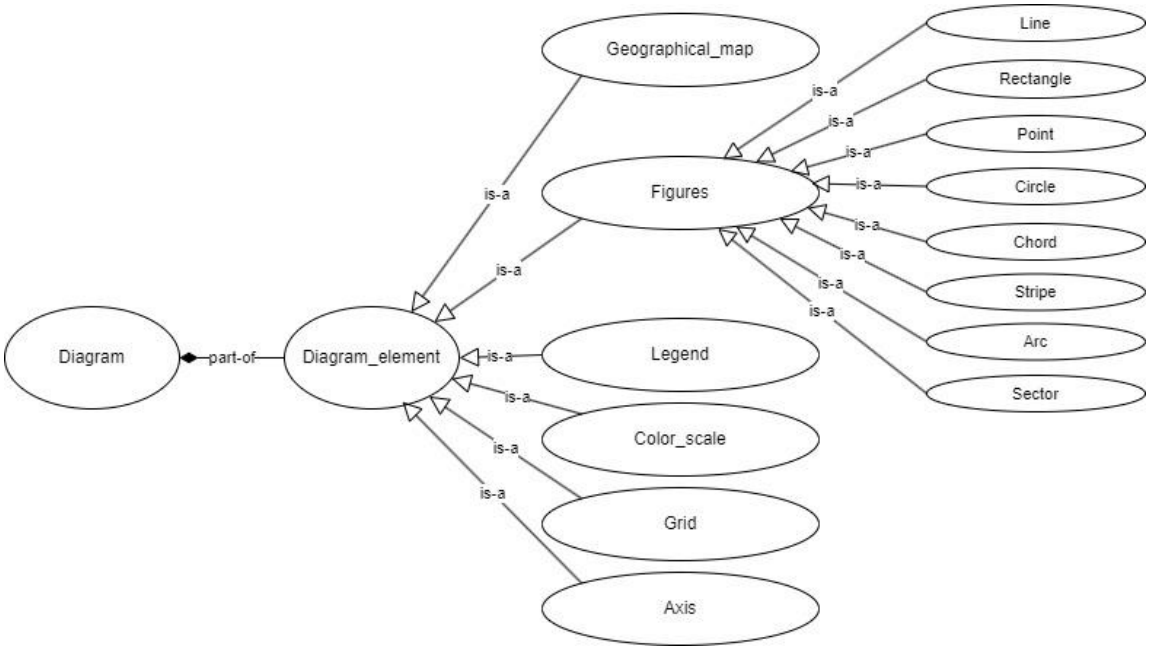


Рис. 6. Фрагмент многоаспектной онтологии: иерархия классов элементов диаграммы в онтологии

После определения основных классов и их уникальных элементов необходимо определить иерархические отношения и отношения «часть-целое» между классами. Отношения «as-is» автоматически создаются между родительским классом и дочерним классом в иерархии. Все элементы имеют собственные наборы атрибутов.

Для отображения отношений между конкретной диаграммой и ее элементами создаются специальные связи.

Для каждого типа диаграмм разрабатывается свой базовый язык. Адаптация базового языка к потребностям пользователей выполняется сопоставлением элементов диаграмм, представленных в метамодели базового языка, с элементами в описании онтологии соответствующей предметной области. На рис. 7 показана метамодель языка (DSL), который настроен на визуализацию рейтингов.

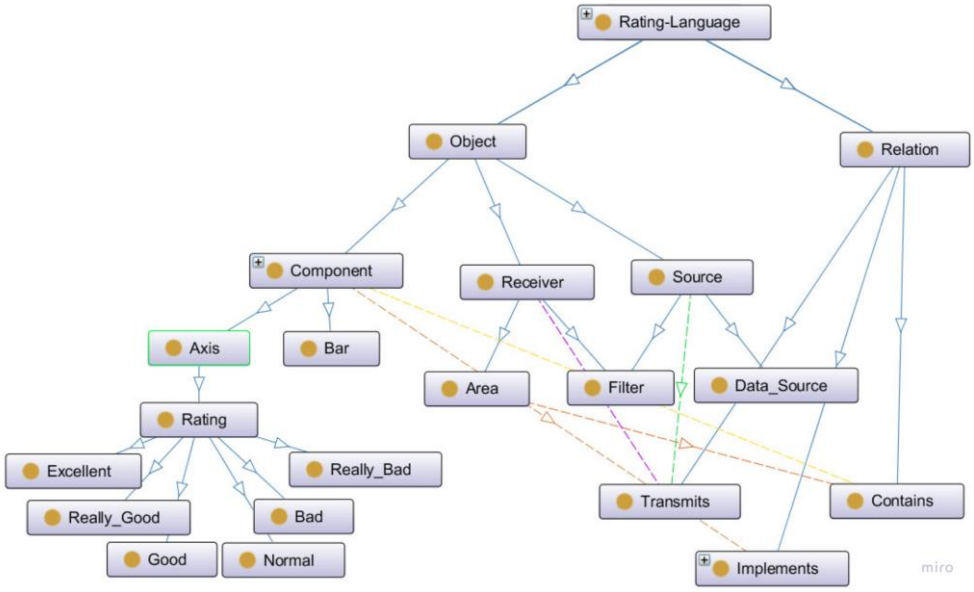


Рис. 7. Метамодель языка «Rating-Language» в онтологии визуальных языков

В разработанной метамодели DSL представлены базовые компоненты разработанной графической нотации, такие как «Axis» (оси), «Bar» (столбец), а также «Area» (область). У каждого из компонентов есть свойства (атрибуты), описание которых опущено, чтобы не загромождать иллюстрацию. С классом оси связан класс «Rating», используемый для отображения области значения допустимых рейтингов, в данном случае «Excellent», «Really_Good», «Good», «Normal», «Bad» и «Readlly_Bad». Такие классы как «Source» и «Receiver» используются для представления источника и приёмника данных соответственно. Можно заметить, что оба класса связаны с другим классом «Filter», позволяющим фильтровать данные.

С помощью описанного языка пользователь может построить свою модель визуализации, разработать диаграмму в визуальном конструкторе, определив значения всех атрибутов, указанных в метамодели языка. На рис. 8 показана диаграмма, разработанная с использованием описанного DSL.

Рис. 8. Диаграмма, построенная с использованием языка визуализации рейтингов «Rating-Language»

3. Трансформация моделей – генерация кода, реализующего пользовательскую визуализацию

С помощью разработанных DSL строятся модели визуализации, соответствующие потребностям пользователей, но для реализации разработанной модели визуализации данных требуется либо генерация кода, либо интерпретация моделей.

Подавляющее большинство современных DSM-платформ использует подход к генерации кода на основе шаблонов, что позволяет эффективно повторно использовать шаблоны [25]. Шаблоны описываются не для

конкретной модели, а для метамодели [26] (в данном случае метамодели языка визуализации). Каждый шаблон состоит из двух частей – статической и динамической. Статическая часть одинакова для всех моделей, разработанных с использованием языка, определяемого метамодельюю, а динамическая использует информацию, «извлеченную» из конкретной модели.

В данной работе мы предлагаем использовать подход, описанный в [21], заключающийся в создании правил трансформации моделей в визуальной среде (в конструкторе правил трансформации). Каждое правило состоит из левой части, в которой указаны объекты метамодели визуального языка, и соответствующей правой части, представляющей конструкции текстовых языков, которые и являются шаблонами.

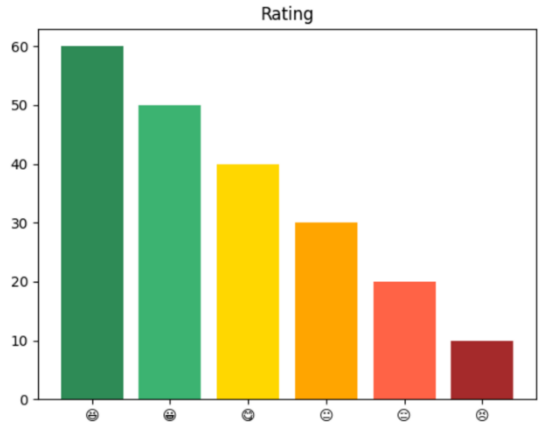
Языки Python и R являются предпочтительными языками программирования для целей визуализации данных. Python широко используется благодаря множеству подходящих библиотек, включая Matplotlib, Seaborn и Plotly, а также простому синтаксису. Хотя R уступает Python, он также обладает богатым арсеналом инструментов визуализации и остаётся популярным выбором в научных кругах. Таким образом, в качестве конструкций текстового языка в шаблонах имеет смысл использовать фрагменты кода на Python или R, содержащие вызовы библиотечных функций для визуализации данных.

После создания правил трансформации их можно применять к конкретным моделям для генерации кода пользовательской визуализации данных. Алгоритм генерации кода основан на обходе внутреннего представления моделей в виде графа.

Применив разработанные правила, можно получить программный код на соответствующем языке программирования для визуализации по правилам, заданным пользователем, разработавшим модель. Если правила трансформации разработаны для базового языка, они будут применимы ко всем моделям, разработанным с использованием всех DSL, созданных на основе этого базового языка. Это снижает трудоёмкость разработки, упрощает работу пользователей.

В качестве примера приведём код функции для построения гистограммы в Matplotlib, сгенерированный на основе приведённой выше модели по правилам трансформации, заданным пользователем в конструкторе правил трансформации, и представленным в онтологии, и результат его выполнения – диаграмму (рис. 9).

```
import matplotlib.pyplot as plt

groups = ['😆', '😀', '😒', '😐', '😔', '😣']
values = [60, 50, 40, 30, 20, 10]
colors = ['seagreen', 'mediumseagreen', 'gold',
'orange', 'tomato', 'brown']

plt.bar(groups, values, color=colors)

plt.title('Rating')

plt.show()
```



а *б*

Рис. 9. Сгенерированный по правилам трансформации код (*а*) и результат его выполнения (*б*)

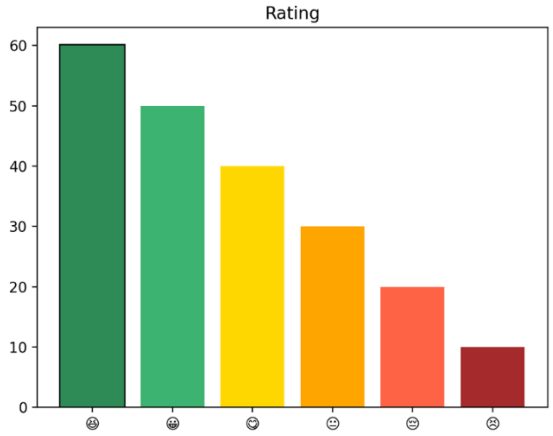
4. *Трансформация моделей и создание интерактивных визуализаций*

Если для языка, разработанного пользователем, описаны также и правила интерпретации, которые могут содержать правила взаимодействия пользователей с построенными визуализациями, то и код, генерируемый в ходе трансформации «Модель–Текст», будет поддерживать описанные правила интерпретации, визуализация становится интерактивной. Таким образом, при добавлении правил интерпретации, правила трансформаций автоматически расширяются данными возможностями без необходимости повторного описания правил трансформации пользователем DSM-платформы.

Код, сгенерированный для того же правила трансформации с учётом правила интерпретации, согласно которому при наведении мыши на конкретный столбец он должен выделяться, представлен на рис. 10 (*а*), а результат выполнения сгенерированного кода – на рис. 10 (*б*). Как видно из листинга, сгенерированный код действительно стал сложнее и теперь включает обработку события наведения мыши на столбцы гистограммы

посредством функции «on_hover». При наведении на столбец его прозрачность уменьшится, а также его граница будет покрашена в чёрный цвет.

```
%matplotlib notebook
import matplotlib.pyplot as plt

groups = ['😂', '😀', '😋', '😐', '😔',
'😣']
values = [60, 50, 40, 30, 20, 10]
colors = ['seagreen', 'mediumseagreen',
'gold', 'orange', 'tomato', 'brown']

fig, ax = plt.subplots()
bars = ax.bar(groups, values)

def on_hover(event):
    for bar in bars:
        if bar.contains(event)[0]:
            bar.set_alpha(1.0)
            bar.set_edgecolor('black')
            bar.set_linewidth(2)
        else:
            bar.set_alpha(0.8)
            bar.set_edgecolor('none')
            bar.set_linewidth(1)
    fig.canvas.draw_idle()

fig.canvas.mpl_connect('motion_notify_event',
on_hover)

plt.bar(groups, values, color=colors)

plt.title('Rating')

plt.show()
```



а *б*

Рис. 9. Сгенерированный по правилам трансформации код интерактивной визуализации (*а*) и результат его выполнения (*б*)

V. Обсуждение результатов

Предложенный в статье подход к разработке инструментов визуализации данных, основанных на DSM, позволяет преодолеть ряд ограничений существующих инструментов визуализации и предоставить пользователям возможность настраивать модели визуализации данных для разных предметных областей за счет создания специальных языков (DSL), не предъявляя при этом высоких требований к компетенциям пользователей в области программирования.

Возможность переиспользования созданных языков, разработанных правил трансформаций снижает трудоёмкость создания визуализаций: все разработанные метамодели языков, правила сохраняются в онтологии и могут быть использованы многократно либо через настройку параметров моделей, определённых в языках, либо через создание новых языков на основе ранее разработанных.

Полученные результаты показывают перспективность предложенного подхода к разработке языков и моделей различного назначения.

Ранее языковой инструментарий был использован для описания графических нотаций, разработки визуальных языков, предназначенных для проектирования информационных систем – для описания структур данных, алгоритмов и пр., для разработки и генерации кода имитационных моделей.

Как показали полученные результаты, представленные в данной статье, возможности применения данного подхода могут быть шире: языковой инструментарий позволяет решать задачи эффективной визуализации данных и может быть интегрирован в системы различного назначения, в частности, может быть полезен при создании аналитических систем и т. п.

VI. Выводы и заключение

Научная новизна подхода к разработке языкового инструментария – в использования онтологии на всех этапах жизненного цикла инструментария, что обеспечивает возможность гибкой настройки на различные предметные области и задачи пользователей, расширение функциональных возможностей системы.

Задача апробации подхода при создании инструментов визуализации данных решена:

- На основе выполненного анализа методов и инструментов визуализации данных были разработаны метамодели базовых языков для разработки стандартных типов диаграмм – основы для генерации пользовательских визуализаций данных, получаемых из различных источников, с настройкой на специфику предметных областей, в которых работают пользователи.
- Для иллюстрации возможностей настройки визуализации на потребности пользователей разработаны на основе базовых языков метамодели DSL, которые включают нестандартные элементы диаграмм, а также элементы, обеспечивающие интерактивную визуализацию; для созданных DSL построены правила трансформации – генерации кода, при выполнении которого создаются пользовательские визуализации с заданными характеристиками.

Практическая применимость данного подхода показана на простых примерах, но в перспективе предполагается расширение возможностей разработанных программных средств (исследовательского прототипа), в частности, возможностей интерактивной визуализации посредством интерпретации созданных моделей. Теоретической основой для разработки этого подхода является Vega-Lite [27] – грамматика высокого уровня, которая позволяет быстро специфицировать интерактивную визуализацию данных. Кроме того, предлагается расширить и возможности генерации новых DSL – реализовать возможность интеграции нескольких языков в качестве базовых, комбинировать их возможности в новом DSL.

Инструменты визуализации могут быть расширены средствами оценки построенных визуализаций на соответствие критериями, рекомендациям, которые предложены в [6, 7], что должно помочь пользователям при разработке избежать ошибок, создавать эффективные визуализации.

Список литературы

1. Sawicki J., Burdukiewicz M. VisQualdex: a Comprehensive Guide to Good Data Visualization // Scientific Visualization. 2023. Vol. 15. Number 1. P. 127–149. DOI: 10.26583/sv.15.1.11.
2. Alyahya S. M. Evaluating Computer Interactions and Infographics Usability: Analyzing Individual's Performance through Viewing Patterns // Scientific Visualization. 2023. Vol. 15. Number 5. P. 111–135. DOI: 10.26583/sv.15.5.10.
3. Шамсутдинова Т. М. Визуализация знаний в учебном процессе // Научная визуализация. 2023. Т. 15. № 1, С. 100–111, DOI: 10.26583/sv.15.1.09.
4. Немтинов В. А., Родина А. А., Борисенко А. В. Морозов В. В., Протасова Ю. В., Немтинов К. В. Комплексное использование различных программных сред для повышения уровня визуализации и восприятия информации // Научная визуализация. 2023. Т. 15. № 2. P. 1–10. DOI: 10.26583/sv.15.2.01.
5. Исаев Р. А., Подвесовский А. Г. Когнитивная ясность графовых моделей: подход к пониманию идеи и способ выявления влияющих факторов с использованием визуального анализа // Научная визуализация. 2022. Т. 14. № 4. P. 38–51. DOI: 10.26583/sv.14.4.04.
6. Midway S. R. Principles of Effective Data Visualization // Patterns, 2020., Vol. 1. Issue 9. Article 100141. DOI: 10.1016/j.patter.2020.100141.
7. Midway S. R., Brum J. R., Robertson M. Show and tell: approaches for effective figures // Limnology and Oceanography Letters (L&O Letters). 2023. Vol. 8. Issue 2. P. 213–219. DOI: 10.1002/lol2.10288.
8. Oral E., Chawla R., Wijkstra M., Mahyar N., Dimara E. From Information to Choice: A Critical Inquiry Into Visualization Tools for Decision Making // IEEE Transactions on Visualization and Computer Graphics. 2024. Vol. 30. No. 1. P. 359–369. DOI: 10.1109/TVCG.2023.3326593.
9. Qin X., Luo Y., Tang N., Li G. Making Data Visualization More Efficient and Effective: a Survey // The VLDB Journal. 2019. Vol. 29. No. 1. P. 93–117. DOI: 10.1007/s00778-019-00588-3.
10. Morgan R., Grossmann G., Schrefl M., Stumptner M, Payne T. VizDSL: A Visual DSL for Interactive Information Visualization // Proc. of the 30th International Conference "Advanced Information Systems Engineering" (CAiSE 2018). 2018. P. 440–455. DOI: 10.1007/978-3-319-91563-0_27.
11. Cepero García M. T., Montané-Jiménez L. G. Visualization to Support Decision-Making in Cities: Advances, Technology, Challenges, and Opportunities // Proc of the 8th International Conference in Software Engineering Research and Innovation (CONISOFT). 2020. P. 198–207. DOI: 10.1109/CONISOFT50191.2020.00037.

12. Shakeel H. M., Iram S., Al-Aqrabi H., Alsboui T., Hill R. A Comprehensive State-of-the-Art Survey on Data Visualization Tools: Research Developments, Challenges and Future Domain Specific Visualization Framework // IEEE Access. 2022. Vol. 10. P. 96581–96601. DOI: 10.1109/ACCESS.2022.3205115.
13. Zayakin V. S., Lyadova L. N., Lanin V. V., Zamyatina E. B., Rabchevskiy E. A. An Ontology-Driven Approach to the Analytical Platform Development for Data-Intensive Domains // Knowledge Discovery, Knowledge Engineering and Knowledge Management. IC3K 2021. Communications in Computer and Information Science. Springer, Cham. 2023. Vol. 1718. P. 129–149. DOI: 10.1007/978-3-031-35924-8_8.
14. Zayakin V., Lyadova L., Rabchevskiy E. Design Patterns for a Knowledge-Driven Analytical Platform // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). 2022. Vol. 34. No. 2. P. 43–56. DOI: 10.15514/ISPRAS-2022-34(2)-4.
15. Smeltzer K., Erwig M., Metoyer R. A transformational Approach to Data Visualization // Proc. of the International Conference on Generative Programming: Concepts and Experiences (GPCE 2014). 2014. P. 53–62. DOI: 10.1145/2658761.2658769.
16. Smeltzer K., Erwig M. A Domain-Specific Language for Exploratory Data Visualization // Proc. of the 17th ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences (GPCE 2018). 2018. P. 1–13. DOI: 10.1145/3278122.3278138.
17. Ledur C., Griebler D., Manssour I., Fernandes L. G. A High-Level DSL for Geospatial Visualizations with Multi-core Parallelism Support // Proc. of the IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). 2017, P. 298–304. DOI: 10.1109/COMPSAC.2017.18.
18. Ladenberger L., Dobrikov I., Leuschel M. An Approach for Creating Domain Specific Visualisations of CSP Models // Software Engineering and Formal Methods. SEFM 2014. Lecture Notes in Computer Science(). Vol. 8938. Springer, Cham. P. 20–35. DOI: 10.1007/978-3-319-15201-1_2.
19. Lyadova L., Sukhov A., Nureev M. An Ontology-Based Approach to the Domain Specific Languages Design // Proc. of the 15th IEEE International Conference on Application of Information and Communication Technologies (AICT2021). 2021. 6 p. DOI: 10.1109/AICT52784.2021.9620493.
20. Kulagin G., Ermakov I., Lyadova L. Ontology-Based Development of Domain-Specific Languages via Customizing Base Language // Proc. of the 16th IEEE International Conference on Application of Information and Communication Technologies (AICT2022). 2022. 6 p. DOI: 10.1109/AICT55583.2022.10013619.
21. Lyadova L., Ermakov I., Lanin V., Proskuryakov K. Approach to the Development of Ontology-Driven Language Toolkits Based on Metamodeling // Proc. of the IEEE 17th International Conference on Application of Information and Communication Technologies (AICT2023). 2023. 6 p. DOI: 10.1109/AICT59525.2023.10313152.
22. Matta N., Ermine J. L., Aubertin G., Trivin J. Y. Knowledge Capitalization with a Knowledge Engineering Approach: The Mask Method // Knowledge Management and Organizational Memories. Springer, Boston, MA. 2002. P. 17–28. DOI: 10.1007/978-1-4615-0947-9_2.
23. Zelazny G. The Say It with Charts Complete Toolkit. New York: McGraw-Hill Professional, 2006. 312 p.
24. Kirk A. Data Visualization: A Successful Design Process. Birmingham: Packt Publishing Ltd, 2012. 189 p.
25. Kahani N., Bagherzadeh M., Cordy J. Survey and Classification of Model Transformation Tools // Software & Systems Modeling. 2019. Vol. 18. P. 2361–2397. DOI: 10.1007/s10270-018-0665-6.
26. Ding J., Lu J., Wang G., Ma J., Kiritsis D., Yan Y. Code Generation Approach Supporting Complex System Modeling Based on Graph Pattern Matching // IFAC-PapersOnLine. 2022. Vol. 55. Issue 10. P. 3004–3009. DOI: 10.1016/j.ifacol.2022.10.189.
27. Satyanarayan A., Moritz D., Wongsuphasawat K, Heer J. Vega-Lite: A Grammar of Interactive Graphics // IEEE Transactions on Visualization and Computer Graphics. 2016. Vol. 23. No. 1. P. 341–350. DOI: 10.1109/TVCG.2016.2599030.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-315-317

Использование средств компьютерной графики в проектировании освещения метро
The use of computer graphics tools in the design of subway lighting

Данила Казаков[1], Денис Макаров[2]
[1] *ООО ТПК «Вартон», Москва, Россия*
[2]*Национальный исследовательский университет «МЭИ», Москва 111250, Россия*

Danila Kazakov[1], Denis Makarov[2]
[1]*LLC TPC "Varton", Moscow, Russia*
[2] *National Research University «Moscow Power Engineering Institute», Moscow 111250, Russia*

Аннотация. В статье рассматривается использование инструментов программ компьютерной графики для решения задач в области светотехники. Авторы описывают, как различные функции и возможности графических редакторов, таких как 3D-моделирование, конструирование, освещение и визуализация, могут быть применены при проектировании и анализе осветительных систем. Обсуждаются преимущества применения компьютерной графики по сравнению с традиционными методами, включая повышение точности, скорость и наглядность результатов. Приводятся конкретные примеры использования графических инструментов для решения задач расчета освещенности. Статья будет интересна специалистам в области светотехники, архитектуры и дизайна, а также разработчикам программного обеспечения для компьютерной графики.

***Ключевые слова*:** Метро, естественное освещение, метро с естественным освещением, световоды, системы освещения, энергоэффективность, компьютерная графика, программы, программные пакеты.

Аннотация. The article discusses the use of computer graphics software tools to solve problems in the field of lighting engineering. The authors describe how various functions and capabilities of graphic editors, such as 3D modeling, construction, lighting and visualization, can be applied in the design and analysis of lighting systems. The advantages of using computer graphics in comparison with traditional methods, including improved accuracy, speed and visibility of results, are discussed. Specific examples of the use of graphical tools for solving problems of calculating illumination are given. The article will be of interest to specialists in the field of lighting engineering, architecture and design, as well as developers of computer graphics software.

***Ключевые слова*:** Metro, natural lighting, metro with natural lighting, light guides, lighting systems, energy efficiency, computer graphics, programs, software packages.

I. Введение

Специалисты в области светотехники часто сталкиваются с необходимости визуализации своих расчётов. В век компьютерных технологий это стало реальностью. Существует не мало инженерных программ, которые могут выводить в понятном для представления виде графические данные для удобного анализа [1]. Возможности программных продуктов позволяют без живых экспериментов и сложных расчетов быть уверенным в возможной реализации проектируемого объекта. Пример такого объекта - станция метро с световодами естественного света, которая будет рассмотрен в рамках работы [2].

II. Постановка задачи

В процессе работы над довольно нестандартным проектом освещения возникла необходимость использования различных программ, косвенно или на прямую связанных со светотехникой. Поставлена задача подобрать по необходимым для решения отдельных задач проекта критериям инженерные программы компьютерной графики.

III. Теория

В первую очередь для разработки концепций освещения станции метро, работающей, как при естественном, так и при искусственном освещении благодаря световодам естественного света, была использована известная программа 3Д моделирования - 3ds max. Помимо удобного создания трехмерных моделей, интерфейс содержит множество инструментов для освещения. В дополнение к этому есть возможность подключить плагин Corona Render, который значительно поднимает уровень визуализаций, делая их максимально реалистичными.

Однако визуализации подойдут только для создания красивой картинки, наброска для будущего проекта. Все светотехнические расчёты необходимо выполнять в специализированных программах. Одна из главных таких программ - DIALux Evo. В ней так же возможно получить световую картину, но уже с технической точки зрения. Проведя расчёт методом радиосити, в результате можно получить распределение освещенности и яркости в фиктивных цветах, что является основной функцией для анализа освещенности.

DIALux подходит и для расчёта естественного света, но задачу затрудняет некорректное пропускание света сквозь конструкции из-за сложной 3Д модели и систем отражения внутри световода, так как все поверхности в программе по умолчанию считаются диффузными. Благодаря импорту различных файлов светильников (ies-файлы) можно вместо 3Д моделей создать световод-источник и загрузить его как ies.

В решении задачи создания световода, как источника, поможет TracePro. Солнце возможно сымитировать обычным rау-файлом с соответствующими характеристиками. После трассировки лучей через световод получаем кривую силы света и экспортируем в нужном формат с нужными размерами источника. В этой же программе были посчитаны все характеристики светильников, работающих в вечернее время суток.

Удобной программой для создания твердотельных конструкций светильника и световода послужил SolidWorks. Простота использования, возможность вывода чертежей, экспорт в любых форматах дают множество преимуществ в работе.

Relux – схожий с DIALux программный продукт, но более продвинутый в расчёте естественного освещения. Благодаря ему можно получить данные о процентном попадании солнечного света на нужную расчётную плоскость в зависимости от места, погоды и других аспектов.

Совместное использование всех перечисленных программ, дополняющих друг друга своим уникальным функционалом, даёт возможность воплотить в жизнь задуманный проект.

IV. Результаты экспериментов

Пройдя через все этапы работы с программами компьютерной графики, в рамках проекта было получено множество графических данных. В 3ds max сделаны визуализации, и на основе концепции проведены расчёты в DIALux Evo (рис. 1).

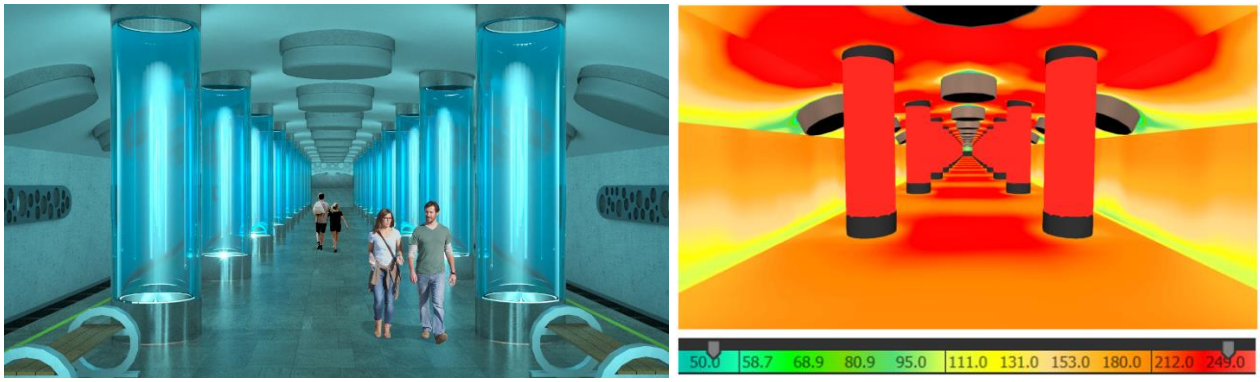


Рис. 1. Визуализация в 3ds max и результат светотехнического расчёта в DIALux evo

Благодаря Relux был определены положение солнца на небосводе, доля попадания солнечного света на станцию и другие данные, благодаря которым удалось провести расчёт освещенности для каждого часа работы метро. По полученным данным построен график и отмечены уровни разброса освещенности (-10%, +20%) от нормированного уровня освещенности (СП 32-105-2004. Свод правил по проектированию и строительству. Метрополитены.) [3] (Рис.2). Из графика можно сделать вывод, что станция в состоянии освещаться 6 часов полностью за счёт естественного освещения. В соответствии с циркадными циклами, этим часам соответствует лучшая координация, время быстрой реакции и повышенное внимание.

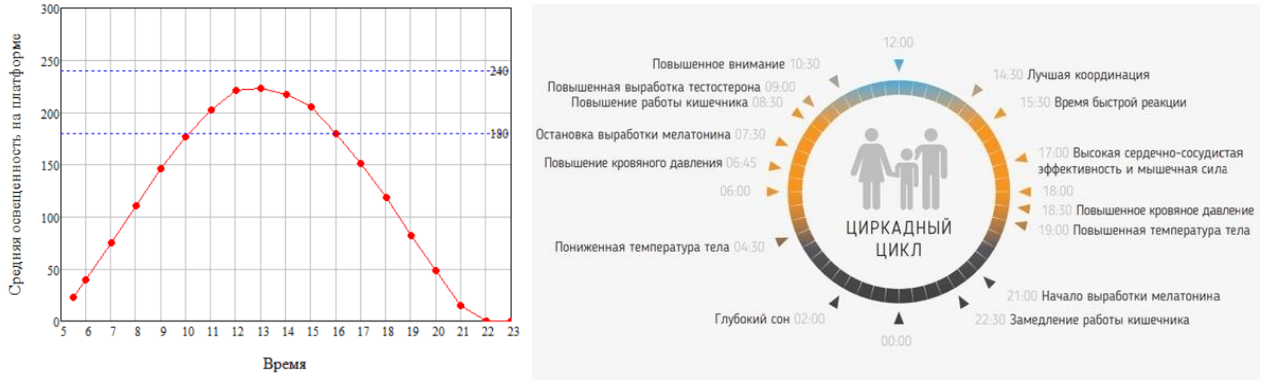


Рис. 2. График зависимости средней освещенности платформы от времени работы метро, синхронизация с циркадными циклами

V. Выводы и заключение

Таким образом, в статье показано, что инструменты компьютерной графики могут быть эффективно применены для решения широкого спектра задач в области светотехники. Было изучены и проанализированы несколько конкретных программных пакетов, которые нашли своё применение для моделирования, расчётов и визуализации станции метро с световодами естественного света.

Список литературы

1. Будак В.П., Макаров Д.Н. Компьютерная графика с приложением в светодизайн // Учебник для высших учебных заведений / Москва, 2022.-264 с.
2. Соловьёв А.К. Естественное и совмещенное освещение // Справочная книга по светотехнике / Под. Ред. Ю.Б. Айзенберга, Г.В. Бооса 4-е изд. перераб. и доп. М.: - С 575-579.
3. СП 32-105-2004. Свод правил по проектированию и строительству. Метрополитены. // docs.cntd.ru [сайт]. – URL: https://docs.cntd.ru/document/1200035937/titles/BSK0PF (дата обращения: 10.07.2024).

УДК 004.93, 531.715.2, 531.717.1, 531.717.82, 62.5, 621.7, 67.05
DOI: 10.25206/978-5-8149-3873-2-2024-318-321

Разработка и применение высокопроизводительных алгоритмов машинного зрения на ПЛИС
Development and application of high-performance machine vision algorithms on FPGA

Ю.А. Сальков[1], Д.Л. Сапрыкин[1], О.Н. Третьякова[1,2], Д.Н. Тужилин[1]
[1] ООО «Лаборатория промышленных исследований» группы компаний НПЦ «Лазеры и аппаратура ТМ», Москва, 124498, Россия
[2] ФГБОУ ВО Московский авиационный институт (национальный исследовательский университет) МАИ, Москва, 125993, Россия

Yu.A. Salkov[1], D.L. Saprykin[1], O.N. Tretiyakova[1,2], D.N. Tuzhilin[1]
[1]*LLC "PROMIS LAB" of the group of companies "Lasers and Equipment TM", Moscow, Russia*
[2]*Moscow aviation Institute (national research university) MAI, Moscow, Russia*

Аннотация. Статья посвящена деталям реализации алгоритмов машинного зрения на ПЛИС. Алгоритмы были реализованы при помощи инструмента высокоуровневого синтеза Vitis HLS. Главным параметром при реализации алгоритмов является их быстродействие, так как необходимо с минимальной задержкой получать новые данные из приходящего кадра.

***Ключевые слова*:** система на кристалле, ZYNQ, методы машинного зрения, обработка видео, фильтрация изображения на FPGA, технология приборостроения, скользящее среднее, центр масс, HLS.

Abstract. The paper is devoted to the details of implementation of machine vision algorithms on FPGA. The algorithms were implemented using the Vitis HLS high-level synthesis tool. The main characteristic in the realization of the algorithms is speed, as it is necessary to process the incoming video stream with minimum delay.

Keywords: system-on-chip, ZYNQ, machine vision techniques, video processing, FPGA image filtering, instrumentation technology, moving average, mass center, HLS.

I. Введение

В процессе размерной обработки деталей полупроводниковых приборов необходимо точно фокусировать луч лазера. Для подстройки фокуса лазера прямо в процессе резке была разработана система оптического слежения (СОС) [1], которая состоит из лазерного диода, линейной камеры и контроллера. Луч лазерного диода под углом светит на поверхность обрабатываемой детали и отражается от нее. Отраженный луч лазерного диода попадает на линейную камеру, тем самым формируя пик на кадре с камеры.

Кадр с камеры поступает в контроллер СОС, где сначала проходит фильтрацию фильтром «скользящее среднее», после чего на отфильтрованном кадре ищется пик по заданным параметрам, формулой поиска центра масс.

Главная задача контроллера СОС заключается в поиске пика на кадре и, в зависимости от положения пика, генерации управляющих сигналов для серводрайвера. Так как подстройка фокуса происходит непосредственно в процессе резки, контроллер должен с минимальной задержкой принимать и обрабатывать данные. Для разработки контроллера была взята система на кристалле Xilinx Zynq, которая в включает в себя как часть с программируемой логикой (PL-часть), так и процессорную часть (PS-часть). Части соединены между собой общей памятью, что позволяет быстро передавать данные из PS-части в PL-часть и наоборот. Реализация алгоритмов на ПЛИС позволяет добиться большей производительности в сравнении с привычными архитектурами вычислительных систем (таких как x86, ARM или RISC-V), так как при реализации алгоритмов на ПЛИС, микросхема становится узкоспециализированным вычислителем для решения конкретных задач.

II. Постановка задачи

Основная задача исследования – разработка и реализация эффективного алгоритма фильтрации и поиска пика на ПЛИС. Существует немного трудов по теме реализации алгоритмов машинного зрения на ПЛИС [2-4]. Основным отличием данной статьи является описание реализации алгоритмов при помощи средства высокоуровневого синтеза Vitis HLS.

III. Теория

При разработке контроллера СОС был использован инструмент высокоуровневого синтеза Vitis HLS [5]. Инструменты высокоуровневого синтеза позволяют разрабатывать отдельные функциональные блоки при помощи написания кода на высокоуровневых языках (C++, Python и т.д.), а не на языках описания аппаратуры (Verilog, HDL и т.д.). Данные инструменты дают возможность разработчику функционального модуля абстрагироваться от уровня регистровых передач (RTL) и разрабатывать код на уровне работы с данными.

Инструменты высокоуровневого синтеза имеют большой потенциал для коммерческой разработки каких-либо обработчиков данных. В частности, Vitis HLS имеет обширный набор библиотек, для различных задач. Например, Vitis-HLS имеет адаптированную библиотеку OpenCV, таким образом, разработка функциональных блоков, связанных с обработкой видео, становится на много проще, но в данной разработке функции OpenCV не применялись, из-за специфики работы контроллера СОС.

IV. Разработка алгоритма фильтрации

Как уже упоминалось выше, контроллер СОС должен быстро принимать и обрабатывать приходящий видеопоток. Перед поиском пика на кадре, кадр необходимо отфильтровать. Для фильтрации кадра был взят модифицированный алгоритм «скользящее среднее» [4]. Данный алгоритм был выбран из-за того, что пик на кадре является явно выраженным и необходимо просто убрать лишние шумы, для правильного определения центра масс пика.

На вход фильтру подается изображение разрешением 2*2048 пикселей, кадр является черно-белым, каждый пиксел кодируется 8 битами. Первая строка кадра записывается в буфер кадра без изменений. С начала прихода второй строки кадра начинается фильтрация с помощью скользящего окна, ширина окна задается пользователем, а высота окна фиксирована – 2 пикселя. С начала фильтрации окно заполнено нулями, из-за этого теряется часть информации в начале изображения, что не критично. Отфильтрованная вторая строка изображения записывается в буфер кадра, таким образом в одном кадре присутствуют, как и «сырые» данные, так и отфильтрованные, за счет этого в визуализирующей программе можно отобразить сразу оба графика. На рис. 1 изображен алгоритм работы фильтра «скользящее среднее».

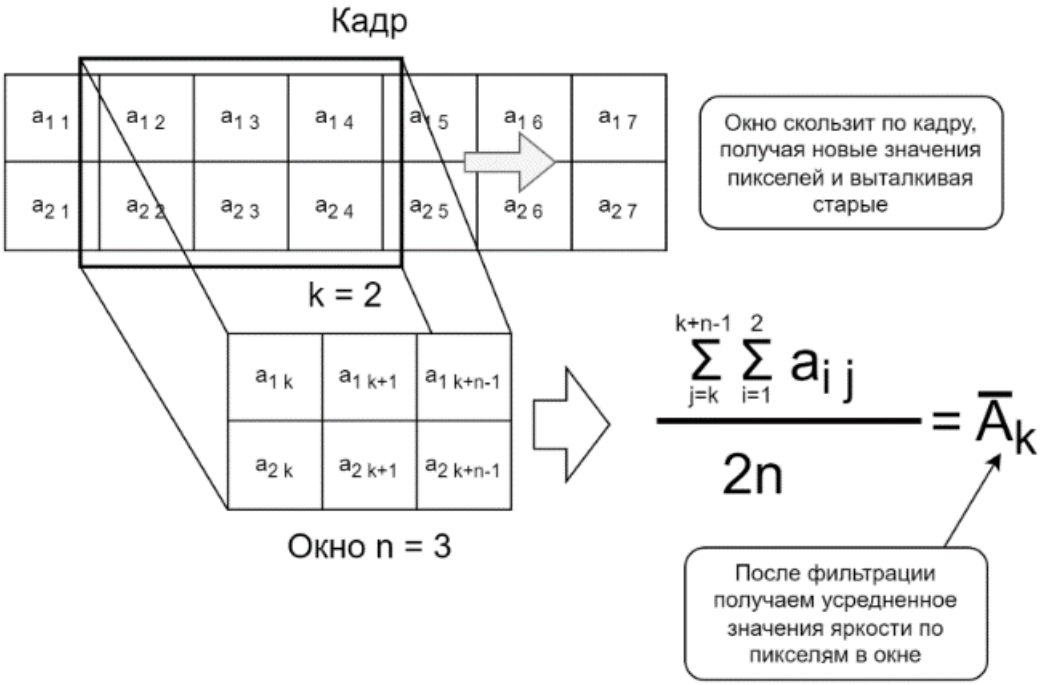


Рис. 1. Алгоритм и формула фильтрации.

В итоге удалось разработать оптимизированный алгоритм для фильтрации приходящего видеопотока. Фильтр выдает результат сразу как получил полный кадр.

V. Разработка алгоритма поиска пика на кадре

Для поддержания быстродействия контроллера было принято решение реализовать алгоритм поиска пика вместе с алгоритмом фильтрации видеопотока. Таким образом в одном функциональном блоке реализованы два алгоритма, работающие друг за другом. После операции фильтрации окна происходит обновление данных о пике, а сразу после получения всего кадра происходит вычисление нового положения пика. Задержка между получением полного кадра и вычислением нового положения пика является минимально возможной.

Сам алгоритм поиска центра масс пика представляет собой формулу поиска центра масс, где вместо массы была взята яркость конкретного пикселя (mi), а вместо координат точки был взят индекс пикселя (ri). Алгоритм поиска центра масс представлен на рис. 2.

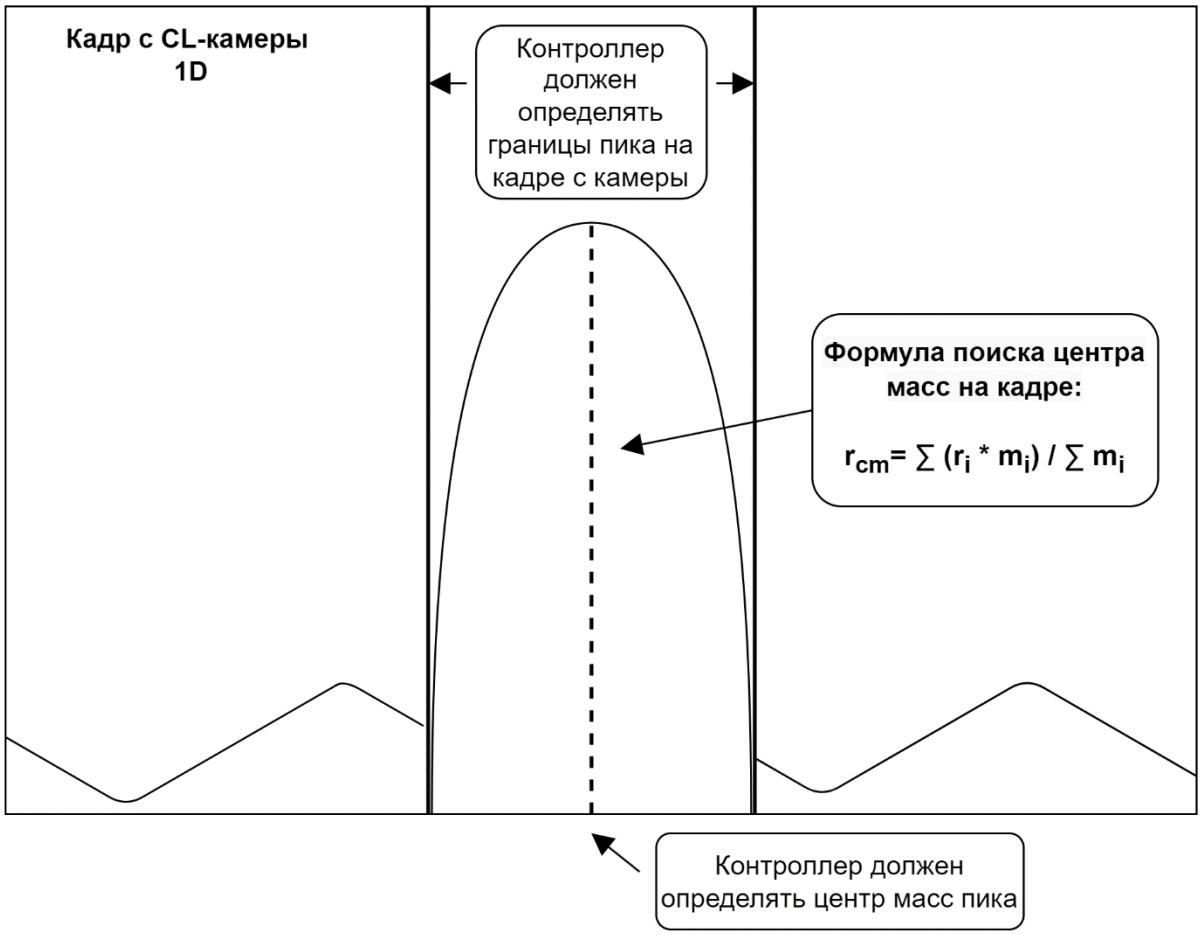


Рис. 2. Алгоритм поиска центра масс.

Поиск пика производится на всех ненулевых отфильтрованных пикселях кадра. Для поиска пика пользователь задает параметры, такие как: максимальная и минимальная ширина пика, порог яркости пикселей пика, максимальная яркость пика и т.д.. Сразу после получения среднего значения с окна усреднения происходит обновление данных, если яркость пикселя равна нулю, но до него пиксели были не нулевые, выполняется расчет масс центра пика и проверка полученных данных о пике с заданными параметрами. Первый пик, удовлетворяющий установленным параметры, считается подходящим и алгоритм заканчивает свою работу и на выход функционального блока поступают данные о положении и размере пика.

В результате разработки алгоритма поиска пика удалось реализовать оптимизированный и высокопроизводительный алгоритм, выдающий результат почти сразу после получения полного кадра. Как и с реализацией алгоритма фильтрации производительность алгоритма также упирается в скорость передачи данных от камеры.

VI. Результаты экспериментов

Для испытания работы контроллера СОС был разработан испытательный стенд, представленный на рис. 3. Лазерный диод (4) светит на линейную камеру (1), видеопоток с линейной камеры обрабатывается контроллером СОС и управляющие сигналы подаются на серводрайвер (2). Серводрайвер управляет Z-осью (3), к которой

прикреплена линейная камера. Таким образом, при смещении пика на кадре от установленного положения контроллер СОС подстраивает положение Z-оси таким, образом, чтобы центр масс пика был в правильной позиции. Алгоритмы фильтрации и поиска пика позволили с достаточным быстродействием находить пик на кадре с камеры и формировать управляющие сигналы для мотора серводрайвера.

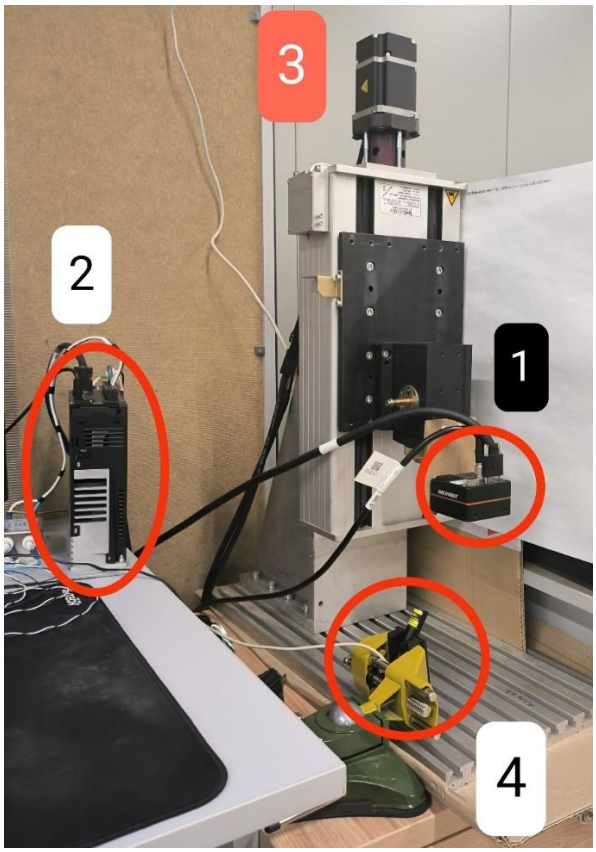


Рис. 3. Испытательный стенд.

VIII. Выводы и заключение

В результате реализации алгоритмов был разработан функциональный блок фильтрации и поиска пика на видеопотоке. Данный блок обрабатывает приходящий кадр с минимальным временем 42 мкс, данное время включает в себя время получения кадра. Функциональный блок имеет гибкие настройки для фильтрации и поиска пика на кадре.

Список литературы

[1]. Кондратенко В.С., Сапрыкин Д.Л., Третьякова О.Н., Тужилин Д.Н. Разработка системы автоматического управления подстройкой фокуса для технологии лазерной микрообработки материалов // Приборы. 2022. № 4. С.26–31.

[2]. Calliope-Louisa Sotiropoulou, Liberis Voudouris, Christos Gentsos, Athanasios M. Demiris, Nikolaos Vassiliadis, Spyridon Nikolaidis Real-Time Machine Vision FPGA Implementation for Microfluidic Monitoring on Lab-on-Chips // IEEE Transactions on Biomedical Circuits and Systems. - 2014. - №Volume: 8. – p.p. 268 - 277.

[3]. Zhonghua Zhou The FPGA Implementation of Computer Vision Algorithm // UC Riverside Electronic Theses and Dissertations. - 2014.

[4]. Орлова И.А., Устюков Д.И., Ефимов А.И. Особенности реализации пространственных фильтров изображений на FPGA // Известия ТулГУ. Технические науки. 2018. №2. С.Y195-206.

[5]. Ryan Kastner, Janarbek Matai, and Stephen Neuendorffer. Parallel Programming for FPGAs [Электронный ресурс] // Kastner Research Group URL: https://kastner.ucsd.edu/wp-ontent/uploads/2018/03/admin/pp4fpgas.pdf (дата обращения: 07.07.2024)

УДК 004.93, 531.715.2, 531.717.1, 531.717.82, 62.5, 621.7, 67.05
DOI: 10.25206/978-5-8149-3873-2-2024-322-326

Разработка методики визуализации и анализа дефектов в аддитивных технологиях
Development of methodology for visualization and analysis of defects in additive technologies

А.В. Суворов[1], О.Н. Третьякова[1,2], Д.Н. Тужилин [1], А.А. Шамордин[1]
[1] *ООО «Лаборатория промышленных исследований» группы компаний НПЦ «Лазеры и аппаратура ТМ», Москва, 124498, Россия*
[2] ФГБОУ ВО Московский авиационный институт (национальный исследовательский университет) МАИ, Москва, 125993, Россия

A.V. Suvorov[1], O.N. Tretyakova[1, 2], D.N. Tuzhilin[1], A.A. Shamordin[1]
[1]LLC “PROMIS LAB” of the group of companies “Lasers and Equipment TM”, Moscow,124498, Russia
[2]Moscow aviation Institute (national research university) MAI, Moscow, Russia

Аннотация. В данной работе рассматривается вопрос визуализации и анализа дефектов, возникающих в процессе селективного лазерного сплавления в целях упрощения подбора технологических параметров работы установки аддитивного производства. В статье представлен подход по визуализации и обнаружению дефектов, возникающих на этапе разравнивания слоя металлического порошка. Реализованный подход способен обнаружить дефекты на начальных этапах производства, что позволяет остановить процесс печати заведомо некачественного изделия. Рассмотрена и реализована методика программно-аппаратного обнаружения дефектов. Представлены преимущества и недостатки проделанной работы и получившихся результатов.

Ключевые слова: SLM, машинное зрение, распознавание дефектов, оптико-электронные приборы.

Abstract. This paper deals with the issue of visualization and analysis of defects arising in the process of selective laser fusion in order to simplify the selection of technological parameters of the additive manufacturing plant. The paper presents an approach for visualization and detection of defects arising at the stage of flattening of the metal powder layer. The realized approach is able to detect defects at the initial stages of production, which allows to stop the process of printing a known low-quality product. The methodology of hardware-software defect detection is considered and realized. Advantages and disadvantages of the work done and the results obtained are presented.

Keyword: SLM, machine vision, defect recognition, optoelectronic devices.

I. Введение

Данная работа является продолжением и развитием наших исследований [1-4,8] по разработке программного обеспечения и лазерного технологического оборудования для практической реализации на отечественных предприятиях технологии послойного лазерного сплавления или SLM (Selective Laser Melting) технологии, также известной как 3D-печать. В основу SLM технологии положен процесс производства изделий путем послойного плавления металлического порошка с использованием лазерного излучения. На первом этапе на рабочую поверхность равномерно наносится тончайший слой металлического порошка, на втором этапе лазерный луч точечно нагревает и плавит материал согласно предопределенным данным из слоя 3D модели. На каждом из двух этапов в результате некорректного подбора технологических параметров возможно возникновение дефектов. На этапе нанесения порошка, далее разравнивание могут возникать продольные полосы ортогональные одной их осей рабочей поверхности (в зависимости от направления разравнивания), в результате повреждения ракеля. На основании этих данных был разработан программный блок, реализующий визуализацию и обнаружение дефектов разравнивания.

II. Общая информация

Программный код написан на языке C++ 17 с применением программной платформы QT 5.15, для работы с промышленной камерой использована библиотека производителя MvCameraControl, для упрощения реализации алгоритмов распознавания и работы с кадром применяется библиотека OpenCV. Для визуализации видеопотока с камеры, а также дефектов разравнивания используются встроенные механизмы QT, предоставляющие удобный интерфейс для работы со сценой OpenGL. В программе реализована поддержка работы с несколькими камерами одновременно. Ввиду кроссплатформенной природы использованных технологий программа может работать, как на ОС Windows, так и Linux.

III. Общая методика распознавания дефектов

Для стабильного распознавания дефектов не всегда достаточно только программных средств иногда необходимо влиять на интенсивность и направление освещённости в области рабочей поверхности. Для решения данной проблемы в конфигурации под которую велась разработка, имеется несколько модулей подсветки, которыми можно управлять из нашей программы, такое решение предоставляет возможность физически влиять на контрастность дефектов на видеопотоке, за счёт изменение угла падения света на рабочую поверхность, благодаря чему возрастает и эффективность программных методов.

Физическое устройства лазерной установки, под которую велась, разработка не позволяет разместить камеру непосредственно над рабочей поверхностью, однако такое расположение является необходимым для точного определения координат обнаруженных дефектов. Чтобы скомпенсировать несоосность камеры с рабочей поверхностью в программе присутствует механизм компенсации перспективы, который позволяет наблюдать поверхность под нужным ракурсом рис. 1, при этом обрезаются области не обязательные для обработки, благодаря чему дополнительно уменьшается размер анализируемого изображения, что в свою очередь благоприятно влияет на скорость работы алгоритмов.

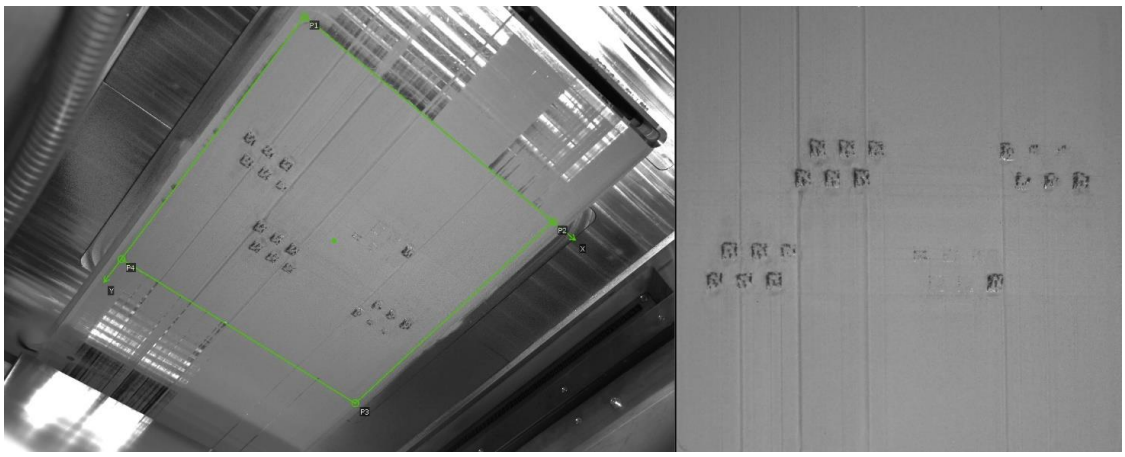


Рис. 1. Программное исправление перспективы.

Распознавание дефектов происходит в строго определённый момент процесса, а именно после разравнивания, когда порошок должен ровным слоем закрывать рабочую поверхность. Чтобы определить момент, когда необходимо включать детектирование, была налажена связь по протоколу TCP с программой, предоставляющей информацию о текущем печатаемом слое, описанной в работах [1-4], программе управляющей последовательностью действий 3D-печати. При поступлении команды на поиск дефектов наша программа последовательно запоминает n изображений с n камер (последовательная обработка в данном случаи обусловлена необходимостью переключать конфигурации подсветки). После завершения получения изображений к ним применяются фильтры и алгоритмы распознавания, по окончанию работы которых происходит отправка найденных дефектов в управляющую программу.

Таким образом, для успешного обнаружения дефектов создана методика, при которой оператору необходимо программно исправить перспективу, настроить конфигурацию подсветки, а также определить конвейер обработки видеопотока.

IV. Конвейер обработки видеопотока

Разнообразие возможных материалов порошка, используемых в технологии селективного лазерного сплавления, диктует необходимость давать оператору возможность настраивать не только параметры самого детектора, но и типы, и порядок фильтров, используемых при обработке входного изображения. Чтобы предоставить пользователю такую возможность был внедрён механизм конвейерной обработки видеопотока рис. 2, по методике, предложенной в статье [4].

Данный механизм оперирует двумя типами примитивов. Первый тип, детекторы – алгоритмы, реализующие поиск дефектов на поступающем кадре. Второй тип, фильтры, осуществляющие предварительную обработку кадра перед его отправкой в детектор.

В программе распознавания дефектов реализована гибкая, настройка фильтров и детекторов, за счёт интерактивной визуализации влияния изменения параметров на состояние кадра рис. 3. Важную роль в настройке играет порядок расположения фильтров в конвейере, так как на каждый фильтр поступает изображение, обработанное предыдущим фильтром (см. рис. 2).

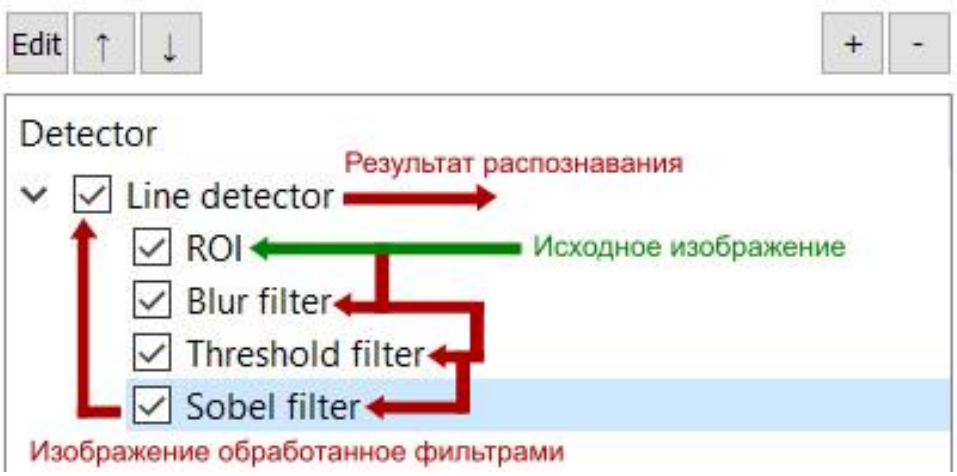


Рис. 2. Конвейер обработки видеопотока.

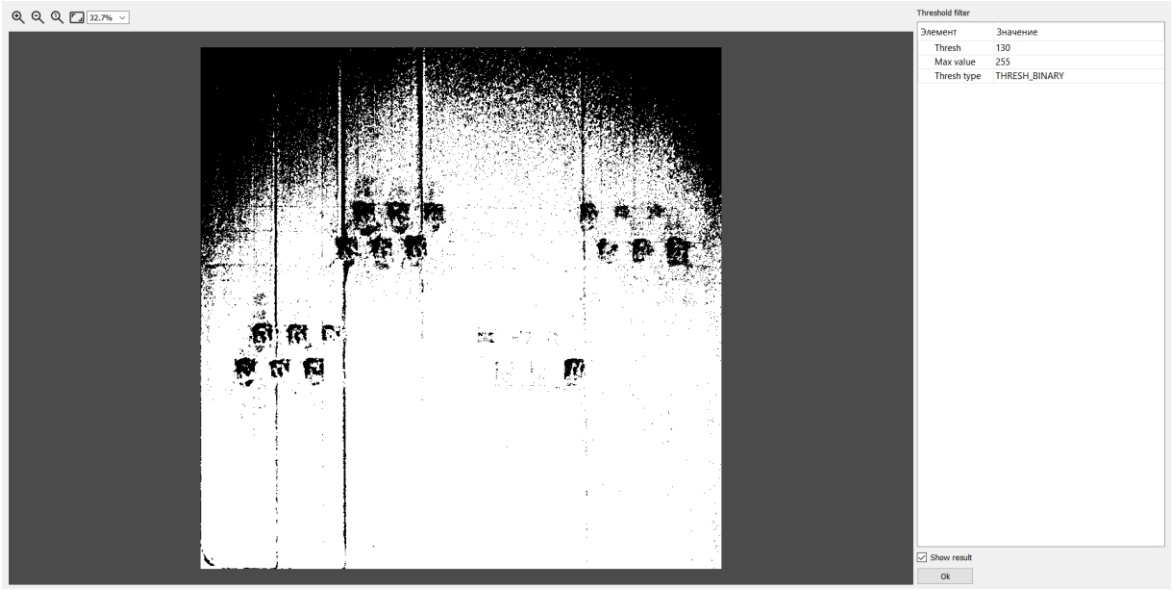


Рис. 3. Пример настройки фильтра бинаризации.

За счёт такой обработки входного кадра достигается возможность нивелировать разницу в физических свойствах порошка, влияющих на его отображение на камере.

V. Распознавание полос

Как уже было сказано ранее, полосы могут возникнуть в результате прохода повреждённого ракеля по рабочей поверхности. Для детектирования линий, оставляемых ракелем, был применён популярный метод на основе преобразований Хафа [5–7], остановимся на принципах его работы с примитивами типа прямая несколько подробнее.

Для начала выразим прямую в полярной системе координат. Следовательно, линейное уравнение можно записать

$$y = \left(-\frac{cos(\theta)}{sin(\theta)}\right)x + \left(\frac{r}{sin(\theta)}\right)$$

Для каждой точки (x_0, y_0) мы можем определить семейство линий, проходящих через эту точку, следующим образом

$$r\theta = x0 \cdot cos(\theta) + y0 \cdot sin(\theta)$$

Это означает, что каждая пара (rθ, θ) представляет каждую линию, проходящую через (x_0, y_0). Если для заданного (x_0, y_0) мы построим семейство линий, которые проходят через него, мы получим синусоиду. Если проделать ту же операцию для всех точек изображения и окажется, что кривые двух разных точек пересекаются в плоскости (θ r), это означает, что обе точки принадлежат одной прямой.

После обхода изображения данным алгоритмом на выходе мы получаем уравнение прямой в полярных координатах, благодаря которому реализована визуализация дефекта на кадре с камеры рис. 4.

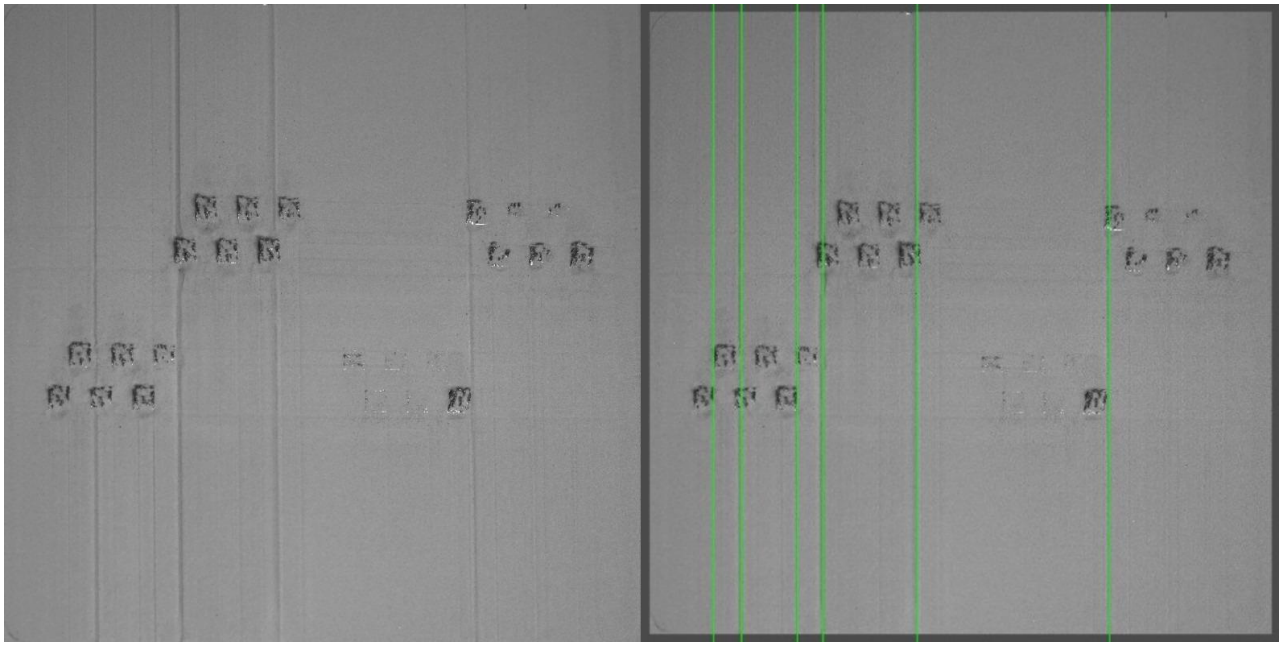


Рис. 4. Обнаруженные дефекты.

VI. Преимущества и недостатки

Аддитивная SLM технология применяется в первую очередь для создания изделий со сложной геометрией, выполнить которую классическими методами механообработки либо кратно сложнее, либо вовсе невозможно. Но мало просто напечатать деталь, она как минимум должна соответствовать механическим свойствам такого же изделия, выполненного классическими методами. Для достижения требуемых механических параметров выходного образца необходимо отладить процесс печати, путём подбора параметров технологического режима [8], задача кратно усложняется с использованием нескольких лазеров, так как все они должны печатать одинаково. Таким образом, на подбор оптимального режима может уйти существенный промежуток времени.

Благодаря предложенному нами программному блоку распознавания дефектов можно существенно сократить время на подбор параметров печати, что достигается за счёт определения дефектов, на любой стадии печати. Это позволяет сразу завершать длительный технологический процесс создания заведомо бракованного изделия. Определение дефектов на ранних стадиях роста, также позволяет сэкономить металлический порошок, что в свою очередь способствует удешевлению процесса производства.

Основными же недостатками представляемой работы, является усложнение процесса на этапе подготовки оборудования к печати, так как оператору необходимо настроить ряд параметров, среди которых ранее описанные: фильтры, детекторы, а также конфигурация подсветки и перспектива.

VII. Выводы и заключение

В результате создания программного блока распознавания дефектов, удалось добиться сокращения общего времени, затрачиваемого на подбор технологических параметров для печати изделия, а также сокращения себестоимости данного процесса за счёт экономии металлического порошка, посредством остановки процесса при возникновении дефектов.

VIII. Благодарности

Авторы выражают благодарность руководству группы компаний «Лазеры и Аппаратура ТМ» за помощь в материально-техническом обеспечении проведения экспериментальных исследований и моделирования рассматриваемого процесса.

Список литературы

[1]. Molotkov A.A., Tretiyakova O.N. On possible approaches to visualizing the process of selective laser melting // Scientific Visualization. 2019. Vol 11. № 4. PP. 1-12.
[2]. Molotkov A.A., Tretiyakova O.N., Tuzhilin D. N. About development and application of a software platform for machine vision for various laser technologies // Scientific Visualization 2022. №5. P.P.108-118.
[3]. Молотков А.А., Третьякова О.Н. **Применение методов машинного зрения и математического моделирования для разработки технологий создания электронных приборов //** Приборы. 2022. № 4. С.55-58.
[4]. Молотков А. А., Сапрыкин Д.Л., Третьякова О. Н., Тужилин Д.Н. Разработка комплекса программ для создания промышленного лазерного технологического оборудования // Приборы. 2022. № 5. С.15-22.
[5]. Кудрина М.А. Использование преобразования Хафа для обнаружения прямых линий и окружностей на изображении // Известия Самарского научного центра Российской академии наук, т. 16, №4(2), 2014
[6]. Кудрина М.А., Кирьяков С.А., Кудрин К.А., Дегтярева О.А. Векторизация растровых изображений с использованием преобразования Хафа // "Надежность и качество" - 2013 - №1 - P. 337-340
[7]. Матвеев И.А., Чинаев Н.Н. Выделение зрачка на изображении глаза преобразованием Хафа границ компонент связности // "Графикон". - 2014 - №25 - P. 115-118
[8]. Молотков А.А., Третьякова О.Н. Отработка технологических режимов при создании SLM технологии. // Приборы. 2023. № 8. С. 44-47.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-327-346

Стохастическая семантика больших данных
Big data stochastic semantics

П. А. Васёв1, Д .В. Манаков2
1ИММ УрО РАН им. Н.Н. Красовского, ул. Софьи Ковалевской 16, г. Екатеринбург, 620108, Россия
2ИММ УрО РАН им. Н.Н. Красовского, ул. Софьи Ковалевской 16, г. Екатеринбург, 620108, Россия

P. A. Vasev1, D. V. Manakov2
[1]1N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, 16 S. Kovalevskaya Str., Ekaterinburg, 620108, Russia
[2]1N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, 16 S. Kovalevskaya Str., Ekaterinburg, 620108, Russia

Аннотация. Рассмотрение задач, связанных с действительно большими данными, неизбежно ведет к использованию блочного подхода, который применяется и в теории информации и стохастических дифференциальных уравнениях. В качестве естественной метафоры выбраны сигнальные графы – это граф в узлах, которого определена спектральная функция, в рассматриваемы примерах это функция цвета (RGB), высоты или количество данных. В параллельных вычислениях блок можно ассоциировать с вычислительным юнитом (процессором) и рассмотреть задачу максимизации энтропии (производительности). В разрабатываемой системе онлайн визуализации и параллельных вычислений для геометрического распараллеливания можно реализовать и сравнить стационарный случайный процесс (равновероятные сообщения, реализованные с использованием широковещания и миксинов) и установившийся случайный процесс (сообщения точка-точка), которые имеют разные аналитические решения, В совокупности это позволяет сделать вывод, что предложенная реализация стационарного процесса имеет определенную новизну, кроме того она задувалась как более удобная для автоматизированного распараллеливания. Так же рассмотрены задачи автоматической балансировки нагрузки (задача интерполяции) и оптимальной масштабируемости параллельных вычислений (задача экстраполяции). В области верификации визуализации сделано не так много – предложена визуализация сеток, рассматриваемая как параметризованная модель белошумного случайного процесса. Конечно, данную работу нельзя считать завершенной, но направление, которое авторы назвали стохастическая семантика очевидно является перспективным..

***Ключевые слова*:** сигнальные графы, динамические системы, балансировка нагрузки, энтропия, визуализация цифровой модели поверхности.

Аннотация. Considering problems involving truly big data inevitably leads to the use of a block approach, which is used in both information theory and stochastic differential equations. As a natural metaphor, the graph signals were chosen. This is a graph in nodes, of which a spectral function is defined, in the examples considered this is a function of color (RGB), height or amount of data. In concurrent computing, a block can be associated with a computing unit (processor) and consider the problem of entropy (performance) maximization. In the developed on-line visualization and concurrent computing system for geometric parallelization, it is possible to implement and compare a stationary random process (equiprobable messages implemented using broadcasting and mixins) and a steady-state random process (point-to-point messages), which have different analytical solutions. Together, this allows concluding that the proposed implementation of a stationary process has a certain novelty; in addition, it was intended to be more convenient for automated parallelization. The problems of automatic load balancing (interpolation problem) and optimal scalability of parallel computing (extrapolation problem) are also considered. Not much has been done in the field of visualization verification - mesh visualization has been proposed, considered as a parameterized model of a white-noise random process. Of course, this work cannot be considered complete, but the direction that the authors called stochastic semantics is obviously promising.

***Ключевые слова*:** graph signals, dynamic systems, load balancing, entropy, visualization of a digital surface model.

I. Введение

Существует такое теоретическое направление как верификация программного обеспечения (ПО). Поскольку визуализация является то же вычислением хотелось бы иметь общую математическую модель верификации как для ПО, так и для визуализации. Верификация визуализации – формальное (математическое) доказательства правильности визуализации. Разрабатываемые математические модели должны ответить на вопрос правильно ли мы решили задачу, поставленную пользователем, оценить качество и эффективность визуализации, перспективы

развития современных тенденций в области визуализации. В частности, виртуальной реальности (VR), Web-визуализация, on-line визуализации, больших данных, параллельных вычислений.

Для верификации в качестве базовой модели выбраны стохастические дифференциальные уравнениями (СДУ), как наиболее общее решение, учитывающие шум, в сочетание с теорией информации. Эту гипотезу назовем стохастической семантикой.

Первоначальной моделью обработки изображения является модель шумоподавления, которая получила широкое распространение, в частности, из-за распространения таких пакетов как Стабильная Диффузия, генерирующих изображение по тексту. В работе [1] описаны три варианта шумоподавления:

1. Denoising Diffusion Probabilistic Models (DDPM) - вероятностная диффузионная модель шумоподавления;
2. Noise Conditioned Score Networks (NCSNs) – Оценочные сети с условными шумами:
3. Stochastic Differential Equations (SDEs) – СДУ, диффузионный процесс.

Как отмечается, первые две модели являются частным случаем диффузионного процесса, вероятно, которые можно получить, воспользовавшись Марковским свойством, чего и не хватает с точки зрения верификации. В принципе и для других работ по этой тематике характерен инженерный подход, то есть показывается на примерах, что модель рабочая, производится валидация при недостаточной верификации (обычно используется правдоподобие). Наибольшее распространение проучил подход оценочные сети с условными шумами. В этой же работе сказано, что он используется для решения транспортной задачи, а именно генерации суперразрешения. Этот подход также известен как baking (запекание (модель отжига)). Также можно сослаться, например, на запекание нормалей в Blender.

В принципе, те же самые подходы используются или могут применятся и для распознавания речи, но в этом случае спектр состоит из фонем, а относительно визуализации спектр – это базис цветов, например, RGB.

Можно порекомендовать к изучению книгу [2] по обработке изображений на сигнальных графах (частный случай СДУ) с достаточно прозрачной математикой. Сигнальный граф – сеть, в узлах которой заданы значения функции, следовательно, можно определить операторы дифференцирования в узлах и ребрах графа и перейти к СДУ в частных производных. Изначально сигнальные графы применялись в электротехнике, рассматриваемые подходы являются задачей автоматического управления, цель которого построить передаточную функцию, например, по формуле Мэйсона. Для балансировки нагрузки так же применяется метод диффузии [3].

В тоже время хотелось бы, чтобы кроме теории присутствовал определенный прагматизм - применение моделирования для решения конкретных задач. В этой работе представлены две задачи, одна в области параллельных вычислений, другая в области визуализации: формализации динамической системы онлайн визуализации и параллельных вычислений (ДС) на сигнальных графах, визуализация сеток как параметризованная модель белошумного случайного процесса.

1. Стохастической семантика

Давайте рассмотрим визуализацию (интерактивный процесс и анимацию) с точки зрения динамических систем. Определим визуальный процесс одновременно как параллельный процесс, рассматриваемый как взаимодействующие последовательные процессы (Хоара) с точки зрения программирования и как случайный процесс с точки зрения математического моделирования. Случайный процесс — параметризованный набор (семейство) случайных величин [4]. Но в отличие от типовых моделей СДУ вводятся два линейно независимых параметра: количество данных и время (две случайные величины) $\omega = \omega_N \times \omega_t$, где ×- декартово произведение или каррирование в зависимости от модели, теория информации как раз нужна для того, чтобы от произведения перейти к сумме, воспользовавшись свойством аддитивности энтропии. То есть случайны процесс определим как композицию функций $f(t, N, \omega) = f_t(t, \omega_t) \circ f_N(N, \omega_N)$, причем количество (объем) данных N может быть сколь угодно большим, например, не входит в оперативную память одного процессора.

Для верификации программного обеспечения, как и для визуализации наравне с денотацинной семантикой, структурами событий можно использовать и СДУ. Стохастическую семантику можно определить, как игровую семантику с шумом, а можно сразу использовать СДУ. Хотя эти два подхода эквивалентны и относительно программирования рассматривается модель потока данных, но в результате получаются разные математические модели.

В основу верификации положен переход от декларативных определений к формальным (верифицируемым) определениям. В работе [5] семиотическое определение метафоры визуализации рассматривается как непрерывное отображение исходного множества на целевое множество. При этом в стандартное определение метафоры по Лакоффу [6] добавлено только свойство непрерывности. Аналогичным и наиболее известным подходом в области верификации программного обеспечения является денотационная семантика Скотта [7]. В результате применения аксиоматического (семиотического) подхода к формализации визуализации и

параллельной фильтрации данных получен некоторый аналог задачи стохастического управления с морфологической неопределенностью. Полученный результат скорее является ментальным (полезным для понимания природы явления), чем математическим. В данной работе рассматриваются задачи в рамках априорной неопределенности, а именно СДУ и теории информации.

Рассмотрение задач, связанных с действительно большими данными, неизбежно ведет к использованию блочного подхода, который применяется и в теории информации и стохастических дифференциальных уравнениях. Большие данные - предельный (на данный момент) случай обработки данных, при котором универсальные подходы к анализу и визуализации не работают или неэффективны. Тогда в качестве больших данных могут рассматриваться многомерные и многокатегориальные данные, данные большого объема, данные с неполной информацией (модель с неопределенностью). Предельный случай формирует вызовы, на которые необходимо ответить, чтобы двигаться дальше. Решение возникающих проблем приводит к тому, что сегодняшние “большие данные” завтра становятся нормой [8]. При анализе и визуализации больших данных рассмотрение предельной неопределенности – неопределенности, которая имеет конечный предел в конкретной метризуемой топологии, является вынужденным. Поскольку разрабатываются вычислимые модели, поясним это определение на примере вычислимой функции, в определение которой используется специальный элемент, означающий неопределённость, соответствующая случаю, когда алгоритм зависает. Поскольку рассматривается процесс алгоритм не может зависнуть, он всегда интерактивно разрешается в результате отладки правильности и эффективности программы. Почему же до сих пор не предложена асимптотическая теория алгоритмов или автоматов, появление которой прогнозировал Р. Л. Стратонович?

II. Постановка задачи

Параллельные вычисления являются генератором больших данных. В ряде случаев, в место того чтобы хранить все данные, предпочтительнее их обработать средствами визуальной аналитики и принять решение нужны ли эти данные, как следствие, возникает потребность в реализации онлайн визуализации. А если необходимо хранить данные, то их надо размешать не на клиенте, а в облаке или на параллельной файловой системе, как следствие, возникает потребность в реализации удаленной визуализации. Задачи онлайн и удаленной визуализации архитектурно очень близки, поэтому должна существовать единая среда разработки таких программ. Как уже отмечалось, визуализация является то же вычислением, поэтому разрабатывается среда онлайн визуализации и параллельных вычислений. В работе [9] основное внимание уделено валидации этой среды, в том числе задаче балансировки нагрузки средствами визуальной аналитики. В данном случае речь пойдет о верификации, а именно о формализации динамической системы онлайн визуализации и параллельных вычислений. Основное внимание уделено оценке эффективности параллельных программ, учитывающая количество данных, а именно рассматриваются задачи максимизации энтропии для блоковых моделей, где блок – процессор имеет физические и логические уровни.

Прежде чем перейти к формализации, перечислим основные спецификации среды онлайн визуализации и параллельных вычислений:

1. Универсальная система онлайн визуализации и параллельных вычислений;
2. Динамическая система:
3. Применимость параметрических (параметризованных) моделей, в том числе управление через параметры.

Во –первых. технически визуализация является вычислением (конечно, с некоторыми особенностями), поэтому можно предложить некий универсальный способ программирования, включая язык программирования, который позволит эффективно реализовывать алгоритмы визуализации, их связь с параллельными вычислительными программами, а также сами эти программы, то есть произвольные вычислительные алгоритмы.

Во-вторых, эта среда построена вокруг идеи представления параллельного процесса вычисления в форме множества зависимых задач [4]. Зависимость задач означает, что входные аргументы для некоторой задачи определяются результатами вычисления других задач. Список задач и зависимостей между ними определяется пользовательским алгоритмом, который может иметь как форму конечного вычисления, так и случайного процесса. Задержки, возникающие перед приемом и передачей данных, будем рассматривать как шум, который в идеале стремится к нулю. Задачи поступают на вход процессу планирования, который определяет, на каком вычислительном узле следует выполнять ту или иную задачу. Затем, по мере разрешения зависимостей, задачи выполняются. Естественно, что от качества используемого алгоритма планирования зависит эффективность всего параллельного вычисления. Соответствие задачи узлу будем называть назначением. Набор таких назначений и зависимости между задачами будем называть планом выполнения. Подчеркнем, что план выполнения строится во времени, по мере поступления задач и их аргументов (данных). В итоге в работе [9] опытным путем предложена формула балансировки нагрузки, одна из задач данной работы показать, что

формула соответствует задачи интерполяции на сигнальных графах [2]. Для которых должно быть определено значение функции в узлах. Это функцию можно рассматривать как зависящею от времени, но реальное значение времени передачи данных можно получить только после определенного шага выполнения программы, то есть офлайн, что неприемлемо. Поэтому в данном случае рассматривается функция, зависящая от количества данных (аналогично вычислительной сложности). Действительно сложность параллельного алгоритма задается иногда как отношение сложности передачи данных к сложности последовательного алгоритма, например, как $O(N^2)/O(N^3)$. Конечно, это грубая экспертная оценка. Возникает вопрос: можно ли уточнить эту оценку, например, определив передаточную функцию через каноническое разложение, например, Тейлора?

В-третьих, наиболее важной особенностью среды онлайн визуализации и параллельных вычислений является возможность управления как вычислениями, так и визуализацией через параметры. Например, параллельный рендеринг можно организовать не как цикл, зависящий от положения камеры, а как реактивное вычисление. Подобные подходы не только экономят память, но и востребованы для исследовательских задач, кроме того они формализуемы с применением СДУ. В качестве примера, будет рассмотрена визуализация сеток как параметризованная модель белошумного случайного процесса. В частности, для визуализация цифровой модели поверхности.

А теперь перейдем к формализация динамической системы онлайн визуализации и параллельных вычислений. Кроме формального описания этой автоматизированной системы программирования, как уже отмечалось, нас интересуют две задачи: автоматическая балансировка нагрузки и оценка эффективности параллельных вычислений, учитывающая количество данных.

III. Теория

1. *Формальное описание динамической системы онлайн визуализации и параллельных вычислений*

Давайте рассмотрим конструкцию универсальной алгебры, которую назовем структурой процессов: $PU(\mathbb{N}, +, \circ, t\{\omega_i\})$, где $\mathbb{N}$ – множество натуральных чисел, на которое отображаются идентификаторы объектов, $t\{\omega_i\}$ – логическое время, зависящее от списка случайных событий.

Цель автоматической балансировки нагрузки определим следующим образом: для любого процесса pu из структуры процессов построить бинарное гомоморфное отображение G – сигнальный граф, который отображает процесс на CO-OPN (Concurrent Object-Oriented Petri Nets [11]) с минимальным шумом (задержками): $G: pu \rightarrow CO - OPN$.

CO-OPN –параллельная объектно-ориентированная сеть Петри, один из вариантов алгебраической сети Петри. Сеть Петри - дискретная динамическая система (направленный двудольный граф), специально придуманная для параллельных вычислений. Поскольку любое дерево, включая синтаксическое дерево, представимо в виде двудольного направленного графа, гомоморфный образ этого отображения G ничем другим кроме модификаций сетей Петри являться не может. Мотивацией применения сигнальных графов является получение автономной системы дифференциальных уравнений. Аналогичным подходом к применению сигнальных графов является графо-символическое программирование [12]. (Можно привести и другие примеры автоматизации программирования на графах, например, информационные графы, вероятно список работ по данной тематике неограничен, но авторы делают акцент не на программирование, а на формализацию, в частности на развитие теории вычислительной сложности для параллельных вычислений). На наш взгляд, введения типа дуги (последовательная, параллельная, терминирующая) является лишним поскольку есть понятие пути графа. С точки зрения автоматизации программирования и с целью избегания излишнего копирования данных, а возможно и попадания в кэш должно выполнятся правило (функция цены) максимизации пути сигнального графа (графа задач) на каждом процессоре. В сети Петри есть понятие перехода аналогично в графо-символическом программирование используется метка предиката, а в ДС динамической системе онлайн визуализации и параллельных вычислений на нижнем уровне реализованы стандартные языковые конструкции фьючерсы (обещания) и миксины (примеси), но в параллельном варианте. Обещание – это предикат, зависящий от типа данных (параметров объекта), в данном случае, есть прямая аналогия с раскрашенными функциональными сетями Петри. Примесь– это предикат, зависящий от номера процессора, следовательно, зависящий от схемы распараллеливания. В контексте больших данных реализация примесей является существенным дополнением к сетям Петри. Далее будет показано, что с точки зрения теории информации эффективная реализация примесей должна основываться на широковещание и векторной маршрутизации потоков (буферизации сообщений). Некоторые примеры примесей также будут приведены.

Логическое время определяет порядок событий. В ДС возможны два типа событий – реактивные вычисления и обмен сообщениями. Поскольку энтропия является аддитивной, то с точки зрения теории информации эти две задачи можно рассматривать независимо, например, можно ввести два логических времени. Хотя авторов

интересует формализация задач, связанных с обменом сообщениями, немного остановимся на реактивных вычислениях.

В реактивных вычислениях случайным событием является (онлайн или офлайн) изменение значения параметра. По аналогии с абстрактными типами данных, такой подход, авторы, называют абстрагированием параметров Самой простой реализацией онлайн визуализации, но вероятно не самой эффективной, является офлайн изменение параметра (по есть все данные текущей итерации находятся на клиенте, пока пользователь занимается визуальной аналитикой, на вычислителе можно посчитать следующую итерацию), поскольку нет проблемы с установлением порядка событий (то есть можно ввести два ортогональных логических времени). Формально абстрагирование параметров можно рассматривать как лямбда аппликацию и применять, например, аппарат теории категорий. Хотя монада определяется как функтор с дополнительной структурой в контексте больших данных между ними существуют принципиальные отличия. Для функторов не существует проблемы асимптотической сходимости, достаточно, чтобы параметры были линейно независимыми или чтобы выполнялось свойство мультипликативности (в теории категорий используют термин каррирование). Для монад можно рассматривать сходимость, если определен повторный интеграл через кратный интеграл, например, для решеток, кривых, заполняющих пространство, мозаик, случайных графов, иначе редукция синтаксического дерева или графа задач – это задача дискретной оптимизации на графах. Фактически в ДС обещание – это монада. Таким образом, структура процессов является не просто конструкцией универсальной алгебры, а алгебраической системой, Ее сигнатурой называется набор функциональных и предикатных символов с их арностями.

Задачи, связанные с обменом сообщениями, будут рассмотрены с позиции априорной неопределенности, хотя авторы и предполагают ответ. Так оценка эффективности параллельных вычислений - это неоднородная задача Марковского управления, где правая часть уравнения равна количеству блоков (процессоров) в блочной модели. А задача автоматической балансировки нагрузки (алгоритм планирования) соответствует задаче интерполяции на сигнальных графах. Алгоритм планирования - это алгоритм, который выбирает, на каком исполнителе (однопоточный процесс уровня ОС) следует выполнять ту или иную задачу. Поскольку объем оперативной памяти на разных классах процессоров разный (заранее не известен), то как решение можно рассматривать транспортную задачу, например, с переменным числом выходов. Вроде бы существует решение этой постановки, как задачи автоматического управления – построения передаточной функции с использованием полинома Гурвица, но в ДС применяется другой подход, если памяти не хватает, то задача переносится на другой исполнитель, в результате возможна разбалансировка нагрузки. В целях обобщения моделирования далее исполнителя будем называть процессором.

2. *Задача автоматической балансировки нагрузки*

Давайте рассмотрим задачу интерполяции на сигнальных графах. Для начала в области визуализации «рис. 1». Пусть функция $f^0: \mathcal{V}_0 \to \mathbb{R}$, где $\mathcal{V}_0 \subset \mathcal{V}$ подмножество вершин графа с известными значениями. Задача интерполяции сводится к рассмотрению уравнения: $\Delta f(v_i) = 0$ на $\mathcal{V} \backslash \mathcal{V}_0$. В первоисточнике [2] Лапласиан задается специфическим образом для изотропного и анизотропного диффузионного процесса, что для дальнейших рассуждений не существенно. В работе рассмотрены простые примеры распараллеливания - изотропные диффузионные процессы. Авторы считают, что такие примеры распараллеливания как: задачи дискретной оптимизации на графах и решение СЛАУ методом Холецкого являются анизотропными диффузионными процессами.

Рис. 1. Схема интерполяции [2]

Алгоритм планирования должен динамически распределить граф задач по процессорам, так чтобы они не простаивали. Для задачи [13] распараллеливания одномерного массива на линейке процессоров с теневой гранью шириной в единицу, то есть на каждом шаге итерации вычислялась функция: $f(mu[i-1], mu[i], mu[i+1], \sigma)$, σ –среднее значение функции на предыдущем шаге. Опытным путем в сочетание с визуальной аналитикой была получена следующая формула балансировки нагрузки [9]:

$$E_i = \left|\mathcal{V}_i \backslash \mathcal{V}_i^0\right| + kln(queue_i),$$

где математическое ожидание нагрузки процессоров - средние значение нагрузки i-процессора E_i (очередная задача назначается исполнителю с наименьшим значением E_i), $\left|\mathcal{V}_i \backslash \mathcal{V}_i^0\right|$ - количество аргументов задачи отсутствующих в кэше исполнителя, $queue_i$ - текущий размер списка назначенных и пока нерешенных заданий (очереди сообщений). Вычислительная сложность пирамидальной (оптимальной) обработки очереди имеет логарифмическую зависимость. Одной из проблем являлось подбор коэффициента $k = 0.1$. Можно предположить, что он является характеристикой архитектуры распараллеливания и равен отношению (физической) пропускной способности канала к скорости вычисления процессора.

Чтобы перейти к общему случаю, давайте рассмотрим другую схему геометрического распараллеливания: задано равномерное распределения данных (двумерного массива) по процессорам, $N \times N$ - количество данных на каждом процессоре, N – переменный параметр, $f_i = N^2$, ∇f_i=2N . Задачу интерполяции для балансировки нагрузки запишем в следующем виде:

$$\nabla E_i = \nabla\left|2N(\mathcal{V}_i \backslash \mathcal{V}_i^0)\right| + \nabla kln(queue_i) = 0.$$

Очевидно, что для рассматриваемого примера 2N – длина сообщения. В соответствии с Марковским свойством диффузионных процессов [4] $\lim_{i\to\infty} \nabla kln(queue_i) = 0$ (то есть рассматривается белый шум, общее название - задача Пуассона, другое слагаемое соответствует задаче Дирихле).

Стохастический процесс обладает Марковским свойством, если условное распределение вероятностей будущих состояний процесса зависит только от нынешнего состояния, а не от последовательности событий, которые предшествовали этому.

Вместо математической формулировки Марковского свойства достаточно ссылки на следующий пример.

Пример 7.3.4. Броуновское n-мерное движение, конечно, является решением стохастического дифференциального уравнения: $dX_t = dB_t$.

Таким образом, производящий оператор A процесса B_t (броуновское n-мерное движение) имеет вид:

$$Af = \frac{1}{2}\Delta f.$$

Для общего случая достаточно правдоподобно выглядит следующая формула балансировки нагрузки:

$$E_i = \left|d * \nabla f_i(\mathcal{V}_i \backslash \mathcal{V}_i^0)\right| + kln(queue_i),$$

где d – ширина теневой грани, $\mathcal{V}_i \backslash \mathcal{V}_i^0$ – количество ожидаемых сообщений, а не. количество аргументов задачи отсутствующих в кэше исполнителя, поскольку в целях оптимизации необходимо использовать буферизацию сообщений.

Актуальным направлением в программирование является векторная маршрутизация потоков. Кратко остановимся только на буферизации сообщений. На практике нужно рассматривать среднюю длину сообщения в количестве буферов (полезный размер буфера является константой). Очевидно, что буферизация коротких сообщений значительно увеличивает скорость обменов, а для длинных сообщений она не хуже, следовательно, для сообщений произвольной длины также выгодно использовать буферизацию. Поскольку командам синхронизации свойственна высокая латентность коротких сообщения, а буферизация не всегда возможна, то возникает желание отказаться от синхронизации, далее будут рассмотрены недетерминированные сообщения.

Хотя формула балансировки нагрузки выглядит достаточно правдоподобно сточки зрения эксперта программиста, авторы обещали рассмотреть поставленные задачи в рамках априорной неопределенности, а именно СДУ и теории информации.

3. Теории информации и модель обмена сообщениями

При любой формализации неизбежна определенная идеализация. Хотя авторы акцентируют внимание на модели обмена сообщениями, эта модель применима и для других параллельных архитектур: для общей памяти и ускорителей, с несущественными дополнениями. Моделирование основано на применение монографии Р. Л. Стратоновича: “Теория информации” [14], который принципиально отказался от использования специальных терминов теории информации с целью обобщения ее и термодинамики. Авторы напротив намерены пояснять некоторые формулы в терминах теории информации и СДУ. Р. Л. Стратонович рассматривал кодирование /(передачу) информации, авторы рассматривают передачу (обмен) сообщениями, что в принципе одно и тоже.

Максимальное значение энтропии называется пропускной способностью (информационной емкостью) канала (вычислителя, параллельной программы) без помех. Рассмотрим задачу максимизации энтропии (производительности) при максимизации пути сигнального графа на каждом процессоре, которая, действительно, напоминает постановку транспортной задачи дробно-линейного программирования: максимизация паросочетаний при максимизации потока.

Понятие “сообщения” без ущерба для теории можно заменить на понятие “случайной величины”, понятие “последовательности сообщений” на “случайный процесс”. Так количество информации в контексте теории вероятности представляется в виде средней энтропии:

$$I = H_\xi = -\sum_\xi P(\xi) lnP(\xi),$$

где ξ – дискретная случайная величина, а $P(\xi)$ – ее распределение вероятности.

Эта формула является следствием (в асимптотическом смыли) формулы Хартли, для неравновероятных событий, которая в виде случайной энтропии (то есть энтропия - случайная величина) представлена как:

$$H(\xi) = -lnP(\xi),$$

с условием нормировки - $\sum_\xi P(\xi) = 1$.

Естественно, средняя энтропия является средним значением случайных энтропий:

$$H_\xi = MH(\xi).$$

Авторы предпочли бы изложение теории информации сразу в терминах СДУ. Так в случае непрерывной случайной величины можно ли вместо суммы использовать интеграл Ито, ведь известно, что дифференциальная энтропия неограничена? И решение тоже известно: надо рассматривать нормализованную энтропию, которую в контексте нечетких множеств называют невероятностной энтропией, или: надо рассматривать производную Радона-Никодима.

При определении структуры процессов ранее упоминалось свойство аддитивности энтропии.

Теорема 1.3. Если случайные величины ξ_1, ξ_2 независимы, то полная (совместная) энтропия H_{ξ_1,ξ_2} распадается на сумму энтропий:

$$H_{\xi_1,\xi_2} = H_{\xi_1,} + H_{\xi_2}.$$

Теорема 1.4. Энтропия обладает свойством иерархической аддитивности:

$H_{\xi_1,\dots,\xi_n} = H_{\xi_1} + H_{\xi_2|\xi_1} + \dots + H_{\xi_n|\xi_1,\dots,\xi_{n-1}}$, где $H_{\xi_2|\xi_1}$ – условная энтропия.

Это свойство в практическом плане используется при реализации миксинов. Рассмотрим задачу для которой надо посчитать Mf_i, где f_i – среднее значение функции на i-процессоре. Исходя из последовательного варианта с целью повышения эффективности можно предложить попарное пирамидальное суммирование, но в этом случае сообщения будут точка-точка, то есть неравновероятными, а канал несимметричным. Возможно для небольшого числа процессоров этот вариант будет более эффективным, чем реализация с учетом свойства иерархической аддитивности: Используя широковещание, суммирование надо вести на каждом процессоре, при этом дерево выбора на каждом процессоре будет свое, но условная энтропия на каждом процессоре все равно будет стремится к нулю. Данная схема распараллеливания близка к схеме мастер-рабочий без синхронизации, для которой каждый процессор является и мастером, и рабочим.

Для реализации дерева выбора каждый процессор как минимум должен иметь два (физических) двунаправленных канала связи., где k - количество каналов процессора аналогично числу букв алфавита (для многоядерной архитектуры k - количество ядер, но дерево выбора направлено в другую сторону, аналогично пирамидальному суммированию). Кроме того, существует не совсем корректное мнение, что результат суммирования зависит от порядка суммирования, в случае энтропийной устойчивости все реализации суммирования будут приблизительно равны.

Закон Амдала иллюстрирует ограничение роста производительности вычислительной системы с увеличением количества вычислителей (процессоров), который по формулировке является хорошо изученным распределением Бернулли. То есть закон Амдала является частным случаем с точки зрения теории информации и не учитывает обмен сообщениями. Благодаря свойству аддитивности энтропии эти две задачи можно рассматривать независимо. Кроме того, для геометрического распараллеливания при фиксированном количестве данных доля последовательных вычислений обратно пропорциональна количеству процессоров, то есть стремится к нулю при количестве процессоров стремящимся к бесконечности, следовательно, при оценке производительности надо учитывать только обмен сообщениями. При рассмотрении общего случая понадобится не только равноправность процессоров, но и равновероятность сообщений (асинхронные недетерминированные сообщения, стационарный процесс по количеству процессоров). Поскольку наибольшие распространение получила кластерная архитектура, в которой обмены осуществляются через общую шину (можно рассматривать и k-дерево), то подобная идеализация оправдана, кроме того каналы являются двусторонними, то можно считать, что длина пути для асинхронных сообщений равна единице, а для синхронизированных обменов она равна двум. Длина пути - это константа, зависящая от типа сообщения и архитектуры вычислителя, на которую надо умножить, среднею длину сообщения.

На самом деле равновероятность сообщений не обязательна. Можно рассматривать детерминированные сообщения (установившийся процесс), например, решетка данных на решетке процессоров, но это достаточно узкий класс задач. Перспективой расширения класса задач является рассмотрение анизотропных диффузионных процессов и кратных интегралов.

Дать общую формулировку свойства асимптотической эквивалентности неравновероятных возможностей (сообщений) равновероятным помогает понятие энтропийной устойчивости семейства случайных величин.

Семейство случайных величин $\{\eta^n\}$ является энтропийно устойчивым, если отношение $H(\eta^n)/H_{\eta^n}$ при $n \to \infty$ сходится по вероятности к единице. Это значит, что каковы бы не были $\varepsilon > 0,\ \eta > 0$ найдется такое, $N(\varepsilon, \eta)$ что будет выполнено неравенство:

$$P\{|\, H(\eta^n)/H_{\eta^n} - 1| \geq \varepsilon\} \leqslant \eta \ ,$$

при любом $n > N(\varepsilon, \eta)$.

В определении подразумевается, что все $0 < H_{\eta^n} < \infty$ и H_{η^n} не убывает сростом n.Обычно $H_{\eta^n} \to \infty$.

Факт асимптотической равновероятности можно сформулировать при помощи понятия энтропийной устойчивости в виде следующей теоремы.

Теорема 1.9. Если семейство случайных величин $\{\eta^n\}$ является энтропийно устойчивым, то множество реализаций каждой случайной величины можно разбить на два подмножества A_n и B_n таким образом, что

1. Суммарная вероятность подмножества A_n исчезает:

$P(A_n) \to 0$ при $n \to \infty$;

2. Реализации второго подмножества B_n становятся относительно равновероятными в смысле соотношения $|\frac{lnP(\eta) - inP(\eta')}{lnP(\eta}| \to 0$ при $n \to \infty, \eta \in B_n, \eta' \in B_n$;

3. Число M_n реализаций множества B_n связано с энтропией H_{η^n} соотношением

$lnM_n/H_{\eta^n} \to 1$ при $n \to \infty$.

Приведем некоторые комментарии относительно энтропийной устойчивости. С точки зрения СДУ определение энтропийной устойчивости можно рассматривать как обобщение интеграла Ито через предел по вероятности: Отношение $H(\eta^n)/H_{\eta^n}$ – относительная энтропия или расстояние Кульбака-Лейблера для равномерного распределения. В этом случае используется обозначение $H_{\xi}^{P/Q}$.

В теории информации основное внимание уделено текущему кодированию, но также рассмотрен и другой подход, который можно назвать блочным. При нем подлежит кодировке конечная совокупность (блок) элементарных сообщений. Если блок является энтропийно устойчивой величиной, то вероятность потери части реализаций сообщений достаточно мала. Как уже отмечалось вместо кодировки сообщений можно рассматривать передачу сообщений, а также максимальное значение энтропии называется пропускной способностью канала без помех. Рассмотрим задачу максимизации энтропии в случае блочного подхода. В качестве комментария отметим, что эта задача эквивалентна задачи минимизации расстояние Кульбака-Лейблера, также ее называют задачей максимального правдоподобия. В нашей интерпретации блок – это процессор, который должен быть энтропийно устойчивым, то есть ПРОЦЕССОР=МАРКОВСКИЙ ПРОЦЕСС, с учетом аппаратной и программной реализаций так же, как и весь вычислитель. Если рассматривать задачу отказоустойчивости для Марковского процесса, то для ее реализации достаточно одного дополнительного процессора (средняя вероятность выхода из строя одного процессора очень мала), причем события, связанные с заменой одного процессора на другой, не влияют на производительность вычислений (без учета сохранения и восстановления данных).

Часть задачи, реализуемая на процессоре (блоке), должна быть Марковским процессом.

Приведем некоторые примеры блочных Марковских процессов.

Авторы постоянно ссылаются на геометрическое распараллеливание, поскольку в этом случае интерпретация является достаточно наглядной. Данные расположенные в блоке являются открытым множеством, поскольку есть теневые грани. Суммарная мера на границе множества должна быть меньше меры внутренности множества, по сути это вариант неравенства Чебышева, активно используемом при доказательстве теорем.

В принципе можно рассматривать граф (задач на процессоре) достаточно произвольной структуры. Без умаления общности можно рассмотреть производную Радона-Никодима, введя две меры, а, следовательно, и случайный процесс:

$$E \leq X \to \frac{1}{X} \leq \frac{1}{E(X)} \ll 1,$$

где X – мера определенная в узлах графа и зависящая от их количества, E – мера, определенная на дугах (границах) графа.

Эти соотношения справедливы и для сигнальных графов, например, количество ближайших соседних вершин. И применимо и в других областях знаний, например, для визуализации графом, в частности антологии знаний. “Для гиперграфа h с заданным набором узлов X, заданным набором ребер E и набором соответствующих значений матрицы инцидентности $\{a\}$ информация I(h) имеет следующий вид [15]”:

$$I(h) = \frac{1}{|X|} log_{|\{a\}|}|E|.$$

Хотя матрица инцидентности используется для редукции графов, включая сети Петри, на взгляд авторов, использование определителя матрицы инцидентности в основание логарифма не обосновано.

Распишем задачу минимизации расстояния Кульбака-Лейблера:

$$\lim_{X\to\infty} D_{KL}\left(\frac{1}{X}\middle|\frac{1}{E(X)}\right) = \sum \lim_{X\to\infty}\frac{1}{X}\ln\left(\frac{E(X)}{X}\right) = \sum \lim_{X\to\infty}\frac{1}{X} ln(E(X)) - \sum \lim_{X\to\infty}\frac{1}{X} ln(X) = \sum (I(h) - 0)$$

Фактически *I(h)* – относительная энтропия, а условная энтропия стремится к нулю.

Известным примером является случай, когда, дерево выбора ограничено сверху k-деревом, в теории информации получившим название – оптимальное кодирование дешифруемых кодов Крафта.

Теорема 2.3. Можно указать такой способ кодирования (передачи) равнораспределённых независимых сообщений, что

$$l_{\text{ср}} < \frac{H_\xi}{lnD} + 1.$$

В нашей интерпретации $l_{\text{ср}}$ – средняя длина пути между узлами вычислителя, D – количество каналов процессора (в оригинале количество букв в алфавите). В случае обмена сообщениями через общую шину можно считать, что $D = \infty$, следовательно, средняя длина пути между узлами вычислителя равна единицы, что соответствует экспертному подходу. В случае блочного подхода применима следующая теорема.

Теорема 2.4. Существуют такие способы кодирования бесконечного сообщения, что средняя длина элементарного сообщения может быть сделана сколь угодно близкой к $\frac{H_\xi}{lnD}$.

Подобные оценки для средней длины пути применимы не только для аппаратной архитектуры вычислителя, но и на логическом уровне для графов в пространстве R^D. Полезными ссылками являются: внеядерные алгоритмы (реструктуризация данных по *k*-дереву) [16]; Полиномиальная приближенная схема для задачи о цикловом покрытии графа фиксированного размера k в евклидовом пространстве произвольной фиксированной размерности (дискретная оптимизация) [17].

Поскольку процессор сам с собой не обменивается сообщениями, можно ввести понятие субэнтропии, например, для равновероятных сообщений, аналогично определению субфакториала. Субфакториал *!n* определяется как количество беспорядков порядка n, то есть перестановок *n*-элементного множества без неподвижных точек В случае асимптотической сходимости равновероятность сообщений не обязательна как и рассмотрения субэнтропии. В работе [18] показано, что в классической теории информации субэнтропия двойственная субэнтропии фон Неймана, определенная через перестановки попарной разности собственных значений, является точной нижней границей пропускной способности канала и ее вычисление соответствует задаче минимизации расстояния Кульбака-Лейблера.

В зависимости от задачи длина сообщения между разными процессорами может быть не одинаковой. В этом случае основной характеристикой является средняя длина сообщения. Рассмотрим прямой способ вычисления максимальной энтропии для этого примера, что соответствует параграфу 3.1. Пусть имеются m процессоров $V_1, \dots, V_m$, которые передают сообщения соответственно длиной $l(1), \dots, l(m)$. Суммарная длина сообщения будет $L = l(1) + \cdots + l(m) = l_{\text{ср}}m$. Зафиксируем эту длину и будем подсчитывать число $M(L)$ различных реализаций данной длины. Максимальная информация достигается, когда все из $M(L)$ возможностей равновероятны. При этом

$$\frac{H_L}{L} = \frac{lnM(L)}{L}.$$

Взяв придел при $L \to \infty$ получим энтропию, рассчитанную на единицу длины. То есть достаточно рассмотреть решение в случае асимптотической сходимости линейного однородного уравнения, которое имеет вид: $M(L) = Ce^{\lambda L}$. Решение имеет единственный корень с максимальной действительной частью λ_m. И так число различных реализаций длины L имеет вид:

$$M(L) \approx C_m e^{\lambda_m L}.$$

Тот же самый результат можно получить из решения первой вариационной задачи. Ограничимся только рассмотрением дискретного канала без помех (общей постановкой задачи). Система $[Y, c(y), \alpha]$ полностью характеризует дискретный канала без помех, $y \in Y$, функция штрафов - $c(y) \le \alpha$ (микроканоническое распределение). В частности $c(y) \le L$. Удобнее рассматривать каноническое распределение: $\sum c(y)P(y) \le \alpha$. Пропускную способность C или информационную емкость канала $[Y, c(y), \alpha]$ определим, как максимальное значение энтропии $\mathrm{C} = \sup_{P(y)} H_y$. Таким образом пропускная способность канала определена как решение вариационной задачи. Отметим, что она эквивалентно задачи минимизации рисков (временных задержек). Эквивалентность микроканонического и канонического распределений доказана. В контексте данной работы закон Амдала – это микроканоническое распределение, а соответствующим ему каноническим распределением является распределение Бернулли. Также подчеркнем, что рассматривается задача максимизации производительности (энтропии) параллельных вычислений.

В качестве иллюстрации важности средней длины сообщения приведем пример поведения производительности параллельной программы при фиксированном количестве данных и переменном количестве процессоров. При

росте количества процессоров средняя длина сообщения в количестве буферов может уменьшится на единицу, в следствие экспоненциальной зависимости от длины сообщения, должен наблюдаться скачек производительности. Таким образом, длина сообщения может зависеть, так же как и доля последовательных вычислений, от количества процессоров (переменного параметра по сути аналогичного времени). Следовательно, надо рассматривать решение линейного неоднородного уравнения, а для начала выписать функцию штрафов.

4. Оптимальная масштабируемость параллельных вычислений (блочный подход)

В контексте высокопроизводительных вычислений существует два показателя масштабируемости:

1. Сильная масштабируемость — показывает, как меняется время решения задачи с увеличением количества процессоров (или вычислительных узлов) при неизменном общем объёме задачи;
2. Слабая масштабируемость — показывает, как меняется время решения задачи с увеличением количества процессоров (узлов) при неизменном объёме задачи для одного процессора (или узла)

Оптимальная масштабируемость (оценка производительности параллельных вычислений, учитывающая как количество данных, так и количество процессоров) - это неоднородная задача Марковского управления, где правая часть уравнения равна количеству блоков (процессоров) в блочной модели. Конечно, это очередная правдоподобная гипотеза

Под оптимальной масштабируемостью будем понимать случай, когда пропускная способность канала определена как решение вариационной задачи, то есть энтропию можно выписать в явном виде. Для начала надо определить функцию штрафов (для дискретного канала), зависящую от двух параметров количества данных и количества процессоров – p:

$$c(N,p) = s\sum_{i=1}^{p} l_i(N,p) + \sum_{i=1}^{p} \alpha_i(N,p),$$

где α_i – доля последовательных вычислений итого процессора, s – средняя длина пути (напомним, что для равновероятных сообщений она равна единице). Кроме того, ранее предполагалось использование сигнальных графов, как развитие идей вычислительной сложности, то есть $l_i(N,p) = \nabla f_i(N,p)$, что приведет к рассмотрению второй вариационной задачи.

Далее постараемся найти подобную функцию штрафов в теории информации. Процитируем параграф 3.5 – метод потенциалов в случае большого числа параметров: “Функция штрафов зависит от числового параметра и является дифференцируемой по этому параметру.” В нашем случае по количеству процессоров (блоков) – p. Авторы не видят большого смысла переписывания известных формул в других обозначениях. Все же приведем одно определение. Функция

$$B(N) = -\frac{\partial c(N,p)}{\partial p}$$

называется случайным внутренним (эндогенный) термодинамическим параметром, сопряженным с внешним (экзогенный) параметром p. Далее в том же параграфе рассмотрен пример с двумя случайными переменными, в нашем случае это длина сообщения и доля последовательных вычислений, которые находятся из решения системы двух уравнений с двумя неизвестными. Таким образом аналитическое решение рассматриваемой задачи известно.

Аналитическое решение может, например, использоваться для определения помех в вычислителе. Следовательно, возникает потребность в разработке адекватных моделей обработки статистических данных производительности параллельных вычислений в том числе и для следующей задачи. До сих пор рассматривались только стационарные процессы. Используя миксины, в ДС для графа задач можно реализовать стационарный процесс с некоторыми накладными расходами относительно доли последовательных вычислений (очевидно, что должно выполнятся неравенство Чебышева, чем больше данных в блоке, тем меньше доля накладных расходов, вероятно, надо рассматривать канал с помехами (параграф 7)) и чтения сообщений (напомним, что размер очереди сообщений стремится к бесконечности, что соответствует белому шуму). В случае геометрического распараллеливания можно реализовать и сравнить стационарный процесс и установившийся процесс (сообщения точка-точка). Аналитическое решение для последнего в теории информации то же известно (хотя, термин установившийся процесс не используется и параллельные вычисления не рассматривались). Для этого вводится понятие информации связи (в нашем случае дуги графа), естественно, через условную вероятность, через нее же определяется симметричный канал, естественно, применяя перестановки, соответственно параграфы 6 и 8. При этом условное распределение на выходе канала при фиксированном входном сигнале предполагается известным (фактически строится передаточная функция).

Пропускной способностью канала $[P(y|x), c(x)]$ называется максимальное значение информации связи между входом и выходом:

$$\mathrm{C} = C[P(y|x), c(x)] = \sup_{P(x)} I_{x^{i,j}}.$$

В случае полностью симметричного канала (решетка процессоров) формула выглядит достаточно простой (8.4.9):

$$\mathrm{C} = lnM\,{}^{1}/_{p_j} - Mln\,{}^{1}/_{p_j}.$$

В случае равномерного распределения формула имеет следующий вид

$$\mathrm{C} = lnM\frac{Q(dy)}{P(dy)} - Mln\frac{Q(dy)}{P(dy)}.,$$

где $Q(dy)$ вспомогательная мера, на основание которой производится перестановка интервалов.

Но эта формула выполняется разве что на ядре (в активной зоне) или в случае общей памяти, поскольку в рассмотренных случаях [14] функция штрафов зависит от одной переменной. Для геометрического распараллеливания можно сократить зависимости до одной переменной, учитывая, что при фиксированном количестве данных доля последовательных вычислений обратно пропорциональна количеству процессоров, то есть стремится к нулю (но не ноль) при количестве процессоров стремящимся к бесконечности. Чем больше данных на процессоре, тем сбалансированность вычислений лучше, что ведет к определенному противоречию.

В практическом плане более важным является соблюдение определенных правил программирования при разработке параллельных программ, начиная с бинарного гомоморфного отображения и заканчивая установившийся процессом. Если график сильной масштабируемости соответствует логарифмической функции, то параллельную программу можно считать установившийся процессом по количеству блоков (если строго, то надо рассматривать линейную задачу фильтрации [4]: системы с шумом и измерений с шумом). С целью повышения производительности можно рассмотреть следующую задачу [5] экстраполяции (прогноза [4]) до момента остановки: $N(p-1)$ известно, надо найти N(p) при максимизации энтропии, которую авторы как раз и называют неоднородной задачей Марковского управления. Поскольку общее решение неоднородного уравнения является суммой фундаментальной системы решений и частного решения, рассмотрение неоднородной задачи автоматически ведет к увеличению энтропии и понижает размерность системы однородных уравнений (в нашем случае до одного). Неоднородную задачу можно получить из случайной замены времени (в [5] рассматривалась функция энтропии $p/N(p)$ или дробная линеаризации на ядре) $L = l_{\text{ср}}p$, дифференцируя по средней длине сообщения. На границе должно выполнятся условие, соответствующие линейному ускорению (максимизации энтропии) $\dot{X} = p$.

Поскольку обрабатываются статистические данные, надо перейти к Марковским цепям (решению линейной системы алгебраических уравнений), что и будет является наилучшей аппроксимацией задачи максимизации эффективности параллельных вычислений модели обмена сообщениями:$H_pX = p$, где H_p – матрица обмена сообщениями (матрица смежности), на главной диагонали указано среднее время последовательных вычислений блока, а другие элементы соответствуют среднему времени передачи сообщений между процессорами, предполагается, что схема распараллеливания не меняется (не зависит от номера итерации).

Матрица обмена сообщениями будет близка к симметричной поскольку ответное сообщение той же длины предположительно должно занимать немного больше времени. В таких случаях рассматривается задача Дирихле со стохастическим управлением вида регулятор с неполной информацией о состоянии системы назовём ее задача стохастического управления информацией. Но зато у этой матрицы будет другая симметрия, связанная с балансированными вычислениями, схемой распараллеливания и определением симметричного канала.

Можно считать, что собственные значения отвечают за способ распараллеливания. Важно, что в итоге получаются матрицы определенных шаблонов, так для схемы мастер-рабочий не нулевыми являются главная диагональ и i -строка и i –столбец (i – номер процессора, на котором расположен мастер). Для конвейера или линейки процессоров получается трехдиагональная матрица, для решетки пятидиагональная и т.д.

Подробнее рассмотрим схему мастер-рабочий. Матрица смежности имеет следующий вид

$$H_p = \begin{vmatrix} a & b & \dots & b \\ b & a & & 0 \\ \vdots & & \ddots & \\ b & 0 & \dots & a \end{vmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ p \end{matrix}.$$

Система не совместна, только когда $a_{1,1} = 0$ - время без учета подготовки и сохранения данных, то есть мастер не занимается счетом. Из последнего уравнения легко найти x_p

$$dx \approx x_p = \frac{p - bx_1}{a}.$$

В пространстве L_2 приращение энтропии (коэффициент сдвига в диффузионном процессе)^

$$dx = \overline{X_p} - \overline{X_{p-1}} \approx x_p.$$

С учетом вычислительной сложности алгоритма $a = f(N)$ (коэффициент диффузии), а $b = \nabla f(N)$, например, $f(N) = N^3$ и балансировки загрузки по количеству процессоров приращение энтропии имеет вид;

$$dx = \left(\frac{P}{N}\right)^3 p - O\left(\frac{1}{N}\right) \rightarrow\ dx \approx \frac{p^4}{N^3}.$$

Получается, что чем сложнее алгоритм, тем он лучше распараллеливается (масштабируется). Если положить, $\frac{p^4}{N^3}$.=1 (максимальное значение энтропии равно единице) то получим выражение N через p/ Конечно, приведенная оценка является грубой, но можно рассмотреть параметризованную модель белошумного случайного процесса, поскольку функция энтропии всегда $p/N(p) \ll 1$ (подробнее в разделе про визуализацию). Эта функция является гармонической функцией (свойство среднего), следовательно, рассмотрение задачи о движущейся границе является обоснованным, которая СДУ называется задачей Якоби [4].

В качестве того, что функция штрафов выписана правильно, можно привести сравнение графиков бета-функции распределения Дирихле и эффективности параллельных вычислений для задачи решения СЛАУ методом Холецкого [19] в случае сбалансированных вычислений «рис. 2».

Пусть $X = (X_1, \dots, X_K) \sim Dir(a)$ – распределение Дирихле и $a_0 = \sum_{i=1}^{K} a_i$, тогда

$$E[X_{i|a}] = \frac{a_i}{a_0}.$$

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

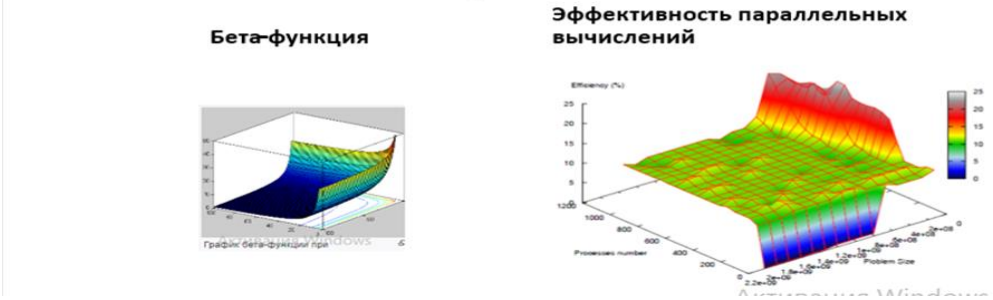


Рис. 2. Бета-функция и эффективность параллельных вычислений

С точки зрения визуальной верификации графики похожи

5. *Примеры применения блочного подхода в других областях знаний*

Блочный подход в СДУ применятся достаточно часто, например, в глобальной модели сейсмичности земли [20], но количество блоков является фиксированным (например, количество литосферных плит). К сожалению, авторам не удалось найти примеров, когда количество блоков является переменным параметром. А эта задача является актуальной, не только для задачи прогноза производительности параллельных вычислений, но и для других задач экстраполяции (в общем случае задач стохастического управления информацией), которые приведены ниже.

Рассмотрим аналогичный пример из области экономики – анализ данных Счетной Палаты. Сгенерирована матрица поставок по (автомобильным) номерам регионов «рис. 3» и определена гармоническая функция - отношение суммы поставок в регионе к количеству фирм в регионе, ее значение между регионами отображены в виде сфер в соответствующих элементах матрицы. Видно, что перестановкой строк из этой матрицы можно получить матрицу близкую к симметричной с учетом того, что заказывают всегда больше, чем надо (определена точная верхняя граница). Хотя значение регулятора является важным в экономике, оно связано и с нормой прибыли и уровнем коррупции, рассмотрим только задачу сокращается ли информационный разрыв между Москвой и другими регионами. Определим меру информационный разрыв, как отношение многомерного расстояния к географическому расстоянию, которое является вспомогательной мерой в производной Радона-Никодима). Если рассматривать отношение информационных разрывов в задаче экстраполяции, то вспомогательная мера сократится. Если мера отношения будет возрастать, то информационный разрыв будет сокращаться и наоборот. А теперь рассмотрим эту же задачу, когда появился новый регион. Она уже была рассмотрена - количество блоков (регионов) является переменным параметром. Можно рассматривать и типовые задачи кластеризации, но с точки зрения диссипативных систем – образования новых кластеров. Эта же модель примени для анализа распространения информации в интернет сетях, где в качестве параметра выступает не количество регионов, а количество информационных каналов (блогов).

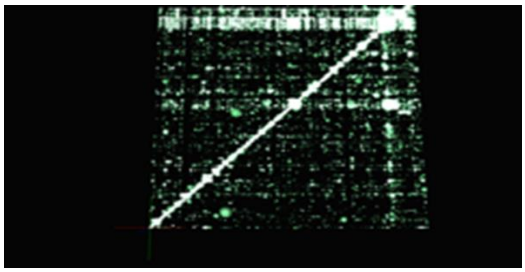


Рис. 3. Матрица поставок по регионам

Далее рассмотрим некоторые примеры задач стохастического управления визуальной информацией, сосредоточив основное внимание на применение блочного подхода в визуализации.

6. *Стохастическое управление визуальной информацией*

Синкхорновские (Sinkhorn) нейронные сети, в основе которых лежит теорема Синкхорна [21], используются для решения широкого класса транспортных задач; суперразрешение, сравнение двух дистрибутивов. Утверждается что, эти сети лучше по быстродействию и числу параметров, чем генеративные сети максимального правдоподобия. Далее будет рассмотрена возможность применения СДУ для задач на (сигнальных) графах как альтернатива традиционному применению нейронных сетей. Для параллельных вычислений функция, определенная на сигнальных графах, зависела от количества данных в этом параграфе зависимость от высоты (карта высот). Сначала будет рассмотрена задачи визуализации цифровой модели (неограниченной, плоской) поверхности (ЦМП) с фиксированным количеством блоков и переменным количеством данных в каждом блоке. Для этой задачи максимальное количество блоков зависит от количества данных, которое ограниченно памятью видеокарты, то есть является константой. В продолжение будет рассмотрена задача распознавания жестов младенцев на одном блоке как сравнение двух дистрибутивов (видеопотоков) здорового младенца и возможно с отклонениями в будущем. В принципе для задач распознавания количество блоков может быть переменным параметром, а не константой. В перспективе ставится цель решения задачи детекции объектов как композиция этих двух задач этих двух случайных процессов с разными видами неоднородности: колличество блоков и высота.

7. *Визуализация сеток как параметризованная модель белошумного случайного процесса*

Блок - это элемент хранения данных ЦМП (аналогично для задачи распознавания), матрица размером NxN, в каждой точке которой определена функция высоты над уровнем моря и константный цвет, соответствующий классу объекта, то есть задан сигнальный граф на решетке. ЦМП представляется набором таких матриц или блочной матрицей. При этом блок имеет иерархическую структуру – квадро-дерево, которая отражает разные уровни детализации по точности. Основной спецификацией визуализации является ее применение в VR (трехмерная графика), то есть в Банаховом пространстве (отображаются не точки, а интервалы), из-за этого между блоками образуются “стыки, дырки”, которые вызывают трудности при реализации алгоритмов рендеринга. Они должны особым образом обрабатывать границы блоков. В Гильбертовом пространстве (для растровой графики) таких проблем нет. Основное отличие трехмерного графического подхода от стандартного, применяемого в СДУ заключается в том, что предел ступенчатой функции должен быть определен не в топологии поточечной сходимости, а в компактно-открытой топологии, что сделано с целью построения непрерывного отображения (визуализации) с точки зрения зрительного восприятия.

При облете ЦМП подгружаются блоки с учетом функции минимума расстояния между камерой и блоком с наихудшей точностью, с дальнейшим улучшение точности. Приложение визуализации выполнено средствами WebGL при этом реализовано абстрагирование шейдеров аналогично абстрагированию параметров, то есть шейдер – это функция зависящая, например, от положения камеры. Фактически реализованы реактивные вычисления на уровне видеокарты. Конечно данное направление интересно, но как же отмечалось рассматриваться не будет.

Приведем виды отображения, реализованные на блоке. Самое простое –это облако точек, которое рассматриваться не будет, поскольку не является непрерывным отображением с точки зрения зрительного восприятия, в преобразование Лапласа оно называется “оригиналом”. На «рис. 4» для сравнения приведены два малоразличимых между собой вида отображения, но с разной моделью визуализации. Слева стандартная полигональная графика – барицентрические координаты. Справа визуализация столбиками. Столбик – это метафора. Этот вид отображения иногда называют статистической призмаграммой (четырехугольная призма) - трехмерный аналог диаграммы. С точки зрения математика – это обратное преобразование для “изображения” в

преобразование Лапласа, что и требуется показать, для начала по построению. (Визуализацию часто рассматривают как решение обратной задачи, скорее надо говорить о решение сопряженной задачи).

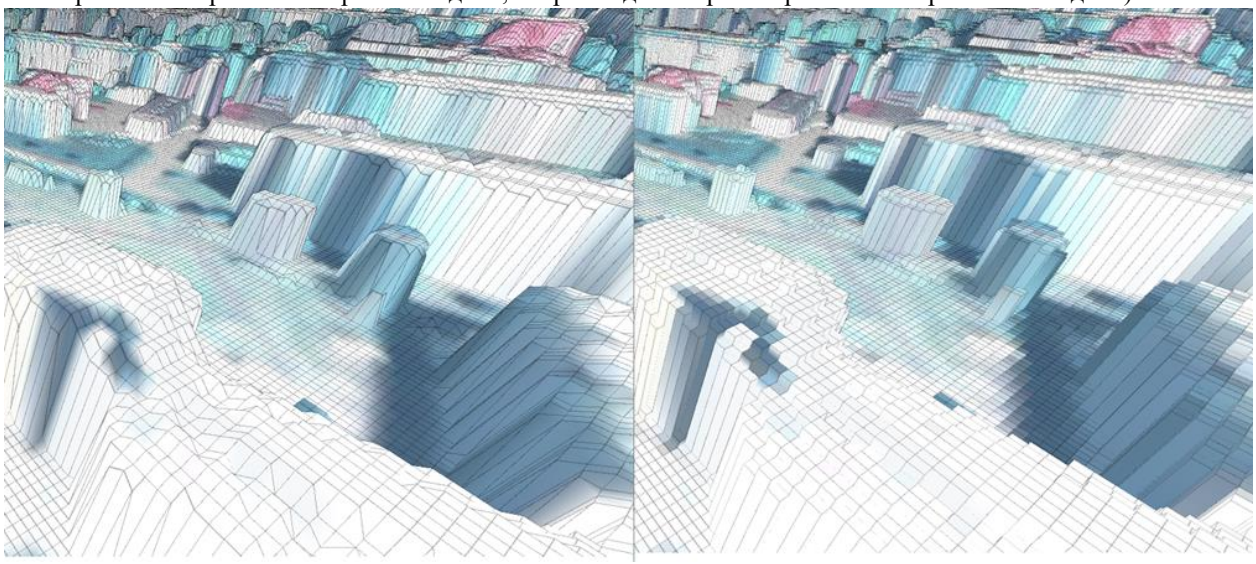


Рис. 4. Задачи визуализации цифровой модели поверхности. Слева поверхность изображена треугольниками, справа – столбиками.

На «рис. 5.» показано как строится статистическая диаграмма для двумерного случая. Формальное описание будет приведено ниже, пока что ограничимся декларативным с использованием когнитивной визуализации. Исторически термин когнитивная визуализации происходит из решения математических задач графическим способом [22]. Статистическая диаграмма - это направленная ступенчатая функция, функция высоты, математическое ожидание значения которой соответствует середина отрезка, интервала (показано красной точкой), что соответствует интегралу Стратоновича. В трехмерном пространстве двумерный интервал с учетом детализации и направления нормалей принято называть микрогранью. Например, в компьютерной визуализации микрограни используются для отображения шероховатой поверхности. Очевидна связь микрограни с конусом нормалей. С учетом уровня детализации (точности) можно рассматривать кратный интеграл Стратоновича по пространственным переменным (стационарный процесс), в пределе равный двойному. В дополнение на виде отображения, так же прорисовываются “ступеньки” - частные производные (показано красным отрезком), что сделано с целью непрерывности отображением с точки зрения зрительного восприятия. Стоит подчеркнуть, что для проекции на плоскость двойной интеграл Ито нарисовать столбиками не получится в отличие от проекции на сферу или цилиндр, применение, которых планируется при реализации волнового уравнения рендеринга как диффузионного процесса. Известно, что для одностороннего преобразования дисперсия стремится к бесконечности. Конечно, можно было бы отображать значение высоты в квадрате, но такой рисунок не имеет смысла. Кроме того, перспективная проекция и аффинные преобразования изображения являются линейными, поэтому не влияют на дисперсию (например, для дисперсии Алона), что важно при непосредственно взаимодействием с ЦМП.

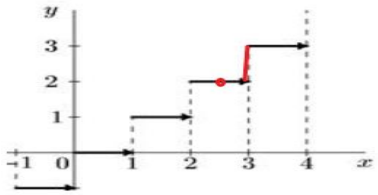


Рис. 5. Статистическая диаграмма - ступенчатая функция

Стоит отметить, что в качестве базового вида отображения выбрана визуализация столбиками, потому что, во-первых, в ЦМП много вертикальных линий –ступенек, во-вторых, меньше площадь дырок и они расположены вертикально, а не в горизонтальной плоскости как в случае с полигональной графикой «рис. 6».

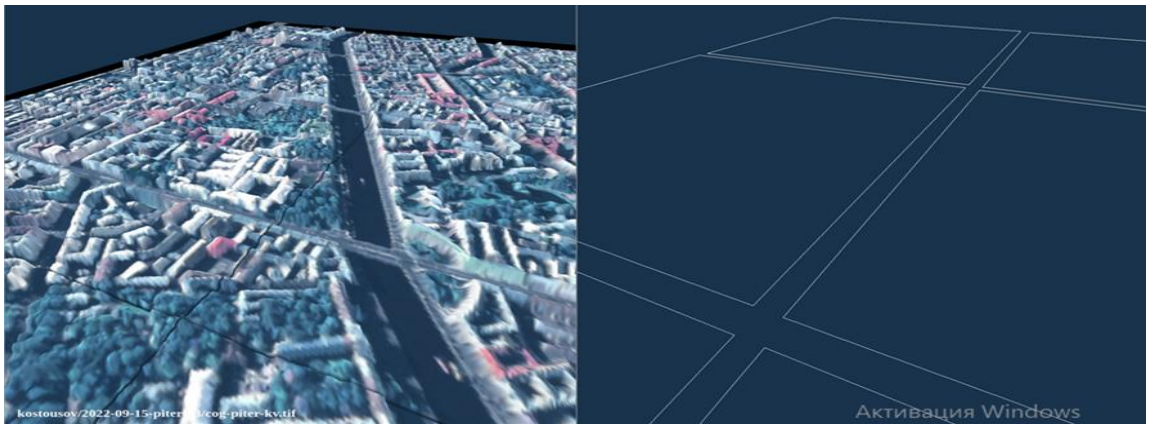


Рис. 6. Слева видны артефакты (синие линии, цвета фона на границах блоков), справа показано разбиение на блоки

Возникает вопрос можно ли убрать артефакты в изображение, например, рассмотрев транспортную задачу на границе блоков. В случае полигональной графики ответ очевиден: можно ввести фиктивную линию (теневую грань) на границе, и взять среднее значение высоты в точке (производную). Конечно, здесь есть определённые трудности в смысле программирования, на которых останавливаться не будем. Но этот подход в случае визуализации столбиками не сработает «рис. 7» приведен для сравнения артефактов эти двух видов отображений.

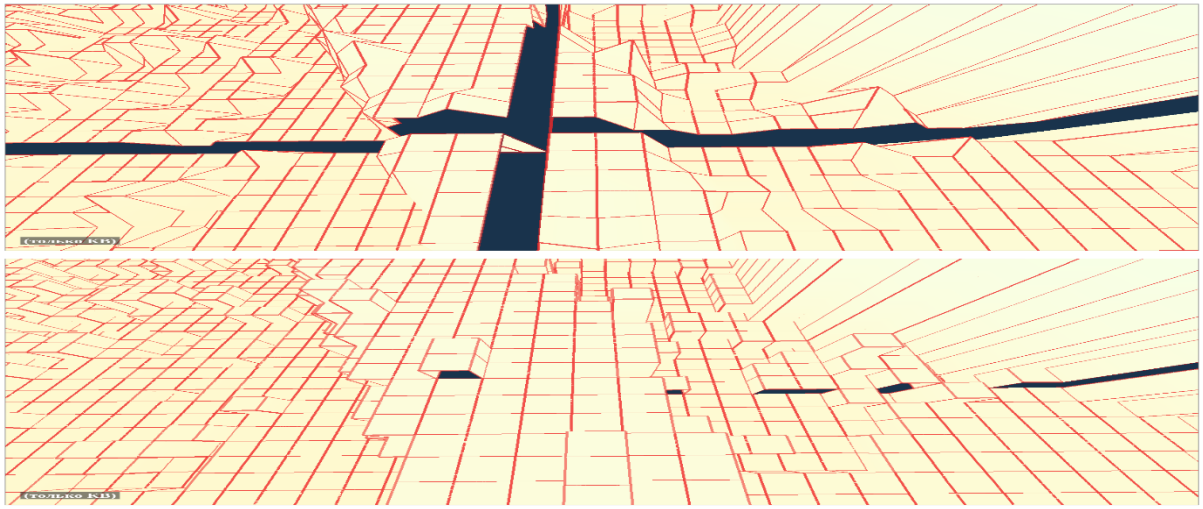


Рис. 7. Вверху артефакты в случае полигональной графики, внизу в случае визуализации столбиками

Чем же являются дырки в случае визуализации столбиками? Это интегральная метрика - метрика Васерштейна для частных производных:

$$\sum_{v_i \in \overline{V_j}} |||\nabla f(v_i)||_2^2,$$

где v_i – узел графа, $f(v_i)$ – высота, $\overline{V_j}$ – внутренняя граница блоков.

Интегральная метрика является более информативной чем векторное поле. Вряд ли математики рассматривали задачу убирания дырок, наверное, здесь можно поколдовать с распределение Дирихле или со стохастическим управлением, но подобное решение все равно приведет к обменам между блоками. Вероятно, можно обойтись без обменов, приблизительно доопределив частную производную на границе симметрично вниз с предыдущих ячеек, поскольку для неравенства Маркова, являющемся частным случаем неравенства Чебышева, выполняется правило трех сигм.

Этот же подход визуализации ЦМП применим и для трехмерных сеток, когда вместо /(графического) фильтра проекция интерактивно применяется фильтр сечения плоскостью (или сечения сферой для волнового уравнения рендеринга), как следствие возникает необходимость определения формулы Ито для фильтров. Так же сетка не обязательно должна быть регулярной, ее всегда можно реструктурировать по окта-дереву. В многомерном случае используется матрица рассеивания, которая на прямую связана с определением полностью заданного случайного процесса «рис. 8», где вместо “оригинала” (облака точек) используется визуализация столбиками.

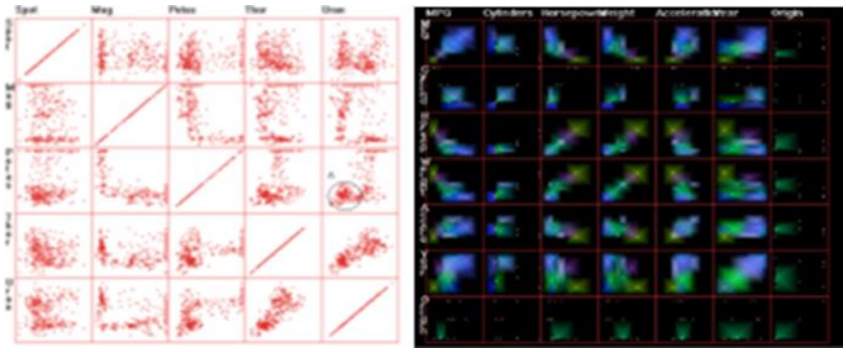


Рис. 8. Матрица рассеивания, слева для облака точек [23], справа для параллельных координат [24], которые можно рассматривать как полный дифференциал [25]

На этом декларативное описание задачи закончим, формализация в основном будет базироваться на монографии [26] Д. Ф. Кузнецова: “Некоторые вопросы теории численного решения стохастических дифференциальных уравнений Ито.”

8. Формализация задачи визуализации сеток

Начнем с обобщение интеграла Ито. Чтобы перейти к интегралу предел суммы определяется особым способом с использованием ступенчатой функции [4]:

$$\sum_i f(t_i^*, \omega)\,(B_{i+1} - B_i)(\omega)$$

Интеграл Ито: $t_i^* = t_i$ - левый конец отрезка. Обозначается:

$$\int_S^T f(t, \omega)\, dB_t(\omega).$$

Функция $f(t, \omega)$ является измеримой, согласованной и:

$$E(\int_S^T f(t, \omega)^2\, dt) < \infty$$

Важное свойство интеграла Ито состоит в том, что он является мартингалом

Интеграл Стратоновича: : $t_i^* = (t_{i+1} - t_i)/2$ - середина отрезка.

Можно определить **обобщение интеграла Ито** через предел по вероятности:

$$\int_0^t f(s, \omega)\, dB_s(\omega) = \lim_{n\to\infty} \int_0^t f_n(s, \omega)\, dB_s(\omega),$$

где $f, f_n \in W_H$,, f_n - ступенчатые функции : $\int_0^t |f_n - f|^2\, ds \to 0$ по вероятности (по мере относительно P).

Очевидно, что определение энтропийной устойчивости так же является обобщением интеграла Ито через предел по вероятности. Аналогичное определение используется в [26]:

Определение 1.2 Последовательность случайных величин $\xi_k(\omega)$ называется сходящейся с вероятностью единица или почти наверное к случайной величине $\xi(\omega)$: $\xi_k \overset{\text{п.н.}}{\longrightarrow} \xi$ при $k \to \infty$, если:

$$P\{\omega\colon \xi_k \to \xi \text{ при } k \to \infty\} = 1.$$

Эта последовательность так же называется фундаментальной сходящейся с вероятностью единица. Фундаментальность последовательности случайных величин является необходимым и достаточным условием существования ее предела, что называется критерием Коши.

Рассмотрим отличие стандартного определения полностью заданного случайного процесса от определения в Банаховом пространстве.

Случайный процесс считается полностью заданным, если заданы его конечномерные распределения - набор функций распределения, которые определены для любого $k \geq 1$ соотношениями:

$$F_\xi(x_1, \dots, x_k, t_1, \dots, t_k) = P\{\cap_{j=1}^k \{\xi(t_j, \omega) < x_j\}\},$$

где $x_j \in R^j$.

Справедливо и обратное утверждение, установленное Колмогоровым:

Если функции $F_\xi(x_1, \dots, x_k, t_1, \dots, t_k)$при всех $k \geq 1$ удовлетворяют условиям:

1. $F_\xi(x_1, \dots, x_k, t_1, \dots, t_k)$- является совместной функцией распределения k случайных величин;
2. справедливо тождественное равенство: $F_\xi(x_1, \dots, x_k, t_1, \dots, t_k) \equiv F_\xi(x_{i_1}, \dots, x_{i_k} t_{i_1}, \dots, t_{i_k})$;для любой перестановки $i_1, \dots, i_k$ чисел $1, \dots, k$;
3. $\lim_{x_k \to +\infty} F_\xi(x_1, \dots, x_k, t_1, \dots, t_k) = F_\xi(x_1, \dots, x_{k-1}, t_1, \dots, t_{k-1})$/

тогда существует такой случайный процесс $\xi(t, \omega)$, совместными функциями распределения которого являются функция :

$$F_{\xi}(x_1, \dots, x_k, t_1, \dots, t_k).$$

Таким образом, совокупность совместных функций распределения значений случайного процесса $\xi(t, \omega)$ является его исчерпывающей характеристикой.

Если функции распределения $F_{\xi}(x_1, \dots, x_k, t_1, \dots, t_k)$ имеют конечную смешанную k-ую производную, то существуют совместные плотности распределения значений случайного процесса ξ_t в соответствующие моменты времени:

$$p_{\xi}(x_1, \dots, x_k, t_1, \dots, t_k) = \frac{\partial^k}{\partial x_1 \dots \partial x_k} F_{\xi}(x_1, \dots, x_k, t_1, \dots, t_k).$$

В Банаховом пространстве матрица частных производных, как и ковариационная матрица, являются несимметричными (производная слева не равна производной справа). Для симметричной матрицы количество перестановок равно $k!$, а для несимметричной - 2^k. В общем случае надо применять теорему Синкхорна [21]. Гильбертово пространство является Банаховым, то можно рассматривать симметричные матрицы в пределе. В этом случае как раз и возникают кратные интегралы. Очевидно, что детализация на одном блоке задает сжимающее отображение, следовательно, можно применять теорему Банаха о неподвижной точке. Стандартно, для численного решения задач применяются Гауссовы процессы, центрирование фундаментальной последовательности и преобразование Лапласа.

Случайный процесс называется гауссовским, если все его совместные плотности распределения являются гауссовскими:

$$p_{\xi}(x_1, \dots, x_k, t_1, \dots, t_k) = \frac{1}{\frac{k}{2\pi^2}|K|^{\frac{1}{2}}} \exp\left(- - \frac{(x-m)^T K^{-1}(x-m)}{2}\right),$$

где $x = (x_1, \dots, x_k)^T$, $m = (m_1, \dots, m_k)^T$, $m_i = M\{m_i\}$, ковариационная матрица $K = K^T > 0$.

Процесс $\overset{\circ}{\xi_t} = \xi_t - M\{\xi_t\}$ называется центрированной составляющей процесса ξ_t .

Функция: $R_{\xi}(t_1, t_2) = M\left\{\overset{\circ}{\xi_{t_1}} \overset{\circ}{\xi_{t_2}}\right\}$ называется корреляционной функцией процесса ξ_t, причем:

$$D\{\xi_t\} = R_{\xi}(t, t),$$

где $D\{\xi_t\}$ -дисперсия случайного процесса ξ_t

Для того, чтобы функция $R_{\xi}(t)$ при $t \in (-\infty, +\infty)$ была корреляционной функцией стационарного в широком смысле случайного процесса ξ_t , удовлетворяющего условию: $M\{(\xi_{t+\tau} - \xi_t)^2\} \longrightarrow 0$ при $\tau \longrightarrow 0$, необходимо и достаточно, чтобы она допускала представление:

$$F_{\xi}(\omega) = \iint_{-\infty}^{\infty} e^{it\omega} dF_{\xi}(\omega), \text{где } F_{\xi}(\omega))$$

- произвольная неотрицательная ограниченная монотонно неубывающая функция, непрерывная слева.

$F_{\xi}(\omega)$ называется спектральной функцией, если она абсолютна непрерывна и

$$F_{\xi}(\omega) = \iint_{-\infty}^{\infty} S(u) du,$$

где $S(u)$спектральная плотность процесса ξ_t.

Очевидно, что частным случаем спектральной функции (например, RGB) является преобразование Фурье. Преобразование Фурье со сдвигом (точное значение не совпадает с его магматическим ожиданием) принято называть преобразованием Лапласа. Таким образом визуализация столбиками есть обратное преобразование Лапласа в Банаховом пространстве.

Очевидно, что задачу визуализации ЦМП можно свести к параметризованной модели белощумного случайного процесса [26] параграф 1.3, положив $\frac{p}{N} = \mu \to 0$, где количество блоков p $= const$, μ – малый параметр. Анализ динамики интерактивной визуализации при случайных внешних воздействиях пользователя сводится к исследованию вероятностных и статистических свойств решений систем дифференциальных уравнений, возмущенных случайными процессами (стохастическому управлению). Система дифференциальных уравнений для параметризованной модели белошумного случайного процесса имеет вид:

$$\dot{x_t} = a(x_t, u(t), t) + \sum(x_t, t) \frac{1}{\mu} \xi_{\frac{1}{\mu^2}}; x_0 = x(0).$$

Задача визуализации ЦМП является стационарной линейной (подробнее [26] параграф 1.3) и потому не очень интересная с точки зрения математики.

Определенную ценность имеет и разработка метафор визуализации и взаимодействия сопряженных с математической моделью. Визуализация столбика не такая уж элементарная метафора с учетом блочного подхода. Кроме того, она порождает другую метафору – интегральную метрику для частных производных как альтернатива векторному полю.

IV. Результаты экспериментов

Разработка системы онлайн визуализации и параллельных вычислений имеет прикладной характер, с ее помощью авторы намерены разрабатывать специализированные системы визуализации, в том числе связанные с результатами параллельных вычислений. Основные результаты экспериментов относительно параллельной части изложены в работе [9]. Для сложных систем валидации недостаточно нужна верификация, чему и посвящена данная работа. Важное направление – разработка языка программирование в данной работе упоминается, но не рассматривается, как и реактивные вычисления, частным случаем которых является параллельная фильтрация данных [5].

V. Обсуждение результатов

В рассматриваемых примерах возмущение в параметризованной модели белошумного случайного процесса связано с количеством данных: вычислительной точностью, длиной очереди задач, длиной сообщений. Задача визуализации ЦМП является стационарной линейной (подробнее [26] параграф 1.3) и потому не очень интересная с точки зрения математики, но сам блочный подход является перспективным, потому что есть более сложные задачи, например, разработка тренажеров с обратной связью, задача детекции на карте высот, рассмотрение которой начнем с задачи распознавания жестов младенцев.

Существует значительное количество работ по применению нейронных сетей для распознавания поведенческих шаблонов во временных рядах, в том числе и для жестов, но их использование не дает гарантии, что поставленная задача решена правильно. Фактически будет рассматриваться задача Коши, то есть при каких условиях задача распознавание жестов младенцев имеет решение.

Многолистный скелет определяется как многолистная фигура (плоская фигура с самопересечениями) [27]. Будет рассматриваться нетривиальная задача, когда многолистные скелеты младенцев анатомически подобны.

Срединная ось плоской фигуры представляет собой множество внутренних точек фигуры, каждая из которых имеет, по меньшей мере, две ближайшие граничные точки. Известны решения задачи коммивояжера для графов Делоне (Евклидово минимальное остовное дерево [17]), так же графы Делоне (диаграммы Вороного) применяются для распознавания жестов (считается, что расстояние между узлами графа (многолистного скелета) не меняется). Для решения поставленной задачи важным предположением является, то что узел графа (сустава) имеет площадь (задача с неопределенностью). В этом случае срединная ось как объединение прямых и параболических отрезков плоской фигуры, например, для сустава локтя, схематично представлена на «рис. 9».

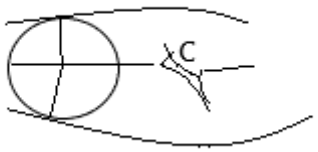


Рис. 9. Срединная ось сустава (с) – вогнутый четырех угольник, вершины которого соединены параболам

Численное, достаточно точное построение субдифференциала в трехмерном пространстве, как и срединной оси проблематично, поэтому будет рассматриваться другая задача – нахождение списка собственных значений линейно независимых подмножеств (окрестностей разных суставов) для карты высот, к которой применено преобразование Лапласа. Список собственных значений фактически определяет субэнтропию. Если два дистрибутива здорового и потенциально больного младенца разделяются, то у них должен быть разный набор передаточных функций (жестов), следовательно, и собственных значений. Дистрибутив представляет собой карту высот, изменяющуюся во времени. Берем два близких по времени кадра вычитаем один из другого получаем матрицу с большим количеством “нулевых” элементов, имеющую блочную структуру. К блоку применяем преобразование Лапласа. Строим по нему передаточную функцию, которая существует по теореме Синкхорна [21]:

Если A — матрица размера $n \times n$ со строго положительными элементами, то существуют диагональные матрицы D_1 и D_2 со строго положительными диагональными элементами такие, что $D_1 A D_2$ дважды стохастическая матрица. Матрицы D_1 и D_2 уникальны по модулю умножения первой матрицы на положительное число и деления второй на то же число. [21].

Простой итерационный метод приближения к двойной стохастической матрице состоит в том, чтобы поочередно масштабировать все строки и все столбцы матрицы A, чтобы в сумме получить 1.

Поскольку рассматривается задача на остовных графах с неопределенностью, возможно, необязательно брать разность кадров, достаточно нахождение собственных значений для бесконечной последовательности кадров (для каждого кадра по отдельности).

Можно сделать вывод, что задача распознавания жестов младенцев является перспективной в плане решения, хотя известного математического аппарата явно недостаточно.
Авторов интересуют и другие постановки задач, связанные с применением стохастической семантики в области визуализации, имеющие как прикладной, так и теоретический характер. Например, формула Ито для графических фильтров (параллельная фильтрация данных), рассмотрение волнового уравнения рендеринга как параметризованной модели белощумного случайного процесса (геометрическое решение).

VI. Выводы и заключение

Для верификации программного обеспечения, как и для визуализации в данной работе использованы СДУ, теория информации, сигнальные графы. Этот подход назван стохастической семантикой. Важно, что визуальный процесс является параллельным процессом (взаимодействующие последовательные процессы Хоара) с точки зрения программирования и случайным процессом с точки зрения математического моделирования.
Расмотрение задач, связанных с действительно большими данными, неизбежно ведет к использованию блочного подхода. В параллельных вычислениях блок можно ассоциировать с процессором и рассмотреть задачу максимизации энтропии (производительности). В разрабатываемой динамической системе онлайн визуализации и параллельных вычислений для геометрического распараллеливания можно реализовать и сравнить стационарный случайный процесс и установившийся случайный процесс, которые имеют разные аналитические решения. Это позволяет сделать вывод, что предложенная реализация стационарного процесса имеет определенную новизну.
В области верификации визуализации сделано не так много – предложена визуализация сеток, рассматриваемая как параметризованная модель белошумного случайного процесса. Авторов интересуют и другие постановки задач, связанные с применением стохастической семантики в области визуализации. Особо хочется отметить исследовательскую серию работ по обобщенному вычислительному эксперименту [28].
Конечно, данную работу нельзя считать завершенной, но направление, которое авторы назвали стохастическая семантика очевидно является перспективным.

Список литературы

1. Croitoru F-A., Hondru V., Tudor Ionescu R., and Shah M. Diffusion models in vision: a survey, // IEEE Transactions on pattern analysis and machine intelligence 14(8): 2022. P. 1-25.
2. L´ezoray O. and Grady L Image Processing and Analysis with Graphs: Theory and Practice, CRC Press, July 2012. Graph Signal Processing and Applications (univ-amu.fr).
3. Cheng-Zhong Xu, Francis C., Lau M. Optimal parameters for load balancing with the diffusion method in mesh networks. // Parallel processing letters, Volume 4, 1994 P. 139-147.
4. Oksendal B. (2003). Stochastic differential equations: an introduction with applications. Berlin: Springer. ISBN 3-540-04758-1.
5. Манаков Д., Авербух В. Верификация визуализации // Научная визуализация 2016. Кв.1. Том 8. N: 1. С. 58 - 94.
6. Lakoff G. The contemporary theory of metaphor // Metaphor and thought. (2nd ed.). Cambridge: Cambridge university press, 1993, P. 202-251.
7. Scott D. S. Data types as lattices // Proceedings of the international summer institute and logic colloquium, Kiel, in Lecture notes in mathematics. Springer-Verlag. 499. P. 579-651.
8. Авербух В.Л., Манаков Д.В. Анализ и визуализация “больших данных” // Труды международной научной конференции “Параллельные Вычислительные Технологии” (ПаВТ'2015). Екатеринбург, 31 марта - 2 апреля 2015. Челябинск, Издательский центр ЮУрГУ. 2015. С.332-340.
9. Васёв П.А. Визуализация работы алгоритма планирования параллельных задач. // ГрафиКон 2023: 33-я Международная конференция по компьютерной графике и машинному зрению, 19-21 сентября 2023, Москва: труды. С. 341-353. DOI: https://doi.org/10.20948/graphicon-2023-341-353
10/ Котов В. Е., Проблемы развития параллельного программирования // Труды Всесоюзного симпозиума “Перспективы системного и теоретического программирования”. Новосибирск, 1979. С. 58-72
11. Biberstein O., Buchs D., and Guelfi N. //Object-oriented nets with algebraic specifications: The CO-OPN/2 formalism. / Agha G., De Cindioand F, and Rozenberg G. editors, Advances in Petri nets on object-orientation, LNCS. Springer-Verlag, 2001
12. Кудрин К.А., Коварцев А.Н., Прохоров С.А. Методы автоматизации отладки в технологии графо-символического программирования //Сборник научных трудов "Информационные системы и технологии",Самара, 1996. С. 75-79.
13. Averboukh Y. Lattice approximations of the first-order mean field type differential games. Nonlinear Differ. Equ. Appl. 28, 65 (2021). DOI: 10.1007/s00030-021-00727-2

14. Стратонович Р. Л. Теория информации. М.: Сов. радио, 1975. 424 с.
15. Baimuratov I., Nguyen T., Golchin R., Mouromtsev D. Developing non-empirical metrics and tools for ontology visualizations evaluation and comparing. Scientific Visualization. 2020. Vol. 12. No. 4. P. 71-84. DOI: 10.26583/sv.12.4.07.
16. Shih M., Zhang Y., Ma K.-L., Sitaraman J., Mavriplis D. Out-of-Core Visualization of TimeVarying Hybrid-Grid Volume Data // IEEE Symposium on Large Data Analysis and Visualization. 2014. P. 93 – 100.
17. Незнахина Е. Д. PTAS для задачи Min-k-SCCP в евклидовом пространствепроизвольной фиксированной размерности //Труды Института математики и механики УрО РАН. 2015. Т. 21. N: 3. С. 268-278.
18. Datta N., Jozsa R., Dorlas T. Benatti F. Properties of Subentropy, // J. Math. Phys. **55** (2014) 062203. **doi: 10.1063/1.4882935**
19. Теплов А.М. Об одном подходе к сравнению масштабируемости параллельных программ // Вычислительные методы и программирование. 2014. Т. 15. Выпуск 4. С. 697-711
20. Розенберг В.Л. Сферическая блоковая модель динамики и сейсмичности литосферы: современное состояние и перспективы развития // Современные методы оценки сейсмической опасности и прогноза землетрясений: III Всероссийская научная конференция с международным участием, 25 -26 октября 2023, Москва: материалы. Москва: ИТПЗ РАН, 2023. С. 224-228.
21. Sinkhorn, Richard. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices // Ann. Math. Statist. **35**. P. 876–879. doi:10.1214/aoms/1177703591
22. Зенкин А.А. Когнитивная компьютерная графика / Под ред. Д.А. Поспелова. М.: Наука, 1991. 192 с.
23. Cui Q., Ward M., Rundensteiner E., and Yang J. Measuring data abstraction quality in multiresolution visualizations // IEEE TVCG, 12(5): 2006. P. 709–716,
25. Манаков Д.В. Модели абстракции данных: выборка (параллельные координаты), фильтрация, кластеризация // Научная визуализация, 2019, Кв.1.Том 11. N: 1. С. 139 - 176, DOI: 10.26583/sv.11.1.11
26. Кузнецов Д. Ф. Некоторые вопросы теории численного решения стохастических дифференциальных уравнений Ито // Дифференциальные уравнения и процессы управления. 1998. N 1. 367 с.
27. Мехедов И. С. Многолистная плоская фигура и её серединная ось // Известия вузов. Математика. 2011. № 12. С. 42–53
28.. Алексеев А.К, Бондарев А.Е., Пятакова Ю.С. Об использовании канонического разложения для визуализации результатов параметрических расчетов // Научная визуализация, 2023, Кв.4.Том 15. N: 4. С. 12 - 23, DOI: 10.26583/sv.15.4.02

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-347-353

Нормативное обеспечение и применение методов визуализации данных
Regulatory support and application of data visualization methods

Л.А. Максименко
Сибирский государственный университет геосистем и технологий, Новосибирск, Российская Федерация

L. A. Maksimenko
Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation

Аннотация. В статье исследуется роль и значение визуализации данных в современном информационном обществе. Рассматриваются нормативные аспекты применения метода визуализации, включая вопросы классификации, безопасности, конфиденциальности и этики использования данных. Особое внимание уделяется использованию методов визуализации для работы с открытой информацией и открытыми данными для представления графиков, диаграмм, больших данных, наборов данных, информационных панелей и интерактивных отчетов. При выполнении работы были приняты во внимание стандарты ЕСКД и СПДС, содержащие правила и рекомендации по графическому оформлению объектов; рассмотрены BI-инструменты для создания дашбордов. При этом подчеркивается, что современные тренды цифровизации и автоматизации выдвигают дополнительные требования к обработке информации, что может трактоваться как новый этап в области визуализации данных.

***Ключевые слова*:** визуализация, BI-инструменты, Power BI Desktop, метод визуализации, открытая информация, открытые данные, графики, диаграммы, большие данные, датасеты, дашборды.

Аннотация. The article examines the role and importance of data visualization in the modern information society. The regulatory aspects of the application of the visualization method are considered, including issues of classification, security, confidentiality and ethics of data use. Special attention is paid to the use of visualization techniques to work with open information and open data to present graphs, charts, big data, datasets, dashboards and interactive reports. When performing the work, the ESCD and SPDS standards were taken into account, containing rules and recommendations for the graphic design of objects; BI tools for creating dashboards were considered. At the same time, it is emphasized that modern trends in digitalization and automation put forward additional requirements for information processing, which can be interpreted as a new stage in the field of data visualization.

***Ключевые слова*:** visualization, BI tools, Power BI Desktop, visualization method, open information, open data, graphs, charts, big data, datasets, dashboards.

I. Введение

В настоящее время происходит стремительное накопление информации, которое исчисляется большими объемами данных. В неформальной обстановке информация и данные могут рассматриваться как синонимы, хотя они не являются идентичными. Информация всегда имеет определенную форму представления и преобразовывается в данные посредством ее обработки. При работе с информацией должны быть определены процедуры сбора и хранения информации; обработки и подготовки информации к размещению; размещения информации; актуализации (регулярного обновления) информации. Цифровая среда открывает новые возможности для визуального моделирования информации, а наличие развитых графических инструментов позволяет эффективно решать задачи интеллектуального анализа данных, с выбором удобных вариантов отображения результатов. Графическая форма отображения информации осуществляется посредством различных методов и инструментов включая графики и диаграммы, карты, интерактивные визуализации, дашборды и инфографику. Инфографика (информационная графика), как правило, представляет собой комбинацию графических элементов, текста и изображений для передачи сложной информации в упрощенной и привлекательной форме. Исследовательская визуализация данных создается алгоритмическим путем, на основе репрезентативных выборок, охватывающих большие объемы данных, и представляет собой законченные аналитические решения от извлечения данных из различных источников до формирования отчетности. Основное различие между концепциями визуализации и инфографики заключается в больших объемах данных, с которыми могут работать визуализаторы; инфографика же, наоборот, создает изображения, не содержащие большого объема данных.

II. Постановка задачи

Целью работы является исследование нормативного обеспечения и применения методов визуализации данных, влияющих на выбор эффективного способа представления данных, в удобном для пользователей формате, на основе современных программ. Нормативное обеспечение метода визуализации данных направлено на установление стандартов и правил использования визуализации данных в различных областях. Некоторые аспекты нормативной поддержки метода визуализации данных могут включать в себя:

- установление стандартов представления и интерпретации данных в графическом формате;
- установление объектов для применения методов визуализации данных;
- регулирование использования конкретных инструментов и технологий визуализации данных;
- определение правил сохранения и безопасности данных, визуализируемых в различных форматах;
- обеспечение конфиденциальности и защиты данных при использовании методов визуализации.

Для достижения максимальной эффективности и понятности в представлении данных и информации важно соблюдать правила оформления и использования соответствующих методов визуализации. К настоящему времени как в учебном, так и в рабочем процессе преобладают вопросы использования специализированного программного обеспечения для создания трехмерных моделей, визуализации и анимации, в то время как правила оформления чертежей, графиков и диаграмм практически не рассматриваются.

Актуальность работы обусловили вопросы комплексной подготовки специалистов не только работе с программным обеспечением, но и навыкам оформления и документирования результатов работы в соответствии с требованиями стандартов. В связи с введением распоряжения Правительства РФ от 21 декабря 2021г. №3759-р «Об утверждении стратегического направления в области цифровой трансформации науки и высшего образования» [1] в статье рассматриваются вопросы о применении метода визуализации информации с использованием графиков, диаграмм и дашбордов в рамках обновления и расширения учебного процесса.

III. Теория

Открытость информации как концептуальное направление развития информационных систем регулируется информационными системами в сфере стандартизации, которые функционируют на основе информационных технологий и технических средств, обеспечивающих сбор, обработку, хранение, размещение, использование и предоставление информации, касающейся планирования работ по стандартизации, в том числе формирования и реализации программ по стандартизации, разработки, редактирования и экспертизы проектов документов по стандартизации в соответствии Федеральным законом "О стандартизации в Российской Федерации" от 29 июня 2015 года N 162-ФЗ "О стандартизации в Российской Федерации". Национальный стандарт российской федерации ГОСТ Р 56907-2016 «Бережливое производство. Визуализация» внесен Техническим комитетом по стандартизации ТК 076 "Системы менеджмента" , утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 31 марта 2016 г. N 232-ст 4. Настоящий стандарт является руководством по использованию метода визуализации на основе рекомендуемых принципов в соответствии с ГОСТ Р 56407. Правила применения настоящего стандарта установлены в ст. 26 Федерального закона от 29 июня 2015 г. N 162-ФЗ "О стандартизации в Российской Федерации". В соответствии с ГОСТ Р 56907: «Метод визуализации (visualization method): Систематизированная совокупность действий по визуализации объектов. Метод визуализации применяется в организации с целью представления информации в наглядной форме (рисунок, фотография, график, диаграмма, схема, таблица, карта и т.п.) и доведение ее до сведения персонала в режиме реального времени для анализа текущего состояния и принятия обоснованных и объективных решений».

Федеральный закон от 27.07.2006 N 149-ФЗ (ред. от 12.12.2023) "Об информации, информационных технологиях и о защите информации" определяет и нормы правового регулирования отношений в сфере информации, информационных технологий и защиты информации [2]. Стандарты графических символов, используемые для технической документации на продукцию представлены в [3].

Объекты метода визуализаций определены в [4] и представлены на рис. 1.

Визуализация данных с точки зрения классификации по предметным областям, как показано на рис. 2. может быть представлена графиками и диаграммами, таблицами и схемами, дашбордами, а с развитием программ компьютерного моделирования отдельным видом может быть представлена «предметная» визуализация.

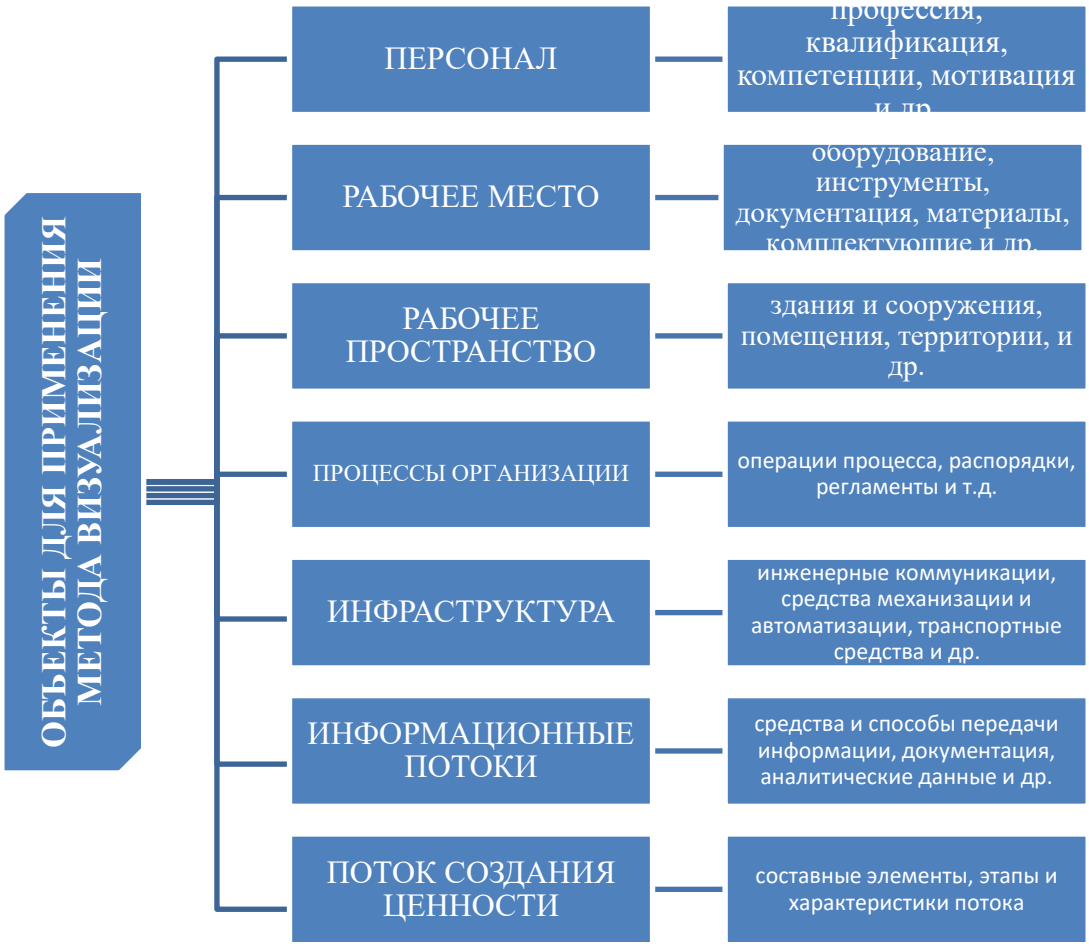


Рис. 1. Объекты метода визуализации данных

Рис. 2. Классификация методов визуализации данных

Предметная или 3D визуализация, создает копию близкую к реальному объекту. Получение качественной модели трудоёмкий процесс, требующий достаточно высоких профессиональных навыков от разработчиков для выбора и настройки текстуры поверхностей, освещения, проведения рендеринга и других параметров при работе с программным обеспечением.

Визуализация и применение условных графических обозначений для визуальной идентификации информации зависят от функциональности, определенной САПР и могут меняться в зависимости от используемой автоматизированной системы, но подготовка 2D графической документации всегда проводится в соответствии с требованиями стандартов [5]. Богатый функционал графических примитивов современных программ [7] для пользователей не знакомых с требованиями ГОСТ создает иллюзию их свободного выбора, что ведет к ошибкам в оформлении графиков, диаграмм, выборе масштаба отображения чертежа, начертания линий др. Поэтому в период обучения, независимо от специальности важно изучать вышеупомянутые стандарты в рамках той или иной дисциплины.

Базовые стандарты ЕСКД, содержащие правила и рекомендации для графической визуализации объектов, рекомендуемые к изучению приведены на рис. 3.

Рис. 3. Базовые стандарты ЕСКД

Среди графических программ, адаптированных к отечественной нормативной базе ЕСКД и СПДС, заметно выделяется графический редактор NanoCAD (разработка от Nanosoft). ПО NanoCAD включено в российскую базу программного обеспечения. Модульная структура ПО NanoCAD позволяет пользователям перейти от базового уровня работы к технологиям BIM проектирования.

IV. Результаты экспериментов

Как следует из рис.1 [4] к объектам визуализации относится документация. В проектной деятельности документация разделяется на графическую и текстовую. К текстовой документации относят все виды таблиц, оформление которых регламентирует ГОСТ 2.105 [6], а к графической документации относят чертежи и схемы. Правила оформления строительной документации регламентированы в стандартах СПДС 21.1101[7]. Новые технологии компьютерного моделирования объектов недвижимости связали в единый блок 3D модель и чертежи, построенные по модели. Комплекс вопросов, связанных с оформлением результатов таких работ в той или иной мере связан с графической визуализацией объекта недвижимости.

В работе [9] рассматриваются вопросы о профессиональной подготовке будущих специалистов в области кадастрового учета и регистрации прав к графической визуализации объектов недвижимости, как вариант, в рамках проведения отдельной учебной дисциплины «Графическая (пространственная) визуализация объектов недвижимости». Практически, вопросы графической визуализации объектов недвижимости могут успешно решаться в рамках таких дисциплин, как «Градостроительство и планировка населенных мест», «Типология объектов недвижимости», «Техническая инвентаризация объектов недвижимости», «Технические средства отображения информации в графическом и семантическом виде», «Разработка и создание цифровых тематических карт (схем)», «Построение 3D моделей местности для целей землеустройства и кадастров» и др. , если в рабочих программах предусмотрены заданию по моделированию, обмену и оформлению чертежей в соответствии с требованиями нормативных документов, с учетом использования открытого формата данных (IFC форматы), которые доступны для всех заинтересованных сторон и не зависят от конкретного производителя программного обеспечения (ArchiCAD, NanoCAD, Tekla Structures, Revit, Renga и др.). Важно отметить, что

формат IFC постоянно совершенствуется. В него добавляются новые атрибуты и свойства, позволяющие передавать дополнительную важную информацию. С каждым годом IFC демонстрирует все большую эффективность в создании проектов с использованием информационного моделирования, являясь основным форматом для обмена данными в области проектирования. Для представления результатов 3D моделирования, в форме необходимой для подготовки графической части технического плана, требуется подготовка чертежей поэтажных планов объекта недвижимости. Файлы формата IFC можно использовать в качестве подосновы для оформления поэтажных планов. Например, если поэтажный план оформляется в Renga, то достаточно открыть файл формата IFC, перейти в раздел «Чертежи» и перенести на чертежный вид нужный план этажа. ПО NanoCAD также предусмотрено выполнение этих действий рис. 4.

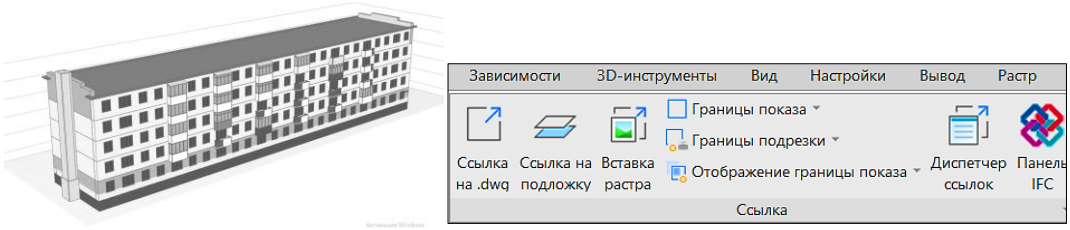


Рис. 4. Панель IFC

На рис. 5 представлен фрагмент поэтажного плана, сформированный по 3D модели в Renga.

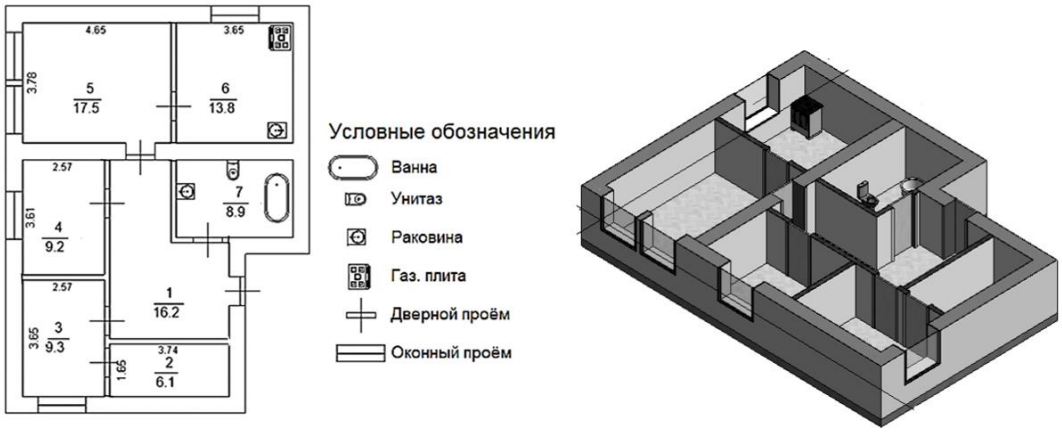


Рис. 5. Формирование и оформление поэтажного плана

Обобщая вышеизложенный материал можно отметить, что процессы подготовки и оформления чертежей, в частности поэтажных планов, в настоящее время вышли на новый уровень цифровой предметной визуализации графической документации.

Другим не менее важным аспектом нормативного обеспечения и применения методов визуализации в кадастровой деятельности являются вопросы визуализации данных. Графики и диаграммы являются неотъемлемой частью практически любого отчета. С увеличением объема данных появляется необходимость их эффективной обработки. Визуализация результатов является ключевым фактором для понимания информации. Вопросам подготовки к такой деятельности в учебном процессе практически времени не отводится. Элементарные навыки работы в Excel и Word позволяют обучающимся выполнять несложные задания по построению графиков и диаграмм, но для работы с большими данными требуется работа в более профессиональных системах, таких как POWER BI и др.

В ГОСТ Р ИСО/МЭК 20546-2021 «Информационные технологии. Большие данные. Обзор и словарь» раскрывается терминологический аппарат, даются понятия таким терминам как датасет, озеро данных и др, поясняющие процессы сопровождающее накопление цифровой информации в мире. Огромное поле деятельности для обновления и дополнения информации в кадастровых базах данных. выявления новых тенденций и закономерностей представляют собой такие ресурсы как: национальная система пространственных данных (НСПД), Национальная система пространственных данных (НСПД), Геоданные OpenStreetMap (OSM), Геопространственные данные NASA. Базы данных по недвижимости, такие как ЦИАН (Центр информации по недвижимости), Avito Недвижимость, Яндекс. могут быть использованы для общего представления о недвижимости на рынке и для анализа тенденций цен и предложений. Кадастровая информация представлена также в источниках [10-17].

С учетом наметившихся тенденций роста данных и новых технологий для их обработки следует отметить наиболее эффективные методики визуализации данных [18]. Дашборды все чаще появляются на страницах web-сайтов государственных органов и учреждений. Для подготовки подобной документации требуются знания, умения и навыки в подготовке и обработке данных, техниках визуализации и публикации информации. Преимущества обработки информации в BI-системах приведены на рис. 6.

Рис. 6. BI -система

Примеры подготовленных дашбордов в рамках изучения дисциплины «Системы интеллектуальной обработки кадастровой информации» приведены в работах [18-20].

V. Обсуждение результатов

Часто специалисты, обладающие навыками работы в специализированном ПО, не умеют оформлять результат работы, в соответствии со стандартами ЕСКД и СПДС. Причины, по которым это происходит могут быть следующими: недостаточное понимание стандартов ЕСКД и СПДС; ориентация на визуальные аспекты, игнорируя стандарты оформления документации; не хватает практики и опыта в оформлении документации; считают оформление документации менее важным, чем создание моделей и визуализаций. Важно также отметить, что возможности современных информационно коммуникационных и цифровых технологий по визуализации объектов недвижимости еще недостаточно используются.

VI. Выводы и заключение

В соответствии с поставленной целью исследования нормативного обеспечения и применения методов визуализации данных, влияющих на выбор эффективного способа представления данных, в удобном для пользователей формате, на основе современных программ были разработаны рекомендации по применению метода визуализации в кадастровой деятельности:

- при подготовке поэтажных планов объектов недвижимости рекомендуется применение базовых стандартов ЕСКД и СПДС, содержащих правила и рекомендации для графической визуализации чертежа;
- при работе с информацией должны быть определены процедуры сбора и хранения данных; обработки и подготовки информации к размещению; размещения информации; актуализации (регулярного обновления) информации ответственным лицом;
- актуализация информации, предполагает периодичность сбора и размещения данных; ответственность за достоверность данных; формат представления данных;
- наиболее эффективными способами представления кадастровой информации являются интерактивные отчеты, доступные для совместного обновления данными и принятия решений.

Рассмотрение теоретических вопросов нормативного обеспечения и применения методов визуализации позволило обобщить и сделать выводы об объектах визуализации данных, рассмотреть современные классификации методов визуализации, влияющих на выбор эффективного способа представления кадастровой информации.

Список литературы

1. Распоряжение Правительства РФ от 21 декабря 2021г. №3759-р «Об утверждении стратегического направления в области цифровой трансформации науки и высшего образования»
2. Об информации, информационных технологиях и о защите информации [Электронный ресурс] : Федеральный закон от 27.07.2006 №149-ФЗ (ред. от 14.07.2022). – Доступ из справ.-правовой системы «Консультант Плюс».
3. Стандарты графических символов, используемые для технической документации на продукцию. https://www.muayene.com/ru/standartlar/genellemeler-terminoloji-standardizasyon-dokumantasyon-standartlari/grafik-semboller-standartlar/teknik-urun-dokumantasyonlari-icin-kullanilan-grafik-semboller-standartlari
4. Национальный стандарт российской федерации ГОСТ Р 56907-2016 «Бережливое производство
5. Единая система конструкторской документации. Общие требования к ГОСТ 2.301 - 2.321
6. Единая система конструкторской документации. Общие требования к текстовым документам. ГОСТ 2.105-2019
7. Каталог визуализации данных [Электронный ресурс]. – Режим доступа: https://datavizcatalogue.com/RU/.
8. ГОСТ Р 21.101-2020 Система проектной документации для строительства. Основные требования к проектной и рабочей документации. Дата актуализации: 01.01.2021
9. Федотова В. С. Профессиональная подготовка будущих специалистов в области кадастрового учета и регистрации прав к графической визуализации объектов недвижимости / В. С. Федотова // Педагогика. Вопросы теории и практики
10. Агентство недвижимости ЦИАН [Электронный ресурс]. – Режим доступа: https://www.cian.ru/analiz-rynka-nedvizhimosti-b2b/.
11. Real Capital Analytics [Электронный ресурс]. – Режим доступа: https://app.rcanalytics.com/.
12. О государственной регистрации недвижимости [Электронный ресурс] : Федеральный закон от 13.07.2015 №218-ФЗ (ред. от 20.10.2022). – Доступ из справ.-правовой системы «Консультант Плюс».
13. Росреестр. Статика и аналитика [Электронный ресурс]. – Режим доступа: https://rosreestr.gov.ru/open-service/statistika-i-analitika/.
14. Портал открытых данных РФ [Электронный ресурс]. – Режим доступа: https://data.gov.ru/.
15. Сайт федеральной службы государственной статистики [Электронный ресурс]. – Режим доступа: https://rosstat.gov.ru/.
16. Набор открытых данных Министерства просвещения РФ [Электронный ресурс]. – Режим доступа: https://opendata.edu.gov.ru/opendata/.
17. Открытые данные на сайте Министерства труда [Электронный ресурс]. – Режим доступа: https://mintrud.gov.ru/opendata.
18. Microsoft Power BI Desktop. Бесплатное приложение для создания отчетов и аналитики данных, разработанное компанией Microsoft [Электронный ресурс]. – Режим доступа: https://powerbi.microsoft.com/ru-ru/desktop/.
19. Евстифеева А. А. Визуализация данных и дашбординг для отображения кадастровой информации [Электронный ресурс] // Сборник тезисов докладов: в 2 ч.– Новосибирск : СГУГиТ, 2023. Ч. 1. – С. 33–34. – Руководитель: Л. А. Максименко. – Режим доступа: https://sgugit.ru/student/research-work/issc/ Сборник РСНК 2023-1.pdf (дата обращения: 04.07.2023).
20. Максименко Л. А. Сбор и обработка кадастровой информации в сфере управления недвижимым имуществом / Л. А. Максименко. – DOI 10.33764/2411-1759-2024-29-1-118-126. – Текст : непосредственный // Вестник СГУГиТ (Сибирского государственного университета геосистем и технологий) – 2024. – Т. 29, № 1. – С. 118–126/
21. Линовская А. В. Динамика развития кадастрового учета линейных объектов / А. В. Линовская, Л. А. Максименко. – DOI 10.33764/2687-041X-2024-2-154-158. – Текст: непосредственный // Регулирование земельно-имущественных отношений в России: правовое и геопространственное обеспечение, оценка недвижимости, экология, технологические решения: Сб. материалов 7 Нац. науч.-практ. конф. с междунар. участием, Новосибирск, 21–24 нояб. 2023 г. : в 3 ч. – Новосибирск : Изд-во СГУГиТ, 2024. – Ч. 2. – С. 160–164. –

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-354-360

Сетевая технология программирования визуальных интерфейсов
A network technology for visual interfaces programming

П. А. Васёв
Институт математики и механики им. Н.Н. Красовского УрО РАН, г. Екатеринбург, Россия

P. A. Vasev
N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia

Аннотация. В работе предлагается технология управления визуальными образами, такими как пользовательские интерфейсы и сцены визуализации. Технология базируется на сетевом взаимодействии. Это позволяет реализовать программные компоненты, отвечающие за визуализацию, на одних языках и библиотеках, а общую компоновочную и логическую части на других. Это в свою очередь обеспечивает возможность запуска проекта в разных средах (веб и настольные приложения), привлечение более широкого круга разработчиков, переносимость проекта между разными поколениями графических технологий.

***Ключевые слова*:** визуализация, архитектура программного обеспечения, пользовательские интерфейсы

Annotation. The paper proposes a technology for managing views, such as user interfaces and visualization scenes. The technology is based on network interaction, so it is possible to describe the software components responsible for direct visualization in some languages and libraries, and the general layout and logical parts in others. This makes it possible to launch the project in different environments (web and desktop applications), attract a wider range of developers, and portability of the project between different generations of graphics technologies.

***Ключевые слова*:** visualization, software architecture, graphical user interfaces

I. Введение и постановка задачи

Технологии графических интерфейсов [1] и библиотек стремительно меняются. Это, например, означает, что проект, написанный 10 лет назад, становится морально устаревшим. Под моральным устареванием понимается то, что появляются технологии с более логически выверенными программными интерфейсами, лучшими привязками к оборудованию, лучшей образностью, и т.п. Например, на смену когда-то впечатляющей технологии QML пришли новые технологии, такие как React или Kotlin Compose Multiplatform.

Но тоже самое произойдёт и с этими новыми технологиями. Важным аспектом многих подобных технологий является то, что они предлагают сквозное решение, которое охватывает все основные аспекты программ. Это означает, что если программу реализовывать с использованием подобных технологий, то она вскоре морально устареет вместе с этими технологиями. Это является проблемой.

Что делать? Человечество решает этот вопрос, постепенно двигаясь к разделению графических программ на слои. В целом, по мнению автора, такое разделение наблюдается следующее:

1. Графический слой – элементы управления, визуальные образы, т. е. компоненты непосредственно взаимодействующие с оконной системой и графическими библиотеками уровня Vulkan/OpenGL/etc.
2. Логический слой – абстракции более высокого уровня, в которых программируются прикладные модели. Этот слой опирается на графический слой, использует его в своих алгоритмах.
3. Компоновочный слой — увязывает элементы графических и логических частей в более крупные абстракции, формирует итоговый вид программы.
4. Дополнительные прикладные части и библиотеки.

Например, следующие технологии опираются на подобное разделение: среда X Window System, парадигма Model-view-controller, клиент-серверные веб-технологии и их вариации, такие как Ruby on Rails, и уже упоминавшийся QML.

Имеющиеся решения не идеальны. Как отмечено выше, многие технологии являются сквозными, т. е. предполагают реализацию всех слоёв в рамках одной экосистемы. Другие являются достаточно сложными для

взаимодействия слоёв, например веб-технологии уровня HTML+Javascript подразумевают для взаимодействия слоёв применение низкоуровневых протоколов (Websocket, Server sent events и т. п.).

Одним из кардинальных методов решения обозначенной проблемы является переход на другой уровень абстракции. Например, в среде SciVi [2] предлагается описывать предметную область в форме онтологии, а затем среда генерирует коды для всех вышеупомянутых слоёв и связи между ними.

Другим возможным решением является подход, когда слои разделены явно, и взаимодействуют с помощью сетевых технологий. Это позволяет описывать слои и даже части слоёв на разных языках программирования. А это, в свою очередь, позволяет составлять программы из разных языков и что самое главное – обновлять слои независимо друг от друга. При этом ключевое дополнительное требование – чтобы взаимодействие слоёв было удобным для программиста.

Этот подход реализуется, например, в платформе Логос и описан в работе [3].

Настоящая работа также посвящена исследованию этого подхода. Необходимо отметить, что каждая реализация подразумевает некоторую собственную программную модель (сущности, их взаимосвязи и взаимодействия). В работе представляется оригинальная модель, которая, по мнению автора, решает обозначенную проблему.

II. Теория: модель управления визуальными образами.

Сущности модели и их взаимодействие следующие. Компонента - это логический процесс в смысле Хоара [4]. Над компонентами вводится отношение "родитель-дети». Исполнитель – процесс операционной системы (ОС), работающий на некотором оборудовании. Каждая компонента выполняется в некотором заданном исполнителе.

Канал — сущность для передачи данных. Вводится операция записи сообщений в канал. Вводится операция подписки (реакции) на сообщения в канале. Каналы сопоставляются пространству имён, которое является общим (глобальным) для всех исполнителей. Каждой компоненте сопоставляется набор входных и выходных каналов. Компонента взаимодействует с внешним миром посредством этих каналов.

Связь между двумя каналами — специальный процесс, который производит передачу сообщений записываемых в первый канал во второй канал. Список операций в модели: Открытие канала: open_channel: channel_id → channel, где channel_id — имя канала в общем для всех исполнителей пространстве имён. Возвращает локальный объект для операций с каналом (локальный в смысле доступный в том исполнителе и в том языке программирования, в котором выполнена операция open_channel).Запись сообщения в канал: put: channel, msg → void, где channel — локальный объект для операций с каналом, msg — сообщение в некоторой форме. Технически в зависимости от реализации операция put может поддерживать передачу дополнительных уведомлений о ходе отправки сообщения, например посредством обещаний (promise).Создание реакции:react: channel, fn → void, где channel — локальный объект для взаимодействия с каналом, fn — функция, которая будет вызываться средой при (после) записи следующего значения в канал. В канал можно записать сообщение, находясь в одном исполнителе, а получать эти сообщения — в этом же или других исполнителях. При записи сообщения его получают все подписавшиеся стороны. Название реакция выбрано не случайно, оно соответствует термину "реакция" из модели Lingua Franca [5]. Есть и аналогичные термины, например "подписка на сообщения". Создание связи: create_link: source_id, target_id → local_link_objectгде source_id это имя канала-источника в общем для всех исполнителе пространстве имён, target_id — имя канала-приёмника, local_link_object — локальный объект связи для последующих операций с ней (удаления). После создания связи все записи в канал-источник приводят к копированию этих сообщений в канал-приёмник.

Создание дерева компонент в заданной точке исполнения: create_component: description, runner_id, target_id → root_id, где runner_id – идентификатор исполнителя, в котором следует создать компонент, target_id – идентификатор компонента, в список "детей" которого добавить создаваемый компонент, а description — описание создаваемого компонента. Это описание рекурсивно и определяет компонент и поддерево его вложенных компонент в рамках иерархии "родитель-дети". Структура описания description: type_id — идентификатор типа. Исполнитель должен уметь создавать компоненты этого типа .name — локальное имя объекта (необязательно) .params — список (словарь) значений параметров функции-конструктора компоненты, а также список значений, которые следует записать во входные каналы компонента после его создания .items — список описаний вложенных компонент (описания такой же структуры как description).Создаваемому компоненту назначается уникальный идентификатор root_id. Входные и выходные каналы, соответствующие компоненту, а также вложенные компоненты, получают идентификаторы вида "root_id/subitem_id". Код, который вызвал операцию create_component, получает в результате идентификатор root_id и может, используя его, обратиться к каналам компоненты и его вложенным компонентам. Компоненты, таким образом, представляются

для внешнего мира как набор входных и выходных каналов. Также извне можно обратиться к вложенным компонентами, которые в свою очередь также представляются как набор каналов, и т. д.

Дополнительно к перечисленным выше операциям в модели вводятся и обратные операции, которые в данном тексте пропущены для сокращения описания: закрыть канал, отменить подписку на сообщения, удалить связь, удалить дерево компонент. Порядок работы подразумевается следующий. Каким-либо образом запускаются процессы исполнители (на локальной машине или на наборе машин, связанных сетью). Часть этих исполнителей подразумеваются графические, способные создавать компоненты пользовательского интерфейса и (или) сцен визуализации. Другая часть — логические. Некоторый логический компонент (или просто функция) инициирует создание необходимых деревьев компонент в заданных исполнителях. Также он создаёт связи между каналами этих компонент. Дополнительно логический слой подписывается на сообщения в каналах графических компонент для получения сигналов от пользователя. Получая эти сигналы, логический слой формирует в итоге сообщения для входных каналов компонент пользовательского интерфейса, создаёт новые компоненты, и так далее.

Работа программы идёт сообразно алгоритму, описанному в логическом слое, а её визуальное представление реализуется в графическом слое. В результате становится возможным, например, написать исполнитель на языке QML в котором реализовать компоненты графического слоя (пользовательского интерфейса), а логический слой – в программе на языке Python. Можно затем заменить графический слой QML на слой на web-технологиях, при этом программа логического слоя останется неизменной. Также можно заменить и программу логического слоя, например, переписав её на другом языке. Можно заменить не весь логический слой, а его часть, и, например, вынести его в другой исполнитель / язык. Необходимо отметить, что связи между каналами это крайне существенный элемент, и он находится на том же уровне, что и компоненты. Программа в представленной модели является по сути направленным графом из процессов, ребра в котором — это связи между каналами процессов.Также необходимо отметить, что возможна высокоэффективная реализация связей. Например, пусть создаётся пара компонент на некотором исполнителе, и два их канала связываются. Эту связь можно технически реализовать таким образом, что запись в канал будет приводить к прямому вызову функций реакции на втором канале без сетевого взаимодействия. Это позволяет описывать наборы компонент и связей между ними внешним образом, а их выполнение при этом будет столь же эффективно как обычное взаимодействие кодов внутри процесса ОС.

III. Результаты экспериментов.

Была разработана прототипная среда, реализующая вышеуказанную модель и обеспечивающая сетевую коммуникацию для программ на языках: Javascript (в версиях для web и для nodejs), Python, C++, QML.

Было реализовано две программы: специализированная ГИС и программа визуализации параллельных вычислений. Рассмотрим результаты их разработки.

Специализированная ГИС была разработана в контексте того, что имелась задача и имелся большой технологический задел кодов на QML и Qt. Однако, по мнению автора, к этим технологиям следует выдвинуть ряд претензий (некорректность терминологии, устаревшая версия языка Javascript в QML, ряд неудачных дизайнерских решений, приводящих к трудновыявляемым ошибкам, например, возможность доступа к свойствам вне текущей области видимости выше по иерархии, другие неудобства), и поэтому нецелесообразно создавать новые продукты с их использованием. Было решено опробовать предлагаемую технологию. При этом для логического слоя был выбран язык Python по причине его распространённости. Ожидается, что к проекту на Python будет относительно легко подключать дополнительных разработчиков.

Графический слой был реализован на QML и Qt: объекты имеющегося технологического задела были обёрнуты в оболочки, и представлены как компоненты модели. Сущности и операции модели были реализованы в форме объектов C++ с мостами в QML посредством способов Qt (слоты, сигналы и прочее). Сетевая часть на C++ была реализована на базе библиотеки Mongoose [6] [7].

Примеры созданных компонент графического слоя (для реализации пользовательского интерфейса):

- Ряд (row) — компонент в роли контейнера, размещает вложенные компоненты горизонтально.

6 Этот метод реализации оказался очень громоздким. Его мотивация была получить "заодно" реализацию и для C++. Этого удалось достичь, но в совокупности с "мостами" в Qt/QML получился внушительный объем кода. Если делать проект заново, то автор предпочёл бы провести всю реализацию внутри QML, например на базе библиотеки QML Websocket.

7 Библиотека Mongoose оказалась крайне неэффективной. Например, в ней принимаемые и передаваемые данные копируются несколько раз в промежуточных буферах, прежде чем быть переданными программе или ОС. Поэтому для C++ рассматривается вариант переписать реализацию на библиотеке C++ Asio.

- Колонка (column) — компонент в роли контейнера, размещает вложенные компоненты вертикально.
- Окно (dialog) – компонент в роли контейнера, отображает вложенные компоненты в отдельном окне.
- Кнопка (button), текстовое поле (field), текстовая надпись (text), и т. п. – компоненты, реализующие элементы графического интерфейса пользователя.
- Набор картографических слоёв (map_group) — компонент для управления отображением картографической информации, взаимодействует с мышкой и клавиатурой пользователя и показывает содержимое вложенных компонент.
- Отображение растровых и векторных карт (show_image) — компонент картографического слоя, применяется как вложенный в map_group.
- Отображение векторных данных (show_vector) — компонент картографического слоя, применяется как вложенный в map_group.

Логический слой был реализован на языке Python 3. Слой определяет состав и группировку компонент графического слоя (см. выше), посылает данные в каналы этих компонент, реагирует на сообщения в их выходных каналах (например, движение мышки, клики по картографическим слоям, нажатие на кнопки, и т.п.)

Для работы представленной модели в среде Python была разработана специальная библиотека на базе модуля asyncio. Сущности и операции модели были реализованы в форме объектов. Пример использования кода Python показан на рис. 1.

Программа была успешно реализована и применяется на практике. Планируется разработка варианта графического слоя с применением веб-технологий, например, на базе библиотеки отображения карт OpenLayers.

```
############  создание таблицы из 10 полей и кнопки
### формирование описания description
lst1 = []
for x in range(0,10):
    d = gui.elem(type="row",params={"spacing":10},items=[
          gui.elem( type="text",params={"text":"Поле "+str(x)}),
          gui.elem( type="field",params={"value":""},"id"="field_"+str(x)),
        ])
    lst1.append(d)
lst1.append( gui.elem(type="button",params={"text":"ПОЕХАЛИ"},id="gobtn"))
d = gui.elem(type="column",params={"spacing":5},items=lst1)


### выполнение операции create_component на исполнителе gui
col2 = await gui.create_component( target_id="app",description=d )


############ реакция на нажатие кнопки
def go(msg):
    print("ЕДЕМ!")
col2.item("gobtn").open_channel("clicked").react( go )
```

Рис. 1. Пример кода на языке Python. Код формирует описание графических компонент, выполняет операцию по созданию компонент согласно этому описанию, и добавляет реакцию к выходному каналу одной из компонент.

Визуализация параллельных вычислений. Разрабатывается параллельная версия одного алгоритма обучения с подкреплением. На начальном этапе разработки оказалось, что версия имеет низкий коэффициент распараллеливания. Чтобы понять причины этого, было принято решение визуализировать её работу. Вид отображения был разработан следующий (см. пример на рис. 2):

- каждая задача (в смысле задач графа задач [6]) отображается отрезком вдоль оси OX. Положение отрезка на оси X соответствует времени начала и конца выполнения задачи, положение на оси Z соответствует номеру параллельного исполнителя;
- дополнительно изображаются передачи данных между процессами, поскольку передача может занимать значительное время, и это хотелось бы видеть. Параллельная версия использует схему мастер-рабочий, и коммуникация осуществляется только между мастером и рабочими. Поэтому чтобы не захламлять изображение, было решено показывать передачу вертикальными отрезками от плоскости рабочих Y=1 к плоскости Y=0, что означает пересылку данных на вход рабочему от мастера и в обратном направлении.

Графический слой реализован на веб-технологиях, на базе Javasript и DOM API. Трёхмерная составляющая реализована на базе библиотеки ThreeJS. Были разработаны компоненты:

- Контейнеры, элементы пользовательского интерфейса — реализованы такие же как в примере выше (для колонок и рядов использована технология CSS Flexbox).

- Рендеринг (render) — компонент, инкапсулирующий процесс рендеринга сцены.
- Сцена (scene) — компонент в роли контейнера, реализует древовидную структуру для компонент трёхмерной графики.
- Точки, отрезки, треугольники (points, lines, mesh) – компоненты трёхмерной графики, предназначены для отображения соответствующих графических примитивов.

Компоненты графических примитивов points, lines, mesh для управления составом векторных данных предоставляют на вход каналы двух видов: 1) по установке данных целиком, и 2) по модификации данных, например, по добавлению части данных. Это позволяет менять состав векторных данных, не пересылая их целиком каждый раз по сети.

Логический слой был реализован на языке Python 3. Это было обусловлено тем, что обучение (однопоточная и параллельная версии) были разработаны именно на этом языке. В уже имеющуюся программу были добавлены операции представленной модели для управления трёхмерной сценой.

А именно, реализация следующая. При начале вычисления задачи исполнитель замеряет текущее время. По завершению он вычисляет время, затраченное на задачу. Затем высылается сообщение в канал «решённые задачи». Специальный модуль логического слоя («рисователь») обрабатывает эти сообщения и формирует координаты вершин для сцены. Затем он пересылает эти координаты посредством сообщений в каналы «модификация данных» (см. выше) компонентам графического слоя, ответственным за рендеринг.

Программа запускается, запускает сетевую среду выполнения, переходит к вычислениям и показывает веб-ссылку для запуска визуализации. Пользователь может в любой момент открыть по этой ссылке визуализацию в браузере, программа по ходу работы обновляет графическое представление. Таким образом, была получена визуализация работы параллельного вычисления в режиме онлайн [7], т. е. по ходу вычисления.

Пример полученной визуализации показан на рис. 2. Эта визуализация была использована для понимания работы и последующей оптимизации параллельного алгоритма.

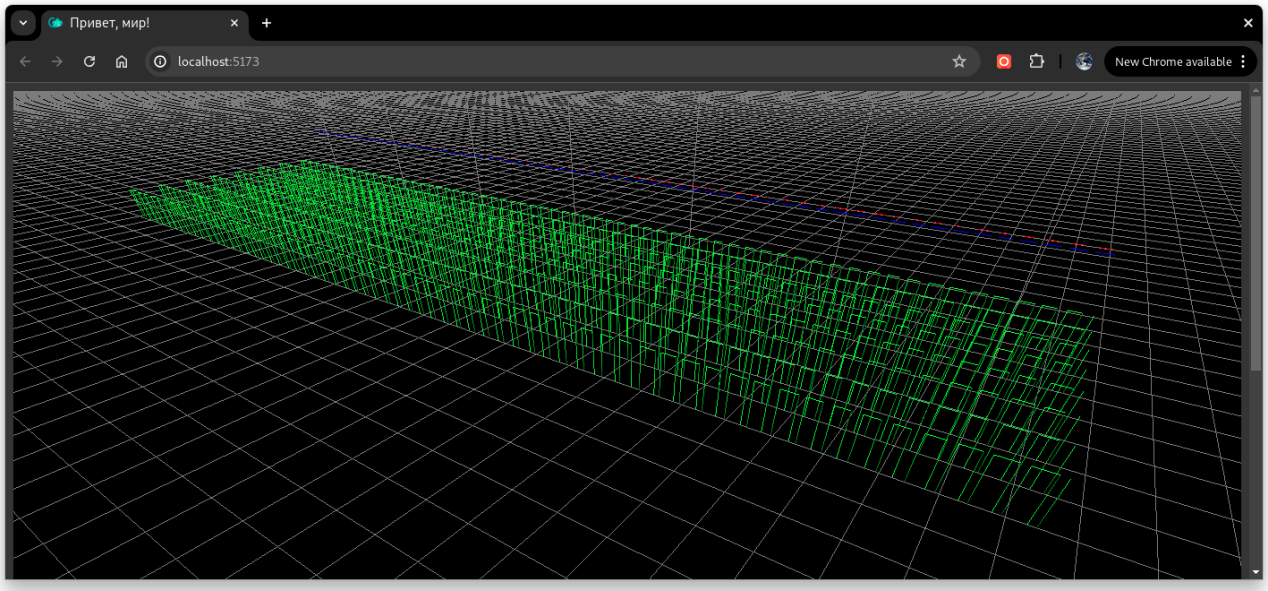


Рис. 2. Пример визуализации, полученной с помощью представленной технологии. Показана работа 8 рабочих процессов. Время идёт слева направо. Рабочие расположены каждый на своей оси Z (ближе-дальше к экрану). Горизонтальная линия на оси рабочего — выполнение задачи. Вертикальная — получение рабочим входных данных от мастера и высылка ему результата. На дальнем плане синие отрезки — ожидание мастером всех рабочих, красные — одновременное выполнение мастером оценки текущей ситуации.

IV. Обсуждение результатов.

Отличие предложенной модели от модели X Window System следующее. В последней компоненты графического слоя выполняются в основном процессе программы, и они отправляют графические примитивы в программы, работающие на графическом оборудовании (на «сервер» в терминологии X Window). В обратном направлении высылаются команды пользователя — сигналы от мыши и клавиатуры. В предложенной модели посылаются не графические примитивы, а высокоуровневые сообщения для компонент графического слоя,

которые работают в процессах на графическом оборудовании. В обратном направлении также движутся высокоуровневые сообщения («нажата кнопка», «изменено значение бегунка» и т. п.).

Среди отличий предложенной модели от модели Логос, представленной в работе [3], в первую очередь необходимо выделить следующие. В работе [3]:

– вместо возможности прямого доступа к каналам компонент, в описание компонент (см. поле description выше) вносятся связи локальных каналов компонент с глобальными каналами. Такой подход позволяет, в том числе, программировать динамику взаимодействия графических компонент сразу на уровне описания посредством передачи сообщений между компонентами опосредовано через глобальные каналы. Направления связей можно указывать в обе стороны — от локальных к глобальным и наоборот.

– Реакции можно создавать только в привязке к глобальным каналам.

– Поддерживаются «теги», которые позволяют включать каждую компоненту в несколько множеств, и поддерживаются групповые операции над элементами таких множеств. Например, команда «установить значение поля «value» для всех компонент, входящих в заданное множество.

– На уровне реализации предоставлена возможность загружать описание компонент в форме текста на языке JSON. Реализация логического слоя сделана для языка Python. Поддерживаются версии графического слоя для веб и настольных платформ, которые реализуют большое количество компонент.

Работа [3] активно применяется на практике, но, к сожалению, она недоступна для сторонних разработчиков.

Отличие предложенной модели от модели Yandex Divkit следующее. В Divkit предложена богатая модель встроенной динамики компонент графического слоя. Это сделано потому, что подразумевается большая автономность этого слоя от логического слоя, подход «загрузи и работай». В модели настоящей работы подразумевается более тесное, постоянное взаимодействие компонент графического и логического слоёв.

V. Выводы и заключение

В работе предложено вынести интерфейс взаимодействия графического и логического слоёв программ на сетевой уровень. Это позволяет:

1. Не зависеть от развития и устаревания технологий. При устаревании технологии, на которой написан слой, его можно переписать. В первую очередь это касается графического слоя. Но и логический слой можно переписывать — постепенно. Предложенная технология такова, что обеспечивает возможность работы частей слоя одновременно в разных процессах ОС и на разных языках программирования.
2. Удобно переключать реализации графического слоя, если это ценно для задачи. Например, менять "веб" на "настольную" версию, или варианты с OpenGL на Vulkan.
3. Применять наиболее удобные и подходящие в контексте задачи языки для реализации слоёв. Например, графический слой можно реализовать на веб-технологиях, а логический на Python.
4. Запускать графические программы в многомашинной среде "естественным" образом, благодаря сетевому протоколу взаимодействия. Например, графический слой может выполняться на исполнителях на машинах с мониторами, а логический на исполнителях на других машинах, в том числе на узлах суперкомпьютера.

Технология протестирована на двух прикладных задачах, и показала свою работоспособность. Вместе с тем её можно развивать, некоторые варианты развития показаны в разделе "Обсуждение результатов".

Прототипная реализация технологии выполнена на базе системы параллельного программирования PPK (https://github.com/pavelvasev/ppk).

Автор благодарит многоуважаемых коллег, сотрудников Академии наук, Корпорации Росатом, Уральского федерального университета и других организаций – за поддержку, плодотворное обсуждение и синтез идей.

Список литературы

1. В.Л. Авербух, Развитие человеко-компьютерного взаимодействия // Научная визуализация 12.5: 130 - 164, DOI: https://doi.org/10.26583/sv.12.5.11

2. Ryabinin K., Chuprina S. Ontology-Driven Edge Computing // Lecture Notes in Computer Science. – Springer, 2020. – Vol. 12143. – P. 312–325. DOI: https://doi.org/10.1007/978-3-030-50436-6_23.

3. Губайдулина Е. А. и др., Инструмент разработки специализированных пользовательских графических интерфейсов для проведения комплексного математического моделирования средствами пакета программ ЛОГОС // Тезисы XIX Международной конференции «Супервычисления и математическое моделирование», 20 - 24 мая 2024г., г. Саров, стр. 68. URL: www.cv.imm.uran.ru/e/3241932

4. C.A.R. Hoare, Communicating Sequential Processes , December 4, 2022.

5. Marten Lohstroh, Christian Menard, Soroush Bateni, and Edward A. Lee. 2021. Toward a Lingua Franca for Deterministic Concurrent Systems. ACM Trans. Embed. Comput. Syst. 20, 4, Article 36 (July 2021), 27 pages. DOI: doi.org/10.1145/3448128

6. Васёв П.А., Визуализация работы алгоритма планирования параллельных задач // Труды 33-ей Международной конференция по компьютерной графике и машинному зрению ГрафиКон 2023, 19-21 сентября 2023 г., город Москва. С. 341-353. DOI: https://doi.org/10.20948/graphicon-2023-341-353

7. Pavel Vasev, Analyzing an Ideas Used in Modern HPC Computation Steering // 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), Yekaterinburg, Russia, 2020, pp. 1-4, doi: https://doi.org/10.1109/USBEREIT48449.2020.9117685

УДК 519.6
DOI: 10.25206/978-5-8149-3873-2-2024-361-367

Визуализация влияния вихревых структур на сверхзвуковое обтекание крыла под углом атаки
Visualization of the vortex structures effect on the supersonic flow around wing at an attack angle

Т.В. Константиновская [1], В.Е. Борисов [1], А.Е. Луцкий [1]
[1] *ИПМ им. М.В. Келдыша РАН, Миусская пл., д. 4, Москва, 125047, Россия*

T.V. Konstantinovskaya [1], V.E. Borisov [1], A.E. Lutsky [1]
[1] *Keldysh Institute of Applied Mathematics, Miusskaya sq. 4, Moscow, 125047, Russia*

Аннотация. В работе рассматривается влияние вихревых структур на обтекание крыла под углом атаки в сверхзвуковом потоке. Представлены результаты анализа и обработки численных данных при помощи методов научной идентификации и визуализации вихревых течений. Крыло-генератор и основное крыло расположены под углом атаки 20º градусов к набегающему потоку. Рассмотрено две конфигурации, отличающиеся длиной вихрегенератора. Моделирование проводилось с помощью URANS уравнений. Численные расчеты проведены на гибридной суперкомпьютерной системе К-60 в ИПМ им. М.В. Келдыша РАН. Научная визуализация результатов расчетов проводилась с использованием преимущественно Лютекс (Liutex) критерия идентификации вихревых структур, относящемуся к третьему поколению таких методов.

Ключевые слова: Визуализация сверхзвукового течения, концевой вихрь, идентификация вихревых структур, взаимодействие вихря с поверхностью

Abstract. The effect of vortex structures on the flow around the wing at an attack angle in a supersonic flow is considered in this paper. The results of the analysis and processing of numerical data using scientific identification and visualization methods for vortex flows are presented. The generator wing and the main wing are located at an attack angle of 20° degrees to the incoming flow. Two configurations differing in the length of the vortex generator are considered. The simulation was carried out using URANS equations. Numerical simulations were performed on the K-60 hybrid supercomputer system at the Keldysh Institute of Applied Mathematics of RAS. The scientific visualization of the numerical results was carried out using mainly the Liutex criterion of the identification of vortex structures, which belongs to the third generation of such methods.

Keywords: Supersonic flow visualization, tip vortex, vortex structures identification, vortex surface interaction

Введение

В современном мире скорости передвижения имеют большое значение и оказывают решающую роль при выборе средства передвижения. В связи с чем вновь возрастает внимание к развитию высокоскоростных летательных аппаратов (ЛА) в различных областях применения.

При разработке и конструировании высокоскоростных ЛА важную роль играет подробное изучение аэродинамики их полета.

Одной из важных задач является изучения влияния вихревых структур на аэродинамику ЛА, т.к. образование вихревых структур сопровождает полет любого ЛА. Изучение взаимодействия вихревых структур с поверхностью ЛА представляет отдельный интерес, в частности, такой структуры, как концевой вихрь. Концевой вихрь образуется на крыльях или на других оребренных частях тел в результате разницы давлений между верхней и нижней поверхностями.

Специальные методы научной идентификации и визуализации вихревых структур могут эффективно применяться для обработки и анализа результатов численного и экспериментального моделирования, предоставляя инструменты не только для визуализации течений, но и для их анализа [1-3]. В частности, ими могут быть выделены положение основных вихревых структур.

В данной работе представлены результаты визуализации численного исследования влияния вихревых структур на крыло при сверхзвуковом режиме. Показано влияние концевого вихря и вихревой пелены с

переднего вихрегенератора на обтекание основного крыла. Рассматривается две компоновки, отличающиеся размахом генератора. Для визуализации используется главным образом Лютекс (Liutex) критерий идентификации вихревых течений [4 – 9]. Для расчетов используется подход URANS с моделью турбулентности SA. Расчеты проводились на многопроцессорной гибридной системе К-60 в ИПМ им. М.В. Келдыша РАН.

Постановка задачи

Исследовалось сверхзвуковое обтекание двух крыльев, расположенных последовательно друг за другом по направлению течения с числом Маха набегающего потока $M_\infty = 3$. Крылья использовались прямоугольные в плане с острыми передней, боковой и задней кромками. Полуразмах основного крыла был $l_1 = 0.095$ м, переднего вихрегенератора в первой компоновке – $l_2 = 0.0475$ м (короткий генератор), во второй – $l_2 = 0.095$ м (длинный генератор), хорда обоих крыльев $b = 0.03$ м. Расстояние между осями крыльев составляло 4 хорды по направлению потока. Ось основного крыла располагалась напротив задней кромки переднего крыла (генератора) по потоку. Угол атаки крыльев составлял 20°, число Маха набегающего потока было $M_\infty = 3$, число Рейнольдса задавалось $\mathrm{Re}_L = 1 \times 10^7$ (L – характерная длина обезразмеривания вычислительной модели, здесь $L = 1$ м). На рис. 1 представлена схема расчетной области.

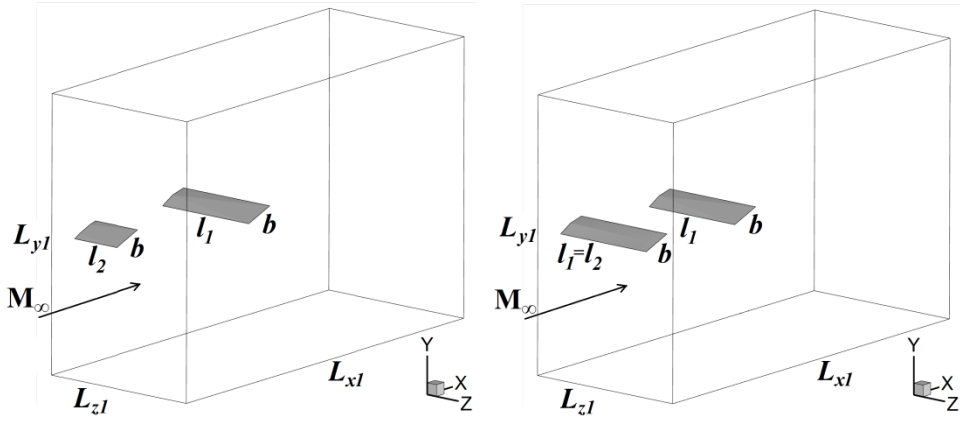


Рисунок 1: Расчетная область, установка крыльев

Течение рассматривалось на расстоянии вплоть до 6.5 хорд крыла вниз по потоку от задней кромки основного крыла.

Здесь и далее в тексте статьи под компоновкой 1 понимается вариант обтекания при коротком генераторе, под компоновкой 2 – при длинном генераторе.

Численный алгоритм

Расчеты трехмерного турбулентного течения сжимаемого газа проводились с использованием системы нестационарных осредненных по Рейнольдсу и Фавру уравнений Навье–Стокса (URANS). Однопараметрическая модель турбулентности Спаларта-Аллмараса (SA) в модификации для сжимаемых течений [9] использовалась для определения значения турбулентной вязкости. Начальные и граничные условия ставились стандартным образом.

Аппроксимация уравнений модели проводилась по пространству с помощью метода конечных объемов со схемой реконструкции TVD 2-го порядка точности. Метод конечных объемов в предположении, что расчетная область покрыта сеткой, состоящей из неперекрывающихся многогранных ячеек, реализуется интегрированием системы уравнений модели по каждой счетной ячейке с последующим преобразованием объемных интегралов от потоков в поверхностные интегралы по граням ячейки. Для вычисления невязких потоков на гранях расчетных ячеек использовался обобщенный метод С. К. Годунова с точным римановским солвером. Для аппроксимации уравнений по времени применялись как явная, так и неявная (на основе метода LU-SGS) схемы. Используемый численный метод описан в [11].

Для проведения численных расчетов использовался авторский программный комплекс ARES [12]. Расчеты проводились на гибридной суперкомпьютерной системе К-60 [13]. Для данной задачи использовались неструктурированные сетки, содержащие порядка 5 млн гексагональных ячеек для каждой из постановок.

Визуализация вихревых структур течения

Внутри авторского программного комплекса ARES был разработан специальный модуль пост-процессинговой обработки, позволяющий определять и анализировать вихревые структуры на гексагональных сетках. В его рамках реализованы некоторые классические методы научной визуализации, такие как λ_2, Q-критерий и др [14, 15]. Также, модуль содержит Лютекс (Liutex) метод научной визуализации – один из наиболее современных критериев идентификации вихревых структур, относящийся к третьему поколению таких методов. Пост-процессинговый модуль формирует выходные данные в формате пакета программ Tecplot.

В данной работе приводятся результаты и их визуализация, полученные с помощью методов идентификации и визуализации вихревых структур, главным образом, при помощи Лютекс критерия. Как показала практика его применения авторами, для ряда задач он позволяет получать значительно более лучшие результаты [16] в сравнении с классическими методами. Приведем далее его описание. Описание других методов идентификации вихревых структур можно найти в литературе (например [14, 15]).

Лютекс (Liutex) метод визуализации

Лютекс (Liutex) метод (или критерий) идентификации и визуализации вихревых структур является одним из наиболее современных, он принадлежит к т.н. третьему поколению таких критериев. Этот критерий свободен по своему построению от сдвиговых и сжимающих компонент тензора скоростей деформации [4], что уменьшает вероятность ложного определения вихревых структур этим критерием. Лютекс критерий позволяет оценивать не только направление, но и силу вихря.

Критерий был предложен в 2018 г. под названием Ортекс (Rortex) критерий [5], впоследствии он был переименован в Лютекс (Liutex) по имени одного из авторов [6].

Согласно этому критерию областью течения с вихревыми структурами считается та, в которой тензор скоростей деформации имеет одно действительное λ_r и два комплексно-сопряженных $\lambda_{cr} \pm \lambda_{ci}$ собственных значения. С использованием этих собственных значений определяется Ортекс-вектор [7], локально совпадающий с осью вращения вихря как твердого тела:

$$\mathbf{R} = \mathcal{R}\,\mathbf{r}, \quad \mathcal{R} = \boldsymbol{\omega}\cdot\mathbf{r} - \sqrt{(\boldsymbol{\omega}\cdot\mathbf{r})^2 - 4\lambda_{ci}^2},$$

где $\boldsymbol{\omega}$ – вектор завихренности, $\mathbf{r}$ – нормированный собственный вектор, отвечающий λ_r с условием $\boldsymbol{\omega}\cdot\mathbf{r} > 0$. На его основе формируется нормированная величина $\Omega_R \in [0,1]$, показывающая интенсивность локального вращения среды:

$$\Omega_R = \frac{\left(\boldsymbol{\omega}\cdot\mathbf{r}\right)^2}{2\left[\left(\boldsymbol{\omega}\cdot\mathbf{r}\right)^2 - 2\lambda_{ci}^2 + 2\lambda_{cr}^2 + \lambda_r^2\right] + \varepsilon_{Lu}}, \quad \varepsilon_{Lu} = k\max\{\lambda_{ci}^2\}.$$

Здесь ε_{Lu} – величина, предназначенная для фильтрации численного «шума», максимум λ_{ci} берется по всей рассматриваемой области, $k = 0.001 \sim 0.002$ [7].

Результаты расчетов и их визуализация

На рис. 2 представлен общий вид полученного течения для обеих рассмотренных конфигураций, а) – для короткого генератора, б) – для длинного.

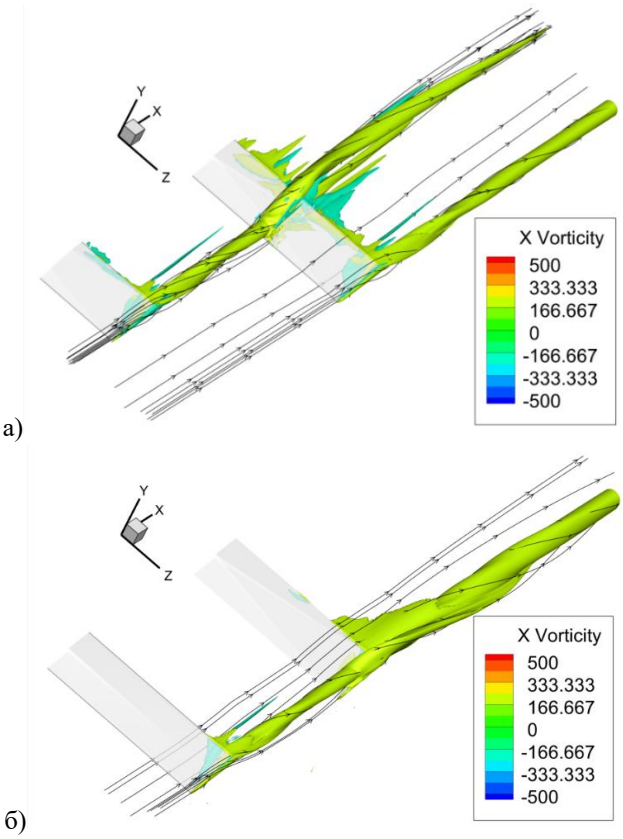


Рисунок 2: Общий вид полученных течений в рассмотренных конфигурациях

Для визуализации результатов течения на рис. 2 используются линии тока и изоповерхности продольной завихренности *XVort* (значение 170 и -170). Использование продольной завихренности позволяет показать направление вращения выявленной вихревой структуры (показано светлым и темным тонами).

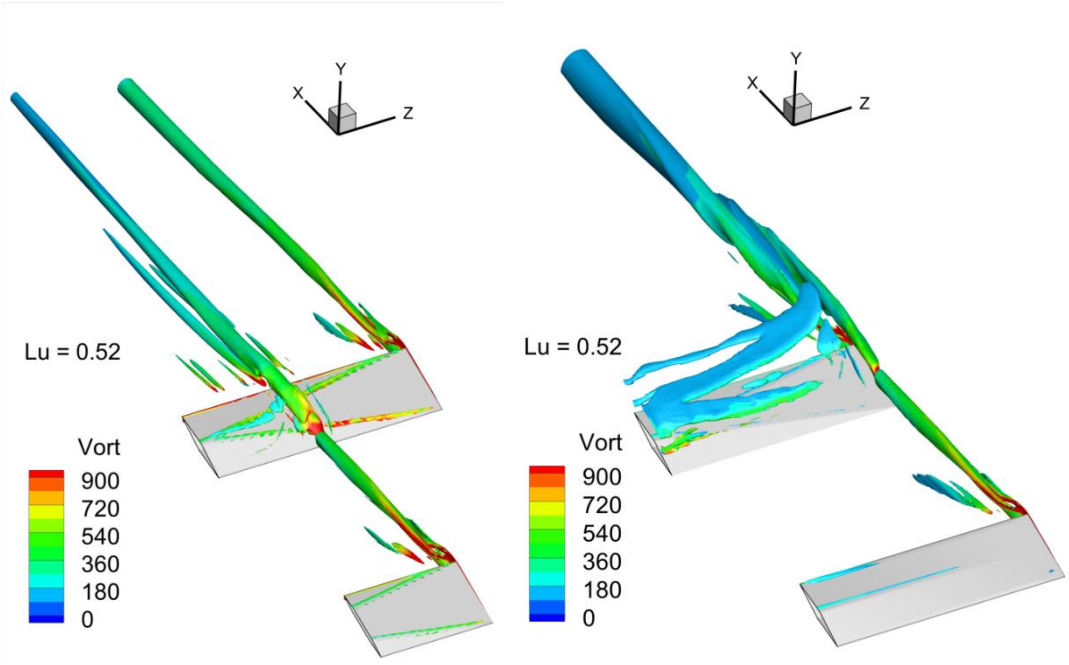


Рисунок 3: Визуализация вихревых структур при помощи Liutex критерия: изоповерхности $\widetilde{\Omega}_R = 0.52$, для конфигурации с коротким генератором (а) и с длинным (б)

При обтекании крыла потоком образуются вихревые структуры, основные из которых это концевой вихрь, а также вихревая пелена.

На рис. 3 показана визуализация течения с помощью Лютекс критерия идентификации вихревых структур для рассмотренных конфигураций с коротким (а) и длинным (б) генератором. Для отображения изолиний использовано стандартное значение параметра $\widetilde{\Omega}_R = 0.52$, достаточное для определения основных вихревых структур для подобных течений. Как видно, при данном значении $\widetilde{\Omega}_R$ хорошо различимы основные концевые вихри, а также вторичные вихри.

Однако на рис. 3 не отображено вихревой пелены за крыльями, которая тоже является важной составляющей, влияющей на обтекание основного крыла. Чтобы отобразить вихревую пелену, необходимо уменьшить значение параметра $\widetilde{\Omega}_R$. На рис. 4 показана визуализация течения с помощью Лютекс критерия при значении параметра $\widetilde{\Omega}_R = 0.15$. Вихревая пелена видна хорошо, однако остальные вихревые структуры течения (концевые и вторичные вихри) становятся практически неразличимыми, что особенно заметно для конфигурации с длинным генератором.

Таким образом, необходимо адаптировать используемые значения параметров критериев визуализации для рассматриваемых конфигураций. На рис. 5 показана визуализация при значении параметра Лютекс критерия $\widetilde{\Omega}_R = 0.4$ для рассмотренных конфигураций с коротким (а) и длинным (б) генератором. На этом рисунке хорошо различимы все вихревые структуры за крылом: и концевые вихри и вихревая пелена. Обе эти структуры взаимодействуют с основным крылом, влияя на его аэродинамические характеристики.

На рис. 3 – 5 при визуализации изолиний Лютекс критерия добавлены значения завихренности (*Vort*), что дает более наглядное представление об интенсивности вихря, которая больше непосредственно за крылом, и уменьшается по мере удаления от крыла вниз по потоку. Таким образом, при комбинации различных способов идентификации и визуализации вихревых структур можно получить дополняющие друг друга результаты.

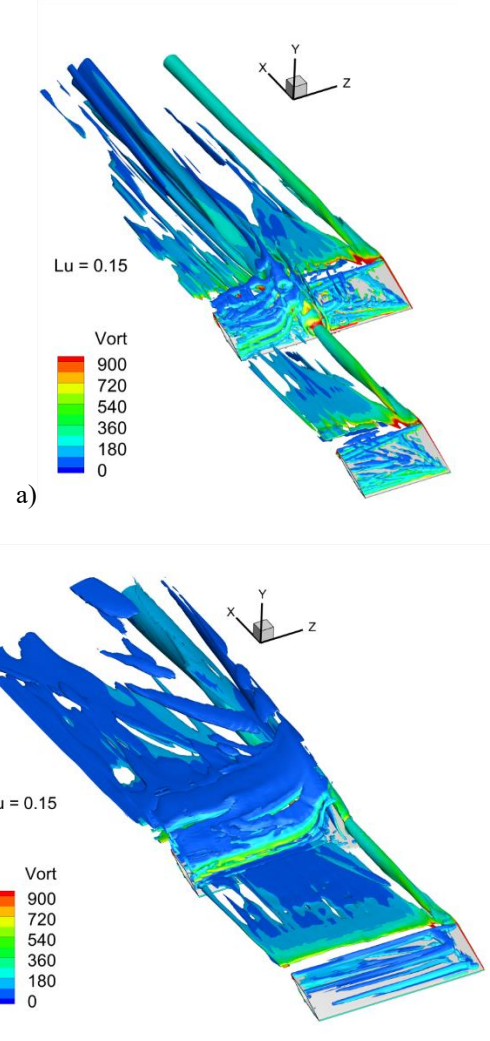


Рисунок 4: Визуализация вихревых структур при помощи Liutex критерия: изоповерхности $\widetilde{\Omega}_R = 0.15$, для конфигурации с коротким генератором (а) и с длинным (б)

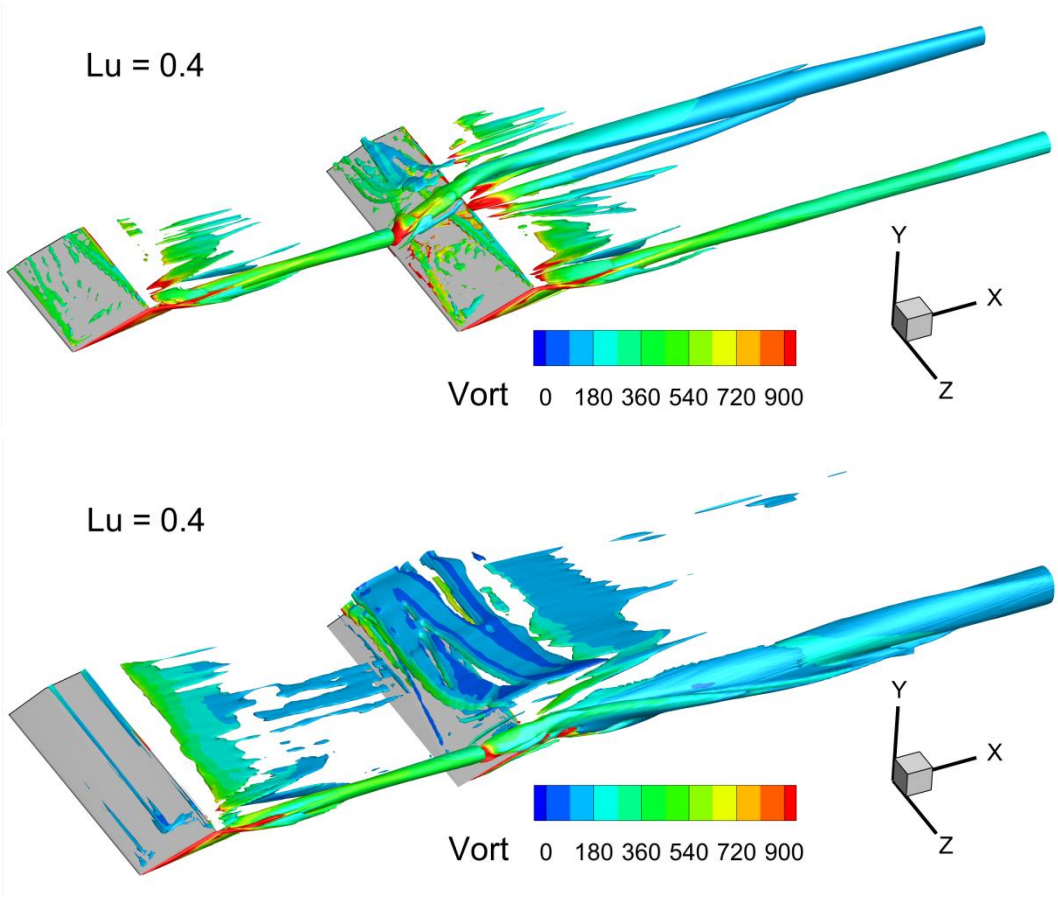


Рисунок 5: Визуализация вихревых структур при помощи Liutex критерия: изоповерхности $\widetilde{\Omega}_R = 0.4$, для конфигурации с коротким генератором (а) и с длинным (б)

Концевой вихрь с короткого генератора в рассмотренной конфигурации попадает на наветренную сторону основного крыла, где взаимодействует с его поверхностью (рис. 3 а – 5 а). При этом взаимодействии происходит образование вторичных вихрей, часть из которых имеет направление вращения противоположное направлению вращения концевого вихря. Факт образования вторичных вихрей качественно согласуется с экспериментальными и численными данными для несжимаемых течений [17, 18].

Концевой вихрь с длинного генератора в соответствующей конфигурации попадает не на поверхность основного крыла, а в зону его концевой хорды (рис. 3 б – 5 б). Происходит своеобразное накручивание одного концевого вихря на другой с последующим их объединением в одну продольную вихревую структуру с большей шириной.

Заключение

В работе представлены результаты анализа и научной визуализации задачи влияния вихревых структур на сверхзвуковое обтекание крыла, расположенного под углом атаки. Рассмотрена зависимость от размаха крыла генератора. Численные расчеты были выполнены с помощью авторского программного комплекса ARES на суперкомпьютере К-60.

Лютекс-критерий (Liutex) применялся главным образом для идентификации и визуализации вихревых структур.

Для конфигурации с коротким генератором было получено, что концевой вихрь попадает на наветренную поверхность крыла, при взаимодействии с которой образуется несколько вторичных вихрей, часть из которых имеет направление вращения противоположное направлению вращения концевого вихря. Форма этого взаимодействия качественно согласуется с экспериментальными и численными данными для несжимаемых течений.

Для конфигурации с длинным генератором получено «наматывание» концевого вихря с крыла генератора на концевой вихрь с основного крыла с последующим их слиянием в одну вихревую структуру.

Показано, что комбинация различных способов идентификации и визуализации вихревых структур позволяет получить дополняющие друг друга результаты.

Благодарности

Работа выполнена при финансовой поддержке РНФ, проект № 24-21-00230.

Список литературы

Yang Wen-Jei. Handbook of flow visualization. CRC Press; 2nd edition. 2001. 724 p.

Волков К.Н. Методы визуализации вихревых течений в вычислительной газовой динамике и их применение при решении прикладных задач // Научно-технический вестник информационных технологий, механики и оптики. 2014. 3 (91)

Epps B.P. Review of Vortex Identification Methods // 55th AIAA Aerospace Sciences Meeting. 2017. AIAA 2017-0989. DOI: 10.2514/6.2017-0989

Liu C., Gao Y., Dong X., Wang Y., Liu J., Zhang Y., Cai X., Gui N. / Third generation of vortex identification methods: Omega and Liutex/Rortex based systems // J. Hydrodyn. 2019. **31** 2, pp. 205–223.

Liu C., Gao Y., Tian S., Dong X. Rortex—A new vortex vector definition and vorticity tensor and vector decompositions // Phys. Fluids. 2018. 30:035103.

Shrestha P., Nottage C., Yu Y., Alvarez O., Liu C. / Stretching and shearing contamination analysis for Liutex and other vortex identification methods // Advances in Aerodynamics. 2021. **3** 8.

Liu J., Liu C. Modified normalized Rortex/vortex identification method // Phys. Fluids. 2019. 31:061704, 6 p.

Kolář V., Šístek J. Stretching response of Rortex and other vortex-identification schemes // AIP Advances. 2019. **9**, 105025. DOI: 10.1063/1.5127178

Gao Y., Yu Y., Liu J., Liu Ch. Explicit expressions for Rortex tensor and velocity gradient tensor decomposition // Phys. Fluids. 2019. 31:081704. DOI: 10.1063/1.5118948

Allmaras S.R., Johnson F.T., Spalart P.R.. Modifications and Clarifications for the Implementation of the Spalart-Allmaras Turbulence Model // Seventh International Conference on CFD (ICCFD7), Big Island, Hawaii, 9-13 July 2012.

В.Е. Борисов, А.А. Давыдов, И.Ю. Кудряшов, А.Е. Луцкий, И.С. Меньшов. Параллельная реализация неявной схемы на основе метода LU-SGS для моделирования трехмерных турбулентных течений // Математическое моделирование, 2014, 26 (10), с. 64–78;
V.E. Borisov, A.A. Davydov, I.Yu. Kudryashov, A.E. Lutsky, I.S. Men'shov. Parallel Implementation of an Implicit Scheme Based on the LU-SGS Method for 3D Turbulent Flows // Mathematical Models and Computer Simulations, 2015, 7 (3), p. 222–232.

Программный комплекс ARES для расчета трехмерных турбулентных течений вязкого сжимаемого газа на высокопроизводительных вычислительных системах. В.Е. Борисов, А.А. Давыдов, И.Ю. Кудряшов, А.Е. Луцкий. Свидетельство о регистрации программы для ЭВМ RU 2019667338, 23.12.2019.

Вычислительный комплекс К-60 // https://www.kiam.ru/MVS/resourses/k60.html .

Jeong J., Hussain F. On the identification of a vortex // Journal of Fluid Mechanics. 1995. Volume 285, pp. 69–94.

Hunt J.C.R., Wray A.A., Moin P. Eddies, stream, and convergence zones in turbulent flows // Technical Report № CTR-S88. Palo Alto: Center for Turbulent Research. 1988. pp 193–208.

Konstantinovskaya T.V., Borisov V.E. , Lutsky A.E. Visualization Problems of a Supersonic Tip Vortex in a Heat Wake // Scientific Visualization. 2022. **14** 4, pp. 71 – 82. DOI: 10.26583/sv.14.4.07

Chen C., Wang Z., Gursul I. Experiments on tip vortices interacting with downstream wings // Experiments in Fluids. 2018. **59** 5. Art. 82. 24 p. DOI: 10.1007/s00348-018-2539-7

Гаджиев Д.А., Гайфуллин А.М. Эволюция двух вихрей вблизи твердой поверхности // Прикладная механика и техническая физика. 2018. **59** 2, с. 31-38.

УДК 001.89
DOI: 10.25206/978-5-8149-3873-2-2024-368-372

Опыт проведения НИР «Научная визуализация и визуальная аналитика» в РТУ-МИРЭА в 2024 году
Experience of the R&D “Scientific visualization and visual analytics” at RTU-MIREA in 2024

А.Е.Бондарев[1]
[1]ИПМ им М.В. Келдыша РАН, Миусская пл., 4, 125047, Москва, Россия
[2]Российский технологический университет – МИРЭА, 119454, Москва, проспект Вернадского, 78, Москва, Россия

A.E. Bondarev[1]
[1] Keldysh Institute of Applied mathematics of RAS, Miusskaya sq. 4, Moscow, 125047, Russia

Аннотация. Данная работа представляет опыт построения и проведения НИР «Научная визуализация и визуальная аналитика» в РТУ-МИРЭА в 2024 году. НИР проводилась для студентов 3-го курса кафедры высшей математики Института искусственного интеллекта РТУ-МИРЭА. Описано построение и организация НИР. Приведены примеры заданий и их выполнения. Данная работа может быть интересной для преподавания аналогичных дисциплин в данной предметной области.

Ключевые слова: Опыт проведения НИР, научная визуализация, визуальная аналитика

Abstract. This paper presents the experience of constructing and conducting the R&D “Scientific visualization and visual analytics” at RTU-MIREA in 2024. The research was conducted for 3rd year students of the Department of Higher Mathematics of the Institute of Artificial Intelligence of RTU-MIREA. The construction and organization of the R&D is described. Examples of tasks and their fulfillment are given. This work may be interesting for teaching similar disciplines in this subject area.

Keywords: Experience in R&D, scientific visualization, visual analytics

1. Введение

Данная работа представляет опыт построения и проведения НИР «Научная визуализация и визуальная аналитика» в РТУ-МИРЭА в 2024 году. В 2024 году мне предложили провести НИР по теме Научная визуализация и визуальная аналитика» для студентов 3-го курса кафедры высшей математики института искусственного интеллекта в РТУ-МИРЭА. В прошлом (более 20 лет назад) подобный авторский курс был мною подготовлен и прочтен на кафедре автоматизации научных исследований ВМК МГУ. Естественно, здесь для меня было интересным все – и современный уровень студентов, и современный уровень систем визуализации, поэтому я безусловно согласился. Курс НИР был построен и проведен. Основные целями и задачами НИР было дать студентам основы обработки и визуального представления расчетных данных и выработать практические навыки пользования системами визуализации. Основным лейтмотивом НИР было следующее: «Решать численно задачи математического моделирования вас научат. Вы будете решать конкретные задачи. А как вы покажете другим и себе, что вы вообще что-то делали?». Подобные установки нашли понимание у студентов. При создании курса НИР был соблазн использовать материалы предыдущего курса 20-летней давности, но от этого пришлось отказаться, так как все программы считывания и обработки данных того времени были реализованы на Fortran, которого современные студенты вообще не изучают и не знают. Поэтому курс был построен следующим образом: сначала читались 3 ознакомительные лекции по основам научной визуализации и визуальной аналитики, затем студентам раздавались индивидуальные задания по визуализации расчетных данных, которые студенты должны были выполнить с попутным освоением систем визуализации. В качестве расчетных данных использовались реальные открытые расчетные данные, полученные мною и коллегами в ИПМ им. М.В. Келдыша РАН.

2. Выбор программы визуализации для освоения

При создании НИР естественно встал вопрос выбора программы визуализации для освоения студентами. Наиболее мощными и популярными программными комплексами, как и 20 лет назад, являются программные комплексы – TecPlot [1], Paraview [2], VisIt[3], Avizo[5]. Ранее также существовало большое количество открытых программ с ограниченной функциональностью (преимущественно, французских), позволяющих, тем не менее, успешно визуализировать 1D-графики, 2D распределения величин и 3D распределения величин в объемах. Мною и сотрудниками ИПМ им. М.В. Келдыша РАН был проведен предварительный анализ современной ситуации с ПО в этой области. Было выявлено следующее. Современная ситуация радикально изменилась. Во-первых, программ с частичной функциональностью в Интернете нет. То есть, тезис «Интернет помнит все» в данном случае неверен. Во-вторых, «большие» программы, перечисленные выше также изменились. Программы Paraview [2] и VisIt[3] существенно изменились в сторону усложнения обработки исходных данных, и их явно не имело смысла давать для обучения студентам. Мы с коллегами потратили несколько вечеров, решая достаточно простые примеры данных для визуализации. В результате в качестве программных комплексов для освоения были выбраны широко распространенный в ИПМ им. М.В. Келдыша РАН комплексе TecPlot [1] (имеющий trial-версию) и gnuplot [4], о котором имелась предварительная информация, что большинство студентов имеет опыт работы с ним.

3. Построение задания

НИР проводилась для 4-х групп студентов, численность каждой составляла в среднем 25 человек. Всего численность студентов составляла около 100 человек. Для этих групп было подготовлено 30 заданий. 26 заданий были ориентированы на использование системы TecPlot и 4 задания были ориентированы на использование системы gnuplot. В качестве исходного набора данных для каждого задания использовались ранее рассчитанные и опубликованные в открытой печати результаты решения аэрогазодинамических задач, полученные в отделе компьютерной графики и технического зрения ИПМ им. М.В. Келдыша РАН. Каждому студенту предлагалось установить программу визуализации, изучить данную программу, визуализировать результаты из исходных данных и написать краткий отчет об основных функциональных свойствах используемой программы и полученных результатах. Примеры заданий и полученных результатов можно увидеть в разделе 5.

4. Трудности и курьезы

Следует сказать, что подавляющее большинство студентов успешно справились с заданием, но при этом выполнение задания потребовало от них определенного упорства и энтузиазма. При выполнении задания они постоянно получали консультации от меня в режиме online. В среднем, выполнение задания занимало 3 итерации. Большинство ошибок были связаны с недостаточно тщательным изучением руководств к программам (manuals).

Так, например, трудности вызвало достаточно элементарное построение кросс-сечений. Если использовать инструмент построения сечений в TecPlot, то при смене координатного направления сечения само сечение не сохраняется. Пример консультационной подсказки: «…Как сделать перпендикулярные сечения на одном рисунке? Сначала строите одно сечение, например по X, затем согласно manual включаете Data>Extract>Current Slices, затем в средстве построения сечений меняете направление на перпендикулярное. Сохраняете рисунок. Можно так и далее менять направления. …».

Надо отметить, что освоение программ вызвало большой энтузиазм от освоения новых умений и навыков. Что иногда приводило к некоторым курьезом. В порыве энтузиазма некоторые студенты пытались построить линии тока для заведомо скалярных величин, таких как температура или давление. Здесь приходилось обращать внимание студентов на физическую природу решенной аэрогазодинамической задачи.

Следует также отметить, что достаточно простые задания с gnuplot вызвали у студентов наибольшее количество трудностей. Это стоит отнести к менее дружественному интерфейсу данной программы. Здесь выполнение заданий требовало на практике до 4 коррекций.

5. Результаты

В данном разделе приводятся характерные примеры заданий и общие результаты выполнения заданий.

Пример задания для TecPlot

1. ЦЕЛЕВАЯ УСТАНОВКА: Освоение и практическое применение универсальной программы визуализации данных TecPlot.

2. СОДЕРЖАНИЕ ПРАКТИКИ:

2.1 Изучить: универсальной программы визуализации данных TecPlot,

, написать краткий отчет по изучению, включающий основные типы данных, и перечень основных функций, предоставляемых пользователю.

2.2 Практически выполнить: Построить на рисунке визуализацию численного массива по данным файла MPrSh.lay в виде набора плоскостей. Файл содержит трехмерный массив данных, представляющий критическу скорость перехода между различными режимами течения пр вариации чисел Маха, Прандтля, Рейнольдса и Струхаля. Представить рисунок в виде набора кросс-сечений, используя инструменты программы. Рисунки предоставить в отчете.

2.3 Ознакомиться: с возможностями, предоставляемыми данной программой.

Результаты выполнения задания представлены на рисунках 1,2,3, где показаны 2 кросс-сечения, 3 кросс-сечения и 4 шага анимации, где одно из кросс-сечений перемещается.

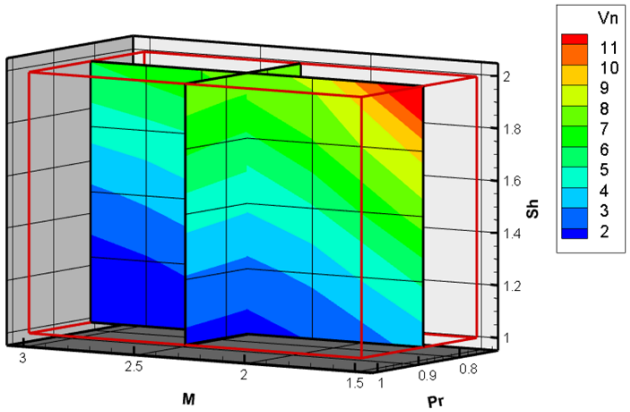


Рисунок 1. Набор двух кросс-сечений (по осям OX и OY).

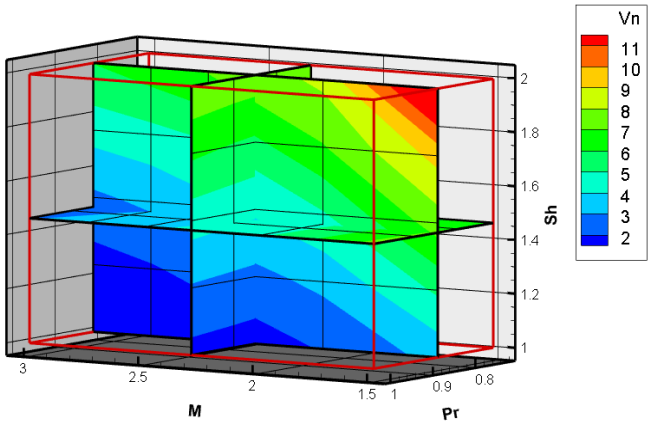
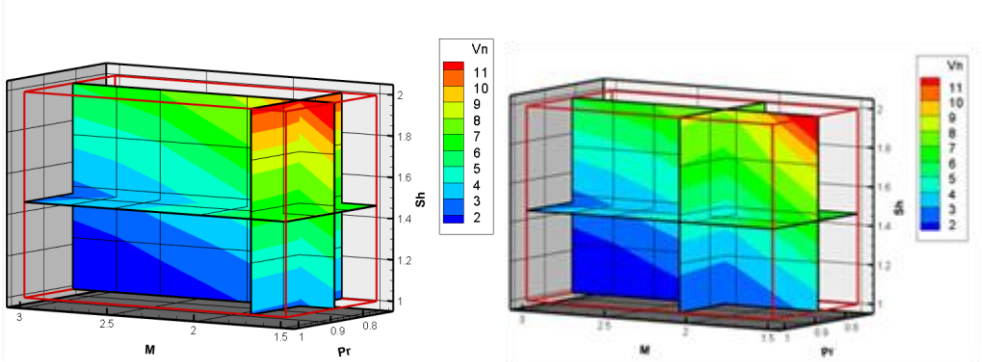


Рисунок 2. Набор трёх кросс-сечений (по осям OX, OY, OZ).



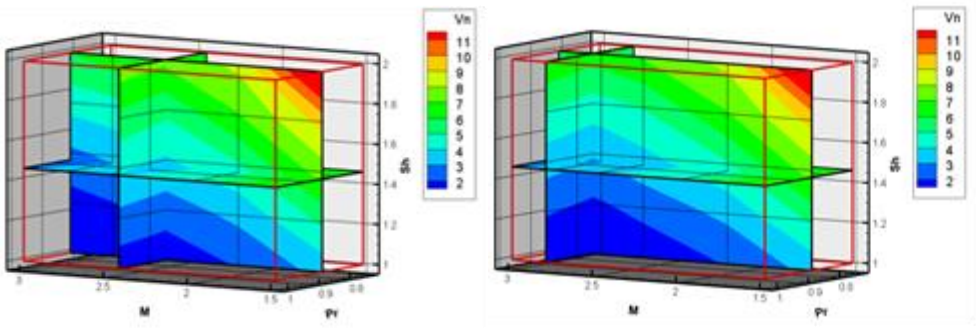


Рисунок 3. 4 шага анимации для сечения, параллельного оси OX.

Пример задания для gnuplot

«**1. ЦЕЛЕВАЯ УСТАНОВКА**: Освоение и практическое применение универсальной программы визуализации данных gnuplot.

2.1 Изучить: универсальной программы визуализации данных gnuplot.

, написать краткий отчет по изучению, включающий основные типы данных, и перечень основных функций, предоставляемых пользователю.

2.2 Практически выполнить: Построить изображение двумерного массива, представленного в файле Test2D.dat. Файл содержит двумерный массив данных, представляющий распределение скалярной величины F в прямоугольной области. В реальности эти данные соответствуют распределению температуры на плоскости. Построить изображение данных в виде трехмероной поверхности. Рисунки предоставить в отчете.

2.3 Ознакомиться: с возможностями, предоставляемыми данной программой.

3. ДОПОЛНИТЕЛЬНОЕ ЗАДАНИЕ:

4. ОРГАНИЗАЦИОННО-МЕТОДИЧЕСКИЕ УКАЗАНИЯ: использовать для визуализации данных gnuplot, версию программы брать с сайта http://www.gnuplot.info/

Результаты построения данных в виде трехмерной поверхности показаны на рис.4

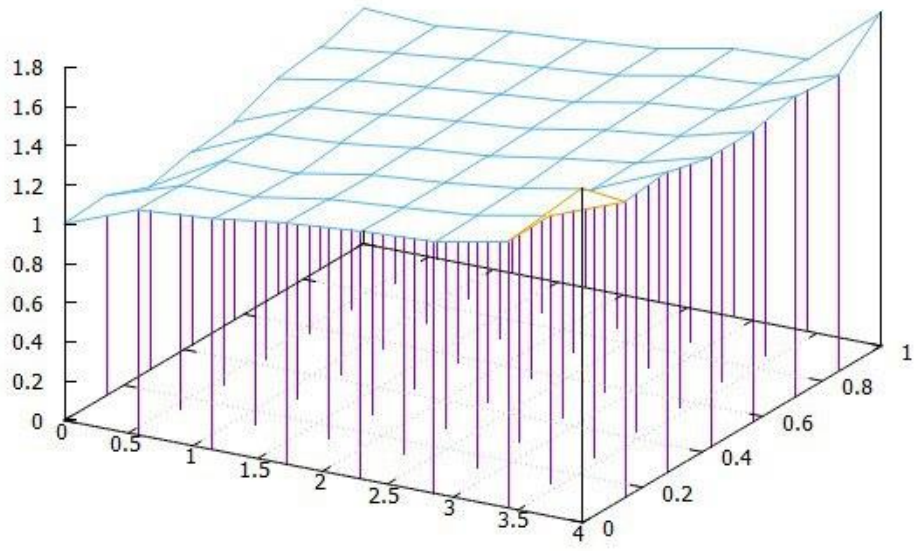


Рисунок 4. Построенная модель с помощью программы gnuplot

В целом, результаты проведения НИР можно оценить как достаточно успешные. Большинство (более 90%) студентов сдало НИР с положительной оценкой. Подобная самостоятельная работа, связанная с получением новых конкретных практических навыков вызвала интерес у студентов, что они и отразили во многих отчетах по НИР.

6. Заключение

В данной работе описан опыт построения и проведения НИР «Научная визуализация и визуальная аналитика» в РТУ-МИРЭА в 2024 году. НИР проводилась для студентов 3-го курса кафедры высшей математики Института искусственного интеллекта РТУ-МИРЭА. Описано построение и организация НИР. Приведены примеры заданий и их выполнения. Данная работа может быть интересной для преподавания аналогичных дисциплин в данной предметной области. Подобная организация работ служит источником получения студентами конкретных практических умений, которые востребованы в науке и промышленности.

Благодарности:

Автор выражает глубокую благодарность руководству Института искусственного интеллекта и заведующей кафедрой высшей математики д.ф.-м.н. Шатиной Альбине Викторовне за помощь и поддержку при создании и проведении НИР, а также м.н.с ИПМ им. М.В. Келдыша РАН Кувшинникову Артему Евгеньевичу и Бондаревой Николь Александровне.

Список источников

[1] Tecplot Visualization and Analysis Tools for CFD Post-pricessing https://tecplot.com/

[2] Paraview https://www.paraview.org/

[3] VisIt visualization system https://hpc.llnl.gov/software/visualization-software/visit

[4] gnuplot homepage http://www.gnuplot.info/

[5] Avizo 3D https://www.thermofisher.com/software-em-3d-vis/customerportal/download-center/amira-avizo-3d-installers/

УДК 004.89
DOI: 10.25206/978-5-8149-3873-2-2024-373-388

Инструментальное средство визуализации динамики событий в условиях неполноты картографического материала
Visualizing the dynamics of events in conditions of incompleteness of source cartographic material by means of special instrumental tool

С. И. Чуприна, М. И. Клишин [1]
1ФГАОУ ВО «Пермский государственный национальный исследовательский университет», ул. Букирева, 15, г. Пермь, 614068, Россия

S. I. Chuprina, M. I. Klishin[1]
[1]Perm State University, Perm, Bukireva Str. 15, 614068, Russia

Аннотация. В работе предложен онтологически управляемый подход к автоматизации труда разработчиков приложений виртуальной реконструкции исторических событий в условиях неполноты и неточности исходных сведений и отсутствия качественного картографического материала. Благодаря онтологически управляемому подходу к разработке инструментального окружения пользователь имеет возможность в автоматизированном режиме с использованием высокоуровневого интерфейса реконструировать произвольные исторические события на виртуальной графической карте с учетом конкретного хронологического порядка, вносить изменения и пополнять реконструкцию новыми сведениями без необходимости внесения изменений в исходный программный код системы. Исходными данными для системы являются текстовые документы с описанием исторических событий и картографическими изображениями. Для автоматизации извлечения знаний из текстовых коллекций и построения онтологий используются методы и средства генеративного ИИ. Изложение ведется на примере построения виртуальной реконструкции Куликовской битвы, картографический материал которой весьма неполон и представляет собой схематичные изображения, часто без указания масштаба, что характерно для карт средневековья. Для визуализации динамики событий в условиях неполноты картографического материала используются методы и средства 2D-анимации.

Ключевые слова: виртуальная реконструкция исторических событий, когнитивная визуализация, 2D-анимация, онтологически управляемое инструментальное окружение, генеративный ИИ, средневековые карты, Куликовская битва

Abstract. The paper proposes an ontology-driven approach to automating the work of application developers of virtual reconstruction of historical events in conditions of incompleteness and inaccuracy of source information and lack of high-quality cartographic material. Thanks to the ontology-driven approach to the instrumental tools development, the user has the opportunity to reconstruct arbitrary historical events on a virtual graphics card in an automated mode using a high-level interface, considering a specific chronological order, make changes and replenish the reconstruction with new information without having to make changes to the source code of the system. The initial information sources for the system are text documents with descriptions of historical events and illustrative cartographic images. Generative AI methods and tools are used to automate the extraction of knowledge from text collections and the construction of ontologies. The presentation is based on the example of building a virtual reconstruction of the Kulikovo battle, the cartographic material of which is very incomplete and consists of schematic images, often without specifying the scale, which is typical for maps of the Middle Ages. 2D animation methods and tools are used to visualize the dynamics of events in conditions of incomplete cartographic material.

Keywords: virtual reconstruction of historical events, cognitive visualization, 2D animation, ontology-driven instrumental environment, generative AI, medieval maps, Battle of Kulikovo.

I. Введение

В настоящее время все большее значение приобретает реконструкция исторических событий, выполняющая множество значимых социальных функций. Такая реконструкция направлена на пополнение культурного наследия нации, что способствует сохранению исторической памяти общества и обеспечивает понимание прошлого в контексте современности. Кроме того, она играет важную роль в образовании, научных исследованиях, музейном деле, туризме, игровой индустрии и во многих других областях. Например, виртуальная реконструкция событий находит свое применение в сфере юриспруденции, позволяя создавать виртуальные реконструкции событий с мест происшествия для выявления неполноты и неточности имеющихся

фактов, обнаружения взаимосвязей между разрозненными фактами, построения цепочек хронологически и пространственно связанных событий, выдвижения и проверки разного рода гипотез и поддержки принятия обоснованных решений. С аналогичными целями визуальная реконструкция событий востребована в военном деле, в т.ч. для проведения ретроспективы прошедших сражений и планирования хода дальнейших действий на поле боя.

Историческая визуализация как деятельность по представлению исторических сведений и знаний в визуальной форме является одним из направлений научной визуализации. Если говорить об исторических исследованиях в контексте научной визуализации, то преимущественно исторические события реконструируются с целью воссоздания и погружения человека в исторический контекст соответствующего периода для приобретения более подробных знаний о данных событиях, выявления неполноты, неточности и даже противоречивости различных исторических свидетельств. Еще одной важной функцией реконструкции исторических событий является возможность использования ее для анализа не только прошлых событий. Путем изучения воссозданных моделей и сценариев исторических событий можно проводить аналитику событий, выявлять взаимосвязи между различными фактами, что позволяет лучше понимать дальнейший ход истории. Этот аспект реконструкции помогает исследователям и историкам более глубоко понять причины и последствия прошлых событий, а также прогнозировать возможные сценарии развития событий в будущем. И здесь роль средств научной визуализации трудно переоценить.

Предоставление доступных визуальных средств исторической реконструкции событий для их углубленного освоения и осознания значимости для современного мира особенно актуально в контексте воспитания подрастающего поколения, включая младший школьный возраст. Поэтому в настоящее время требуются новые формы подачи материала, соответствующие уровню современных ИТ-технологий. Новые инструментальные средства должны быть доступны не только профессиональным программистам как с т.зр. их цены и открытости использования, так и с т.зр. простоты адаптации к специфике конкретного исторического события, но и не предъявлять повышенных требований к уровню ИТ-квалификации пользователя, а также не требовать наличия обязательного доступа к высокопроизводительным вычислительным кластерам.

Для автоматизации процесса качественной реконструкции исторических событий в рамках нашего подхода разрабатывается инструментальное окружение, которое удовлетворяет указанным требованиям и позволяет создавать виртуальные модели и воспроизводить прошлые события средствами 2D-анимации, что обосновано в случае отсутствия качественного картографического материала. Мы разрабатываем онтологически управляемое инструментальное окружение, позволяющее создавать виртуальные реконструкции событий на принципах когнитивности и интерактивности в условиях неполноты и нечеткости информации.

Это актуально, в первую очередь, в случае визуальной реконструкции давних исторических событий, например, эпохи средневековья. В средние века исторических карт в традиционном понимании этого слова не существовало: информация о соответствующих некоторому событию местности и объектах представлялась в виде схематичных карт или зарисовок, зачастую без указания масштаба, которые с современной точки зрения весьма условно можно назвать картами. В рамках данной работы мы будем называть их «схем-картами». Именно поэтому, создавая приложения для виртуальной реконструкции исторических событий (ВРИС), надо учитывать, что исходные для реконструкции данные представляют собой не векторные, а растровые изображения и могут не иметь масштаба.

Растровые изображения, используемые в качестве основы для виртуальной реконструкции, могут быть неполными, неоднородными и содержать искажения, что затрудняет точное восстановление пространственных и даже временных координат. Наш подход предполагает сохранение соответствующих метаданных об авторстве и первоисточниках средневековых карт, что позволяет в интерактивном режиме визуализировать несколько вариантов реконструкции одного и того же исторического события в привязке к соответствующим метаданным и выполнять, в случае необходимости, их семантическую фильтрацию с сохранением комментариев исследователя.

В настоящее время для ВРИС наиболее распространенными и доступными являются редакторы векторной графики. Однако, большой объем исторических картографических материалов представляет собой растровые изображения, поэтому требуется преобразование картографического материала в векторное представление, для чего имеется большой арсенал известных и доступных средств. Однако, как показывает практика, качество такого преобразования в условиях описанных выше ограничений средневекового картографического материала оставляет желать лучшего.

Динамическая виртуальная реконструкция исторического события представляет собой последовательную визуализацию ряда взаимосвязанных событий, причем эти взаимосвязи носят не только временной и пространственный характер, но и разнообразный семантический, в т.ч. политический, результат которых так или иначе влияет на ход последующих событий. Поэтому с когнитивной точки зрения очень важно демонстрировать

пользователю переходы от одного события к другому в контексте указанных взаимосвязей, предоставляя возможность интерактивной визуализации всех взаимосвязанных данных. Так, например, военный совет, созванный Дмитрием Донским для определения дальнейшего места сражения, оказал решающее влияние на весь ход Куликовской битвы. Однако, как это демонстрируется далее в разделе с описанием концепции предлагаемого подхода, не существует точных исторических данных об участниках, месте и времени проведения этого военного совета.

Необходимо отметить, что динамика событий на схем-картах средневековья обычно представляется статически в виде одного рисунка с использованием всякого рода стрелок, поэтому в современных системах ВРИС разработчик, хотя и в автоматизированном режиме, но практически с нуля создает на основе статичного изображения «раскадровку» динамики событий. Мы предлагаем автоматизировать этот процесс с учетом того, что вновь поступившие исторические факты о том или ином событии могут опровергать ранее заложенные в систему факты и знания, либо, наоборот, свидетельствовать в их пользу. Для того, чтобы такая ситуация не приводила к необходимости вносить изменения в ранее отлаженный программный код системы, мы реализуем нашу систему на основе модельно-ориентированной архитектуры (англ., MDA – Model Driven Architecture), используя для этих целей онтологическую модель. Онтологически управляемый подход имеет целый ряд преимуществ, аналогично тому, как и для чего онтологии используются в качестве промежуточного слоя знаний (графа знаний, англ., Knowledge Graph) в технологиях фабрик данных (англ., Data Fabric) [1].

Таким образом, проблема исторической визуальной реконструкции событий носит комплексный характер. Наш подход к решению указанных проблем акцентируется на создании онтологически управляемых инструментальных средств автоматизации труда разработчиков, включая разработчиков-непрограммистов, которые хотят строить виртуальные карты исторических событий и визуализировать их в динамике при описанных выше ограничениях с точки зрения картографической составляющей. Во-вторых, мы предлагаем способ учета неполноты и противоречивости исходных данных посредством автоматизированной адаптации системы к постоянно пополняющейся базе знаний (БЗ) за счет онтологически управляемого решения и визуализации всех доступных метаданных о соответствующих источниках исторических событий.

II. Виртуальная историческая реконструкция событий как направление научной визуализации

В контексте тематики данной работы научную визуализацию можно рассматривать с точки зрения как научно-исследовательской, так и научно-образовательной. Как следует из [2], научно-исследовательская визуализация направлена на представление текущих данных исследований в форме, наиболее удобной для понимания исследователями, проведения анализа и оценки, а также для коммуникации между учеными. Научно-образовательная (научно-популярная) визуализация позволяет доступно и наглядно иллюстрировать результаты исследований для широкой аудитории, способствуя распространению и популяризации знаний.

К основным категориям объектов исторической визуализации можно отнести:

- действующие лица (визуализация ключевых исторических фигур и персонажей);
- артефакты (моделирование и представление исторических объектов);
- документы (цифровое представление исторических документов и источников);
- время (использование хронолент для отображения временных последовательностей событий);
- пространство (визуализация географических и пространственных аспектов исторических событий);
- явления (моделирование исторических процессов и событий);
- связи (отображение связей и взаимодействий между элементами исторических событий).

В контексте цифровой исторической визуализации перспективными считаются следующие ее формы:

- антропологическая реконструкция (воссоздание облика исторических личностей);
- моделирование артефактов (цифровая реконструкция и визуализация исторических объектов);
- архитектурная реконструкция (воссоздание исторических зданий и сооружений);
- цифровое представление документов (оцифровка и визуализация исторических текстов и источников);
- хронолента (лента времени для отображения последовательности исторических событий);
- «живая» карта (интерактивные карты, показывающие изменения в пространстве и во времени);

- реконструкция процессов (визуализация динамики исторических событий);
- сети и генеалогические деревья (отображение родственных и социальных связей);
- викторины и игры (образовательные игры для взаимодействия с историческим материалом);
- виртуальные туры (интерактивные экскурсии по историческим местам).

Создание качественной исторической визуализации требует слаженной работы команды специалистов из различных областей. Ученые-историки собирают достоверные факты и составляют описания, дизайнеры разрабатывают графические решения, а разработчики выбирают технологии и создают программное обеспечение. Особое внимание уделяется верификации визуализаций профессиональными историками, чтобы избежать неточностей и ошибок, которые могут привести к искажению исторических фактов.

Для автоматизации труда всех перечисленных категорий пользователей требуется комплексное решение, что и обусловило использование для проектирования и реализации инструментального окружения OntoRecon методов и средств онтологического инжиниринга. Эти средства позволяют сохранять в виде онтологий метаданные о категориях визуализируемых на виртуальной карте объектов в привязке к их пиктографическим изображениям, относительным временным и пространственным координатам схем-карты, а также к информации о соответствующих элементах исторических событий, извлекаемых из первоисточников. Автоматизация процессов извлечения исторических фактов и знаний, построения онтологий, управляющих визуализацией динамики исторических событий, представленных на схем-картах, выполняется как отдельный этап работы методами генеративного ИИ вне среды разрабатываемого инструментального средства.

В контексте приведенной выше категоризации объектов и форм исторической визуализации при разработке OntoRecon основной акцент делается на создании средств виртуальной реконструкций процессов (динамики исторических событий), построения «живых» карт и хронолент.

На данный момент для реализации ВРИС на практике наиболее распространен подход, ориентированный на создание 3D моделей для последующего встраивания в виртуальную среду. Далее конструируются сценарии прохождения виртуального тура пользователем. К наиболее распространенным в России программным средствам- реконструкции исторических событий в 3D можно отнести: Unity3D, Unreal Engine (направлены на реконструкцию с помощью виртуальной реальности, ArcGis (позволяет интегрировать геопространственные данные), Blender (используется для создания 3D моделей), Google Tour Creator и др.

Использование методов и средств 3D-реконструкции событий, виртуальной реальности широко освещено в литературе, в том числе и для моделирования исторических событий. Эта тематика уже давно и достаточно полно отражена как в монографиях (см., например [3]), так и на страницах научных журналов (см., например, публикации журнала «Историческая информатика» на сайте https://nbpublish.com/e_istinf/), и выходит за рамки данной работы.

Конечно, указанные подходы обеспечивает глубокое погружение в эпоху того времени, что оказывает положительный эффект на впечатления пользователя. Но, к сожалению, для создания таких 3D реконструкций требуются большие ресурсные вложения как временные, так и финансовые, а также требуется привлечение специалистов из разных областей, а именно: дизайнеров, сценаристов, 3D-моделлеров, аниматоров, инженеров-программистов. Создание виртуальной реконструкции в 3D требует большого количества рабочего времени всей привлеченной команды, что также увеличивает итоговую стоимость проекта. Таким образом, создание виртуальной реконструкции исторического события в 3D до сих пор является весьма трудоемкой задачей, которая не может быть решена одним человеком и требует больших вложений.

Поэтому в нашем подходе к автоматизации построения ВРИС, ориентированном на широкий круг пользователей, в т.ч. не являющихся высококвалифицированными ИТ-специалистами, в условиях отсутствия качественного картографического материала, где входных данные не являются традиционными картами в современном понимании этого слова, а представляют собой простые схемы в растровом формате, мы используем 2D-анимацию..

Еще одним из факторов в пользу разработки новых средств ВРИС, отличных от применения приложений виртуальной реальности, является необходимость встраивания соответствующих сервисов в сторонние среды визуализации с целью сокращения стоимости проекта. Поэтому разрабатываемое нами онтологически управляемое инструментальное средство OntoRecon реализуется в виде платформы веб-сервисов с механизмом встраивания в инструментальную среду графического редактора Студии креативных карт – Creative Maps Studio (https://creativemaps.studio/).

III. О понятии исторического события и неполноте исторических данных

Мы солидарны с мнением известных представителей исторической науки (см., например, [4, 5]), что понятие «историческое событие» достаточно близко к понятию «история» и что, даже более, — историю можно интерпретировать как событие о чем-то или рассказ, так как историческое событие предполагает подход скорее не сугубо объективного описания исторических фактов историком, а совокупность всех сопутствующих комментариев, вставок, рассуждений. Поэтому в процессе разработки модели представления событий в задачах визуализации динамики событий в условиях неполноты картографического материала мы руководствовались следующим.

Все сущности («нечто»), связанные с историческими событиями, можно отнести к одной из трех основных категорий верхнего уровня: «Объект», «Ситуация», «Событие». Понятия, принадлежащие этим категориям, именуют вершины онтологии верхнего уровня. Именование является важной составляющей одного из начальных этапов моделирования – этапа идентификации понятий, который предполагает также выбор наиболее точных определений понятий с учетом других близких по смыслу понятий в контексте их взаимосвязей в соответствии с целями и задачами построения онтологии. Т.е. из всех возможных определений некоторого понятия в онтологии сохраняется то, которое наиболее адекватно отражает роль этого понятия в контексте целей и задач построения онтологии, поэтому наиболее близкие в этом контексте понятия именуют соседние вершины онтологии. Подробное описание механизма учета контекста близких по смыслу понятий и соответствующих метрик выходит за рамки данной работы. В ходе моделирования для отнесения того или иного понятия к одной из указанных выше категорий, необходимо учитывать связанные с ними понятия верхнего уровня моделирования и дать им более четкие определения.

Объект – это понятие, воспринимаемое индивидуально (оно может быть однозначно идентифицируемо), представляющее интерес само по себе (может рассматриваться вне связи с другими объектами) и имеющее определенную протяженность во времени. Событие – это понятие, воспринимаемое индивидуально, представляющее интерес само по себе и происходящее мгновенно. Историческое событие уже произошло: в этом смысле оно, в отличие от объекта, – «мгновенно», хотя у него есть характеристики типа «время-дата» начала и конца. Значение – это понятие, воспринимаемое индивидуально и представляющее интерес только в связи с некоторым объектом или событием. Говоря о взаимосвязях между сущностями, выделим такие категории как «отношение» и «атрибут», к которым в конечном счете можно свети все виды взаимосвязей предметного уровня.

Отношение – это временная или постоянная связь между сущностями категории Объект и/или Событие. Атрибут – это временная или постоянная связь между сущностями категории Объект или Событие и значением. В контексте моделирования динамических событий очень важно представление того, как изменяются объекты и связи во времени (объекты не обязательно существуют в любой момент времени, а связи с течением времени могут меняться). Состояние, образуемое объектами и связями в отдельные моменты времени, фактически и определяет то или иное событие. Событие может быть комплексным (составным), т.е. состоять из нескольких других событий (соответствующие этим событиям вершины связаны с вершиной онтологии, представляющей составное событие, отношением «часть-целое», что в графе онтологии отображается при помощи дуг с именем «a_part_of»).

Ситуацию мы рассматриваем как особый вид объекта и, в отличие от события, ситуация, как и любой другой объект, имеет определенную протяженность во времени. Например, ситуация, характеризующая то, что в контексте такого события как Куликовская битва Московским княжеством правил Дмитрий Донской. Эта информация важна безотносительно конкретных сроков правления: важно, что Дмитрий Донской правил Московским княжеством во время Куликовской битвы, некоторое время до и после Куликовской битвы, т.е. такая ситуация сохранялась в течение продолжительного, по сравнению со сроком любого события, времени. В нашей модели ситуация в отличие от обычного объекта именуется глаголом действия, характеризующего эту ситуацию, а в качестве атрибутов ситуации выступают не произвольные характеристики обычных объектов, отражающие их свойства, а часть падежей падежных рамок, предназначенных для описания событий.

Обязательным является представление таких семантических категорий как А́генс и (от лат. agens — действующий) и Па́циенс (от лат. patiens — «принимающий, терпящий» действие). Агенс отвечает на вопросы «Кто?» или «Что?» и обозначает объект (одушевленный или неодушевленный), который играет активную роль в ситуации, т.е. выполняет действие, указанное в названии ситуации и/или контролирует ее. В нашей модели онтологии связь с агенсом представляется дугой с именем «agent», исходящей из неименованного центра падежной рамки предиката ситуации. Пациенс представляет собой пассивного участника ситуации, который вовлечен в ситуацию, но не контролирует и не исполняет ее, т.е. – это объект (одушевленный или неодушевленный), на который направлено действие, выраженное глаголом, именующим ситуацию. В нашей модели онтологии связь с пациенсом представляется дугой с именем «object», исходящей из неименованного центра падежной рамки предиката ситуации.

С учетом описанных категорий моделирования верхнего уровня была разработана модель онтологии предметного уровня. При этом в ходе онтологического моделирования мы стремились минимизировать базовый набор поддерживаемых типов связей для повышения эффективности реализации механизмов интерпретации онтологий. Несмотря на то, что большой набор поддерживаемых типов связей облегчает моделирование, для эффективной реализации указанных механизмов могут потребоваться значительные вычислительные ресурсы и применение технологий высокопроизводительных вычислений. Мы же, как уже отмечалось, ориентируемся на доступность разрабатываемых нами инструментальных средств обычным пользователям и небольшим исследовательским группам, а автоматизацию процесса построения онтологий при помощи сервисов генеративного ИИ выполняем отдельно и вне рамок инструментального окружения.

Под неполнотой исторического и картографического материала в данной работе мы понимаем такой вид нечеткости, который в классической теории БЗ и экспертных систем характеризуется как «неполные знания и немонотонная логика»). Этот вид нечеткости обусловлен тем, что знания в БЗ неполны, поступают в систему фрагментарно (база знаний время от времени пополняется новыми фактами и знаниями), а вновь поступившие знания могут приводить к отрицанию ранее накопленных фактов и знаний, противоречить ранее сделанным выводам. То есть в ходе логического вывода на текущей БЗ и фактов для логической формулы P может одновременно стать истинным и P, и¬P, что, как известно, в классической логике является противоречием. Система, работающая в условиях неполноты БЗ, в ходе символических рассуждений должна быть способна работать в таких нечетких условиях, например, строить цепочку логического вывода на основе концепции построения в рамках текущей БЗ и фактов т.н. «возможного мира», в котором нет противоречий.

В дальнейшем изложении примерами источников данных, иллюстрирующих применение предлагаемой концепции на практике, будут служить ресурсы интернет, посвященные Куликовской битве, из которых в автоматизированном режиме извлекаются необходимые картографические данные и исторические сведения. Выбор исторической реконструкции Куликовской битвы для демонстрации возможностей разрабатываемой системы обусловлен тем, что доступные для этой реконструкции исходные исторические документы в полной мере отражают проблемы, связанные с виртуальной реконструкцией исторических событий в условиях неполноты исторических данных и отсутствия качественного картографического материала.

IV. Концепция предлагаемого подхода

На вход системы поступают документы в текстовом формате с описанием исторического события и иллюстрациями в виде схем-карт. Инструменты OntoRecon, управляемые знаниями и фактами из онтологии соответствующего исторического события, позволяют в автоматизированном режиме распознать на иллюстрации необходимые для визуализации элементы исторических событий и строить методами 2D- анимации «живую карту» событий с учетом их хронологического порядка. Визуализация процессов в виртуальной среде позволяет пользователю в интерактивном режиме простым кликом по пиктограмме, изображающей некоторый элемент исторического события, получать дополнительные исторические данные об этом элементе, которые хранятся в предметной онтологии и могут постоянно пополняться/обновляться без внесения изменений в программный код системы визуализации. Эта дополнительная информация хранится в онтологии с указанием источника и/или авторства, что позволяет при наличии средств семантической фильтрации учесть персональные предпочтения пользователей и в динамике исключить из процесса реконструкции те или иные источники исторических сведений.

1. Основные этапы решаемой задачи

Как отмечалось выше, несмотря на наличие большого числа продвинутых средств визуальной реконструкции исторических событий, таких как Unity3D, в настоящее время автоматизация труда разработчика таких средств по-прежнему остается проблемой и требует высокого уровня квалификации и больших временных затрат, особенно в условиях неполноты и неточности исходных данных. Актуальным является разработка таких средств реконструкции, которые не только выполняют в автоматическом режиме часть работ по реконструкции исторических событий в графической форме непосредственно из исходных источников данных, не являющихся в полном смысле слова картографическими, но и позволяют без привлечения значительных ресурсов, используя высокоуровневые средства, обогащать визуальную реконструкцию по мере поступления дополнительных и уточняющих сведений. Эта проблема учета неполноты и неточности исходных данных носит общий характер безотносительно к специфике используемого метода реконструкции. На рис. 1 приведен ответ нейросети ChatGPT на запрос о местоположении проведения военного совета 6 сентября 1380г., из которого видно, что даже хорошо обученные на огромном массиве данных большие лингвистические модели не могут предоставить точной информации с силу специфики исторических данных.

Концепция предлагаемого подхода заключается в разработке отдельных сервисов, которые позволяют автоматизировать процесс создания ВРИС в условиях неполноты исторических данных и отсутствия качественного картографического материала. Рис. 2 в обобщенном виде иллюстрирует последовательность основных этапов, которые необходимо реализовать для создания виртуальной реконструкции исторического события в рамках предлагаемого подхода. Для каждого этапа виртуальной исторической реконструкции, кроме этапа №6, в среде разрабатываемого инструментального окружения OntoRecon создается отдельный сервис, автоматизирующий данный процесс. Для реализации этапа №6 используется сервис рендеринга хронологической ленты в среде стороннего графического редактора Creative Maps Studio [6].

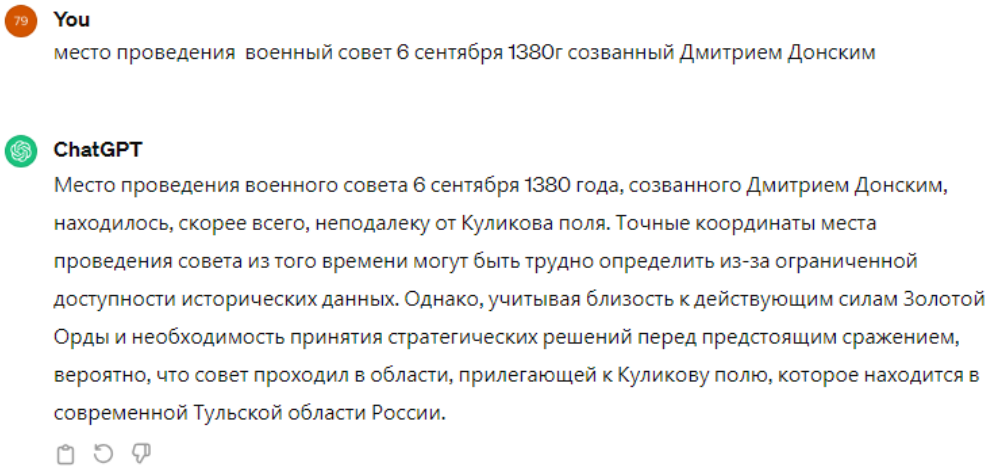


Рис. 1. Свидетельство неточности исторических данных

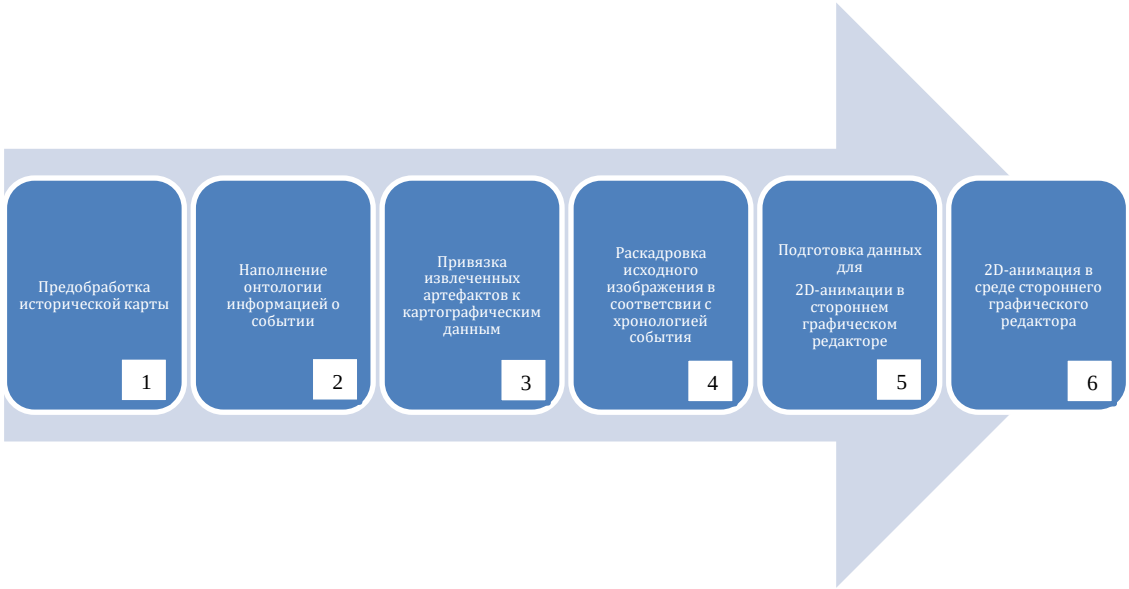


Рис. 2. Основные этапы решаемой задачи в рамках предлагаемого подхода

В случае реконструкции динамических исторических событий и других, которые могут быть сведены к категории исторических в том смысле, как это описано в разделе III, помимо этапов идентификации и концептуализации предметного содержания их первоисточников, привязки к картографическим данными и пр., важная роль в задачах 2D-анимации отводится определению величины, характеризующей отношение времени демонстрации кадров карты к реальному времени процесса, т.е. установлению т.н. временного масштаба. Однако для многих карт, особенно исторических карт средневековья и более ранних, временной масштаб установить достаточно сложно. Поэтому, для того чтобы демонстрировать динамику развития событий, необходимо извлекать нужную информацию из дополнительных информационных источников, зачастую текстового характера. Извлекая такую информацию и сохраняя ее в виде онтологии на основе некоторой семантической модели, мы имеем возможность анимировать статичное изображение 2D-карты, подбирая соответствующий временной масштаб. В отсутствии традиционных масштабных карт временной и пространственный масштаб в процессе визуализации измеряется в относительных величинах: относительно выбранной на изображении точки

начала координат и визуально подобранного масштаба карты. Наложение масштабной сетки на растровое представление схем-карты выполняется специальными средствами визуализации на этапе предобработки.

2. Роль и место онтологий

Предлагается многоуровневая гибридная модель онтологии. Онтология верхнего уровня определяет основные понятия и связи между ними, характерные для описания любого события на основе данных исторической картографии и исторических текстовых ресурсов. Предметные онтологии служат для описания конкретных событий и метаданных, связанных со способами визуализации категориальных понятий (все понятия одной категории визуализируются на карте одной и той же пиктограммой, но маркируются в соответствии с конкретным названием), и наследуют знания онтологии верхнего уровня.

На рис. 3 представлен фрагмент онтологии верхнего уровня, которая представляет собой базовую семантическую структуру для описания любого исторического события. В этой онтологии определена таксономия основных понятий, таких как «Объект», «Ситуация», «Событие» и др. Концепция описания ситуации приведена в разделе III. Фрагмент онтологии с описанием ситуации приведен на рис. 4. Фрагмент онтологии, описания исторического события «Куликовская битва» приведен на рис. 5.

Для представления событий используется модель на основе падежных рамок Ч. Филлмора. Как отмечается в [7] «Ч. Филлмор представил план содержания предложения как пропозицию – вневременной набор семантико-синтаксических функций («ролей») для именных компонентов, чей состав и взаимоотношения задаются лексическим значением предикатного знака, как правило, глагола». Семантические роли (семантические падежи) представляют те понятия, «в терминах которых человек судит о происходящих вокруг него событиях». Каждому глаголу соответствует т.н. падежная рамка, т.е. определенный набор падежей, а каждому из падежей соответствует определенный участник ситуации, или партиципант. Каждый семантический падеж предполагает ответ на определенный тип вопроса.

Существую разные трактовки падежных рамок и смысла самих падежей (см., например, [7]). Мы предлагаем «гибридную» модель онтологии, в которой разные сущности, относящиеся к разным категориям (классам): объекты, ситуации, события, – моделируются по-разному, но основные виды парадигматических отношений типа «класс-подкласс», «экземпляр-класс» «часть-целое» имеют общеизвестный смысл и трактовку.

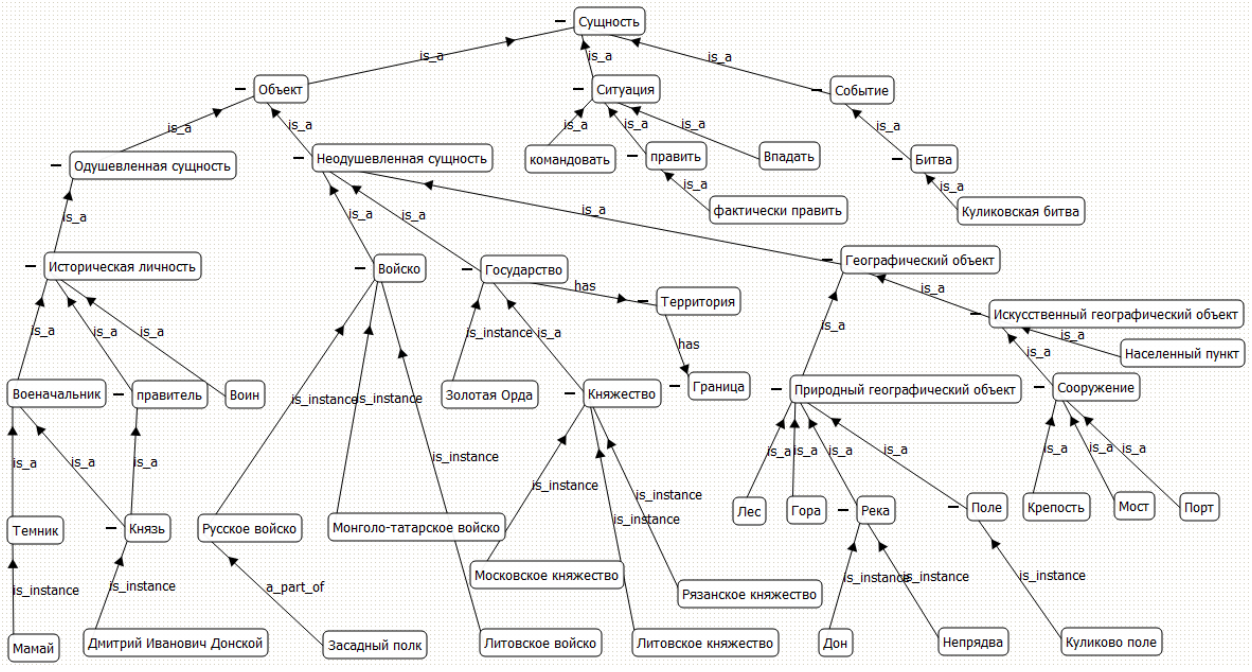


Рис. 3. Фрагмент онтология верхнего уровня

Помимо описанных ранее падежей, соответствующих таким семантическим категориям, как «Агенс» и «Пациенс» (на рис.5 в соответствующие вершины онтологии входят связи «agent» «object», соответственно), поддерживаются связи «goal» для описания цели события, «date» для описания даты события (времени действия), «result» для описания результата события, «location» для описания места, где произошло событие, и «instrument» для описания орудия/движущей силы, с помощью которой был получен описываемый результат того или иного события.

N-арные связи устанавливаются посредством введения неименованных вершины онтологии.

Наш подход предполагает, что процесс визуализации динамики событий выполняется не просто на основе онтологии, т.е. когда онтология пассивна, а сервисы системы обращаются к ней за фактами и знаниями, а является онтологически-управляемым. Онтологически управляемый подход к разработке программных систем является разновидностью модельно-ориентированного подхода, где в качестве модели выступает модель онтологии, и предполагает, что именно механизм интерпретации онтологии в процессе обхода графа онтологии, управляет функционирования системы, вызывая нужные функции в нужном порядке с учетом текущего контекста онтологии и хранящихся в ней метаданных.

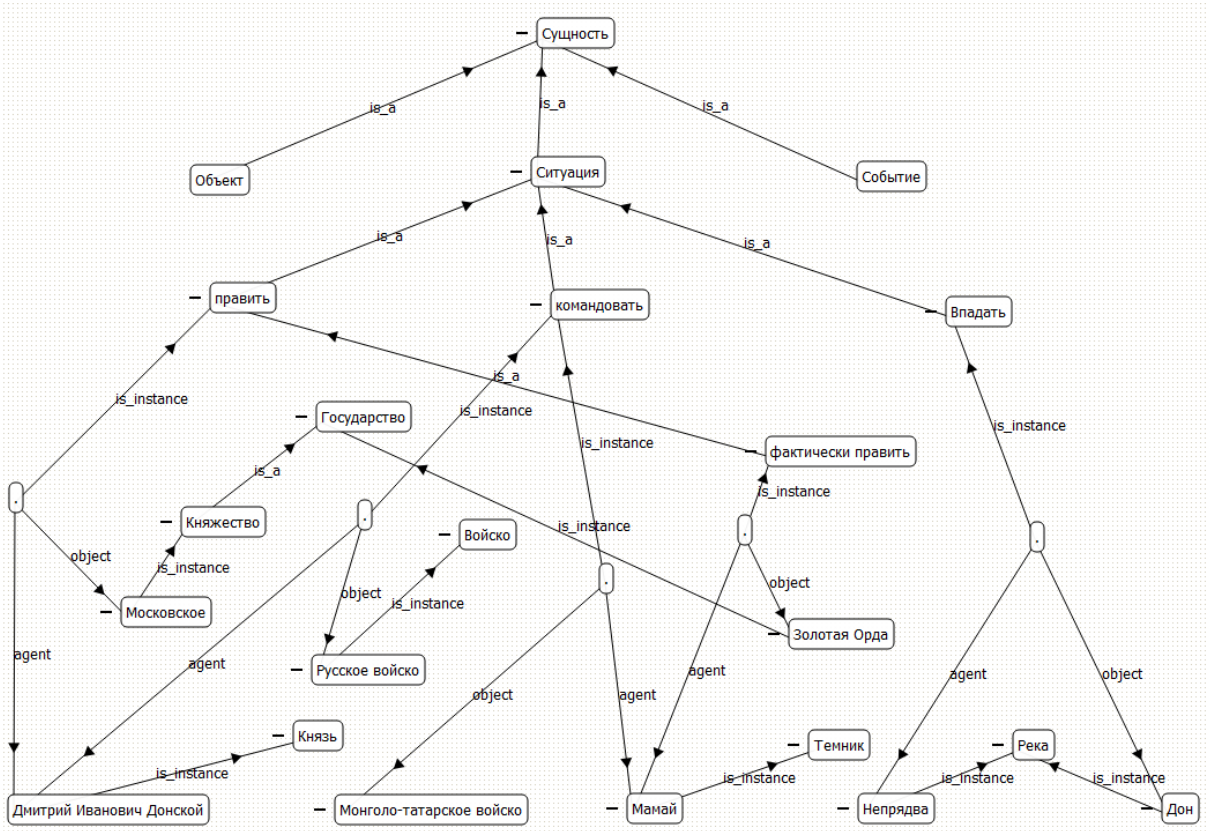


Рис. 4. Фрагмент онтология описания ситуации

Поэтому основным содержанием текущего этапа реализации инструментального окружения, предназначенного для автоматизации работ по визуализации динамики событий в условиях неполноты картографического материала и ориентированного на широкий круг пользователей, в т.ч. непрофессионалов в сфере ИТ, является исследование адекватности предлагаемой модели онтологии решаемым задачам и потребностям целевой аудитории.

1. *Учет неполноты исторического материала*

В соответствии с целями разработки инструментального окружения и целевой аудиторией общая идея нашего подхода к учету описанного выше понятия неполноты исторического и картографического материала заключается в предоставлении пользователям возможности различной интерпретации одних и тех же исторических событий, включая их временные и пространственные характеристики, а также учета персональных предпочтений пользователя посредством исключения из процесса анимации тех кадров, которые построены на основе данных из первоисточников, к которым он по тем или иным причинам не испытывает достаточного доверия. Это возможно благодаря тому, что используемый нами визуальный редактор онтологий ОНТОЛИС [9, 10] позволяет сохранять в атрибутах вершин и дуг графа онтологии необходимые метаданные о первоисточниках и их авторстве. Метаданные хранятся в формате JSON и система поддерживает унифицированный механизм их обработки. Если какие-то представленные в онтологии знания и факты являются результатом генерации внешних или внутренних относительно инструментального окружения сервисов, например, GPT, сведения об этом также сохраняются в метаданных соответствующих вершин онтологии (см. рис. 6).

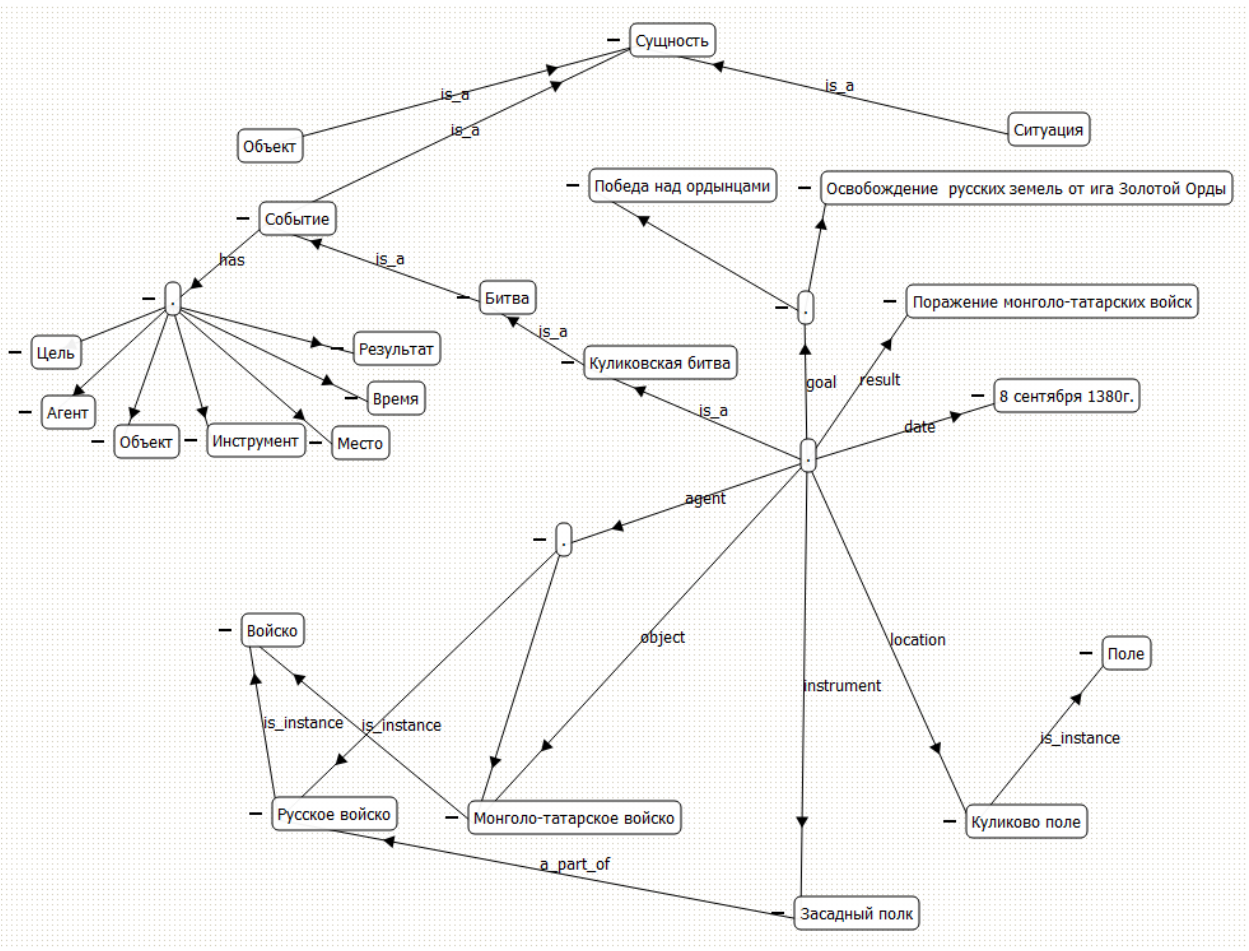


Рис. 5. Фрагмент онтология описания события

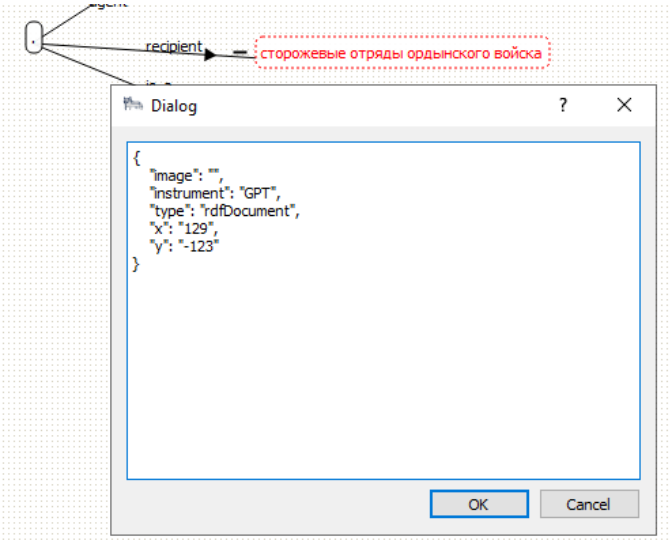


Рис. 6. Метаданные об источнике информации в атрибутах вершин онтологии

2. Автоматизация построения онтологий

Для создания виртуальных исторических реконструкций, в т.ч. в условиях неполноты картографического материала необходимо работать с большим объемом текстовых документов, содержащих историческую информацию. Для автоматизации построения онтологий мы используем возможности генеративного ИИ – сервисы OpenAI. Ввиду ограничений на объем публикации мы приведем только некоторые результаты автоматизированного построения онтологий. На рис. 7 представлен пример запроса к сервисам Open AI для автоматического создания RDF-триплетов на основе текста, посвященного «Куликовской битве» с официального

сайта ГОСУДАРСТВЕННЫЙ МУЗЕЙ-ЗАПОВЕДНИК «КУЛИКОВО ПОЛЕ» (https://kulpole.ru/bitva/hod-bitvy) и соответствующий ответ чат-бота GPT-4o до и после уточнения запроса.

Наша методика использования сервисов генеративного ИИ предполагает последовательное задание запросов сначала на построение RDF-троек, и только затем – модели онтологии в формате падежных рамок Ч. Филлмора. На рис. 8-7 представлен результат автоматического создания средствами GPT-4o падежных рамок, причем онтологическая модель падежных рамок была расширена семантической ролью (падежом,) отвечающим за описание источника (англ., source) – исходного ресурса с историческим описанием события.

Имея такие результаты, сервисы OntoRecon осуществляют автоматическую генерацию онтологий в соответствии с описанной ранее гибридной моделью (см. графическое представление онтологического описания ситуаций и событий на рис. 4-5.

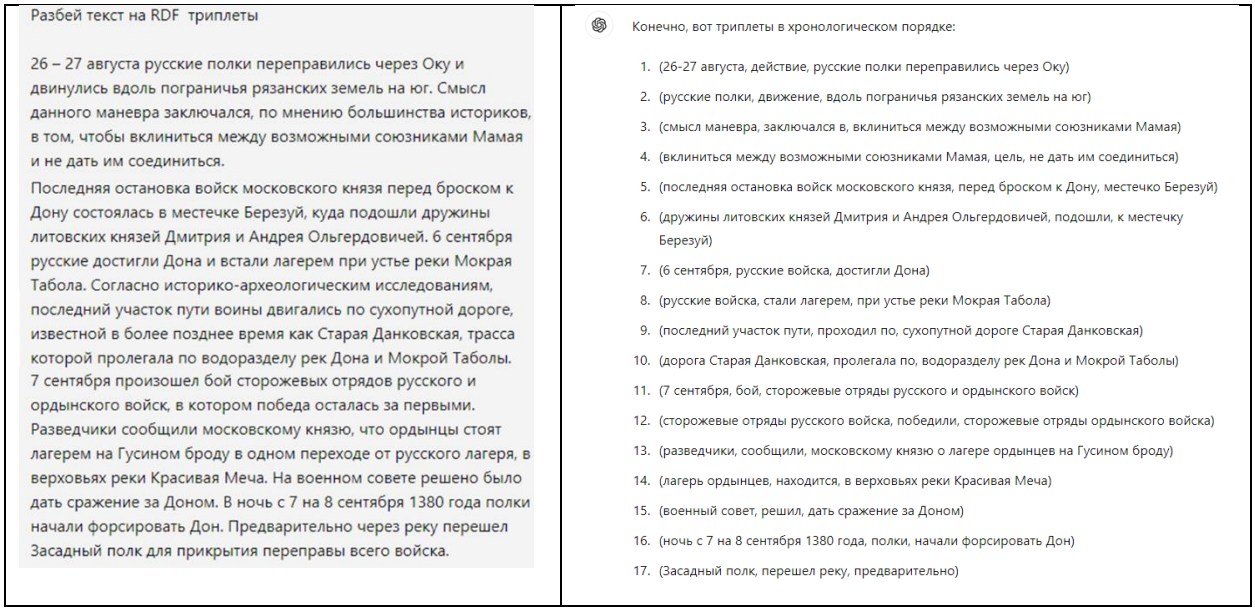


Рис. 7. Автоматизация построения RDF-триплетов

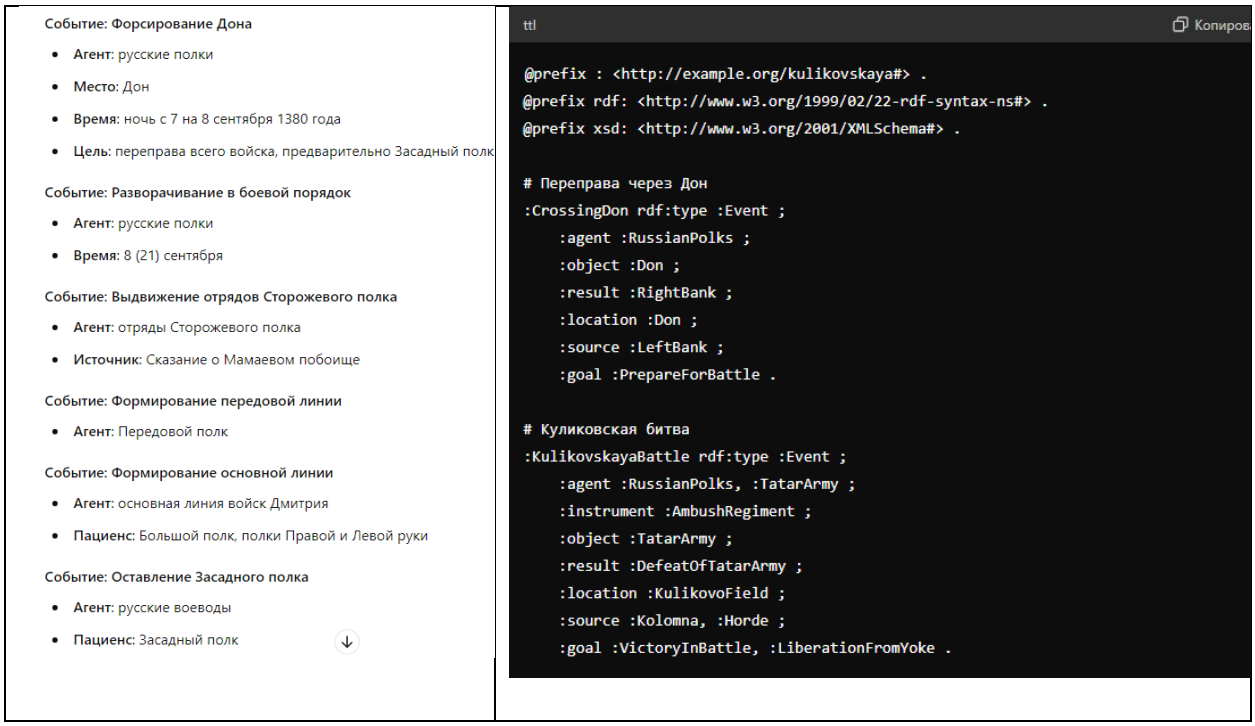


Рис. 8. Автоматизация построения падежных рамок

V. Описание реализации предложенного подхода

1. *Обобщенная архитектура демо-прототипа OntoRecon*

Исходя из концепции предлагаемого подхода, в состав архитектуры OntoRecon входит несколько сервисов (см. рис. 9):

- сервис предобработки исторических карт – отвечает за сбор текстовой информации с изображения, очистку изображения от статической реконструкции, определение контуров географических объектов и их координат;
- сервис пополнения (генерации) онтологии историческими данными, извлеченными при помощи средств генеративного ИИ, в привязке к изображениям соответствующих кадров;
- сервис меппинга онтологии для преобразования построенной онтологии в формат, необходимый для визуализации в студии креативных карт и подготовки к автоматизации построения хронологической ленты событий средствами 2D-анимации (выполняется в студии креатвных карт).

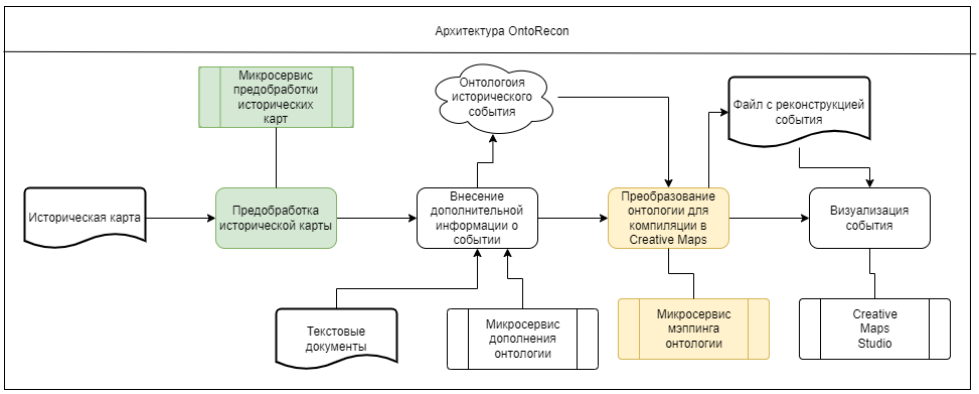


Рис. 9. Обобщенная архитектура OntoRecon

Предобработка изображения является первым шагом в процессе создания виртуальной реконструкции исторического события. Этап предобработки исторической карты включает в себя решение следующих задач:

- извлечение существующего на карте текста,
- очистка изображения от негеографических объектов (статической реконструкции),
- определение контуров и координат географических объектов,
- дополнение онтологии сведениями об извлеченных из изображения схем-карты артефактах.

Удаляемые в процессе очистки изображения от негеографических объектов значимые артефакты, например, стрелки, маркируются и сохраняются в метаданных системы вместе с указанием их относительных/абсолютных координат. Это необходимо для генерации последовательных кадров 2D-анимации, представляющих динамику исторического события. Ведь статичная по своему формату схем-карта тем не менее иллюстрирует не одно событие, а динамику составляющих его отдельных событий. На рис. 10, например, стрелки иллюстрируют динамику продвижения русских полков в ходе Куликовской битвы.

2. Распознавание текста на исторической схем-карте

Для реализации распознавания текста на изображениях был выбран известный движок Tesseract и готовая предобученная модель для распознавания на изображении русскоязычных букв и слов. Один из примеров, демонстрирующих результат распознавания текстовых данных на схем-карте Куликовского сражения с указанного выше сайта, приведен на рис. 10.

Как видно, сервисы Tesseract не смогли с первого раза качественно распознать русскоязычный текст, представленный прописью или курсивом. Однако эта проблема была легко устранена посредством дальнейшего дообучения. Кроме того, в OntoRecon предметное содержание онтологий, автоматически построенных на основе текстовых исторических ресурсов средствами генеративного ИИ, служит подсказкой на этапе валидации результатов распознавания. В перспективе предполагается полностью автоматизировать этот этап за счет онтологически-управляемой валидации и средств устранения «ошибок набора», которые посредством сопоставления распознанных терминов и понятий из эталонной онтологии исправляют неверно распознанные символы.

Рис. 10. Пример результата распознания русскоязычного текста на средневековой схем-карте

3. *Очистка изображения схем-карты от артефактов статической реконструкции*

Прежде чем извлечь из изображения необходимые артефакты и увязать их представление с понятиями из онтологии с учетом временной и пространственной составляющих, требуется получить «подложку» – слой карты без исторических артефактов, т.е. как бы воспроизвести 1 этап ручного создания карты – создать схем-карту местности, содержащую только географические объекты без исторических артефактов, связанных с событиями. Для краткости, мы называем этот этап «очисткой» и используем для его реализации нейросеть U-Net, которая позволяет получить качественные результаты при использовании обучающих множеств даже небольшого объема. На рис. 11 слева представлена исходная историческая схем-карта Куликовской битвы, справа – результат ее очистки от артефактов исторических событий статической реконструкции.

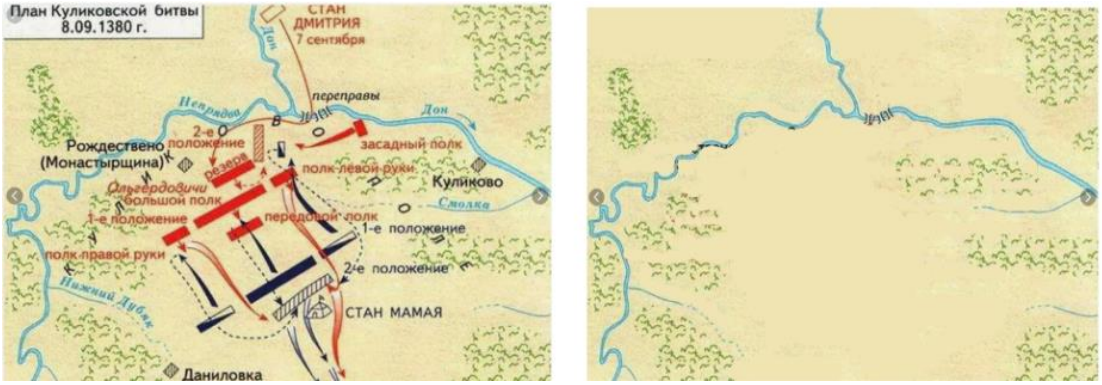


Рис. 11. Результат очистки схем-карты от негеографических объектов

4. *Пополнение онтологии сведениями об извлеченных из изображения схем-карты артефактах*

Занесение в онтологию сведений об извлеченных из изображения схем-карты артефактах и текстовых данных реализуется с помощью вопросно-ответной подсистемы, управляемой онтологией верхнего уровня и таксономией основных понятий (см. рис.3). Активную роль в диалоге играет система: в процессе обхода онтологии верхнего уровня сервис генерирует соответствующие содержанию этой онтологии вопросы и предлагает варианты ответов. Пример работы вопросно-ответной подсистемы представлен на рис. 12.

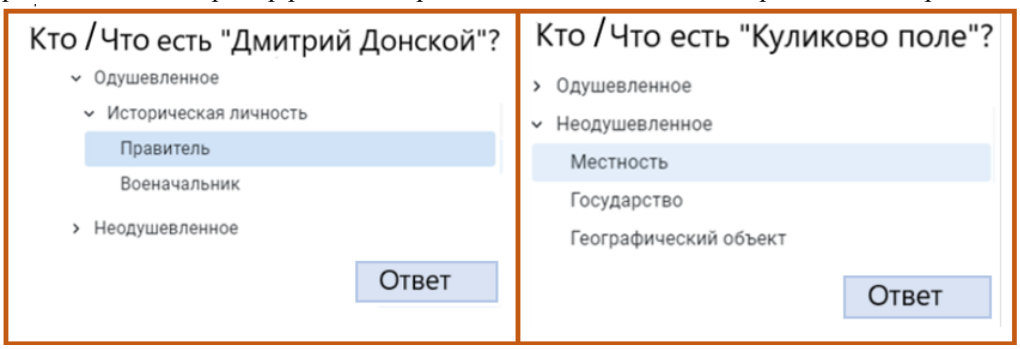


Рис. 12. Результат онтологически управляемой генерации интерфейса вопросно-ответной подсистемы

Важно отметить, что если на верхнем уровне онтологии будут внесены изменения, то эти изменения автоматически отобразятся на интерфейсе в генерируемых вопросах и вариантах ответов. Например, в

представленном на рис. 3 фрагменте онтологии представлена усовершенствованная версия таксономии понятия «Куликово поле»: «Неодушевленная сущность»/»Географический объект»/»Природный географический объект»»/Поле»»/Куликово поле», и эти изменения автоматически отобразятся на интерфейсе сервиса вопросно-ответной подсистемы, представленного на рис. 12.

5. Подготовка схем-карты для целей 2D-анимации в студии креативных карт

После очистки схем-карты от негеографических объектов и пополнения онтологии новыми сведениями на основе извлеченных из изображения текстовых данных, наступает этап пространственной и временной привязки к очищенной схем-карте извлеченных текстовых и иллюстративных артефактов. Выполняется своего рода восстановление исходного изображения, но покадрово, с целью его последующей трансформации в формат студии креативных карт и 2D-анимации. Описание сервиса меппинга онтологии для преобразования ее в формат, необходимый для визуализации в студии креативных карт и автоматизации построения хронологической ленты событий имеющимися там средствами 2D-анимации, выходит за рамки данной работы ввиду ограничений на объем публикации. Остановимся лишь на описании интерфейса, позволяющего выполнить все необходимые действия для подготовки кадров («раскадровки» статичного изображения) для целей последующей 2D-анимации.

Указанный интерфейс также управляется знаниями системы об извлеченных в ходе очистки схем-карты артефактах и носит активный характер со стороны системы, в ходе которого пользователю предлагается выбрать место на карте для пространственной привязки очередного артефакта. Для указания координат сущности пользователю необходимо нажать на кнопку «Указать на карте», после чего карта будет представлена вместе с наложенной на нее масштабной сеткой, на которой необходимо выбрать соответствующую ячейку (см. рис. 13).

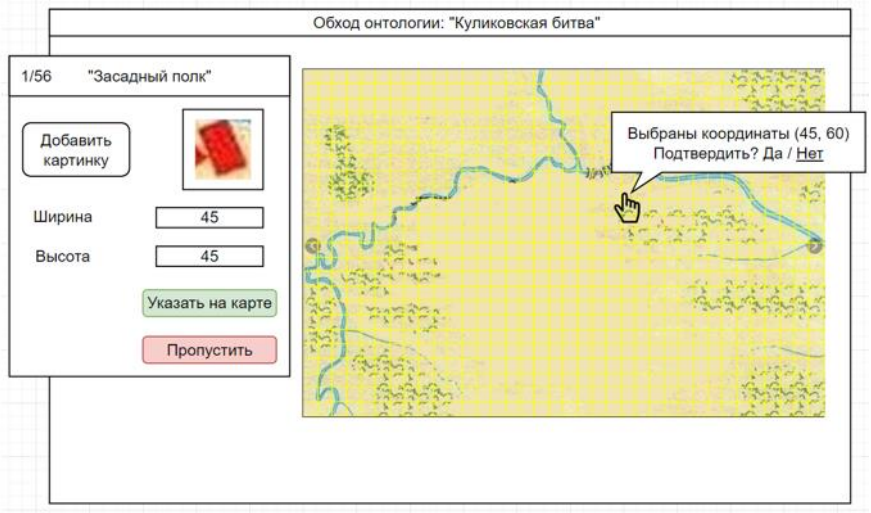


Рис. 13. Определение пространственных координат сущностей

Размер ячеек масштабной сетки можно менять интерактивно, подбирая такой вариант, при котором как можно меньше различных артефактов попадают в одну ячейку. Если же выбран вариант, когда в одну ячейку попадают сразу несколько артефактов, сервис предложит установить между ними один из вариантов пространственных типов связей: «Быть слева», «Быть справа», «Быть сверху», «Быть снизу», «Пересекаться в пространстве слева», …, «Быть наискосок слева», …, «Касаться слева», …, «Находиться на». В результате данные об относительных координатах автоматически пополняют атрибутивный состав тех вершин онтологии, к которым имеет отношение та или иная сущность (ранее на рис.6 был продемонстрирован пример некоторых атрибутов вершин онтологии). Полученные таким образом и сохраненные в онтологии относительные пространственные координаты в дальнейшем будут автоматически переведены в абсолютные для компиляции в среде Creative Maps Studio.

Следующим шагом в подготовке к 2D-анимации является так называемая «раскадровка», когда на основе одного или нескольких изображений строятся с учетом хронологии отдельные кадры одного и того же события (составляющие его отдельные события, например, переправа через реку и др.). Если предположить, что нам доступно для обработки только одно изображение схем-карты Куликовской битвы, например, представленное ранее на рис. 10, 11, 13, то такой вариант предполагает специальную обработку соответствующих артефактов, предназначенных для иллюстрации на статичном изображении динамики соответствующих событий. В демо-прототипе OntoRecon к таким «динамичным» артефактам относятся изображения всякого рода стрелок. На рис. 14 представлен фрагмент «раскадровки» указанного изображения (первые 2 кадра) в относительном временном масштабе «до» и «после» переправы русского войска в месте слияния рек Дона и Непрядвы.

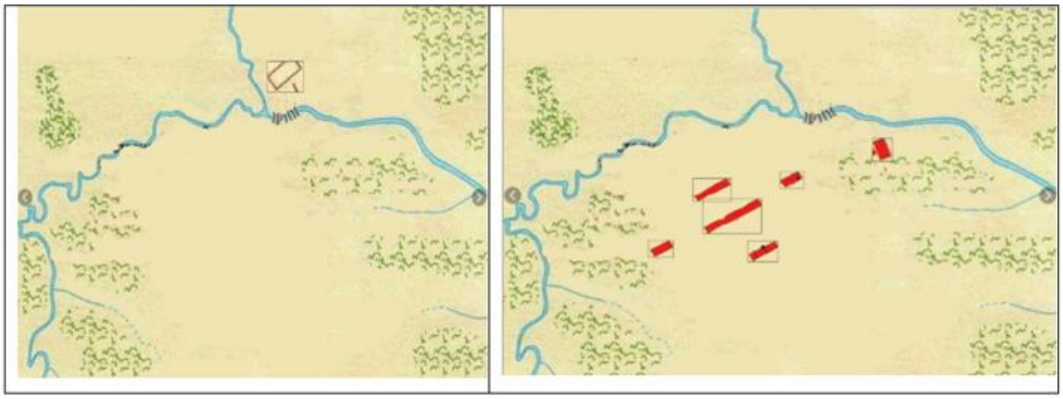


Рис. 14. Фрагмент раскадровки статичного изображения для целей 2D-анимации

VIII. Заключение

В работе описан и обоснован онтологически управляемый подход к разработке инструментального средства OntoRecon визуализации динамики событий в условиях неполноты картографического материала методами 2D-анимации, ориентированный на широкий круг пользователей, включая неспециалистов в области ИТ. В ходе работы был введен новый термин – онтологическая реконструкция событий, подразумевающий использование онтологически управляемых решений для ВРИС. Важно подчеркнуть, что наш подход применим не только для реконструкции исторических событий, но для виртуальной реконструкции средствами 2d-анимации широкого круг других категорий событий, которые условно можно свести к историческим, учитывая те определения понятий «событие» и «ситуация», которые даны в рамках данной работы. Это могут быть визуальные реконструкции индивидуальных и групповых туристических маршрутов, визуализация хода выполнения бизнес-процессов в проектной команде (здесь в качестве «исторических личностей» выступают участники проекта, а пространственные характеристики связаны с характеристиками конкретных проектов) и ряд других.

Разработанные средства могут быть использованы также для когнитивного онтологически управляемого способа решения недавно появившихся проблем, связанных с пространственными и временными «нестыковками» в раскадровке видеопотока и многократного повтора отдельных кадров в отечественных средствах построения и визуализации маршрутов продвижения автотранспорта, входящих в состав сервисов навигации. Но описание этих возможностей выходит за рамки данной работы и является перспективой дальнейших исследований.

В ближайшей перспективе планируется использовать разработанный инструментарий для визуализации хода работ по построению коллективом прикладных лингвистов так называемых «наивных карт» [11], которые являются разновидностью ментальных карты и отражают индивидуальные представления людей и общества о различных типах пространства, а также связанные с ними знания, представления, стереотипы, ценности, и способы представления. Исследование затрагивает графические, языковые, культурные и социально-демографические аспекты. «Наивные» карты являются субъективными визуализациями индивидуальных представлений о пространстве, которые в свою очередь служат основой для научного обобщения ментальных представлений о геопространстве и геоконцепциях.

Список литературы

1. Chuprina S.I. Using Data Fabric Architecture to Create Personalized Visual Analytics Systems in the Field of Digital Medicine // Scientific Visualization, 2023. Vol. 15 (5). P. 50–63. DOI: 10.26583/sv.15.5.05.

2. Быстрицкий Н.И. Историческая визуализация // Информационный бюллетень Ассоциации "История и компьютер". Т. 44. БФУ им. И. Канта. Калининград, 2015. С. 78–81.

3. Бородкин Л.И. Моделирование исторических процессов:от реконструкции реальности к анализу альтернатив. СПб.: Алетейя, 2016. 304 с.

4. Козеллек, Р. Словарь основных исторических понятий: Избранные статьи в 2-х т. Т. 1 / Пер. с немецкого К. Левинсон; сост. Ю. Зарецкий, К. Левинсон, И. Ширле; научн. ред. перевода Ю. Арнаутова. — М.: Новое литературное обозрение, 2014. 736 с.

5. Кудряшов, К. О. Историческое событие как один из основных механизмов историописания // Молодой ученый, 2020. № 18 (308). С. 367—369.

6. Chumakov R., Ryabinin K., Belousov K., Duan, J. Creative Map Studio: A Platform for Visual Analytics of Mental Maps //Scientific Visualization, 2021. Vol. 13 (2). P. 79–93.

7. Рядинская С.В., Мигачев В.А. Падежная грамматика Ч. Филлмора и современное понимание ролевой семантики глагола [Электронный ресурс]. URL: https://core.ac.uk/download/pdf/151220395.pdfl (дата обращения 29.07.2024).

8. Курсанина Е. Е. К вопросу о падежной грамматике // Проблемы романо-германской филологии, педагогики и методики преподавания иностранных языков, 2007. №5. URL: https://cyberleninka.ru/article/n/k-voprosu-o-padezhnoy-grammatike (дата обращения: 29.07.2024).

9. Чуприна, С.И., Зиненко Д.В. ОНТОЛИС: адаптируемый визуальный редактор онтологий // Вестник Пермского университета. Математика. Механика. Информатика, 2013. № 3(22). С. 106–110.

10. Chuprina S., Nasraoui O. Using Ontology-Based Adaptable Scientific Visualization and Cognitive Graphics Tools to Transform Traditional Information Systems into Intelligent Systems // Scientific visualization, 2016. Vol. 8 (1). P. 23-44.

11. Белоусов, К. И. Лингвосемиотическое моделирование обыденной географической картины мира / К. И. Белоусов, Н. Л. Зелянская // Вопросы когнитивной лингвистики, 2013. № 2. С. 73–85.

УДК 004.89
DOI: 10.25206/978-5-8149-3873-2-2024-389-402

Применение методов визуального анализа данных для выявления потребности в семантической интеграции данных
Using visual data analysis methods to identify the need for semantic data integration

С. И. Чуприна, К. В. Гимашева [1]
[1]*ФГАОУ ВО «Пермский государственный национальный исследовательский университет», ул. Букирева, 15, г. Пермь, 614068, Россия*

S. I. Chuprina, K. V. Gimasheva [1]
[1]*Perm State University, Perm, Bukireva Str. 15, 614068, Russia*

Аннотация. В работе предлагается подход, основанный на применении методов визуального анализа для построения и совершенствования моделей онтологически управляемых средств виртуальной семантической интеграции данных в рамках применения технологии фабрик данных. Это в конечном счете способствует повышению уровня пертинентности поисковой выдачи в случае комплексных запросов, предполагающих наличие мультипредметных информационных ресурсов. Сервисы разрабатываемого инструментального окружения NuCoBoShell позволяют выявить, имеются ли среди доступных ресурсов мультипредметные информационные ресурсы, удовлетворяющие информационную потребность пользователя, или же требуется семантическая интеграция данных на принципах виртуальной интеграции. Необходимость интеграции данных на семантическом уровне связана с распределенностью и разнородностью информационных ресурсов. В нашем подходе унифицированным образом извлекаются и семантически интегрируются необходимые факты и знания из информационных ресурсов различных текстовых форматов, находящихся как в открытом доступе, так и на локальном компьютере пользователя. Используя онтологические модели, методы семантического веба и средства визуального анализа, удается объединить разнородные данные в единое виртуальное интегрированное представление сразу о нескольких предметных областях безотносительно местоположения исходных ресурсов. Для автоматизации построения онтологий используются инструменты генеративного искусственного интеллекта.

***Ключевые слова*:** семантическая интеграция данных, семантический веб, фабрика данных, онтология, онтологический инжиниринг, визуальный анализ, визуальный редактор онтологий ОНТОЛИС.

Аннотация. The paper proposes an approach based on the use of visual analysis methods to build and improve models of ontology-driven tools of virtual semantic data integration within the framework of Data Fabric technology. This ultimately contributes to an increase in the level of persistence of search results in the case of complex queries involving the presence of multi-subject information resources. The services of the NuCoBoShell tool environment under development allow us to identify whether there are multi-subject information resources among the available resources that satisfy the user's information need, or whether semantic data integration is required and based on the principles of their virtual integration. The data integration at the semantic level is needed because the distribution and heterogeneity of information resources. In our approach, the necessary facts and knowledge are extracted and semantically integrated from of various textual information resources in a unified way. These resources may be not only public but also private accessible within user's local computer. Using ontological models, semantic web methods and visual analysis tools, it is possible to combine heterogeneous data into a single virtual integrated representation of several subject areas at once, regardless of the location of the resources. Generative artificial intelligence tools are used to automate the construction of the ontologies.

***Ключевые слова*:** semantic data integration, Semantic Web, Data Fabric, ontology, ontology engineering, visual analytics, visual ontology editor ONTOLIS.

I. Введение

Развитие цифровых информационных технологий приводит к появлению новых источников данных и способов их сбора, загрузки, хранения и передачи, что в свою очередь способствует непрерывному росту объема разнородной информации. Необходимость эффективно обрабатывать данные любой природы приводит к тому, что возникает потребность в разработке новых методов и средств для выполнения качественной автоматизированной и унифицированной обработки данных. Здесь на помощь приходят методы и средства генеративного ИИ. Однако, несмотря на увеличение объема обрабатываемых данных, решение аналитических

задач на основе сравнительно небольших объемов исходных данных, которые зачастую не находятся в открытом доступе и хранятся в локальных хранилищах или на локальных компьютерах отдельных пользователей или небольших исследовательских групп, по-прежнему остается актуальным направлением развития информационных технологий.

Обработка данных включает в себя широкий спектр процессов, направленных на преобразование и анализ данных с целью извлечения фактов и знаний из разнородных источников данных, их интеграции, интерпретации, генерации ответов на ЕЯ-запросы, поддержки в принятии решений. Для качественной обработки текстовых данных необходимо привлечение методов автоматической обработки текстов (АОТ; англ., NLP – Natural Language Processing) для предобработки, учета синонимов и смыслового контекста распределенных источников данных, а также решения целого ряда других задач. Среди таких задач выявление потребности в смысловой интеграции данных из разнородных источников и собственно автоматизация такой интеграции занимают особое место [1]. Основная цель интеграции данных состоит в том, чтобы в автоматизированном режиме объединить информацию из различных источников с учетом ее смыслового контекста, что позволяет пользователю получать более полные ответы на поисковые запросы без необходимости учета реального местоположения источников данных, их внутренней структуры и формата представления.

Это особенно актуально в случаях, когда запросы носят комплексный характер (будем называть их «комплексными»), т.е. когда они по своему содержанию предполагают обращение к ресурсам сразу из нескольких предметных областей, которые лишь в относительно небольшой степени пересекаются с точки зрения используемой терминологии. В таких случаях отсутствуют или недоступны отдельные ресурсы, содержащие сразу ответ на все смысловые части запроса с учетом определенного контекста их взаимосвязей, и требуется их предварительная семантическая интеграция, ведь традиционные поисковые механизмы в этих случаях ранжируют поисковую выдачу по отдельности для каждой части запроса. Примером такой ситуации могут служить запросы типа найти «рецепты с содержанием витамина B12, кальция и магния»: пока в открытом доступе нет сайта с ответами на такой комплексный вопрос, но при этом имеет место большое число отдельных ресурсов как про рецепты с указанием состава продуктов, так и про нутриенты в составе различных продуктов.

Семантический анализ при обработке естественного языка выходит за рамки поверхностного анализа слов и фокусируется на обнаружении и извлечении семантических связей между заданными концептами, которые позволяют более углубленно понять смысл текста [2]. В настоящее время значительное внимание в области обработки естественно-языковых текстов занимают большие языковые модели (англ., LLM – Large Language Model), построенные на основе методов глубинного машинного обучения (англ., ML – Machine Learning), в том числе генеративного ИИ. С помощью предобученной языковой модели можно генерировать связный осмысленный текст как на естественном, так и на искусственном языках [3]. Однако при обработке локальных файлов пользователя или закрытых источников корпоративной информации в некоторых узких предметных областях, содержащих зачастую специфическую информацию, использование открытых общедоступных сервисов для интегрированного анализа данных не всегда оправдано и не гарантируют защищенности данных. Кроме того, остаются открытыми вопросы автоматизированной валидации полноты и качества текстов, сгенерированных генеративным ИИ.

Использование больших языковых моделей в случае, когда загрузка и обработка данных выполняются на локальном компьютере пользователя, требует значительных вычислительных ресурсов. И это касается не только высоких требований относительно объема памяти. Важно учитывать, что точность сгенерированных системой ответов напрямую зависит от доступного контекста запроса. Поэтому требуется привлечение дополнительных ресурсов, в т.ч. квалифицированных специалистов, для дообучения модели. На данный момент перспективным направлением исследования для решения различных задач, в том числе задачи семантической интеграции данных, является интеграция подходов на основе методов машинного обучения и инженерии знаний, в частности, онтологического инжиниринга, который занимает первостепенное место в современных технологиях семантического веба (англ., Semantic Web).

Сложность обработки неструктурированных разнородных источников и создания такой структуры данных, которая была бы пригодна для машинной интерпретации, привела к появлению W3C, разрабатывающей единые принципы и стандарты для представления данных в интернете. Создание семантического веба [4, 5] как расширения World Wide Web, способствовало развитию систем виртуальной интеграции данных. В основе стандартов семантического веба лежит стандарт RDF, который задает способ представления моделей и описания ресурсов, язык OWL для описания конструкций классов, их свойств и связей, а также SPARQL – стандартизованный язык запросов, с помощью которого можно в унифицированном виде выполнять поиск и интеграцию данных, хранящихся в RDF-графах. Основная идея семантического веба заключается в том, что содержимое информационных ресурсов должно быть выражено не только на естественном языке, но и на формальном языке, позволяя тем самым автоматизировать и унифицировать процессы обработки и интеграции

данных из разнородных источников, повысить семантическую мощность поисковых механизмов. В системах интеграции, основанных на технологиях семантического веба, можно интегрировать большое количество распределенных разнородных ресурсов посредством подзапросов к RDF-тройкам, построенным в результате автоматического парсинга этих ресурсов. Базовым строительным блоком модели данных RDF является утверждение, представляющее собой тройку: ресурс, именованное свойство и его значение (в терминах RDF: субъект, предикат и объект). Под свойством следует понимать некий аспект, характеристику, атрибут или отношение, используемое для описания ресурса.

Технологии семантического веба ориентированы на обработку преимущественно веб-ресурсов в сети интернет. К отличительной особенности данной работы можно отнести то, что решается задача семантической интеграции неструктурированных текстовых данных, представленных не только в веб-форматах, но и в любых текстовых форматах, причем текстовые коллекции документов могут храниться не только в общем доступе, но и локально. Для решения этой задачи нами используются методы семантического веба и технологии фабрик данных (англ., Data Fabric) [6, 7], предназначенные для виртуальной интеграции и предполагающие наличие между пользовательским интерфейсом поисковых запросов и источниками данных промежуточного слоя в виде графа знаний – онтологии, хранящей необходимые метаданные, факты и знания в терминах соответствующей предметной области.

Для эффективной и качественной реализации нашего подхода, помимо чисто технических вопросов эффективной реализации программных сервисов, требуется не только адекватная модель онтологии и качественное автоматизированное построение самих онтологий, но и наличие сервиса, который еще до выполнения виртуальной семантической интеграции данных в автоматизированном режиме определял бы необходимость в указанной предварительной интеграции информационных ресурсов. И здесь необходимость применения методов визуального анализа данных трудно переоценить.

Главное внимание в данной работе уделяется методам и средствам визуальной аналитики для выявления потребности в семантической интеграции данных из распределенных разнородных ресурсов с целью повышения уровня пертинентности поисковой выдачи (т.е. уровня ее соответствия не только тексту запроса, но и информационной потребности пользователя) и эффективности реализации самих поисковых механизмов за счет исключения из процесса поиска этапа семантической интеграции данных в случаях, когда доступны отдельные ресурсы, в достаточно полной мере удовлетворяющие поисковому запросу (т.е. когда применение традиционных поисковых механизмов без средств семантической интеграции вполне достаточно). Указанные средства визуальной аналитики в то же время наглядно демонстрируют потребность и обоснованность применения средств семантической интеграции, т.е. играют роль своего рода объяснительной компоненты для обоснования необходимости временных и ресурсных затрат на интеграцию данных. Кроме того, они являются инструментом разработчиков для валидации качества построенных онтологий и визуальным аналитическим средством для автоматизации исследовательской работы по выявлению и совершенствованию метрик семантической близости онтологий в задачах предварительной оценки неполноты поисковой выдачи на основе отдельных ресурсов (последнее также способствует повышению качества ранжирования поисковой выдачи).

Конечной целью нашего исследования является полная автоматизация процесса выявления потребности в семантической интеграции данных. Описание текущих результатов исследований адекватности различных метрик оценки неполноты поисковой выдачи в полностью автоматическом режиме без средств визуального анализа остается вне рамок данной работы. Основной акцент делается на описании средств автоматизации этой исследовательской работы, а также валидации автоматически построенных онтологий.

II. Обоснование применения технологии фабрик данных для семантической интеграции данных

Задача семантической интеграции данных является нетривиальной и ресурсоемкой [8]. Сложность и характер используемых методов и средств интеграции существенным образом зависят от специфики отдельных источников данных и всего множества источников в целом, а также от специфики самих способов интеграции. На начальном этапе выполняется автоматическая предобработка текстов для преобразования их в стандартный вид, в т.ч. их нормализация, когда слова из текста приводятся к их словарной форме: к именительному падежу и к единственному числу (если есть) для имен существительных, к инфинитиву для глаголов и т. п. Именно в таком виде термины, представляющие факты и знания о соответствующей предметной области, хранятся в промежуточном слое графа знаний в виде онтологии.

На практике для выполнения семантической интеграции данных зачастую используется традиционный подход к консолидации данных. При консолидации создается единое хранилище, в котором выполняется интеграция различных источников данных путем извлечения, преобразования и загрузки данных специальными средствами (англ., ETL – Extract, Transform, Load), что приводит к проблеме «свежести» данных. На текущий момент наиболее перспективные методы и средства интеграции данных базируются на принципах

федерирования (виртуализации) данных. За счет того, что создается виртуальное хранилище, которое не предполагает копирование и загрузку исходных данных в единое физическое хранилище, появляется возможность оперативно учитывать изменения в источниках данных и настраивать ограничения прав доступа как к отдельным распределенным разнородным ресурсам, так и ко всему виртуальному пространству данных. Виртуальная интеграция позволяет решить проблему вариативности данных, а извлечение данных из различных источников, их предобработка и смысловая интеграция выполняются в режиме реального времени в ходе выполнения запроса пользователя. Результат запроса генерируется унифицированным образом для множества распределенных разнородных источников вне зависимости от их местоположения и формата представления.

На рис. 1 представлена обобщенная архитектура системы виртуальной интеграции данных, где унифицированным образом обрабатываются как структурированные, полуструктурированные так и неструктурированные данные. Центральный элемент системы интеграции – посредник – выполняет преобразование запросов, которые автоматически трансформируются в подзапросы к отдельным информационным ресурсам (источникам данных). Выполнение подзапросов реализуется специальными модулями (адаптерами), которые позволяют оптимизировать взаимодействие посредника и источника данных, не требуя при этом физического копирования данных.

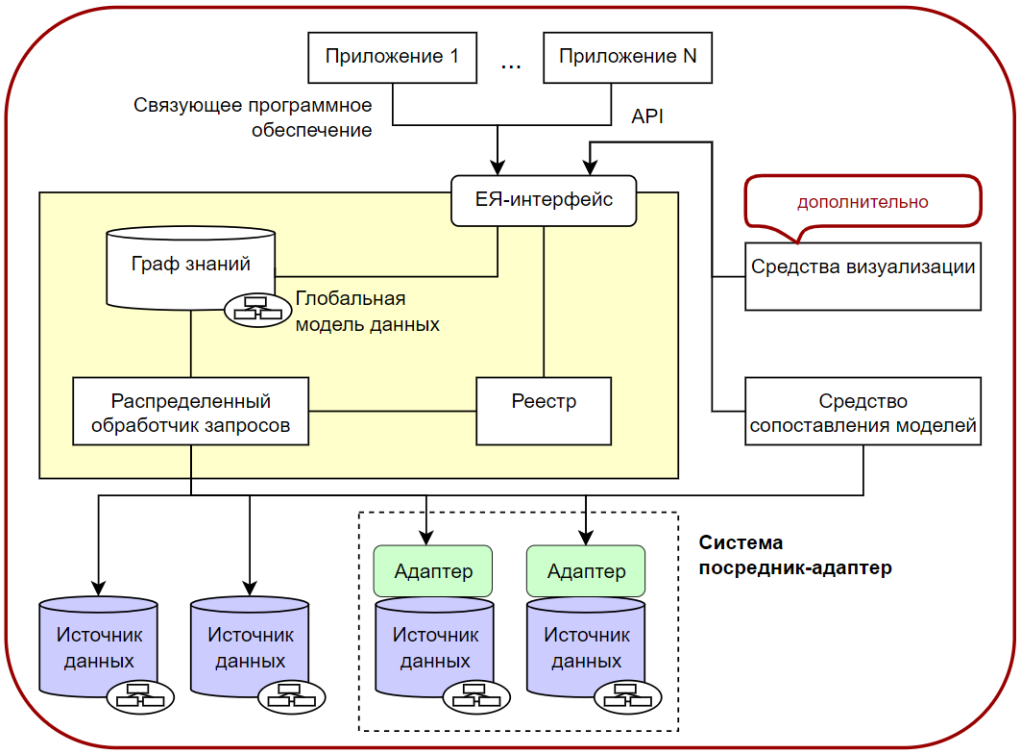


Рис. 1. Обобщенная архитектура системы виртуальной интеграции данных

В последнее время наибольший интерес в области технологий виртуальной интеграции данных получили технологии построения интеллектуальных фабрик данных – нового поколения хранилищ данных [6, 7]. Фабрика данных представляет собой интегрированное виртуальное хранилище, архитектура и сервисы которого обеспечивают смысловую связь между данными из разнородных источников вне зависимости от формата их хранения и технологии создания систем [9].

В основе построения фабрики данных заложено использование промежуточного онтологического слоя, который хранит как метаданные о предметном содержании распределенных разнородных источников данных, их структуре и местоположении, так и сведения о различных программных сервисах, которые используются для обработки данных независимо от их местоположения и формата. Промежуточный слой онтологий выступает в качестве графа знаний (англ., Knowledge Graph) и позволяет настроить интерфейс конечного пользователя на интегрированную обработку разнородных источников данных без их физического перемещения. При этом не предъявляется повышенных требований к уровню квалификации конечного пользователя за счет того, что обращение к данным осуществляется в едином формате через поисковую строку по аналогии с поисковыми сервисами интернет, а реальное местоположение и структура данных скрыты. Так как онтологический слой знаний хранит метаданные в том числе и об используемых методах и средствах, то становится возможным унифицированным образом обращаться и к программным ресурсам – сервисам фабрики данных.

В нашем проекте в качестве источников данных выступают только полуструктурированные и неструктурированные данные, представляющие коллекции текстовых документов в различных форматах и веб-ресурсы.

III. Концептуальные и технологические аспекты реализации семантической интеграции данных

Предлагаемый нами онтологически управляемый подход к семантической интеграции данных позволяет повысить качество данных за счет установления смысловых связей между понятиями из разных предметных областей. В результате появляется возможность выполнять запросы на естественном языке к разнородному пространству данных, скрывая от пользователя не только их формат, структуру и местоположение, но и обогащать их предметное содержание за счет генерации новых виртуальных документов, которые, с одной стороны, интегрируют нужный контент из разных источников, а, с другой стороны, выполняют его семантическую фильтрацию в соответствии с контекстом запроса.

Важно отметить, что при нашем подходе такая семантическая агрегация текстовых данных выполняется унифицированным образом вне зависимости от того, расположены информационные ресурсы в сети интернет или на локальном компьютере, причем тексты локальных документов не передаются на вход сервисам интернет, как это обычно имеет место при использовании LLM в случаях, когда отсутствует возможность их локального хранения на компьютере пользователя. Последнее позволяет автоматизировать процесс аналитики данных в контексте персонального виртуального пространства пользователя, учитывать его личные предпочтения и доверие к различным источникам данных, агрегируя и сопоставляя данные из его личных локальных текстовых документов и общедоступных интернет-ресурсов или ресурсов корпоративного хранилища, обеспечивая при этом конфиденциальность данных из локальных ресурсов.

Предлагаемая концепция ориентирована на разработку инструментального окружения, позволяющего автоматизировать и унифицировать семантическую интеграцию распределенных разнородных данных, посредством построения виртуального хранилища данных. В качестве технологий для реализации виртуальной интеграции данных были выбраны коротко описанные выше технологии построения интеллектуальных фабрик данных, основанные на установлении между пользовательским интерфейсом и различными распределенными источниками данных промежуточного онтологического слоя, и технологиях семантического веба. На рис. 2 представлена упрощенная схема работы инструментального окружения NuCoBoShell, ориентированного на смысловую интеграцию данных из распределенных разнородных ресурсов и позволяющего адаптироваться к новым источникам данных и/или к новым предметным областям за счет пополнения общего репозитория онтологий, управляющих функционированием всей системы.

Важным аспектом в разработке онтологий является соблюдение концептуального баланса, то есть представление знаний о предметной области в однозначно интерпретируемом и логичном виде с учетом семантической близости между двумя и более понятиями предметной области. Онтология должна быть прозрачна для чтения и не перегружена избыточной информацией. В рамках предлагаемой концепции достаточно использования так называемых «легковесных» (англ., lightweight) онтологий, которые состоят из описания понятий предметной области с указанием их свойств, а также взаимосвязей с другими понятиями, но без аксиоматики.

Создание легковесных онтологий требует меньше времени и усилий, чем создание тяжеловесных онтологий, что делает их построение более доступными для широкого круга разработчиков. Кроме того, как показала практика, такие онтологии достаточно качественно генерируются с использованием сервисов генеративного ИИ, что обеспечивает автоматизацию их построения и упрощает подключение к системе виртуальной интеграции новых ресурсов. Использование сервисов генеративного ИИ, в частности GPT-4o, не нарушает концептуального единства предлагаемого подхода, так как эти сервисы используются только для автоматизации построения онтологий и только на базе общедоступных ресурсов интернет. Кроме того, эти сервисы не подключаются к инструментальной среде NuCoBoShell, а используются отдельно для пополнения ее онтологической базы знаний. В качестве среды для визуального построения и анализа онтологий предлагается использовать визуальный редактор онтологий ОНТОЛИС [10], который позволяет строить легковесные онтологии и конвертировать их в формат стандарта OWL. Кроме того, в этом редакторе поддерживаются атрибутивные свойства вершин и дуг графа онтологии, что позволяет унифицированным образом расширять возможности по интерпретации онтологий, их визуальному анализу и организации логического вывода.

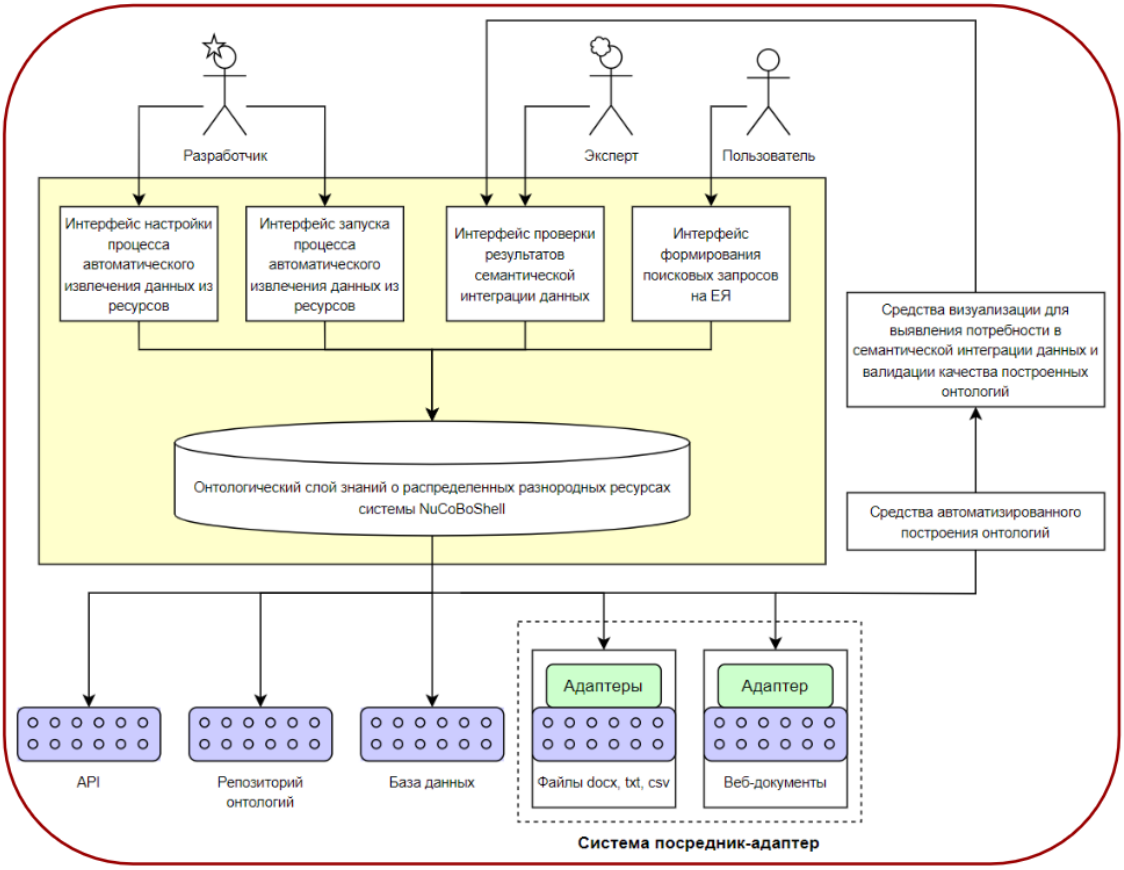


Рис. 2. Схема работы NuCoBoShell

Для исследования вопросов эффективности решения задачи семантической интеграции данных и автоматизации процесса определения необходимости в такой интеграции в составе инструментального окружения NuCoBoShell разработаны соответствующие сервисы визуальной аналитики, позволяющие на основе анализа поискового запроса пользователя и доступных ресурсов автоматизировать исследовательскую работу и выполнять валидацию полученных результатов. На рис. 3 схематично представлен порядок применения таких сервисов, где $R_1, \ldots, R_N$ – доступные/выбранные для анализа ресурсы, $ONT_1, \ldots, ONT_N$ – соответствующие им онтологии в контексте семантики запроса. Ресурсы по своей тематике в нашем случае (учитывается специфика комплексных запросов) соответствуют, минимум, двум слабопересекающимся предметным областям $ПрО_1, \ldots, ПрО_M$, $M \geq 2$. Соответственно, и $N \geq 2$, но это – крайний случай: обычно количество анализируемых ресурсов значительно превышает число предметных областей.

Построенные на разных этапах визуального анализа онтологии и метаданные о них, включающие в том числе сведения об исходных ресурсах и графическое представление соответствующих ЕЯ-запросов, сохраняются в смарт-репозитории системы и индексируются для последующего переиспользования. Сам результат семантической интеграции N текстовых ресурсов из M предметных областей остается виртуальным. На рис. 3 звездочками отмечены те сервисы, которые в случае повторного обращения к одним и тем же ресурсам одних и тех же предметных областей (что определяется системой автоматически на основе указанных метаданных) повторно не вызываются: происходит прямое обращение к результатам соответствующих этапов. Это повышает эффективность анализа. Соответствующие сервисы используются не только для автоматизации труда разработчиков-исследователей, но и частично входят в средства пользовательской подсистемы поиска ответов на комплексные вопросы, где семантическая интеграция выполняется уже не в автоматизированном, а в полностью автоматического режиме. Эти же сервисы можно использовать как объяснительную компоненту для дальнейшего совершенствования и оптимизации процесса семантической интеграции, а также как средство валидации построенных онтологий. Более подробно результаты визуального анализа описаны в следующем разделе.

На рис. 4 схематично представлен процесс семантической интеграции данных. Описание правил отображения является предметом отдельной статьи. Термином «эталонная модель» предметной области здесь обозначается онтология отдельной предметной области, построенная экспертом/инженером по знаниям, либо непосредственно разработчиком с учетом специфики предметного содержания этой области безотносительно конкретных

задаваемых запросов в среде визуального редактора онтологий ОНТОЛИС, либо автоматически сгенерированная на базе средств генеративного ИИ, но прошедшая валидацию при помощи средств визуального анализа данных.

Рис. 3. Порядок применения сервисов визуального анализа данных в ходе виртуальной семантической интеграции данных

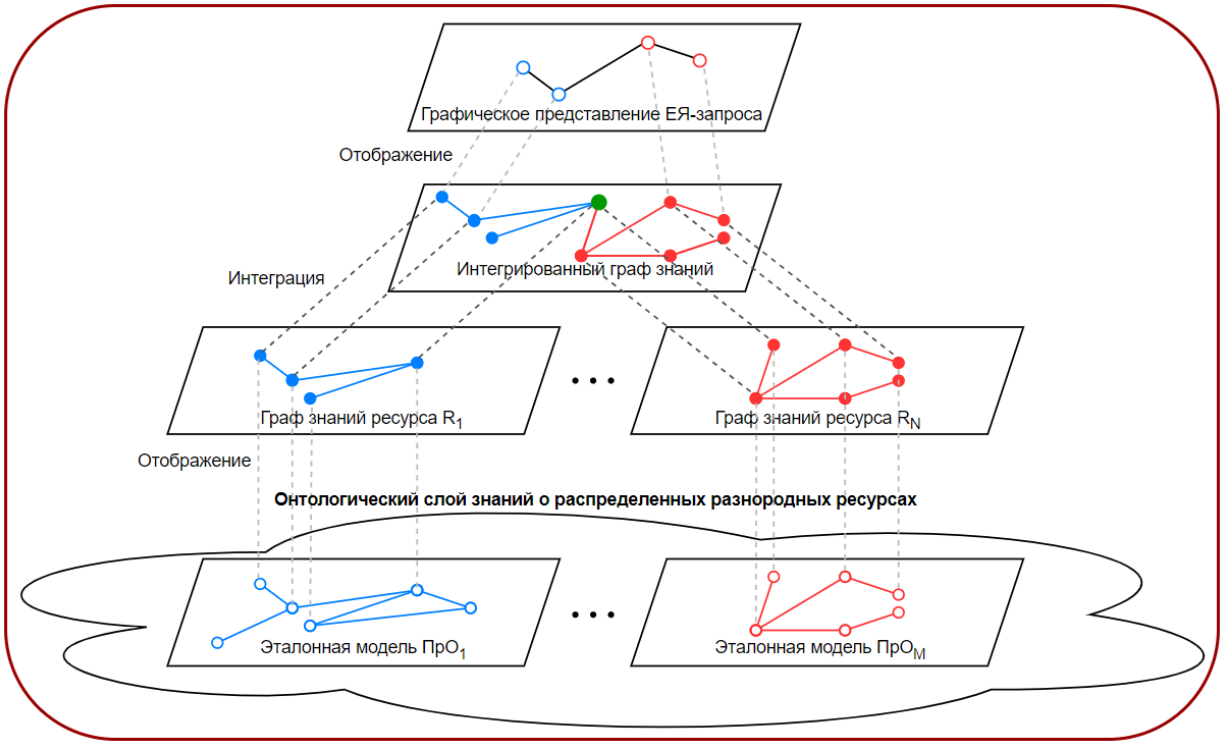


Рис. 4. Процесс семантической виртуальной интеграции данных

III. Автоматизация построения онтологий методами генеративного ИИ

Для автоматизации построения онтологий в виде RDF-триплетов в рамках данного проекта было исследовано применение методов генеративного ИИ с использованием универсального решения – GPT-4o – многоязычного мультимодального чат-бота на базе LLM, способного принимать в качестве входных данных любую комбинацию текста, звука и изображения и генерировать ЕЯ-ответы в различных форматах. Так как предварительно обученные большие модели не могли не учитывать существующие ресурсы семантического веба, то предполагалось, что задача автоматического построения онтологий ресурсов в виде RDF-триплетов на основе таких моделей с высокой степенью уверенности гарантирует качественное решение на базе таких моделей. Кроме того, мы исследовали возможности GPT-4o также и для решения основной задачи нашего исследования – получения пертинентных ответов на комплексные запросы.

В качестве предметной области были выбраны ресурсы с описанием продуктов и их нутриентным составом *($ПрО_1$)*, а также ресурсы с описанием рецептов блюд, содержащих список ингредиентов и способ их приготовления *($ПрО_2$)*. Выбор предметной области обусловлен тем, что на данный момент нет открытых общедоступных ресурсов, которые содержали бы сведения одновременно и о продуктах с указанным составом их нутриентов, и о рецептах с данными продуктами в качестве ингредиентов. Как указывалось во введении, в ответ на комплексные поисковые запросы типа «рецепты с продуктами, содержащими витамин B12 и кальций» традиционные поисковые механизмы интернет выдают отранжированный список ссылок, где на первых позициях представлены ссылки на ресурсы с данными о продуктах и их нутриентах, а после них – ссылки на сайты с рецептами, причем в составе их ингредиентов могут вообще не содержаться указанные нутриенты – «витамин B12» и «кальций».

Так как семантическая интеграция данных должна унифицированным образом выполняться для ресурсов, представленных в различных форматах, в качестве исходных данных для автоматического построения онтологий были выбраны как коллекции рецептов в текстовом формате, так и в формате JSON.

Продемонстрируем результат автоматического построения чат-ботом RDF-триплетов на нескольких примерах рецептов из источников, представленных как в традиционных текстовых форматах, так и в формате JSON. Без дополнительных уточняющих запросов и установок чат-бот разбивает текст рецептов на тройки, правильно выделяя состав продуктов, но не выявив при этом нужную последовательность действий и не увязав их с количеством отдельных ингредиентов в рецепте (в одном и том же рецепте один и тот же ингредиент может использоваться по-разному и в разных количествах для разных целей). На рис. 5 представлен фрагмент RDF-триплетов, полученных в ходе автоматической генерации ответа с помощью GPT-4o. Поэтому на основе только таких результатов построения RDF-триплетов выполнить качественную семантическую интеграцию данных невозможно и требуются дополнительные действия со стороны пользователя, в частности, ряд уточняющих запросов.

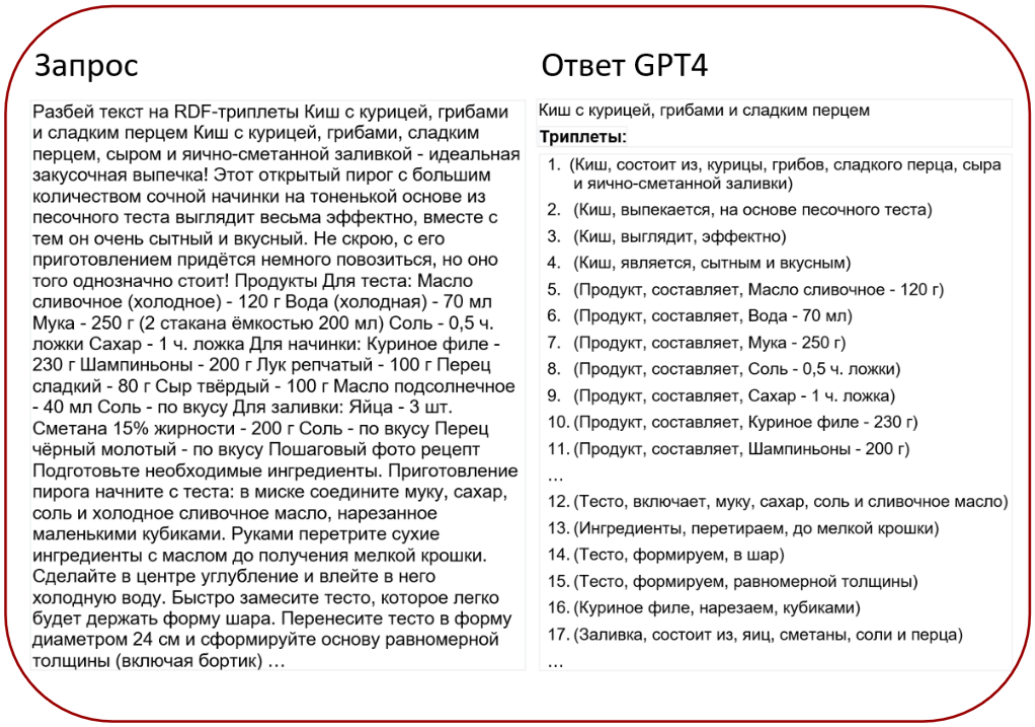


Рис. 5. Фрагмент RDF-триплетов, построенных непосредственно из текста рецепта в формате docx

Приведем пример разбиения на RDF-триплеты описания рецептов, представленных в формате JSON (см. рис. 6). В данном примере разбиение текста в формате JSON на RDF-триплеты ожидаемо выполнено более качественно по сравнению с разбиением произвольного неразмеченного текста рецепта (см. рис. 7), потому что JSON – стандартный текстовый формат для хранения структурированных данных и обмена ими. Однако такой вариант представления исходных текстовых данных также требует дополнительной обработки, так как связь конкретного ингредиента с его количеством в контексте выполнения разных действий с одним и тем же ингредиентом в процессе приготовления блюда по конкретному рецепту по-прежнему явно не выявлена. Из фрагмента ответа GPT-4o, приведенного на рис. 7, видно, что даже для простого определения количества конкретного ингредиента в рецепте нужно дополнительно учитывать относительный порядок триплетов, т.к. в RDF-тройках, описывающих ингредиенты рецептов, в качестве ресурса для свойства «количество» *(<amount>)* и соответствующих ему значений указывается URI рецепта, а не ингредиента.


```
{
    "recipes": [{
            "name": "Брауни (brownie)",
            "description": "Шоколад разломать на кусочки и вместе со сливочным маслом растопить на водяной бане, не переставая все время помешивать лопаткой или деревянной ложкой...",
            "category": "Выпечка и десерты",
            "image": "https://eda.ru/images/RecipePhoto/4x3/brauni-brownie_20955_photo_29166.jpg",
            "url": "https://eda.ru/recepty/vypechka-deserty/brauni-brownie-20955"
        }
    ],
    "ingredients": [{
            "name": "Темный шоколад"
        }, {
            "name": "Сливочное масло"
        }
    ],
    "recipe_ingredients": [{
            "recipe_name": "Брауни (brownie)",
            "ingredient_name": "Темный шоколад",
            "amount": "100 г"
        }, {
            "recipe_name": "Брауни (brownie)",
            "ingredient_name": "Сливочное масло",
            "amount": "180 г"
        }
    ]
}
```


Рис. 6. Фрагмент описания рецептов в формате JSON

Запрос

Разбей JSON на RDF триплеты { "recipes": [{ "name": "Брауни (brownie)", "description": "Шоколад разломать на кусочки и вместе со сливочным маслом растопить на водяной бане, не переставая все время помешивать лопаткой или деревянной ложкой…", "category": "Выпечка и десерты", "image": "https://eda.ru/images/RecipePhoto/4x3/brauni-brownie_20955_photo_29166.jpg", "url": "https://eda.ru/recepty/vypechka-deserty/brauni-brownie-20955" }], "ingredients": [{ "name": "Темный шоколад" }, { "name": "Сливочное масло" }, { "name": "Коричневый сахар" }, { "name": "Куриное яйцо" }, { "name": "Пшеничная мука" }, { "name": "Грецкие орехи" }], "recipe_ingredients": [{ "recipe_name": "Брауни (brownie)", "ingredient_name": "Темный шоколад", "amount": "100 г" }, { "recipe_name": "Брауни (brownie)", "ingredient_name": "Сливочное масло", "amount": "180 г" }, { "recipe_name": "Брауни (brownie)", "ingredient_name": "Коричневый сахар", "amount": "200 г" }, { "recipe_name": "Брауни (brownie)", "ingredient_name": "Куриное яйцо", "amount": "4 штуки" }, { "recipe_name": "Брауни (brownie)", "ingredient_name": "Пшеничная мука", "amount": "100 г" }, { "recipe_name": "Брауни (brownie)", "ingredient_name": "Грецкие орехи", "amount": "100 г" }] }

Ответ GPT4

1. **recipes**:
   - `<recipe_id_1> <rdf:type> <Recipe>`
   - `<recipe_id_1> <name> "Брауни (brownie)"`
   - `<recipe_id_1> <description> "Шоколад разломать на кусочки и вместе со сливочным маслом растопить на водяной бане..."`
   - `<recipe_id_1> <category> "Выпечка и десерты"`
   - `<recipe_id_1> <image> https://eda.ru/images/RecipePhoto/4x3/brauni-brownie_20955_photo_29166.jpg`
   - `<recipe_id_1> <url> "https://eda.ru/recepty/vypechka-deserty/brauni-brownie-20955"`
2. **ingredients**:
   - `<ingredient_id_1> <rdf:type> <Ingredient>`
   - `<ingredient_id_1> <name> "Темный шоколад"`
   - `<ingredient_id_2> <rdf:type> <Ingredient>`
   - `<ingredient_id_2> <name> "Сливочное масло"`
   - …
3. **recipe_ingredients**:
   - `<recipe_id_1> <hasIngredient> <ingredient_id_1>`
   - `<recipe_id_1> <ingredient_name> "Темный шоколад"`
   - `<recipe_id_1> <amount> "100 г"`
   - `<recipe_id_1> <hasIngredient> <ingredient_id_2>`
   - `<recipe_id_1> <ingredient_name> "Сливочное масло"`
   - `<recipe_id_1> <amount> "180 г"`
   - …

Рис. 7. Фрагмент RDF-триплетов, построенных по описанию рецептов в формате JSON

Так как подобные GPT-4o сервисы способны учитывать контекст диалога с пользователем и накапливать полученный в ходе дообучения опыт, то при повторном запросе на построение RDF-триплетов непосредственно из текста запроса в формате docx (см. рис. 5), будет автоматически учтен результат, полученный в ходе реализации запроса, представленного на рис. 7 (эти запросы выполняются как бы в обратном порядке), и получен более качественный результат. Однако если при использовании чат-бота дополнительно не уточнять запросы, а только поменять их последовательность выполнения, то проблема правильного установления взаимосвязи между конкретным ингредиентом и его количеством в конкретном рецепте так и не находит своего решения. Поэтому использовать GPT-4o для полностью автоматического решения задачи семантической интеграции данных для реализации комплексных запросов, по крайней мере, пока, не представляется возможным.

На основе проведенных исследований, можно сделать вывод, что использование чат-бота позволяет автоматизировать построение онтологий ресурсов и предметных областей, при этом исходные данные могут быть представлены в разном формате и/или иметь разную структуру, однако для решения каждой отдельной задачи требуется индивидуальная доработка запроса и дообучение системы, что требует высоких ресурсных затрат. Кроме того, важен контекст и порядок вопросов. Такая работа требует привлечения специалистов, причем чат-бот запоминает полученные результаты и делает их общедоступными при решении аналогичных задач другими пользователями. А если пользователь подает на вход свои личные персональные коллекции документов, то результат их обработки фактически тоже становится общедоступным.

Аналогичные проблемы возникают и в случае использования GPT-4o для генерации ответов на комплексные запросы. Поэтому мы в контексте решаемой нами задачи рекомендуем использовать этот чат-бот для автоматизации построения онтологий соответствующих предметных областей с последующей их валидацией и подключением к смарт-репозиторию системы, представляющему собой распределенную онтологическую базу знаний, причем выполнять этот этап предварительно в ходе пополнения репозитория системы, а не в динамике в ходе выполнения запросов. При решении персональных задач пользователь настраивает свой профиль на определенные предметные области, онтологии которых представлены в смарт-репозитории системы, что сокращает объем и количество используемых онтологий на уровне промежуточного слоя знаний системы. В перспективе эта персональная настройка на доступные ресурсы будет учитывать и степень доверия конкретного пользователя к тому или иному ресурсу.

IV. Выявление потребности в семантической интеграции данных средствами визуальной аналитики

Как уже отмечалось, интеграция онтологий играет значимую роль для решения целого ряда задач, связанных с обработкой распределенных разнородных источников и доступом к данным, содержащимся в этих источниках. В работах [1, 11] уделяется большое внимание этим вопросам. В данной статье мы акцентируем внимание на необходимости управляемой онтологиями смысловой интеграции разнородных ресурсов с целью получения более пертинентных ответов на комплексные запросы пользователей. В терминологии [11] такие онтологии и ресурсы называются «многоаспектными», в терминологии [1] – «мультипредметными».

В рамках современной парадигмы семантического доступа и интеграции данных, называемой доступом к данным на основе онтологий (англ. OBDA – Ontology-Based Data Access), онтология определяет высокоуровневую глобальную схему разнородных источников данных и предоставляет словарь для формирования пользовательских запросов, в том числе на естественном языке [12]. После получения запроса OBDA-подсистема выполняет его трансформацию в отдельные запросы к источникам данных, тем самым выполняя виртуальную семантическую интеграции данных без физического перемещения исходных данных в рамках подхода, описанного в разделах II и III

Текущий этап работы над проектом NuCoBoShell связан с совершенствованием используемых моделей онтологий и применением средств визуального анализа данных для автоматизированного выявления потребности в семантической интеграции данных, а также исследовании подходящих метрик семантической близости с тем, чтобы одновременно с высоким уровнем пертинентности поисковой выдачи повысить эффективность семантического поиска за счет обоснованного исключения этапа интеграции, если находятся отдельные ресурсы, содержащие ответы на комплексный запрос с высоким уровнем пертинентности. И здесь на помощь приходят средства визуального анализа. Кроме того, визуальные средства помогают оценить качество построения онтологий по таким характеристикам, как:

- полнота и точность словаря предметной области;
- адекватность структуры онтологии с точки зрения таксономии понятий и других, поддерживаемых моделью, отношений между понятиями;
- степень простоты восприятия содержимого онтологии.

Управляющие семантической интеграцией, анализом и визуализацией данных онтологии частично были построены вручную, а затем обогащены в результате автоматического построения онтологий на основе RDF-

триплетов, сгенерированных сервисами GPT-4o. Визуальный анализ, а также валидация построенных онтологий выполнялись в среде отечественного визуального редактора онтологий ОНТОЛИС [10], который предназначен для создания легковесных онтологий и способен автоматически интерпретировать атрибуты вершин (концептов предметной области) и дуг (связей между концептами) онтологии в ходе логического вывода. Преимуществом нашего редактора онтологий является то, что он не требует специальной подготовки и ориентирован на пользователя с любым уровнем ИТ-квалификации, что подтверждено практикой его использования при разработке большого числа разнообразных, управляемых онтологиями, приложений (см., например, [3, 9]).

На рис. 8 представлен фрагмент результата интеграции онтологии ресурсов, посвященных двум разнородным предметным областям: часть ресурсов содержит описание рецептов с указанием их ингредиентов и способа приготовления, другая часть описывает нутриентный состав продуктов безотносительно рецептов приготовления блюд, в состав которых входят эти продукты. Графическое представление онтологий указанных ресурсов наглядно демонстрирует необходимость виртуальной семантической интеграции соответствующих ресурсов для получения пертинентных ответов на комплексные вопросы о составе конкретных нутриентов в доступных ресурсах с рецептами приготовления блюд, т.к. имеет место относительно небольшое число пересечений понятий из разных онтологий (во фрагменте, представленном на рис. 8, – это понятие «Бекон»). Напоминаем, что интеграция ресурсов является виртуальной, без физического копирования данных в единое хранилище, а интеграция управляющих этим процессом онтологий – это реальный ETL-процесс, который сохраняет результаты интеграции онтологий в смарт-репозитории системы.

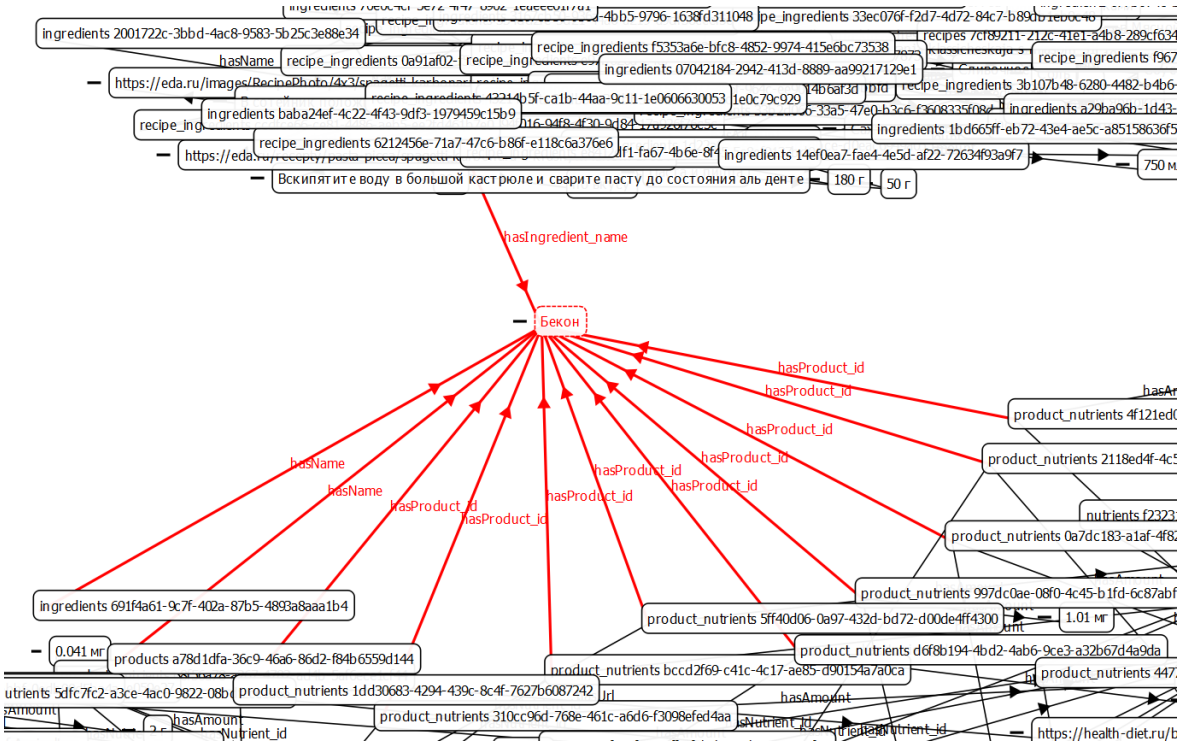


Рис. 8. Фрагмент результата интеграции онтологий

Результаты построения RDF-троек для отдельных ресурсов с рецептами и нутриентным составом продуктов были проанализированы средствами визуальной аналитики и использовались в ходе усовершенствования как онтологической модели представления графов знаний, описывающих предметное содержание конкретных ресурсов, так и эталонных онтологий соответствующих предметных областей (см. рис. 4).

Модель онтологии включает в себя такие категории ресурсов как «Рецепт», «Продукт», «Нутриент» и т.п. Модель поддерживает следующие типы связей:

- hasAmount – для представления свойства «Количество»;
- hasCategory – для представления свойства «Категория»;
- hasDescription – для представления свойства «Описание рецепта»;
- hasImage – для представления свойства «Изображение» (URI соответствующего ресурса с изображением);
- hasIngredient_name – для представления свойства «Название ингредиента»;
- hasName – для представления свойства «Название»;
- hasNutrient_id – для представления свойства «Название нутриента»;
- hasProduct_id – для представления свойства «Название продукта»;

– hasRecipe_name – для представления свойства «Название рецепта»;
– hasSource – для представления свойства «Источник» (URI соответствующего исходного ресурса);
– hasUrl – для представления свойства «URL».

На рис. 9 представлен фрагмент описания рецептов и входящих в их состав ингредиентов в виде автоматически сгенерированной онтологии на основе полученных RDF-триплетов. Рис. 10 демонстрирует онтологию нутриентного состава продуктов, автоматически построенную сервисами NuCoBoShell на основе одного из авторитетных веб-ресурсов, к которому в настоящее время нет открытого доступа (извлеченные факты и знания сохранены в смарт--репозитории системы и по мнению экспертов-нутрициологов являются достоверными и представляют ценный информационный ресурс). Как видно из демонстрируемых фрагментов онтологий, благодаря онтологически управляемому решению и модели онтологии, построенной в результате исследовательской работы с использованием средств визуального анализа данных, решена проблема явного и корректного представления взаимосвязей и свойств ингредиентов в составе рецептов, которая обсуждалась нами ранее в разделе III при анализе ответов, сгенерированных чат-ботом GPT-4o. Как часть онтологической базы знаний в составе описанной выше модели предметной онтологии была автоматически построена также эталонная таксономия продуктов и их нутриентов с использованием парадигматических связей типа «класс-подкласс» и «часть-целое».

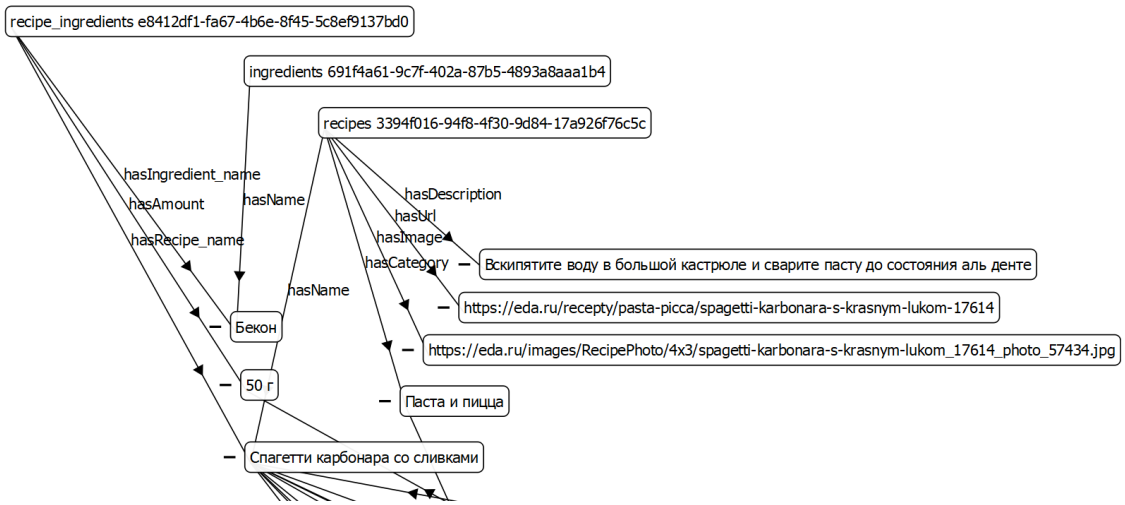


Рис. 9. Фрагмент автоматически построенной онтологии рецептов и ингредиентов (ресурс: https://eda.ru/)

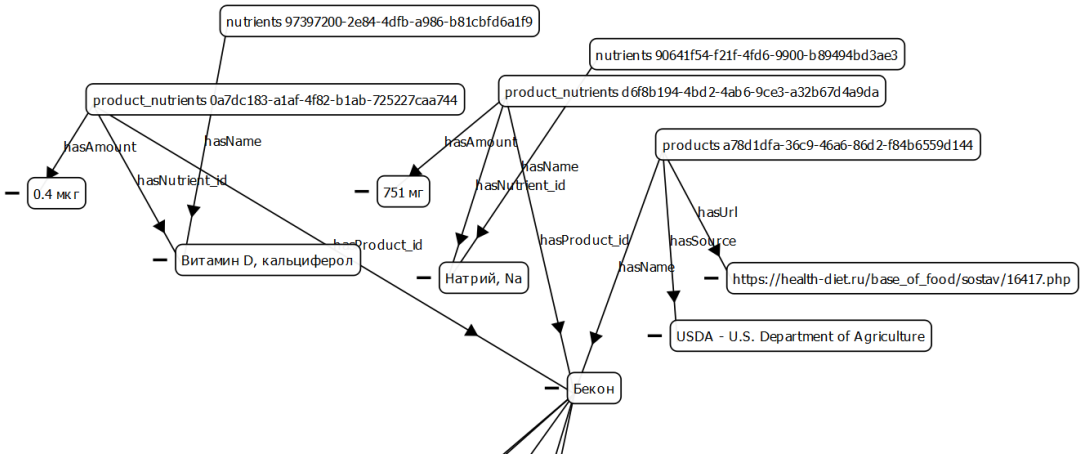


Рис. 10. Фрагмент автоматически построенной онтологии нутриентов продуктов

Средства визуального анализа позволяют выделить в графе онтологии все вершины, соответствующие продуктам, наименования которых автоматически извлечены из ресурсов с описанием нутриентного состава продуктов, и сопоставить их с выделенными вершинами, поименованными названиями продуктов, извлеченных из ресурсов с описаниями рецептов, где эти продукты выступают в качестве ингредиентов. Как уже было сказано ранее, полное пересечение по названиям продуктов в демонстрируемом примере произошло только в одном случае, однако можно наглядно видеть (см. рис. 11), что часть наименований ингредиентов и продуктов являются

синонимами, а часть – входит подстрокой в другие наименования. В этом случае подключаются дополнительные лингвистические ресурсы и происходит обогащение онтологий предметных областей новыми синонимами и сокращениями. Таким образом, разработчик/аналитик наглядно видит, что для определения нутриентного состава продуктов в тех или иных рецептах нет доступных мультипредметных ресурсов и для получения пертинентных результатов поиска в ответ на соответствующие комплексные запросы необходима семантическая интеграция данных, о чем в смарт-репозиторий системы заносятся соответствующие метаданные. Такая проверка выполняется повторно в соответствии с определенным регламентом работы аналитика, в частности, в случае обновления доступных ресурсов, она выполняется автоматически.

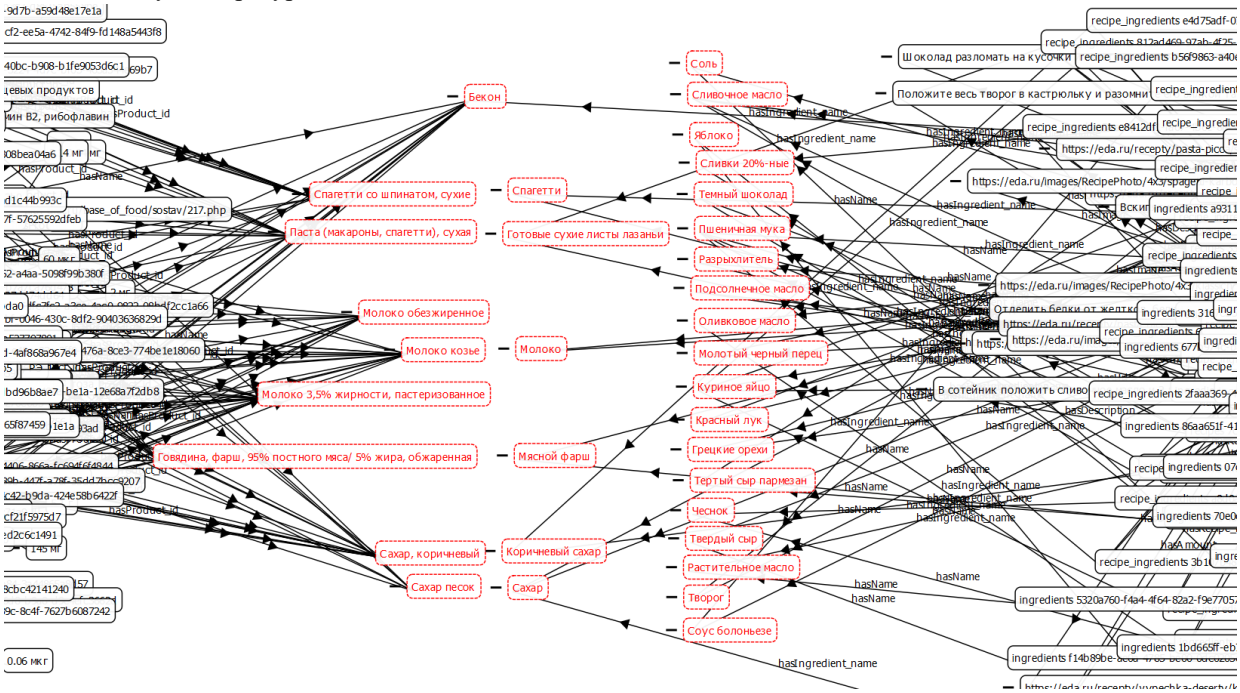


Рис. 11. Визуальный анализ извлеченных наименований продуктов и ингредиентов из разнородных ресурсов

V. Заключение

В данной работе основное внимание было уделено использованию средств визуального анализа данных для совершенствования моделей онтологий, управляющих процессом виртуальной семантической интеграции данных в рамках применения технологии фабрик данных, автоматизации построения графов знаний исходных ресурсов, эталонных онтологий предметных областей, их валидации и выявления потребности в смысловой интеграции данных. Изложение велось на примере семантической интеграции предметного содержания разнородных ресурсов, одна часть которых содержит описание рецептов и их ингредиентов, а другая – описание продуктов и их нутриентов. Предлагаемые методы и средства легко адаптируются к любой другой мультипредметной области за счет автоматизированного построения онтологий без необходимости внесения изменений в исходный программный код разрабатываемого инструментального средства.

Построение онтологий автоматизировано благодаря определенной методике использования современных сервисов генеративного ИИ, результаты применения которой также демонстрируются в данной работе. Сервисы генеративного ИИ служат исключительно для целей автоматизации построения онтологий и не включены в состав разрабатываемой инструментальной среды NuCoBoShell. Использование средств визуального анализа, предоставляемых средой отечественного визуального редактора онтологий ОНТОЛИС, позволило усовершенствовать модели указанных выше онтологий и решить ряд описанных в статье проблем семантической интеграции разнородных данных, представленных в разнообразных текстовых форматах. Подлежащие смысловой интеграции текстовые коллекции документов могут храниться как в общем доступе, так и локально на компьютере пользователя, что повышает уровень защищенности данных.

В перспективе планируется совершенствовать высокоуровневые средства анализа данных в среде визуального редактора онтологий ОНТОЛИС для автоматизации работ по подбору наиболее адекватных специфике конкретной предметной области метрик семантической близости понятий и комплексной проверке качества построенных онтологий.

Список литературы

1. Диковицкий В.В., Ломов П.А., Сепеда-Эррера Р.Р., Шишаев М.Г. Современные методы создания мультипредметных веб-ресурсов на базе визуализации и обработки формализованной семантики // Вестник Кольского научного центра РАН. 2011. №3. С. 63–73.

2. Semantic Analysis In NLP Made Easy, Top 10 Best Tools & Future Trends [Электронный ресурс]. – URL: https://spotintelligence.com/2023/10/16/semantic-analysis-in-nlp/ (дата обращения: 10.07.2024).

3. Гимашева К.В., Гладких Е.А., Чуприна С.И. Унифицированное применение методов онтологического инжиниринга в задачах обработки текстовых данных // Вестник Пермского университета. Математика. Механика. Информатика. 2022. №3(58). С. 49–58. DOI: 10.17072/1993-0550-2022-3-49-58.

4. Тузовский А.Ф., Ямпольский В.З. Интеграция информации с использованием технологий semantic web // Проблемы информатики. 2011. №2. С. 51–58.

5. Semantic Web W3C [Электронный ресурс]. – URL: https://www.w3.org/standards/ (дата обращения: 27.06.2024).

6. Patel A., Debnath, N.C., Bhushan, B. (Eds.). Semantic Web Technologies: Research and Applications (1st ed.). CRC Press, 2022. 404 p. DOI:10.1201/9781003309420.

7. Ballard C., Davies N., Gavazzi M., et al. IBM Informix: Integration through data federation. IBM Intern. Tech. Support Organizat. [Электронный ресурс]. 2008. 270 p. – URL: http://www.iiug.org/library/ids/technical/sg247032.pdf (дата обращения: 10.07.2024).

8. Когаловский М.Р. Методы интеграции данных в информационных системах // Сб. трудов Третьей Всероссийской конференции «Стандарты в проектах современных информационных систем». Москва. 23-24 апреля 2003 г.

9. Chuprina S.I. Using Data Fabric Architecture to Create Personalized Visual Analytics Systems in the Field of Digital Medicine // Scientific Visualization, 2023. Vol. 15(5). P. 50–63. DOI: 10.26583/sv.15.5.05.

10. Чуприна, С.И., Зиненко Д.В. ОНТОЛИС: адаптируемый визуальный редактор онтологий // Вестник Пермского университета. Математика. Механика. Информатика. 2013. № 3(22). С. 106–110.

11. Dzheiranian A.D., Ermakov I.D., Proskuryakov K.A., Lyadova L.N. Designing Data Visualization System Based on Language-Oriented Approach. Trudy ISP RAN/Proc. ISP RAS, Vol. 36(2), 2024. P. 127–140. DOI: 10.15514/ISPRAS-2024-36(2)-10.

12. Chuprina S., Postanogov I., Nasraoui O. Ontology Based Data Access Methods to Teach Students to Transform Traditional Information Systems and Simplify Decision Making Process // Procedia Computer Science. 2016. Vol. 80. P. 1801–1811. DOI:10.1016/j.procs.2016.05.458.

13. Гаврилова Т.А., Горовой В.А., Болотникова Е.С. Оценка когнитивной эргономичности онтологии на основе анализа графа // Искусственный интеллект и принятие решений. 2009. №3. С. 34–39.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-403-410

Сравнительный анализ точности солверов OpenFOAM для задачи обтекания конуса, имеющего сферическое затупление
Comparative Analysis of the Accuracy of OpenFOAM Solvers for the Problem of Flow Over a Spherically Blunted Cone

А. Е. Бондарев1, А. Е. Кувшинников1
1ИПМ им. М.В. Келдыша РАН, Москва, Россия

A. E. Bondarev1, A. E. Kuvshinnikov1
[1]Keldysh Institute of Applied Mathematics Russian Academy of Sciences, Moscow, Russia

Аннотация. Статья рассматривает вопрос о том, насколько точны различные солверы из программного пакета OpenFOAM. В качестве примера используется задача обтекания осесимметричного сферически затуплённого конуса несжимаемым газом при нулевом угле атаки. Результаты моделирования с использованием разных солверов сравниваются с данными из таблиц. Во время решения задачи угол полураствора конуса и скорость набегающего потока меняются с определённым шагом для каждого солвера. Сравнение проводится в широком диапазоне этих параметров. Анализ и визуализация полученных данных помогают определить, насколько точен численный метод для такого типа задач.

***Ключевые слова*:** сравнительная оценка точности, вычислительная газовая динамика, сверхзвуковой поток, сферически затупленный конус, OpenFOAM.

Аннотация. The article focuses on a comparative assessment of the accuracy of various solvers within the OpenFOAM software package. An axisymmetric flow problem over a spherically blunted cone, utilizing a non-viscous compressible gas at a zero angle of attack, serves as the benchmark case for evaluation. The computational outcomes from different solvers are juxtaposed against established tabulated solutions. In addressing the benchmark problem, each solver is subjected to variations in the cone's half-angle and the incoming flow velocity at specified increments. The comparative analysis spans a broad spectrum of these critical parameters. Detailed analyses and graphical representations of the acquired results yield insights into the relative accuracy of the numerical methods applied to this specific class of aerodynamic problems.

***Ключевые слова*:** comparative assessment of accuracy, computational fluid dynamics, supersonic flow, spherically blunted cone, OpenFOAM.

I. Введение

Ученых, работающих в программном пакете OpenFOAM [1] волнует вопрос сравнения точности и эффективности различных солверов. Использование пакета OpenFOAM для решения задач вычислительной газовой динамики связано с важной проблемой выбора подходящего солвера, который соответствует ожиданиям по точности, эффективности и устойчивости. Для этого необходимо провести анализ работы солвера на основе некоторых референтных решений. Однако существует не так много исследований, предлагающих детальное сравнение солверов OpenFOAM по указанным характеристикам. Некоторые примеры сравнений можно найти в работах [2] и [3], но они не предоставляют четких рекомендаций по выбору решателя. Более подробная попытка сравнения решателей была сделана в серии работ, где в качестве референтных рассматривались: задача сверхзвукового обтекания конуса под углом атаки [4, 5], моделирования скачка уплотнения [6], задача об образовании двумерной волны разрежения [7]. Также стоит отметить обзорную работу [8]. Метод, предложенный в этих работах, получил распространение в [9], [10]. С использованием техники обобщённого вычислительного эксперимента [11–14] были получены результаты, позволяющие расчётчику ориентироваться в многообразии разработанных численных методов и выбирать для расчетов наиболее точные и эффективные. Обобщенный вычислительный эксперимент предполагает разбиение определяющих параметров задачи внутри некоторого диапазона с последующим параметрическим исследованием и визуализацией многомерных результатов [15].

OpenFOAM (Open Source Field Operation And Manipulation CFD Toolbox) — популярный программный пакет для вычислительной гидродинамики (CFD), который активно используется для моделирования течений жидкостей и газов, теплопереноса и множества других задач, связанных с физическими процессами. OpenFOAM представляет собой open-source пакет, что значит, что он доступен бесплатно и его исходный код открыт. Тем

самым пользователь может модифицировать код под свои задачи, добавлять собственные уравнения, модели турбулентности или другие физические процессы. Допустим, если у вас есть специфическая задача, например, моделирование газодинамических ударных волн с учетом химических реакций, вы можете самостоятельно создать солвер, учитывающий данную особенность. Такая гибкость отличает OpenFOAM от коммерческих программ, где многие возможности закрыты для модификации. OpenFOAM поддерживает параллельные вычисления, что позволяет решать задачи с большой размерностью и сложной физикой. Это критично, когда нужно провести расчет на сетке, состоящей из миллионов ячеек. В OpenFOAM можно легко разделить расчетную область на несколько процессоров и тем самым ускорить вычисления. В сравнении участвовали четыре солвера: два стандартных солвера – rhoCentralFoam и sonicFoam и два авторских – pisoCentralFoam [16] и QGDFoam [17]. Последние два солвера разработаны коллективами Института системного программирования РАН и Института прикладной математики им. М.В.Келдыша РАН.

Солвер rhoCentralFoam использует центрально-противопотоковую схему годуновского типа, предложенную Кургановым и Тадмором (КТ) [18], а позже модифицированную в статье Курганова, Ноэль и Петровой (KNP) [19]. Схема Курганова — Тадмора обходит решение задачи Римана, используя аппроксимацию локальных скоростей распространения волн и центральный подход. Вместо решения задачи Римана для определения потока на границе между ячейками, КТ-схема использует оценку максимальной скорости распространения волн (обычно через оценку максимальной скорости характеристик уравнений) с обеих сторон ячейки. Это скорость, с которой информация распространяется через границу ячейки. Например, для уравнений Эйлера это может быть скорость звука плюс скорость потока, поскольку это максимальная скорость распространения информации в сжимаемом газе. Такой подход позволяет избежать многих трудностей, связанных с вычислением ударных волн и резких градиентов в полях плотности или давления, которые могут привести к осцилляциям в численных решениях. Схема KNP была реализована для OpenFOAM Гриншилдсом [20].

Солвер sonicFoam использует алгоритм PIMPLE [21] для связи давления и скорости. Этот алгоритм объединяет методов элементы методов PISO и SIMPLE. В основе алгоритма PIMPLE лежит его итерационная структура, предназначенная для уточнения полей скорости и давления в области жидкости. Алгоритм начинается с начального предположения о поле скоростей, которое часто берется из предыдущих временных шагов или оценивается с помощью более простых подходов. Это начальное предположение имеет решающее значение, поскольку оно задает основу для последующего итерационного процесса. На первом этапе алгоритм предсказывает поле скоростей на основе существующего распределения давления, применяя явную схему, которая облегчает прямой расчет скорости. После получения первоначального прогноза скорости алгоритм переходит к обновлению поля давления. Именно здесь проявляется сила метода PISO. На этом этапе алгоритм рассчитывает поправку на давление, которая обеспечивает исполнение закона сохранения массы. Поправка на давление очень важна, поскольку она позволяет устранить расхождения между прогнозируемым и фактическим поведением газа, особенно в областях, где скорости потока могут быстро меняться или где существуют сложные граничные условия. После коррекции давления поле скоростей обновляется снова с учетом новой информации о давлении. Этот итерационный процесс повторяется несколько раз, как правило, в течение одного временного шага, что позволяет непрерывно уточнять поля скорости и давления. Каждая итерация помогает повысить точность решения, что приводит к большей стабильности.

При использовании явных методов, основанных на решении задачи Римана или расщеплении вектора потоков, возникают проблемы: при низких числах Маха решение теряет устойчивость, либо необходим крайне малый временной шаг, что увеличивает вычислительные затраты и делает такие методы неэффективными для рассматриваемых задач. С другой стороны, неявные методы расщепления, такие как PISO и SIMPLE, могут вызывать нефизические осцилляции в численном решении и потерю устойчивости при высоких числах Маха. Подобные явления значительно ограничивает эффективность этих методов в задачах, где сильную роль играет сжимаемость, и делает их неподходящими для изучения высокоскоростных потоков или акустических явлений. Одним из решений может быть использование гибридного подхода, где алгоритмы PISO/SIMPLE для неявного интегрирования уравнений сохранения массы, количества движения и энергии сочетаются с неосциллирующими методами для дискретизации конвективных компонентов. Авторами [16] была выбрана схема Курганова — Тадмора в качестве неосциллирующего метода; она уже реализована в виде явной схемы в OpenFOAM и прошла успешные испытания, а также достаточно проста для интеграции в гибридный подход. Одно из ключевых преимуществ этой схемы — независимость аппроксимации потоков физических величин от характеристик системы уравнений, что устраняет необходимость поиска Римановых инвариантов и разложения решений по характеристикам. Основная идея гибридного метода заключается в добавлении функции-переключателя, которая будет переводить аппроксимацию потоков от вида КТ-схемы к форме, применяемой в методе PIMPLE, в зависимости от приближения к дозвуковым скоростям [22]. Все эти гибридные решатели не включены в OpenFOAM, но доступны в публичном репозитории авторов [23].

Система квазигазодинамических уравнений была разработана коллективом под руководством Б.Н. Четверушкина [24]. Она основана на идее, что газ можно представить не как непрерывную среду, а как набор частиц, которые движутся и взаимодействуют друг с другом на микроскопическом уровне. КГД-модель является промежуточной между классической гидродинамикой и более сложными моделями, которые учитывают микроскопические эффекты. Введение дополнительных членов в квазигазодинамические уравнения осуществляется с целью учета мелкомасштабных эффектов, которые не могут быть разрешены напрямую в рамках традиционных уравнений Навье-Стокса [25]. Эти дополнительные члены часто называются диссипативными, и они вводятся путем различных усреднений и аппроксимаций, которые отражают физические процессы. Добавление слагаемых позволяет более точно описывать сложные явления и взаимодействия в газах, особенно в условиях, когда важны неравновесные процессы и трансформация энергии. На основе системы был разработан солвер QGDFoam. Параметр τ играет ключевую роль в учете мелкомасштабных эффектов и определяет масштаб времени для диссипативных процессов. Управляемый параметр α (связанный с τ) в диссипативных членах обеспечивает этому солверу регулируемую схемную вязкость, что позволяет уменьшать нежелательные осцилляции на разрывах.

II. Постановка задачи

В данной работе для сравнения солверов используется двумерная невязкая задача образования установившегося течения, полученного при обтекании затупленного кругового конуса с углом полураствора β сверхзвуковым потоком газа с числом Маха *M* при нулевом угле атаки. Определяющими параметрами задачи в терминах обобщенного вычислительного эксперимента здесь служат число Маха *M* и угол полураствора конуса β. Выбор диапазонов изменения переменных параметров и шага их изменения производился следующим образом: число Маха M равнялось 2, 4 и 6, угол полураствора конуса β составлял 5°, 10°, 15° и 25°. Общая схема течения представлена на Рис. 1. Для всех выбранных параметров у задачи существует табличное решение [26]. Для расчета взята система уравнений Эйлера, замыкаемая уравнением состояния идеального газа.

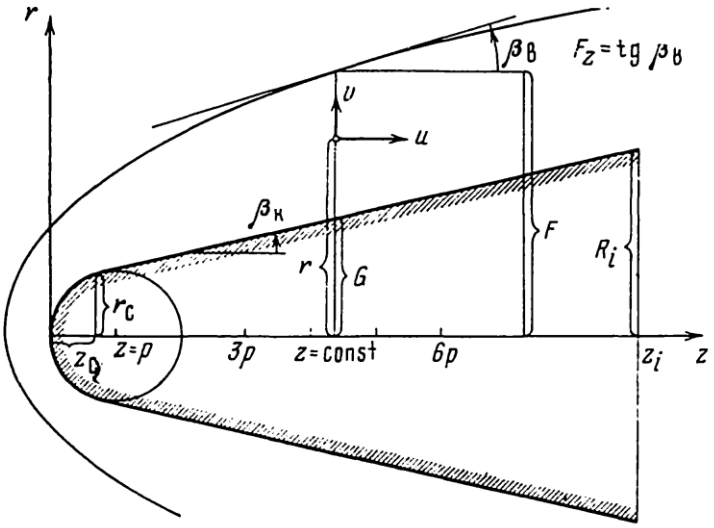


Рис. 1. Схема течения

III. Организация расчётов

Расчётная область, как показано на Рис. 2, разбивается на ячейки. Для работы с пакетом OpenFOAM необходимо определить граничные и начальные условия. На входной границе, обозначенной как «inlet», задаются характеристики невозмущенного потока: давление P составляет 101325 Па, температура T равна 300 K, а x-компонента скорости Ux варьируется от 694.5 м/с до 2083.5 м/с, при этом y-компонента скорости Uy равна 0 м/с. На выходной границе «outlet» и верхней границе «top» устанавливаются условия, при которых производные газодинамических функций по нормали к границе равны нулю. На границе конуса «cone» для давления и температуры применяется условие нулевого градиента, а для скорости используется условие «slip», что соответствует непротеканию в уравнениях Эйлера. Для моделирования осесимметричной геометрии в OpenFOAM на передней «front» и задней «back» границах применяется специальное условие «wedge». На оси «axis» в OpenFOAM вводится особое граничное условие «empty», которое используется, когда вычисления в данном направлении не производятся.

Расчётная область (Рис. 2) разбивается на ячейки. Пакет OpenFOAM для решения требует задания граничных и начальных условий. На входной границе «inlet» задаются параметры невозмущенного набегающего потока (давление P = 101325 Па, температура T = 300 K, x-компонента скорости изменяется от 694.5 м/с (2 Маха) до 2083.5 м/с (6 Маха), y-компонента скорости равна 0 м/с). На выходной границе «outlet» и на верхней границе «top» задаются граничные условия равенства нулю производных газодинамических функций по нормали к границе, определяемое в терминах OpenFOAM как «zeroGradient». На границе конуса «cone» для давления и температуры также задается условие нулевого градиента, для скорости задается условие «slip», соответствующее условию непротекания для уравнений Эйлера. Для моделирования осесимметричной геометрии для передней «front» и задней «back» границ используется специальное условие «wedge». Для оси «axis» задается специальное граничное условие «empty». Оно используется в случаях, когда вычисления в заданном направлении не проводятся.

Схема расчётной области для конуса с углом полураствора β = 15°представлена на Рис. 2. Стоит отметить, что на указанном изображении для наглядности сетка крупнее, чем в реальных расчётах.

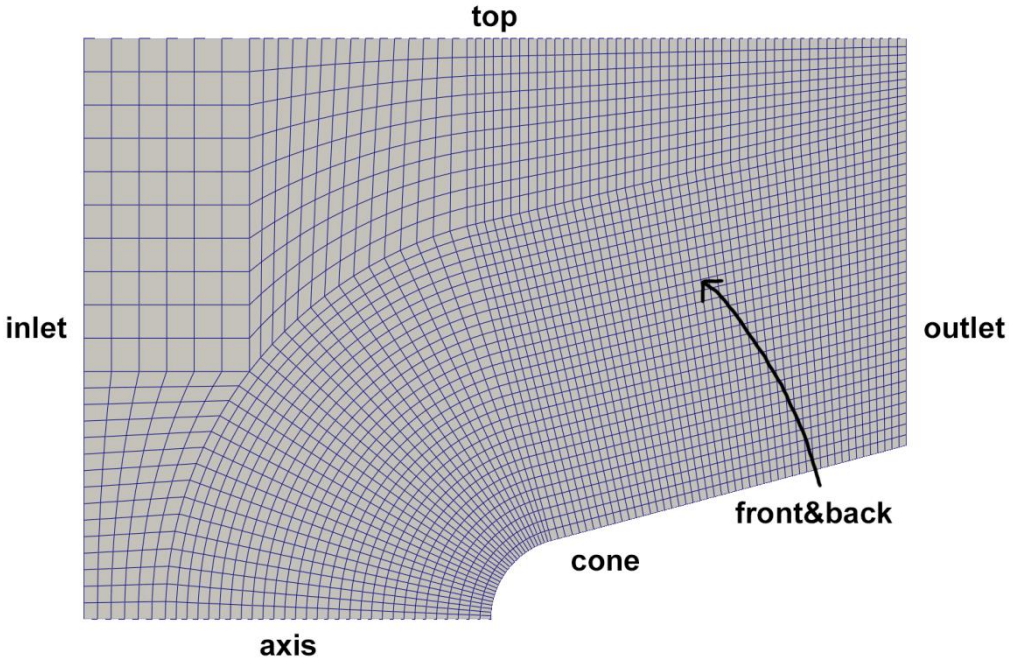


Рис. 2. Схема расчётной области

Число ячеек сетки — 35950. Начальные условия соответствуют граничным на грани inlet, то есть параметры набегающего потока являются и начальными условиями. В солвере QGDFoam на всей расчетной области устанавливался коэффициент сглаживания α = 0.1. Также задавались значения молярной массы M = 28.96 и удельной теплоемкости при постоянном давлении C_p = 1004.

В отличие от многих других пакетов, в OpenFOAM управление моделированием происходит через текстовые файлы. Это обеспечивает гибкость, так как легко можно автоматизировать запуск задач, изменять параметры симуляции и обрабатывать результаты. Унификация расчётов играет ключевую роль в организации сравнения солверов, так как обеспечивает однородные условия для оценки их производительности и точности. При наличии стандартизированных методик, сеток, граничных условий и физических моделей результаты становятся сопоставимыми и надежными. Это позволяет исследователям исключить влияние внешних факторов и сосредоточиться на характеристиках конкретного солвера. Кроме того, унификация способствует лучшему пониманию сильных и слабых сторон каждого солвера, что может быть полезно для выбора оптимального инструмента в зависимости от специфики задачи. В пакете OpenFOAM мы использовали те же параметры файлов конфигурации fvSchemes и fvSolution, что и в [5].

IV. Результаты экспериментов

Расчёты для всех солверов обеспечили получение хорошо известной качественной картины течения для рассматриваемой задачи. Один из вариантов представлен на Рис. 3 в виде распределения давления в расчетной области. Представленное распределение давления получено с помощью солвера *rhoCentralFoam.* Разрушения

решения не наблюдалось ни для одного из солверов, что свидетельствует о высоких стабилизирующих свойствах всех солверов, участвующих в исследовании.

Построим оценки отклонения от точного решения для всей расчетной области в норме L_2. Для этого определим относительную погрешность Err для нормы L_2 следующим образом:

$$L_2: \; Err = \sqrt{\sum_m |y_m - y_m^{exact}|^2 V_m} \Big/ \sqrt{\sum_m |y_m^{exact}|^2 V_m}$$

Здесь y_m — давление p, V_m — объём ячейки. Значения y_m^{exact} получены интерполяцией табличного решения задачи. В анализе сравнительной точности участвовали солверы sonicFoam, QGDFoam, rhoCentralFoam и pisoCentralFoam. Далее в таблицах для солверов используются сокращенные обозначения: rCF (*rhoCentralFoam*), pCF (*pisoCentralFoam*), sF (*sonicFoam*), QGDF (*QGDFoam*). Значения отклонения от точного решения по всей расчетной области рассчитаны для давления p и приведены в Табл. 1–3. Жирным выделены наименьшие значения в каждой строке.

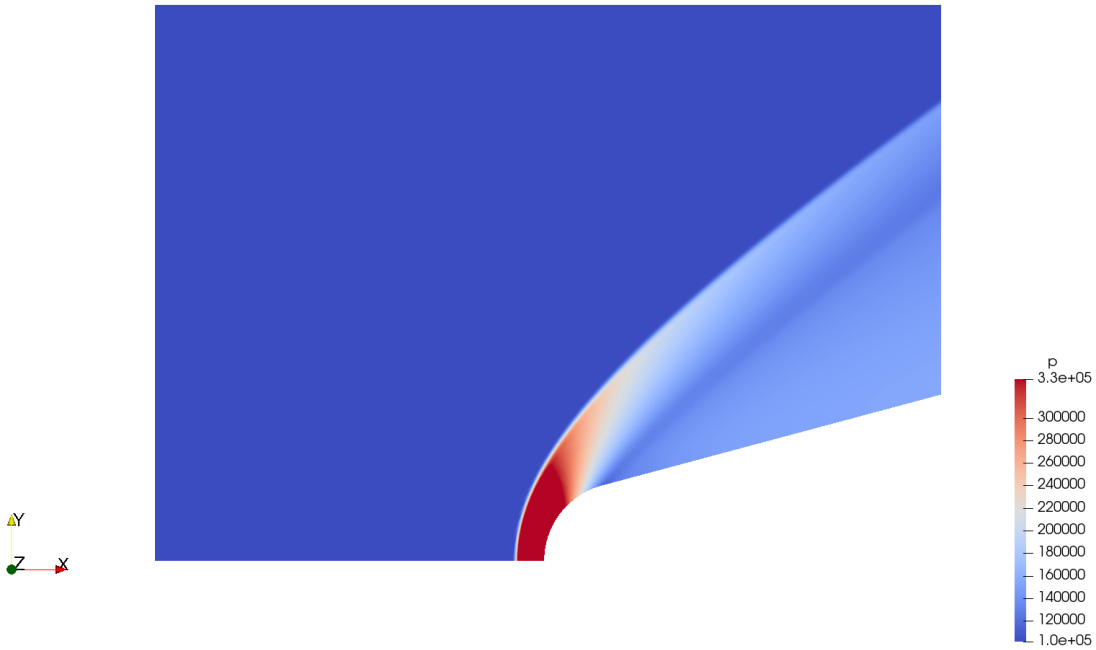


Рис. 3. Поле давление установившегося течения для солвера rhoCentralFoam

ТАБЛИЦА 1
ОШИБКИ ДЛЯ M = 2

Угол полураствора	rCF	pCF	sF	QGDf
5	**0.017028**	0.022472	0.031921	0.025685
10	**0.027280**	0.033699	0.055463	0.042351
15	**0.034960**	0.040031	0.080514	0.046906
25	**0.036960**	0.043399	0.085460	0.049201

ТАБЛИЦА 2
ОШИБКИ ДЛЯ M = 4

Угол полураствора	rCF	pCF	sF	QGDf
5	**0.031943**	0.040018	0.067521	0.051875
10	**0.045202**	0.053764	0.125038	0.062389
15	**0.051245**	0.060396	0.154412	0.074341
25	**0.051458**	0.061273	0.161013	0.077917

ТАБЛИЦА 3
ОШИБКИ ДЛЯ M = 6

Угол полураствора	**rCF**	**pCF**	**sF**	**QGDf**
5	**0.057095**	0.064405	0.174251	0.083892
10	**0.065626**	0.067459	0.183761	0.095616
15	**0.068654**	0.071534	0.194212	0.099098
25	**0.071054**	0.072618	0.204263	0.100022

Визуализация в виде поверхности ошибок размещена на Рис. 4.

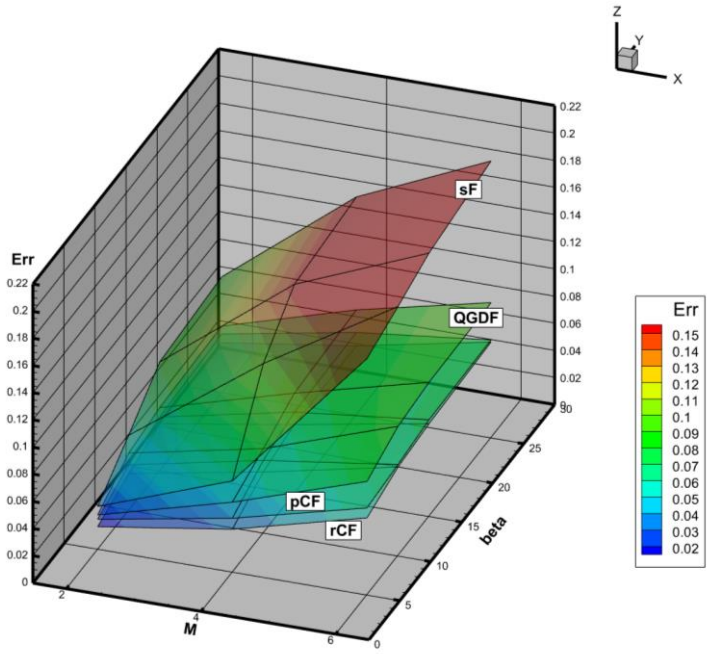


Рис. 4. Изменение отклонения от точного решения для давления в зависимости от числа Маха и угла полураствора конуса для всех солверов в норме L2

Анализ рисунка показывает, что наименьшее отклонение от точного решения показывает солвер rhoCentralFoam, наибольшее отклонение показывает солвер sonicFoam. Солвер psioCentralFoam чуть хуже солвера rhoCentralFoam. Отклонение солвера QGDFoam больше, чем у rhoCentralFoam и psioCentralFoam. Но стоит учитывать, что в этом исследовании у солвера QGDFoam не менялся параметр сглаживания α, а, как показано в работе [27], изменение этого параметра влияет на точность решения. Вполне возможно, что α = 0.1 не является оптимальным значением для подобной задачи. Этот вопрос будет рассмотрен в наших следующих работах.

Также наглядно показано увеличение отклонения при увеличении определяющих параметров. Сильнее всего, от 0.067521 до 0.174251, изменилось отклонение у солвера sonicFoam для β = 5° при увеличении *M* от 4 до 6. Это связано с образованием осцилляций, так что использование солвера sonicFoam не рекомендуется.

V. Выводы и заключение

Для задачи обтекания конуса, имеющего сферическое затупление, сверхзвуковым потоком идеального газа под нулевым углом атаки проведено сравнение численного решения четырёх солверов программного пакета OpenFOAM с решением, полученным из таблиц [21]. Исследование охватывало широкий диапазон вариаций ключевых параметров, таких как число Маха набегающего потока и угол полураствора конуса. Сравнение результатов с точным решением осуществлялось по всему расчетному полю с использованием разностного аналога нормы L_2.

Результаты, представленные в табличной и графической формах, демонстрируют, что солвер rhoCentralFoam имеет наименьшую норму погрешности. Солвер pisoCentralFoam занимает второе место по точности. Для всех значений числа Маха солвер sonicFoam вызывает осцилляции на переднем фронте ударной волны, что отрицательно влияет на норму погрешности. Солвер QGDFoam показывает большее отклонение от точного решения по сравнению с солверами rhoCentralFoam и pisoCentralFoam, но в данном исследовании не менялся параметр сглаживания, который, исходя из наших прошлых работ, влияет на точность решения. В перспективе авторы рассматривают расширение сравнительной оценки для других величин, а именно, плотности и компонент скорости. Также планируется решение оптимизационной задачи для солвера QGDFoam.

Благодарности

Вычисления проведены с помощью гибридного суперкомпьютера К100, установленного в Суперкомпьютерном Центре коллективного пользования ИПМ им. М.В. Келдыша РАН.

Список литературы

1. OpenFOAM Foundation: [Электронный ресурс]. URL: http://www.openfoam.org (Дата обращения: 02.07.2024).

2. Gutierrez L. F., Tamagno J. P., Elaskar S. A. High speed flow simulation using OpenFOAM // Mecanica Computacional. 2012. Vol. XXXI. P. 2939–2959.

3. Lorenzon D., Elaskar S. A. Simulacion de flujos supersonicos bidimensionales y axialmente simetricos con OpenFOAM // Revista de la Facultad de Ciencias Exactas, Fisicas y Naturales. 2015. Vol. 2. no. 2. P. 65–76.

4. Bondarev A. E., Kuvshinnikov A. E. Analysis of the behavior of OpenFOAM solvers for 3D problem of supersonic flow around a cone at an angle of attack // CEUR Workshop Proceedings, 2020, V. 2763, p.48–51, CPT2020, Proceedings of the 8th International Scientific Conference on Computing in Physics and Technology, Moscow region, Russia, November 09-13, 2020.

5. Alekseev A. K., Bondarev A. E., Kuvshinnikov A. E. On uncertainty quantification via the ensemble of independent numerical solutions // Journal of Computational Science. 2020. Vol. 42 101114.

6. Alekseev A. K., Bondarev A. E., Kuvshinnikov A. E. Comparative analysis of the accuracy of openfoam solvers for the oblique shock wave problem // Mathematica Montisnigri, 2019, Vol. XLV. P. 95-105.

7. Bondarev A. E., Kuvshinnikov A. E. Analysis and Visualization of the Computational Experiments Results on the Comparative Assessment of OpenFOAM Solvers Accuracy for a Rarefaction Wave Problem // Scientific Visualization. 2021. Vol. 13. № 3. P. 34–46.

8. Бондарев А.Е., Кувшинников А.Е. Задачи сравнительной оценки численных методов на референтных решениях // Труды Семнадцатой Международной научно-технической конференции «Оптические методы исследования потоков». М.: Научно-технологический центр уникального приборостроения РАН, 2023. С. 517–528.

9. Canteros M.A., Polanský J. Review and comparison of two OpenFOAM® solvers: rhoCentralFoam and sonicFoam // EPJ Web of Conf. 2024. Vol. 299 01005.

10. Analysis of the oscillations induced by a supersonic jet applied to produce nanofibers / Quintero F., Doval A.F., Goitia A. et al. // International Journal of Mechanical Sciences. 2022. Vol. 238. 107826.

11. Бондарев А. Е., Галактионов В. А. Построение и обработка результатов вычислительного эксперимента на основе параллельных решений для оптимизационных и параметрических исследований в газовой динамике (пленарный доклад) // Оптические методы исследования потоков: Труды XIV Международной научно-технической конференции. М.: Издательство «Перо», 2017. С.7–15.

12. Bondarev A. E. On visualization problems in a generalized computational experiment //Scientific Visualization. 2019. Vol. 11.2 P. 156–162.

13. Alekseev A. K., Bondarev A. E., Galaktionov V. A., Kuvshinnikov A.E. On the construction of a generalized computational experiment in verification problems // Matematica Montisnigri. 2020. Vol. XLVIII. P. 19–31.

14. Alekseev A. K., Bondarev A. E., Galaktionov V. A., Kuvshinnikov A. E. Generalized Computational Experiment and Verification Problems // Program. Comput. Soft. 2021. Vol. 47 P. 177–184.

15. Zakharova A. A, Korostelyov D. A., Podvesovskii A .G., Bondarev A. E., Galaktionov V. A. Generalized Computational Experiment State Analysis Using Three-Dimensional Visual Maps // Scientific Visualization. 2022 Vol. 14.4 P. 12–23.

16. Kraposhin M., A. Bovtrikova A., Strijhak S. Adaptation of Kurganov-Tadmor numerical scheme for applying in combination with the PISO method in numerical simulation of flows in a wide range of Mach numbers // Procedia Computer Science. 2015. Vol. 66. P. 43–52.

17. Истомина М. А. О реализации одномерного квазигазодинамического алгоритма в открытом программном комплексе OpenFOAM // Препринты ИПМ им. М.В. Келдыша. 2018. № 001.

18. Kurganov A., Tadmor E. New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations // J. Comput. Phys. 2000. Vol. 160. № 1. P. 241–282.

19. Kurganov A., Noelle S., Petrova G. Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton–Jacobi equations // SIAM J Sci Comput. 2001. Vol. 23 P. 707–740.

20. Greenshields C.J., Wellerr H.G., Gasparini L., Reese J.M. Implementation of semi-discrete, non-staggered central schemes in a colocated, polyhedral, finite volume framework, for high-speed viscous flows // Int. J. Numer. Meth. Fluids. 2010. Vol. 63. № 1. P. 1–21.

21. Issa R. Solution of the implicit discretized fluid flow equations by operator splitting // J. Comput. Phys. 1986. Vol. 62. № 1. P. 40–65.

22. Kraposhin M. V., Banholzer M., Pfitzner M., Marchevsky I. K. A hybrid pressure-based solver for nonideal single-phase fluid flows at all speeds // Int. J. Numer. Meth. Fluids. 2018. Vol. 88. № 2. P. 79–99.

23. United collection of hybrid Central solvers — one-phase, two-phase and multicomponent versions: [Электронный ресурс]. URL: https://github.com/unicfdlab/hybridCentralSolvers (Дата обращения 02.07.2024).

24. Четверушкин Б. Н. Кинетические схемы и квазигазодинамическая система уравнений М.: МАКС Пресс, 2004. 328 с.

25. Elizarova T.G. Quasi-Gas Dynamic Equations. Springer, Berlin, Heidelberg, 2009. 286 p.

26. Любимов А.Н., Русанов В.В. Течения газа около тупых тел. Часть II: Таблицы газодинамических функций. М.: Наука, 1970. 380 с.

27. Bondarev A. E., Kuvshinnikov A. E. Comparative Estimation of QGDFoam Solver Accuracy for Inviscid Flow Around a Cone // IEEE The Proceedings of the 2018 Ivannikov ISPRAS Open Conference (ISPRAS-2018). P. 82–87.

УДК 620.1, 621.7, 004.942
DOI: 10.25206/978-5-8149-3873-2-2024-411-417

Применение технологий 3D-печати при изготовлении инструментов для микродеформирования листовой заготовки
Application of 3D-printing technologies in manufacturing of tools for microforming of sheet metal blanks

М. А. Петров[1], Ч. Д. Хынг[1]
[1]ФГАОУ ВО «Московский политехнический университет», кафедра «Обработка материалов давлением и аддитивные технологии», г. Москва, Российская Федерация

M. A. Petrov[1], T. D. Hung[1]
[1]FSAEI HE "Moscow Polytechnic University", Department "Material Forming & Additive Technologies", Moscow, Russian Federation

Аннотация. В представленной статье рассмотрены результаты изготовления инструмента для микродеформирования, полученные по технологиям аддитивного производства или 3D-печати. В качестве метода оценки точности выбран способ бесконтактного 3D-сканирования, на основе структурированного подсвета. Установлено, что технологии 3D-печати позволяют изготовить инструмент, но для совпадения с чертёжными размерами необходимо закладывать допуски на обработку. Проведена оценка поверхности прототипов. Подвергаясь цикличным нагрузкам, инструмент изнашивается. На основании полученных испытаний по схеме «стержень по пластине», была определена склонность напечатанных образцов к износу.

***Ключевые слова*:** 3D-печать, FFF, SLM, LCD, микродеформирование, медь М1, латунь Л63, износ поверхности.

Аннотация. The presented paper considers the results of tools' manufacturing for microforming obtained by additive manufacturing or 3D-printing technologies. The method of non-contact structured-light 3D-scanning was chosen as methods for accuracy assessment. It was found out, that not all considered 3D-printing technologies guarantees the matching drawing dimensions, so the tolerances for post-treatment should be taken into consideration. Surface characteristics have been carried out. Repeating the same function follows to loading in cyclic modes even for small batches, that subjected to wear of the tools' surfaces. Based on the tests according to scheme "rod-on-plate", the wear tendency of the printed samples was determined.

***Ключевые слова*:** 3D-printing, FFF, SLM, LCD, microforming, copper M1, brass L63, fatigue wear.

I. Введение

Представление о прикладных, производственных технологиях трансформируются по мере развития фундаментальных исследований. С появлением высокотехнологичных отраслей, например, микроэлектронной промышленности, в технологиях заготовительного производства стали выделять направление по изготовлению микродеталей и изделий. Лавируя между размером изделий и их требуемым количеством, общая доля заготовительного производства не уменьшилась. Наоборот, массовое производство микродеталей заставляет более тщательно подходить к разработке производственной технологии, чтобы не допустить большого процента брака. Малоизученным направлением остаются процессы микродеформирования или микроштамповки. Основной фундаментальной предпосылкой для разработки технологии микроштамповки остаётся уравнение Холла-Петча, основывающегося на теории дислокаций, согласно которому с уменьшением размера зерна происходит существенный рост прочностной характеристики изделия, согласно уравнению (1), для определённой степени деформации (ε), что приводит к снижению пластических свойств [1, 2]. Иными словами, существует минимальный размер зерна, разрушение которого потребует достаточно высоких напряжений. Соответственно, пластичность металлического образца обусловлена дефектами поликристалла, являющегося совокупностью большого числа кристаллитов или зёрен, имеющих разный размер. Так как размеры зёрен могут изменяться в диапазоне макро- (до 1 мм), микро- (от 1 мм до 1 мкм) и нано-шкал (от 1 мкм до 1 нм), а также иметь промежуточные (мезо-) уровни, то встаёт вопрос об изучении технологических особенностей процесса деформирования на каждом из уровней. В связи с этим появляются отдельные научные направления по изучению материалов и технологий на разных размерных уровнях. Для изделий микро-уровня в англоязычном пространстве существуют такие термины, как microforming (микродеформирование), microstamping

(микроштамповка), miniaturized (миниатюрный) и другие [3, 4, 5]. Сложность изготовления малоразмерных изделий заключается в необходимости применения дополнительных средств объективного контроля процесса и манипуляции/позиционирования, а также специализированного оборудования. Изготавливаемые микродеформированием детали могут быть как листовыми, так и объёмными [6]. На рис. 1а и 1б показаны примеры профилированных пластин и стаканчиков (рис. 1в), полученных из разных материалов с разным коэффициентом масштабирования (подобия) исходной геометрии (λ). Отчётливо видно, что конструкционные материалы ведут себя отлично друг от друга в процессе их деформирования, в зависимости от коэффициента подобия. Для исследования использовались листовые заготовки, толщиной 0,1 мм, из меди М1, латуни Л63, алюминиевых сплавов АД1 (сложенная пищевая фольга, толщиной 25 мкм) и ДПРНМ. Гофры в области фланца указывают на недостаточную силу прижима фланцевой области заготовки в процессе деформирования (рис. 1в).

$$\sigma_т = \sigma_0 + \frac{k}{\sqrt{d}} \tag{1}$$

где $\sigma_т$ – предел текучести материала; σ_0 – напряжение, необходимое для инициализации движения дислокаций, напряжение, которое характеризует меру межзёренного трения (например, для меди σ_0 = 20…25 МПа, для алюминия σ_0 = 20 МПа, для титана σ_0 = 80 МПа [7, 8]); k – коэффициент упрочнения или константа материала (например, для меди k = 0.11…0.14 МПа×м$^{0.5}$, для алюминия k = 0.04 МПа×м$^{0.5}$, для титана k = 0.4 МПа×м$^{0.5}$ [7,8]); d – средний размер зерна или кристаллита.

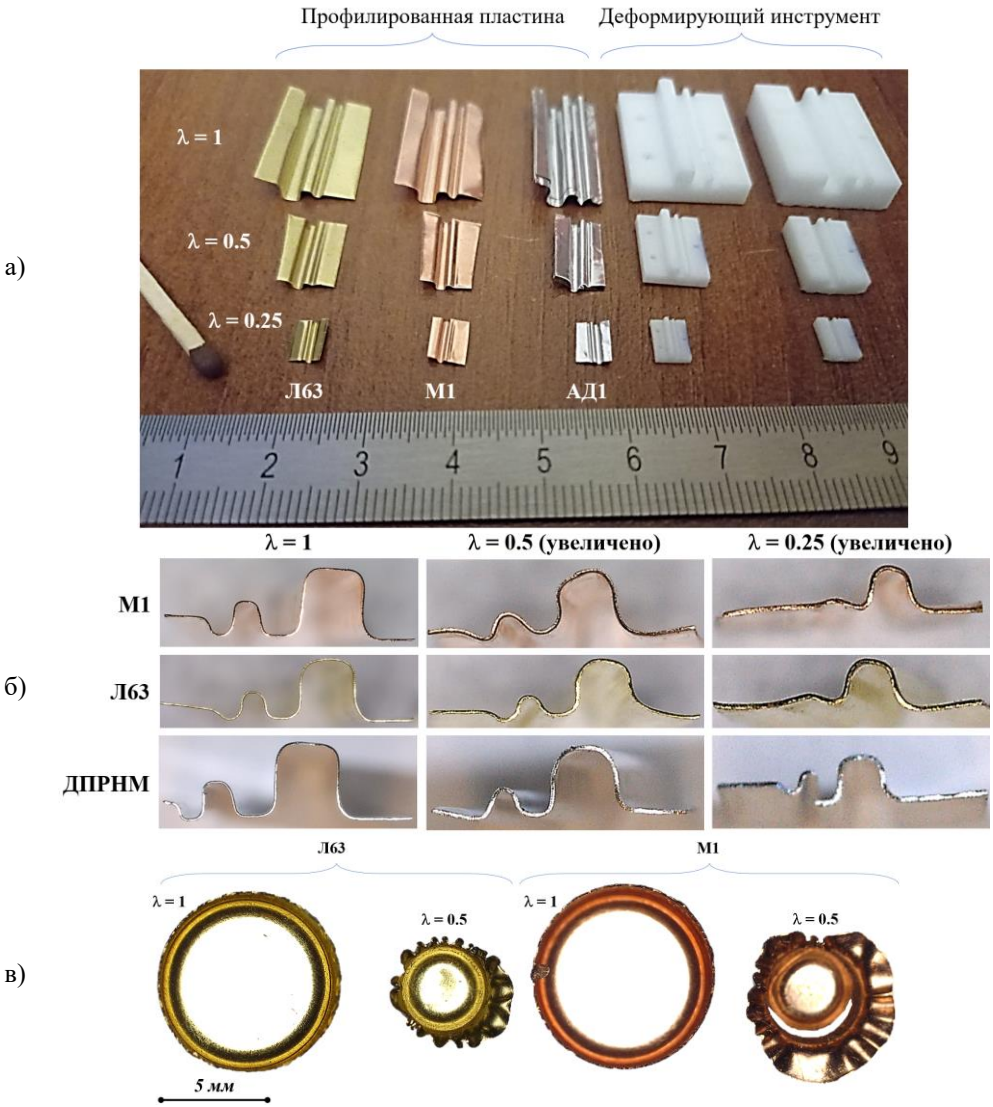


Рис. 1. Примеры полученных микроизделий: профилированные пластины (а, б); осесимметричные детали – стаканчики, вид сверху (в)

Для реализации процесса микроштамповки требуется наличие исходной заготовки заданной толщины и диаметра, деформирующего инструмента, а также совокупности технологических параметров, определяющих граничные условия, например, температура окружающей среды, коэффициент трения на контактной поверхности и т.д. Инструмент является дорогостоящим компонентным звеном в этой цепи. Рабочая часть инструмента обладает сложным профилем и не всегда традиционным способом, механической обработкой (субтрактивные технологии), его изготовление экономически целесообразно. В XXI веке, благодаря развитию аппаратной и программной частей, стало возможным послойное создание прототипов изделий, что снижает расход материала и позволяет изготавливать сложные изделия. Эти технологии, получившие название аддитивные технологии или 3D-печать, работают как с металлическими, так и с полимерными материалами. На рис. 2 показан график, отражающий преимущество изготовления сложных прототипов методами аддитивного производства по сравнению с методами субтрактивного производства. Получаемые прототипы обязательно подвергаются пост-обработки, в большей или меньшей степени. Именно это обстоятельство позволяет сделать выбор либо в пользу традиционных, либо в пользу аддитивных технологий.

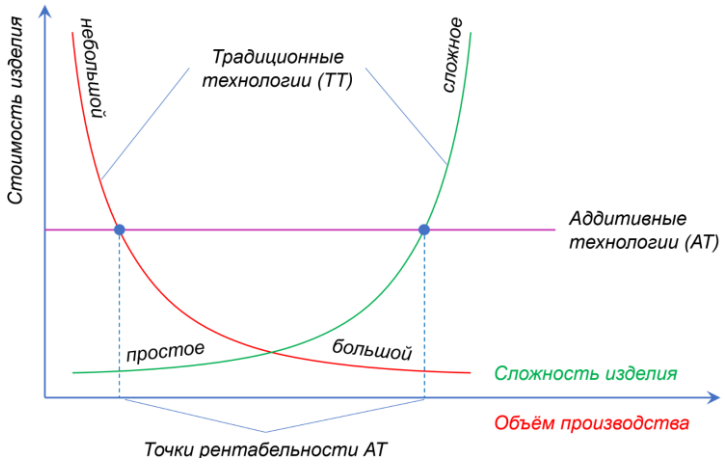


Рис. 2. График сравнения АТ и ТТ

Вместе с преимуществом построения сложной геометрии прототипа возникает недостаток аддитивных технологий, заключающийся в том, что прототип получает текстурированную поверхность, обусловленную различным способом построения объекта, что зависит как от технологии 3D-печати, так и от параметров управляющей процессом печати программы, и не позволяет применять прототип сразу после печати, а требует пост-обработки. Текстурирование поверхности может как улучшить, так и ухудшить технологичность процесса микроштамповки. Так, топография поверхности влияет на коэффициент контактного трения [9] и, соответственно, на минимальный и/или максимальный коэффициент микровытяжки листовой заготовки, силу прижима фланца и другое. Также, в зависимости от технологии 3D-печати имеет место отклонение геометрии от размеров исходного чертежа [10].

II. Изготовление деформирующего инструмента

Были рассмотрены три технологии 3D-печати. Для изготовления металлического инструмента применялись технологии DMLS и SLM. Для изготовления полимерного инструмента применялись технологии FFF и LCD. Краткая сводная информация представлена в табл. 1. В качестве материалов для металлической 3D-печати рассматривались титановый сплав (ВТ6), нержавеющая сталь (03Х17Н12М2) и алюминиевый сплав (AlSi10Mg). Для изготовления полимерных прототипов деталей инструмента методом экструзионной печати использовался полиэтилентерефталат-гликоль (ПЭТГ), а для метода масочной фотополимерной печати – базовый фотополимер (БФП), а также керамоподобный фотополимер (КПФП).

ТАБЛИЦА 1
КРАТКАЯ ХАРАКТЕРИСТИКА ТЕХНОЛОГИЙ 3D-ПЕЧАТИ

Название технологии	**Производитель оборудования**	**Материал**
DMLS	3D Systems	ВТ6, AlSi10Mg (аналог АК9)
SLM	3DLAM	03Х17Н12М2
FFF	Anycubic	ПЭТГ
LCD	Creality-Halot	БФП, КПФП

Далее исследование проходило по схеме, представленной на рис. 3. Технология 3D-печати экструзионного типа FFF (Fused Filament Fabrication) заключается в послойном добавлении полимерного материала, подающегося из катушки с полимерной проволокой ∅1.75 мм, проходящего через горячую зону и выдавливающегося через сопло экструдера, диаметр которого равен ∅0.4 мм, что на ~77% меньше исходного диаметра проволоки. Выращивание прототипа ведётся по направлению снизу-вверх. Фотополимерная 3D-печать по масочной технологии (Liquid-Crystal Display) подразумевает вырост прототипа из вязкой фотополимерной композиции с отверждение УФ-излучением, проходящего через жидкокристаллическую матрицу, на которой генерируются сечения профиля, согласно управляющей программе 3D-принтера, работает по схеме сверху-вниз. Технологии лазерного сплавления металлических порошков SLM (Selective Laser Melting) и DMLS (Direct Metal Laser Sintering) позволяют получать металлические прототипы по схеме выроста снизу-вверх за счёт расплавления лучом лазера материала и его кристаллизации по сечениям, согласно управляющей программе.

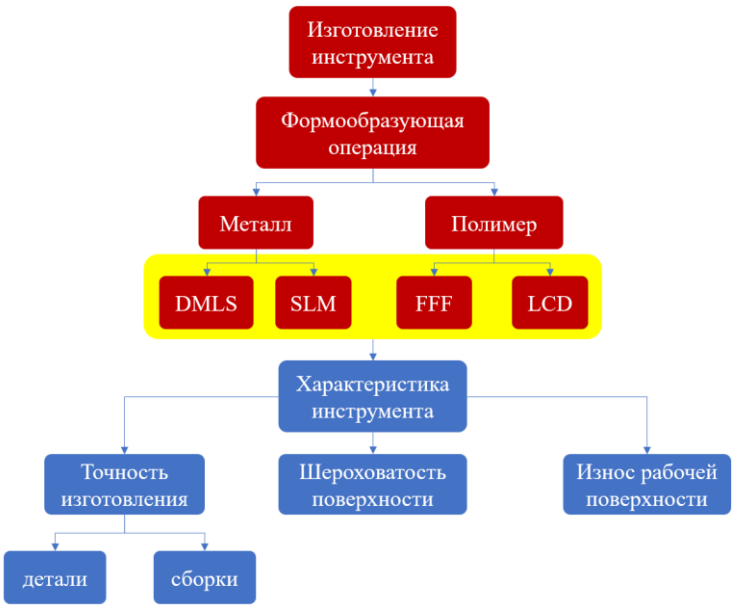


Рис. 3. Схема проведения исследования

По результатам любой 3D-печати выполнялась пост-обработка, заключавшаяся в удалении поддержек, промывки (только LCD), снятии заусенцев и фасок, нарезании резьбы и/или установки резьбовых вставок и втулок, термообработки (только DMLS) и другое.

III. Определение точности изготовления

После 3D-печати прототипов выполнялся контроль размеров отдельных деталей, а также сборочных узлов (рис. 4). Для получения информации о размерах изделий применялся бесконтактный метод 3D-сканирования на основании структурированного подсвета. Перед проведение сканирования все объекты покрывались матирующим самоисчезающим спреем Ateco Ghost для получения равномерного коэффициента отражающей способности.

После получения серии сканов 3D-сканером RangeVision Neopoint, их обработки, финального совмещения и экспорта полигональной (stl-) модели в программе RV 3D Studio, полученный результат сравнивался с исходной CAD-геометрией в программе GOM Inspect. Точность финального совмещения для матрицы составила 0.036 мм, а для подсборки нижней части инструмента составила 0.085 мм. Совмещение CAD-модели и фактической полигональной stl-модели выполнялось в два этапа. На первом этапе выравнивание осуществлялось либо по одной точке, либо по трём точкам соответствия, выбранных на каждой из 3D-моделей. На втором этапе выполнялось локальное наилучшее совпадение геометрии. Точность совмещения для матрицы на первом и втором этапах составила 0.0524 мм. Точность совмещения для подсборки на первом этапе составила 0.61 мм, а на втором этапе составила 0.1618 мм.

Было установлено, что размеры по боковой внешней поверхности матричной вставки лежат вне симметричного допуска ±0.1 мм, соответствующего точности 3D-печати, напоминает дефект типа овальности (рис. 4а).

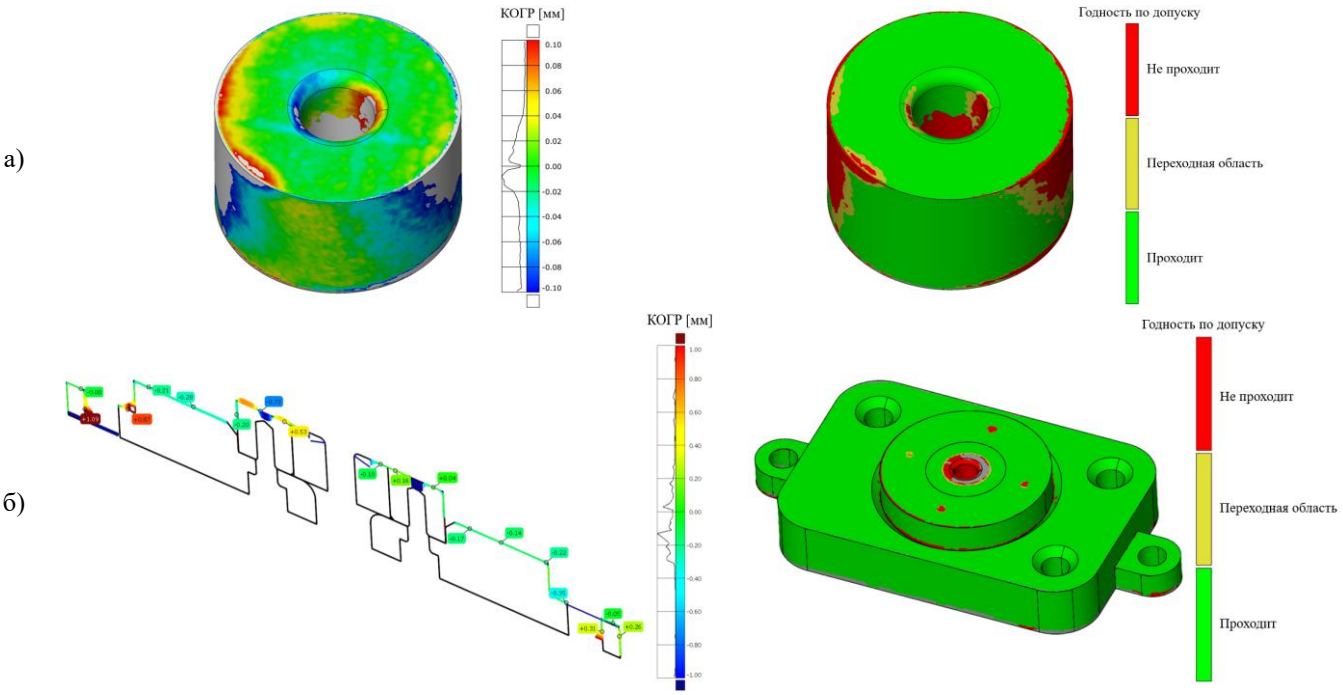
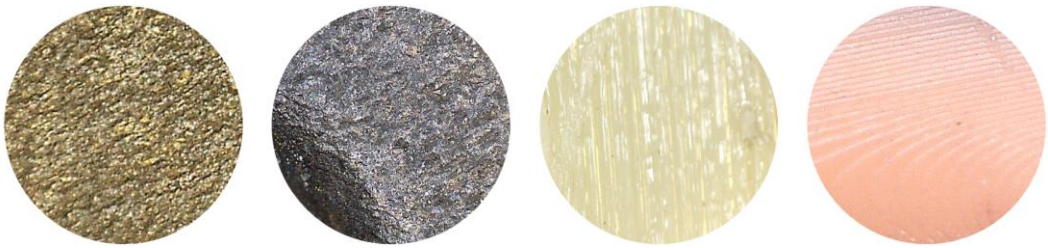


Рис. 4. Схема проведения исследования: матричная вставка (а) и подсборка нижней части инструмента (б)

В центральной части (заходная часть матрицы и цилиндрическая стенка) красная область указывает на отсутствие информации о геометрии по результатам 3D-сканирования. Подсборка нижней части инструмента при установленном симметричном допуске ±1 мм почти вся проходит по полю допуска. Однако, по гистограмме распределения размеров на карте отклонения геометрических размеров (КОГР), построенной для центрального сечения, видно, что разность размеров между исходной CAD-моделью и фактической полигональной stl-моделью лежит в интервале ±0.2, что превышает точность печати.

IV. Шероховатость и износ рабочей поверхности

Послойное построение объектов оставляет характерную для метода 3D-печати топографию или текстуру (рис. 5). Для её описания требуются специальные инструменты, определяющие шероховатость поверхности (ГОСТ 2789-73, ГОСТ Р ИСО 4287-2014). Так согласно ГОСТ Р 70117-2022 при плоском шлифовании, которое используется, как одна из итоговых операций при изготовлении деформирующих инструментов, шероховатость поверхности изменяется в диапазоне 6.3…0.05 мкм. Шероховатость напечатанных прототипов изменяется в другом диапазоне. Это обстоятельство позволяет думать о том, что и технологические параметры процесса обработки давлением будут не соответствовать тем, которые приняты при разработке традиционных технологий, т.е. с применением инструментов, полученных механообработкой.



Вид поверхности				
Материал	ВТ6	AlSi10Mg	ПЭТГ	БФП
Технология	DMLS		FFF	LCD

Рис. 5. Фотографии поверхностей прототипов

При определении шероховатости применялись измеритель шероховатости TR210 (Time Group Inc., Китай), на длине измерения 0.25 и 0.8 мм, и профилометр MarSurf M400 (Mahr GmbH, Германия), на длине измерения 2.5 мм. Шероховатость измерялась на поверхностях прототипов отдельных деталей-представителей инструмента из ВТ6, ПЭТГ, БФП и КПФП. По результатам измерений получены значения шероховатости, представленные в

табл. 2. Определение износа поверхности осуществлялось на основании результатов экспериментов по схеме «стержень по пластине» на трибометре CSM TRB (рис. 6). В процессе эксперимента субстрат получал трансляционное возвратно-поступательное перемещение для общей длины пути, равной 500 м.

ТАБЛИЦА 2
ЗНАЧЕНИЕ ШЕРОХОВАТОСТИ ПОВЕРХНОСТИ ПРОТОТИПОВ

Материал	Ra			Rz		
	0.25	0.8	2.5	0.25	0.8	2.5
ВТ6	1.423	3.679	6.629	7.559	20.260	22.180
ПЭТГ	0,657	1.331	6.765	3.058	6.428	22.730
БФП	0.286	0.744	0.310	1.442	4.553	1.629
КПФП	1.095	1.267	0.641	5.054	9.244	3.017

После прохождения заданного пути материал субстрата и стержня был снят с испытательной машины. Определение износа проводилось по формуле Арчарда, уравнение (2), для которого требуется характеристика получившейся канавки по данным профилометра, а также полусферического наконечника стержня, на основании измерений под оптическим микроскопом. На рис. 7 показан износ полусферической части стержня под электронным (а) и цифровым (б) микроскопами.

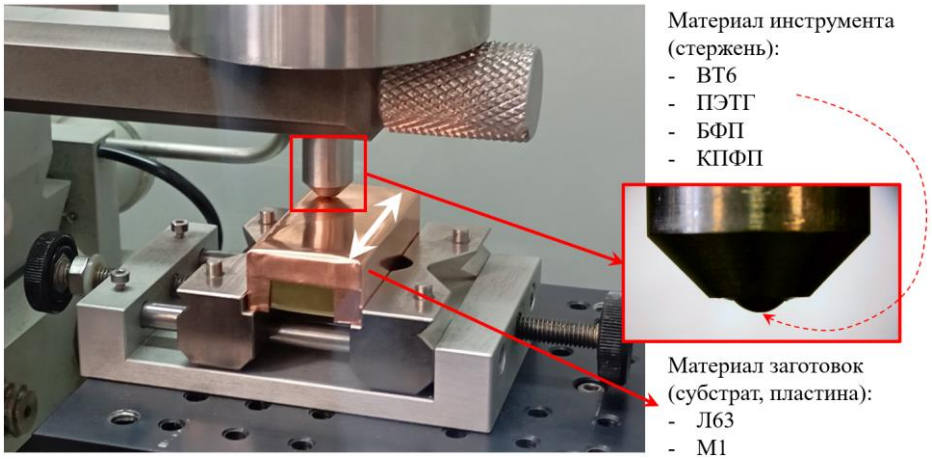
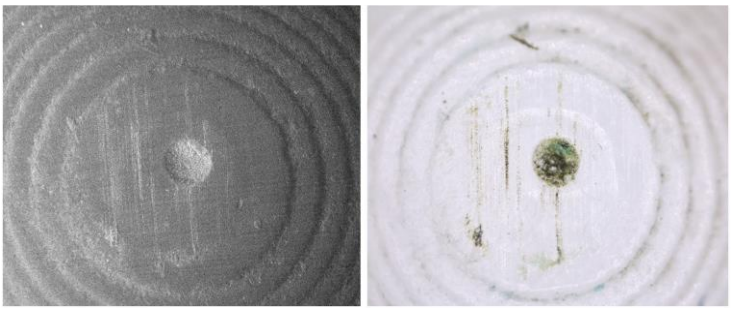


Рис. 6. Испытания на износ по схеме «стержень по пластине»

$$W = \frac{KF_N}{H}S \tag{2}$$

где W – объём износа, [мм^3]; K – коэффициент износа, [-]; F_N – сила, создаваемая грузом, 1…5 [Н]; S – длина пути, [мм]; H – твёрдость более мягкого материала, [Н/мм^2].



а) б)

Рис. 7. Визуальный контроль износа под микроскопом: электронным (а) и цифровым (б)

V. Обсуждение результатов

Полученные методами 3D-печати деформирующие инструменты позволяют изготавливать миниатюрные детали. При помощи измерителя шероховатости и профилометра дана оценка шероховатости прототипов.

Применяя трибометр и микроскопы дана оценка износа рабочих поверхностей инструмента, а именно заходной части матрицы для изготовления осесимметричной детали. Устранение неточностей изготовления могут привести к минимизации преимуществ технологий аддитивного производства, по сравнению с традиционными технологиями, поэтому требуется заранее учитывать допуски на обработку.

VI. Выводы и заключение

Аддитивные технологии являются многообещающим способом изготовления инструмента с высокой точностью. Поставленные задачи по 3D-печати из металлических и полимерных материалов инструментов для деформирования листовой заготовки могут быть выполнены, но нуждаются в обосновании оценки точности изготовления и эксплуатационной характеристики, влияющих на точность и качество детали, и настройку технологических параметров процесса деформирования, соответственно. Применяя известные методы оценки поверхности и износа для макроуровня необходимо также обосновать распространение этих результатов на микроуровень, что требует также практического подтверждения на специализированном оборудовании, работающего на микроразмерных образцах.

Список литературы

1. Hall, E. O. (1951). The Deformation and Ageing of Mild Steel: III Discussion of Results. Proceedings of the Physical Society. Section B, 64(9), 747–753. doi:10.1088/0370-1301/64/9/303
2. Petch, N.J. The cleavage strength of polycrystals. J. Iron Steel Inst. 1953, 174, 25–28
3. Geiger, M., Kleiner, M., Eckstein, R., Tiesler, N., & Engel, U. (2001). Microforming. CIRP Annals, 50(2), 445–462. doi:10.1016/s0007-8506(07)62991-6
4. Fu, M. W., & Chan, W. L. (2014). Micro-scaled Products Development via Microforming. Springer Series in Advanced Manufacturing. doi:10.1007/978-1-4471-6326-8
5. Jiang Z., Zhao J., Xie H. Microforming Technology: Theory, Simulation, and Practice, Academic Press, 2017. — 458 p. — ISBN: 978-0-12-811212-0.
6. Xu, J., Wang, X., Wang, C., Yuan, L., Chen, W., Bao, J., … Guo, B. (2020). A Review on Micro/Nanoforming to Fabricate 3D Metallic Structures. Advanced Materials, 33(6), 2000893. doi:10.1002/adma.202000893
7. Hansen, N. (2004). Hall–Petch relation and boundary strengthening. Scripta Materialia, 51(8), 801–806. doi:10.1016/j.scriptamat.2004.06.002
8. Smith W.F., Hashemi J. Foundations of Materials Science and Engineering, 7th Edition. — McGraw-Hill, 2023. — 1137 p. — ISBN 978-1260597709.
9. Wang, C., Guo, B., & Shan, D. (2014). Friction related size-effect in microforming – a review. Manufacturing Review, 1, 23. doi:10.1051/mfreview/2014022
10. Kuhfuss, B. *et al.* (2020). Micro Forming Processes. In: Vollertsen, F., Friedrich, S., Kuhfuß, B., Maaß, P., Thomy, C., Zoch, HW. (eds) Cold Micro Metal Forming. Lecture Notes in Production Engineering. Springer, Cham. doi: 10.1007/978-3-030-11280-6

УДК 004.925.8
DOI: 10.25206/978-5-8149-3873-2-2024-418-427

Аддитивная технология изготовления персонализированных каркасов сухих электродов для интерфейсов мозг-компьютер
Additive manufacturing technology for personalized dry electrode headsets for brain-computer interfaces

Д. А. Чирухин1, К. В. Рябинин2
1 Пермский государственный национальный исследовательский университет, Пермь, Россия
2 Институт вычислительной астрономии Астрономического центра Университета г. Хайдельберг, Хайдельберг, Германия

D. A. Chiruhin1, K. V. Ryabinin2
[1]Perm State University, Perm, Russia
[2]Astronomisches Rechen-Institut, Centre for Astronomy of Heidelberg University, Heidelberg, Germany

Аннотация. Комфорт в ношении и анатомическая точность каркаса электродов играют важную роль при организации человеко-машинных интерфейсов типа мозг-компьютер, основанных на электроэнцефалографии. В статье предложена технологическая цепочка и соответствующий программный инструментарий для обеспечения полного цикла изготовления каркаса сухих электродов под конкретного человека. Цепочка включает 3D-сканирование головы, интерактивное редактирование схемы расстановки электродов и автоматическую генерацию модели разборного каркаса, оптимизированной для печати на 3D-принтере. Работа цепочки проверена на практике. Точность расстановки электродов оценена путем сопоставления с каркасом электродов от профессионального медицинского оборудования и признана достаточной для задач организации интерфейсов мозг-компьютер.

***Ключевые слова*:** Интерфейс мозг-компьютер, 3D-сканирование, 3D-печать, сухие электроды, электроэнцефалография, когнитивная графика.

Abstract. Comfort in wearing and anatomical accuracy of the electrodes head cap play an important role in the organization of brain-computer human-machine interfaces based on electroencephalography. This paper proposes a technological chain and corresponding software tools for a complete cycle of dry electrodes head cap manufacturing for a particular person. The pipeline includes 3D scanning of the head, interactive editing of the electrodes' placement scheme, and automatic generation of a collapsible head cap model optimized for 3D printing. The performance of the pipeline has been validated in practice. The accuracy of electrodes' placement has been evaluated by comparison with the head cap from professional medical equipment and is recognized as sufficient for the tasks of organizing brain-computer interfaces.

***Keywords*:** Brain-computer interface, 3D scanning, 3D printing, dry electrodes, electroencephalography, cognitive graphics.

I. Введение

При проведении исследований в области человеко-машинного взаимодействия, а именно в области построения интерфейсов мозг-компьютер (англ. Brain Computer Interface, BCI), удобно использовать т.н. сухие электроды, непосредственно прилегающие к коже головы и не требующие наполнения токопроводящим гелем [1-2]. Для их размещения необходим жесткий каркас, фиксирующийся на голове наподобие каски. Существуют стандартные модели таких каркасов, но они рассчитаны на среднестатистический размер головы и часто плохо подходят для конкретного человека, не обеспечивая надежной фиксации электродов в нужных местах (соответствующих выбранной схеме их расстановки). В рамках предыдущих исследований [3] был разработан прототип системы генерации 3D-модели каркаса для сухих электродов, имеющий, однако, ряд недостатков. В частности, получаемый каркас мог иметь только схему расстановки электродов 10-20, а процесс построения модели головы человека был слабо автоматизирован. В данной работе предлагаются улучшения разработанной системы, включающие программные средства для описания схем расстановки электродов, управления их использованием в процессе генерации каркаса, а также автоматизации процесса построения модели головы на основе 3D-скана. Генерируемая системой модель каркаса верифицирована путем сопоставления с фабричным каркасом, входящим в состав профессионального медицинского оборудования для электроэнцефалографии (ЭЭГ).

II. Разработанный ранее прототип системы генерации моделей каркасов

В рамках предыдущей работы был проведен обзор существующих решений, который показал, что на данный момент не существует системы для полной персонификации каркаса. С целью решения этой проблемы был разработан прототип для генерации модели каркаса. Он имел следующий набор возможностей:

1. Создавать эллипсоидную модель головы человека на основе расстояний между ключевыми точками: Назион (самая глубокая точка носолобной впадины), Инион (затылочный бугорок), Ушные точки, или на основе обхватов головы по трем осям.
2. Рассчитывать координаты электродов схемы размещения 10-20 для данной модели головы.
3. Размещать модели гнезд электродов, формируя каркас.
4. Соединять модели гнезд мостиками.
5. Разделять мостики по форме соединения типа “Ласточкин хвост” для последующей сборки каркаса из отдельных небольших частей, оптимизированных для печати на 3D-принтере с минимум «поддержек».
6. Размещать на мостиках текстовые метки для упрощения сборки каркаса.
7. Экспортировать детали в отдельные STL файлы для облегчения печати.

У прототипа были некоторые недостатки, а именно:

1. Невозможность изменения модели расстановки электродов.
2. Необходимость ручного измерения расстояний между ключевыми точками, даже при наличии готового 3D-скана головы.

Прототип разработан на языке Python с использованием геометрического ядра FreeCAD [4].

III. Моделирование головы

В предыдущей версии системы использовалась эллипсоидная модель головы. Известно, что такая модель головы хорошо подходит при расстановке электродов [5], к тому же она упрощает многие вычисления. Для генерации эллипсоидной модели предлагалось два варианта: измерить обхват головы человека по трем осям или сделать 3D-скан головы и измерить расстояния между ключевыми точками напрямую на нем. В первом случае необходимо было решать систему уравнений периметров эллипсов (1) для нахождения длин полуосей эллипсоида,

$$L \approx \pi \left[3(a+b) - \sqrt{(3a+b)(a+3b)}\right], \quad (1)$$

где a, b – длины полуосей, во втором – введенные значения использовались непосредственно как длины осей эллипсоида. Первый вариант не требовал дополнительного оборудования, но его точность сильно зависела от точности измерений длин обхватов, а также аппроксимативная формула периметра эллипса добавляла небольшую погрешность. Второй вариант был более точным, но его точность также зависела от точности измерений, к тому же, несмотря на наличие 3D-скана, приходилось измерять размеры вручную с использованием интерактивных инструментов системы автоматизированного проектирования.

В данной работе представлено развитие второго варианта, при котором построение модели производится непосредственно на основе 3D-скана без промежуточных действий со стороны пользователя, что увеличивает степень автоматизации процесса и позволяет избежать ошибок при измерениях.

Предлагается следующий алгоритм построения модели:

1. Загрузить 3D-скан.
2. Определить положение ключевых точек на скане.
3. Построить систему координат, основанную на положении ключевых точек.
4. Провести кластеризацию точек на 3D-скане.
5. Отфильтровать лишние кластеры.
6. Построить эллипсоид по облаку точек.

Этот алгоритм можно разбить на 3 основных этапа: выделение ключевых точек на скане и построение системы координат (1-3), выделение позиций электродов на скане (4-5) и нахождение длин полуосей эллипсоида как принципиальных компонент полученного облака точек (6).

Первый этап в настоящий момент полностью основан на сторонней программной системе [6-8]. Она реализована в среде MATLAB и предлагает набор функций для загрузки модели 3D-скана, получения двумерных проекций всех ключевых точек на скане, нахождения их координат, преобразования координат точек двумерной проекции в глобальную систему координат скана, а также построение новой системы координат головы, оси абсцисс и ординат которой проходят через ключевые точки, а ось аппликат перпендикулярна им и указывает на

макушку. Программная система предлагает средства визуализации полученных результатов, в частности визуализацию построенной системы координат головы можно увидеть на рисунке 2.

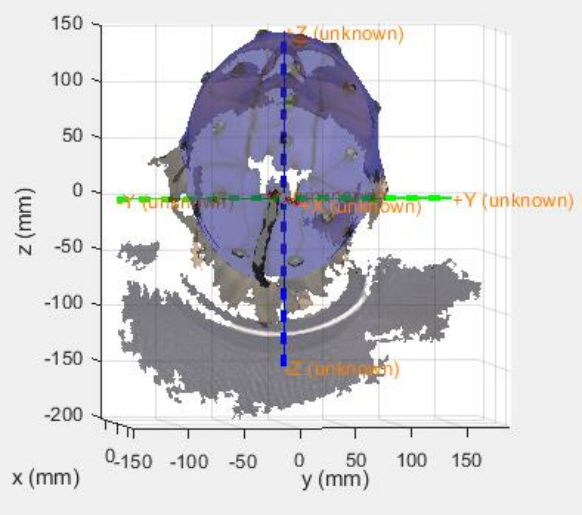


Рис. 2. Система координат на 3D-скане головы

MATLAB имеет API для Python, что позволило интегрировать данную программную систему в разрабатываемый прототип. Проблемой, однако, является то, что MATLAB не распространяется бесплатно, что усложнит распространение разрабатываемого прототипа.

Второй этап основан на результатах исследований Chen S. et al. [8], Shirazi S. et al. [9], Mazzonetto I. Et al [10], в которых была проведена локализация электродов на скане путем кластеризации. В данной работе проводится кластеризация вершин 3D-модели скана методом DBSCAN [11], который основан на плотности точек. После применения метода центры кластеров будут указывать на местоположение электродов. Данный метод плохо применим для схем с высокой плотностью размещения электродов, но в нашем случае для задач BCI планируется в основном использование стандартной схемы 10-20 и производных от нее, так что это не является проблемой.

Далее преобразуем координаты полученных точек к найденной на первом этапе системе координат, что позволит на третьем этапе алгоритма упростить нахождение эллипсоида. Среди полученных точек избавляемся от тех, что выходят за границы, заданные ключевыми точками, оставляя те, что лежат внутри области между ключевыми точками. Для этого строим простые плоскости отсечения, определяющие область между ключевыми точками, и оставляем только те точки, которые находятся внутри этой области. Основная цель отсечения – удалить кластеры, указывающие на посторонние объекты, попавшие на скан. Кроме того, избавляемся от выбросов: слишком больших и маленьких кластеров. В качестве диапазона используется одно стандартное отклонение от среднего размера кластера. Основная цель данного этапа – удалить кластеры, указывающие на положение посторонних объектов, попавших в область отсечения («хвост» из проводов от электродов), и кластеры, указывающие на артефакты сканирования. На выходе получаем облако точек, описывающих положение электродов на поверхности головы на скане.

На третьем этапе для полученного облака точек необходимо найти эллипсоид, который бы наилучшим образом его аппроксимировал. Для данной задачи можно воспользоваться методом главных компонент (англ. Principal Component Analysis, PCA [12]). Данный метод позволяет определить направления максимальной дисперсии данных. Мы можем спроецировать точки на эти направления, найти главные компоненты и соответствующие им дисперсии. Направление осей эллипсоида соответствует направлению главных компонент, а длина осей вычисляется как квадратный корень соответствующих дисперсий.

IV. Редактор описания схем расстановки электродов

В зависимости от задач при создании BCI могут понадобиться каркасы с различными схемами расстановки. Существует множество стандартных схем расстановки, кроме того, могут понадобиться гибридные или уникальные схемы. В связи с этим, разрабатываемая система моделирования должна предоставлять редактор, позволяющий в интерактивном режиме задать любую схему расстановки электродов. Важными требованиями к такому редактору являются эргономичность и когнитивность графики, чтобы обеспечить возможность использования людьми, не обладающими квалификацией ИТ-специалистов (целевая аудитория разрабатываемой

системы включает, в том числе, нейрофизиологов, заинтересованных в проведении исследований в области BCI с использованием ЭЭГ).

Для обеспечения когнитивности графики редактора, в первую очередь необходимо разработать концептуальную модель схемы расстановки, определить компоненты и связи между ними. Концептуальную модель удобно представить в виде прикладной онтологии [13].

Для построения онтологии редактора выделены и описаны следующие понятия:

1. Точка – абстрактный объект, имеющий название.
2. Электрод – точка, на месте которой при генерации каркаса будет размещено гнездо для электрода.
3. Крепление – элемент каркаса, позволяющий собирать цельный каркас из деталей. Может быть двух типов: гнездо-гнездо и гнездо-мостик-гнездо.
4. Мостик – элемент каркаса, имеющий относительную длину и соединяющий две точки. Может быть реальным (имеющим крепление) или мнимым (не имеющим крепления).
5. Арка – список мостиков.

Между вышеперечисленными понятиями выделены связи следующих типов:

1. is_a – связь дочернего понятия с родительским.
2. a_part_of – связь “части” с “целым”.
3. has – связь “владелеца” с “принадлежащей” ему сущностью.

Связь “a_part_of” соединяет точку с мостиком и мостик с аркой. У мостика может быть только две связанные с ним точки. Связь “has” соединяет мостик с креплением. Связь “is_a” выходит из понятия “Электрод” в понятие “Точка".

На рисунке 3 приведена прикладная онтология, описывающая связи между ключевыми понятиями, построенная в программной системе Ontolis [14].

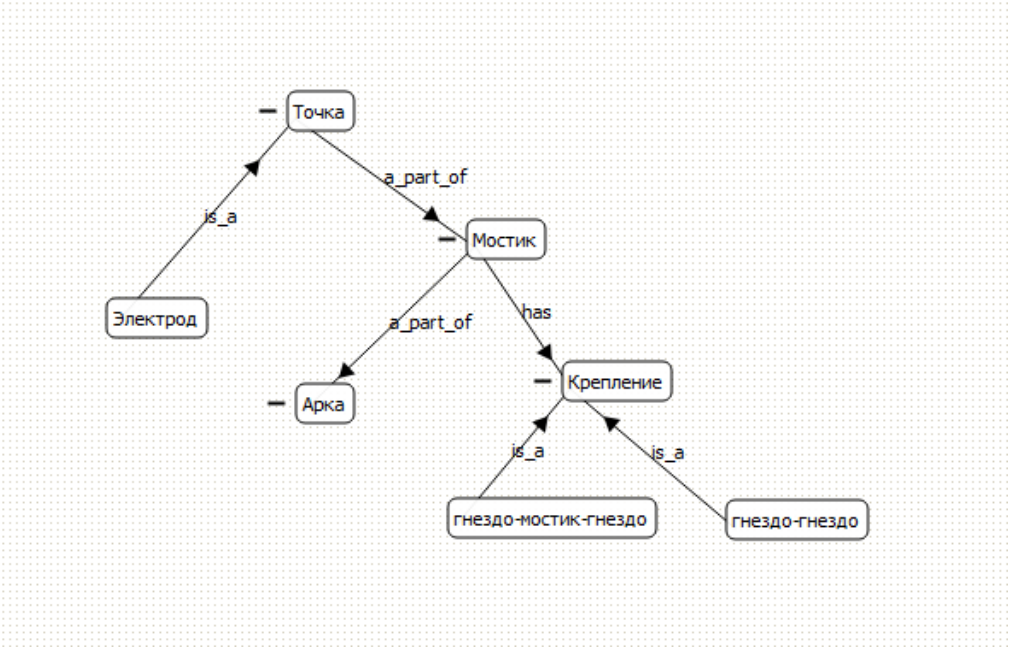


Рис. 3. Онтология понятий схемы для размещения электродов

Для реализации редактора необходимо определить соответствующие программные средства. В качестве языка выбран Python, так как существующая система уже написана на нем. Для удобства использования редактор должен предоставлять графический пользовательский интерфейс, так что необходимо выбрать соответствующий фреймворк для организации интерфейсов. В таблице 1 приведено сравнение популярных фремворков для разработки настольных приложений для Python.

ТАБЛИЦА 1
СРАВНЕНИЕ ФРЕЙМВОРКОВ PYTHON ДЛЯ ГРАФИЧЕСКИХ ПРИЛОЖЕНИЙ

Фреймворк	**Вхождение в стандартную библиотеку**	**Хорошая документация**	**Мощные возможности для рисования на холсте**	**Простота использования**	**Свободная лицензия**
TKinter	+	+	+	+	+
Kivy	-	+	-	+	+
PyQT	-	+	+	-	+/-
wxPython	-	+	+	-	+
PySide	-	+	+	+	+

Для реализации редактора был выбран фреймворк TKinter, так как он входит в стандартную библиотеку, хорошо документирован, имеет мощные возможности для использования виджета холста (англ. Canvas), прост в использовании и распространяется свободно.

При реализации были выделены 3 основных класса: Point, Electrode и Bridge. Из них создается модель схемы. Класс Point имеет поля x, y для размещения на холсте и radius для задания радиуса точки для отображения, id для идентификации точки, name – смысловое поле содержащее имя точки на схеме и base – логическое поле, которое отображает, является ли объект базовой точкой или нет. Класс Electrode наследуется от класса точки, но не добавляет никаких новых полей. Объекты данного класса имеют другое значение поля radius и не имеют поля base. Объекты данного класса будут иметь другое отображение на холсте, но логически данный класс нужен только для семантического разграничения точек и электродов. У класса Bridge присутствуют поля start_point, end_point – это ссылки на точки/электроды, которые соединяют мостик; bridge_id для идентификации объектов; length для указания длины мостика относительно длины арки; arc для указания, к какой арке принадлежит мостик; connection_type для указания типа соединения (гнездо-гнездо, гнездо-мостик-гнездо); imaginary для указания реальный мостик или мнимый. Поля length, arc, connection_type и imaginary используются при расчете координат гнезд, соединении гнезд мостиками и генерации креплений.

Объекты данных классов должны иметь набор ограничений.

Ограничения, накладываемые на значения полей объектов классов Point/Electrode:

1. Имена точек и электродов уникальны.
2. В системе всегда должны быть точки с именами Nasion, Inion, LeftEar, RightEar, которые создаются при создании нового документа и не могут быть удалены или переименованы.

Ограничения, накладываемые на значения полей объектов класса Bridge:

1. У мостика может как быть, так и не быть арка.
2. У мостика с аркой поле length должно содержать число в отрезке [0, 100].
3. Сумма значений полей length всех мостиков любой арки должна быть равна 100.
4. Мостик не может быть в арке, в которой он не имеет общих крайних точек с другими мостиками, если он не единственный
5. Если у мостика поле imaginary имеет значение False, то его поле connection_type должно быть отлично от None.

Соответствующий программный модуль представляет собой графический редактор для схем расстановки электродов. Основная часть интерфейса – это холст для рисования схемы. Справа от холста (рис. 4) представлена панель свойств выбранного элемента, а ниже – список инструментов: “Выбор”, “Точка” “Электрод”, “Мостик”. Для инструментов настроены горячие клавиши 1-4 соответственно.

При выборе инструментов “Точка” и “Электрод” и нажатии на холст на месте нажатия появляется большой белый круг (при выборе электрода) или маленький черный круг (при выборе точки). При выборе инструмента “Мостик” и нажатии на любой из кругов он подсветится синим цветом; затем, при нажатии на любой другой, с которым данный круг еще не связан, между ними создастся мостик.

При выборе инструмента “Выбор” и нажатии на любой из кругов он подсветится красным цветом, а в панели свойств отобразится текстовое поле для свойства “Имя”. При вводе имени в это поле оно сразу отобразится на холсте. Круги можно перемещать по холсту при помощи мыши (жестом Drag’n’Drop).

При нажатии на мостик он также подсветится красным цветом, и отобразятся его свойства. На текстовое поле для свойства “Длина” действует ограничение на ввод только чисел и только в диапазоне от 0 до 100, при вводе текста, не подходящего под ограничение, значение вернется к исходному. Для задания свойства “Арка” используется комбобокс с возможностью дополнения списка. В качестве вариантов предлагаются арки соседних

мостиков. Если мостик не имеет арки, то его длина не задается. Если у мостика не задана длина, или она равна 0, это значение не отображается на холсте. Для задания свойства “Мнимый” используется чекбокс. Изначально установленное значение – False, при данном значении среди свойств появляется элемент для ввода поля “Тип крепления”. В качестве элемента используется непополняемый список с вариантами “Гнездо-гнездо” и “Гнездо-мостик-гнездо”. Мнимый мостик отображается пунктиром, реальный – сплошной линией.

При нажатии клавиши Delete выбранный объект удаляется. Если были удалены электрод или точка, удаляются все инцидентные им мостики. В верхнем меню представлены кнопки для действий “Создать новую схему”, “Загрузить”, “Сохранить”, “Сохранить как”. При вращении колесика мыши происходит зуммирование холста, при зажатии колесика мыши и ее перемещении происходит панорамирование (перемещение по холсту). Схема сохраняется в файл в формате JSON.

На рисунке 4 приведен интерфейс редактора после создания новой схемы.

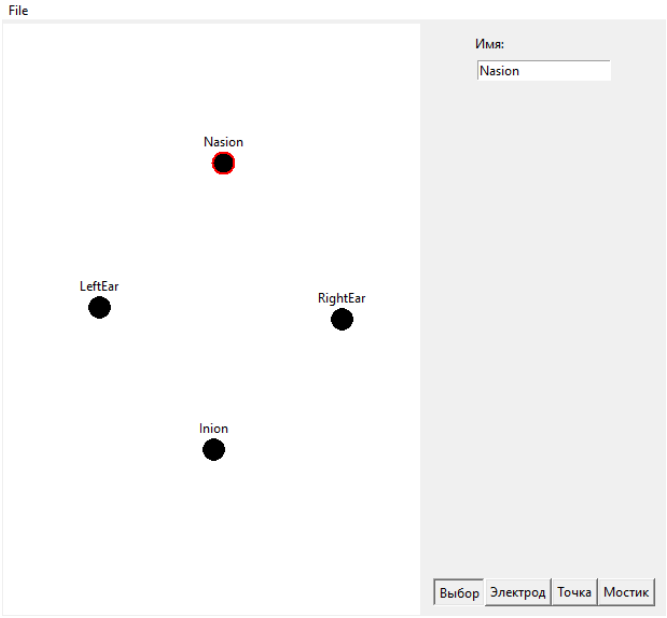


Рис. 4. Интерфейс редактора схем расстановки

С использованием разработанного редактора были созданы схемы 10-20 и ее модификации, которые представлены на рисунках 5–7.

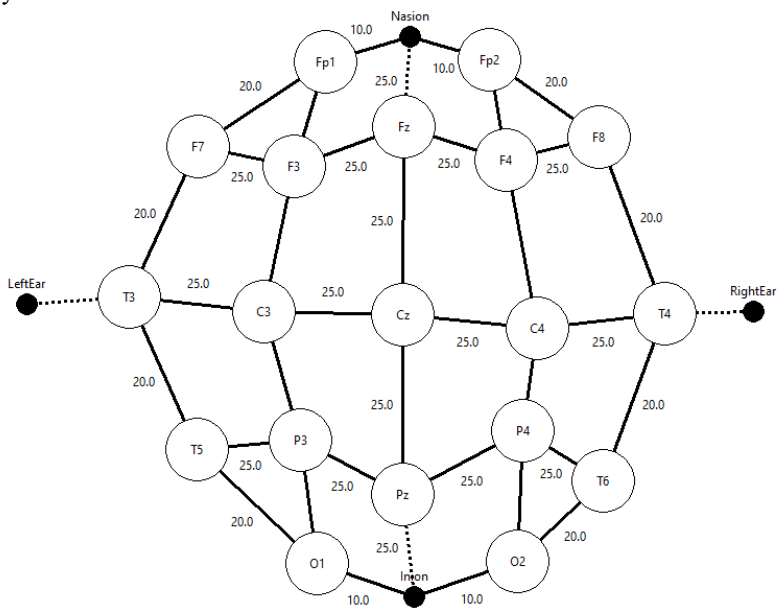


Рис. 5. Схема 10-20 в среде разработанного редактора

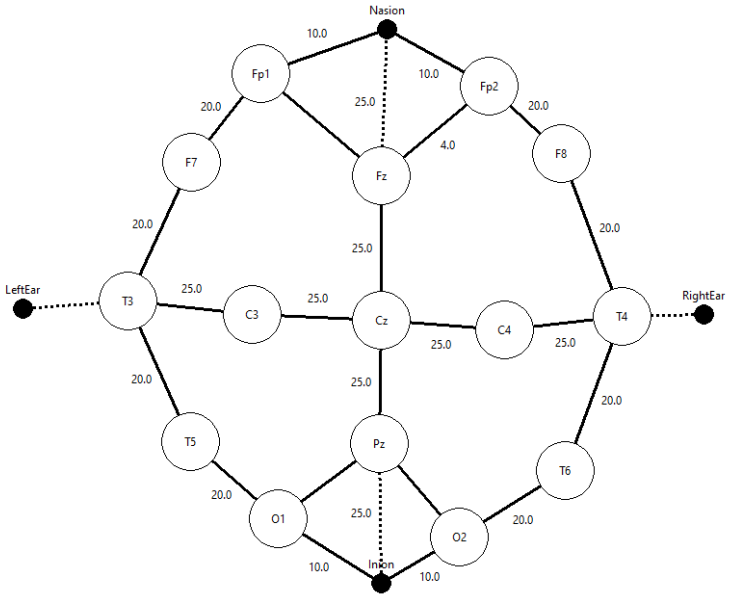


Рис. 6 Схема 10-20 без электродов F3, F4, P3, P4 в среде разработанного редактора

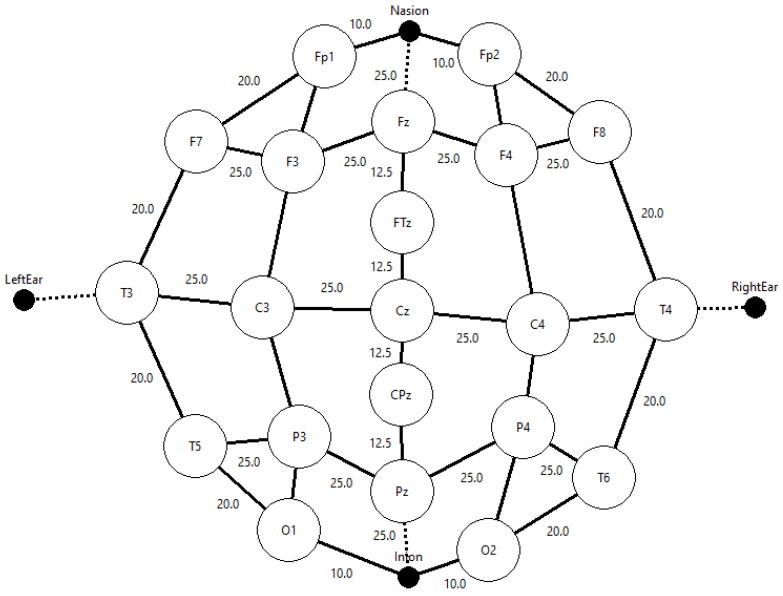


Рис. 7. Схема 10-20 с электродами FTz, CPz в среде разработанного редактора

V. Генерация каркаса

В предыдущей версии системы генерация модели каркаса опиралась на схему 10-20, которая была реализована напрямую в программном коде. Теперь генерация должна опираться на схему, составленную пользователем. Процесс генерации можно разделить на следующие основные этапы: расчет координат гнезд для электродов, их размещение, соединение гнезд мостиками, разделение мостиков по форме крепления и размещение на мостиках текстовых меток для сборки. Размещение моделей гнезд, добавление креплений и текстовых меток не зависит от схемы размещения, поэтому никак не изменятся.

Процесс расчета координат основывается на арках, содержащих ключевые точки. В любой схеме расстановки выделяются 4 ключевые арки: продольная арка от Nasion к Inion через макушку, перпендикулярная ей арка от LeftEar к RightEar (точки перед левой и правой ушными раковинами), еще две арки обхватывают голову от Nasion к Inion через точки LeftEar и RightEar. Данные арки могут быть найдены по следующим условиям:

1. Арка содержит точки Nasion, Inion и не содержит LeftEar, RightEar.
2. Арка содержит LeftEar, RightEar и не содержит Nasion, Inion.

3. Арка содержит Nasion, Inion и LeftEar.
4. Арка содержит Nasion, Inion и RightEar.

Ключевые точки могут выполнять роль метки, если они связаны с электродом мостиком с длиной, равной нулю. В таком случае ключевая арка может содержать не ключевую точку, а электрод, для которого данная точка является меткой. Для данной схемы точки LeftEar и RightEar выполняют роль меток, так как связаны с электродами T3 и T4 мостиками с нулевой длиной. Для электродов, расположенных на данных четырех арках, позиции считаются по формуле (2) параметрического уравнения эллипса как точки на эллипсах в главных сечениях эллипсоида.

8.
$$\begin{cases} x = a\ cos\ (t) \\ y = b\ cos\ (t)' \end{cases} \quad (2)$$

где a, b – длины полуосей, t – параметр.

Координаты точек, лежащих на остальных арках, могут быть рассчитаны, опираясь на уже полученные точки. Для каждой из них находится арка среди инцидентных мостиков, на арке находятся точки с уже известными координатами по обе стороны от текущей. Если таких точек нет, то выдается ошибка с пояснением, для какой точки не получилось найти опорные. Если такие точки удается найти, то находятся относительные расстояния от данной точки до найденных и вычисляется пропорция, в которой данная точка делит часть арки между двумя известными. После чего координаты данной точки могут быть найдены путем интерполяции на поверхности эллипсоида. Для решения задачи интерполяции используется метод minimize пакета scipy.optimize [15].

Общий алгоритм расчета координат приведен в листинге 1 с использованием псевдокода.

Листинг 1. Алгоритм расчета координат гнезд

```
def calculate_coordinates(scheme, ellipsoid):
    nasion, inion, left_ear, right_ear = get_key_points(scheme)
    key_arcs = get_key_arcs(scheme)
    coordinates = {}
    for arc in arcs:
        for point, length in arc.enumerate_points():
            coordinates[point] = arc.get_coordinates(ellipsoid, length)
    for point in scheme.get_points_not_in_list(coordinates):
        coordinate_calculated = False
        for arc in point.get_arcs():
            point1, length1 = arc.get_known_neighbour_1(point)
            point2, length2 = arc.get_known_neighbour_2(point)
            if point1 and point2:
                coordinate_calculated = True
                coordinates[point] = interpolate(ellipsoid, coordinates[point1],
coordinates[point2], length1 / (length1 + length2))
                break
        if not coordinate_calculated:
            raise Exception()
    return coordinates
```

Для соединения гнезд мостиками нужно только заменить список связей электродов для схемы 10-20 на список мостиков используемой схемы.

VI. Результаты

Для созданных схем расположения электродов были сгенерированы модели каркасов, которые можно увидеть на рисунках 8–10. Из рисунков видно, что все особенности схем учтены и гнезда расположены и соединены в соответствии с ними.

В таблице 2 можно увидеть ошибку расположения для наиболее отклоняющихся от положения на 3D-скане гнезд при использовании старой и новой версии системы. Из таблицы видно, что ошибка размещения значительно уменьшилась. Стоит учитывать, что даже ошибки старой версии были допустимы для задач BCI. Для более точного тестирования необходимо собрать больше 3D-сканов и провести больше подобных тестов.

ТАБЛИЦА 2
ОШИБКА РАСПОЛОЖЕНИЯ ЭЛЕКТРОДОВ

Электрод	**Отклонение на старой версии, см**	**Отклонение на новой версии, см**
F4	0.5	0.42
Fz	0.6	0.46
T4	0.68	0.5
T8	0.71	0.45

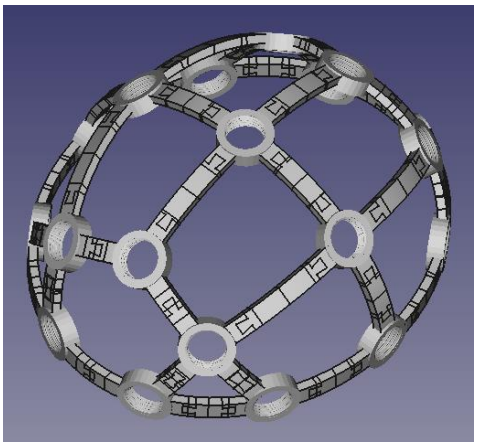


Рис. 8. Модель для схемы 10-20 в среде FreeCAD

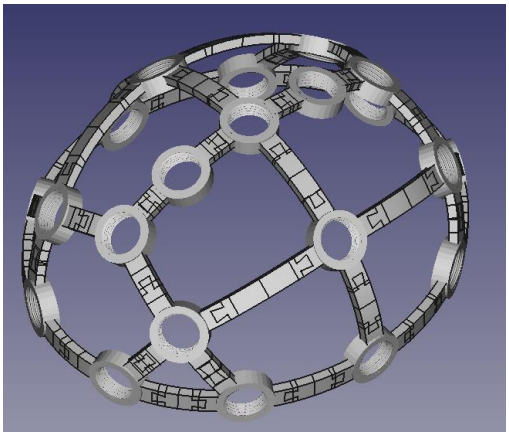


Рис. 10. Модель для схемы 10-20 с электродами FTz, CPz в среде FreeCAD

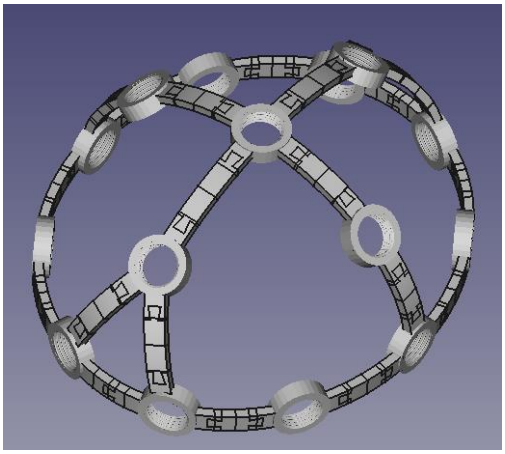


Рис. 9. Модель для схемы 10-20 без электродов F3, F4, P3, P4 в среде FreeCAD

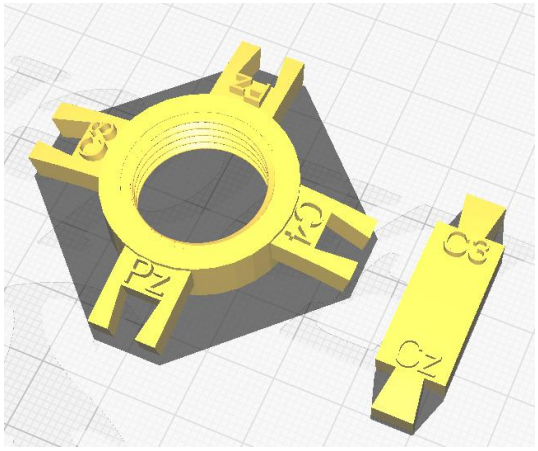


Рис. 11. STL модели, готовые к печати в среде UltiMaker Cura

Сгенерированные модели экспортируются в набор файлов в формате STL, готовых для 3D-печати. Пример STL моделей можно увидеть на рисунке 11.

VII. Заключение

В данной работе предложено развитие разработанной ранее системы генерации каркаса сухих электродов для ЭЭГ. Предложен и реализован метод для автоматизации построения модели головы человека с целью размещения электродов на основе 3D-скана. Предложена онтологическая модель визуального редактора схем для расстановки электродов; редактор реализован в виде отдельного программного модуля в составе разработанной системы. Будучи основанным на формальной модели соответствующей предметной области, интерфейс разработанного редактора обладает свойствами когнитивной графики и интуитивно понятен пользователям, в том числе не обладающим квалификацией ИТ-специалистов. С помощью этого редактора описаны некоторые часто используемые схемы расстановки электродов и проведена генерация соответствующих каркасов электродов. На выходе получен набор моделей, автоматически оптимизированных для 3D-печати. Оптимизация включает автоматическое разделение каркаса, имеющего форму, близкую к эллипсоиду, на разборные элементы, близкие к планарным. Это позволяет минимизировать расход материала за счёт устранения необходимости печати больших поддерживающих конструкций. Таким образом обеспечивается производственный цикл персонализированных каркасов электродов для ЭЭГ, основанный на аддитивных технологиях, эффективный с точки зрения финансовых и временных затрат.

Для дальнейшего развития необходимо провести более тщательное тестирование с большим числом входных данных и выявить точность размещения электродов, также необходимо провести тестирование каркаса в лабораторных условиях при решении задач организации BCI.

Список литературы

1. Mridha M., Chandra S., Moshin M., Akter A., Rashedul M., Watanobe Y. Brain-Computer Interface: Advancement and Challenges // Sensors. 2021. Vol. 17. P. 1–46.
2. Jamil N., Belkacem A., Ouhbi S., Lakas A. Noninvasive Electroencephalography Equipment for Assistive, Adaptive, and Rehabilitative Brain–Computer Interfaces: A Systematic Literature Review // Sensors. 2021. Vol. 14. P. 1--31.
3. Чирухин Д. А., Рябинин К. В. Генерация 3D-модели персонализированного каркаса для сухих электродов электроэнцефалографа // Труды конференции ГрафиКон – 2023. 2023. С. 417–426.
4. FreeCAD [Электронный ресурс] URL: https://www.freecad.org/ (дата обращения 22.07.2024).
5. Koessler L., Cecchin T., Eric T., Maillard L. 3D handheld laser scanner based approach for automatic identification and localization of EEG sensors // Annual International Conference of the IEEE Engineering in Medicine and Biology. 2010. P. 211–252.
6. Martinez E., González A., Garea-Llano E., Bringas-Vega M., Valdes-Sosa P. Automatic Detection of Fiducial Landmarks Toward the Development of an Application for Digitizing the Locations of EEG Electrodes: Occipital Structure Sensor-Based Work // Frontiers in Neuroscience. 2021. Vol. 15. P. 1–14.
7. Homölle S., Oostenveld R. Using a structured-light 3D scanner to improve EEG source modeling with more accurate electrode positions // Journal of Neuroscience Methods. 2019. Vol. 326. P. 1–8.
8. Chen S., He Y., Qiu H., Yan X. Spatial localization of eeg electrodes in a TOF+CCD camera system // Frontiers in Neuroinformatics. 2019. Vol. 13. P. 21–27.
9. Shirazi S., Huang H. More Reliable EEG Electrode Digitizing Methods Can Reduce Source Estimation Uncertainty, but Current Methods Already Accurately Identify Brodmann Areas // Frontiers in Neuroscience. 2019. Vol. 13. P. 1–14.
10. Mazzonetto I., Castellaro M., Cooper R., Brigadoi S. Smartphone-based photogrammetry provides improved localization and registration of scalp-mounted neuroimaging sensors // Scientific Reports. 2022. Vol. 12. P. 1–14.
11. Schubert E., Sander J., Ester M., Kriegel H. P., Xu X. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN // ACM Transactions on Database Systems. 2017. Vol 42. P. 1–21.
12. Gorban A., Kegl B., Wunsch D., Zinovyev A. Principal Manifolds for Data Visualisation and Dimension Reduction. Berlin: Springer, 2008. 300 p.
13. Gruber T. R. The role of common ontology in achieving sharable, reusable knowledge bases // Principles of Knowledge Representation and Reasoning. 1991. P. 601–602.
14. Чуприна, С. И., Зиненко, Д. В. Онтолис: адаптируемый визуальный редактор онтологий // Вестник Пермского университета. Серия: Математика. Механика. Информатика. 2013. № 22. С. 106–110.
15. Virtanen P., Gommers R., Oliphant T. E., Haberland M. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python // Nature Methods. 2020. Vol 17. P. 261–272.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-428-437

Онтологически управляемый инструментарий интеграции нейроинтерфейсов, дружелюбный к пользователю
A user-friendly ontology-driven neurointerface integration toolset

И. А. Лабутин, С. И. Чуприна
Пермский государственный национальный исследовательский университет, 614068, Россия, г. Пермь, ул. Букирева, 15

I. A. Labutin, S. I. Chuprina
Perm State University, 15 Bukireva Str., Perm, 614068, Russian Federation

Аннотация. В последнее время заметно возрос интерес к использованию нейрокомпьютерных интерфейсов для реализации контуров управления различными системами, особенно применительно к различным инфраструктурам Интернета вещей. Однако из-за низкоуровневой природы этих устройств и связанных с ними инструментов, интеграция нейроинтерфейсов с широким спектром устройств IoT представляет собой сложную задачу, требующую специальных знаний в областях нейронаук и обработки сигналов. Мы предлагаем концепцию и прототип системы на основе онтологий для решения задач по бесшовной интеграции интерфейсов мозг-компьютер (ИМК) в экосистемы IoT. Этот высокоуровневый подход позволяет исследователям и инженерам, не имеющим обширного опыта работы с ИМК, автоматизировать интеграцию нейроинтерфейсов с различными инфраструктурами Интернета вещей. В данной работе мы стремимся улучшить предыдущую реализацию нашего подхода, путем создания высокоуровневого графического пользовательского интерфейса и тестирования его с помощью нового конвейера обработки данных на портативном устройстве.

***Ключевые слова*:** нейроинтерфейс, ИМК, онтологический инжиниринг, интернет вещей

Annotation. Recently, there has been a remarkable increase in enthusiasm for utilizing neurocomputer interfaces to develop control systems, particularly within the context of various Internet of Things (IoT) ecosystems. Despite this growing interest, integrating these low-level devices and their functional tools with diverse IoT infrastructure components poses significant hurdles. Achieving successful integration demands a deep understanding of both neuroscience and signal processing. To address these complexities, we propose an innovative, ontology-driven framework designed to enable the effortless incorporation of brain-computer interfaces (BCI) into varying IoT environments. By harnessing the principles of ontology engineering, our high-level approach equips researchers and engineers—with varying levels of BCI experience—to automate the integration between neurocomputer interfaces and multiple IoT infrastructures seamlessly. In this paper, we will detail several advancements made to our previous implementation of this strategy. Notable enhancements include the development of an intuitive graphical user interface and the introduction of a new data processing pipeline that utilizes a portable device to validate the effectiveness of our system.

***Keywords*:** ontology engineering, brain-computer interface, Internet of Things, neurointerface.

I. Introduction

A exponentially growing interest in developing tools and methods for controlling software systems via neurocomputer interfaces have been fueled in recent times due to the rapid advancement of digitization and the exponential growth of applications in the Internet of Things (IoT) as well as augmented and virtual realit. However, a major challenge remains: there is currently no universal standard for integrating different IoT ecosystems with neurocomputer interfaces. The diversity of industrial standards and communication protocols governing device interactions within IoT, combined with frequent instances of incompatibility, complicates the integration of various IoT devices into a cohesive system. This highlights an urgent need for new unified concepts that can improve device interoperability, particularly when incorporating neurocomputer interfaces into existing software ecosystems.

A review of the literature [1, 2, 3, 4] reveals a consensus regarding the insufficient investigation of the challenges associated with creating unified methods and tools for the automatic integration of neurocomputer interfaces into IoT environments, especially concerning the control mechanisms for their components (target systems and subsystems). Generally, BCI devices are tailored to operate effectively only within a limited range of scenarios dictated by proprietary software from their manufacturers. As a result, applying BCIs in contexts or workflows not supported by the original software presents considerable hurdles, which require substantial expertise in both computer science and neuroscience.

The primary aim of this work is to build on our previous studies [5, 6, 7, 8, 9, 10, 11] that introduced an ontology-driven framework for integrating neurointerfaces into IoT ecosystems. The central concept involves creating a "smart mediator," which serves as a new node within the existing IoT infrastructure, facilitating the connection between IoT frameworks and neurocomputer interfaces. The firmware for this "smart mediator" is created based on a collection of ontological descriptions pertaining to the IoT ecosystem, BCI, and data transformation pipeline. A formal model and implementation details were outlined in our earlier work [11], and this paper aims to highlight recent advancements that offer comprehensive improvements.

It's important to note that this paper does not attempt to provide a complete overview of the current state of brain-computer interfaces; interested readers are encouraged to refer to a well-written summary paper [12] for more detailed information.

II. Formal model of a system

Following [11] let's recite the following formal model of the system described in set-theoretic notation:

$$\Xi = \langle \Omega, \Delta, O_b, O_F, O_i, \Gamma, M, E, \Lambda, O_L, S \rangle, \qquad (1)$$

where Ω is the operator for generating source code for mediator, $\Omega : O_D \rightarrow S$;

Λ is the operator for building an ontology of module connections, $\Lambda : O_b \times O_F \times O_i \times D \rightarrow O_L$;

O_b is the ontological description of a specific neurointerface;

O_D is the ontological description of the domain area;

O_L is an ontological description of the links between modules included in the mediator's firmware;

$O_F = \bigcup_{i=1}^{k} O_{fi}$ – the set of k ontological descriptions of modules that implement data transformation;

O_i is an ontological description of the infrastructure, management of which is assumed to be implemented (module for controlling this infrastructure);

Δ is the operator for building an ontology from parts, $\Delta : O_b \times O_F \times O_i \times O_L \rightarrow O_D$;

Γ is an interaction operator, $\Gamma : E \times M \rightarrow D$;

E is a set of supported user actions;

M is a set of supported control elements;

D is a user-created formal description of the interconnection of modules in the form of DFD diagrams in SciVi toolbox;

S is the source code for the mediator's software.

We shall mention that operators Γ, M and E were created by our colleagues within the framework of the SciVi platform [13, 14, 15, 16] and are out of the scope of this work.

III. Operators as SciVi nodes

System prototype designed by us earlier [11] was implemented as a command-line application consisting of few Python modules tied together with a Shell script. Although suitable for demonstrating a proof of concept, this implementation had several drawbacks:

1. **Lack of User-Friendliness**: While a command-line interface might be very powerful and efficient when used by a seasoned IT professional, it poses a significant wall for users that aren't very well-versed in computer science. For such users, a graphical user interface might be much simpler and more intuitive to operate.

2. **Semantic gap**: A few operations (selecting required nodes, connecting and configuring them) were performed in SciVi interface, but then the user had to switch into a terminal or a console emulator to perform a rest of the work. This disrupts a user experience and requires the user to master both parts of the interaction.

3. **Limited Scope for Improvement and Iteration**: This solution wasn't very good for further improvement. Due to the structure and design of the prototype, it posed challenges in terms of enhancing its features, addressing issues, and incorporating feedback for iterative development.

4. **Tight Coupling**: Modules of the system was combined in a single way and was unable to be reused in other situations without writing a corresponding code for the task.

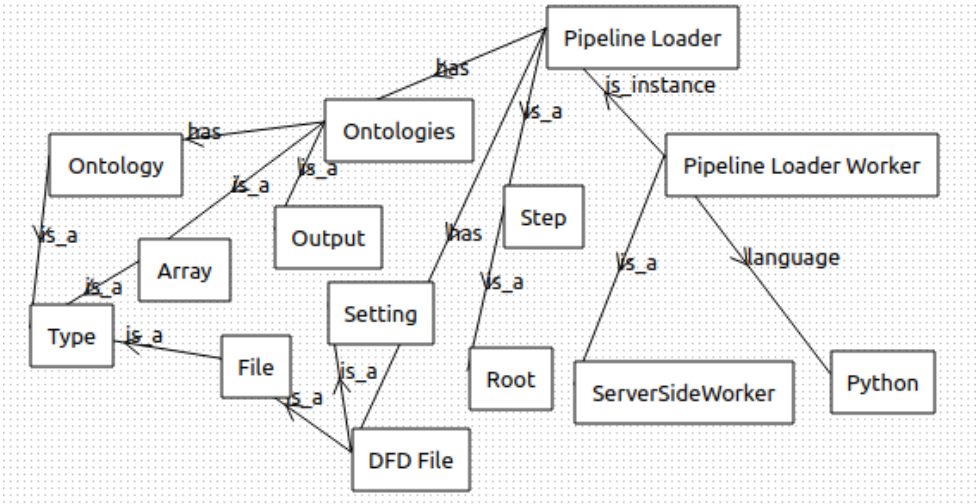


Fig. 1. „Pipeline Loader" operator

To address these challenges effectively, we decided on a strategic approach: designing a comprehensive set of SciVi operators that encapsulate the system's functionality. By implementing all operations within the SciVi framework, users will be presented with the ability to interact consistently through a unified interface, streamlining the creation and management of data processing pipelines. This approach not only enhances user experience by maintaining continuity in interaction methods but also promotes modular reuse of system components. Modules developed under this paradigm can be easily adapted and integrated into various contexts without requiring extensive reconfiguration or redevelopment, thus facilitating smoother iteration and improvement cycles for the system as a whole.

To cover all the functionality of a system, we made a few new SciVi operators:

1. *Pipeline Loader*: This is a utility operator that fetches ontological descriptions from the SciVi storage.
2. *Ontology Merger*: This is an implementation of the Δ operator from (1).
3. *DFD2Link*: This is an implementation of the Λ operator from (1).
4. *Generator*: This is an implementation of the Ω operator from (1).

„Pipeline Loader" operator (fig. 1) is responsible for analyzing the firmware's data pipeline created in SciVi by the user and fetching all required ontological resources from the SciVi repository for their latter use in the generator's pipeline.

„Ontology Merger" operator (fig. 2) accepts a set of ontologies as an input and performs a fusing (merging [14, 15]) of them together. It respects namespaces, but doesn't perform alignment procedure on them, thus not solving the problem of semantic equivalency – ontologies are supposed to be aligned beforehand. This is largely thanks to them being stored in SciVi repository and therefore using the same set of concepts and relations.

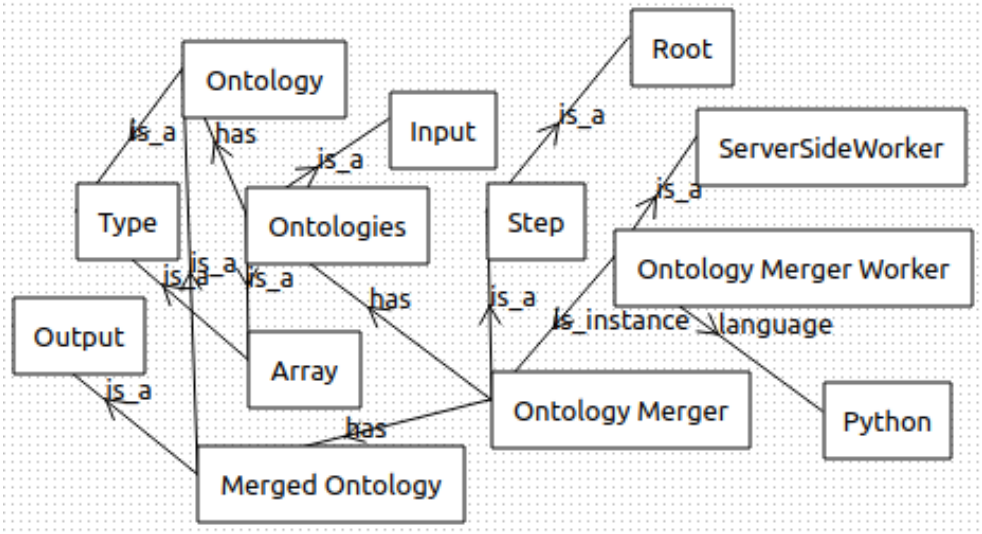


Fig. 2. An implementation of Δ operator

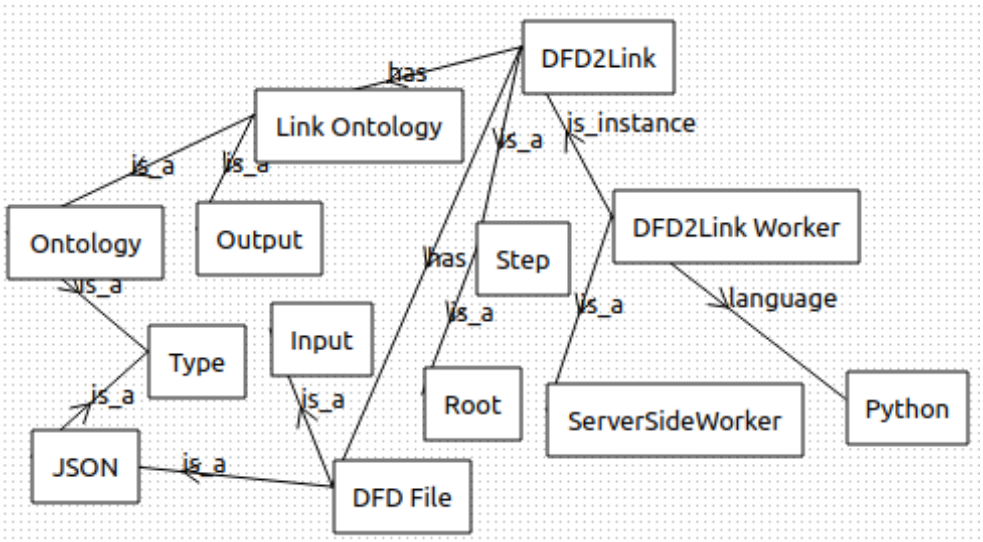


Fig. 3. An implementation of Λ operator

„DFD2Link" operator (fig. 3) has two inputs. One of them is a DFD diagram created in a SciVi editor that describes a data pipeline (module interconnection graph). Another one is a merged domain ontology that has all required module descriptions. The operator then analyses the domain ontology and DFD to produce an ontological description of module's connections.

„Generator" operator (fig. 4) is the one responsible for producing the firmware's source code for the „smart mediator". This operator generates a directed graph that represents the modules and their connections for data transfer by utilizing knowledge about the modules and their interdependencies as outlined in the ontology. In the subsequent step, the operator performs a topological sort to establish a partial order among the vertices. This partial order is defined such that node V_i which corresponds to a module M_i, precedes node V_j, representing module M_j, if and only if there is no path leading from V_j to V_i. If it proves impossible to establish this partial order for a subset of modules, the generator will halt its processing and return an error message indicating a potential cyclic data dependency between the modules. Conversely, if the transformation occurs without any errors, the generator continues to produce source code that specifies all necessary data structures, initializes them, and subsequently calls the entry points of the modules following the sequence established by the dependency graph. This process takes into account the types of parameters and the direction of data flow—whether it is being sent to or received from the module.

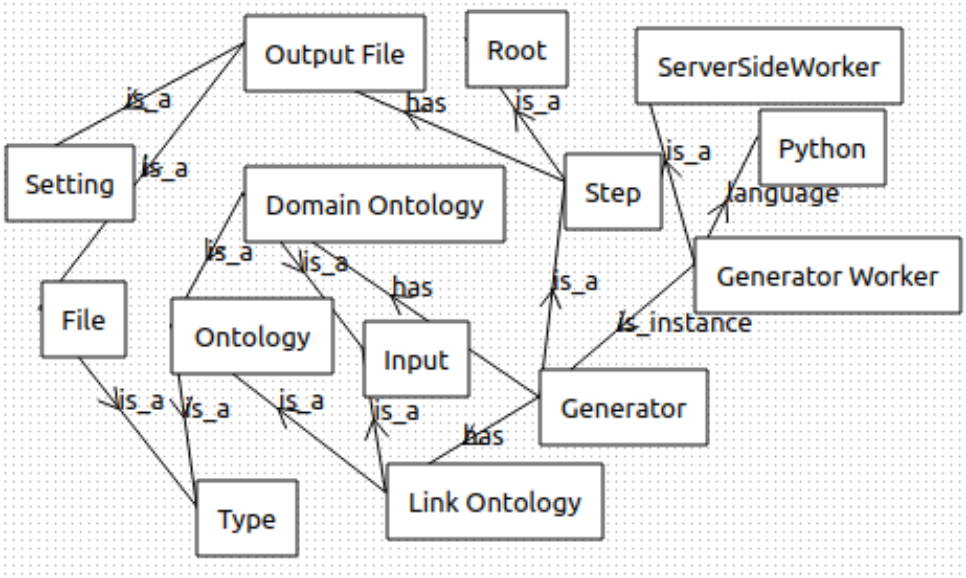


Fig. 4. An implementation of Ω operator

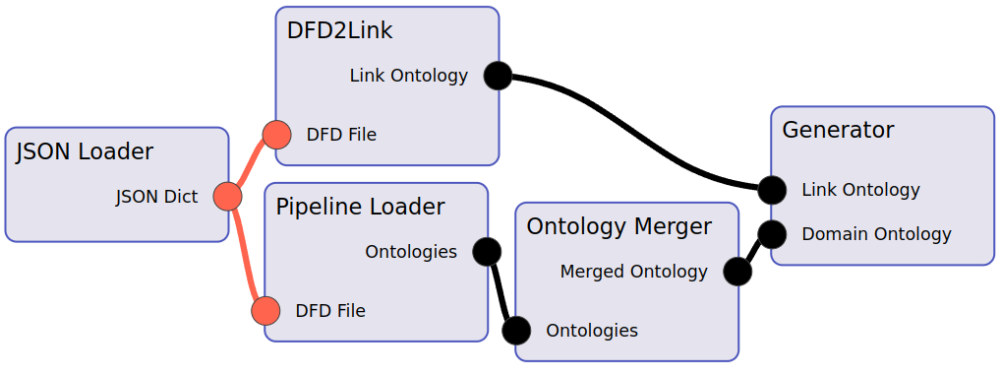


Fig. 5. An overview of a pipeline

A complete pipeline with all interconnected operators are presented in fig. 5.

IV. Controlling Media Playback via BCI

As a test for the generation pipeline created anew, we decided to perform a BCI-powered control over the media playback on the smartphone. As in our previous work [11], the control signal will be derived from the imaginary movement detection classifier we already have as a module in our system, but this time this signal will be used to issue a play / pause command on the smartphone playing a media file. For this task we'll need to create a new module that will allow us to interface with a smartphone's media playback capabilities. We decided to target an Aurora OS-powered devices.

9. Aurora OS

Aurora OS[8] is a mobile operating system that is based on the Linux kernel and a set of GNU utilities, thus being POSIX-compliant, incorporating both open-source projects and components with closed source code, developed by Russian company "Open Mobile Platform". The project is a proprietary fork of Sailfish OS[9] and is designed to build a trusted mobile infrastructure and protect sensitive information in government organizations, educational and healthcare institutions, as well as large and medium-scale commercial companies. It contributes to the technological sovereignty of Russia through a secure mobile environment and complies with all requirements of regulators in the field of national security and information protection.

The OS has been in active development since 2016 by the Russian company „Open Mobile Platform". On November 7, 2023, devices running Aurora OS version 4.0.2, produced by FPlus, were released for a retail sale.

The primary development tools for applications on the Aurora OS are the C++ programming language, Qt framework, and Silica module. The Silica module provides essential components for building the user interface of Aurora OS applications, as well as tools for styling Aurora OS applications and managing their behavior. The Silica module not only facilitates the construction of application interfaces but also ensures stylistic consistency between many Aurora OS applications.

The main reason of choosing Aurora OS for our demonstration purposes is the fact that applications for it are basically a regular GNU/Linux programs (with some quirks) written in C++, which makes interfacing with a C code generated by our system transparent and allows very efficient debugging. Other mobile operating systems, like Android, require much more sophisticated setup for that.

10. Media Playback Operator

Aurora OS, being a descendant of Sailfish OS, behaves very much like any standalone GNU/Linux distribution, and provides same APIs for programmers to employ. Specifically, it follows a set of FreeDesktop.Org[10] specifications that were created to ensure interoperability of a desktop software in the Unix-like environments.

One of such specifications that interest us is a Media Player Remote Interfacing Specification (MPRIS)[11] that is used to control multimedia applications. It's built on top of a D-Bus message bus system which is supported by Qt framework. We can use it to issue a command to a running mediaplayer instance for running or pausing playback.

[8] https://auroraos.ru/
[9] https://sailfishos.org/
[10] https://freedesktop.org/
[11] https://specifications.freedesktop.org/mpris-spec/latest/

For this purpose, a simple operator based around QDBusConnection and QDBusInterface classes was created and added to the system's repository. Fig 6 displays an ontology for this operator.

General data processing pipeline is presented at Fig. 8. It consists of several modules:

• „EEG Settings" is a simple operator that allows for the user to set a few constants for later use in the data processing pipeline.

• „Med Sampling" abstracts away EEG connection and data acquisition. It communicates with EEG device and retrieves a raw data from it in a form of two-dimensional array of floats each tick pipeline is being executed.

• „Band Powers" is an operator built upon Brainflow's library implementation of power spectrum analysis module that was adapted by us to meet the limitations of an embedded and portable systems. It accepts a window of raw EEG data and outputs a number of power spectrum's density values.

• „Concentration Predict" is a (probably a bit misnamed) threshold-based classifier that deduces the presence or absence of a specific pattern in the EEG signal (in our case, an imaginary movement).

• „Band Visualize" operator (fig. 7) is used to display band powers as well as a current playback status on the smartphone's screen while the app is running. It was introduced as a simple way to get a nice and easy to understand feedback on the mediator's current internal state. This operator can be useful both to a system's operator (the one who created pipeline) as a debug tool and to the end user of the smart mediator, who can be a completely obvious person – a hospital's patient, for example – to get a general idea of what's happening.

• At the final stage of the data processing pipeline, our new „Media Playback" operator is introduced to send a command to the running media player instance based on the output from the classifier.

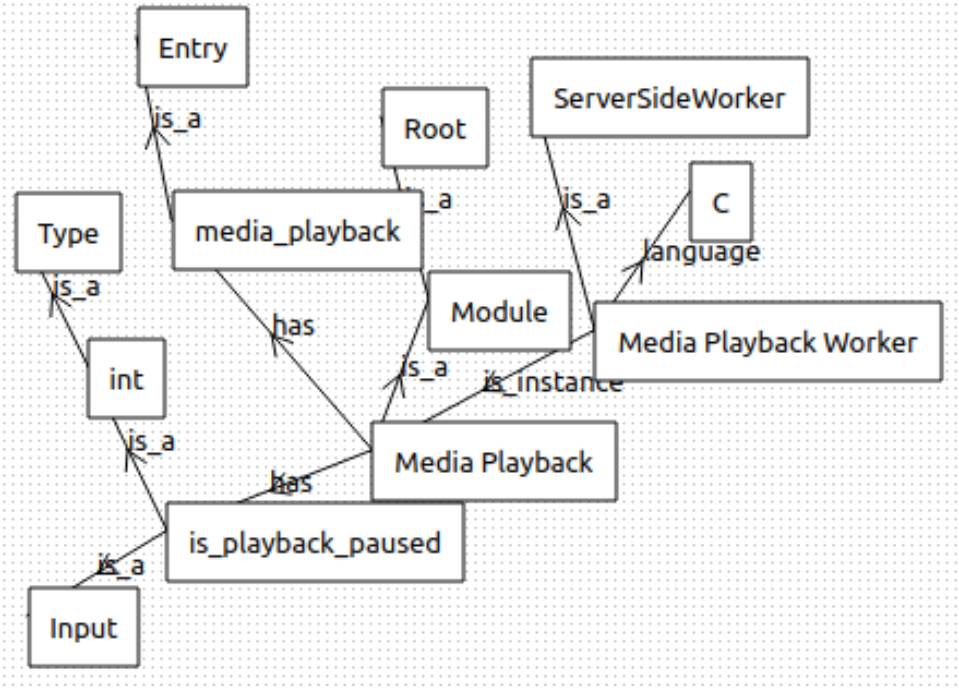


Fig. 6. A „Media Playback" operator

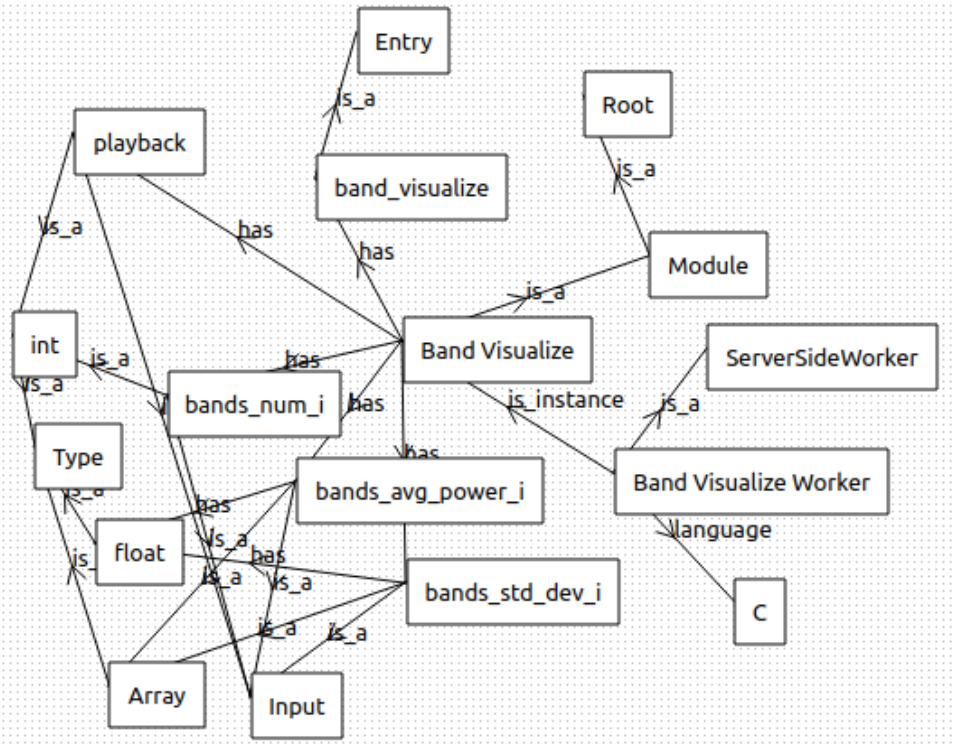


Fig. 7. A „Band Visualize” operator

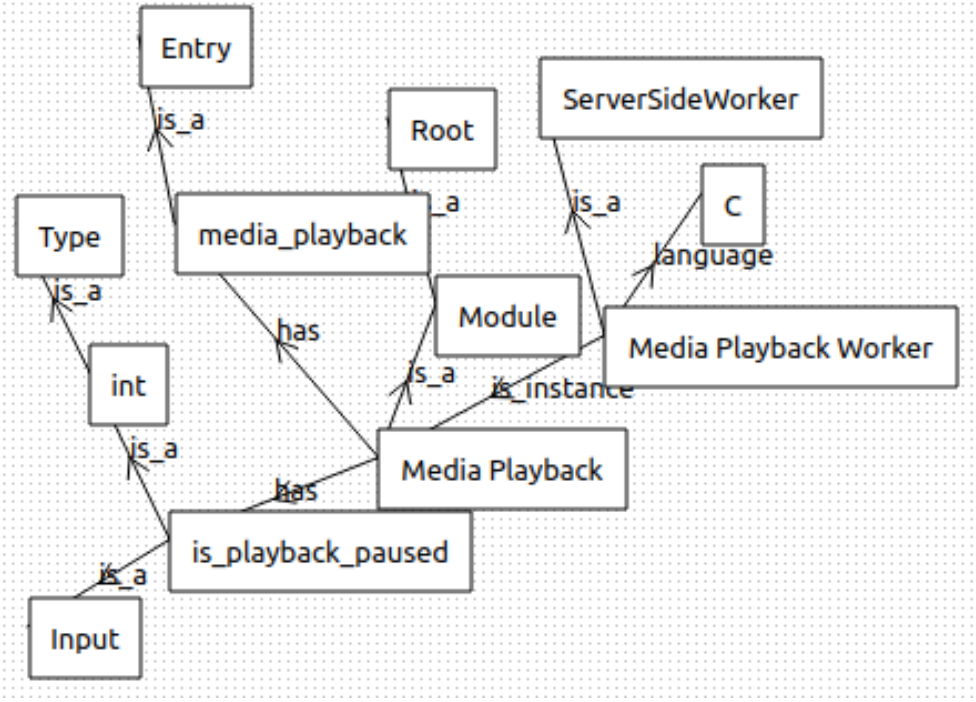


Fig. 8. Media playback pipeline

After this pipeline was created, it was saved in a form of JSON file and then reused as an input to the firmware generation pipeline 5. The resulting domain ontology (fig. 9) and generated source code (fig. 10) are presented below.

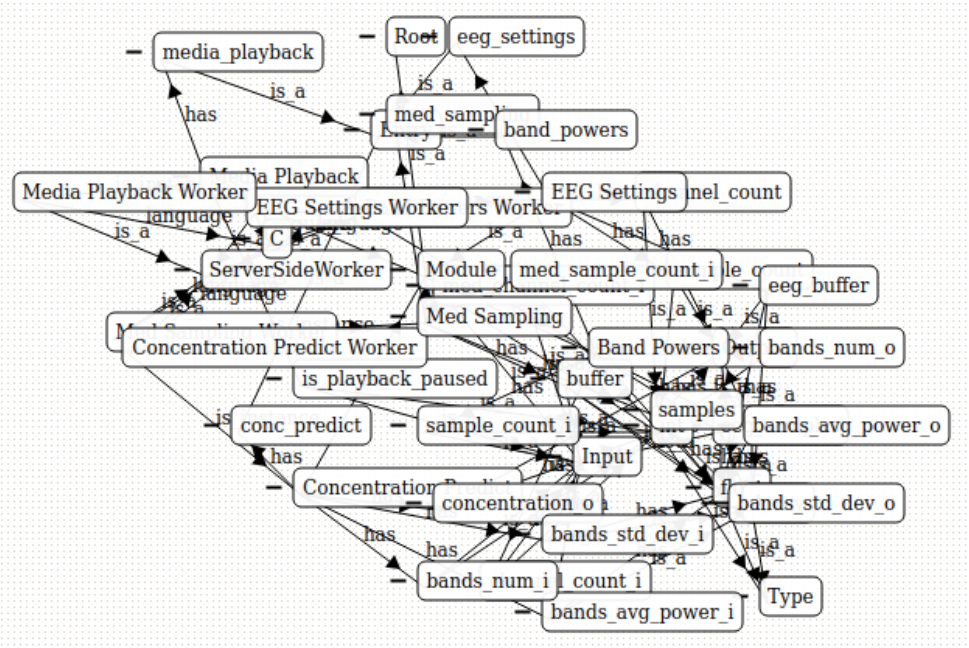


Fig. 9. Resulting domain ontology

The generated source code was then used in an application build process as a glue that calls each module entry point in the specified order and thus processes data from the EEG to achieve the intended result – a media playback control. A screenshot of the application is presented at fig. 11 displaying information about current relative band powers and playback status.

```
// Variable declarations
int bands_num_o;
float* bands_avg_power_o;
float* bands_std_dev_o;
int concentration_o;
float* eeg_buffer;
int channel_count;
int sample_count;
float* samples;

// Module entrypoint declarations
void eeg_settings(float** eeg_buffer, int* channel_count, int* sample_count);
void med_sampling(float* buffer, int med_sample_count_i, int med_channel_count_i,
        float** samples);
void band_powers(float* eeg_samples, int channel_count_i, int sample_count_i,
        int* bands_num_o, float** bands_avg_power_o, float** bands_std_dev_o);
void conc_predict(int bands_num_i, float* bands_std_dev_i, float* bands_avg_power_i,
        int* concentration_o);
void media_playback(int is_playback_paused);
void band_visualize(int bands_num_i, float* bands_std_dev_i,
        float* bands_avg_power_i, int playback);

// Processing code
eeg_settings(&eeg_buffer, &channel_count, &sample_count);
med_sampling(eeg_buffer, sample_count, channel_count, &samples);
band_powers(samples, channel_count, sample_count, &bands_num_o, &bands_avg_power_o,
        &bands_std_dev_o);
conc_predict(bands_num_o, bands_std_dev_o, bands_avg_power_o, &concentration_o);
media_playback(concentration_o);
band_visualize(bands_num_o, bands_std_dev_o, bands_avg_power_o, concentration_o);
```

Fig. 10. Example of generated code

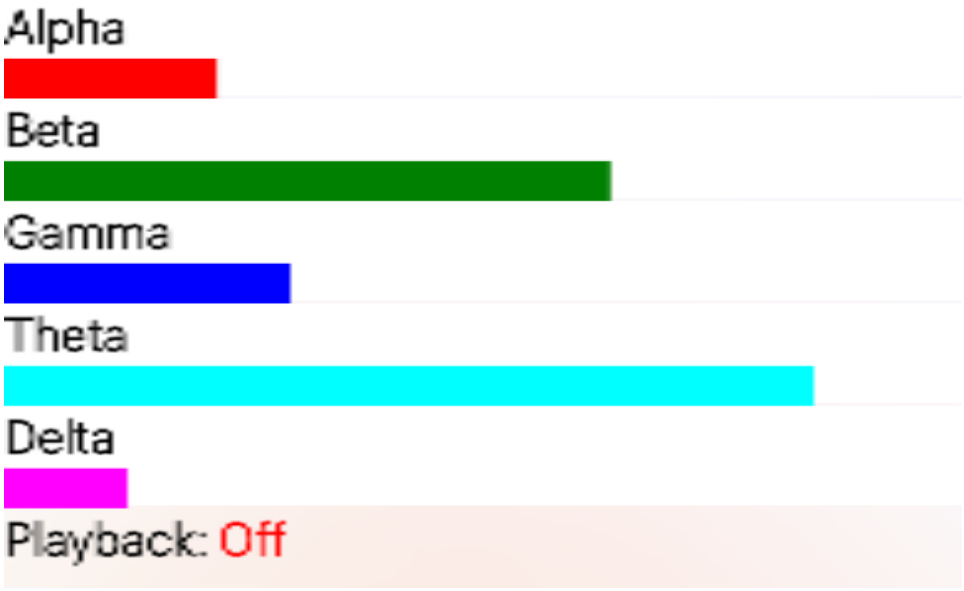


Fig. 11. Application screenshot displaying bands power and playback status

The pipeline presented here (fig 6) can be extended further with a set of modules that will allow more control signals and actions to be performed on a device, and extend a visualization part do display other characteristics of a signal. For example, this pipeline could be extended to be used in a hospital for a patient to call for a doctor or to be employed by a neuroresearcher to visualize some aspect of the EEG signal or it's analysis on their smartphone as the experiment goes on.

V. Conclusion and Future Work

This paper tackles the significant challenge of automating the integration of various neurocomputer interfaces into IoT infrastructure by building upon a previously proposed solution. It emphasizes the need for adaptable methods and tools that facilitate the integration of neurocomputer interfaces with third-party software systems through ontology engineering.

An innovative approach is introduced, utilizing the SciVi platform as a comprehensive user interface for constructing a data processing and firmware generation pipeline. This pipeline, designed to control media playback on a portable device, has been successfully developed and tested using the new user interface.

The effectiveness of this proposed method has been validated through practical testing at the Educational and Scientific Laboratory of Sociocognitive and Computational Linguistics at PSU. The possible next directions of our research include aiming towards improving both the functional characteristins of the system as well as usability of the developed tools. Key areas of focus will include enhancing the integration with build systems and resolving cyclic data dependencies within the firmware of the "smart mediator" between various program modules.

Acknowledgements

The authors wish to express their sincere gratitude to the Educational and Scientific Laboratory of Sociocognitive and Computational Linguistics at Perm State University's Faculty of Philology. Their appreciation extends especially for the invaluable equipment and steadfast support that were critical to the successful execution of this research.

References

1. S. Huang, E. Tognoli, Brainware: Synergizing software systems and neural inputs, in: Companion Proceedings of the 36th International Conference on Software Engineering, ICSE Companion 2014, Association for Computing Machinery, New York, NY, USA, 2014, p. 444–447. URL: https://doi.org/10.1145/2591062.2591131. doi:10.1145/2591062.2591131.

2. P. McCullagh, M. Ware, A. McRoberts, G. Lightbody, M. Mulvenna, G. McAllister, J. L. González, V. Cruz Medina, Towards standardized user and application interfaces for the brain computer interface, in: C. Stephanidis (Ed.), Universal Access in Human-Computer Interaction. Users Diversity, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 573–582.

3. J. Huggins, C. Guger, E. Aarnoutse, B. Allison, C. Anderson, S. Bedrick, W. Besio, R. Chavarriaga, J. Collinger, A. Do, C. Herff, M. Hohmann, M. Kinsella, K. Lee, F. Lotte, G. Müller-Putz, A. Nijholt, E. Pels, B. Peters, T. Zander, Workshops of the seventh international brain-computer interface meeting: not getting lost in translation, Brain-Computer Interfaces (2019) 1–31. doi:10.1080/2326263X.2019.1697163.

4. B. Allison, The i of bcis: Next generation interfaces for brain–computer interface systems that adapt to individual users, in: J. A. Jacko (Ed.), Human-Computer Interaction. Novel Interaction Methods and Techniques, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 558–568.

5. I. A. Labutin, S. I. Chuprina, Kontseptsiya postroeniya ontologicheski upravlyaemykh neirointerfeisov, Intellektual'nye sistemy v nauke i tekhnike (2020) 105–111.

6. K. Ryabinin, S. Chuprina, I. Labutin, Ontologydriven toolset for audio-visual stimuli representation in eeg-based bci research, in: Proceedings of the International Conference on Computer Graphics and Vision "Graphicon", CEUR, volume 31, Keldysh Institute of Applied Mathematics, 2021, pp. 223–234. URL: https://keldysh.ru/papers/2021/prep_vw.asp?pid=9273&lg=e. doi:https://doi.org/10.20948/graphicon-2021-3027-223-234. arXiv:http://ceur-ws.org/Vol-3027/paper21.pdf.

7. K. V. Ryabinin, S. I. Chuprina, L. I. A., Ontologydriven tools for eeg-based neurophysiological research automation, Scientific Visualization 13 (2021) 93–110. doi:10.26583/sv.13.4.08.

8. S. Chuprina, I. Labutin, Scientific visualization tools to improve utilizing neural interface, in: Proceedings of the International Conference on Computer Graphics and Vision "Graphicon", volume 32, Keldysh Institute of Applied Mathematics, 2022, pp. 391–402. URL: https://keldysh.ru/papers/2022/prep_vw.asp?pid=9592&lg=e. doi:https://doi.org/10.20948/graphicon-2022-391-402.
arXiv:https://www.graphicon.ru/html/2022/papers/paper_038.pdf.

9. K. Ryabinin, S. Chuprina, I. Labutin, Tackling iot interoperability problems with ontology-driven smart approach, in: A. Rocha, E. Isaeva (Eds.), Science and Global Challenges of the 21st Century - Science and Technology, Springer International Publishing, Cham, 2022, pp. 77–91.

10. S. I. Chuprina, L. I. A., How to improve utilizing neural interface by means of ontology-driven scientific visualization tools, Scientific Visualization 14 (2022) 83–96. doi:10.26583/sv.13.4.08.

11. S. Chuprina, I. Labutin, Ontology-based neurointerface IoT integration approach, in: Accepted on 28.06.2024, volume 0, Keldysh Institute of Applied Mathematics, 2024.

12. D. Lunev, S. Poletykin, D. O. Kudryavtsev, Brain-computer interfaces: technology overview and modern solutions, Sovremennye innovatsii, sistemy i tekhnologii - Modern Innovations, Systems and Technologies 2 (2022) 0117–0126. URL: http://dx.doi.org/10.47813/2782-2818-2022-2-3-0117-0126. doi:10.47813/2782-2818-2022-2-3-0117-0126.

13. K. Ryabinin, Metody i sredstva razrabotki adaptivnykh mul'tiplatformennykh sistem vizualizatsii nauchnykh eksperimentov, Dissertatsiya, IPM im. M.V.Keldysha, Moskva, 2015. URL: https://library.keldysh.ru/diss.asp?id=2015-ryabinin. arXiv:https://keldysh.ru/council/1/2015-ryabinin/diss.pdf.

14. S. I. Chuprina, K. V. Ryabinin, D. V. Koznov, K. A. Matkin, Ontologicheski upravlyaemye sredstva avtomatizatsii razrabotki prilozhenii vizual'noi analitiki, Programmirovanie (2022) 70–77.

15. S. I. Chuprina, Using data fabric architecture to create personalized visual analytics systems in the field of digital medicine, Scientific Visualization 15(5) (2023) 50–63. URL: https://api.semanticscholar.org/CorpusID:266356696. doi:10.26583/sv.15.5.05.

16. S. I. Chuprina, Adaptatsiya tekhnologii fabrik dannykh k razrabotke sistem vizual'noi analitiki v oblasti tsifrovoi meditsiny, in: Trudy 33 Mezhdunarodnoi konferentsii po komp'yuternoi grafike i mashinnomu zreniyu "GrafiKon-2023", Institut problem upravleniya im. V.A. Trapeznikova RAN, 2023, pp. 405–416. doi:10.20948/graphicon-2023-405-416.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-438-450

Ивент-анализ как инструмент социального прогнозирования
Event analysis as a social forecasting tool

Т. В. Коренькова, А. А. Артамонов, М. С. Улизко
[1]НИЯУ МИФИ, Москва, Россия

T. V. Korenkova, A. A. Artamonov, M. S. Ulizko
[1]National Research Nuclear University MEPhI, Moscow, Russia

Аннотация. Мониторинг взаимосвязей между социальными событиями и явлениями и прогнозирование динамики их изменений необходимы в условиях нестабильности современного мира. Существует много отдельных методов анализа для социального прогнозирования, однако для работы выбрана методика ивент-анализа, которая в научной литературе в рамках данной задачи рассмотрена недостаточно, но имеет высокий потенциал. Целью статьи является адаптация методики ивент-анализа для ее использования в качестве инструмента социального прогнозирования. Основные данные для исследования собраны с российского информационно-новостного ресурса в период 2020-2023 года. Опираясь на классическую методику ивент-анализа, в работе определены классификаторы, представленные в работе в виде социальных сфер. В рамках этапа аналитического сравнения проведен графовый анализ (построены графы связей между категориями, выделены центральные узлы-категории); осуществлен анализ временных рядов (сегментация временных рядов алгоритмом PELT, кластеризация временных рядов алгоритмом k-средних); определены ключевые термины для событий-инфоповодов. Итоговым продуктом является аналитическая панель с фильтрами, статистикой и интерактивными графиками. Аналитическая панель позволяет сравнивать данные в статическом и динамическом состоянии, делать выводы о прошлых и будущих состояниях объектов социального прогнозирования. Главным результатом работы является разработанная автором методика проведения ивент-анализа, которую можно использовать для комплексного анализа новостных потоков, адаптировать под нужные категории, представляющие определенную сущность или сферу, и применять в различных социальных организациях или мониторинговых службах.

***Ключевые слова*:** социальное прогнозирование, ивент-анализ, анализ новостей, анализ графов, вычисление центральности, анализ временных рядов, сегментация временных рядов, кластеризация временных рядов, PELT

Аннотация. Monitoring the interrelationships between social events and phenomena and forecasting the dynamics of their changes are necessary in the conditions of instability of the modern world. There are many separate methods of analysis for social forecasting, however, for this work, the method of event analysis has been chosen, which is insufficiently considered in the scientific literature within the framework of this task, but has high potential. The purpose of the article is to adapt the event analysis methodology for its use as a social forecasting tool. The main data for the study was collected from the Russian information and news resource in the period 2020-2023. Based on the classical methodology of event analysis, the classifiers presented in this paper in the form of social spheres are defined in the work. As part of the analytical comparison stage, a graph analysis was carried out (graphs of relationships between categories were constructed, central nodes-categories were identified); time series analysis was performed (segmentation of time series by the PELT algorithm, clustering of time series by the k-means algorithm); key terms for press events were defined. The final product is an analytical dashboard with filters, statistics and interactive graphs. The analytical dashboard makes it possible to compare data in a static and dynamic state, to draw conclusions about the past and future states of objects of social forecasting. The main result of the work is the event analysis methodology developed by the author, which can be used for a comprehensive analysis of news streams, adapted to the necessary categories representing a certain entity or sphere, and applied in various social organizations or monitoring services.

***Ключевые слова*:** social forecasting, event analysis, news analysis, graph analysis, centrality calculation, time series analysis, time series segmentation, time series clustering, PELT

I. Введение

В условиях быстро меняющегося мира, эффективное прогнозирование социальных процессов становится одной из ключевых задач для исследователей и практиков [1]. Ивент-анализ, как метод, предлагающий системный подход к изучению значимых событий и их последствий, представляет собой инструмент для выявления закономерностей и тенденций в социальном развитии.

А. В. Мальцевой [1] предложен алгоритм применения ивент-анализа на примере изучения процессов на рынке труда, состоящий из этапов сбора данных, выделения классификаторов, подсчета результатов, сравнения

полученных величин и верификации результатов. При этом выявлено, что не существует однозначной методологии проведения данного вида анализа, поэтому становится возможным модифицировать некоторые этапы. М. В. Новоселов [2] описал некоторые дополнительные статистические, математические и графические методы, которые можно применять для проведения расширенного ивент-анализа, в частности, анализ временных рядов, кластерный анализ.

Количество как русскоязычных, так и зарубежных научных работ на стыке социального прогнозирования и ивент-анализа несущественно. А. А. Азаров [3] исследовал термин социального компьютинга — междисциплинарной области исследований, включающей изучение социального поведения средствами вычислительных систем. Одной из важных задач социального компьютинга является моделирование, анализ и прогнозирование социального поведения акторов с помощью таких методов, как интент-анализ, ивент-анализ, психографический анализ. В. И. Кудрявцевой [4] описаны основные подходы к проведению социального прогнозирования: общенаучные, интуитивные, формализованные и комплексные методы прогнозирования, и среди их числа упоминается применение ивент-анализа.

Гораздо больше исследований посвящено исследованию динамики развития социальных сущностей в новостях и их кластеризации, что связано с конкретными инструментами ивент-анализа. Дж. Л. Ортега [5] предложил и реализовал подход к построению сетевых графов на основе частоты, с которой СМИ совместно упоминают исследовательские работы. И. Блох и В. Александров [6] построили временные ряды, которые отображают распределение популярности кластеров, представляющих определенные социальные явления, с течением времени.

Таким образом, в отдельности тематики социального прогнозирования и ивент-анализа изучаются в научных публикациях довольно обширно, но работ, описывающих их совместное применение, ограниченное число. На основе обзора литературы определено, что анализ графов и временных рядов событий могут быть хорошими инструментами для изучения связей между социальными сущностями и отслеживания динамики их изменений. Стоит отметить, что эти инструменты редко описываются другими авторами на этапах ивент-анализа. Это свидетельствует о новизне и актуальности работы и необходимости изучать ивент-анализ и социальное прогнозирование в связке.

II. Постановка задачи

Цель работы заключается в адаптации методики ивент-анализа для использования в качестве инструмента социального прогнозирования. В ходе исследования рассмотрены ключевые аспекты социального прогнозирования, выявлены его основные сферы применения и описаны этапы классической методики ивент-анализа. Особое внимание уделено разработке модификаций, которые позволят усовершенствовать данную методику, а также её практическому применению на основе реальных данных.

В процессе работы исследованы как теоретические основы ивент-анализа, так и его практическая значимость в различных социальных контекстах. Результаты исследования направлены на создание эффективного инструмента, который не только повысит точность анализа, но и позволит более надежно предсказывать социальные изменения.

III. Теория

Социальное прогнозирование – это процесс определения ожидаемых и желаемых изменений в различных социальных институтах, таких как государство, образование, рынок труда и т.д. Основными особенностями социального прогнозирования являются отсутствие четких целей и направлений прогнозирования, сложность формализации социальной сферы и необходимость сочетания качественных и количественных методов [7]. Социальное прогнозирование, в отличие от обществоведческого – это прогнозирование только в области социологии. Однако социология с каждым годом все более активно вторгается в другие отношения со стороны их социальной организации: экономические, национальные, морально-этические и т.д. На основе этих тенденций, некоторые авторы выделяют перечень областей и сфер, в которых возможно проводить социальное прогнозирование, см. Табл. 1 [7].

ТАБЛИЦА 1
СФЕРЫ ПРИМЕНЕНИЯ СОЦИАЛЬНОГО ПРОГНОЗИРОВАНИЯ

Сфера	**Области применения социального прогнозирования**
Наука	Перспективы развития научных кадров, исследовательских учреждений, финансирования научных открытий
Технологии	Перспективы развития новых технологий, информатизации общества, сохранения конфиденциальности

Экономика	Перспективы развития социальной организации труда, борьбы с безработицей и инфляцией
Политика	Перспективы развития государственных и международных отношений, мониторинг отношения народа к власти
Право	Перспективы внедрения новых законов в социальной сфере, мер поддержки населения, сохранения прав человека
Народонаселение	Перспективы изменения структуры общества, миграционных процессов
Образование	Перспективы развития различных учреждений образования, повышения квалификации и стимулирования кадров
Здравоохранение	Перспективы развития медицинских учреждений, открытий в области медицины, здорового образа жизни
Культура	Перспективы развития культурных благ, туризма, влияния СМИ, сохранения культурного наследия
Экология	Перспективы освоения Земли и космоса, сохранения окружающей среды, развития городов и транспорта
Социальная структура	Перспективы развития социально-производственной, профессиональной, образовательной и половозрастной структур
Общественная жизнь	Перспективы развития общественного порядка, социальных потребностей, борьбы с антиобщественными явлениями, неравенством и бедностью

Существует множество методов социального прогнозирования: как правило, применяются в отдельности либо экспертные, либо фактографические. Под критерии комплексности, гибкости, универсальности и изученности в научных работах, подходят не так много методов: в частности, экспликативные [3]. На основе анализа научных работ выдвинута гипотеза, что ивент-анализ обладает достаточной гибкостью и набором инструментов для его применения в задаче социального прогнозирования.

Ивент-анализ как метод политической науки зародился в 1960-х годах в научных трудах Чарльза Макклелланда [8]. Ивент-анализ является количественным методом изучения политической реальности, фокусирующимся на систематическом анализе сообщений о событиях. Объектом ивент-анализа являются не сами события, а сообщения о событиях, преимущественно из СМИ. Эти события систематизируются, анализируются, классифицируются и обрабатываются с использованием программно-математических методов [1].

В настоящее время ивент-анализ находит применение в конфликтологии, социологии, политологии и естественных науках. Широкая сфера его применения объясняется возможностью сравнения различных событий, их анализа по количеству участников, продолжительности и масштабам взаимодействия. Это позволяет не только сравнивать события, но и строить многовариантные сценарии, в связи с чем повышается точность не только тактического, но и стратегического прогнозирования. Таким образом, ивент-анализ обеспечивает более детальное представление об изменениях в политической и социальной обстановке по сравнению с традиционными исследовательскими методами.

Практическую реализацию ивент-анализа можно разделить на две основные фазы. Первая фаза включает формализованное представление сообщений о событиях с использованием определенной кодировочной схемы, что создает «данные о событиях» (Event Data). Вторая фаза включает использование данных для формулирования содержательных гипотез и выводов относительно изучаемых политических процессов, а также для построения и тестирования моделей. В современной политической науке на этой стадии используется широкий набор статистических методов и математических подходов, таких как факторный, дискриминантный, корреляционный, кластерный анализ и т.д. На заключительном этапе происходит валидация результатов с последующим составлением прогнозов и экспертных оценок [1]. Модификации в виде внедрения графовых алгоритмов, сегментации и кластеризации временных рядов, выделения ключевых слов внесены на этапе аналитического исследования, а все остальные этапы выполнены в соответствии с установленной методологией. Описание каждого этапа ивент-анализа, реализованного в этой работе, приведено далее.

11. Составление информационного массива

В качестве источника сбора данных выбран информационно-новостной ресурс, поскольку новости отражают практически все аспекты социальных взаимодействий и событий. Сбор данных проведен автоматизированно методом веб-скрапинга на языке программирования Python с помощью библиотек BeautifulSoup и requests. Для этой цели использован только один российский информационно-новостной ресурс, чтобы избежать проблем дублирования и агрегирования данных из разных источников. Новости собраны за полных четыре года в период с 2020 по 2023 года вместе с такими метаданными, как дата, рубрика, заголовок исходной новости, а также заголовки новостей, упомянутых в контексте [9].

12. Построение системы классификаторов

Второй этап – разработка системы классификации новостных сообщений о социальных событиях для формализации событий и явлений и анализа их взаимодействий. Прежде всего определены социальные сферы, которые могут требовать разработки социальных прогнозов, см. Табл. 1. При необходимости, некоторые из них можно расширять или объединять. В научных трудах при подготовке систем классификации предполагалось составление таких категорий, которым возможно дать конкретную эмоциональную окраску – либо позитивную, либо негативную: например, в сфере изучения конфликтов. В работе также перенята данная особенность, и к каждой из вновь образованных категорий добавлена эмоциональная составляющая. Некоторые из категорий изначально предполагают ту или иную эмоциональную окраску – например, бедность и преступление – это негативные явления, а сотрудничество и импортозамещение – позитивные. Некоторые категории представлены более обобщенно путем добавления приставок «развитие», «достижения», «проблемы» – например, «развитие образования». Итоговый список категорий представлен в результатах.

Следующим важным шагом является классификация новостного массива по выбранным категориям. Для этой цели выбран подход тематического моделирования с помощью Zero-Shot классификации. Тематическое моделирование — одно из современных приложений машинного обучения к анализу текстов, которое позволяет определить, к каким тематикам относится каждый документ и какие слова образуют каждую из них [10]. При этом одна из главных трудностей – формирование обучающих данных под каждую категорию, что в рамках исследования очень проблематично, так как сделать качественную разметку данных под каждую категорию не представляется возможным в связи с несбалансированностью всей выборки. В связи с этим выбран подход Zero-Shot классификации для тематического моделирования, который позволяет обойти все эти ограничения.

Zero-Shot классификация текста – задача классификации, в которой модели могут классифицировать текст, не обучаясь при этом на наборе данных, созданном для этой задачи классификации [11]. Модель способна предсказать, к какому из предложенных классов вероятнее всего относится текст на основе анализа ключевых слов и контекста. Для классификации новостей по новым категориям выбрана готовая мультиязычная модель тематического моделирования с помощью подхода Zero-Shot классификации [12]. Данный этап выполнен на языке программирования Python с помощью библиотеки transformers.

13. Подсчет результатов заполнения матричного классификатора

Третьим этапом в классической методике ивент-анализа является подсчет результатов заполнения матричного классификатора. Количественные данные для аналитического сравнения качественных характеристик ситуации выражают через определение их относительных значений, а также через построение индексов. Определение относительных значений целесообразно, если требуется статистическая обработка данных, особенно при использовании событийной информации. Построение индекса применяется для объединения различных количественных данных в единый комплексный показатель с целью последующего отслеживания ситуации [1].

Поскольку основной этап ивент-анализа состоит из совокупности методов анализа, то для каждого из них требуется свой подход к подсчету итоговых результатов. В большинстве случаев предполагается подсчет абсолютного количества новостных сообщений по каждой тематике за определенный период времени.

14. Проведение аналитического исследования

Самый трудоемкий этап исследования – проведение аналитических сравнений полученных значений показателей, описывающих типы событий или их аспекты на различных временных этапах. Весь анализируемый период поделен на интервалы, а события, наблюдаемые в каждом из них, сравнены по различным критериям в рамках данных периодов.

Методы анализа выбраны таким образом, чтобы после выполнения данного этапа можно было оценить параметры, предложенные Ч. Макклелландом, в соответствии с которыми осуществляется обработка данных, см. Табл. 2 [8].

ТАБЛИЦА 2
ПАРАМЕТРЫ ИВЕНТ-АНАЛИЗА

Параметр	**Вопрос, на который отвечает параметр**
Оценка сюжета	Что происходит?
Оценка субъекта-инициатора	Кто за этим стоит?
Оценка объекта	По отношению к кому?
Оценка времени события	Когда?

Этот этап разделен на три основные части: анализ графов, анализ временных рядов, выделение ключевых слов. Выбор этих подходов обоснован параметрами, выделенными выше. Графовый анализ используется для оценки объекта («по отношению к кому/чему?») и частично субъектов-инициаторов («кто/что за этим стоит?»), анализ временных рядов – для оценки времени события («когда?»), выделение ключевых слов – для оценки сюжета («что происходит?»).

Графовый анализ в рамках ивент-анализа описывается очень редко, однако он наилучшим образом позволяет увидеть силу взаимосвязей между различными социальными категориями [13]. Граф представляет собой пару G = (V, E), где V – множество, элементы которого называются вершинами, а E – набор неупорядоченных пар вершин, элементы которых называются ребрами [14]. Графы взаимосвязей построены на основе появления категорий в контексте одной новости. Узлы – сами категории, а ребра показывают наличие связи между ними в рамках одного новостного текста. Ребра также имеют веса, и чем больше новостей связывает две категории, тем больше значения этих весов. Чтобы проследить динамику взаимосвязей между социальными событиями и явлениями, исследователю представляется возможность построить графы за годовой интервал и сравнивать значения в разрезе четырех лет.

Для алгоритма ивент-анализа вычисление центральности также является важным шагом, поскольку позволяет выявить самые межотраслевые и связующие категории. Центральность – это один из важнейших показателей, который используется для того, чтобы показать релевантность или структурную важность узла в сети. Для каждой категории в разрезе лет возможно рассчитать показатель центральности по степени: согласно этому подходу, большее значение центральности получают узлы с большим количеством связей, а сам показатель считается как отношение количества узлов, с которым у рассматриваемого узла есть связи, к общему количеству узлов [14]. Графовый анализ проведен на языке программирования Python с использованием библиотек networkx и plotly для визуализации результатов.

Для оценки времени события выбран метод анализа временных рядов. Временные ряды построены отдельно для каждой категории, где на оси x – дата, на оси y – количество новостей по данной категории за эту дату. В научных работах по ивент-анализу временные ряды, как правило, используются для построения частотного распределения классификаторов по заданным периодам [6]. Однако такой подход может быть не очень информативным, и здесь главная цель – определить ключевые события, в результате которых появились экстремальные точки или выбивающиеся из общего ряда периоды, а для решения этого вопроса ручная обработка может быть трудоемкой и неэффективной. Из-за этого было решено добавить этап сегментации временных рядов, в результате которого такие события будут определяться автоматически и показывать изменения более наглядно.

Для этой цели использован алгоритм Pruned Exact Linear Time (PELT). Этот алгоритм ищет набор точек «перегиба» для заданного временного ряда таким образом, чтобы их количество и местоположение минимизировали заданную «стоимость» сегментации. Основные шаги алгоритма заключаются в определении функции «стоимости» для сегмента, затем итерации по всем возможным начальным и конечным точкам сегмента и проверке того, уменьшает ли разделение на новые сегменты значение функции стоимости по сравнению с сегментом без разделения. Одним из часто используемых подходов к идентификации нескольких точек изменений является сведение к минимуму суммы, представленной в формуле ниже [15].

$$\sum_{i=1}^{m+1}\left[C_{\left(y_{(\tau_{i-1}+1):\tau_i}\right)}+\beta f(m)\right]$$

Здесь C – функция «стоимости» сегмента, t – точка «перегиба», m – общее количество точек «перегиба», $\beta f(m)$ – регуляризатор для предотвращения переобучения.

Выявлено, что некоторые точки и периоды сниженной новостной активности на временных рядах совпадают с праздничными днями и выходными. Поэтому данные по этим дням удалены из выборки, чтобы временные ряды получились более сглаженными и не находилось ложных корреляций. Данный этап реализован на языке программирования Python с использованием библиотеки ruptures для сегментации и библиотеки plotly для визуализации результатов.

Еще один метод, позаимствованный из научных работ – это кластеризация временных рядов, которая позволяет обнаруживать корреляции между временными рядами, представляющими различные категории. Для проведения кластеризации построены временные ряды не по дням, а по месяцам, так как это позволяет сгладить

временные ряды и избавиться от шума. Для данной задачи взят метод k-средних, по которому считается евклидово расстояние между векторами несмещенных временных рядов, для них ищутся центроиды и наконец определяются кластеры в результате перемещения центроид по количеству итераций [16]. Сам алгоритм k-средних кластеризует временные ряды, построенные по месяцам за интересующий год, так как это помогает определить корреляции между ними лучше, чем по датам. Для реализации этой задачи использовался алгоритм k-средних, реализованный на языке программирования Python с использованием библиотек sklearn для нормализации временных рядов, tslearn для кластеризации и plotly для визуализации результатов.

Для оценки сущности событий, участвующих в них субъектов и определения более содержательной корреляции между категориями, добавлен еще один метод – выделение ключевых терминов из новостных заголовков для экстремальных значений на временных рядах. Под такими значениями подразумевается даты, совпадающие с максимальной частотной активностью по конкретным категориям, а также точки «перегиба», которые обозначают появление нового сегмента на временном ряде: некоторых переломных событий, в результате которых новостная активность в определенные периоды стала резко отличаться.

Для реализации данного шага использовался подход для определения ключевых терминов в тексте, основанный как на лингвистических инструментах (определение части речи, токенизация), так и на относительной частотности [17]. Для каждой экстремальной даты в рамках одной категории и временного периода выводится список наиболее частотных ключевых терминов, которые позволяют делать более информативные выводы о ключевых событиях, их акторах и эмоциональной составляющей, которые повлияли на изменение новостной активности. Для этой задачи взята модель из библиотеки TermExtractor, код реализован на языке программирования Python.

15. Верификация результатов и формулирование выводов

Поскольку в данном варианте методики не предполагается прогнозирование программно-математическими методами, то этот этап является полностью ручным, а главный результат реализуемой методики – проанализированные данные, оформленные в виде интерактивных графиков на итоговой аналитической панели [18]. Выбирая различные фильтры, такие как интересующий временной период и категория, можно сравнивать результаты анализа по годам, определять изменение поведения анализируемых объектов в динамике и на основе этого строить социальные прогнозы.

Выполнение финального этапа методики ивент-анализа – валидация результатов и социальное прогнозирование – проводится на основании итоговой аналитической панели, которая включает в себя:

- совокупность интерактивных графиков, построенных по перечисленным ранее методам: графовый анализ, сегментация и кластеризация временных рядов, распределение ключевых терминов;
- выбор фильтров – «Категория» и «Год»;
- статистику по абсолютному и относительному количеству новостей по «Категории» и «Году», а также центральные категории за взятый период.

IV. Результаты экспериментов

Фрагмент собранного массива данных с российского информационно-новостного ресурса представлен на рисунке (рис. 1).

дата	рубрика	заголовок	новости, упомянутые в контексте
2023/02/28	Россия	Песков рассказал о планах Путина участвовать в президентских выборах в 2024 году	[В Кремле ответили на вопрос об участии Путина в президентских выборах 2024 года]
2021/12/10	Экономика	Белоруссия собралась погашать долги за счет займов у России	[США оставили Белоруссию без долларов, В Белоруссии вступил в силу указ о конфискации валюты, Всемирный банк «рассекретил» должников России]
2023/06/13	Мир	В США заявили о неусвоенных уроках спецоперации	[Американский ветеран указал на обман Запада Украиной]
2023/06/26	Мир	Стало известно о финансовом источнике Украины	[Евросоюз расширит военную помощь Украине на 3,5 миллиарда евро]
2022/07/20	Путешествия	Россиянам облегчили планирование отдыха на южных курортах	[Россиян предупредили о многочасовых пробках по пути на южные курорты, На юге России в 24-й раз продлили запрет на выполнение полетов]

Рис. 1. Фрагмент массива новостей

Результаты по второму и третьему этапу методики объединены на одном рисунке – продемонстрирован список сформированных социальных категорий, а также их абсолютное количество в тысячах в разрезе четырех лет (рис. 2).

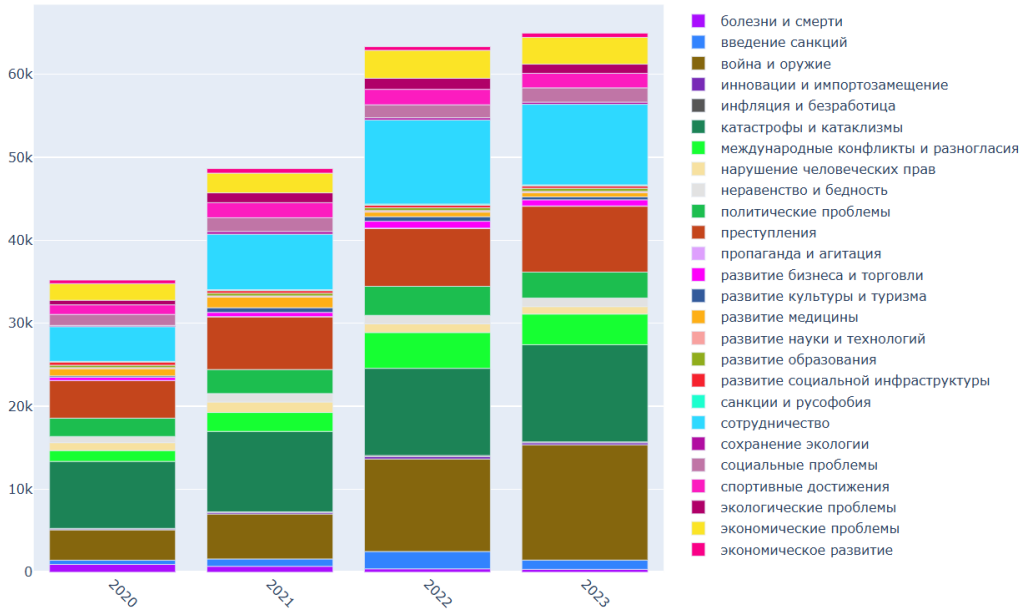


Рис. 2. Распределение количества новостей по категориям и годам

На рисунке представлен граф, построенный для 2021 года: при наведении на узлы на аналитической панели можно увидеть название соответствующих категорий (рис. 3). Размер узлов-категорий пропорционален количеству новостей по данной категории за взятый период. Толщина ребер пропорциональна количеству новостей, в которых упоминались пары новостей, принадлежащие соответствующим категориям.

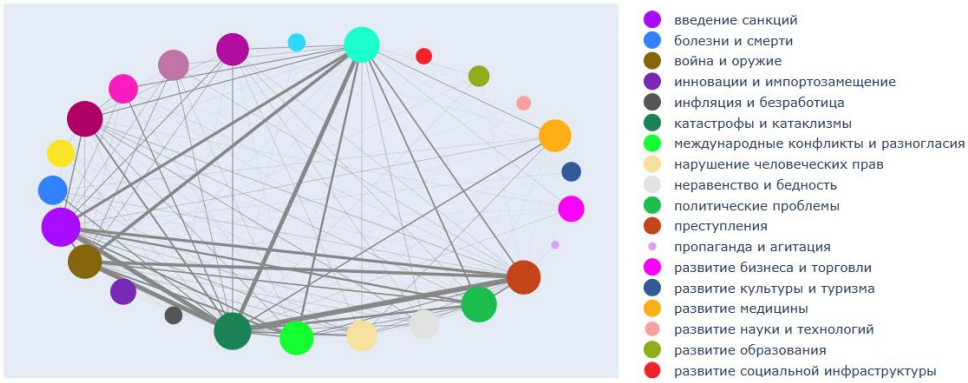


Рис. 3. Граф категорий за 2021 год

Результат определения топ-5 центральных узлов-категорий для 2021 года представлен в таблице, см. Табл. 3.

ТАБЛИЦА 3
ТОП-5 САМЫХ ЦЕНТРАЛЬНЫХ УЗЛОВ-КАТЕГОРИЙ ЗА 2021 ГОД

Категория	**Центральность**
Введение санкций	1
Катастрофы и катаклизмы	0.958
Политические проблемы	0.917
Сотрудничество	0.917
Экономические проблемы	0.917

На рисунке представлен пример графика для сегментации временного ряда по категории «развитие науки и технологий» за 2023 год (рис. 4), где на оси x указана дата, а на оси y – количество новостей. В данном случае выбрано объединение в сегменты размером не меньше недели. Параметр регуляризации подбирается экспериментально, чтобы предотвратить переобучение алгоритма. Один из популярных подходов, который

использован в работе – брать регуляризацию как два логарифма от длины исходного ряда. Чем меньше значение параметра регуляризации, то есть меньше «штраф», тем больше сегментов выделяется.

Каждый сегмент представляет собой отдельный период особой новостной активности: есть сегменты с маленькой амплитудой разброса частотной активности, а есть с высокой. Можно предположить, что в первом случае не было сильных инфоповодов. Вертикальной красной полосой отмечены те точки перегиба, то есть даты, которые являются границами сегмента и обозначают изменение поведения временного ряда на определенном промежутке. Помимо этого, красным цветом выделены топ-5 точек с максимальными значениями новостной активности за весь взятый период. Поскольку график интерактивный, нет необходимости выводить даты, которые совпадают с точками максимума, и их можно увидеть, наведя курсор на соответствующую точку. По данному графику можно определить те даты, в которые произошли самые интересные с точки зрения анализа события, которые привели к появлению инфоповода. Более информативное описание этих событий в указанные даты приведено на этапе анализа ключевых слов.

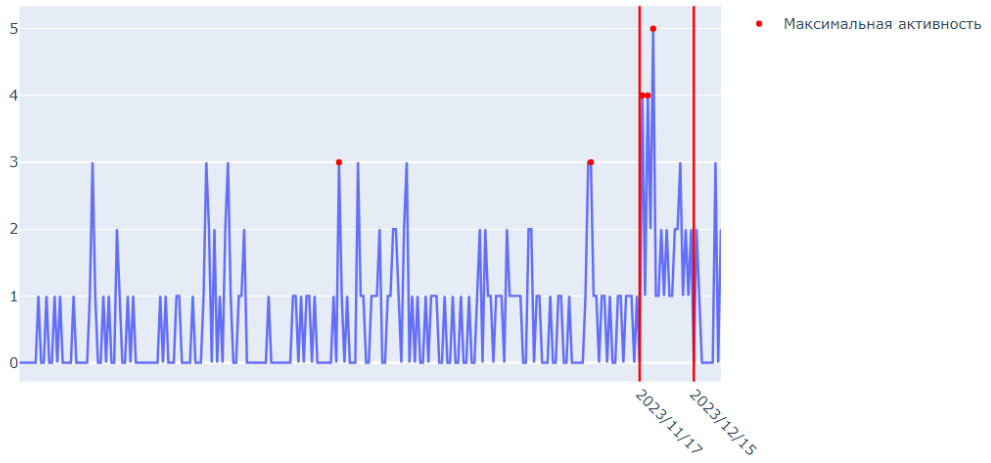


Рис. 4. График с результатами сегментации временного ряда для категории «развитие науки и технологий» за 2023 год

В результате работы алгоритма кластеризации на графике выводится тот кластер временных рядов, который содержит выбранную в фильтре аналитической панели категорию. На рисунке представлен пример кластера для категории «развитие науки и технологий» за 2023 год), где на оси x указана дата, а на оси y – количество новостей (рис. 5). Благодаря построению данного графика возможно проследить в разрезе лет то, как меняется корреляция временных рядов различных категорий, определить причины, по которым поведение одних категорий в определенный период схоже с выбранной, а в другой период разительно отличается.

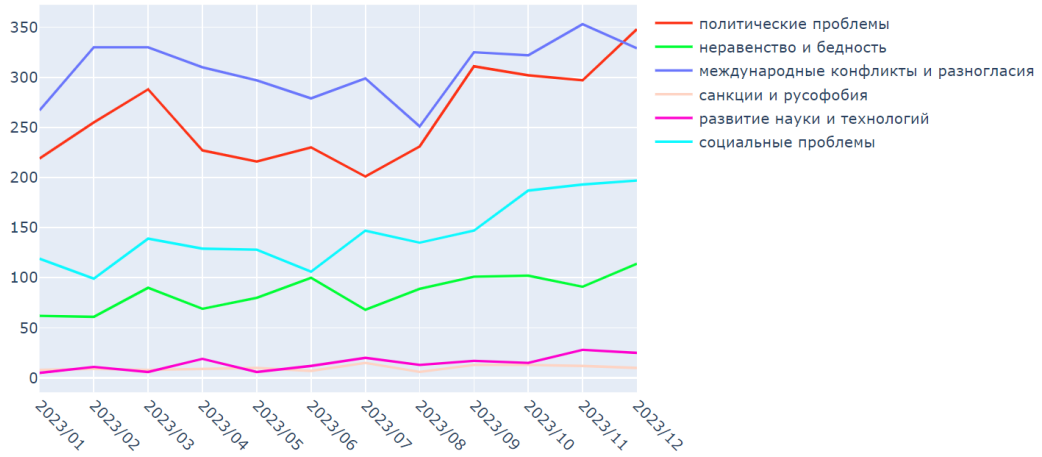


Рис. 5. График с результатами кластеризации временного ряда для категории «развитие науки и технологий» за 2023 год

Метод, применяемый для более информативного описания событий, произошедших в даты, совпадающий с точками экстремума или перегиба на временном ряде – выделение ключевых терминов. На рисунке представлен

результат работы алгоритма для категории «развитие науки и технологий» за 2023 год, где на оси x указана дата, на оси y – суммарное количество ключевых терминов (рис. 6).

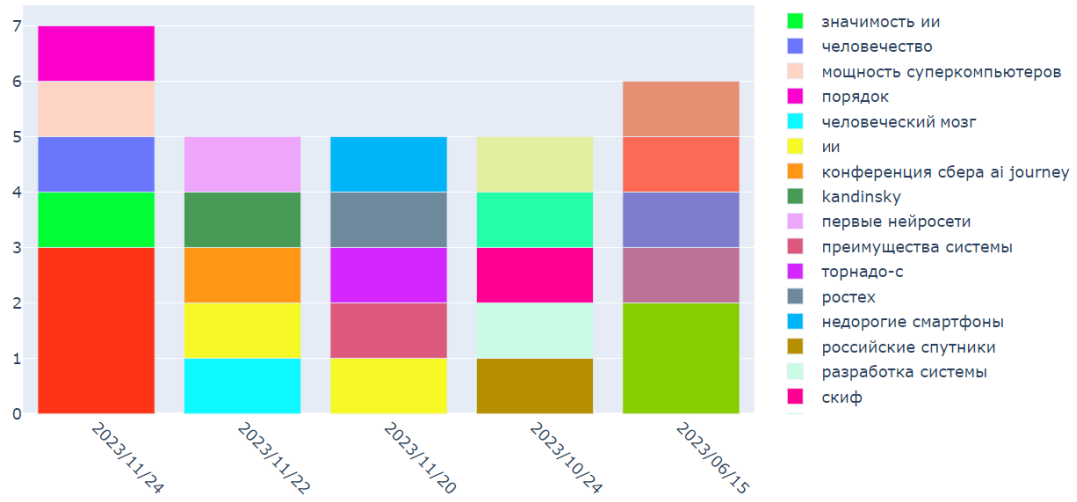


Рис. 6. График с результатами определения ключевых терминов для категории «развитие науки и технологий» за 2023 год

Аналитическая панель реализована на языке программирования Python с использованием библиотеки dash, результаты выводятся на локальном сервере. Пример аналитической панели для категории «развитие науки и технологий» за 2023 год представлен на рисунке (рис. 7).

В качестве метода построения социального прогноза выбрано написание аналитической записки по тенденциям развития сферы «развитие науки и технологий» в России на ближайшие годы на основе анализа ретроспективных данных из аналитической панели. Для удобства аналитическая записка представлена в виде таблицы ниже с анализом тенденций, выдвижением наиболее вероятных сценариев и рекомендаций как для государства, так и для бизнеса, см. Табл. 4.

V. Обсуждение результатов

Ивент-анализ является адаптируемой технологией для использования в качестве инструмента социального прогнозирования, поскольку выбор системы классификации данных и методов на этапе аналитического исследования зависит от поставленных задач. Основным результатом работы стала авторская модификация алгоритма ивент-анализа, в частности, для его использования в качестве инструмента социального прогнозирования. Для выдвижения конкретных прогнозов необходима экспертная оценка, однако она должна опираться на проанализированные с различных сторон данные, что подтверждает актуальность и значимость проведенного исследования.

По сравнению с классической методикой ивент-анализа, где главным методом на этапе аналитического сравнения, как правило, выступает вычисление частотного распределения категорий в определенные периоды, добавлены дополнительные подходы, которые позволяют рассмотреть взаимосвязи и тенденции более подробно. Модифицированные этапы сделали возможным более обоснованно отвечать на главные вопросы, поставленные в классической методике: об оценке сюжета, субъектов-инициаторов, объектов, времени события.

Выводы, сделанные на основе данных из интерактивных графиков, позволили определить тенденции поведения сферы «развитие науки и технологий» по соответствующим годам. Достоверность полученных результатов оценена посредством ручной верификации выявленных тенденций путем сравнения с ранее составленными экспертами выводов. Это доказывает, что адаптированная методика ивент-анализа обладает достаточной степенью достоверности и полноты, что позволяет использовать ее возможности как одного из этапов в проведении масштабных социальных исследований и прогнозов. Разработанные на ее основе наиболее вероятные сценарии развития объекта социального прогнозирования и рекомендации позволят государственным органам вовремя предпринимать меры по поддержке актуальных исследований частных организаций, а бизнесу – искать новые и прибыльные ниши для развития. Это приведет к синхронизации действий государства и бизнеса, их взаимовыгодному сотрудничеству и ускоренному развитию различных сфер жизнедеятельности.

Основное преимущество разработанной методики в том, что ее можно адаптировать под некоторые задачи социального прогнозирования, где требуется классификация данных по различным категориям и их разносторонний анализ как в статическом, так и в динамическом состоянии. В качестве дальнейшей работы планируется увеличение количества источников сбора данных, улучшение качества классификации данных по категориям и усовершенствование методов, используемых на этапе аналитического сравнения.

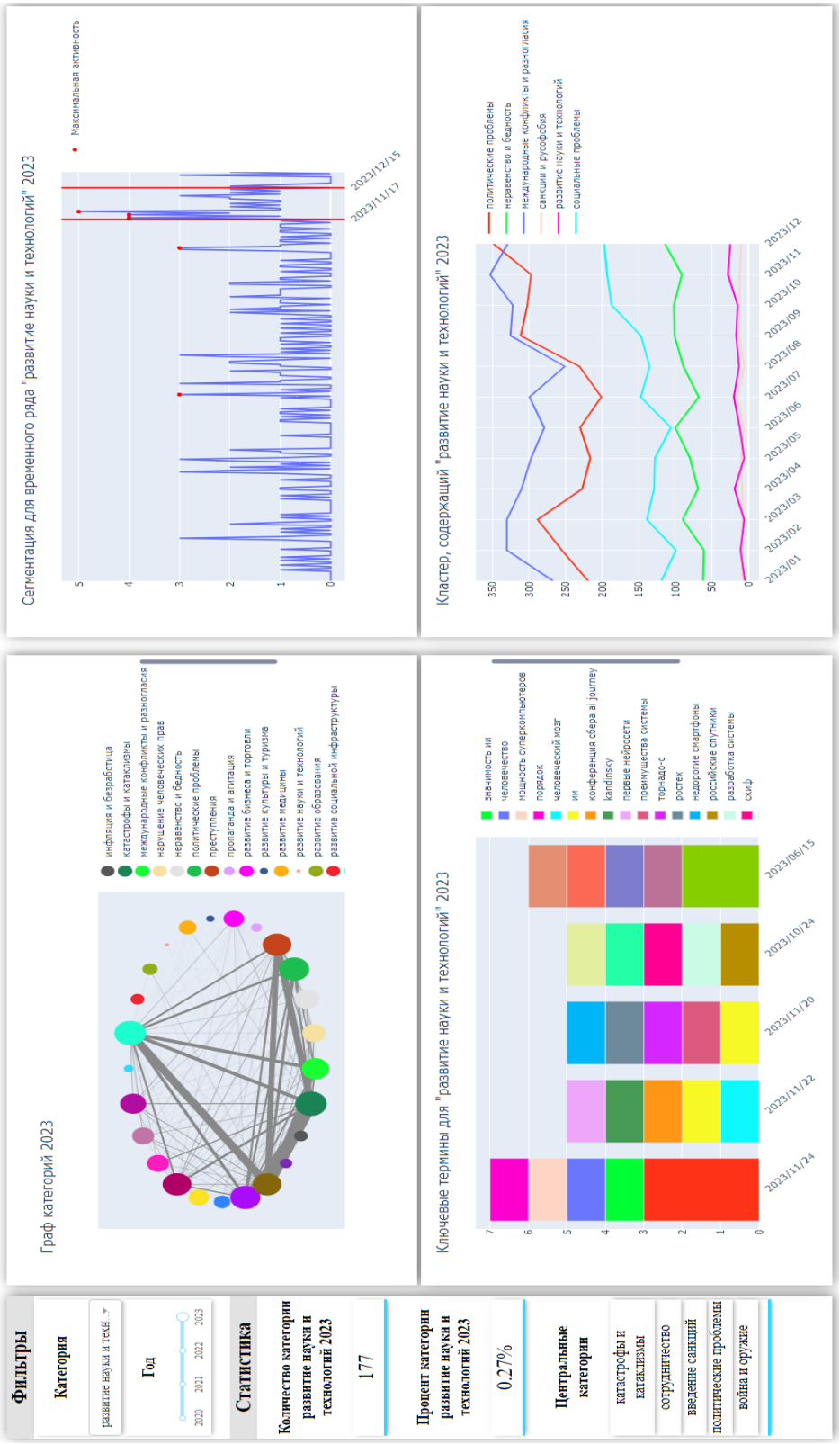


Рис. 7. Результат построения итоговой аналитической панели с фильтрами «Категория» – «развитие науки и технологий», «Год» – 2023.

ТАБЛИЦА 4
СОЦИАЛЬНЫЙ ПРОГНОЗ ПО КАТЕГОРИИ «РАЗВИТИЕ НАУКИ И ТЕХНОЛОГИЙ» В РОССИИ

Тенденции по годам			**Прогнозы и рекомендации**		
	Тенденции	**Источники**	**Прогноз**	**Государство**	**Бизнес**
2020	Медицина: развитие вакцинационных технологий. Экология: технологии для экологически чистого строительства, инновации в материаловедении, освоение Арктической зоны.	Ключевые термины и сегментация временных рядов (ВР): вакцины, зондирование ионосферы, морозостойкий бетон. Графы и кластеризация ВР: экологическое развитие, развитие медицины, сотрудничество.	Активное развитие своих технологий, направленных на укрепление национальной безопасности и суверенитета.	Повышение государственного финансирования в области технологий безопасности, киберзащиты и разработки высокотехнологичных продуктов. Внедрение мер поддержки национальных производителей и исследовательских центров.	Инвестирование в разработку и инновации, направленные на укрепление технологической базы страны. Сотрудничество с государственным и заказчиками и научными учреждениями.
2021	Образование: цифровизация образовательных процессов, развитие технопарков и инновационных кластеров. Медицина: Расширение применения ИИ для диагностики и лечения, продолжение работы по усовершенствованию и распространению российской вакцины.	Ключевые термины и сегментация ВР: вакцины, технопарки, искусственный интеллект. Графы и кластеризация ВР: развитие социальной инфраструктуры, развитие медицины, развитие образования.	Ускоренное развитие цифровых технологий, включая искусственный интеллект, в ответ на потребности в цифровой трансформации экономики и общества.	Формирование национальных стратегий цифровизации, включая разработку законодательства и инфраструктуры для поддержки цифровых технологий. Поддержка исследований в области искусственного интеллекта и его приложений.	Инвестирование в разработку цифровых платформ, развитие инновационных продуктов на основе искусственного интеллекта. Активное использование цифровых технологий для улучшения эффективности бизнес-процессов.

2022	Промышленность: развитие высокотехнологичных отраслей, включая ВПК, разработка новых технологий для повышения производительности. Инновации и сотрудничество: укрепление технологического суверенитета через развитие отечественных технологий.	Ключевые термины и сегментация ВР: российские беспилотники и радиолокационные спутники, технологический суверенитет России, искусственный интеллект. Графы и кластеризация ВР: экономическое развитие, развитие бизнеса и торговли, инновации и импортозамещение.	Активное развитие международных научных и технологических инициатив, особенно с участием стран Восточной Азии.	Поддержка международных научных проектов и программ обмена, стимулирование технологического трансфера и совместных исследовательских проектов.	Активное участие в международных научных и технологических партнерствах, создание совместных исследовательских и инновационных лабораторий с зарубежными партнерами.
2023	Безопасность: развитие технологий в области безопасности и обороны, включая кибербезопасность и радиолокационные технологии. IT: инвестиции в разработку искусственного интеллекта, особенно в областях кибербезопасности и автоматизации процессов.	Ключевые термины и сегментация ВР: искусственный интеллект, суперкомпьютеры, российские радиолокационные спутники. Графы и кластеризация ВР: политические проблемы, введение санкций, международные конфликты и разногласия.	Усиление экологических проблем и климатических изменений стимулирует развитие экологически чистых технологий и устойчивых инноваций.	Введение стимулов и льгот для разработки и внедрения экологически чистых технологий. Финансирование исследований в области экологии и устойчивого развития.	Инвестирование в разработку зеленых технологий и инноваций. Активное участие в экологических инициативах и проектах по устойчивому развитию.

VI. Выводы и заключение

В ходе проведенного исследования разработана и протестирована адаптированная методика ивент-анализа, которая продемонстрировала свою эффективность в контексте социального прогнозирования. Основные выводы, сделанные в результате работы, можно обобщить следующим образом. Адаптированная методика ивент-анализа показала высокую степень универсальности, позволяя применять ее в различных областях социального исследования. Она может быть адаптирована под специфические задачи, что делает ее подходящей для анализа как краткосрочных, так и долгосрочных социальных явлений.

Внедрение новых аналитических инструментов, таких как анализ графов и временных рядов, значительно повысило уровень достоверности получаемых данных. Ручная верификация выявленных тенденций подтвердила, что результаты анализа соответствуют реальным событиям и их контексту.

Создание интерактивных графиков с возможностью фильтрации данных улучшило восприятие информации и обеспечило возможность динамического сравнения различных сценариев. Это позволяет не только визуализировать результаты, но и углубить анализ, выявляя скрытые взаимосвязи и закономерности.

Результаты исследования могут быть использованы как для государственных органов, так и для бизнеса. Сценарии развития событий, сформулированные на основе полученных данных, могут служить основой для принятия стратегических решений. Это способствует синхронизации действий различных субъектов и обеспечению взаимовыгодного сотрудничества.

В дальнейшем планируется расширение источников данных, включая новые информационные потоки и социальные медиа. Это позволит улучшить классификацию событий и усовершенствовать аналитические методы. Также стоит рассмотреть возможность интеграции машинного обучения для автоматизации прогнозов и повышения его точности.

Таким образом, адаптированная методика ивент-анализа представляет собой мощный инструмент для социального прогнозирования, который сочетает в себе традиционные подходы с современными аналитическими технологиями. Ее использование может значительно улучшить качество прогнозов и способствовать более эффективному взаимодействию между государственными структурами и бизнесом в области научных исследований и технологий.

Список литературы

1. Мальцева, А. В. Использование методики event-анализа для изучения процессов на рынке труда / А. В. Мальцева и др. // Вестник евразийской науки. – 2012. – № 3 (12). – С. 12.

2. Новоселов, М. В. Горизонты социологического применения EVENT-анализа / М. В. Новоселов // Социальные и гуманитарные науки: теория и практика. – 2018. – № 1 (2). – С. 497-506.

3. Азаров, А. А. Predictor Mining: применение методов интеллектуального анализа данных в задачах социального компьютинга / А. А. Азаров // Информатика и автоматизация. – 2013. – № 26. – С. 136-161.

4. Кудрявцева, В. И. Особенности социального прогнозирования в международных отношениях / В. И. Кудрявцева // Вестник БГУ. – 2004. – № 2. – С. 67-71.

5. Ortega, J. L. «How do media mention research papers? Structural analysis of blogs and news networks using citation coupling» / J. L. Ortega // Journal of Informetrics. – 2021. – № 15 (3). – P. 101175. – doi.org/10.1016/j.joi.2021.101175.

6. Blokh, I. «News clustering based on similarity analysis» / I. Blokh, V. Alexandrov // Procedia Computer Science. – 2017. – № 122. – P. 715-719. – doi.org/10.1016/j.procs.2017.11.428.

7. Нехамкин, А. Н. Социальное прогнозирование: достижения, недостатки, пути совершенствования / Нехамкин А. Н., Нехамкин В. А. // Вестник Московского государственного областного университета. Серия: Философские науки. – 2020. – №. 2. – С. 57-68.

8. McClelland, C. A. Let the user beware / C. A. McClelland // International Studies Quarterly. – 1983. – № 27(2). – P. 169-177. – doi: 10.2307/2600544.

9. Лента.ру: сайт / учредитель ООО «МИНС». – Москва. – Обновляется в течение суток. – URL: https://lenta.ru/ (дата обращения: 20.08.2024).

10. A, Daud. Knowledge discovery through directed probabilistic topic models: a survey / Daud A., Li J., Zhou L., Muhammad F. // Frontiers of Computer Science in China. — 2010. — № 2 (4). — P. 280-301. – doi:10.1007/s11704-009-0062-y.

11. Z, Ji. Zero-shot classification with unseen prototype learning / Ji Z., Cui B., Yu Y., Pang Y., Zhang Z. // Neural computing and applications. – 2023. – P. 1-11. – doi: 10.1007/s00521-021-05746-9

12. Hugging Face: сайт / MoritzLaurer/mDeBERTa-v3-base-mnli-xnli. – Обновляется в течение суток. – URL: https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli (дата обращения: 20.08.2024).

13. Ulizko, M. S. Visual analytics of twitter and social media dataflows: a casestudy of covid-19 rumors / M. S. Ulizko, E. V. Antonov, M. A. Grigorieva, E. S. Tretyakov, R. R. Tukumbetova, A. A. Artamonov // Scientific Visualization. – 2021. – № 4 (13). – P. 144-163.

14. Camacho, D. The four dimensions of social network analysis: An overview of research methods, applications, and software tools / D. Camacho, Á. Panizo-LLedot, G. Bello-Orgaz, A. Gonzalez-Pardo, E. Cambria // Information Fusion. – 2020. – № 63. – P. 88-120. – doi: 10.1016/j.inffus.2020.05.009.

15. Killick, R. Optimal detection of changepoints with a linear computational cost / R. Killick, P. Fearnhead, I. A. Eckley // Journal of the American Statistical Association. – 2012. – № 107(500). – P. 1590-1598. – doi: 10.1080/01621459.2012.737745.

16. Ulizko, M. S. Clustering Thematic Information in Social Media / M. S. Ulizko, A. A. Artamonov, J. E. Fomina, E. V. Antonov, R. R. Tukumbetova // Proceedings of the International Conference on Computer Graphics and Vision «Graphicon». – 2022. – № 32. – P. 403-413.

17. Мотовских, Л. В. Выделение ключевых слов для классификации текстов / Л. В. Мотовских // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2020. – № 9 (838). – С. 235-242.

18. Dabbas, E. Interactive Dashboards and Data Apps with Plotly and Dash: Harness the power of a fully fledged frontend web framework in Python–no JavaScript required. / E. Dabbas. – Packt Publishing Ltd, 2021. – 336 p. – ISBN 978-1-80056-891-4.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-451-465

Автоматизация визуального контроля печатных плат
Automation of visual inspection of printed circuit boards

Т.С. Ходатаева[1]
[1]Федеральное государственное бюджетное образовательное учреждение высшего образования Марийский государственный университет, г.Йошкар-Ола, Российская Федерация
tkhodataeva@gmail.com
T. S. Khodataeva[1]
[1]Mari State University, Federal State Budgetary Educational Institution of Higher Education, Respublika Marii El, g.Ioshkar-Ola, Russian Federation

Аннотация.Обнаружение дефектов поверхности печатных плат необходимо для управления качеством производства печатных плат. Однако автоматическое обнаружение дефектов токопроводящего рисунка печатных плат по-прежнему остается сложной задачей, поскольку даже в пределах одной категории дефекты имеют большие различия в морфологии и рисунке. Хотя для решения этих проблем было создано множество детекторов на основе компьютерного зрения, остаётся актуальной проблема подбора детектора для конкретной задачи, сочетающего в себе показатели высокой точности обнаружения, высокой скорости обнаружения и низкого потребления памяти одновременно. Для изучения этих вопросов было проведено исследование на основе глубокого обучения. Для проверки эффективности и результативности исследовались разные версии свёрточной нейронной сети архитектуры YOLO, обученные на собственном наборе данных. Необходимость создания набора данных для обучения сети потребовало разработки устройства получения изображений. Формирование набора данных происходило на заводе по производству печатных плат с использованием специально разработанного устройства получения изображений. В результате работы получены результаты сравнения классической модели четвёртой версии YOLO и одной из последних восьмой версии семейства YOLO, обученных на цветных изображениях и изображениях в градациях серого.

Ключевые слова: автоматическая оптическая инспекция, обработка изображений, сверточные нейронные сети, глубокое машинное обучение, компьютерное зрение, обнаружение дефектов поверхности

Abstract. Detection of printed circuit board surface defects is essential for quality management of printed circuit board production. However, detection of defects in the conductive pattern of the printed circuit board still remains a serious challenge, since even within the same category, defects have large differences in morphology and pattern. Although many computer vision-based detectors have been created to solve the problems, the problem of choosing a detector for these tasks that combines high detection accuracy, high detection speed and low memory consumption at the same time remains relevant. To study these issues, a study based on deep learning was conducted. Different versions of the YOLO convolutional neural network were studied on a proprietary dataset to test the efficiency and effectiveness. The need to create a dataset for training the network required the development of an image acquisition device. Formation of a dataset at a printed circuit board manufacturing plant using a specially designed image acquisition device. As a result of the work, the results of comparing the classic model of the fourth version of YOLO and one of the latest eighth versions of YOLO, trained on color and grayscale images, were obtained.

Keywords: automatic optical inspection, image processing, convolutional neural networks, deep machine learning, computer vision, surface defect detection

I. Введение

К настоящему времени было проведено и опубликовано значительное количество прикладных исследований [1] с целью использования технологии обработки изображений для обнаружения дефектов печатных плат. Начиная с 2014 года в практику детектирования дефектов печатных плат стали активно внедряться алгоритмы глубокого обучения, в основе которых в качестве метода извлечения признаков лежат сверточные нейронные сети CNN (Convolutional Neural Network)[2] различной архитектуры.

Обнаружение объектов на основе глубокого обучения можно разделить на две категории: двухэтапное и одноэтапное обнаружение. Серия RCNN (Region Based Convolutional Neural Networks) [3], Fast RCNN [4] и Faster RCNN [5]) представляет собой двухэтапный алгоритм, точность которого значительно превосходит точность

многих других алгоритмов обнаружения. Однако этот тип подхода требует более высоких вычислительных затрат, что увеличивает время обработки.

В одноэтапной категории можно выделить SSD (Single Shot MultiBox Detector) [6] и алгоритм YOLO (You Only Look Once) [7]. YOLO решает задачу обнаружения объектов как задачу регрессии и отображает положение и классификацию объекта в сквозной сети за один шаг. Благодаря скорости обнаружения в настоящее время это один из наиболее часто используемых алгоритмов.

Для решения проблемы обнаружения дефектов печатных плат многие исследователи брали за основу именно CNN YOLO, архитектура которой сформировалась к 2019 году.

В статье [8] авторами был предложен экономичный детектор на основе YOLOv4[9] под названием YOLOv4-MN3. Авторы пытались достигнуть более высокой точности и скорости обнаружения дефектов и повысить производительность вычислений. Они заменили магистральную сеть CSPDarknet-53 на более лёгкую MobileNetV3[10] с небольшим количеством параметров, оптимизировали функции активации и потерь в сети шеи и предиктивной сети. Улучшенный детектор достиг высокой производительности при обнаружении поверхностных дефектов печатных плат, набрав 98,64% по средней точности (mAP@50) при 56,98 кадрах в секунду (FPS). Улучшенный YOLOv4-MN3 сократил пространство параметров YOLOv4 с 63,96 M до 39,59 M. Авторы создали набор данных из 2008 изображений, содержащий шесть типов дефектов поверхности такие как неровные или прерывистые линии, загрязнения, царапины, повреждения, потеря отверстий и избыточное заполнение маслом.

В работе [11] авторы предложили улучшенную сеть под названием YOLO-MBBi, взяв за основу YOLOv5s[12]. Целью работы было устранение недостатков существующих методов обнаружения поверхностных дефектов печатных плат, таких как их низкая точность и плохая производительность в реальном времени. Авторы улучшили пять аспектов YOLOv5s: магистральную сеть, модуль внимания, слияние признаков, функцию потерь и изменили блок свертки. Авторы заменили модули CBS и C3 магистральной сети YOLOv5s на модуль MBConv, заимствованный из CNN EfficientNet [13]. Для улучшения обучения признакам объектов в магистральной сети добавили блок внимания CBAM (Convolutional Block Attention Module) [14]. В сетях пирамид признаков вместо FPN [15] использовали BiFPN[16]. BiFPN был добавлен в сеть слияния признаков детектора головы. Некоторые свертки в сети обнаружения головы заменены свертками по глубине [17], а функцию потерь CIoU[18], которая изначально использовалась в YOLOv5, заменили функцией потерь SIoU [19] для улучшения обучения. Для проверки эффективности YOLO-MBBi при обнаружении дефектов поверхности печатной платы проводились сравнения с другими архитектурами CNN: TDD-Net [20], YOLOv4, YOLOv5s и YOLOv7[21] на общедоступном наборе данных DeepPCB[22], содержащим 1500 изображений печатных плат с шестью типами дефектов поверхности: отверстия, короткие замыкания, мышиные укусы, шпоры, ложная медь и точечные отверстия.

В работе [23] исследователи сконструировали модель обнаружения дефектов Double-Cross-Residual[24] на основе архитектуры YOLO. Все усилия авторов были направлены на извлечение признаков как относительно крупных, так и мелких целевых дефектов проводящего рисунка печатных плат. При создании модели для улучшения способности извлечения признаков небольших целевых дефектов и устранения проблемы исчезновения градиента для магистральной части сети использовалась идея "пропущенных соединений" - Residual block [24], реализованная как структура под названием Cross-Residual blockbody. Для улучшения механизма слияния признаков в разных масштабах между сетью извлечения признаков и модулем слияния признаков была разработана дополнительная объединенная сверточная структура PCR (Pooled Convolutional Residual Structure). Для повышения точности прогнозирования как относительно крупных, так и мелких целевых дефектов был доработан модуль слияния признаков - разработана структура SDDT-FPN с изотропной двойной пирамидой признаков. Для уменьшения фокуса на второстепенной информации о признаках и усиления фокуса на важной информации была добавлена структура C5ECA на базе сети Effificient Channel Attention (ECAnet) [13]. Для проверки эффективности обнаружения и прогнозирования модели DCR-YOLO были проведены сравнительные эксперименты с шестью основными моделями обнаружения целей: YOLOv3[26], YOLOv4, YOLOv4-tiny, YOLOv5s, YOLOv5m и YOLOv7-tiny [21]. В работе использовался общедоступный набор данных дефектов печатных плат из открытой лаборатории интеллектуальной робототехники.

В работе [27] был предложен улучшенный алгоритм на основе YOLOx[28] и представлен быстрый метод обнаружения дефектов печатных плат на основе ПЛИС [29]. Предложенный алгоритм, названный YOLOx-Plus, имеет более сложную сетевую структуру, чем YOLOx. Во-первых, YOLOx-Plus вводит обобщенный модуль внимания SimAM (Simple Attention Mechanism) [30] в магистральную сеть, который позволяет сети улучшить обучение модели без увеличения количества параметров и подавлять несущественные признаки. Во-вторых, его шея становится сложной структурой, состоящей из двух сетей улучшения признаков: FPN (Feature pyramid network) [15] и PANet (Path aggregation network) [**Ошибка! Источник ссылки не найден.**], значительно у лучшающие возможности слияния и извлечения признаков. Наконец, YOLOx-Plus добавляет в существующую

сеть улучшения признаков FPN+PAN модуль SimOTA[**Ошибка! Источник ссылки не найден.**], что позволяет п олучать признаки более высокого порядка, улучшать пространственные взаимодействия задач нисходящего потока, объединять признаки разных масштабов и улучшать результаты предиктивной классификации модели. Путем квантования параметров сети и проектирования ускорителя ПЛИС алгоритм ускоряется для достижения точного и быстрого обнаружения.

Практически во всех работах сравнительные эксперименты обнаружения дефектов поверхностного рисунка печатных плат проводились на ранних версиях YOLO, стремясь улучшить скорость и точность детектирования мелких дефектов. Для обучения сети использовались общедоступные наборы данных, где зачастую дефекты создавались с помощью графических редакторов. Наборы данных формировались на основе изображений печатных плат с финишным защитным покрытием, когда платы приобретают зелёный или синий матовый цвет. Исследователи ограничивались шестью классами дефектов.

В условиях реального производства оптический контроль производится на начальном этапе изготовления печатных плат, после травления и снятия фоторезиста. Оптический контроль обнаруживает порядка 200 видов дефектов. Отсутствие в открытом доступе больших наборов данных печатных плат с дефектами, нужными для обучения CNN, вызвало необходимость создания собственного набора данных. Необходимо было провести исследования и предложить метод определения дефектов медного проводящего рисунка печатных плат на начальном этапе их изготовления после травления и снятия фоторезиста. Требовалось также расширить количество классов детектируемых дефектов.

Содержание статьи организовано следующим образом. Раздел II содержит постановку задачи. В Разделе III кратко рассмотрены архитектуры сетей YOLO, которые будут использованы в экспериментах. Раздел IV знакомит с созданием собственного набора данных, который включает в себя устройство получения изображений дефектов поверхности, сбор изображений дефектов, аугментацию данных и маркировку. В разделе V представлены результаты обучения и обнаружения поверхностных дефектов печатных плат. Раздел IV содержит выводы.

II. Постановка задачи

Целью исследовательской работы является выработка требований, проведение технического проектирования и создание опытно-конструкторского образца системы оптического контроля проводящего рисунка печатных плат с использованием относительно недорогого оборудования.

Работа проводится в несколько этапов:

a) анализ современных подходов, реализованных в различных системах АОИ (автоматической оптической инспекции), используемых в электронике. Это было сделано в работе [**Ошибка! И сточник ссылки не найден.**];
b) апробация выбранного подхода с использованием сверточной нейронной сети архитектуры YOLO на наборе данных для обучения из открытых источников;
c) сбор реальных изображений дефектов печатных плат на предприятии при участии технологов и создание собственного набора данных;
d) обучение CNN на собственном наборе данных;
e) анализ результатов.

Требовалось оценить затраты по времени и мощность компьютера для решения поставленных задач и специально разработать устройство получения изображений. Оценить работу различных архитектур CNN YOLO на полученных изображениях.

III. Теория

Основной идеей использования CNN при реализации обнаружения дефектов проводящего рисунка является быстрое нахождение областей на плате, имеющих дефект. Основываясь на проведённых исследованиях статья [**Ошибка! Источник ссылки не найден.**], для экспериментов были выбраны две версии YOLO. Одну из самых р анних, но уже с устоявшейся архитектурой, версию YOLOv4 и одну из последних и самых популярных версий YOLOv8[**Ошибка! Источник ссылки не найден.**]. Принятое решение по выбранным архитектурам потребовало у глублённого знакомства с этими архитектурами сетей. Необходимо было понять, за счёт каких структур сеть этой архитектуры может идентифицировать дефекты разных масштабов.

Архитектура YOLO (You Only Look Once) [7] была предложена Джозефом Редмоном в 2015 году и завоевала популярность. Эта архитектура позволяет моделировать процесс обнаружения объектов в реальном времени. Модель основана на разделении задачи обнаружения объектов на две части, то есть на регрессию для определения положения объекта с помощью ограничивающих рамок и на классификацию для определения

класса объекта. Это одноэтапный детектор объектов, который использует возможности сверточных нейронных сетей для обнаружения нескольких объектов на данном изображении. Он накладывает на изображение сетку и для каждой ячейки сетки изображения прогнозирует значения вероятности классов и координаты рамки.

К 2018 году сформировался устоявшийся подход к архитектуре детекторов объектов. Архитектура детекторов объектов стали описывать тремя частями: магистральная сеть (Backbone), шея (Neck) и предиктивная сеть - голова (Head).

Магистральная сеть отвечает за извлечение полезных признаков из входного изображения. Обычно это сверточная нейронная сеть (CNN) обученная на крупномасштабной задаче классификации изображений ImageNet [**Ошибка! Источник ссылки не найден.**].

Шея — это промежуточный компонент, соединяющий магистральную сеть с детектором объектов. Он объединяет и уточняет признаки, извлеченные магистральной сетью, часто фокусируясь на улучшении пространственной и семантической информации в разных масштабах.

Предиктивная сеть (голова) — это последний компонент детектора объектов. Она отвечает за создание прогнозов на основе характеристик, предоставляемых магистральной сетью и шеей, выполняет классификацию и локализацию, генерируя прогнозы для каждого кандидата на объект. В конце отфильтровывает перекрывающиеся прогнозы и сохраняет только самые надежные обнаружения с помощью механизма NMS non-max suppression.

Архитектура модели YOLOv4 [9], анонсированная в 2020 году, базировалась на архитектуре YOLOv3[26] и включала в себя все последние усовершенствования в области детектирования объектов. Магистральная сеть Darknet-53 YOLOv3 была дополнена кросс-ступенчатыми частичными соединениями CSPResNet *[**Ошибка! Источник ссылки не найден.**]*. В качестве функции активации выбрана функция Mish *[**Ошибка! Источник ссылки не найден.**]*. YOLOv4 использует усовершенствованную функция потерь. Схема архитектуры YOLOv4 представлена на рисунке (рис.1).

Модифицированный блок, называемый **SPP (Spatial Pyramid Pooling)***[**Ошибка! Источник ссылки не найден.**]* добавлен между магистралью CSPDarknet-53 и сетью агрегаторов признаков PANet [**Ошибка! Источник ссылки не найден.**]. Блок **SPP** используется для увеличения рецептивного поля и изоляции наиболее важных контекстных особенностей. Блок PANet спользуется для улучшения передачи информации о функциях и повышения производительности обнаружения за счет различных уровней объединения функций, особенно при обнаружении целей небольшого размера.

В качестве предиктивной части сети YOLOv4 остались подходы, реализованные в YOLOv3, предсказания на нескольких размерах сетки. Предиктивный алгоритм работает следующим образом: первый выход использует размер сетки 19×19. Второй выход состоит из конкатенации выхода после (CSPResNet×4) CSPDarknet-53[9] с выходом после (CSPResNet × 8). Карты объектов имеют разные размеры 19×19 и 38×38, поэтому перед конкатенацией есть операция повышения дискретизации (Upsample). Наконец, используя операцию повышения дискретизации (Upsample), третий выход конкатенирует карты объектов размерностью 38×38 и 76×76. Это помогает получить более мелкие детализированные блоки и значительно улучшило предсказание мелких объектов.

Функция потерь YOLOv4 является ключевым компонентом эффективности обнаружения целей и включает три части: достоверность, классификацию и потерю регрессии ограничивающей рамки. YOLOv4 использует перекрестную энтропийную потерю для оценки того, содержит ли ограничивающая рамка, предсказанная моделью, объект, и измерения точности его прогноза. Для классификации YOLOv4 также использует перекрестную энтропийную потерю для расчета потерь категорий. Для более точного пределения координаты поля прогнозирования модель YOLOv4 вводит потерю CIoU [188], заменив традиционную потерю IoU(Intersection over Union). IoU количественно определяет перекрытие между предсказанным ограничивающим прямоугольником областью и истинным ограничивающим прямоугольником. CIoU одновременно учитывает область перекрытия, расстояние от центральной точки и соотношение сторон.

В YOLOv4 реализован продвинутый подход к обучению. Помимо обычных дополнений, таких как случайная яркость, контрастность, масштабирование, обрезка, переворачивание и вращение, авторы реализовали мозаичное дополнение, которое объединяет четыре изображения в одно, позволяя обнаруживать объекты вне их обычного контекста, а также уменьшая необходимость в большом размере мини-пакета для пакетной нормализации. Для регуляризации они использовали DropBlock *[**Ошибка! Источник ссылки не найден.**]*, который работает как замена Dropout *[**Ошибка! Источник ссылки не найден.**]* но для сверточных нейронных сетей, а также для сглаживания меток классов *[**Ошибка! Источник ссылки не найден.**,**Ошибка! Источник ссылки не найден.**]*. Для детектора они добавили потерю CIoU *[18]* и перекрестную мини-нормализация (CmBN) для сбора статистики по всей партии, а не по отдельным мини-партиям, как при обычной нормализации партии *[**Ошибка! Источник ссылки не найден.**]* .

YOLOv4-tiny — это облегченная версия YOLOv4, на основании которой создаётся структура сети YOLOv4-tiny. В YOLOv4-tiny устранены узкие места, которые требуют большего объема вычислений в модуле CSP-блока (рис.2.). В модуле CSP-блока YOLOv4-tiny использует функцию LeakyReLU в качестве функции активации для упрощения процесса вычислений.

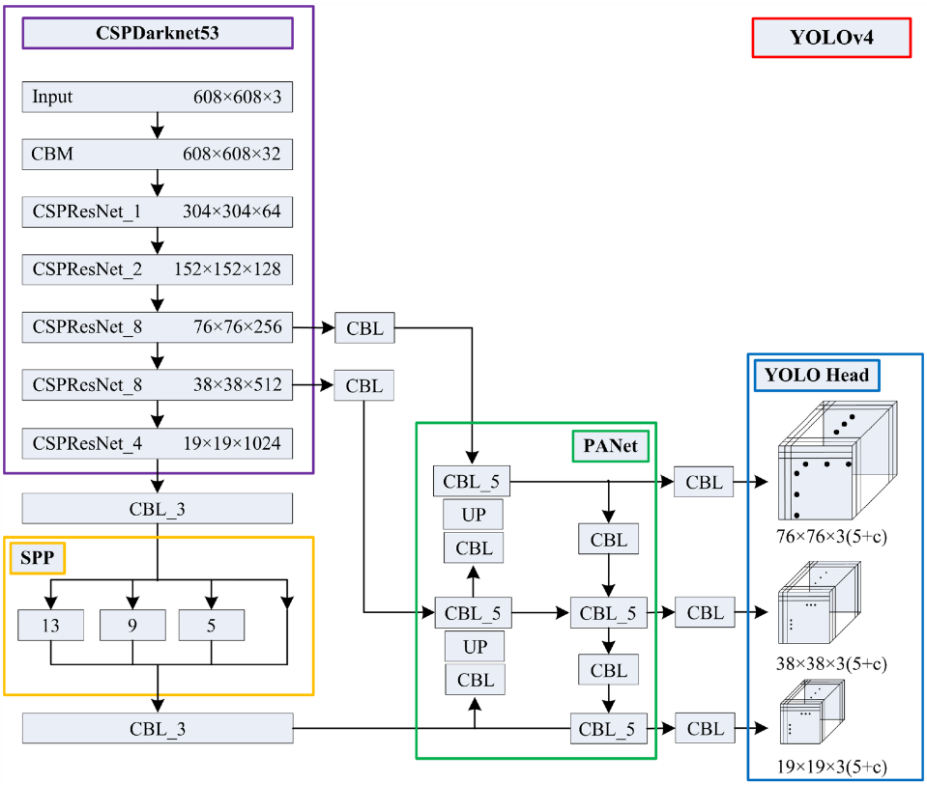


Рис. 1. Схема архитектуры YOLOv4. Модули на схеме: CBM : свертка + пакетная нормализация + активация Mish, **CBL** : свертка + пакетная нормализация + LeakyReLU, UP : повышающая дискретизация (Upsample); **SPP** : пространственное пирамидальное объединение и PANet : сеть агрегации путей.
(Схема заимствована *[**Ошибка! Источник ссылки не найден.**]*).

Архитектура модели YOLOv8 (Детальную схему архитектуры YOLOv8 надо смотреть на сайте [**Ошибка! Источник ссылки не найден.**].)

Основой для модели YOLOv8 (выпуск январь 2023 года) послужила YOLOv5 *[12]*. YOLOv5 *была выпущена через пару месяцев после выхода YOLOv4 и воплотила с некоторыми улучшениями все подходы и наработки YOLOv4. В* YOLOv5 *в области* магистральной сети блоки свёртки CSPLayer, которые аналогичны блокам CSPResNet модели *YOLOv4, были усилены конструкцией bottleneck, предложенной в* Resnet50*[45]*, добавлена пакетная нормализация (BN) и функция активации Mish заменена на функцию активации SiLU *[**Ошибка! Источник ссылки не найден.**]*. В версии YOLOv8 произошли изменения в CSPLayer. В этот модуль была добавлена вторая свёртка. CSPLayer стал называться модулем C2f. Модуль C2f (CSP с двумя свертками) объединяет высокоуровневые функции с контекстной информацией для повышения точности обнаружения.

Слой SPPF (Spatial Pyramid Pooling Fusion, быстрое пространственное пирамидальное объединение), расположенный между магистральной сетью и сетью детектора объектов, направлен на ускорение вычислений сети путем объединения признаков разных масштабов в карту признаков фиксированного размера. К этому блоку применили те же самые подходы, что и для блоков CSPLayer. За каждой сверткой следует пакетная нормализация (BN) и функция активации SiLU*[**Ошибка! Источник ссылки не найден.**]*. Шея использует SPPF и модифицированный CSP-PANet.

В качестве предиктивная части сети, аналогично YOLOv4, реализованы предсказания на нескольких размерах сетки. YOLOv8 использует модель без якорей. Это означает, что она напрямую предсказывает центр объекта, а не смещение от известного блока привязки. В выходном слое YOLOv8 используется сигмоидная функция[47] в качестве функции активации для оценки вероятности того, что ограничивающий прямоугольник содержит объект, и функцию softmax для преобразования выходов модели в вероятности принадлежности объектов каждому возможному классу. Для локализации области детектируемого объекта YOLOv8 использует CIoU *[18]*

и DFL *[**Ошибка! Источник ссылки не найден.**]* в качестве функций потерь. Эти функции потерь улучшили производительность обнаружения объектов, особенно при работе с более мелкими объектами.

В категории моделей YOLOv8 имеется пять моделей (YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, YOLOv8x) для обнаружения, сегментации и классификации. YOLOv8 Nano самая быстрая и маленькая, а YOLOv8 Extra Large (YOLOv8x) самая точная, но самая медленная среди них.

Индикаторы оценки

FPS (Frame Per Second, количество кадров в секунду) используется для оценки скорости обнаружения модели, как показано в уравнении (1), где Fn обозначает количество обнаруженных изображений, а T обозначает общее время обнаружения изображений.

$$FPS = \frac{F_n}{T} \tag{1}$$

В исследовании использованы mAP, R и P [**Ошибка! Источник ссылки не найден.**] для количественной о ценки эффективности модели обнаружения объектов. *P* (Precision, точность) количественно определяет долю истинных положительных результатов среди всех положительных предсказаний, оценивая способность модели избегать ложных срабатываний. R (Recall, отзыв) - рассчитывает долю истинных положительных прогнозов среди всех прогнозов, оценивая способность модели обнаруживать все экземпляры класса. Формула средней точности (AP)[48] показана в уравнении (2) и относится к размеру области, заключенной между двумя кривыми, кривой точности P и кривой точности R, на интервале (0,1). mAP расширяет концепцию AP, вычисляя средние значения AP по нескольким классам объектов.

Расчеты для P, R и AP проводились следующим образом:

$$\begin{aligned} P &= \frac{TP}{(TP + FP)} \\ R &= \frac{TP}{(TP + FN)} \\ AP &= \int_0^1 P(R)dR \\ mAP &= \frac{\sum_{i=1}^{k} AP_i}{k} \end{aligned} \tag{2}$$

где TP — является истинно положительным предсказанием, FP — является ложно положительным предсказанием, FN — является ложно отрицательным предсказанием.

IV. Эксперимент

Экспериментальная среда:

Операционная платформа — Windows 10Pro,
процессор — Intel(R) Xeon(R) CPU E5-2683 v4@2.10 GHz (2);
объем памяти — ОЗУ 256ГБ;
графический процессор — INVIDIA Quadro P6000;
фреймворк — PyTorch-2.0.1+cu118, версия CUDA — 10. 0, cuDNN 8.0 (для обучения YOLOv8)
фреймворк — darknet +cu118, версия CUDA — 10. 0, cuDNN 8.0 (для обучения YOLOv4).

Для экспериментов были выбраны наиболее распространенные поверхностные дефекты проводящего рисунка печатных плат (рис. 2) – это микроотверстия (проколы, 1.pin-hole), обрыв проводника (2.open), короткие замыкания (3.short), отклонение толщины проводников (4.mousebite), выступы (5.spur), медные включения (6.copper), искажение проводника (7.lineNick) и дефекты отверстий: отклонение толщины металлизации отверстия (8.nick), отсутствие металлизации отверстия (9.circleVioletion), неправильное расположение отверстия (10.fineShort), посторонний материал внутри отверстия (11.clearance), дефект внутренней металлизации отверстия (12.etching). Англоязычные названия дефектов были заимствованы из системы АОИ Fusion фирмы Orbotech Ltd.(www.orbotech.com), которая в данный момент эксплуатируется на предприятии ООО «Технотех».

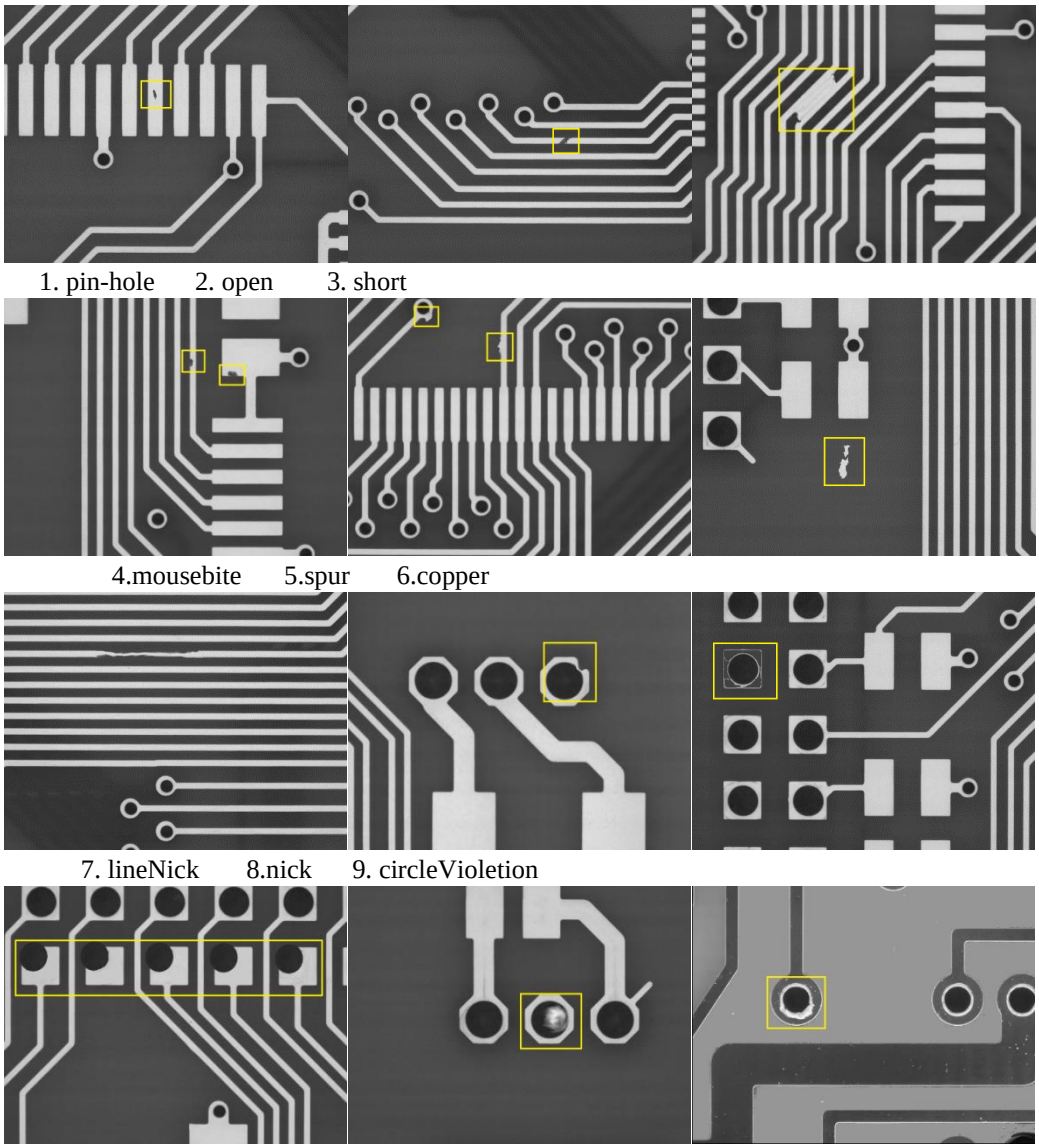


Рис. 2. Типы дефектов проводящего рисунка и отверстий печатных плат

Первый этап исследований. Цель первых исследований состояла в том, чтобы провести проверку, сможет ли выбранная архитектура нейронной сети не потерять и, главное, различить мелкие объекты, которыми являются дефекты печатных плат. Для обучения нейронной сети использовалась упрощённая версия YOLOv4 — YOLOv4-tiny, предобученная на наборе данных Microsoft COCO (Common Objects in Context – 100 тысяч изображений с 80 классами) [49]. Обучение YOLOv4-tiny происходило на собственном наборе данных из 60 цветных изображений размерности 416×416 пикселов с двумя классами дефектов [mousebit, open], функцией активации LeakyRelu, в качестве оптимизатора был использован алгоритм стохастического градиентного спуска.

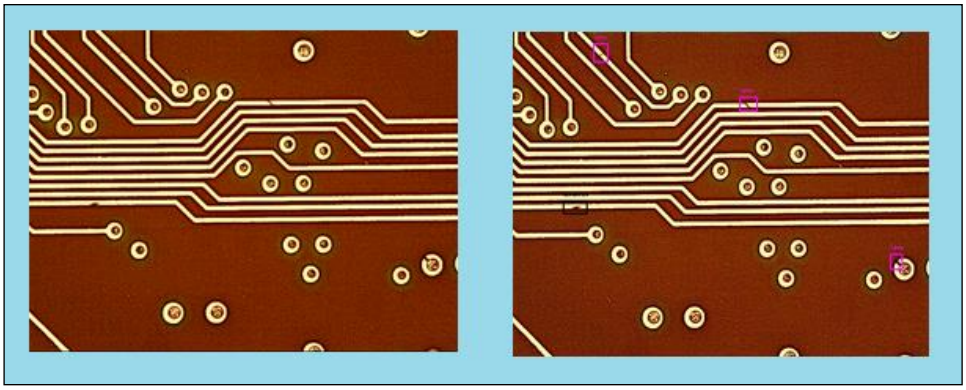


Рис. 3. Распознавание дефектов YOLOv4-tiny, 416x416

Результат на тестовых данных оказался вполне удовлетворительным. Средняя точность распознавания при mAP@0.50 (mean average precisions/средняя точность) составила 95.82%. Сеть смогла различить очень мелкие дефекты со столь незначительными различиями, как те, что показаны на рисунке (рис. 3).

Второй этап исследований. Следующие эксперименты проводились на наборе данных PCBData, который был взят с платформы Kaggle и адаптирован к формату YOLO. Размеченный набор данных PCBData состоит из 1500 чёрно-белых изображений размера 640×640 и содержит шесть классов наиболее распространённых видов дефектов: pin-hole, open, short, mousebite, spur, copper. Набор был разбит на части: 1200 изображений выделено для обучения, 200 изображений составил тестовый набор и 100 изображений оставлено для валидации.

На этом наборе данных тестировались две предобученные версии архитектуры YOLO: YOLOv4-tiny и YOLOv8 Nano. YOLOv8 [**Ошибка! Источник ссылки не найден.**] восьмой версии, которая является одной из п оследней на сегодняшний день версией архитектуры семейства YOLO. YOLOv4-tiny – время обучения составило 4 часа, Средняя точность распознавания при mAP@0.50 - 98.02%. YOLOv8n – время обучения составило 30 минут, 100 эпох. Средняя точность распознавания при mAP@0.50 составила 92,00%. Обе архитектуры YOLOv4-tiny и YOLOv8n смогли детектировать шесть различных дефектов на одном изображении (Таблица 1, Таблица 2).

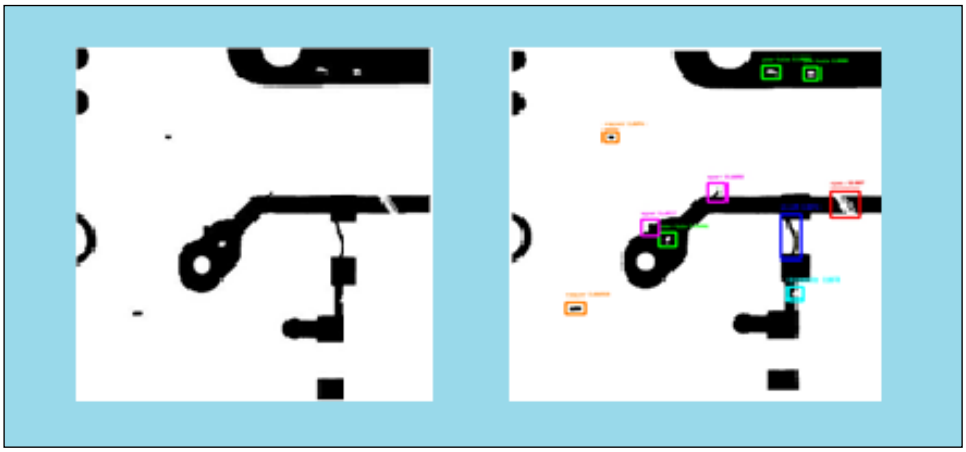


Рис. 4. Результат распознавание шести дефектов YOLOv4-tiny/ YOLOv8n.

ТАБЛИЦА 1 РАЗМЕР МОДЕЛИ, ВРЕМЯ ОБУЧЕНИЯ, СКОРОСТЬ ОБНАРУЖЕНИЯ И РЕЗУЛЬТАТ ОЦЕНКИ mAP@0.50m CNN YOLO НА НАБОРЕ ДАННЫХ PCBData.

Архитектура	**Размер модели (МБ)**	**Время обучения (часы)**	**Скорость обнаружения (кадров в секунду)**	**mAP@0.50 (%)**
Yolov4-tiny	23,070	4	30,2	98.02
Yolov8n	6,107	0.5	32	92.00

Таблица 2. РЕЗУЛЬТАТ ОЦЕНКИ mAP@0.50 ПО ДЕФЕКТАМ.

Архитектура	all	pin-hole	open	short	mose-bit	spur	cooper
YOLOv4-tiny	0.958	0.991	0.929	0.914	0.968	0.981	0.964
YOLOv8n	0.920	0.934	0.952	0.960	0.890	0.830	0.952

Третий этап исследований. Третий этап исследований включил в себя сбор реальных изображений дефектов проводящего рисунка печатных плат на предприятии при участии технологов и создание собственного набора данных. Обучение CNN YOLO на собственном наборе данных. Анализ результатов.

Изображения дефектов печатный плат собираются специально разработанным устройство получения изображений, после чего проводятся аугментация и аннотационная маркировка программой labelImg[50]. Затем версии YOLOv4-tiny, YOLOv4, YOLOv8n, YOLOv8x обучается на основе настроенного набора данных. Для обучения модели были выбраны двенадцать категорий наиболее распространенных дефектов поверхности и отверстий (рис. 2). Затем результаты эксперимента оцениваются и сравниваются.

Создание набора данных. Специально разработанное устройство получения изображений состоит из модуля управления движением, модуля освещения и модуля получения изображений (рис. 5). Отдельные детали конструкции были изготовлены с применением аддитивных технологий, напечатаны на 3D принтере.

Модуль освещения (5) обеспечивает подходящее освещение для модуля получения изображений камеры для фотографирования. Модуль получения изображений (5) отвечает за сбор изображений печатных плат и подключается к компьютеру для хранения и подготовки изображений. Он состоит из опорной рамки камеры, промышленной камеры ToupCam U3CMOS14000KPA, микроскопа и внешнего компьютера. Камера последовательно захватывает изображения печатных плат под стабильным источником света. Максимальное поле зрения составляет 130 мм × 100 мм. Для печатных плат большего размера использовалась многоточечная съемка, при которой камера перемещается компонентами управления движением. Движение камеры зигзагообразное, начинается из нижнего левого угла. В имени полученных изображений сохраняются координаты - номер изображения по вертикали и горизонтали. После определения изображения, содержащего дефект, можно переместить камеру к требуемой области на плате и подсветить лазерной указкой.

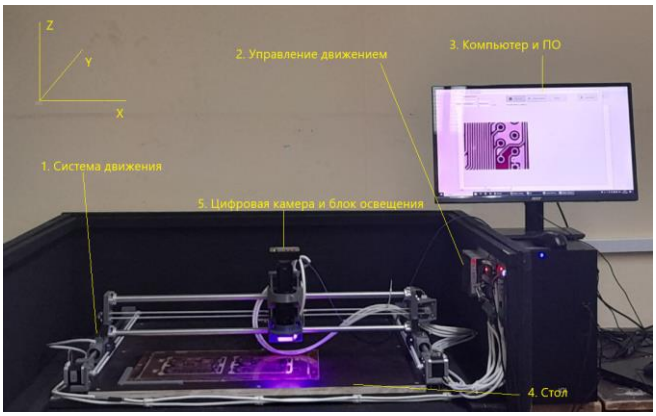
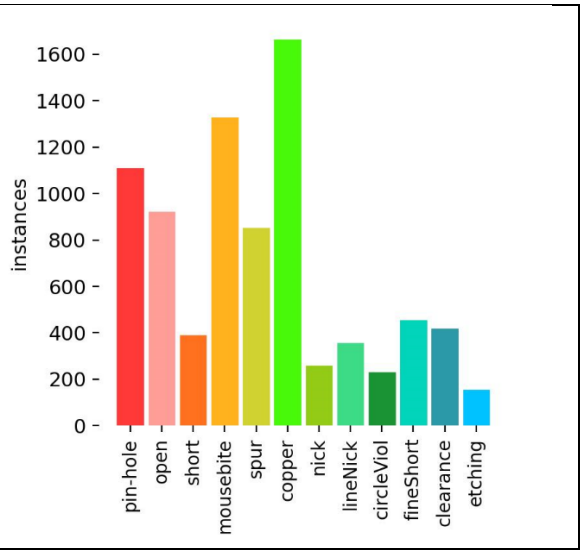


Рис. 5. Устройство получения изображения печатной платы. (1) система движения (ось сканирования X оптической головки, ось Y стола, ось Z оптической системы); (2) электронная система управления движением на базе контроллера Arduino UNO; (3) компьютер и ПО для обработки изображений и выявления дефектов; (4) стол, поддерживающий различные размеры панелей (max размер 610x610); (5) цифровая камера с различными источниками освещения.

На данный момент создан размеченный набор данных, содержащий 2912 цветных изображений и дополнительный вариант этого же набора данных в градациях серого. Размер изображений - 416×416 пикселей. Модели архитектуры YOLO различных версий прошла обучение на этих наборах данных. Тренировочный набор состоит из 2200 изображений, тестовый набор – 512 изображений, набор для валидации – 200 изображений. Распределение дефектов на наборе данных показано в таблице (см. табл.3). Результаты обучения представлены в следующих таблицах (см. табл. 4, 5):

ТАБЛИЦА 3. РАСПРЕДЕЛЕНИЕ КЛАССОВ ДЕФЕКТОВ В НАБОРЕ ДАННЫХ.

Классы дефектов	Количество дефектов
All	11784
pin-hole	1656
open	1344
short	432
mousebite	1864
spur	1185
copper	2392
nick	584
lineNick	568
circleViol	304
fineShort	568
clearance	577
etching	310



Обучение модели. На рисунке 6 (см. рис. 6) представлен процесс обучения YOLOv8n на наборе данных из изображений в градациях серого. На графиках траектории потерь отражают производительность модели обнаружения объектов по всему спектру фаз обучения и проверки. Эти фазы исследуют различные измерения потерь: точность ограничивающего поля (‘box_loss’), классификацию категорий ('cls_loss') и уверенность обнаружения объектов ('dfl_loss'-используется для регрессии ограничивающей рамки). По мере постепенного увеличения числа итераций значение потерь модели постепенно уменьшается. После 50 итераций кривая потерь обучения медленно сходится и стабилизируется, когда число итераций обучения достигает примерно 100.

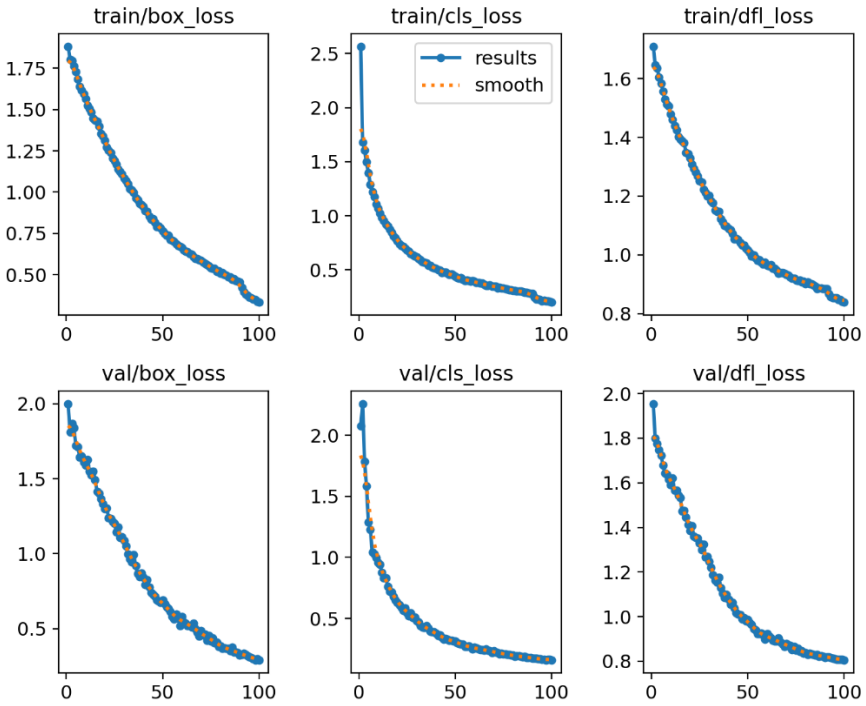


Рис. 6**.** Кривая потерь во время обучения YOLOv8n (100 эпох) на наборе данных из цветных изображений.

Этот анализ (см. рис. 7) охватывает эпохи обучения, описывая производительность моделей по основным метрикам, таким как точность, полнота, средняя средняя точность при пороге пересечения по объединению (IoU) 0,5 (mAP_0.5) и средняя средняя точность, охватывающая спектр порогов IoU от 0,5 до 0,95 (mAP_0.5:0.95). Траектория Precision, показатель достоверности положительных обнаружений, показывает колеблющуюся, но в целом стабильную тенденцию модели YOLOv8n. Обращаясь к Recall, показателю полноты модели в определении соответствующих экземпляров в наборе данных, график указывает на хорошую способность модели захватывать полный набор экземпляров объектов. Метрика mAP@0.5, оценивающая гармонизацию точности и полноты при пороговом значении IoU 0.5, показывает, что YOLOv8n достигает отличной точности обнаружения при этом

дискретном пороговом значении. В строгой оценке mAP@0.5-0.95, охватывающей диапазон пороговых значений IoU от 0,5 до 0,95, модель показывает способность поддерживать эффективность обнаружения в разнообразном диапазоне контрольных показателей IoU.

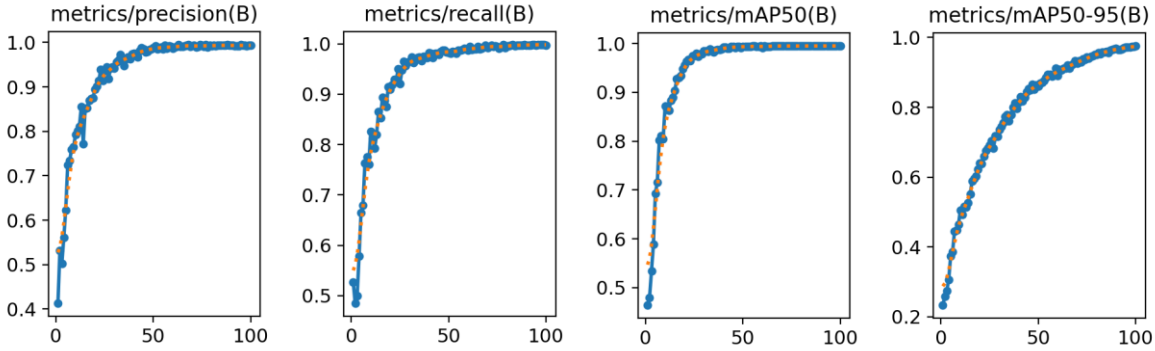


Рис. 7**.** Метрики обучения YOLOv8n (100 эпох, изображения в градациях серого). Точность, полнота, средняя средняя точность mAP@50 при пороге 0,50 пересечения по объединению (IoU) и средняя средняя точность mAP@50-95, охватывающая спектр порогов IoU от 0,5 до 0,95.

Эти показатели служат барометрами надежности модели обнаружения объектов, где точность и комплексное обнаружение имеют первостепенное значение. Описанные тенденции наделяют YOLOv8n потенциалом для эффективной работы со различными, сложными, разномасштабными дефектами проводящего рисунка печатных плат, позиционируя YOLOv8n как перспективного претендента на развертывание в реальных приложениях, требующих высокой точности обнаружения объектов.

На рисунке (рис.8.) представлена визуализация работы модели YOLOv8n на цветных изображениях и изображениях в градациях серого.

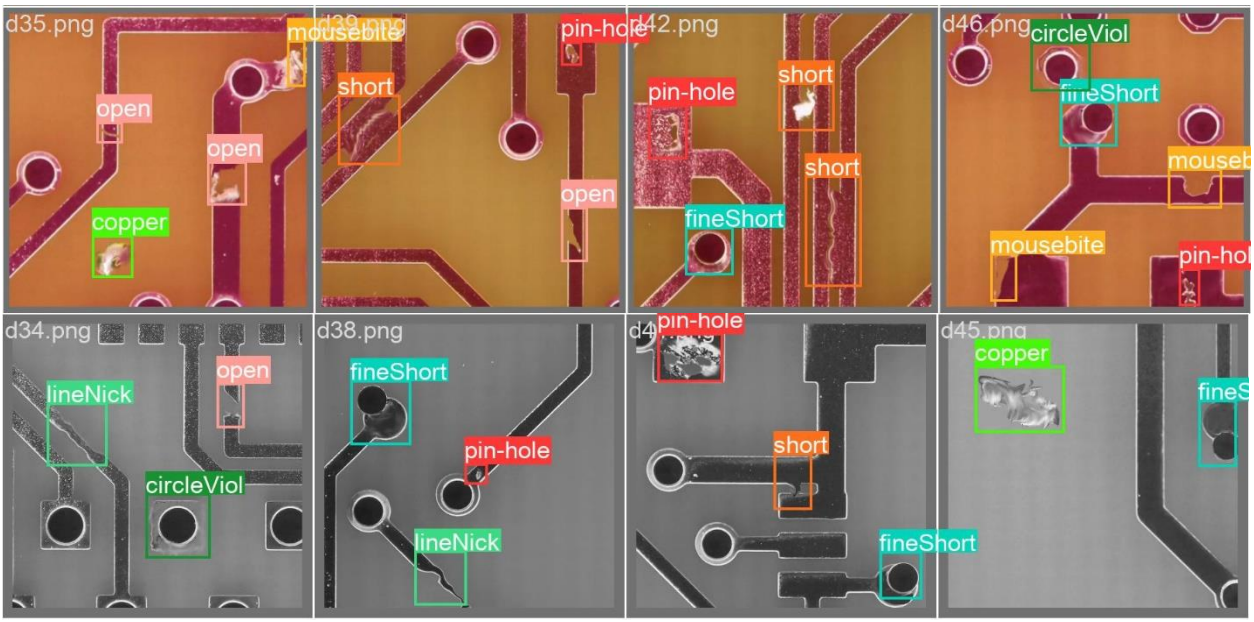


Рис. 8. Визуализация результатов теста. YOLOv8n обученная на цветных изображениях и изображениях в градациях серого.

Сравнение производительности различных алгоритмов обнаружения. Сравнительные экспериментальные результаты различных алгоритмов показаны в таблицах (см. табл. 4, 5)**.** Все исследуемые алгоритмы на цветных изображениях и изображениях в градациях серого имеют близкую среднюю точность mAP@50 обнаружения — более 98%. Yolov4-tiny имеет самую высокую скорость обнаружения 300 кадров в секунду на всех типах изображений. Обучение архитектуры семейства Yolov4 четвёртой версии занимает достаточно длительное время. YOLOv8n — сочетает в себе точность, высокую производительность обнаружения и самое короткое время обучения. Размер модели YOLOv8n является самым оптимальным, составляет 6.08 МБ.

ТАБЛИЦА 4. РАЗМЕР ПАРАМЕТРОВ МОДЕЛИ, ВРЕМЯ ОБУЧЕНИЯ СКОРОСТЬ ОБНАРУЖЕНИЯ И РЕЗУЛЬТАТ ОЦЕНКИ mAP@0.50m CNN YOLO НА НАБОРАХ ДАННЫХ ЦВЕТНЫХ ИЗОБРАЖЕНИЙ И ИЗОБРАЖЕНИЙ В ГРАДАЦИЯХ СЕРОГО.

Архитектура	**Цвет изображений**	**Размер параметров модели (МБ)**	**Время обучения (часы)**	**Скорость обнаружения (кадров в секунду)**	**mAP@0.50 (%)**
Yolov4-tiny	Серый	23.070	5	300	98.02
Yolov4	Серый	250.247	9.3	32	99.44
Yolov4-tiny	Цветные	23.070	6.3	**302**	98.38
Yolov4	Цветные	250.247	7.3	32	**99.86**
Yolov8n	Серый	**6.082**	**0.5**	84	98.90
Yolov8x	Серый	133.509	2.3	14	99.50
Yolov8n	Цветные	6.082	0.5	83	99.00
Yolov8x	Цветные	133.539	6.9	8	99.50

ТАБЛИЦА 5. РЕЗУЛЬТАТ ОЦЕНКИ mAP@0.50 ПО ДЕФЕКТАМ

YOLO	**all**	**pin-hole**	**open**	**short**	**mose-bit**	**spur**	**cooper**	**nick**	**line-nick**	**circle-viol**	**fine Sort**	**clea-rance**	**etching**
Y_v4-tiny_G	0.981	0.967	0.986	0.954	0.992	0.968	0.959	0.985	0.990	0.990	0.990	0.990	0.950
Y_v4_Gray	0.991	0.989	0.994	0.998	0.997	0.994	0.998	0.992	0.988	0.996	0.992	0.983	0.952
Y_v4-tiny_C	0.983	0.985	0.971	0.990	0.990	0.986	0.960	0.967	0.990	0.990	0.990	0.994	0.936
Y_v4_C	0.998	0.990	0.989	0.990	0.999	0.990	0.994	0.990	0.990	0.990	0.990	0.990	0.985
Y_8n_Gray	0.989	0.988	0.989	0.995	0.994	0.989	0.979	0.995	0.995	0.995	0.995	0.995	0.958
Y_8x_ Gray	0.994	0.995	0.995	0.995	0.995	0.995	0.994	0.995	0.995	0.995	0.995	0.995	0.990
Y_8n_Color	0.990	0.992	0.995	0.995	0.994	0.994	0.977	0.993	0.995	0.995	0.995	0.995	0.956
Y_8x_Color	**0.995**	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995

где

Y_v4-tiny_G – YOLOv4- tiny обучалась на изображениях в градациях серого;
YOLOv4_Gray – YOLOv4 обучалась на изображениях в градациях серого;
Y_v4-tiny_C – YOLOv4- tiny обучалась на цветных изображениях;
YOLOv4_C – YOLOv4 обучалась на цветных изображениях;
Y_8n_Gray – YOLOv8n обучалась на изображениях в градациях серого;
Y_8x_Gray – YOLOv8x обучалась на изображениях в градациях серого;
Y_8n_Color – YOLOv8n обучалась на цветных изображениях;
Y_8x_ Color – YOLOv8x обучалась на цветных изображениях;

V.Обсуждение результатов

Полученные результаты показали, что все выбранные для эксперимента модели архитектуры YOLO отлично справляются с задачей обнаружения разномасштабных дефектов проводящего рисунка печатных плат. Для выбора конкретной модели критическими показателями являются точность и скорость обнаружения объектов. Наиболее сбалансированные параметры имеет модель Yolov8n.

Следует отметить, что обучение проводилось на цветных изображениях и изображениях в градациях серого. Это было вызвано проблемами получения изображений требуемого качества пригодных для бинаризации. Полученные изображения печатных плат цветной цифровой камерой сильно отличались друг от друга по цвету. Выяснилось, что отражающие/поглощающие свойства текстолита и медного рисунка напрямую зависят от тонкостей технологического процесса при производстве печатной платы. На данный момент детектирование дефектов проводится на изображениях в градациях серого, как наиболее универсальных.

При получении изображений были выявлены трудности, связанные с освещением металлизированной поверхности. Отражение от медных поверхностей относительно больших площадей было неравномерным, несмотря на то, что использовалось рассеянное освещение. Это создавало трудности при получении изображений требуемого качества для бинаризации. В текущий момент проводятся дополнительные эксперименты с использованием поляризованного освещения.

Опытно-конструкторский образец автоматизированной системы оптического инспекции прошёл предварительные испытания на предприятии ООО «ТЕХНОТЕХ». В ходе исследовательской работы было получено порядка 6000 изображений. По результатам испытаний оптической системы и анализу полученных результатов было сделано заключение о необходимости доработок в устройстве получения изображений. Возникла необходимость увеличить скорость получения изображений. Для этого планируется заменить цветную 2D камеру на линейную монохромную камеру.

Классификация дефектов требует уточнения, например, дефекты, помеченные как “nick” часто распознаются как «mousebite». Некоторые виды дефектов невозможно правильно идентифицировать только по топологии проводящего рисунка с использованием свёрточных нейронных сетей. Представляет определённые затруднения определение таких дефектов как вариативность расположения отверстий, неоднозначность определение дефекта «замыкания», и возможных и сложных их комбинаций, что приводит к ложным определениям дефектов на плате. Примеры таких дефектов приведены на рисунке (рис. 9). В таких случаях требуется дополнительная информация из gerber-файла чертежа.

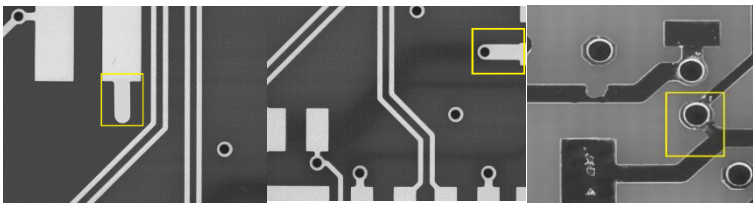


Рисунок 9. Примеры изображений на которых невозможно определить дефект.

VI. Выводы и заключения

Полученные в ходе экспериментально-опытной работы результаты показали, что модели YOLOv4-tiny и YOLOv8 можно эффективно использовать для быстрого обнаружения мелких объектов, какими являются дефекты токопроводящего рисунка печатной платы. CNN архитектуры YOLO справляется с задачей быстрого и точного нахождения областей на изображении печатной платы имеющих дефект проводящего рисунка. Архитектура YOLOv8n — сочетает в себе точность, высокую производительность обнаружения и самое короткое время обучения. Размер модели YOLOv8n является самым оптимальным, составляет 6.08 МБ.

Для уменьшения количества ложных срабатываний при детектировании дефектов необходимо использовать информацию gerber-файла чертежа.

В планах дальнейшей работы намечена детальная разработка классификации дефектов под руководством технологов предприятия и дальнейшее формирование набора данных для обучения сети. По оценкам для обучения нейронной сети требуется не менее 700 изображений на каждый потенциально выявляемый дефект.

На следующих этапах работ планируется дальнейшее проведение сравнения результатов точности обнаружения дефектов с импортными промышленными аналогами АОИ, которые используются в данный момент на предприятии ООО «ТЕХНОТЕХ» г. Йошкар-Ола.

Благодарности

Работа выполняется в рамках государственного задания на оказание государственных услуг (выполнение работ) №075-01252-22-03 от 26.10.2022.

Список литературы

1. Moganti M., Ercal F., Dagli C. H., Tsunekawa S. Automatic PCB Inspection Algorithms: A Survey. // Computer Vision and Image Understanding. – 1996. – Vol. 63, No. 3. – P. 287–313. –DOI:10.1006/cviu.1996.0020
2. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet Classification with Deep Convolutional Neural Networks. // Communications of the ACM. – 2017. – Vol. 60, No.6. – P. 84–90. – DOI: 10.1145/3065386
3. Girshick, R., Donahue, J., Darrell, T., Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2015.– Vol. 38, No.1. – P.142–158. – DOI:10.1109/tpami.2015.2437384.
4. Girshick R. Fast R-CNN.// IEEE International Conference on Computer Vision. – 2015. – DOI: 10.1109/ICCV.2015.169.
5. Ren S., He K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2016. – Vol. 39, No.6. – P.1137–1149. – DOI:10.1109/tpami.2016.2577031.

6. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C. SSD: Single Shot MultiBox Detector. // Lecture Notes in Computer Science. – 2016. – P.21–37. – DOI:10.1007/978-3-319-46448-0_2.
7. Redmon, J., Divvala, S., Girshick, R.,Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. // IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – DOI:10.1109/cvpr.2016.91.
8. Liao, X.; Lv, S.; Li, D.; Luo, Y.; Zhu, Z.; Jiang, C. Yolov4-mn3 for pcb surface defect detection. // Applied Sciences. – 2021. – Vol.11, No. 24. – P. 11701. – DOI: 10.3390/app112411701.
9. Bochkovskiy A., Chien-Yao Wang, Hong-Yuan Mark Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection. // IEEE Conference on Computer Vision and Pattern Recognition.– 2020. – DOI:10.48550/arXiv.2004.10934.
10. Howard, A.; Sandler, M.; Chen, B. Searching for MobileNetV3. //In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea – 2019. – P. 1314–1324. –DOI:10.48550/arXiv.1905.02244
11. Bowei Du Fang Wan, Guangbo Lei, Li Xu,Chengzhi Xu and Ying Xiong, YOLO-MBBi: PCB Surface Defect Detection Method Based on Enhanced YOLOv5. // Electronics. –2023. – Vol.12, No. 13. – P 2821. – DOI:10.3390/electronics12132821.
12. Ultralytics Yolov5. Available at: https://github.com/ultralytics/yoloV5. – 2020. Accessed 20.07.2024
13. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. // In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA. – 2019. – P. 6105–6114. – DOI:10.48550/arXiv.1905.11946.
14. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional block attention module. // IEEE Conference on Computer Vision and Pattern Recognition. –2018. – Vol.10. – P. 973–978. –DOI:10.48550/arXiv.1807.06521.
15. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. // IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA. – 2017. – DOI:10.48550/arXiv.1612.03144.
16. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. // IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA. – 2020. – P. 10781–10790. – DOI:10.48550/arXiv.1911.09070.
17. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. // IEEE Conference on Computer Vision and Pattern Recognition. – 2017.–DOI:10.48550/arXiv.1704.04861.
18. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. // In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA. – 2020. – Vol. 34. – P. 12993–13000. – DOI:10.48550/arXiv.1911.08287.
19. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. – 2022. – DOI:10.48550/arXiv.2205.12740.
20. Ding, R.; Dai, L.; Li, G.; Liu, H. TDD-net: A tiny defect detection network for printed circuit boards. //CAAI Transactions on Intelligence Technology. – 2019. –Vol. 4. – P.110–116. – DOI:10.1049/trit.2019.0019.
21. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. //IEEE Conference on Computer Vision and Pattern Recognition. – 2022. – Vol.116. – DOI:10.48550/arXiv.2207.02696.
22. DeepPCB https://github.com/Charmve/Surface-Defect-Detection/tree/master/DeepPCB (дата обращения: 3 августа 2024 г.).
23. Jiang, Y.; Cai, M.; Zhang, D. Lightweight Network DCR-YOLO for Surface Defect Detection on Printed Circuit Boards. // Sensors.– 2023. – Vol.17. – P. 7310. – DOI:10.3390/s23177310.
24. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. // IEEE Conference on Computer Vision and Pattern Recognition. – 2016 . – DOI:10.48550/arXiv.1605.07146
25. He, K., Zhang, X., Ren, S., & Sun, J. Deep Residual Learning for Image Recognition.//IEEE Conference on Computer Vision and Pattern Recognition. - 2016. - DOI:10.1109/cvpr.2016.90.
26. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. // IEEE Conference on Computer Vision and Pattern Recognition. – 2018. – DOI:10.48550/arXiv.1804.02767.
27. Rapid Detection of PCB Defects Based on YOLOx-Plus and FPGA. //IEEE Access. – 2024. –Vol. 1. –No.1. – P. 99. – DOI:10.1109/ACCESS.2024.3387947.
28. Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun YOLOX: Exceeding YOLO Series in 2021. // IEEE Conference on Computer Vision and Pattern Recognition. – 2021. – DOI:10.48550/arXiv.2107.08430.
29. Mittal S., A Survey of FPGA-based Accelerators for Convolutional Neural Networks. // Neural Computing and ApplicationsArchive. – 2020. – Vol. 32, No. 06. – P. 1109-1139. – DOI:10.1007/s00521-018-3761-1.
30. Yang, Lingxiao and Zhang, Ru-Yuan and Li, Lida and Xie, Xiaohua, SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. // Proceedings of the 38th International Conference on

Machine Learning. – 2021. –Vol.139. – P. 11863—11874. URL:http://proceedings.mlr.press/v139/yang21o.html (дата обращения 2.09.2024).
31. Liu S., Qi L., Qin H., Shi J., Jia J. Path aggregation network for instance segmentation // IEEE Conference on Computer Vision and Pattern Recognition. –2018. – P. 8759–8768. – DOI:10.48550/arXiv.1803.01534
32. CSPHB Zheng Ge1,Songtao Liu, Zeming Li, Osamu Yoshie, Jian Sun, OTA: Optimal Transport Assignment for Object Detection. // IEEE Conference on Computer Vision and Pattern Recognition. – 2021. – DOI:10.48550/arXiv.2103.14259
33. Подходы к разработке системы обнаружения дефектов печатных плат на основе технологии АОИ / Т. С. Ходатаева, Н. В. Каширин, А. И. Аверина, А. Е. Гурьянов // Труды Института системного программирования РАН. – 2023. – Т. 35, № 4. – С. 109-120. – DOI 10.15514/ISPRAS-2023-35(4)-5.
34. G. Jocher, A. Chaurasia, and J. Qiu, YOLO by Ultralytics. URL:https://github.com/ultralytics/ultralytics. (дата обращения 2.09.2024).
35. Deng J., Dong W., Socher R., Li K. Li, Li F.-F. ImageNet: A large-scale hierarchical image database. // IEEE Conference on Computer Vision and Pattern Recognition. – 2009. – P.248-255. – DOI: 10.1109/CVPR.2009.5206848.
36. C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN. // IEEE Conference on Computer Vision and Pattern Recognition. – 2019. – P. 390–391. – DOI:10.48550/arXiv.1911.11929.
37. D. Misra, Mish: A self regularized non-monotonic neural activation function. // IEEE Conference on Computer Vision and Pattern Recognition. – 2019. – Vol. 4, No. 2. – P. 10–48550. – DOI:10.48550/arXiv.1908.08681.
38. K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition. // IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – Vol. 37, No. 9– P. 1904–1916. – DOI:10.48550/arXiv.1406.4729.
39. G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Dropblock: A regularization method for convolutional networks," Advances in neural information processing systems. – 2018. – Vol. 31. – DOI:10.48550/arXiv.1810.12890.
40. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. // The journal of machine learning research. – 2014. – Vol. 15, No. 1. – P. 1929–1958.
41. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision. // IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – P. 2818–2826. – DOI:10.48550/arXiv.1512.00567.
42. M. A. Islam, S. Naha, M. Rochan, N. Bruce, and Y. Wang, Label refinement network for coarse-to-fine semantic segmentation.//IEEE Conference on Computer Vision and Pattern Recognition. – 2017. – DOI:10.48550/arXiv.1703.0055.
43. S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. // Conference on Machine Learning. – 2015. – P. 448–456. – DOI:10.48550/arXiv.1502.03167.
44. Архитектура YOLOv8, визуализация сделана пользователем GitHub RangeKing https://github.com/open-mmlab/mmyolo/blob/main/configs/yolov8/README.md (дата обращения 2.09.2024)
45. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. // IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – DOI:10.48550/arXiv.1512.03385.
46. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Netw. // Conference on Machine Learning. – 2018. – Vol.107. – P. 3–11. – DOI:10.48550/arXiv.1702.03118.
47. X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. // IEEE Conference on Computer Vision and Pattern Recognition. – 2020. – Vol. 33. – P.21002–21012.– DOI:10.48550/arXiv.2006.04388.
48. J. Davis, M. Goadrich The Relationship Between Precision-Recall and ROC Curves. // Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA. – 2006. – P.233–240. – DOI:10.1145/1143844.1143874
49. Lin T.-Y., Maire M., Belongie S., et al. Microsoft coco: common objects incontext. // IEEE Conference on Computer Vision and Pattern Recognition. Springer. – 2014. – P. 740–755. – DOI:10.48550/arXiv.1405.0312.
50. Tzutalin. LabelImg. Git code (2015). https://github.com/tzutalin/labelImg (дата обращения 2.09.2024)

Секция 3

КОМПЬЮТЕРНОЕ ЗРЕНИЕ

Computer Vision

GraphiCon 2024
17 − 19 сентября 2024
Омск, Россия

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-467-475

Координатная привязка автономного подводного аппарата к объектам подводного добычного комплекса по стереоизображениям
Coordinate referencing of an autonomous underwater vehicle to objects of a subsea production system using stereo images

В.А. Бобков1, А.П. Кудряшов 1
1Институт автоматики и процессов управления ДВО РАН, ул.Радио,5, Владивосток, 690041, РФ

V.A. Bobkov1, A.P. Kudryashov 1
[1]Institute of Automation and Control Processes Far Eastern Branch of Russian Academy of Sciences,690041, Vladivostok, Russia

Аннотация. В последнее время возросла актуальность проблемы инспекции подводной инфраструктуры (газовые и нефтяные трубопроводы, подводные добычные комплексы и др.) с использованием автономных необитаемых подводных аппаратов (АНПА), как альтернативы традиционно применяемой технологии удаленного управления подводным аппаратом. Точная локализация АНПА относительно объектов является ключевой задачей инспекции. Однако трудности, характерные для подводной среды, не позволяют обеспечить высокоточную навигацию, основываясь только на сенсорах штатной навигационной системы (гидроакустика, инерционно измерительный блок и др.). Альтернативой/дополнением к упомянутым средствам навигации являются технологии, основанные на применении компьютерного зрения, которые потенциально могут обеспечить более высокую точность навигации АНПА в сравнении с обработкой акустических измерений в условиях локальных сцен при небольших расстояниях до объектов. Поэтому в настоящей работе, в контексте решения проблемы инспекции объектов подводного добычного комплекса, предлагается новый подход к решению задачи распознавания подводных объектов и координатной привязки к ним АНПА, основанный на обработке стереоизображений. Отличительными особенностями подхода являются применение нестандартной автоматизированной методики формирования геометрической модели объекта по изображениям обзорной траектории, использование при распознавании объектов разнотипных характерных геометрических элементов (точки, отрезки, углы) и применение оригинальных алгоритмов сопоставления получаемых визуальных данных с априорной моделью объекта.

***Ключевые слова*:** Автономный необитаемый подводный аппарат(АНПА), подводный добывающий комплекс (ПДК), компьютерное зрение, стерео, визуальная навигация.

Аннотация. Recently, the urgency of the problem of inspection of underwater infrastructure (gas and oil pipelines, subsea production systems (SPS), etc.) using autonomous underwater vehicles (AUVs) has increased as an alternative to the traditionally used technology of remote control of an underwater vehicle. Accurate localization of AUVs relative to objects is a key inspection task. However, the difficulties characteristic of the underwater environment do not allow for high-precision navigation based only on the sensors of the standard navigation system (hydroacoustics, inertial measuring unit, etc.). An alternative/addition to the mentioned navigation tools are technologies based on the use of computer vision, which can potentially provide higher accuracy of AUV navigation in comparison with the processing of acoustic measurements in local scenes at short distances to objects. Therefore, in this work, in the context of solving the problem of inspecting objects of an SPSs, we propose a new approach to solving the problem of recognizing underwater objects and the coordinate reference of AUVs to them, based on stereo image processing. Distinctive features of the approach are the use of a non-standard automated technique for generating a geometric model of an object from images of a survey trajectory, the use of different types of characteristic geometric elements (points, segments, angles) when recognizing objects, and the use of original algorithms for comparing the resulting visual data with an a priori model of the object.

***Ключевые слова*:** Autonomous underwater vehicle (AUV), subsea production systems (SPS), computer vision, stereo, visual navigation.

I. Введение

Одной из практически важных задач применения АНПА является инспекция объектов подводной промышленной инфраструктуры, включая подводные трубопроводы, коммуникации, подводные добычные комплексы и др. Точная локализация аппарата, по отношению к конкретному объекту комплекса является ключевой задачей инспекции.

Однако трудности в обеспечении точной навигации АНПА в подводной среде (невозможность использования GPS, отсутствие карт местности, недостаточная точность акустического зондирования) обусловливают более активное использование АНПА при выполнении задач обследования. Поэтому для определения точности навигации применяются различные средства и технологии в зависимости от задач миссии АНПА, характера сцен и внешних предварительных данных. Часто применяют интеграцию штатных средств с дополнительными сенсорами. Например, в статье [1] представлен проект «CView», который решает проблему инспекции порта. Для обнаружения препятствий и досмотровых целей используется сканирующий гидролокатор. Изображения гидролокатора автоматически обрабатываются с помощью алгоритмов обнаружения границ и выделения линий для получения упрощенного описания окружающей среды, которое используется методами наведения, представленными в этой статье. В статье [2] представлены современные тенденции исследований в области АНПА. Обсуждается применение различных сенсорных технологий, включая и системы стереовидения для локализации, навигации и картографирования.

Альтернативой/дополнением к вышеупомянутым средствам навигации являются технологии, основанные на обработке оптических изображений, получаемых моно- и стереокамерами. Эти методы позволяют обеспечить более высокую точность навигации АНПА по сравнению с обработкой акустических измерений в условиях локальных/компактных сцен на небольших расстояниях до объектов.

В [3] описан метод локализации АНПА с использованием визуальных измерений подводных сооружений и искусственных ориентиров. В ряде работ, например, в [4, 5] рассматриваются методы отслеживания искомой траектории с использованием визуальных измерений характеристик точек и адаптивного управления, в том числе нейронных сетей. В [6] применяется визуальный метод для приведения АНПА к док-станции. В [7] предложен метод монокулярной визуальной одометрии с оптическим отслеживанием потока, который, по мнению авторов, более пригоден для подводной съемки, чем классические подходы, основанные на дескрипторах.

Задача определения положения аппарата относительно инспектируемых объектов основывается на решении задачи распознавания объекта. Поскольку речь идет о подводных объектах промышленной инфраструктуры, целесообразно учитывать геометрическую специфику искусственных объектов (явное присутствие в форме объекта стандартных геометрических элементов).

В статье [8] рассматривается проблема распознавания объектов по бесцветным трехмерным облакам точек в подводной среде. В нем представлено сравнение производительности современных методов определения подводных объектов (глобальных дескрипторов). Изученные методы предназначены для оказания помощи АНПА в выполнении автономных вмешательств при подводных проверках, техническом обслуживании и ремонте. Предполагается, что 3D-модель объекта из трехмерных точек доступна априори.

В [9] авторы разрабатывают свою предыдущую работу, включающую метод распознавания трехмерных объектов для облаков точек с использованием точечных признаков. Метод предназначен для таких сценариев применения, как инспекция промышленных подводных сооружений, состоящих из труб и соединительных объектов (таких как клапаны, колена и соединители). Алгоритм распознавания использует доступную априори базу данных частичных видов объектов, хранящихся в виде облаков точек. При разработке технологий инспекции может учитываться предварительная информация об объектах - цифровые карты местности [10] или наличие координатных измерений характерных точечных особенностей объектов во внешней системе координат (СК) [11].

В настоящем докладе предлагается новый подход к решению задачи распознавания подводных объектов и координатной привязки к ним АНПА в контексте проблемы инспекции объектов подводной промышленной инфраструктуры. Подход основан на применении нестандартной автоматизированной методики формирования геометрической модели объекта по изображениям обзорной траектории, использовании при распознавании объектов разнотипных характерных геометрических элементов и применении оригинальных алгоритмов сопоставления получаемых визуальных данных с априорной моделью объекта.

II. Постановка задачи

Проблема регулярной инспекции различных подводных инженерных сооружений, включая подводный добывающий комплекс (ПДК), требует точной координатной привязки АНПА к подводным искусственным объектам при выполнении инспекции (модель ПДК показана на рис.1). Привязка АНПА к СК объекта необходима, как для планирования траектории АНПА, так и для выполнения операций инспекционной миссии (фотографирование критичных элементов конструкций, возможная посадка и ремонт) непосредственно в координатном пространстве инспектируемых объектов. Предполагается, что АНПА, наряду со штатным навигационным оборудованием, оснащен оптической стереокамерой, которая смотрит строго вниз по отношению к продольной оси АНПА. Задача расчета точной координатной привязки АНПА к собственной СК подводного искусственного объекта в контексте этой проблемы решается с использованием видеоинформации.
Структурно под ПДК понимается множество функционально объединенных объектов, распределенных на ограниченной площади морского дна, и координированных в СК ПДК. При движении АНПА по запланированной траектории в районе размещения ПДК осуществляется регулярная проверка возможности идентификации объектов ПДК на основе обработки фиксируемого камерой видеопотока стереоизображений. Распознавание инспектируемого объекта выполняется сравнением извлекаемой из стереоизображений геометрической информации с априорно заданной 3D моделью объекта. Модель строится с участием оператора с применением автоматизированной методики. В отличие от традиционной методики, когда модель формируется на основе использования натурных измерений или документальной информации, описывающей объект, в предлагаемой методике используются изображения нескольких видов (ракурсов) предварительно выполненной обзорной траектории АНПА и полученная на их основе 3D информация. Обзорная траектория выполняется таким образом, чтобы камера могла зафиксировать изображения всех объектов (или частей объекта, если объект в силу своих больших размеров не может полностью попасть в поле зрения камеры) распределенного ПДК. Тогда под моделью ПДК будем понимать объединение моделей всех входящих в ПДК объектов. В качестве геометрических элементов модели рассматриваются характерные 3D точечные особенности объекта, прямолинейные отрезки и макроэлементы, составленные из отрезков, характеризующие форму объекта. Согласно методике, оператор формирует для каждого вида характерные элементы (ХЭ), исходя из визуального анализа используемых снимков обзорной траектории и вычисляемых 3D данных. ХЭ координируются, как в СК, связанной с конкретным видом, так и в собственной СК объекта, которую строит оператор. Для каждого использованного вида обзорной траектории вычисляется матрица преобразования координат, связывающая СК этого вида с собственной СК объекта. Матрица формируется известным способом - заданием ортов СК объекта в СК вида. В качестве СК ПДК фиксируется одна из СК объектов. Соответственно, для каждого объекта формируется матрица координатных преобразований, связывающая СК объекта с СК ПДК.

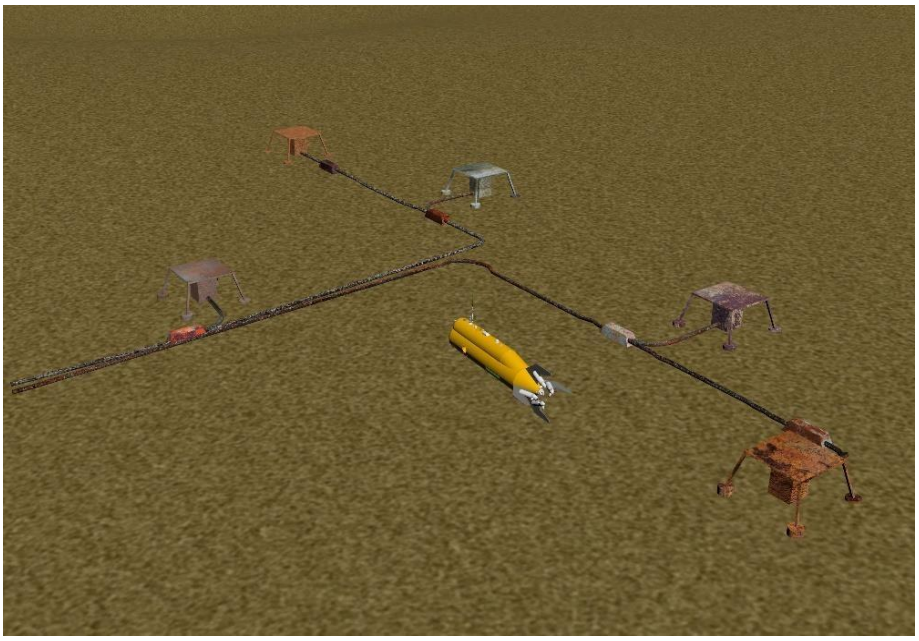


Рис. 1. Виртуальная сцена: подводный добычный комплекс

Использование нескольких видов объекта и ХЭ разных типов повышает достоверность идентификации конкретных объектов и точность навигации АНПА по отношению к подводному объекту.

Идентификация объекта и координатная привязка АНПА к собственной СК объекта в общем случае реализуется последовательной трехэтапной вычислительной схемой. На первом этапе выполняется непосредственное распознавание 3D точечных особенностей объекта по изображениям стереопары, полученной камерой АНПА в анализируемой позиции рабочей (инспекционной) траектории. Это делается с помощью алгоритма сопоставления множества точечных особенностей на изображениях стереопары рабочей траектории с множеством точечных особенностей на изображениях стереопары модели (для каждого используемого вида обзорной траектории) (см. описание алгоритма ниже). Результатом работы алгоритма сопоставления являются два сопоставленных множества 3D точек, одно из которых определено в СК модели, а другое - в СК камеры в позиции рабочей траектории АНПА. По этим двум сопоставленным 3D множествам вычисляется матрица преобразования координат между СК АНПА и СК модели объекта. Полученная матрица преобразования координат может рассматриваться как приемлемое решение проблемы привязки АНПА к объекту, или использоваться в качестве начального приближения при получении более точного решения на втором и третьем этапе с учетом распознанных ХЭ всех типов.

На втором этапе выполняется распознавание прямолинейных отрезков и макроэлементов с использованием полученной матрицы преобразования координат на первом этапе. Отметим, что в этом подходе не используются векторизованные формы исходных изображений (требующие значительных вычислительных затрат), но при сравнении образов отрезка в модели и на изображениях рабочей стереопары используются корреляционные оценки. Распознавание дополнительных к точкам элементов повышает надежность идентификации объектов и точность координатной привязки к ним.

На третьем этапе вычисляется уточненная матрица координатного преобразования из СК позиции рабочей траектории АНПА в СК объекта с учетом точечных данных по всем видам и типам ХЭ.

В дальнейшем изложении множества ХЭ и матрицы координатных преобразований будут именоваться с использованием англоязычной аббревиатуры. ПДК именуется как SPS (subsea production system).

M_2D_POINT_V^kV^k_LR и S_2D_ POINT_V^jV^j_RL – сопоставленные множества точек на левом и правом изображениях «модельной» стереопары k-вида (напомним, что модель рассматривается как совокупность нескольких используемых видов обзорной траектории АНПА). Здесь символ «M» подразумевает, что множество относится к модели, а символ «S» будет использоваться в идентификаторах множеств, относящихся к изображениям «рабочей» стереопары. V^k обозначает, что используется вид «k» обзорной траектории. Мнемоника LR подразумевает, что множество на левом изображении стереопары сопоставляется с множеством на правом изображении стереопары. Аналогичным образом используется мнемоника RL. Используемое далее обозначение V^kV^k означает, что сопоставление относится к изображениям стереопары k-вида. Аналогичным образом мнемоника обозначения WW подразумевает, что сопоставляются множества, относящиеся изображениям «рабочей» стереопары. Обозначение V^kW подразумевает, что множество на изображении рабочей стереопары сопоставляется с множеством модели вида k (вид k обзорной траектории). И, соответственно, обозначение WV^k подразумевает, что множество модели вида k сопоставляется с множеством на изображении рабочей стереопары. Идентификатор объекта - object_n. Множества, относящиеся к модели объекта, формируются с помощью автоматизированной методики с участием оператора. CS^{view_k} – координатная система камеры АНПА для k-вида обзорной траектории. M_3D_POINT_V^kV^k_CS^{view_k} – множество 3D точек в системе координат CS^{view_k} (принадлежит 3D модели k-вида обзорной траектории), полученное по сопоставленным множествам. LINE - множество 3D «отрезков», CORNER - множество 3D «углов». $H_{CSwork,CSview\_k}$ – матрица преобразования координат из СК позиции рабочей траектории в СК к-вида обзорной траектории.

III. Распознавание объектов ПДК

Распознавание объектов ПДК с привязкой АНПА к СК объекта необходимо для планирования и реализации траектории АНПА непосредственно в координатном пространстве инспектируемого объекта. Задача координатной привязки АНПА к объекту решается сравнением получаемых в позиции инспекционной траектории АНПА визуальных данных с моделью объекта. Реализуется трехэтапная вычислительная схема с использованием нескольких видов (ракурсов) объекта, представляющих модель объекта (рис.2). Выбор видов для сравнения определяется критерием близости анализируемой позиции рабочей траектории АНПА и позиции обзорной траектории выбранного вида.

На первом этапе для каждого выбранного вида k выполняется распознавание искомого объекта на основе сопоставления точечных особенностей на рабочей (инспекционной) и модельной стереопарах. Если распознавание успешное (используется пороговый критерий), то результатом является: а) множество точек S_3D_POINT_V^k_CS^{work}, сопоставленное с множеством M_3D_POINT_V^k_CS^{view_k} (принадлежащим модели вида k); и б) вычисленная матрица $H_{CSwork,CSview\_k}$ координатного преобразования из СК позиции инспекционной

траектории в СК модели используемого вида k объекта. А поскольку для каждого вида имеется матрица $H_{CSview\_k\ ,CSobject\ \_\ n}$ преобразования координат из СК этого вида в СК объекта n (полученная при формировании модели вида k), то искомое преобразование из CS^{work} в CS^{object_n} можно получить последовательным выполнением двух указанных преобразований. Каждое полученное преобразование (число которых равно числу используемых видов) может рассматриваться как приемлемое решение поставленной задачи. Полученная матрица может также использоваться в качестве начального приближения в методе оптимизации при вычислении уточненной матрицы координатного преобразования из CSwork в CSobject_n с учетом точечных данных по всем видам и всем типам ХЭ (на третьем этапе). Полученное множество распознанных «точек» $S\_3D\_POINT\_V^kW\_CS^{work}$ заносится в накапливаемое по всем видам множество $S\_3D\_POINT\_CS^{object\_n}$, координированное в $СК^{object\_n}$ (применяется имеющаяся матрица $H_{CSview\_k\ ,CSobject\ \_\ n}$).

На втором этапе на рабочей стереопаре распознаются «отрезки» и «углы» (с использованием полученного преобразования $CSwork\ ,CSview\_k\ H$), сопоставленные элементам модели вида k.

Алгоритм распознавания и вычисления 3D координат ХЭ типа «отрезки». Осуществляется последовательная выборка отрезков из 3D множества $M\_3D\_LINE\_V^kV^k\_CS^{view\_k}$, и для каждого отрезка выполняется распознавание его 2D образа на изображениях стереопары анализируемой позиции рабочей (инспекционной) траектории. А именно, для каждого отрезка li (заданного концевыми точками p1i и p2i):

1. Извлекаются указатели на 2D образы его двух точек на изображениях стереопары модели.

2. Отрезок помещается в координатное пространство CS^{work} с помощью матрицы $(H_{CSwork,CSview\ k})^{-1}$ (вычислена на этапе 1), и проецируется на изображения стереопары рабочей траектории. Под выражением «отрезок проецируется» понимается проецирование концевых точек отрезка.

3. Для каждой концевой точки вычисляется коэффициент корреляции сравниваемых образов точки на изображении стереопары модели и изображении рабочей стереопары. При сравнении концевых точек отрезков используется ориентация аппарата по компасу как модельной, так и рабочей траектории, для обеспечения одинаковой ориентации сравниваемых окрестностей.

4. В случае неполного соответствия образов выполняется попытка найти соответствие в небольшом радиусе, в предположении, что проецирование (шаг 2) выполнено с некоторой погрешностью (шаг 3).

5. Принимается решение о соответствии или о несоответствии образов. В конечном счете, предлагаемый способ сопоставления без использования векторизованной формы изображений существенным образом снижает вычислительные затраты.

Распознавание и вычисление 3D координат ХЭ типа «углы». Выделение ХЭ «углы» осуществляется на полученном множестве «отрезков». Алгоритм выделения достаточно простой, поскольку ХЭ «угол» конструируется из двух отрезков, лежащих в одной плоскости в пространстве. За счет указания принадлежности отрезка конкретному «углу», исключается неоднозначность в случаях одинаковой геометрии углов.

Результатом являются множества $S\_3D\_LINE\_V^kW\_CS^{work}$ и $S\_3D\_CORNER\_V^kW\_CS^{work}$. Эти множества заносятся в соответствующие накапливаемые по всем видам множества $S\_3D\_LINE\_CS^{object\_n}$, $S\_CORNER\_CS^{object\_n}$, координированные в CS^{object_n} (применяется имеющаяся матрица $H_{CSview\_k\ ,CSobject\ \_\ n}$). Таким образом, эти множества вместе с множеством точек $S\_3D\_POINT\_СК^{object\_n}$ являются полным объединением распознанных 3D ХЭ объекта всех типов со всех видов в собственной СК объекта. Их визуализация может дать наглядное представление о том, насколько полно и достоверно был распознан / идентифицирован анализируемый объект.

На третьем этапе вычисляется уточненная матрица $H_{CSwork\ ,CSobject\ \_\ n}$. Уточнение (повышение точности координации АНПА в координатном пространстве объекта) реализуется за счет учета распознанных «точек» на всех используемых видах, а также учета точек, принадлежащих распознанным «отрезкам» и «углам».

IV. Вычисление координатной привязки АНПА к объекту ПДК

Рассматриваются два способа вычисления координатной привязки АНПА к объекту ПДК:

1.Привязка осуществляется, как было сказано выше, с помощью матрицы координатного преобразования из CS^{work} в $СК^{object\_n}$, которая вычисляется с использованием модели вида k – как последовательное выполнение двух имеющихся преобразований:

$$H_{CSwork\ ,CSobject\ \_\ n}\ =\ H_{CSview\_k\ ,CSobject\ \_\ n}\ \cdot\ H_{CSwork\ ,CSview\_k} \tag{1}$$

В этом случае мы получаем k матриц (по числу используемых видов объекта). Для каждого варианта оценивается погрешность вычисления точности на тестовом наборе сопоставленных точек.

2. Матрица преобразования координат $H_{CSwork\ ,CSobject\ \_\ n}$ из CS^{work} в CS^{object_n} вычисляется на объединенном

множестве сопоставленных точек (по всем видам объекта). Применяется стандартный метод минимизации целевой функции, построенной на оценке суммарного расхождения распознанных точек множества $S\_3D\_POINT\_V^kW\_CS^{work}$ (на изображениях рабочей стереопары) и сопоставленных им точек множества $M\_3D\_POINT\_V^k\_CS^{view\_k}$ (модель вида k объекта). Суммарное расхождение подсчитывается в координатном пространстве CS^{object_n}. Сравниваются два варианта получения объединенного множества точек: а) учитываются только ХЭ типа «точки» на всех используемых видах; б) к ХЭ типа «точки» в каждом виде добавляются еще точки, принадлежащие распознанным ХЭ типа «отрезки» и «углы».

Для компактной записи целевой функции введем промежуточные обозначения. Обозначим множество $M\_3D\_POINT\_V^k\_CS^{view\_k}$ как $MP^k\left(mp_1^k,\ldots,mp_{np\_k}^k\right)$, а сопоставленное ему множество $S\_3D\_POINT\_V^kW\_CS^{work}$ обозначим как $SP^k\left(sp_1^k,\ldots,sp_{np\_k}^k\right)$. Тогда задача минимизации целевой функции в варианте а формулируется как:

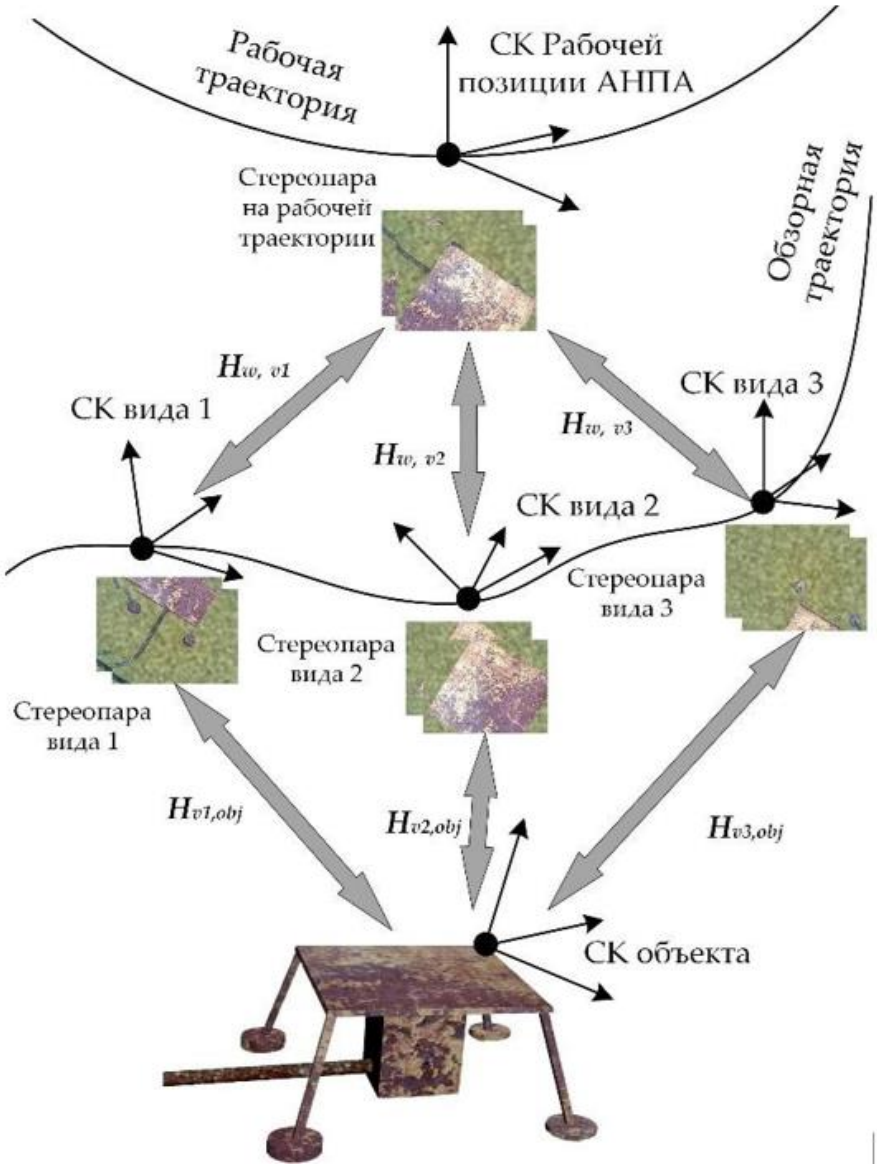


Рис. 2. Распознавание объекта ПДК по стереоизображениям с использованием нескольких видов объекта обзорной траектории (показаны 3 вида)

$$\min \sum_1^k \sum_{i=1}^{np\_k} R1_i^k \text{, здесь } R1_i^k = \left\| mp_i^k \cdot H_{CSview\_k,CSobject\_n} - sp_i^k \cdot H_{CSwork\_k,CSobject\_n} \right\| \quad (2)$$

здесь используется матрица $H_{CSview\_k,CSobject\_n}$, построенная на этапе формирования модели вида k.

В варианте **б** дополнительно учитываются, как было выше сказано, точки, принадлежащие ХЭ-ам «отрезки» и «углы» (эти точки определяют координатное положение ХЭ). Аналогичным образом упростим запись, введя промежуточные обозначения. Обозначим множество M_3D_LINE_V^k_CS^{view_k} точек, принадлежащих ХЭ-ам типа «отрезки» модели вида k, как ML^k $(ml_1^k,\ldots,\ ml_{nl\_k}^k)$, а множество точек сопоставленного ему множества распознанных отрезков S_3D_LINE_V^{k}W_CS^{work} обозначим как $SL^k(sl_1^k, \ldots, sl_{ml\_k}^k)$. Здесь элементы и обозначают начальную и конечную точку отрезка i.

Тогда учет точек «отрезка» i в целевую функцию можно записать как:

$$R2_i^k = \left\| ml_i^k \cdot H_{CSview\_k,CSobject\_n} - sl_i^k \cdot H_{CSwork\_k,CSobject\_n} \right\| \quad (3)$$

Аналогичным образом введем упрощающие обозначения для точек, принадлежащих ХЭ типа «углы». Обозначим множество M_3D_CORNER_V^k_CS^{view_k} точек, принадлежащих ХЭ-ам типа «углы» модели вида k, как $MC^k(mc_1^k, \ldots, mc_{nc\_k}^k)$, а множество точек сопоставленного ему множества распознанных углов S_3D_CORNER_V^{k}W_CS^{work} обозначим как $SC^k(sc_1^k, \ldots, sc_{nl\_k}^k)$..

Соответственно, учет точек «угла» i в целевую функцию можно записать как:

$$R3_i^k = \left\| mc_i^k \cdot H_{CSview\_k,CSobject\_n} - sc_i^k \cdot H_{CSwork\_k,CSobject\_n} \right\| \quad (4)$$

Тогда задача минимизации целевой функции в варианте **б** формулируется как:

$$\min \sum_1^k \left(\sum_{i=1}^{np\_k} R1_i^k + \sum_{i=1}^{nl\_k} R2_i^k + \sum_{i=1}^{nc\_k} R3_i^k \right) \quad (5)$$

V. Результаты экспериментов

Для оценки эффективности предложенного метода были выполнены эксперименты с виртуальной сценой и с реальными данными.

Модельная сцена. Для получения сравнительной оценки точности координатной привязки АНПА к объекту по одному виду и по нескольким видам объекта была проведена серия экспериментов на виртуальной сцене, которая была сгенерирована средствами моделирующего комплекса [12]. Высота АНПА над грунтом изменялась от 6.4 м до 7.1 м. Высота до крышки скважины 3.6 метра, частота съемки 10 кадров/сек, разрешение снимков 1200 x 900. Расчет координат АНПА осуществлялся в позициях траектории через каждые 10 кадров фотосъемки. Координатная привязка осуществлялась согласно нашему методу с помощью вычисляемой матрицы преобразования координат $H_{CSwork,CSobject\_n}$. При использовании одного вида обзорной траектории (вид 1 - число сопоставленных точек 26) точность координатной привязки AUV к объекту составила 32 мм. При добавлении точек видов 2, 3, и удаления дубликатов, число сопоставленных точек увеличилось до 76, а точность, соответственно, повысилась до 2.8 мм. Дополнительно были проведены еще три таких же эксперимента, но с другими тройками видов (имеющими меньшую общую область видимости объекта камерой АНПА в рассматриваемой позиции инспекционной траектории). Были получены точности, ожидаемым образом, немного ниже – 3,4 мм, 4,2 мм, 12,3 мм – но, все равно они выше, чем при использовании одного вида.

Таким образом, использование нескольких видов обзорной траектории для вычисления матрицы $H_{CSwork,CSobject\_n}$ повысило точность координатной привязки АНПА к объекту в этих экспериментах в 3-11 раз.

Точность привязки была подтверждена также высоким значением вычисленного коэффициента корреляции (до 94%) концевых точек при перепроецировании отрезков.

Реальные данные. Эксперимент проводился в условиях, максимально приближенных к реальным – использовался сооруженный макет объекта, помещенный на морское дно на глубине 1.5 м; съемка осуществлялась камерой Karmin2 (Nerian's 3D Stereo Camera, 25 cm), помещенной в герметичный бокс с её подключением к ноутбуку на сопровождающей лодке; непосредственное фотографирование (согласно вышеописанной методике получения разных видов) осуществлял оператор-дайвер. Стереокадр приведен на рис. 3.

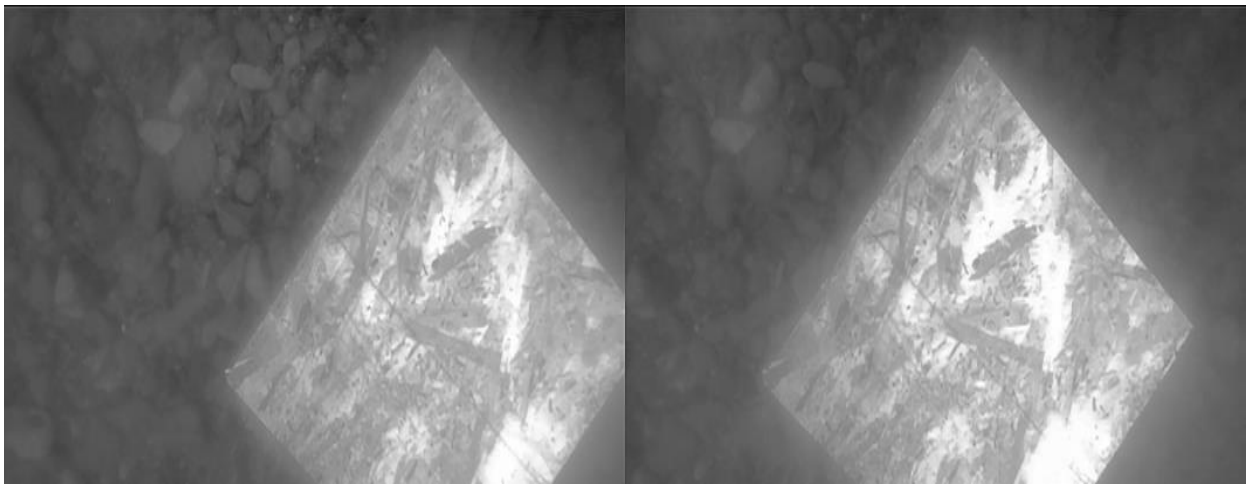


Рис. 3. Стереокадр макета объекта в реальных условиях на морском дне.

Оценка точности метода в этом эксперименте основывалась на сравнении вычисленного тестируемым методом расстояния между двумя заданными 3D точками с «истинным значением», измеренным на макете объекта в его СК. В качестве «истинного значения» рассматривалось значение измеренного расстояния непосредственно на макете до его помещения на морское дно.

При использовании одного вида средняя погрешность составила 18,1 мм. При использовании 3-х видов погрешность снизилась до 7,6 мм. Таким образом, подтвердился вывод, полученный на виртуальной сцене, о повышении точности координатной привязки АНПА к подводному объекту за счет использования нескольких видов обзорной траектории.

VI. Выводы и заключение

В докладе представлен новый подход к обеспечению точной координатной привязки АНПА к объектам ПДК, использующий только обработку стереоизображений. Его отличием от аналогов является использование для формирования моделей объектов ПДК предварительно выполненной обзорной траектории АНПА, разнотипных геометрических элементов, повышающих степень распознавания объекта, а также применение оригинальных алгоритмов сопоставления получаемых визуальных данных с априорной моделью объекта.

Источник финансирования. Благодарности

Работа выполнена при финансовой поддержке за счет гранта Российского научного фонда № 22-11-00032, https://rscf.ru/project/22-11-00032/.

Научный руководитель: Бобков В.А., Институт автоматики и процессов управления ДВО РАН.

Список литературы

1. Jacobi, M. Autonomous inspection of underwater structures. Robot. Auton. Syst. 2015, 67, 80–86, https://doi.org/10.1016/j.robot. 2014.10.006.
2. Avilash Sahoo; Santosha K. Dwivedy; P.S. Robi. Advancements in the field of autonomous underwater vehicle. Ocean Engineering, Volume 181, 2019, Pages 145-160.
3. Jung, J.; Li, J.-H.; Choi, H.-T.; Myung, H. Localization of AUVs using visual information of underwater structures and artificial landmarks. Intell. Serv. Robot. 2016, 10, 67–76, https://doi.org/10.1007/s11370-016-0210-9.
4. Gao, J.; Wu, P.; Yang, B.; Xia, F. Adaptive neural network control for visual servoing of underwater vehicles with pose estimation. J. Mar. Sci. Technol. 2016, 22, 470–478, https://doi.org/10.1007/s00773-016-0426-6.
5. Xu, H.; Oliveira, P.; Soares, C.G. L1 adaptive backstepping control for path-following of underactuated marine surface ships. Eur. J. Control. 2020, 58, 357–372, https://doi.org/10.1016/j.ejcon.2020.08.003.
6. Fan, S.; Liu, C.; Li, B.; Xu, Y.; Xu, W. AUV docking based on USBL navigation and vision guidance. J. Mar. Sci. Technol. 2018, 24, 673–685, https://doi.org/10.1007/s00773-018-0577-8.
7. Ferrera, M.; Moras, J.; Trouvé-Peloux, P.; Creuze, V. Real-Time Monocular Visual Odometry for Turbid and Dynamic Underwater Environments. Sensors 2019, 19, 687, https://doi.org/10.3390/s19030687.
8. Himri, K.; Ridao, P.; Gracias, N. 3D Object Recognition Based on Point Clouds in Underwater Environment with Global Descriptors: A Survey. Sensors 2019, 19, 4451, https://doi.org/10.3390/s19204451.
9. Himri, K.; Ridao, P.; Gracias, N. Underwater Object Recognition Using Point-Features,

Bayesian Estimation and Semantic Information. Sensors 2021, 21, 1807. https://doi.org/10.3390/s21051807.

10. Bertrand Chemisky, Fabio Menna, Erica Nocerino, Pierre Drap. Underwater Survey for Oil and Gas Industry: A Review of Close Range Optical Methods. Remote Sensing, MDPI, 2021, 13 (14), pp.2789. 10.3390/rs13142789. hal-03361951.
11. Bobkov, V.; Kudryashov, A.; Inzartsev, A. Method for the Coordination of Referencing of Autonomous Underwater Vehicles to Man-Made Objects Using Stereo Images. J. Mar. Sci. Eng. 2021, 9, 1038. DOI:10.3390/jmse9091038.
12. V.A. Bobkov, A.P. Kudryashov, S.V. Mel'man, A.F. Shcherbatyuk. Autonomous Underwater Navigation with 3D Environment Modeling Using Stereo Images // Gyroscopy and Navigation, 2018, Vol. 9, No. 1, pp. 67–75.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-476-483

Гибридные алгоритмы обнаружения объектов в системах компьютерного зрения реального времени
Hybrid algorithms for object detection in real-time computer vision systems

С.О Власов[1], А.А. Богуславский[1]
[1] Институт Прикладной Математики им.М.В. Келдыша РАН, 125047, Москва, Россия

S.O. Vlasov[1], *A.A. Boguslavsky*[1]
[1] *Keldysh Institute of Applied Methematics, Moscow, Russia*

Аннотация. В работе рассматривается задача распознавания объекта снимках с использованием комбинированного (гибридного) алгоритма, объединяющего классический метод компьютерного зрения выделения контуров и нейросетевые технологии. Дается классификация гибридных алгоритмов, предлагаются возможные структуры исследуемого алгоритма. Представлено описание этапов работы реализованного алгоритма, исследований на наборе тестовых изображений известного объекта. Рассматривается вопрос эффективности перевода исходного изображения к контурному представлению с последующим поиском объекта интереса сверточной нейронной сетью.

***Ключевые слова*:** обнаружение объектов на изображениях, системы компьютерного зрения, гибридные алгоритмы

Abstract. The paper examines the problem of object recognition in photographs using a combined (hybrid) algorithm that combines the classical computer vision method of contour extraction and neural network technologies. A classification of hybrid algorithms is given, and possible structures of the algorithm under study are proposed. A description of the stages of operation of the implemented algorithm and research on a set of test images of a known object is presented. The issue of the efficiency of converting the original image to a contour representation with subsequent search for an object of interest by a convolutional neural network is considered.

Keywords: object detection in images, computer vision systems, hybrid algorithms.

I. Введение

На сегодняшний день основной целью использования систем компьютерного зрения является дополнение или замена человека в различных областях науки и техники, что является причиной разнообразия задач распознавания объекта интереса на изображении: от поиска лица на изображении, полученном обычной фотокамерой, до классификации антропогенных объектов на спутниковых снимках.

Повышение требований к точности и надежности методов компьютерного зрения, ограниченность вычислительных и временных ресурсов, увеличение числа решаемых задач приводит к необходимости разработки новых методов и алгоритмов распознавания объектов на изображениях. Особенно сильно требование к быстродействию выражается в системах реального времени, в которых необходима обработка порядка десятков кадров в секунду на универсальном компьютере.

Как правило, для задач распознавания объектов интереса предлагается три направления повышения скорости и точности обработки изображений:

- повышение точности работы «традиционных» алгоритмов компьютерного зрения;
- модификация методов машинного обучения, в частности, архитектур сверточных нейронных сетей;
- создание алгоритмов и методов, включающих в себя комбинацию различных подходов к решению поставленной задачи.

Можно сформировать следующую классификацию методов распознавания объектов на изображениях:

1) классические алгоритмы распознавания (подходы, не основанные на глубоком обучении, такие как поиск по шаблону, метод Виолы-Джонса, гистограмма направленных градиентов и др.);

2) алгоритмы поиска объектов интереса на основе машинного обучения (распознавание объектов на изображении с использованием сверточных нейронных сетей и глубокого обучения);

3) гибридные алгоритмы (методы, основанные на последовательном или параллельном использовании алгоритмов компьютерного зрения, ранее применявшихся по отдельности):

- алгоритмы, представляющие из себя комбинацию классических методов компьютерного зрения;
- алгоритмы, представляющие из себя комбинацию методов машинного обучения или их

дополнение;

- алгоритмы, представляющие из себя комбинацию традиционного компьютерного зрения и методов машинного обучения.

Целью комбинированного подхода к построению алгоритмов распознавания являются компенсация недостатков применения конкретных методов по отдельности и повышение эффективности системы при их совместном использовании.

Для решения данной задачи уже ведутся исследования в различных направлениях. В [1] решается задача повышения точности работы с видеопотоком, в котором присутствуют большие смещения объекта интереса, за счет нового алгоритма сопоставления дескрипторов, «глубокого сопоставления», что должно повысить точности работы в системах реального времени. Для обнаружения объектов малых размеров, с чем могут не справиться современные методы обнаружения, разрабатываются многоэтапные алгоритмы [2]. Одним из направлений повышения точности при сохранении времени работы существующих архитектур нейронных сетей без их комбинированного использования либо разработки новых моделей является в том числе применение дополненных конвейеров обучения нейронных сетей [3] либо работа с обучающими данными [4]. В работах [5, 6, 7] решается задача увеличения точности локализации объектов. В [8] предлагается оценивать вероятность полученного взаимного расположения обнаруженных на изображении объектов и убирать те, возникновение которых маловероятно.

В данной работе рассматривается решение задачи распознавания объекта интереса на изображении гибридным алгоритмом, представляющим из себя комбинацию классического компьютерного зрения и методов машинного обучения. Данное направление уже рассматривалось в других исследованиях при решении задач поиска объектов дорожной инфраструктуры [9], распознавания пешеходов [10], распознавании объектов камерой БПЛА [11].

В данной работе предлагаются структуры алгоритмов, основанных на обработке сверточными нейронными сетями изображений, предварительно обработанных алгоритмами поиска контуров.

II. Постановка задачи

Для повышения качества работы систем реального времени предлагается использование гибридных алгоритмов обнаружения и классификации объектов на изображениях. Гибридный алгоритм включает в себя элементы как нейросетевых, так и «традиционных» методов обнаружения и распознавания объектов.

На вход поступает изображение, содержащее в себе не менее одного известного объекта. Формально решаемая в данной работе задача состоит из следующих шагов:

- Формирование наборов изображений для обучения и тестирования сверточных нейронных сетей (используемый в работе набор изображений собран из открытых источников и размечен авторами самостоятельно с использованием программы LabelImg [12]);
- Обработка набора изображений алгоритмами обнаружения контуров, подбор настраиваемых параметров;
- Обучение сверточных нейронных сетей различных архитектур на необработанных изображениях;
- Обучение сверточных нейронных сетей различных архитектур на предварительно обработанных алгоритмами поиска контуров изображениях;
- Распознавание объектов на тестовых обработанных и необработанных изображениях, которые нейронная сеть видит впервые;
- Сравнение полученных результатов.

Подробное описание работы сверточных нейронных сетей, алгоритмов поиска контуров и процесса определения предпочтительных настраиваемых параметров в данной работе не приводятся.

III. Теория

Сверточные искусственные нейронные сети при правильном обучении модели позволяют выполнять распознавание различных объектов на сложном фоне. В то же время данный метод обнаружения и классификации может приводить к возникновению ложных обнаружений или пропускать объект интереса на изображении. Архитектуры нейронных сетей, позволяющие достичь высокой точности, требуют больших вычислительных мощностей технической системы. В то же время модели, позволяющие выполнять решение поставленной задачи с малой затратой ресурсов и за меньшее время, как правило, обладают меньшей точностью работы и могут обнаружить не все объекты на обрабатываемом изображении.

Для решения рассматриваемой задачи применялись модели обнаружения объектов (предварительно обученные классификаторы с определенной архитектурой нейронных сетей) из "зоопарка моделей" для

библиотеки машинного обучения TensorFlow. Этот репозитарий программных решений на основе библиотеки TensorFlow представлен в открытом доступе [13, 14].

В ходе работы были рассмотрены следующие архитектуры сверточных искусственных нейронных сетей:

- Efficientdet_d0;
- Centernet_resnet50_v1_fpn;
- SSD_mobilenet_v1, SSD_mobilenet_v2;
- Faster_rcnn_inception_v2.

Однако, как характерный пример различий получаемых результатов, представлены в данной работе результаты использования двух архитектур: Faster_rcnn_inception_v2 и SSD_mobilenet_v1.

В качестве примера в данной работе использовался объект типа «мост», образ которого содержит прямолинейные признаки, что отличает его от большого количества объектов, описываемых точечными признаками, но при этом характерно для объектов искусственного происхождения с простой геометрической формой образа.

На рис. 1 представлен результат обработки одного и того же входного изображения двумя сверточными сетями различной архитектуры. Обучение проводилось на одинаковых наборах данных и оборудовании до начала переобучения моделей. Двухэтапная нейронная сеть архитектуры Faster_rcnn_inception_v2, сочетающая в себе сеть предложений регионов (RPN) для генерации потенциальных предложений по ограничивающим рамкам объектов и сверточную нейронную сеть для классификации и уточнения этих предложений, показывает более высокую точность, чем одноэтапная нейронная сеть SSD_mobilenet, использующая одну нейронную сеть для одновременного прогнозирования классов объектов и смещений ограничивающих рамок для нескольких ограничивающих рамок по умолчанию различных размеров и соотношений сторон. Однако применение нейронной сети второго типа архитектуры позволило получить результат за приблизительно в восемь меньший промежуток времени так как она предназначена для обработки данных на устройствах с малой вычислительной мощностью. Стоит отметить, что в настоящее время ведутся работы по разработке систем обучения одноступенчатых моделей с подходом имитации двухэтапных функций, целью чего является преодоление разрыва в точности между данными типами моделей [15].

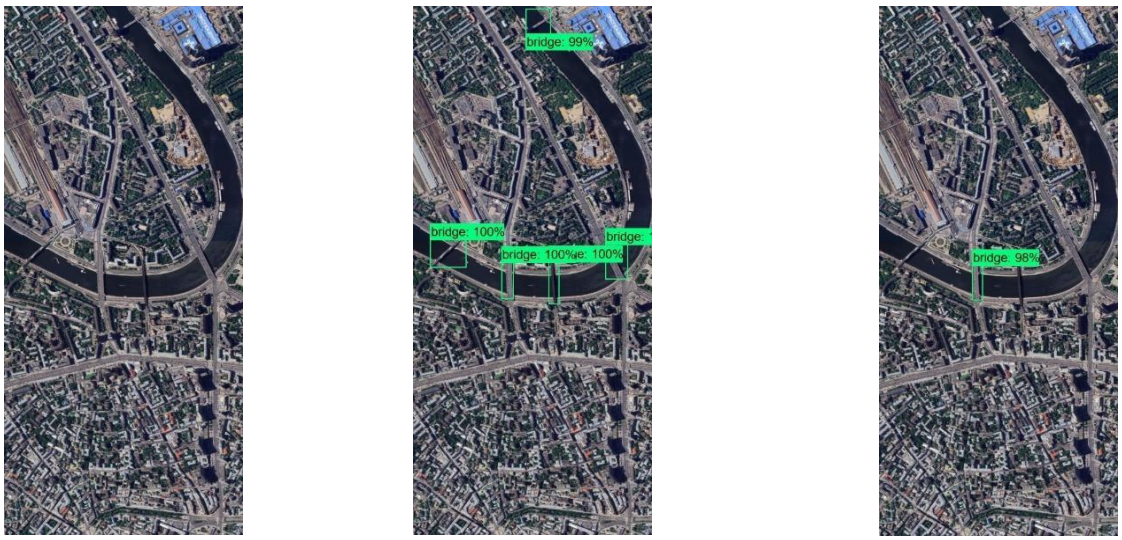


Рис. 1. Результаты обработки исходного изображения (слева) нейронными сетями Faster_rcnn_inception_v2 (по центру) и SSD_mobilenet (справа)

В рамках проведенных исследований предполагается, что для определенного набора объектов интереса предварительная обработка изображения алгоритмом поиска контуров может увеличить вероятность их дальнейшего успешного распознавания. Как показано на рис. 2, объект интереса типа «мост» четко выделается на фоне при переходе к контурному представлению изображения. Помимо существующих алгоритмов распознавания прямых линий на изображении, ведутся исследования по разработке алгоритмов, позволяющих обнаруживать прямые линии контура объектов в реальном времени [16].

Рассматриваемые алгоритмы выделения контуров объектов на изображениях:

- алгоритм Хафа;
- алгоритм Хюккеля.

Работа оператора Хюккеля нахождения контуров на изображении подробно описана в работах [17, 18, 19]. Задачей данного оператора является обнаружение перепадов яркости на изображении и определение параметров, аппроксимирующих их прямых. Оператор Хюккеля разбивает исходное изображение на малые области, в каждой из которых решается локальная задача нахождения границы перехода фона в объект. При помощи специальных функций, определенных в работах [17, 18], выполняется поиск наличия края или линии в каждой обрабатываемой области. При обнаружении признаков перепада яркости достаточной величины и уровня зашумленности, эмпирический элемент контура представляется в виде аппроксимирующего «идеального» элемента контура.

Алгоритм Хафа предназначен для поиска прямых линий на изображении, предварительно обработанном детектором границ Кэнни. Преобразование Хафа поиска прямых заключается в формировании аккумуляторной плоскости, каждая точка которой указывает на вероятность наличия на изображении прямой-края, проходящей через определенную точку исходного изображения. Подробное описание алгоритмов детектора границ Кэнни и поиска прямых Хафа, а также примеры их использования рассматриваются в монографиях [20, 21].

Было определено, что алгоритм Хюккеля применим для задач, требующих получения общего описания контуров объекта интереса на изображении большого разрешения за малый промежуток времени в условиях, когда не требуется большая точность их описания. Также искомый контур должен обладать четкими линейными признаками, размер которых сопоставим с размером области обработки алгоритма.

При обработке изображений малого разрешения или обнаружения объекта сложной формы, а также при необходимости точного описания контура объекта предпочтительно совместное использование алгоритмов Кэнни и Хафа.

Настраиваемые параметры обработки для обоих алгоритмов получают исходя из зашумленности и перепадов яркости, присутствующих на изображении.

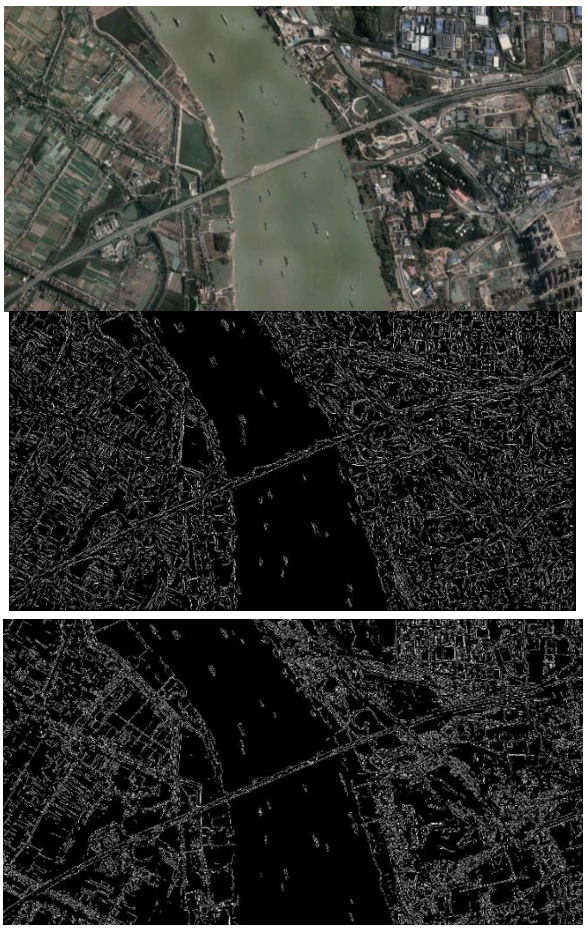


Рис. 2. Результаты обработки исходного изображения (сверху) алгоритмами поиска контуров Хюккеля (по центру) и Хафа (снизу)

На основе раздельного применения сверточных нейронных сетей и алгоритмов поиска контуров объектов предложены три алгоритма распознавания объектов на изображениях, представленные на рис. 3, целью которых

является уменьшение времени обработки изображений и требований к вычислительной мощности при условии сохранения высокой точности распознавания.

Рис. 3. Предложенные структуры гибридных алгоритмов

Для реализации алгоритмов, связанных с предварительным определением областей возможного расположения объекта, предполагается также использование такого метода классического компьютерного зрения, как поиск по правилам. При поиске объектов, содержащих выраженные признаки простой геометрической формы, возможно их выделение при переходе к контурному представлению изображения, как показано на рис. 4.

Рис. 4. Обработанное изображение (слева) и исходное изображение с выделенной областью дальнейшего поиска объекта интереса (справа)

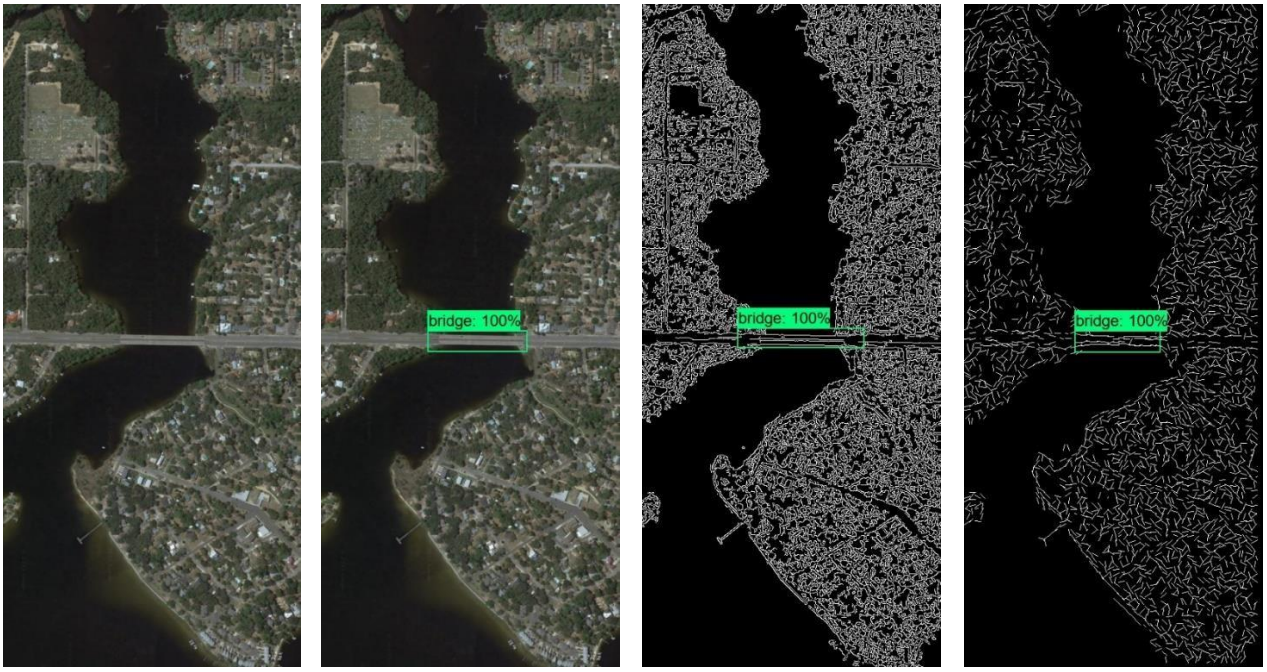
В рамках данной работы рассматривается первый тип гибридного алгоритма: оценивается эффективность предварительной обработки исходного изображения алгоритмом выделения прямолинейных контуров с последующим поиском объекта интереса на всем изображении.

IV. Результаты экспериментов

Для оценки влияния предварительного обнаружения контуров объектов на изображениях была проведена серия расчетов для двух типов искусственных сверточных нейронных сетей: на первом этапе выполнялась разметка данных и классификация объектов на изображениях без дополнительной обработки, на втором этапе процесс повторялся для заранее обработанных изображений алгоритмами Хафа и Хюккеля.

Поиск объектов сверточной искусственной нейронной сетью проводился на всем полученном изображении.

В ходе решения тестовых задач на отдельных изображениях определено, что переход к контурному изображению позволяет выполнять успешное распознавание объектов, как показано на рис. 5.

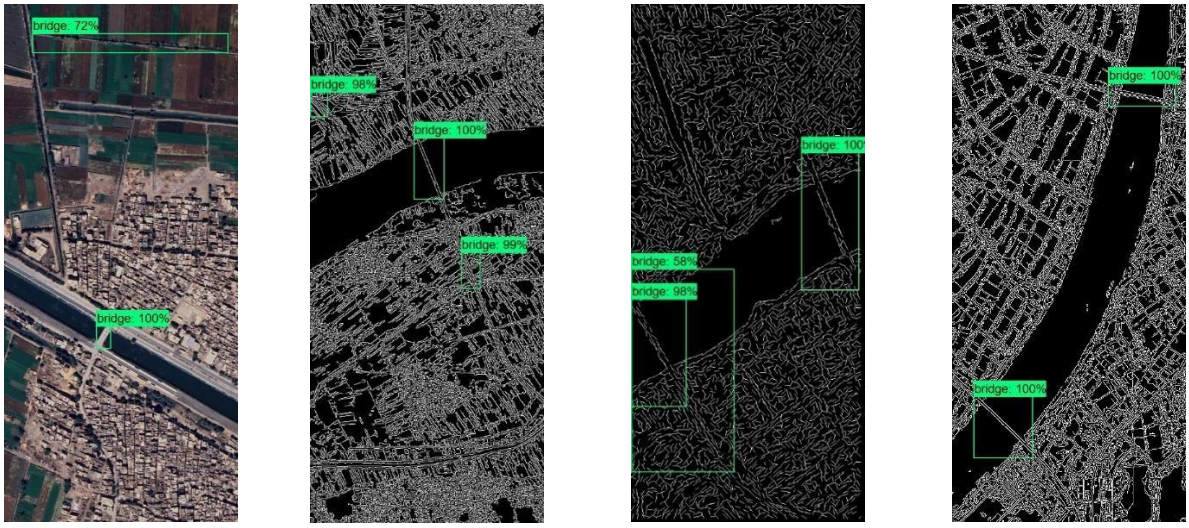


А) Б) В) Г)

Рис. 5 – Распознавание объекта на изображении: А) – исходное изображение; Б) – результат распознавания без обработки; В) – результат распознавания на изображении, обработанном алгоритмом Хафа; Г) - результат распознавания на изображении, обработанном алгоритмом Хюккеля

При обработке большого набора изображений выявлены недостатки предложенного метода: возможно возникновение дополнительных ложных распознаваний при работе с предварительно обработанными изображениями. Однако, при правильном подборе настраиваемых параметров и отсутствии на изображении объектов, похожих на объект интереса (в рассматриваемой задаче дорога может быть подобна мосту), задача решается корректно. Также стоит отметить, что ложное распознавание дороги как объекта интереса присутствует и при отсутствии предварительной обработки изображений алгоритмами выделения контуров.

Пример ложного распознавания при поиске объекта на необработанном изображении, а также примеры корректных и ложных распознаваний представлены на рис. 6.



А) Б) В) Г)

Рис. 6 – Примеры ложного распознавания на исходном изображении (А), на обработанном изображении (Б), успешного распознавания на предварительно обработанных изображениях (В и Г)

Полученные временная и точностная оценки предварительной обработки изображений алгоритмами поиска контуров представлены в таблице 1. Вычисления проводились на процессоре Intel(R) Core(TM) i5-10400F CPU @ 2.90GHz, размер поступающих на вход изображений 450 на 1000 пикселей. Обучающая выборка сверточных

нейронных сетей содержала 200 размеченных изображений, тестовая – 50. Оценка времени и точности обработки проводилась на ранее используемых, а также 35 новых изображениях. Каждое изображение содержало не менее одного объекта интереса. Пороговое значение вероятности, при котором выполнялся вывод области распознавания объекта, было принято равным 50%.

ТАБЛИЦА 1
ОЦЕНКА ВЛИЯНИЯ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ НА ВРЕМЯ И ТОЧНОСТЬ РАСПОЗНАВАНИЯ

	faster_rcnn_inception_v2		ssd_mobilenet_v1	
	Исходные изображения	Обработанные изображения	Исходные изображения	Обработанные изображения
Время предварительной обработки, сек	-	0,14…0,56	-	0,14…0,56
Время обработки одного изображения сетью, сек	0,36	0,35…0,37	0,045	0,042…0,044
Время обнаружения объектов интереса на изображении, сек	0,36	0,51…1,05	0,045	0,144…0,602
% успешного обнаружения	92	75…91	79	61…74
% ложного обнаружения от всех найденных	15	15…18	1,5	5…7

Полученные результаты показывают, что использование одноэтапных сверточных нейронных сетей позволяет достичь времени обработки изображений, удовлетворяющего требованиям систем реального времени. Точность сетей данного класса при обработке всего входного изображения уступает модели двухэтапной архитектуры. Применение искусственной нейронной сети высокой скорости работы только к областям изображения, на которых, предположительно, находится объект интереса, должно привести к уменьшению количества ошибок распознавания.

V. Выводы и заключение

Для распознавания на изображениях объектов искусственного происхождения в условиях ограниченного времени обработки, свойственных системам реального времени, предлагается использование гибридных алгоритмов, представляющих комбинацию нейросетевых и классических методов обнаружения и классификации объектов.

Предложены три структуры данных алгоритмов, включающих в себя выделение прямолинейных контуров на изображениях, а также использование сверточных нейронных сетей.

Реализована одна из описанных структур, основанная на предварительном переходе к контурному представлению входного изображения с последующим поиском объекта интереса. Выполнен численный эксперимент на универсальном компьютере, целью которого являлось определение возможности использования данных алгоритмов в системах реального времени. На основе проведенных оценок времени и точности работы описанного алгоритма в сравнении с применением только сверточных нейронных сетей сделан вывод, что за время обучения и обработки отвечает тип сети, а не наличие предварительной обработки изображений. Применение искусственных нейронных сетей ко всему контурному изображению не дает выигрыша в точности или времени обнаружения и классификации объектов.

Время работы сетей одноэтапной архитектуры позволяет решать задачи в реальном масштабе времени, обеспечивая распознавание объектов более чем на десяти кадрах в секунду. При определении областей- кандидатов расположения объекта интереса классическими методами компьютерного зрения возможно достичь меньшего времени обработки, чем при использовании двухэтапных архитектур сверточных нейронных сетей при сравнимых значениях точности.

Предполагается исследование двух оставшихся предложенных в работе вариантов реализации гибридного алгоритма и последующее их сравнение между собой и с существующими алгоритмами в части возможности их применения в задачах реального времени.

Список литературы

1. N. Ardalani, S. Pal, and P. Gupta, DeepFlow: A cross-stack pathfinding framework for distributed ai systems, 2022. URL: https://arxiv.org/abs/2211.03309.

2. D. Rathgamage, R. Aygun, M. Karakaya, A Multistage Framework for Detection of Very Small Objects, 2023, 9-14. 10.1145/3589572.3589574.

3. M. Hao, Y. Liu, X. Zhang, J. Sun, Labelenc: A new intermediate supervision method for object detection. In European Conference on Computer Vision, 2020, Springer, pp. 529–545.

4. B. Zoph, E.D. Cubuk, G. Ghiasi, T.Y. Lin, J. Shlens, Q.V. Le, Learning data augmentation strategies for object detection, 2019, URL: https://arxiv.org/abs/1906.11172.

5. H. Qiu, Y. Ma, Z. Li, S. Liu and J. Sun, BorderDet: Border feature for dense object detection, 2020, Proc. Eur. Conf. Comput. Vis., pp. 549-564.

6. D. Gaddam, J. Lahoud, F.S. Khan, R.M. Anwer, H. Cholakkal, CMR3D: Contextualized Multi-Stage Refinement for 3D Object Detection, 2022, Proceedings of the 4th ACM International Conference on Multimedia in Asia.

7. L. Zeming, C. Peng, G. Yu, X. Zhang, Y. Deng, J. Sun, DetNet: Design Backbone for Object Detection, 2018, 15th European Conference, Munich, Germany, September 8–14, Proceedings, Part IX. 10.1007/978-3-030-01240-3_21.

8. J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, Graph r-cnn for scene graph generation, 2018, ECCV.

9. Смольянинов, В. А. Проектирование программного комплекса обнаружения и распознавания дорожных знаков в потоковом видео / В. А. Смольянинов, Ю. С. Белов // Научное обозрение. Технические науки. – 2021. –
№ 4. – С. 16-21.

10. Стротов, В. В. Комбинирование нескольких алгоритмов для увеличения точности обнаружения и локализации пешеходов / В. В. Стротов, П. Е. Жгутов // Труды Международной конференции по компьютерной графике и зрению "Графикон". – 2022. – № 32. – С. 628-635. – DOI 10.20948/graphicon-2022-628-635.

11. L. A. Varga, A. Zell, Tackling the Background Bias in Sparse Object Detection via Cropped Windows, 2021, IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 2768-2777.

12. LabelImg — image annotation tool — 2023. — https://github.com/heartexlabs/labelImg

13. TensorFlow 1 Detection Model Zoo // Google. — https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md

14. TensorFlow 2 Detection Model Zoo // Google. — https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md

15. X. Lu, B. Li, Q. Li, J. Yan, MimicDet: Bridging the Gap Between One-Stage and Two-Stage Object Detection, 2020, URL: https://arxiv.org/abs/2009.11528.

16. S. Huang, F. Qin, P. Xiong, N. Ding, Y. He, X. Liu, Tp-lsd: Tri-points based line segment detector, 2020,ECCV.

17. Интегральные роботы: сборник статей : пер. с англ. и яп. / под ред. Г. Е. Поздняка. - Москва : Мир, 1973.

18. Manfred H. Hueckel. A Local Visual Operator Which Recognizes Edges and Lines. Journal of the Association for Computing Machinery, vol. 20, No. 4, October 1973.

19. Солодовщиков А.Ю., Платонов А.К. Исследование метода Карунена-Лоэва // Препринты ИПМ им. М.В.Келдыша. 2006. № 19. 29 с.

20. Кэлер А., Брэдски Г. Изучаем OpenCV 3 / пер. с анг. А. А. Слинкина. – М.: ДМК Пресс, 2017. – 826 с.: ил.

21. Компьютерное зрение. Современные методы и перспективы развития / под ред. Р.Дэвис, М.Терк; пер. с англ. В.С. Яценкова. — М.: ДМК Пресс, 2022. — 690 с.: ил.

УДК 004:932
DOI: 10.25206/978-5-8149-3873-2-2024-484-490

Алгоритмы и программное обеспечение для оценки точности совмещения изображений Algorithms and Software for Evaluating the Accuracy of Image Alignment

Е. С. Жгутова[1], П. В. Бабаян[1]
[1]ФГБОУ ВО РГРТУ им. В.Ф. Уткина, г. Рязань, ул. Гагарина, 59/1, 390005, Россия

E. S. Zhgutova[1], P. V. Babayan[1]
[1]Ryazan State Radio Engineering University named after V.F. Utkin, Ryazan, Gagarina str., 59/1, 390005, Russia

Аннотация. В данной работе рассмотрена проблема оценки ошибок совмещения изображений. Цель работы заключается в разработке алгоритмов и программного обеспечения для определения зависимости точности оценивания параметров различных геометрических преобразований от расположения опорных точек на изображении и оценки точности совмещения изображений. Объектом исследования являются изображения движущихся и неподвижных объектов (автомобилей, строений), сформированные камерами видимого и инфракрасного диапазонов, подверженные геометрическим преобразованиям. В работе приведен общий алгоритм нахождения корреляционной матрицы для каждого вида преобразований, программное обеспечение написано на языке Matlab. Научная новизна работы заключается в том, что впервые исследована зависимость точности оценки параметров геометрических преобразований от расположения опорных точек на изображении. Проведены экспериментальные исследования разработанных алгоритмов с целью оценить ошибки совмещения изображений, и приведены результаты их работы применительно к типовым изображениям на примере евклидова преобразования. Показана важность учета этих ошибок при разработке алгоритмов обработки изображений. Полученные результаты могут использоваться в алгоритмах стабилизации изображений, обнаружения изменений сцен, выделения и обнаружения движущихся объектов.

***Ключевые слова*:** совмещение изображений, опорные точки, геометрические преобразования, стабилизация изображений, евклидово преобразование, корреляционная матрица

Аннотация. In this paper, the problem of estimating image alignment errors is considered. The purpose of the work is to develop algorithms and software to determine the dependence of the accuracy of estimating the parameters of various geometric transformations and evaluating the accuracy of image alignment on the location of reference points in the image. The object of the study is images of moving and stationary objects (cars, buildings) formed by cameras of the visible and infrared ranges, subject to geometric transformations. The paper presents a general algorithm for finding the correlation matrix for each type of transformation, the software is written in Matlab. The scientific novelty of the work lies in the fact that for the first time the dependence of the accuracy of estimating the parameters of geometric transformations on the location of reference points in the image was investigated. Experimental studies of the developed algorithms have been carried out in order to assess the errors of image alignment and the results of their work are presented in relation to typical images using the example of the Euclidean transformation. The importance of taking these errors into account when developing image processing algorithms is shown. The results obtained can be used in image stabilization algorithms, scene change detection, extraction and detection of moving objects.

***Keywords*:** image alignment, reference points, geometric transformations, image stabilization, Euclidean transformation, correlation matrix

I. Введение

В настоящее время существующие алгоритмы показывают неплохие результаты при оценке параметров геометрических преобразований, но при этом ошибки оценивания параметров все равно существуют [1, 2]. Во многих работах исследователи предполагают, что эта ошибка незначительна, либо ее статистические характеристики постоянны для всего кадра, либо не зависят от расположения опорных участков. Степень корректности этих предположений зависит от выбранного алгоритма определения параметров геометрических преобразований.

Для совмещения и сравнения изображений, полученных в условиях геометрических преобразований, необходимо предварительно вычислить оценки этих параметров, чтобы попытаться скомпенсировать искажения и стабилизировать изображение [3]. Тогда возникает задача выбора подхода к оценке параметров геометрических преобразований, который выдавал бы точные оценки и требовал бы невысоких вычислительных затрат.

Анализ литературных источников показывает, что существующие методы оценки параметров геометрических искажений можно разделить условно на 3 группы:

1) Глобальные методы;
2) Локальные методы;
3) Методы, основанные на использовании сверточных нейронных сетей.

К глобальным методам оценки параметров геометрических преобразований относятся алгоритмы, которые вычисляют параметры преобразования, сравнивая изображения полностью. Среди таких алгоритмов можно выделить: пространственно-временные, корреляционные, разностные и спектральные [3, 4]. Недостатками корреляционных подходов являются: большой объем вычислений, который замедляет работу алгоритмов, требование алгоритмов к значительному количеству априорной информации, чувствительность алгоритмов к шумам. Кроме того, применение данных алгоритмов дает хорошие результаты в основном для оценки глобальных преобразований, например, поворот, параллельный сдвиг и т.д.

Существует альтернативный подход, позволяющий существенно уменьшить требуемые вычисления. Локальные методы основаны на поиске элементов на изображении, определении их координат, по которым потом оценивают параметры геометрических преобразований. Такие алгоритмы еще называют структурными из-за необходимости проведения предварительного анализа изображений [5-9]. Преимуществом такого подхода является оценка любых многопараметрических преобразований (аффинное или проективное), а также работа с изображениями разных спектральных диапазонов. Недостатком является сложная процедура поиска некоторых опорных элементов, в частности на однородном фоне, а также их автоматическое сопоставление.

Множество задач компьютерного зрения решается с помощью сверточных искусственных нейронных сетей (ИНС) [10]. Для работы систем на основе ИНС требуется достаточно мощная аппаратная составляющая оборудования. Для обучения сетей используют обучение с учителем (например, метод обратного распространения ошибки), либо без учителя (подаются входные данные, и сеть сама пытается найти между ними связи). Прежде чем запустить процесс обучения, необходимо загрузить довольно большое количество входных данных – те изображения, по которым ИНС будет учиться распознавать объекты или оценивать параметры геометрических преобразований. Среди недостатков ИНС можно отметить наличие большого числа меняющихся параметров. Использование для оценки параметров геометрических преобразований методов, основанных на использовании ИНС, затруднено подбором качественной обучающей базы. Это возникает из-за того, что в открытых и доступных базах изображений не всегда можно найти уникальные объекты [11].

II. Постановка задачи

Анализ существующих подходов и методов, решающих задачу оценки параметров геометрических преобразований, позволяет сделать вывод о том, что в настоящее время до сих пор не существует какого-то универсального подхода, дающего существенный выигрыш в скорости, качестве работы, в аппаратной реализации и др. В конечном итоге оценивать параметры геометрических преобразований целесообразнее методом на основе выбора опорных элементов/точек. Работа выбранного подхода заключается в поиске опорных точек на изображении, которые соответствуют друг другу. Самыми простыми элементами являются точки [3]. Данные алгоритмы показывают неплохую точность работы и не требуют сложных вычислений.

Постановка задачи заключается в следующем: пусть имеется два изображения, полученные в разные моменты времени. Из-за разницы в расположении датчиков или их движения в пространстве получается, что положения точек одного изображения не соответствуют положениям точек другого. Имеют место различные геометрические преобразования координат. Требуется найти параметры преобразования, с помощью которых удастся максимально точно установить соответствие между двумя изображениями.

Пусть нам известны координаты опорных точек на исходном (x_1, y_1) и текущем (x_2, y_2) изображении. Если подставить эти значения в выражение для любого выбранного преобразования, тогда получим систему уравнений. Решение этой системы позволит определить численные значения параметров геометрических преобразований, присутствующих на изображении [3].

Авторы предлагают увеличить точность совмещения изображений за счет учета ошибок совмещения при разработке алгоритмов выделения движения, стабилизации изображений и др. В некоторых задачах присутствие таких ошибок может серьезно влиять на корректность работы алгоритмов. Для этого необходимо разработать математическую модель алгоритмов совмещения изображений, а также программное обеспечение для проведения таких исследований.

III. Теория

1. Общий алгоритм решения задачи

Общий алгоритм решения задачи оценки параметров сводится к следующим шагам:

1) Имеется математическая модель преобразования, которая связывает точки текущего изображения (x_i', y_i') и исходного (x_i, y_i). Параметры преобразования из математической модели образуют вектор **U**, размерность которого зависит от выбранного преобразования.

2) Координаты точек текущего изображения (x_i', y_i') образуют вектор **Z**, а **H** – матрица, состоящая из координат точек исходного изображения.

3) Если геометрические преобразования можно представить линейной моделью, то чтобы найти параметры преобразования (параметры вектора **U**) необходимо решить систему уравнений **HU**=**Z**. Решение системы уравнений получим из соотношения $\hat{\mathbf{U}} = (\mathbf{H}^{\mathbf{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{Z}$. Чтобы оценить точность оценивания параметров вектора $\hat{\mathbf{U}}$ необходимо определить корреляционную матрицу $\mathbf{K}_{\mathbf{U}}$. Из предположения о том, что координаты точек текущего изображения (x_i', y_i') некоррелированные и имеют одну и ту же дисперсию σ_Z^2, следует, что выражение для определения корреляционной матрицы выглядит следующим образом:

$$\mathbf{K}_{\mathbf{U}} = (\mathbf{H}^{\mathbf{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{H}(\mathbf{H}^{\mathbf{T}}\mathbf{H})\cdot\sigma_z^2 = (\mathbf{H}^{\mathbf{T}}\mathbf{H})^{-1}\cdot\sigma_z^2. \quad (1)$$

Конкретный вид матрицы $\mathbf{K}_{\mathbf{U}}$ определяется типом преобразований. В работах [12-14] авторами уже рассмотрены преобразования: масштаб-смещение, подобия, аффинные преобразования. Полученные корреляционные матрицы показывают, что точность оценивания параметров геометрических преобразований зависит от разброса опорных точек по изображению. Чем дальше друга от друга выбираются опорные точки, тем точнее оцениваются параметры преобразования. Кроме того, для некоторых преобразований, например, для аффинного и преобразования подобия, количество опорных точек **N** тоже влияет на точность оценивания параметров: чем их больше, тем точнее оцениваются параметры преобразования.

Таким образом, чтобы добиться высокой точности совмещения изображений, необходимо выбирать больше опорных точек, и располагаться они должны по всей площади изображения.

2. Алгоритм поиска корреляционной матрицы для евклидова преобразования

Рассмотрим алгоритм решения задачи оценки параметров геометрических преобразований на примере евклидова преобразования. Математическая модель для евклидовых преобразований получается при комбинации преобразований смещения и поворота. Выражение для евклидовых преобразований имеет вид:

$$\begin{aligned} x_i' &= \cos\varphi\cdot x_i + \sin\varphi\cdot y_i + \alpha; \\ y_i' &= -\sin\varphi\cdot x_i + \cos\varphi\cdot y_i + \beta, \end{aligned} \quad (2)$$

где x_i', y_i' – координаты опорных точек текущего изображения, x_i, y_i – координаты опорных точек исходного изображения, φ – параметр изменения угла поворота, α, β – смещение по оси *x* и *y*.

При малых углах поворота $\cos\varphi \approx 1$ и $\sin\varphi \approx \varphi$, тогда выражение 2 принимает вид:

$$\begin{aligned} x_i' &= x_i + \varphi\cdot y_i + \alpha; \\ y_i' &= -\varphi\cdot x_i + y_i + \beta, \end{aligned} \quad (3)$$

где вектор $\mathbf{U} = [\varphi \quad \alpha \quad \beta]^{\mathbf{T}}$, вектор $\mathbf{Z} = [x_i' - x_i \quad y_i' - y_i \quad \ldots]^{\mathbf{T}}$ и матрица $\mathbf{H} = \begin{bmatrix} y_1 & 1 & 0 \\ -x_1 & 0 & 1 \\ \cdots & \cdots & \cdots \end{bmatrix}^{\mathbf{T}}$.

Корреляционная матрица:

$$\mathbf{K}_{\mathbf{U}} = \begin{bmatrix} \sum(y_i^2 + x_i^2) & \sum y_i & -\sum x_i \\ \sum y_i & N & 0 \\ -\sum x_i & 0 & N \end{bmatrix}^{-1}\cdot\sigma_Z^2 = \frac{\sigma_Z^2}{N^2\sum(x_i^2 + y_i^2)}\begin{bmatrix} N^2 & -N\sum y_i & N\sum x_i \\ -N\sum y_i & N\sum(x_i^2 + y_i^2) - \sum x_i^2 & -\sum x_i y_i \\ N\sum x_i & -\sum x_i y_i & N\sum(x_i^2 + y_i^2) - \sum y_i^2 \end{bmatrix},$$

где *N* – количество выбранных опорных точек, σ_Z^2 – дисперсия оценки координат опорных точек.

После центрирования координат получим:

$$\mathbf{K_U} = \frac{\sigma_Z^2}{\sum(\overset{\circ}{x}_i^2 + \overset{\circ}{y}_i^2)} \begin{bmatrix} 1 & -M_y & M_x \\ -M_y & \mu_{2,x} + \sigma_x^2 & -M_x M_y \\ M_x & -M_x M_y & \mu_{2,y} + \sigma_y^2 \end{bmatrix}, \quad (4)$$

где $\overset{\circ}{x}_i = x_i - M_x$, $\overset{\circ}{y}_i = y_i - M_y$ – центрированные координаты опорных точек, M_x, M_y – средние значения величин x_i, y_i, $\mu_{2,x}, \mu_{2,y}$ – второй начальный момент, σ_x^2, σ_y^2 – дисперсия величин *x* и *y*.

Для удобства вычислений приведем матрицу (4) к диагональному виду. Для этого необходимо переместить систему координат в точку (M_x, M_y). Получим выражение для корреляционной матрицы, в котором по главной диагонали располагаются дисперсии параметров евклидова преобразования.

$$\mathbf{K_U} = \frac{\sigma_Z^2}{\sum(\overset{\circ}{x}_i^2 + \overset{\circ}{y}_i^2)} \operatorname{diag}(1, 2\sigma_x^2, 2\sigma_y^2).$$

IV. Программное обеспечение для оценивания точности совмещения изображений

Программное обеспечение (ПО) для исследования оценивания точности совмещения изображений и оценки параметров геометрических преобразований было разработано на языке Matlab.

На рис. 1 представлен интерфейс разработанного ПО. Программа позволяет загружать как изображения, так и видеопоследовательности, полученные с помощью ТВ и ИК-датчиков изображений, сохранять изменения, загружать сохраненные настройки, создавать новый проект, сохранять текущие настройки с возможностью его загрузки при повторном запуске программы и т.д.

С помощью меню «Настройка» можно задавать шаг чтения кадров из видеосюжета, формат сохранения файлов (AVI, BMP), метод интерполяции и вид преобразования.

Если установить флаг «Преобразованное», то можно увидеть преобразование одного изображения относительно другого.

В программе предусмотрена возможность выбора соответствия точка-точка. Кнопка «+» означает, что точку можно разместить на изображении, а кнопка «-» удаляет точку с изображения, кнопка «<->» позволяет двигать точку по изображению.

Кнопка «Параметры» выводит на экран окно, в котором отображаются значения параметров преобразования, матрица параметров преобразования, невязка и значение СКО невязки.

Кнопка «Корреляционная матрица изображения» выводит на экран изображения дисперсии по оси *X* и *Y*, а также значение дисперсии в каждой точке изображения.

Кнопка «Показать совмещение изображений» выводит на экран разницу после наложения одного изображения на другое.

На рис.1 крестиками и цифрами обозначаются выбранные опорные точки, синим цветом отмечены точки на исходном изображении, красным – на преобразованном.

V. Результаты экспериментов

Цель проведения экспериментальных исследований заключалась в следующем: необходимо было сравнить точность оценки параметров евклидова преобразования при различных фактических расположениях опорных точек на изображении. Для проведения исследований использовались кадры из видеосюжета, снятого на камеру ТВ-диапазона (рис.2). Размер изображения составлял 692x1024 пикселей. На рис.2,а представлен кадр №1, который будем считать исходным, на рис.2,б – кадр №60, к которому уже применено евклидово преобразование. Для преобразования рассчитывались: корреляционная матрица ошибок (при этом значение дисперсии σ_Z^2 принималось равным 1), значение дисперсии в каждой точке изображения. Помимо этого, на основе полученных оценок параметров геометрического преобразования текущее изображение возвращалось в систему координат исходного изображения с целью их дальнейшего совмещения.

Эксперимент повторялся 2 раза для каждого варианта выбора опорных точек. Первый вариант разрешает выбирать опорные точки только в определенной зоне изображения (в эксперименте выбираем только в правом нижнем углу), и второй вариант не ограничивает выбор, соответственно, точки выбираются по всей площади изображения.

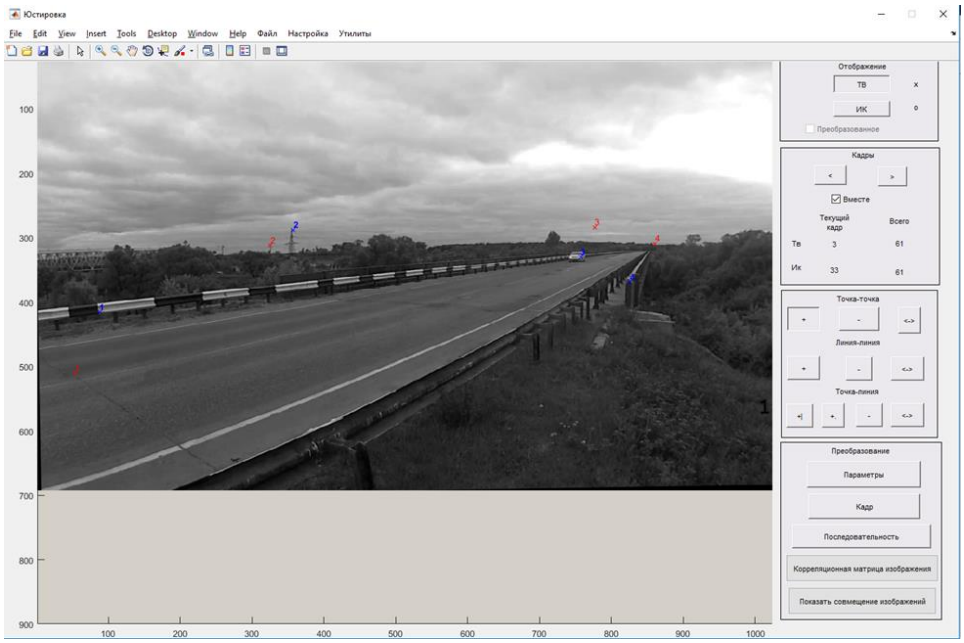


Рис. 1. Интерфейс программы

а б

Рис. 2. Сюжет для экспериментов: а) исходный кадр; б) преобразованный

Корреляционная матрица для каждого эксперимента представлена ниже:

$\mathbf{K_U} = \mathrm{diag}(2.3270 \cdot 10^{-4}, 10.2026, 77.8139).$

$\mathbf{K_U} = \mathrm{diag}(2.6663 \cdot 10^{-6}, 0.9545, 1.3181).$

Значения матриц показывают, что более точных оценок удается добиться при втором эксперименте.

На рис.3 представлены изображения дисперсии в каждой точке, рис.3,а – для первого эксперимента, рис.3,б – для второго.

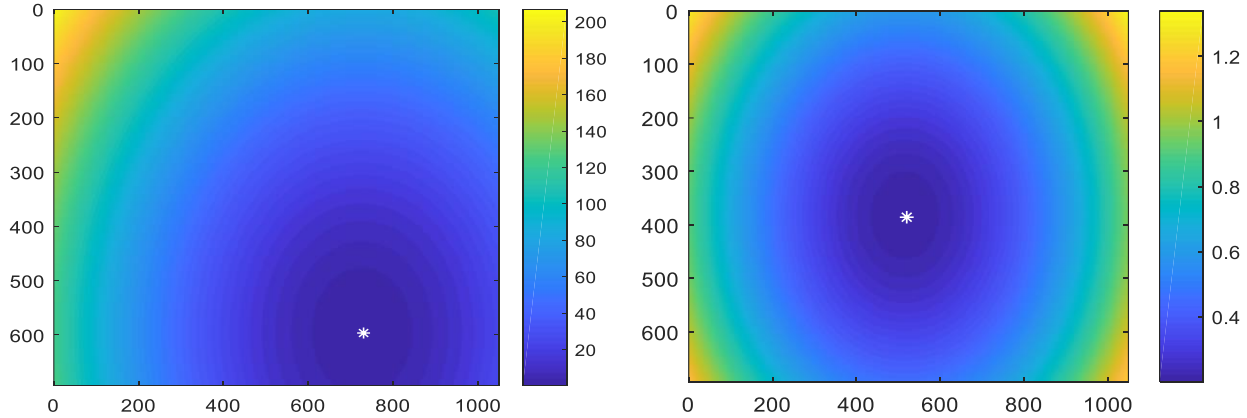


Рис. 3. Значение дисперсии в каждой точки изображения: а) для первого эксперимента; б) для второго эксперимента

Рис.3 показывает значения дисперсии для каждой точки изображения, из рисунка видно, что при втором эксперименте разброс значений дисперсии намного меньше, соответственно, параметры удалось определить более точно. Остаточные разности при совмещении изображений представлены на рис.4.

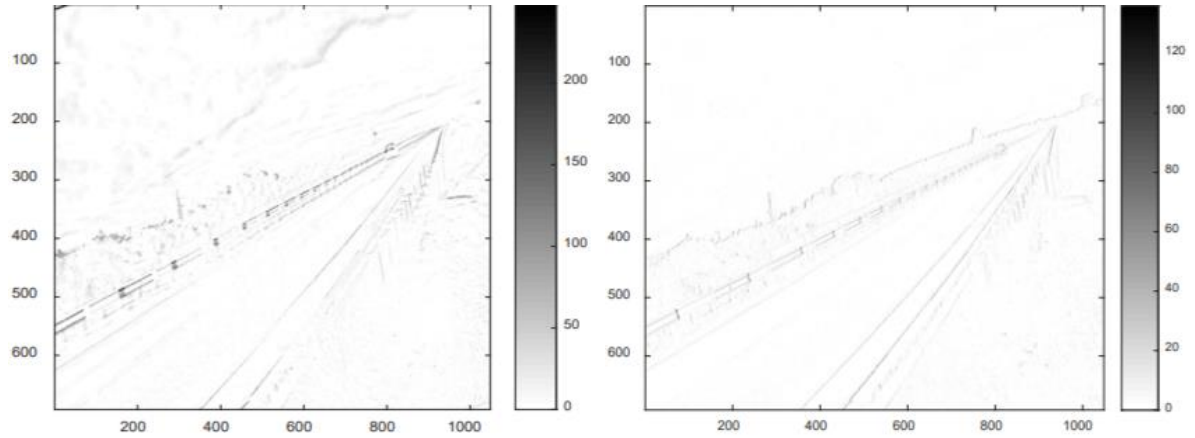


Рис. 4. Совмещение изображений: а) первый эксперимент; б) второй эксперимент

Из рисунка следует, что совмещение изображений для второго эксперимента намного точнее, чем для первого. Это видно по яркости и количеству различных линий на изображении. На рис. 4,б их намного меньше (примерно в 2 раза, что можно наблюдать также по шкале справа), и они светлее.

Представленные эксперименты показывают, что расположение и выбор опорных точек напрямую оказывает влияние на точность совмещения изображений. Чем дальше разбросаны друг от друга и по всему изображению опорные точки, тем точнее оцениваются параметры преобразования, а точность совмещения оказывается выше. Помимо этого, и количество опорных точек вносит дополнительную точность. Чем больше точек выбрано, тем выше точность и оценки параметров преобразования, и дальнейшего совмещения.

VI. Выводы и заключение

Проведенные экспериментальные исследования подтвердили сделанные ранее теоретические выводы [12-14]. На точность оценивания параметров преобразований в первую очередь влияет расположение опорных точек: чем дальше они разнесены относительно друг друга и чем больше охватывают площадь изображения, тем точнее будут оценки параметров геометрических преобразований. В работах [12-14] также сказано, что точность оценки параметров для некоторых преобразований зависит и от количества выбранных опорных точек. Для получения более точных оценок следует увеличивать их количество. Если же по какой-то причине опорные точки группируются в какой-то зоне изображения (в центре, в углах, на какой-то прямой и др.), то достаточно точно оценить параметры геометрических преобразований не удастся, даже при увеличении количества выбранных опорных точек. В свою очередь, более точные оценки позволяют совместить изображения с наименьшей ошибкой.

В заключение следует отметить, что исследование алгоритмов, позволяющих оценить точность оценивания параметров геометрических преобразований, является важной задачей для дальнейшей разработки алгоритмов стабилизации изображений, обнаружения изменений сцен, выделения и обнаружения движущихся объектов. Учет ошибок совмещения изображений позволит сократить ложные выделения и повысить точность работы данных алгоритмов.

Список литературы

1. Han, Y., Javed, A., Jung, S., Liu, S. Object-Based Change Detection of Very High-Resolution Images by Fusing Pixel-Based Change Detection Results Using Weighted Dempster–Shafer Theory. Remote Sens, 2020. 12. 983. https://doi.org/10.3390/rs12060983

2. Babayan P.V., Kozhina E.S. A Study of the Automatic Object Extraction Algorithm on the Image Sequence Under Scaling Ttransformations // 2023 25-th International Conference on Digital Signal Processing and its Applications (DSPA). – IEEE, 2023. – PP. 1-4.

3. Бабаян П.В., Смирнов С.А., Муравьев В.С., Стротов В.В. Обработка изображений в системах обнаружения и сопровождения объектов. Классические методы. Учебное пособие для вузов. – М.: Горячая линия – Телеком, 2024. – 168 с: ил.

4. Ефимов А.С., Козлитин И.А. Критерий перезахвата на ложный объект при корреляционном методе сопровождения // Графикон-2012. – 2012. –С. 266-268

5. Стротов В.В. Сравнение двух структурных алгоритмов оценки параметров геометрических преобразований изображений // Цифровая обработка сигналов. – 2013. – №. 3. – С. 21-25.

6. Амосов О.С. и др. Система автоматического управления шлагбаумом с использованием технологий нечеткого вывода и компьютерного зрения // Вестник евразийской науки. – 2017. – Т. 9. – №. 1 (38). – С. 42

7. Xiang Y., Wang F., You H. OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas // IEEE Transactions on Geoscience and Remote Sensing. – 2018. – Т. 56. – №. 6. – С. 3078-3090

8. Xu S. et al. Binocular measurement method for the continuous casting slab model based on the improved BRISK algorithm //Applied Optics. – 2022. – Т. 61. – №. 11. – pp. 3019-3025

9. Luo C. et al. Overview of image matching based on ORB algorithm //Journal of Physics: Conference Series. – IOP Publishing, 2019. – Т. 1237. – №. 3. – P. 032020

10. Бучацкий П.Ю., Онищенко С.В., Теплоухов С.В. Разработка метода определения геометрических параметров объектов на изображении на основе нейросетевого подхода // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. – 2022. – №. 3 (306). – С. 57-62.

11. Владимиров Д.М., Алексеев А.А. Применение нейросетей в современных военных технологиях: возможности и вызовы // Вестник науки. – 2023. – Т. 2. – №. 12 (69). – С. 1152-1159

12. Бабаян П.В., Кожина (Жгутова) Е.С. Влияние точности оценивания координат опорных участков на точность совмещения изображений // GraphiCon-2023: труды 33-й Междунар. конф. по компьютерной графике и машинному зрению (Москва, 19–21 сент. 2023 г.). – М.: Институт прикладной математики им. М.В. Келдыша РАН, 2023. – 474-481 с DOI: 10.20948/graphicon-2023-474-481

13. Бабаян П.В., Кожина (Жгутова) Е.С. Влияние точности оценивания координат опорных участков на точность совмещения изображений в системах дистанционного зондирования Земли // 8-ая международная научно-техническая конференция «В.Ф. Уткин – 100 лет со дня рождения. Космонавтика. Радиоэлектроника. Геоинформатика»: Мат. докл. / Рязанс. гос. радиотехн. Университет им. В.Ф. Уткина. Рязань, 2023, 360-363 с. – ISBN 978-5-7722-0388-0

14. Babayan P.V., Kozhina (Zhgutova) E.S. Estimation of the Accuracy of Determining the Parameters of Affine Transformations // 2024 26th International Conference on Digital Signal Processing and its Applications (DSPA). – IEEE, 2024. – С. 1-5. DOI: 10.1109/DSPA60853.2024.10510022

УДК 004.853
DOI: 10.25206/978-5-8149-3873-2-2024-491-499

Применение методов обучения с подкреплением для задачи управления роботом-манипулятором в виртуальной среде
The application of reinforcement learning methods for the task of controlling a robot manipulator in a virtual environment

Н.Е. Залогин[1], Д.С. Григорьев[1]
[1]Национальный исследовательский Томский политехнический университет, г. Томск, Россия

N.E. Zalogin[1], D.S. Grigoriev[1]
[1]National Research Tomsk Polytechnic University, Tomsk, Russian Federation

Аннотация. Разработка эффективных методов для обучения интеллектуальных агентов в задачах управления роботами-манипуляторами в виртуальных средах имеет важное значение для улучшения точности и эффективности промышленных процессов, а также дальнейших научных исследований. Настоящая работа направлена на исследование и совершенствование алгоритмов обучения с подкреплением, таких как Deep Q-Network (DQN) и Proximal Policy Optimization (PPO), для управления агентами в виртуальной среде KukaDiverseObjectEnv на платформе PyBullet, на основе графической информации, получаемой с виртуальной камеры робота. Результаты показали, что обученные агенты могут решать поставленную задачу, а модификации алгоритмов снижают время обучения и количество необходимых шагов. В ходе тестирования, проведённого на 1000 эпизодах, алгоритмы продемонстрировали возможность управления манипулятором. Данные разработки имеют потенциал для применения в промышленности и дальнейшего развития уже в реальных условиях, что подчёркивает их значимость для современной автоматизации и робототехники.

***Ключевые слова*:** обучение с подкреплением, виртуальная среда, Deep Q-network, Proximal policy optimization, робот-манипулятор, робототехника.

Аннотация. The development of effective methods for training intelligent agents in tasks involving the control of robotic manipulators in virtual environments is essential for improving the accuracy and efficiency of industrial processes, as well as for further scientific research. This paper focuses on research and improvement of reinforcement learning algorithms, such as Deep Q-Network (DQN) and Proximal Policy Optimization (PPO), for managing agents in the KukaDiverseObjectEnv virtual environment on the PyBullet platform, based on graphical information obtained from the robot's virtual camera. The results showed that the trained agents are capable of solving the given task, and algorithm modifications reduce both training time and the number of required steps. During testing conducted over 1000 episodes, the algorithms demonstrated the ability to control the manipulator. These developments have the potential for industrial applications and further advancement in real-world conditions, highlighting their significance for modern automation and robotics.

***Ключевые слова*:** reinforcement learning, virtual environment, Deep Q-Network, Proximal Policy Optimization, robot manipulator, robotics.

I. Введение

Инновационные решения в робототехнике находят всё большее применение в различных секторах, таких как промышленное производство, логистика и медицина. Роботы-манипуляторы, обладающие высокой точностью и прецизионностью, становятся незаменимыми для выполнения сложных задач. Одной из ключевых проблем в этой области остаётся обучение роботов новым навыкам. Традиционные методы программирования требуют значительных временных и трудовых ресурсов и часто не справляются с задачами, требующими адаптации к изменяющимся условиям [1–3].

Алгоритмы глубокого обучения с подкреплением (Reinforcement Learning, RL) представляют собой перспективный подход для создания интеллектуальных агентов, способных к самообучению через метод проб и ошибок. Эти агенты могут принимать решения в сложных и динамических средах, как виртуальных, так и реальных. Так, виртуальные платформы, такие как PyBullet [4] и Gazebo [5], предоставляют гибкие и адаптируемые среды для обучения интеллектуальных агентов, позволяющие быстро моделировать различные сценарии, тестировать разработанные алгоритмы и оценивать их эффективность. Несмотря на отличия от реального мира, виртуальные среды исключают проблему «безопасного обучения», актуальную для физических роботов [6].

Исследования в области RL-агентов, работающих в виртуальных средах с манипуляторами, стремительно развиваются. Важной задачей является разработка и исследование алгоритмов, которые могут эффективно обучаться в виртуальных средах, имитирующих реальные задачи манипулирования объектами. Сложность обучения в таких средах заключается в высокой размерности пространства состояний и действий, графических входных данных, а также в необходимости обеспечения стабильности и эффективности обучения агентов. Ряд работ рассматривают применение RL-алгоритмов для решения задач манипулирования объектами, а также множество вопросов, включающих безопасность [6], результативность алгоритмов и стратегии для различных сложных виртуальных сред с визуальной информацией [7–10], создание собственных виртуальных сред с использованием компьютерного зрения [11, 12].

Одними из применяемых методов в этой области являются алгоритмы обучения с подкреплением, такие как Deep Q-Network (DQN) [13] и Proximal Policy Optimization (PPO) [14]. Алгоритм DQN использует глубокие нейронные сети для аппроксимации функции награды и позволяет агенту эффективнее обучаться оптимальной стратегии. Алгоритм PPO улучшает стабильность обучения за счёт использования доверительных границ для обновления политик, что делает его одним из передовых методов для обучения агентов в условиях непрерывных и дискретных пространств действий.

II. Постановка задачи

Основной целью данного исследования является разработка, тестирование и модификация RL-агентов в виртуальной среде, обучаемых на основе графических данных с виртуальной камеры робота, а также оценка их производительности и эффективности.

В данной работе использована платформа симуляции PyBullet [4], основанная на физическом движке Bullet Physics, которая зарекомендовала себя как эффективный инструмент для физического моделирования и симуляции роботов. В качестве виртуальной среды используется KukaDiverseObjectEnv [15], специально разработанная для обучения агентов управлению роботами-манипуляторами. Среда ставит перед исследователями задачу, способствующую испытанию и совершенствованию алгоритмов RL.

Основная задача в среде состоит в обучении робота-манипулятора корректно захватывать и поднимать объекты различной сложной формы из контейнера, опираясь лишь на изображения с виртуальной камеры. При каждом шаге симуляции манипулятор автоматически движется по оси z вниз, а нейросетевой агент принимает решения относительно перемещения по осям x и y, а также угла поворота захвата. Награда в среде является бинарной и выдается агенту, только если объект был успешно захвачен и поднят выше определенного уровня к концу эпизода. Пространство действий представлено тремя параметрами с размерностью от –1 до 1: перемещение по осям x и y, а также управление углом захвата. Входные данные включают в себя трёхканальные изображения состояния среды, полученные с виртуальной камеры размерностью (48, 48, 3), пример которых показан на рис. 1.

Рис. 1. Пример изображений с виртуальной камеры

Сложность среды заключается в редкой бинарной награде, необходимости обработки графических входных данных и общей сложности задачи, что делает её подходящей для исследования и тестирования алгоритмов RL. Все это позволяет оценивать и улучшать эффективность агентов в условиях, приближенных к реальным сценариям манипулирования.

III. Теория

DQN [13] является *off-policy* [16] алгоритмом и использует нейронные сети для приближения Q-функции, оценивающей ожидаемое суммарное дисконтированное вознаграждение для каждой пары состояния и действия в среде. Это позволяет агенту принимать оптимальные решения на основе текущего состояния.

В данной работе алгоритм DQN использует входные данные в градациях серого, что, несмотря на усложнение процесса обучения, существенно снижает затраты вычислительных ресурсов. Визуализация изображений в трех промежуточных сверточных слоях приведена на рис. 2. Для извлечения визуальных признаков, на основе которых агент принимает решения, используются сверточные нейронные сети (CNN) [17] с функцией активации ReLU. Эти слои позволяют обрабатывать и интерпретировать визуальные данные, поступающие с виртуальной камеры робота, что критично для задач, связанных с компьютерным зрением.

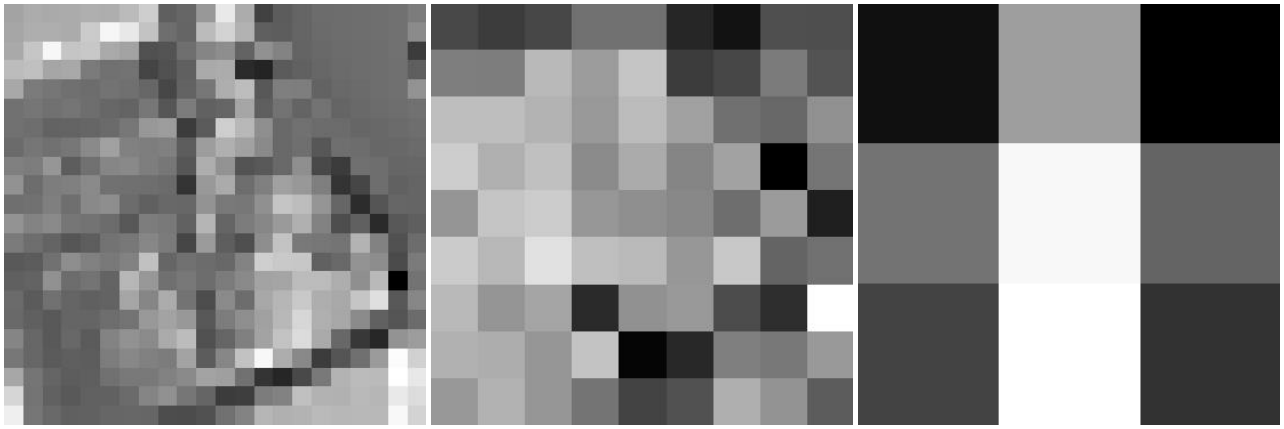


Рис. 2. Визуализация промежуточных сверточных слоев

Для обучения DQN используется метод воспроизведения опыта (*Experience Replay*), который сохраняет последовательности переходов $et=(S_t,A_t,R_t,S_{t+1})$ (где e – эпизод обучения, S – множество состояний среды, A – множество действий, R – множество наград) в реплей буфер $Dt=\{e_1,\ldots,e_t\}$. Это позволяет повторно использовать эти данные позже. Случайная выборка из памяти приводит к декорреляции переходов, из которых формируется пакет данных (batch), что позволяет несколько стабилизировать и улучшить процесс обучения.

Дополнительно в алгоритме используется периодическое обновление целевой сети: Q-функция оптимизируется по целевым значениям, обновляемым с заданным интервалом. Q-сеть клонируется и «замораживается» как целевая сеть каждые C шагов (где C – гиперпараметр). Эта модификация предотвращает краткосрочные колебания в значениях Q-функции, делая процесс обучения более стабильным [18].

Топология нейронной сети, использующейся в DQN, представлена на рис. 3.

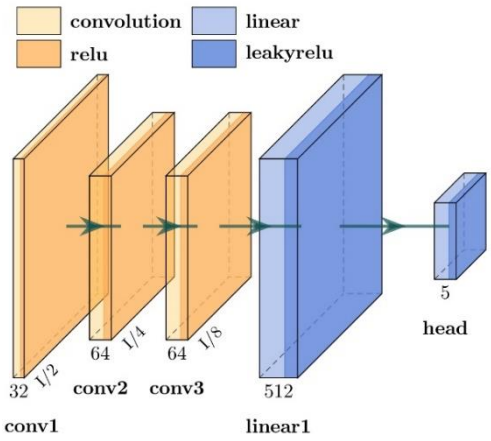


Рис. 3. Топология нейронной сети агента DQN

PPO [14] представляет собой метод на основе политики (*on-policy*) [16], который применяет архитектуру *actor-critic* [19] для обучения. В данном алгоритме используются трехканальные изображения, что позволяет агенту не только различать объекты по цвету и форме, но и принимать более обоснованные решения на основе визуальных данных. Но вместе с этим увеличивается и объем обрабатываемых данных, что, соответственно, влияет на вычислительную сложность.

PPO обеспечивает стабильность и эффективность обучения благодаря двум ключевым идеям. Первая идея заключается в обновлении политики на основе отношения вероятностей между текущей и новой политиками, чтобы ограничить их изменения. Это отношение вычисляется по формуле (1) [20], что гарантирует постепенные и контролируемые обновления политики, предотвращая резкие изменения, которые могут дестабилизировать процесс обучения. Вторая идея – использование значений "преимуществ" (advantages), которые выравнивают обновления политики, обеспечивая более сбалансированное и стабильное обучение в отличие от абсолютных вероятностей.

$$r(\theta) = \frac{\pi_\theta(a \mid s)}{\pi_{\theta_{old}}(a \mid s)}, \quad (1)$$

где π – политика; θ – параметры политики; θ_{old} вектор параметров политики до обновления; a – действие в среде; s – состояние среды, $\pi(a|s)$ нейросеть определяющая вероятность выбора действия a в состоянии s.

Без ограничений на изменения политик максимизация целевой функции может привести к нестабильности с чрезмерными обновлениями параметров и большими коэффициентами изменения политики. PPO решает эту проблему, накладывая ограничение с помощью функции clip(), которая удерживает значения в интервале от 1 – ε до 1 + ε. Целевая функция с этими изменениями вычисляется по формуле (2) [19]:

$$J^{CLIP}(\theta) = E[min(r(\theta)\hat{A}_{\theta_{old}}(s,a), clip(r(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_{\theta_{old}}(s,a)], \quad (1)$$

где $A(s,a)$ – функция преимуществ; ε – гиперпараметр.

Помимо ограниченного вознаграждения, целевая функция включает в себя функционал ошибки при оценке значений и энтропию, что стимулирует достаточное исследование среды агентом. Итоговая целевая функция определяется по формуле (3) [19]:

$$J^{CLIP}(\theta) = E[J^{CLIP}(\theta) - c_1(V_0(s) - V_{target})^2 + c_2 H(s, \pi_\theta(.))], \quad (1)$$

где c_1 и c_2 – константные гиперпараметры; $V(s)$ – ожидаемый возврат состояния s.

Вместо того чтобы обновлять политику на основе всех данных сразу, PPO использует мини-пакеты данных (mini-batch) и метод стохастического градиентного спуска (SGD) [21]. Это позволяет более эффективно использовать вычислительные ресурсы и улучшает обобщающую способность модели. Архитектура нейронной сети, используемой в PPO, представлена на рис. 4.

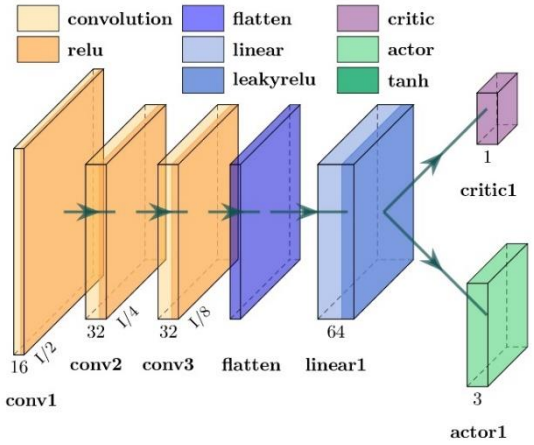


Рис. 4. Топология нейронной сети агента PPO

Модификация PPOv1 представляет собой первую адаптацию алгоритма PPO для работы в среде KukaDiverseObjectEnv. В этой версии алгоритма, помимо изменений в цикле обучения, были внесены коррективы в пространства наблюдений и действий. Размер входных изображений был увеличен с (48, 48, 3) до (84, 84, 3), что позволило повысить качество визуальных данных и улучшить распознавание объектов. Пространство действий также было расширено и теперь включает пять параметров: по два для перемещения по осям x и y соответственно, а также одно для управления углом захвата. В архитектуру нейронной сети были добавлены дополнительные скрытые слои. Параметры сверточных слоев, такие как размеры ядра и шаг, также были изменены для повышения эффективности обучения. Также была применена пакетная нормализация (BatchNorm2d) и метод инициализации весов нейронной сети Ксавье (*xavier-uniform*), что должно стабилизировать и ускорить процесс обучения агента. Структура нейронной сети PPOv1 представлена на рис. 5.

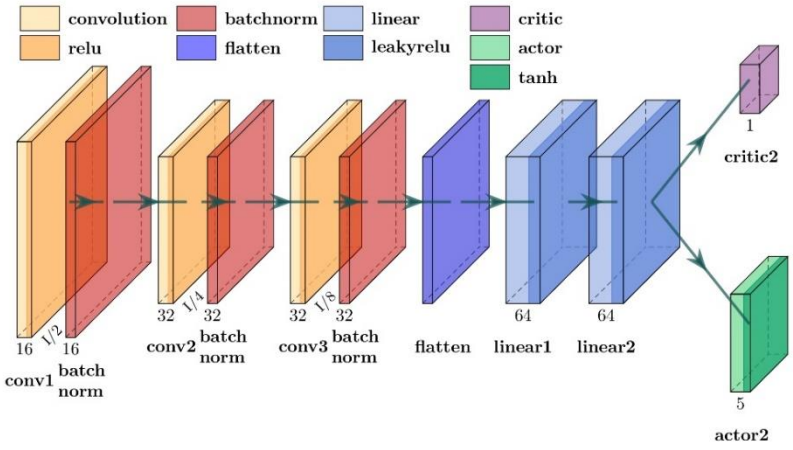


Рис. 5. Топология нейронной сети агента PPOv1

В модификации PPOv2 были добавлены механизмы синхронного параллельного выполнения действий, что позволяет ускорить процесс сбора данных о траекториях агента в среде. Этот подход позволяет эффективнее использовать вычислительные ресурсы и значительно уменьшает время обучения. Для реализации параллельных сред использовалась библиотека «multiprocessing» [22]. Это позволило одновременно запускать до 20 параллельных сред, что соответствует количеству потоков процессора на используемой рабочей станции.

PPOv3 продолжает развитие предыдущей версии, вводя оптимизации гиперпараметров благодаря модульной системе обучающего цикла. Среди ключевых изменений можно отметить уменьшение размера пакета данных и увеличение скорости обучения, что сокращает время на обучение и ускоряет обновление весов модели. Также была увеличена глубина сверточных слоев, включая общие слои (*shared layers*) и слои actor-critic, что улучшает способность модели к извлечению сложных связей признаков из входных изображений. Архитектура нейронной сети PPOv3 представлена на рис. 6.

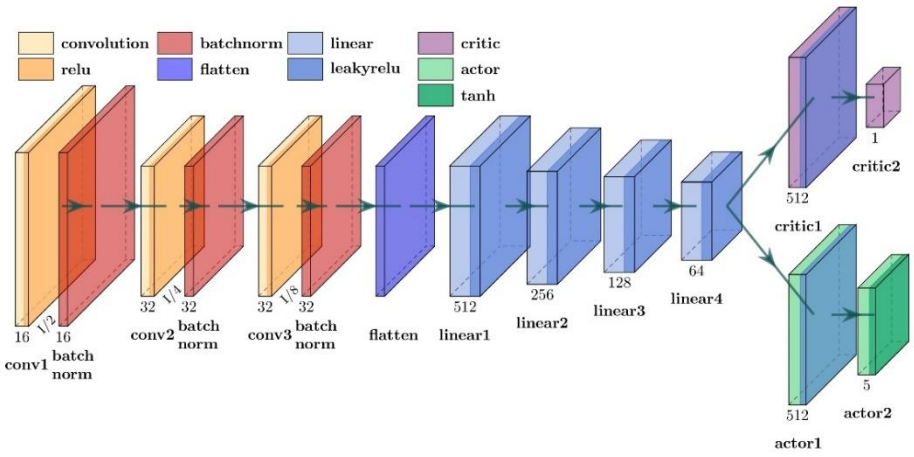


Рис. 6. Топология нейронной сети агента PPOv3

Последняя модификация, PPOv4, включает в себя механизм внимания (*attention*) [23], который позволяет модели фокусироваться на различных частях входных данных, оценивая их значимость для текущей задачи. В контексте управления роботом-манипулятором это позволяет агенту выделять ключевые признаки объектов и их окружения, что критически важно для выполнения точных манипуляций. Механизм внимания призван улучшить способность модели к обработке визуальной информации, обеспечивая более надежное и точное выполнение задач. Структура нейронной сети PPOv4 представлена на рис. 7.

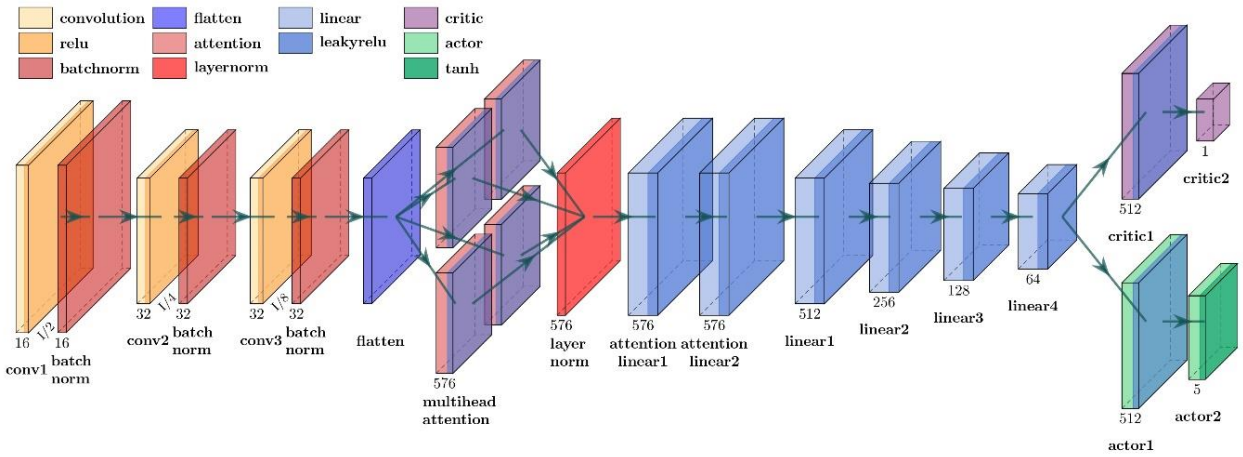
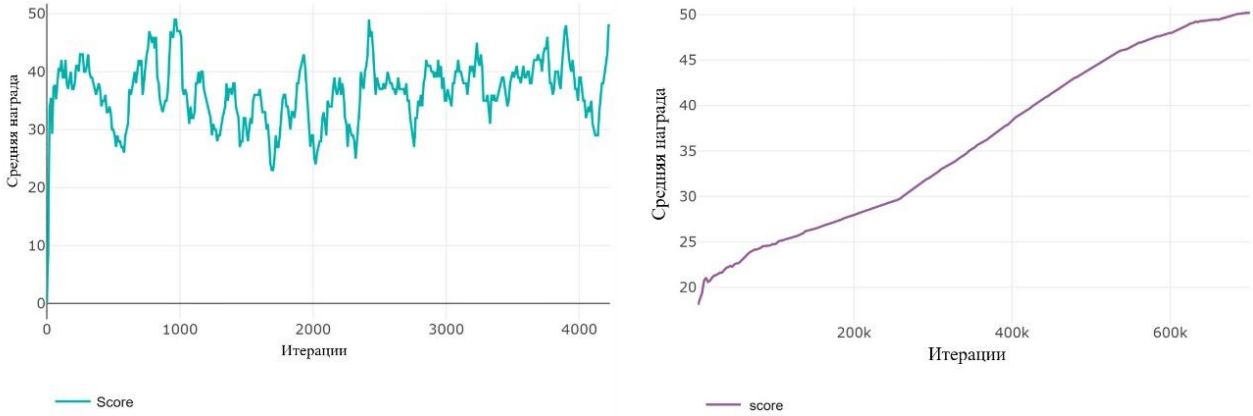


Рис. 7. Топология нейронной сети агента PPOv4

IV. Результаты экспериментов

В процессе обучения агентов в среде KukaDiverseObjectEnv была проведена серия экспериментов для оценки их производительности. Средние значения наград за 100 эпизодов обучения представлены на рис. 8–10. Для уменьшения времени обучения и обеспечения репрезентативности полученных результатов критерием ранней остановки для всех экспериментов было установлено достижение средней награды в 50 за последние 100 эпизодов, что эквивалентно 50% точности. Это позволяет существенно сократить время обучения и избежать переобучения, которое часто наблюдается в задачах с высоким уровнем сложности среды.



(а) (б)

Рис. 8. Средняя награда агента DQN (а), средняя награда агента PPO (б)

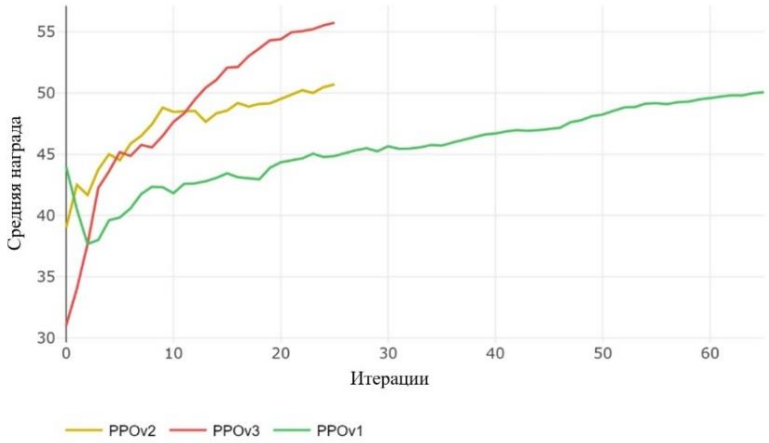


Рис. 9. Средняя награда агентов PPOv1, PPOv2, PPOv3

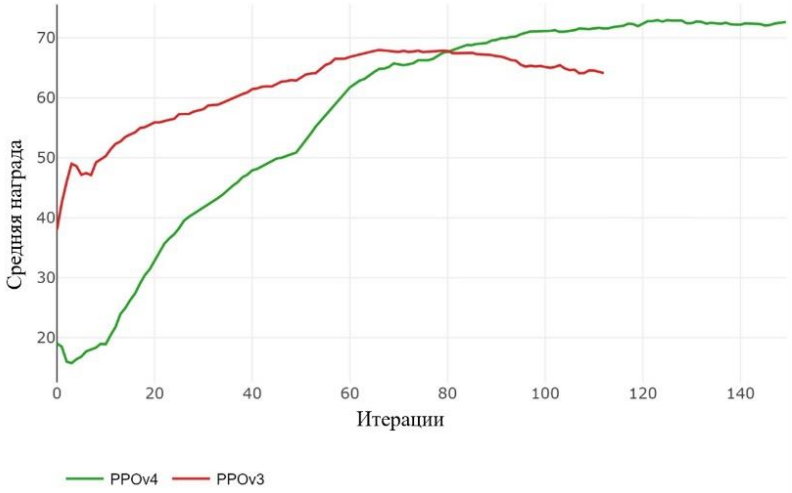


Рис. 10. Средняя награда агентов PPOv3 и PPOv4

Помимо графического представления средней награды, были зафиксированы ключевые показатели, такие как конечная средняя награда, количество шагов обучения и общее время обучения, см. Табл. I.

ТАБЛИЦА 1.
СРАВНЕНИЕ МЕТРИК ОБУЧЕНИЯ ВСЕХ ПОЛУЧЕННЫХ АГЕНТОВ

Агент	Средняя награда	Шаги обучения	Время обучения (ч:мм:сс)
DQN	51.000	4239	2:06:54
PPO	50.202	696320	5:12:04
PPOv1	50.690	65 (66560)	1:52:07
PPOv2	50.061	26 (26624)	0:35:12
PPOv3	50.067	13 (13312)	0:09:34
PPOv4	50.229	47 (48128)	0:32:01

Для проверки точности обученных агентов было проведено тестирование на 1000 тестовых эпизодах в виртуальной среде. Результаты тестирования представлены в Табл. II.

ТАБЛИЦА 2
ТЕСТИРОВАНИЕ ОБУЧЕННЫХ АГЕНТОВ

Агент	DQN	PPO	PPOv1	PPOv2	PPOv3	PPOv4
	%					
Положительный эпизод	47.8 ± 2*2.13	49.7 ± 2*1.97	50.2 ± 2*1.83	49.1 ± 2*1.91	52.4 ± 2*1.78	50.7 ± 2*1.84
Отрицательный эпизод	52.2 ± 2*2.13	50.3 ± 2*1.97	49.8 ± 2*1.83	50.9 ± 2*1.91	47.6 ± 2*1.78	49.3 ± 2*1.84

Как видно из Табл. II, все агенты решают поставленную задачу с заданной точностью в окрестности 50 % за 1000 тестовых эпизодов среды.

V. Обсуждение результатов

Анализ данных, представленных на рис. 8–10 и в таблицах I-II, показывает, что базовые версии алгоритмов DQN и PPO достигают заданной цели, однако процесс обучения требует больших временных затрат. PPO затрачивает примерно 2.5 раза большее количество времени на обучение по сравнению с DQN вследствие использования трехканальных входных данных. Введение модифицированных версий алгоритмов сократило время обучения и количество шагов в среде, необходимых для достижения поставленного результата. Модификация PPOv4, включающая механизм внимания добавленный после сверточных слоев способствовал улучшению стабильности процесса обучения. Но, несмотря на это, не оказал влияния на конечную точность модели. Заметим, что на ранних этапах обучения, модели с механизмом внимания демонстрировали более низкие значения средней награды, однако впоследствии их производительность сравнивалась с PPOv3.

VI. Выводы и заключение

Результаты работы имплементаций стандартных версий алгоритмов DQN и PPO подтверждают их способность решать поставленные задачи, однако они требовательны к вычислительным ресурсам. В отличие от них, последующие модификации алгоритма PPO сокращают время, необходимое для обучения агентов. Каждая из этих модификаций исследует применение различных эвристик: в PPOv1 проводятся изменения в архитектуре нейронной сети, в PPOv2 реализуется параллелизация сред для ускорения процесса сбора траекторий, в PPOv3 внедряются дополнительные оптимизации, а в PPOv4 добавляется механизм внимания для повышения эффективности обучения. Все разработанные агенты, включая оригинальные версии DQN и PPO, а также их модификации v1–v4, демонстрируют способность решать поставленную задачу с заданной точностью около 50 % за 1000 тестовых эпизодов.

Настоящее исследование продолжает развиваться и предполагает дальнейшие эксперименты с модификациями алгоритмов RL, направленными на улучшение процесса обучения и повышение общей эффективности агентов. Дальнейшие исследования могут быть также сосредоточены на интеграции современных методов компьютерного зрения, которые позволят улучшить качество входных данных, используемых агентами для принятия решений. Это включает в себя использование более сложных архитектур нейронных сетей, внедрение дополнительных слоев обработки изображений и применение новых подходов к обучению. Логичным шагом для будущих исследований станет замена текущей среды для проведения симуляций более комплексных задач, что позволит провести более глубокий анализ работы алгоритмов в различных условиях. В перспективе может быть разработана собственная виртуальная среда, что откроет новые возможности для исследований в области обучения с подкреплением.

Список литературы

1. Arulkumaran K., Deisenroth M.P., Brundage M., Bharath A.A. Deep reinforcement learning: a brief survey. *IEEE Signal Processing Magazine*, 2017, vol. 34, no. 6, pp. 26–38. DOI: https://doi.org/10.1109/MSP.2017.2743240
2. Hao-nan Wang, Ning Liu, Yi-yun Zhang, Da-wei Feng, Feng Huang, Dong-sheng Li, Yi-ming Zhang. Deep reinforcement learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 2020, vol. 21, no. 12, pp. 1726–1744. DOI: https://doi.org/10.1631/FITEE.1900533
3. Wang X., Wang S., Liang X., Zhao D., Huang J., Xu X., Dai B., Miao Q. Deep reinforcement learning: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, vol. 35, Iss. 4, pp. 5064–5078. DOI: 10.1109/TNNLS.2022.3207346
4. *Bullet Real-Time Physics Simulation*. Available at: https://pybullet.org/ (accessed: 23 June 2024).
5. *Gazebo*. Available at: https://gazebosim.org/ (accessed: 23 June 2024).
6. Kovač L., Farkaš I. Safe reinforcement learning in a simulated robotic arm. *International Conference on Artificial Neural Networks*. Cham, Springer Nature Switzerland, 2023. pp. 585–589.
7. Aumjaud P., McAuliffe D., Rodríguez-Lera F.J., Cardiff Ph. Reinforcement learning experiments and benchmark for solving robotic reaching tasks. *Advances in Physical Agents II. WAF 2020. Advances in Intelligent Systems and Computing*. Cham, Springer International Publ., 2021, vol. 1285, pp. 318–331. DOI: https://doi.org/10.1007/978-3-030-62579-5_22
8. Kumar S., Sampson H., Behera A. *Benchmarking deep reinforcement learning algorithms for vision-based robotics*. DOI: https://doi.org/10.48550/arXiv.2201.04224 Available at: https://arxiv.org/abs/2201.04224 (accessed: 23 June 2024).
9. Chen H. Robotic manipulation with reinforcement learning, state representation learning, and imitation learning (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 18, pp. 15769–15770.
10. Luo S., Kasaei H., Schomaker L. Accelerating reinforcement learning for reaching using continuous curriculum learning. *Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, Article 9207427. DOI: https://doi.org/10.1109/IJCNN48605.2020.9207427
11. Zeng A., Song Sh., Lee J., Rodriguez A., Funkhouser Th. TossingBot: learning to throw arbitrary objects with residual physics. *IEEE Transactions on Robotics*, 2020, vol. 36, no. 4, pp. 1307–1319.
12. Zeng A., Florence P., Tompson J., Welker S., Chien J., Attarian M., Armstrong T., Krasin I., Duong D., Sindhwani V., Lee J. Transporter networks: rearranging the visual world for robotic manipulation. *4th Conference on Robot Learning (CoRL 2020)*. Cambridge MA, USA, 2020. pp. 726–747. Available at: https://transporternets.github.io/ (accessed: 23 June 2024).
13. Mnih V., Kavukcuoglu K., Silver D., Rusu A.A., Veness J., Bellemare M.G., Graves A., Riedmiller M., Fidjeland A.K., Ostrovski G., Petersen S., Beattie Ch., Sadik A., Antonoglou I., King H., Kumaran Dh., Wierstra D., Legg Sh., Hassabis D. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, pp. 529–533. DOI: https://doi.org/10.1038/nature14236
14. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. *Proximal policy optimization algorithms*. DOI: https://doi.org/10.48550/arXiv.1707.06347. Available at: https://arxiv.org/abs/1707.06347 (accessed: 23 June 2024).
15. kuka_diverse_object_gym_env.py. *bulletphysics*. Available at:

https://github.com/bulletphysics/bullet3/blob/master/examples/pybullet/gym/pybullet_envs/bullet/kuka_diverse_object_gym_env.py (accessed: 23 June 2024).

16. OpenAi Spinning Up Algorithms. *OpenAI*. Available at: https://spinningup.openai.com/en/latest/user/algorithms.html (accessed: 23 June 2024).
17. Introduction to Convolution Neural Network. *Geeksforgeeks*. Available at: https://www.geeksforgeeks.org/introduction-convolution-neural-network/ (accessed: 23 June 2024)
18. Weng L. *A (Long) Peek into Reinforcement Learning*. Available at: https://lilianweng.github.io/posts/2018-02-19-rl-overview/#deep-q-network (accessed: 23 June 2024).
19. Actor-critic algorithm in reinforcement learning. *Geeksforgeeks*. Available at: https://www.geeksforgeeks.org/actor-critic-algorithm-in-reinforcement-learning/ (accessed: 23.06.2024)
20. Weng L. *Policy Gradient Algorithms*. Available at: https://lilianweng.github.io/posts/2018-04-08-policy-gradient/#ppo (accessed: 23 June 2024).
21. Kingma D.P., Ba J. *Adam: a method for stochastic optimization*. DOI: https://doi.org/10.48550/arXiv.1412.6980. Available at: https://arxiv.org/abs/1412.6980 (accessed: 23 June 2024).
22. multiprocessing – process-based parallelism. *Python Software Foundation*. Available at: https://docs.python.org/3.10/library/multiprocessing.html (accessed: 23 June 2024).
23. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. *Attention is all you need*. DOI: https://doi.org/10.48550/arXiv.1706.03762. Available at: https://arxiv.org/abs/1706.03762 (accessed: 23 June 2024)

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-500-508

Программная технология слежения за подвижным объектом на основе алгоритмов CSRT и SAM

Software technology for mobile object tracking based on the fusion of CSRT and SAM algorithms

Цао Инь[1], А.А. Богуславский[2]
[1]МГУ имени М. В Ломоносова, ГСП-1, Ленинские горы, д. 1, Москва, 119991, Россия
[2]ИПМ им. М.В. Келдыша РАН, Миусская пл., д.4, Москва, 125047, Россия

Cao Yin1, A. A. Boguslavsky2
[1]*Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*
[2]*Keldysh Institute of Applied Mathematics RAS, Moscow, Russia*

Аннотация. В настоящее время разработано большое количество алгоритмов прослеживания объектов на видеопоследовательностях. В данной работе приводится сравнение алгоритмов отслеживания, реализации которых представлены в библиотеке OpenCV: MOOSE, TLD, KCF и CSRT. По результатам сравнения на примере видеопоследовательностей городских сцен наилучшие результаты были получены с помощью алгоритма CSRT. Однако при его непосредственном применении возникает типичная проблема, связанная с постепенным смещением и уменьшением прослеживаемой области вокруг объекта интереса. Для решения этой проблемы предлагается усовершенствованная программная технология CSRT_SAMf, которая по сравнению с алгоритмом CSRT обладает большей точностью. В этой технологии нейросетевая модель сегментации Mobile_SAM применяется для предварительной оценки размера объекта, чтобы предотвратить уменьшение прослеживаемой области, а также вводится параметр скорости изменения для корректировки избыточного смещения прослеживаемой области. Кроме того, в представленной программной технологии с целью повышения устойчивости прослеживания интегрировано применение алгоритма вычисления точечных признаков FREAK (Fast retina keypoint). Приводятся результаты апробации представляемой программной технологии CSRT_SAMf на наборе данных UAV123. Результаты экспериментов показывают, что предложенная технология CSRT_SAMf превосходит исходный алгоритм CSRT по точности и устойчивости отслеживания.

***Ключевые слова*:** прослеживание объекта на видеопоследовательности, сегментация изображений, алгоритм CSRT, библиотека OpenCV.

Abstract: At present, a large number of objects tracking algorithms have been developed for video sequences. This paper presents a comparison of tracking algorithms whose implementations are available in the OpenCV library: MOSSE, TLD, KCF, and CSRT. Based on the comparison results using video sequences of urban scenes, the best results were obtained using the CSRT algorithm. However, its direct application presents a typical issue of gradual drift and shrinkage of the tracked region around the object of interest. To address this issue, an improved software technology called CSRT_SAMf is proposed, which offers greater accuracy compared to the CSRT algorithm. In this approach, the Mobile_SAM neural network segmentation model is used for initial estimation of the object's size to prevent the shrinkage of the tracked region, and a speed change parameter is introduced to adjust excessive drift of the tracked region. Additionally, to enhance the robustness of tracking, the proposed software technology integrates the use of the FREAK (Fast Retina Keypoint) feature point extraction algorithm. The results of testing the proposed CSRT_SAMf technology on the UAV123 dataset are presented. The experimental results demonstrate that the proposed CSRT_SAMf technology demonstrates superior performance to the original CSRT algorithm.

***Keywords*:** object tracking in video sequences, image segmentation, CSRT algorithm, OpenCV library.

I. Введение

Распознавание и прослеживание объектов на видеопоследовательности являются одними из ключевых задач компьютерного зрения. С развитием технологий глубокого обучения и машинного обучения, а также повышением производительности оборудования, системы компьютерного зрения значительно продвинулись в обработке изображений и видеоданных. Исследования в области распознавания и отслеживания объектов на видеопоследовательностях имеют значительное значение для безопасности и мониторинга, взаимодействия человека и компьютера, а также дополненной реальности.

В настоящее время ряд алгоритмов прослеживания доступны в библиотеке алгоритмов компьютерного зрения с открытым исходным кодом OpenCV. Однако, из-за необходимости большого количества исходных данных об образах объектов интереса, эти модели не всегда обеспечивают хорошие результаты, особенно в случаях, когда прослеживаемый объект находится в условиях перекрытия, затенения, изменения освещения, или когда образ объекта интереса визуально сливается с фоном. В таких ситуациях происходит часто происходит потеря прослеживаемого объекта, а также неверный захват объекта при прослеживании.

В данной работе представлена улучшенная программная технология прослеживания объекта на видеопоследовательности, основанная на алгоритме CSRT в сочетании с оптическим потоком Лукаса-Кэнада и нейросетевой моделью сегментации Mobile_SAM. В результате удалось повысить точность и устойчивость прослеживания по сравнению с применением алгоритма CSRT.

II. CSRT алгоритм

Алгоритм CSRT (Channel and Spatial Reliability Tracker) — это алгоритм прослеживания объектов, основанный на корреляционных фильтрах (Correlation Filter), предложенный в 2017 году [1]. По сравнению с другими алгоритмами прослеживания на основе корреляционных фильтров, такими как KCF (Kernelized Correlation Filter) [2], MOSSE (Minimum Output Sum of Squared Error filter) [3], CSRT демонстрирует лучшие результаты в отслеживании, особенно в сложных сценариях, в которых наблюдается частичное перекрытие объектов или их быстрое движение.

Алгоритм CSRT, получив шаблон объекта на изображении ***frame*** и (***x, y, w, h***), вычисляет фильтр для поиска этого шаблона (рис. 1). На следующем кадре с применением операции свёртки в области поиска на изображении, алгоритм находит область, наиболее схожую с шаблоном, и таким образом определяет положение объекта. Алгоритм CSRT может одновременно применять корреляционные фильтры на отдельных цветовых каналах RGB-изображения и вычислять оценку надежности каждого канала, т.н. доверительную вероятность канала (channel reliability). Эта величина влияет на распределение весов на карте отклика для каждого канала, усиливая отклик важнейших каналов и подавляя шумовые каналы.

Чтобы адаптироваться к изменениям объекта в пространстве, например, к частичному перекрытию или изменению внешнего вида в процессе отслеживания, CSRT создаёт карту пространственной доверительной вероятности (spatial reliability), которая оценивает надёжность объекта в пространстве, определяя, принадлежит ли каждый пиксель переднему плану или фону. В конечном счёте, доверительные вероятности каналов и пространства используются для взвешивания фильтра, а сам фильтр модифицируется на основе предыдущего кадра.

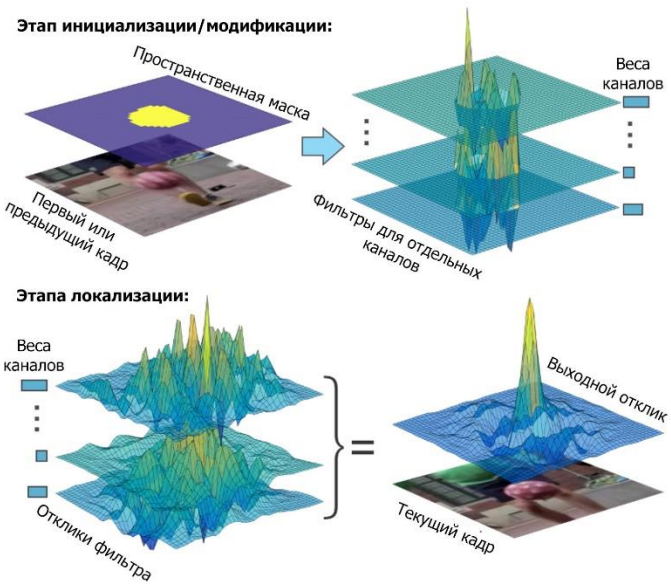


Рис. 1. Схема этапов обработки в алгоритме прослеживания CSRT [1]

Из вышесказанного можно заключить, что, если фон и объект имеют схожие характеристики, такие как цвет и текстура, то фильтр CSRT может быть подвержен помехам и ошибочно воспринимать границы объекта как меньшие. Это приводит к деградации алгоритма прослеживания – к постепенному уменьшению области изображения, внутри которой находится прослеживаемый объект (рис. 2). В случаях, когда характеристики

объекта и фона схожи, CSRT может прекратить различение границ между объектом и фоном. При значительных изменениях внешнего вида объекта, освещения, угла обзора или в случае слабой информации о краях или текстурах объекта, это может привести к тому, что оценка масштаба будет выполнена неверно, так что область прослеживаемого объекта будет постепенно уменьшаться. Модификация фильтра CSRT с обновлением, который учитывает изменение размеров объекта (по сравнению с фиксированной областью прослеживания), позволяет получить более точные результаты, хотя перечисленные проблемы полностью не устраняются [4].

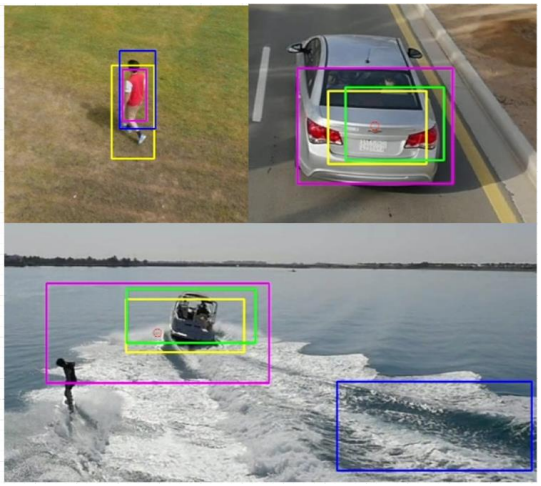


Рис.2. Примеры прослеживания объекта с помощью алгоритма CSRT (фиолетовая рамка).
(вверху слева: область сужена, внизу: область расширена).

III. Предлагаемая программная технология CSRT_SAMf

В данной работе предлагается программная технология CSRT_SAMf, которая содержит дополнительные этапы обработки результатов, полученных с помощью алгоритма CSRT. Эта технология предполагает выполнение следующих этапов после получения выходных данных алгоритма CSRT:

1) Выходные данные алгоритма CSRT подвергаются сегментации с помощью нейронной сети Mobile_SAM [5].
2) Устанавливается порог скорости изменения размеров области объекта (по умолчанию принимается равным 0.10) для предотвращения произвольных быстрых изменений размеров области.
3) Алгоритмы SURF [6] и FREAK [7] применяются для фильтрации признаковых точек, чтобы отфильтровать характерные точки. Фильтрация точек производится на основе расстояния от центра прослеживаемой области. Предпочтительными считаются точки, более близкие к центру области. Этот прием позволяет повысить устойчивость обнаружения признаковых точек объекта и в целом повысить надежность прослеживания объектов.
4) Признаковые точки FREAK в центральной части прослеживаемой области используются для расчета среднего смещения оптического потока, чтобы вычислить или скорректировать координаты и размеры прослеживаемой области (т.е., координаты ограничивающей рамки прослеживаемого объекта).

Нейронная сеть сегментации изображений SAM (Segment Anything Model) была представлена в работе [3]. Кодировщик изображений модели сегментации SAM, разработанный компанией Meta, содержит более 600 миллионов параметров. Представленные в литературе результаты применения этой нейронной сети характеризуются высокой точностью сегментации, но вычислительные затраты применительно к обработке в реальном времени на универсальных компьютерах оказываются высокими. Поэтому была выбрана модификация нейронной сети SAM – упрощенная модель MobileSAM, разработанная Чаонином Чжаном [7]. Нейронная сеть MobileSAM имеет структуру, аналогичную SAM, но SAM использует VIT-H (вариант модели Vision Transformer с 632 миллионами параметров), а MobileSAM – небольшую модель TinyViT для мобильных устройств или встроенных систем (содержит около 5,78 миллионов параметров). Схема обучения нейронной сети MobileSAM приведена на рис. 3.

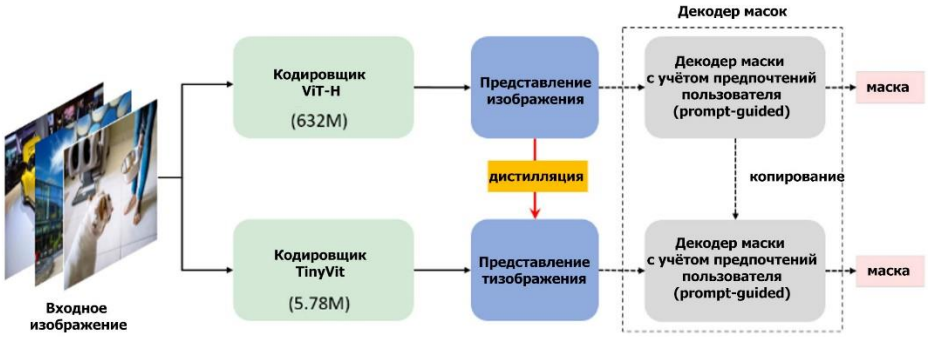


Рис. 3. Схема обучения нейронной сети MobileSAM [5].

Нейронная сеть Mobile_SAM применяется для обработки результатов алгоритма CSRT для вычисления масок – областей сегментации, построенных по данным CSRT:

$$\{masks, scores, logits\} = \text{MobileSAM.predictor.predict}(frame, points, labels, box) \quad (1)$$

В процессе прослеживания для прослеживаемой области применяется порог в форме предельной скорости изменения theta = 0.10 с целью предотвращения неограниченно больших изменений. Для быстрого расчета скорости изменения границы используются значения (x, y, w, h), полученные из масок сегментации. Порог применяется к величине площади прямоугольной прослеживаемой области w*h.

Алгоритмы SURF (Speeded-Up Robust Features) [1, 4] и FREAK [7] предназначен для обнаружения признаковых точек. Получаемые координаты признаковых точек и их дескрипторы будем обозначать следующим образом (2)-(5):

$$\{p_i^{test}\} = \text{SURF\_Detector}(I_i^{test}). \quad (2)$$

$$\{d_i^{test}\} = \text{FREAK\_Descriptor}(I_i^{test}). \quad (3)$$

$$\{p^{target}\} = \text{SURF\_Detector}(I_{target}). \quad (4)$$

$$\{d^{target}\} = \text{FREAK\_Descriptor}(I_{target}). \quad (5)$$

Здесь I_i^{test} обозначает i-й кадр видеопоследовательности TEST, I_{target} представляет изображение прослеживаемой области (т.е., изображение прослеживаемого объекта), p_i^{test}, d_i^{test} и p^{target}, d^{target}, соответственно, представляют характерные точки и дескрипторы, полученные алгоритмами SURF и FREAK.

Схема выборки при построении дескриптора в алгоритме FREAK является равномерной, как в разработанном ранее алгоритме BRISK [8], в котором выборка выполняется с равными интервалами на окружности вокруг признаковой точки. В алгоритме FREAK используется допущение, что надежность дескриптора признаковой точки можно повысить, если для его построения преимущественно использовать более близкие точки окрестности признака. Выборка точек для построения дескриптора FREAK поясняется на рис. 4. Выборка производится в нескольких областях в окрестности признака, расположенных по аналогии с восприятием изображения сетчаткой человеческого глаза. На изображении на рис .4 показана область «Fovea», предназначенная в основном для выборки вблизи центральной области, а также области «Peri» и «Para» для выборки в периферийных областях окрестности признака с пониженной точностью. По сравнению с выборкой для построения дескриптора SURF [6], FREAK позволяет снизить объем выборки более чем на 50%, при этом генерируя более надежные ключевые точки (по критерию стабильности на видеопоследовательности).

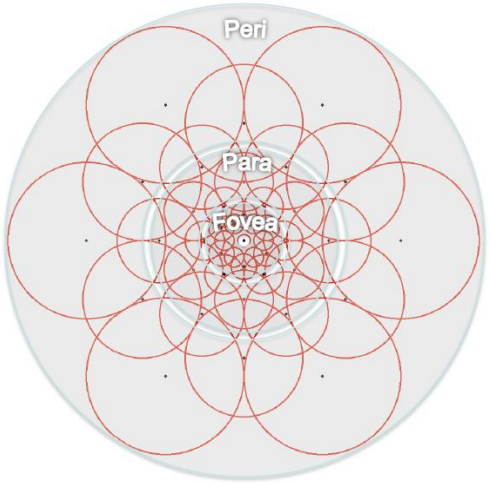


Рис. 4. Схема выборки точек в окрестности признаковой точки при построении дескриптора FREAK [7].

Для сопоставления объектов по признаковым точками используется алгоритм FLANN_Matcher [9]. Этот алгоритм позволяет получить для каждого дескриптора двух ближайших соседей matches в массиве дескрипторов:

$$matches = \text{FLANN\_Matcher}(\{d_i^{test}\}, \{d_{target}\}, k = 2)\,. \tag{6}$$

Для фильтрации сопоставленных признаковых точек применяется критерий Лоува (Lowe's ratio test) [10]:

$$good\_matches = \{matches \,|\, m.distance \; < \; ratio\_thresh * n.distance)\,. \tag{7}$$

В выражении (7) величины $m.distance$ и $n.distance$ – это расстояния до первого и второго ближайших дескрипторов из массива $matches$ (6), а $ratio\_thresh$ – это порог критерия Лоува (в проведенных экспериментах использовалось значение 0,7).

Далее выполняется вычисления оценки качества сопоставления признаков:

$$matching\_score = \frac{|good\_matches|}{\min(|\{d_i^{test}\}, \{d_{target}\}|)}\,. \tag{8}$$

В выражении $|good\_matches|$ представляет размер массива сопоставленных признаков.

В случае, если пространственный масштаб на текущей паре кадров видеопоследовательности с прослеживаемым объектов существенно не изменяется (менее 10%), а образ объекта содержит достаточное количество успешно сопоставленных признаковых точек (несколько десятков), то можно получить относительное положение прослеживаемого объекта на этих кадрах. Однако, если пространственный масштаб существенно изменяется, то количество корректно сопоставленных признаковых точек имеет тенденцию быстро сокращаться или даже сопоставление выполнить не удается. По этой причине предлагается выполнять вычисление оптического потока между парой кадров видеопоследовательности, чтобы обеспечить дополнительный способ нахождения соответствующих признаковых точек.

В алгоритме разреженного оптического потока Лукаса-Кэнада [11] выполняется оценка положения признаковых точек, обнаруженных на кадре i, на следующем кадре i+1. Межкадровое смещение изображения представляется в виде приближения:

$$I(\mathrm{x}_i + \delta \mathrm{x}_i, y_i + \delta y_i, t + \Delta t) \approx I(\mathrm{x}_i, y_i, t) + \frac{\delta I}{\delta x}\delta x_i + \frac{\delta I}{\delta y}\delta y_i + \frac{\delta I}{\delta t}\delta t\,. \tag{9}$$

Здесь $\frac{\delta I}{\delta x}, \frac{\delta I}{\delta y}, \frac{\delta I}{\delta t}$ – частные производные функции яркости изображения видеопоследовательности по координатам и времени.

При вычислении оптического потока используется допущение о постоянстве яркости $I(\mathrm{x}_i, y_i, t) \approx I(\mathrm{x}_{i+1} + \delta x_{i+1}, y_{i+1}, t + \delta t)$, в силу которого:

$$\frac{\delta I}{\delta x}\delta x+\frac{\delta I}{\delta y}\delta y+\frac{\delta I}{\delta t}\delta t=0. \quad (10)$$

В алгоритме Лукаса-Кэнада предполагается, что заданный пиксель x_i, y_i смещается в $x_{i+\Delta t}, y_{i+\Delta t}$ внутри окна W (обычно предполагаются размеры этого окна 3x3). Вектор оптического потока вычисляется посредством минимизации функции ошибки *E*:

$$E \approx \sum_{(x_i,y_i)\in W}[\frac{\delta I}{\delta x}\delta x_i+\frac{\delta I}{\delta y}\delta y_i+\frac{\delta I}{\delta t}\Delta t]^2. \quad (11)$$

Если обозначить набор признаковых точек на i-м кадре $good\_matches\_points_i$, то положение этих признаковых точек на следующем кадре, полученное с помощью алгоритма Лукаса-Кэнада, будем обозначать $lucas\_kanade\_points_{i+1}$:

$$lucas\_kanade\_points_{i+1}=\{(x_i^{'},y_i^{'})\left|(x_i^{'},y_i^{'})=(x_i+\delta x_i,y_i+\delta y_i),i=1,2,...,N\right.\}. \quad (12)$$

Чтобы исключить ошибки, применяется прием обратного прослеживания. Он заключается в повторном применении алгоритма оптического потока, когда *i*+1 используется в качестве первого кадра, а предыдущий кадр *i* – в качестве второго кадра. В результате алгоритм Лукаса-Кэнада позволяет получить оптический поток с кадра *i*+1 на кадр *i*.

$$(x_i^{''},y_i^{''})=(x_i^{'}+\delta x_i^{'},y_i^{'}+\delta y_i^{'}). \quad (13)$$

Для проверки корректности оптического потока используется разность евклидовых расстояний между положением (x_i, y_i) и положением, рассчитанным обратным оптическим потоком $(x_i^{''}, y_i^{''})$.

$$\left\|(x_i^{''},y_i^{''})-(x_i,y_i)\right\|<\varepsilon. \quad (14)$$

Если разность оказывается меньше установленного порога ϵ (использовалось значение ϵ = 1), то результат расчета оптического потока считается надежным.

Пусть $trackinglist$ – это множество признаковых точек прослеживаемого объекта. Тогда на следующем кадре множество признаковых точек будет сформировано следующим образом:

$$trackinglist_{i+1}=good\_matches\_points_{i+1}+lucas\_kanade\_points_{i+1}. \quad (15)$$

Однако, при изменении масштаба прослеживаемого объекта, например, когда объект удаляется от камеры, то количество $good\_matches\_points$ при прослеживании алгоритмом оптического потока точек будет уменьшаться. Это может привести к ошибке прекращения прослеживания. Поэтому для поддержания количества признаковых точек прослеживаемого объекта при необходимости выполняется добавление новых характерных точек. Эти точки вычисляются с помощью алгоритма SURF.

IV. Результаты экспериментов

Был проведен эксперимент по обработке видеопоследовательности, полученной с видеорегистратора в салоне автомобиля в городской среде. Длительности видеопоследовательности составляет порядка 8000 кадров. Эта видеопоследовательности несколько типичных сцен движения автомобиля в городской среде. Представленная технология прослеживания CSRT_SAMf позволяет обнаруживать на большей части кадров видеопоследовательности (более 90%).

Также был проведен эксперимент с использованием видеопоследовательности БПЛА на наборе данных UAV123 [12]. Для оценки качества прослеживания использовались критерии accuracy, precision и OPE (16).

В выражениях (16) используются также следующие общепринятые обозначения: TP (true positive) – истинно положительный результат обработки, FP (false positive) – ложноположительный, TN (true negative) – истинно отрицательный, FN (false negative) — ложноотрицательный.

Величина IoU (отношение пересечения и объединения обнаруженной и истинной части образа объекта) используется для оценки точности и достоверности результатов, величина OPE (one pass evaluation) используется для оценки производительности алгоритмов прослеживания на основе однократной обработки видеопоследовательности.

$$IoU = \frac{true\_area \cap detect\_area}{true\_area \cup detect\_area}$$
$$Precision = TP \ / \ (TP \ + \ FP)$$
$$Accuracy = \frac{(TP \ + \ TN)}{All} \quad . \qquad (16)$$
$$\text{OPE} = \frac{sum(IoU > threshold)}{sum(IoU)}$$

Результаты, полученные при использовании порога IoU (Intersection over union) = 0.6, показаны на рис. 5 и 6, при IoU = 0.8 – на рис. 7.

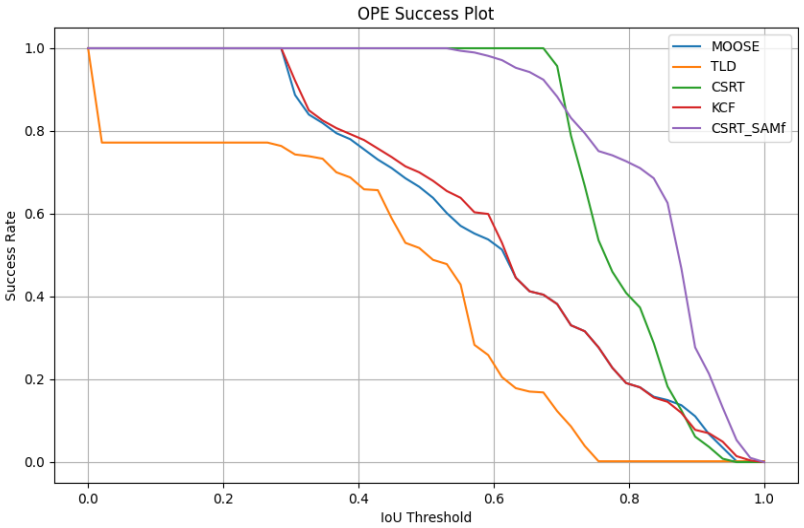


Рис. 5. График оценки OPE для различных алгоритмов прослеживания в зависимости от порога IoU [11].

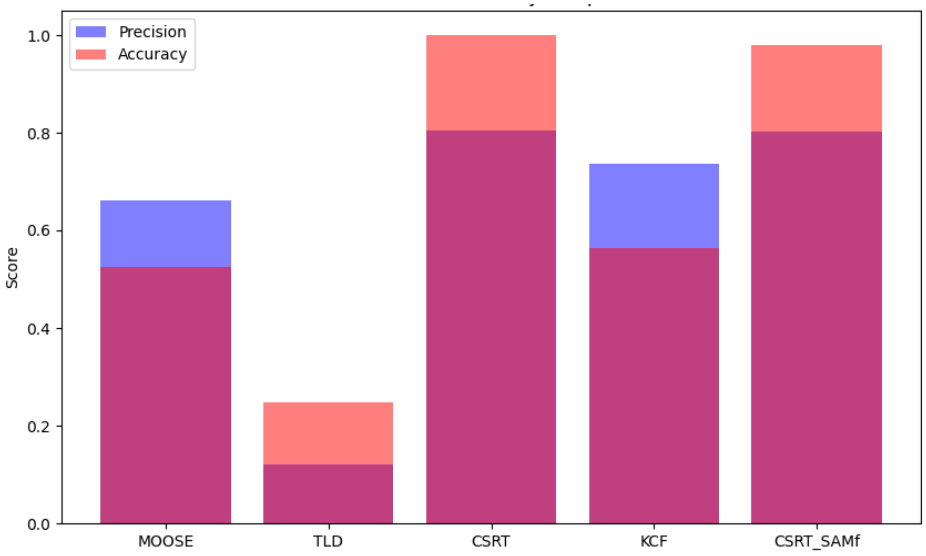


Рис. 6. Сравнение оценок accuracy/precision, порог IoU = 0.6.

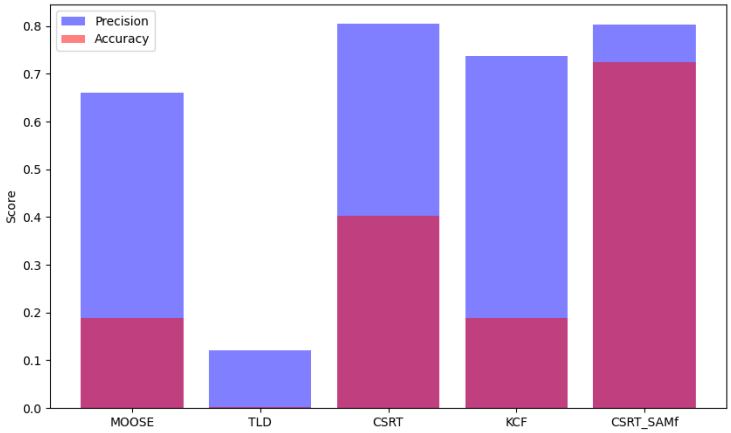


Рис. 7. Сравнение оценок accuracy/precision, порог IoU = 0.8.

В проведенных экспериментах было установлено, что при повышении порога по коэффициенту IoU для точности слежения за объектами, значения ACC алгоритмов, выбранных для сравнения, начинают уменьшаться, но предложенный метод CSRT_SAMf существенно не снизился и составляет более 70% выше, чем исходный алгоритм CSRT (рис. 8, 9).

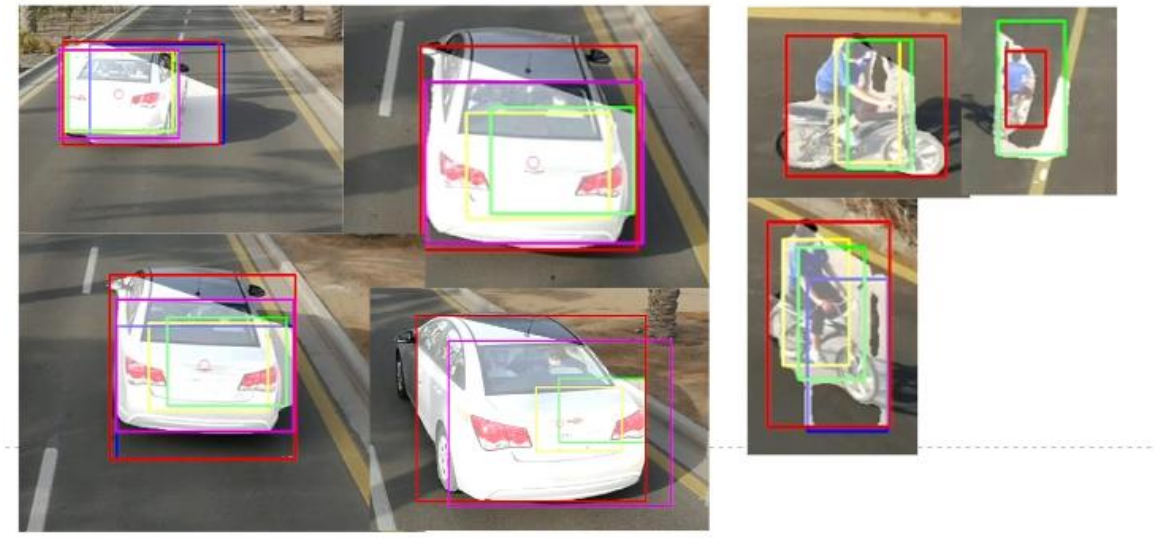


Рис. 8. Результаты, полученные различными алгоритмами прослеживания: MOOSE (зеленая рамка), KCF желтая рамка), CSRT (фиолетовая), TLD (синяя), предлагаемый CSRT_SAMf (красная).

Рис. 9. Результаты прослеживания объекта с помощью предлагаемой технологии CSRT_SAMf

V. Обсуждение результатов и заключение.

Результаты экспериментов на видеопоследовательностях городских сцен показали, что предложенная технология прослеживания CSRT_SAMf позволяет улучшить результаты алгоритма CSRT по точности и устойчивости отслеживания объектов. В процессе прослеживания применение упрощенной нейросетевой модели MobileSAM позволяет выделять маску прослеживаемого объекта.

В целом, использование традиционных алгоритмов компьютерного зрения в сочетании с современными моделями нейронных сетей может повысить точность и надежность прослеживания объектов, особенно, применительно к задачам обработки в реальном времени. В проведенных экспериментах использовалась реализация на языке Python 3.11 и библиотека OpenCV 4.6.0. Скорость обработки цветных изображений размерами 1600x1200 пикселов составляет 10 кадров в секунду на ноутбуке с процессором Intel i7-11800H и видеокартой RTX 3060.

В рамках дальнейших работ будет проводиться оптимизация программной реализации в расчете на применение в режиме реального времени на универсальных компьютерах. Также планируется исследование способов интеграции других легковесных нейросетевых моделей глубокого обучения для решения отдельных подзадач в процессе прослеживания объектов на видеопоследовательности.

Благодарности

Работа была выполнена при поддержке гранта Китайского стипендиального совета №202108090230.

Список литературы

1. Lukezic A. et al. Discriminative correlation filter with channel and spatial reliability //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2017. – С. 6309-6318. Baker S., Matthews I. Lucas-kanade 20 years on: A unifying framework //International journal of computer vision. – 2004. – Т. 56. – С. 221-255.
2. Henriques J. F. et al. High-speed tracking with kernelized correlation filters //IEEE transactions on pattern analysis and machine intelligence. – 2014. – Т. 37. – №. 3. – С. 583-596.Oyallon E., Rabin J. An analysis of the SURF method //Image Processing On Line. – 2015. – Т. 5. – С. 176-218.
3. Bolme D. S. et al. Visual object tracking using adaptive correlation filters //2010 IEEE computer society conference on computer vision and pattern recognition. – IEEE, 2010. – С. 2544-2550. Shafiee M. J. et al. Fast YOLO: A fast you only look once system for real-time embedded object detection in video //arXiv preprint arXiv:1709.05943. – 2017.
4. Portalés C. et al. Mixed reality annotation of robotic-assisted surgery videos with real-time tracking and stereo matching //Computers & Graphics. – 2023. – Т. 110. – С. 125-140.
5. Zhang C. et al. Faster segment anything: Towards lightweight SAM for mobile applications //arXiv preprint arXiv:2306.14289. – 2023.
6. Bay H. et al. Speeded-up robust features (SURF) //Computer vision and image understanding. – 2008. – Т. 110. – №. 3. – С. 346-359.
7. Alahi A., Ortiz R., Vandergheynst P. Freak: Fast retina keypoint //2012 IEEE conference on computer vision and pattern recognition. – Ieee, 2012. – С. 510-517.
8. Leutenegger S., Chli M., Siegwart R. Y. BRISK: Binary robust invariant scalable keypoints //2011 International conference on computer vision. – Ieee, 2011. – С. 2548-2555.
9. Muja M., Lowe D. G. Fast approximate nearest neighbors with automatic algorithm configuration //VISAPP (1). – 2009. – Т. 2. – №. 331-340. – С. 2.
10. Lowe D. G. Distinctive image features from scale-invariant keypoints //International journal of computer vision. – 2004. – Т. 60. – С. 91-110.
11. Leonida K. L., Sevilla K. V., Manlises C. O. A Motion-Based Tracking System Using the Lucas-Kanade Optical Flow Method //2022 14th International Conference on Computer and Automation Engineering (ICCAE). – IEEE, 2022. – С. 86-90.
12. Mueller M., Smith N., Ghanem B. A benchmark and simulator for UAV tracking //Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. – Springer International Publishing, 2016. – С. 445-461.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-509-518

Полноростовой сканер для 3D-печати статуэток: статический и динамический подходы
Full Body Scanner for 3D Figurine Printing: Static and Dynamic Approaches

М.А. Кривов[1], М.А. Федюков[2], А.В. Посконин[2]
[1]МГУ им. М.В. Ломоносова, Москва, Россия
[2]ООО «ТЕКСЕЛ»

M.A. Krivov[1], M.A. Fedyukov[2], A.V. Poskonin[2]
[1]Lomonosov Moscow State University
[2]TEXEL LLC

Аннотация. Датчики глубины и 3D-принтеры стали массовыми, доступными и широко распространенными устройствами. В то же время, наблюдается относительный недостаток программных решений для простого и недорогого изготовления на базе этих устройств персонализированных фигурок людей. Подобные статуэтки, обычно общей высотой 10...20 см, востребованы в индустрии развлечений как оригинальные подарки или как хорошее дополнение к классическим фотографиям. В данной работе мы рассматриваем два алгоритмических подхода, как такие фигурки можно легко изготовить с использованием свободно продаваемых камер глубины ASUS Xtion PRO и Microsoft Azure Kinect, а также 3D-принтеров Raise3D Pro2 и FlashForge Dreamer. Первый предлагаемый подход основан на идее классического алгоритма KinectFusion и предназначен для сканирования неподвижно стоящего человека (так называемый «жесткий» случай). Во втором способе сканирования за основу взято развитие этого же алгоритма, получившее название DynamicFusion, которое позволяет работать со свободно двигающимися людьми («нежесткий» сценарий). Наши реализации и доработки этих классических алгоритмов продемонстрировали, что оба подхода имеют сильные и слабые стороны, поэтому их стоит использовать независимо с учётом потребностей пользователя. Представлено аналитическое и визуальное сравнение нескольких фигурок, изготовленных с помощью рассмотренных методов.

***Ключевые слова*:** Сенсор глубины, Kinect, Восстановление поверхности, SDF, KinectFusion, DynamicFusion, 3D-печать.

Аннотация.
Nowadays, depth sensors and 3D printers have become common, affordable and widespread devices. At the same time, there are only a little software solutions for simple and inexpensive fabrication of personalized human figurines. Such statuettes with a total height of 10...20 cm are in demand in the entertainment industry as novel gifts or good complements to classic photographs. In the work, we consider two algorithmic approaches to how such figurines can be easily produced using mass-market depth sensors such as ASUS Xtion PRO and Microsoft Azure Kinect coupled with Raise3D Pro2 and FlashForge Dreamer 3D-printers. The first one is based on the ideas of the classic KinectFusion algorithm and is designed to scan a statically standing person (rigid case). The second way is inspired by the successor algorithm, called DynamicFusion, and works with freely moving people (non-rigid case). After re-implementing and enhancing these classic algorithms, we demonstrate that both approaches have strong and weak sides, so it is worth to use them independently in accordance with the user's needs. The analytical and visual comparison of several figurines produced by these methods are presented as well.

***Ключевые слова*:** Depth sensor, Kinect, Surface reconstruction, SDF, KinectFusion, DynamicFusion, 3D printing.

I. Введение

Сенсоры глубины Microsoft Kinect (далее просто Kinect) разрабатывались в первую очередь для видеоигр, однако намного большее влияние оказали на классические задачи компьютерного зрения. Массовое распространение этих недорогих, доступных и универсальных камер глубины, по сути, привело к временному разделению алгоритмов на две группы — «до появления Kinect» и «для Kinect». Случай «без Kinect» в тот период рассматривался скорее как редкость, а авторы таких немногочисленных работ (например [1]) добавляли пояснения, как лучше адаптировать их алгоритмы под новые сенсоры от Microsoft.

Наибольший импульс к развитию получила задача одновременной локализации и построения карты (Simultaneous Localization and Mapping, или SLAM), когда по набору кадров с перемещающейся камеры требуется воссоздать геометрическую форму видимых ей объектов и, одновременно с этим, отследить траекторию движения самой камеры. Популярный алгоритм для SLAM, получивший название KinectFusion [2], позволяет создавать весьма точные цифровые копии небольших объектов (книга, цветок, ботинки), просто поводя вокруг

них сенсором Kinect. Ключевым преимуществом метода стала возможность работы в реальном времени, а также заметное повышение детализации за счёт усреднения шумов по многим картам глубины.

Последующие эволюционные доработки этого алгоритма велись в независимых направлениях. В задачах робототехники было важно на порядок расширить размер сканируемой области, для которой в первоисточнике выделялось всего 64^3...512^3 вокселя. В качестве технических решений было предложено перейти на иерархический формат сетки [3] или хранить только заполненные воксели, но в хэш-таблице [4]. Другие методы рассматривали эту воксельную сетку как небольшой перемещающийся вместе с камерой буфер, а объекты, покидающие эту область (и, соответственно, сам буфер), выгружались в виде облака точек [5], сохранялись на диск [6] или просто игнорировались [7].

Ещё одна группа улучшений затрагивала качество реконструируемой поверхности. Согласно мнению многих авторов, основной проблемой KinectFusion является накопление погрешности, из-за которой этап с отслеживанием камеры со временем перестаёт работать корректно. Изящным решением оказалась замена используемого итеративного алгоритма ближайших точек (Iterative Closest Point, или ICP) на оптимизационную задачу о том, как с минимальной погрешностью поместить новую карту глубины в имеющуюся воксельную сетку [8]. Независимо было предложено использовать информацию о цветных кадрах совместно с алгоритмом ICP [9], или же просто прикрепить к сенсору Kinect качественный инерциальный датчик [10].

Альтернативным направлением развития стало решение более фундаментальной задачи — как адаптировать KinectFusion для случаев, когда сканируемые объекты перемещаются и/или меняют форму? Первым и действительно универсальным решением, которое не требовало априорной информации о сцене, стал алгоритм DynamicFusion [11]. Основная идея заключается в добавлении поля трансформаций, которое отображает воксели из текущего момента времени в начальное, и наоборот. Как следствие, воксельная сетка, описывающая сканируемую сцену, подвергается точно таким же трансформациям, как реальные объекты, попавшие в область видимости сенсора глубины. Практически все последующие улучшения этой идеи сводились к повышению стабильности поля трансформаций за счёт работы с цветными кадрами [12], учёта направленного освещения [13], более сложного оптимизационного алгоритма [14] или перехода на промышленные сенсоры с частотой 500 Гц вместо исходных 30 Гц [15].

Помимо этих универсальных подходов, подходящих для съёмки произвольных объектов, было предложено множество идей для сканированию исключительно людей. Самым простым «сканером» стоит признать вращающийся стул и сенсор Kinect, снабжённые цветной версией метода KinectFusion [16]. В других работах предлагались собственные алгоритмы, когда человек должен оставаться неподвижным [17-19], двигаться, но замирать в определённый моменты [20-22] или перемещаться как угодно [23]. Лишь некоторые из них [16, 19, 21] используют идеи KinectFusion, в то время как остальные [17, 18, 20, 22, 23] основаны на выравнивании небольшого числа карт глубины с их последующим склеиванием через преобразование Пуассона [24].

Целью настоящей работы является оценка применимости уже ставших классическими алгоритмов KinectFusion и DynamicFusion для сканирования людей и последующей 3D-печати статуэток. К нашему удивлению, этот вопрос так и не получил должного освещения в научных работах. Алгоритмы сканирования людей типа [20], с помощью которых создавались качественные статуэтки, являются уникальными и вряд ли могут быть воспроизведены без участия их авторов. В то же время, универсальные методы KinectFusion, DynamicFusion и их последующие модификации многократно применялись для сканировании людей [12-15], однако в этих работах практически не оценивалась детализация и, самое главное, стабильность сборки подобных сканограмм. Основной вклад нашей работы заключается в экспериментальном подтверждении, что оба классических подхода вполне применимы для массового производства несложных фигурок людей (см. Раздел V), однако каждый из них обладает своими особенностями и ограничениями (Табл. I).

II. Статическое сканирование

Достаточно подробное описание алгоритма KinectFusion содержится в статье-первоисточнике [2]. Также для него имеется независимая реализация в библиотеке Point Cloud Library, в которую дополнительно встроена поддержка цвета. Поэтому в данном разделе мы рассмотрим только основные этапы алгоритма и некоторые наши изменения, которые потребовалось внести для учёта специфики решаемой задачи.

Поступающая на вход карта глубины представляет собой двумерный массив, каждый ненулевой элемент которого содержит расстояние до объекта с точностью до миллиметров. К сожалению, уровень шума и ошибок оказывается достаточно высоким, что, к примеру, также было подтверждено независимым тестированием [25]. Как следствие, нулевым этапом построения сканограммы можно считать фильтрацию каждой карты глубины. В классической версии KinectFusion [2] предлагается применять билатеральный фильтр, однако имеются успешные примеры использования более сложных методов сглаживания, например, по цветным кадрам [1]. Важно отметить, что в процессе сборки попеременно участвуют обе карты глубины, т. е. исходная и сглаженная. Первая

встраивается непосредственно в воксельную сетку, что обеспечивает эффект супер-разрешения (SuperResolution, SR), а вторая нужна для решения задачи определения местоположения камеры.

В каждом датчике Kinect содержатся его персональные коэффициенты, измеренные в результате калибровки сенсора глубины (фокусное расстояние, оптический центр, дисторсия), поэтому получаемые карты глубины могут быть легко превращены в облако точек, и наоборот [8]. В ряде работ отмечалось [26], что подобные «заводские» коэффициенты недостаточно точны и, как следствие, требуется повторная калибровка, однако в нашем случае эта проблема не наблюдалась. Далее, соседние пиксели карты глубины можно рассматривать не просто как трёхмерные точки, а как вершины полигона, благодаря чему удается построить, помимо самого облака точек, ещё и полигональную сетку.

Самая первая карта глубины загружается в трёхмерный массив вокселей в соответствии с алгоритмом усеченной функции расстояния со знаком (Truncated Signed Distance Function [27], TSDF). Таким образом, каждый воксель содержит два значения — расстояние до ближайшей границы объекта (однако не превышающее порог отсечения порядка нескольких сантиметров), а также вес, который можно рассматривать как коэффициент уверенности. Чем больше карт глубины попало в конкретный воксель, тем больше становится его вес и вклад в итоговую поверхность. В отличие от классической версии KinectFusion, мы интегрируем в TSDF не просто облако точек, а соответствующую ему полигональную сетку, что позволяет варьировать размеры вокселей и разрешение сенсора практически в неограниченном диапазоне.

Для извлечения из TSDF текущей геометрической формы объекта применяется алгоритм марширующих кубов (Marching Cubes [28]). Согласно этому методу, необходимо рассмотреть все комбинации соседних восьми вокселей. Если функция расстояния меняет знак, то для таких соседних вокселей строится набор полигонов по заранее определённому шаблону. В итоге появляется возможность создать полигональную сетку, весьма точно описывающую поверхность, которая хранится в TSDF в неявном виде. Если же требуется лишь некоторая часть этой поверхности, видимая с определённого ракурса, то предпочтительным оказывается классический метод бросания лучей (Ray Casting [29]).

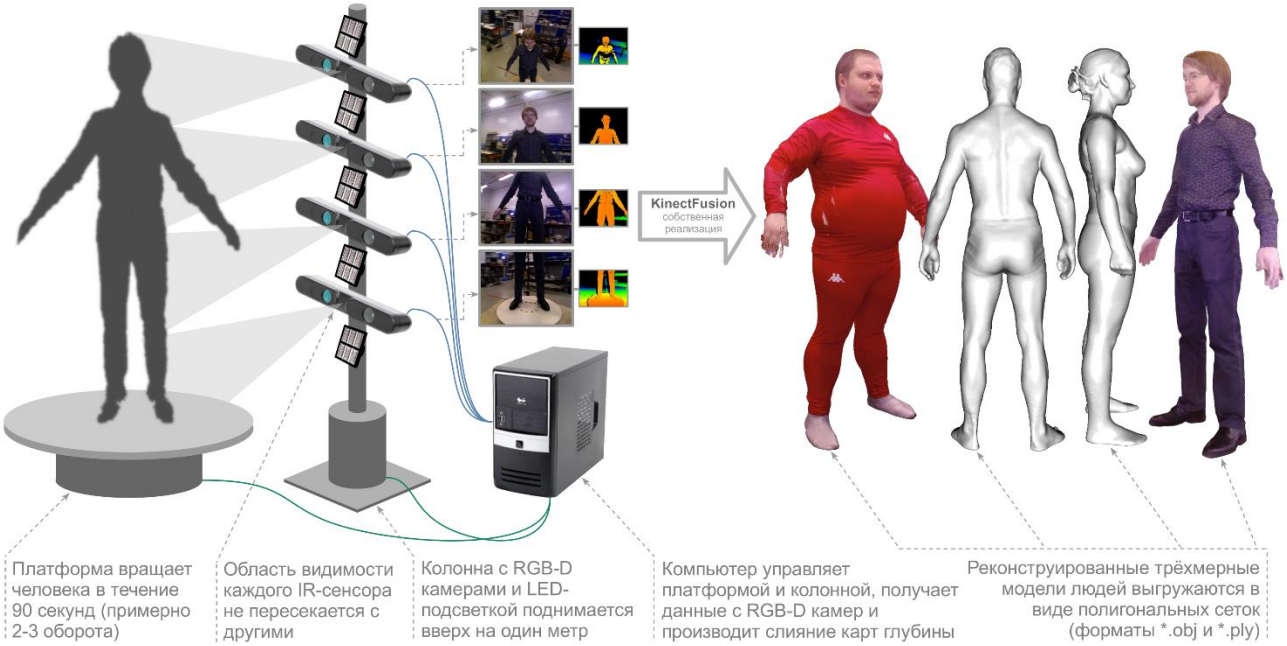


Рис. 1. Схема сканирования людей в рамках статического подхода, который основан на базе алгоритма KinectFusion. Используются четыре сенсора ASUS Xtion Pro и компьютер с графическим ускорителем GeForce 3060.

Чтобы аналогичным образом обработать и загрузить в TSDF вторую и все последующие карты глубины, необходимо отследить сдвиг и поворот соответствующих им виртуальных камер. В нашем случае физическая камера может оставаться статичной, однако подиум вращает неподвижного человека, что, по сути, является полностью аналогичной задачей, хотя и с меньшим числом степеней свободы. В классическом алгоритме KinectFusion для этого используется итерационный метод ICP [30], в котором предполагается, что смешение между камерами из двух последовательных кадров относительное небольшое.

В итоге, чем больше карт глубины удаётся получить и загрузить в TSDF (но в разумных пределах), то тем выше детализация и точность поверхности, соответствующей сканируемому объекту. На рис. 1 представлена схема, как именно мы адаптировали этот алгоритм для случая статичного сканирования. Во-первых, мы

используем сразу четыре сенсора Xtion (аналоги Kinect первого поколения), что позволяет запечатлеть сканируемого человека более крупным планом. Из-за этого возникает известная проблема наложения проецируемой инфракрасной сетки от разных сенсоров друг на друга, что приводит к искажению геометрии в участках с подобными пересечениями. В сторонних работах предлагалось включать сенсоры попеременно [19], ставить их с разных сторон человека [18] или комбинировать камеры типа ToF и IR [31], так как они основаны на разных физических принципах. В нашем случае эта проблема решена за счёт разнесения сенсоров по времени. Сначала строится TSDF по картам глубины от первого сенсора, что соответствует участку тела от ступней и примерно до пояса. Потом, зная расстояние между камерами, в воксельную сетку добавляются данные от второго сенсора, и поверхность достраивается примерно до груди. Аналогичным образом загружаются карты глубины с двух оставшихся камер Xtion, в результате чего мы получаем полностью замкнутую модель человека, от ступней до макушки.

Во-вторых, мы используем ещё одну, но уже более мелкую воксельную сетку, в которой схожим образом аккумулируется информация о цвете и его весе. Когда процесс сканирования завершается, эта сетка применяется для текстурирования выгружаемой из TSDF финальной поверхности. Стоит пояснить, что цветовую информацию вполне можно хранить непосредственно в TSDF [9], однако в нашем случае такая текстура окажется сильно смазанной. Действительно, размер вокселей неявно определяется разрешением карты глубины и границами сканируемой областью, поэтому сетка оказывается слишком крупной. Дополнительным преимуществом нашей схемы является тот факт, что этап раскрашивания можно проводить уже после слияния всех карт глубины, благодаря чему мы можем освободить TSDF и, тем самым, понизить требования к памяти графического ускорителя. Подготовленная с помощью нашего метода цветная поверхность оказывается действительно высокого качества, как это показано на рис. 1.

III. Динамическое сканирование

Динамический подход основан на тех же принципах, но допускает медленные движения человека. Подробное описание этого алгоритма, имеющего название DynamicFusion, может быть найдено в первоисточнике [11]. Основной его идеей является компенсация трансформаций, которым подвергается текущая карта глубины, после чего она может быть загружена в TSDF в полной аналогии с KinectFusion. Для этого вводится несколько сотен вершин поля трансформаций (warp-field, WF), каждая из которых содержит сдвиг и вращение, характерные для соответствующей ей окрестности пространства. Далее, для каждого вокселя сетки отбирается четыре или пять ближайших вершин WF, после чего, проведя их линейную интерполяцию, можно определить трансформацию этого вокселя, т. е. его отображение из нулевого момента времени в текущий, и наоборот.

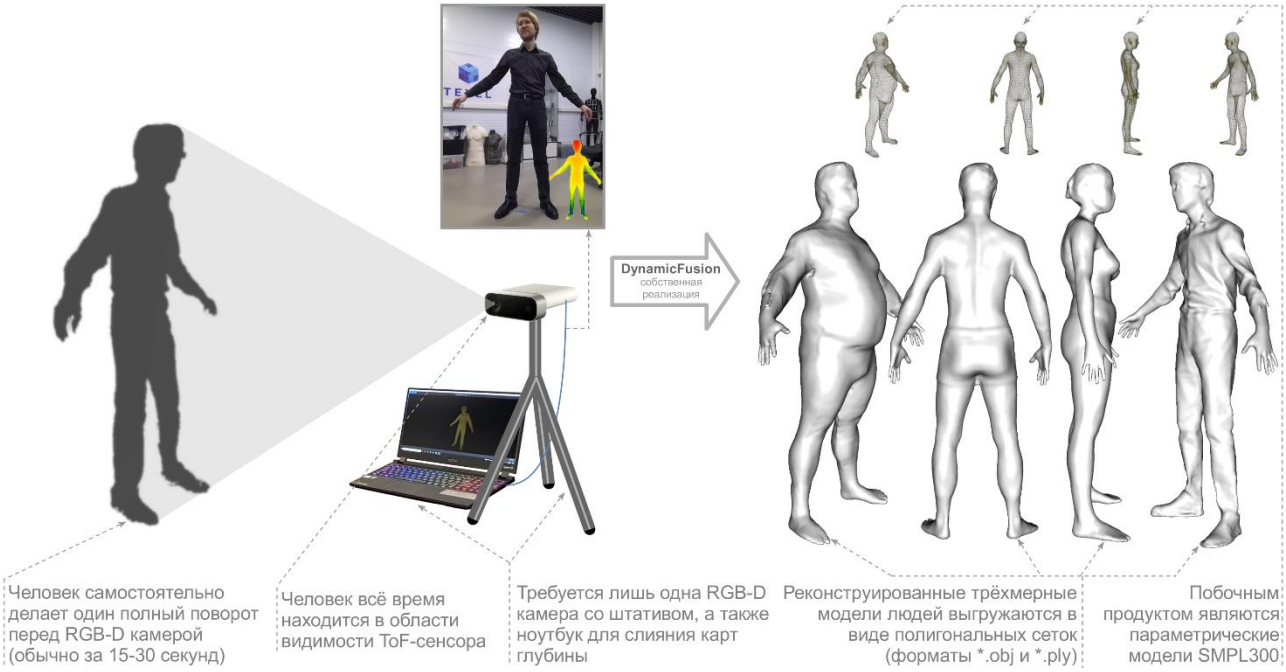


Рис. 2. Схема сканирования людей в рамках динамического подхода, который основан на базе алгоритма DynamicFusion. Используется только один сенсор Microsoft Azure Kinect и ноутбук с графическим ускорителем GeForce 3070. Побочным продуктом работы выбранной вариации алгоритма являются параметрические модели телосложения.

Основной проблемой становится поиск актуального положения вершин WF на текущем кадре. С некоторым допущением можно считать, что алгоритм DynamicFusion основан на идеях из работы [32], в которой решаемая задача заменена на оптимизацию многомерного функционала энергии, описывающего «удачность» требуемого отображения этих вершин из начального момента времени в текущий. Его аргументами являются сдвиг и поворот каждой рассматриваемой вершины (т. е. он имеет более тысячи скалярных аргументов), а общий вид состоит из нескольких слагаемых, точные формулы которых можно найти в первоисточнике [32].

В роли первой части этого функционала энергии обычно выступает сумма минимальных расстояний от каждой вершины WF до ближайшей ей точки карты глубины в рамках некоторой окрестности. При сканировании несложных объектов функционал такого вида уже позволяет построить корректное поле трансформаций, однако сами объекты со временем начинают «терять» свою форму. Например, в нашем случае это означает, что лицо человека может плавно переместиться на его затылок. Чтобы избежать проявления подобных артефактов, принято добавлять второе слагаемое, представляющее собой идейно близкую сумму, но уже не просто расстояний, а скалярных произведений нормали карты глубины и векторной разности между положением вершины и ближайшей к ней точки карты глубины. Другой независимой проблемой становится искажение пропорций объекта, когда, например, руки или ноги неестественно удлиняются или сокращаются. Артефакты этого типа удаётся исправить путём введения третьего регуляризационного слагаемого, в котором сопоставляется расстояние между ключевыми вершинами WF (например, теми, что соответствуют скелету) до и после целевой трансформации пространства.

Подходящий локальный минимум энергии находится методом Гаусса-Ньютона за небольшое и обычно фиксированное число итераций, а в качестве начального приближения проще всего использовать поле трансформаций от предыдущей карты глубины. Наиболее сложной идейной частью этого этапа является аналитическое построение Якобиана функционала энергии, а в вычислительном плане большая часть времени уходит на решение сопутствующей ему системы линейных уравнений. В работах [11, 31-33] для этого используется особая версия метода сопряжённых градиентов [34], однако в нашем случае, в первую очередь благодаря наличию намного более современных графических ускорителях, эта система легко решается прямым методом Холецкого.

Чтобы данный подход формализовать окончательно, необходимо также определить схему размещения вершин WF в начальный момент времени, т. е. по первой карте глубины. За основу мы взяли метод DoubleFusion [33], в котором используются параметрические модели семейства SMPL [35]. В результате работы этого алгоритма создаётся полигональная поверхность абстрактного мужчины или женщины, образы которых весьма точно соответствуют первой карте глубины как в плане телосложения, так и позы (рис. 2, верхняя строчка). Таким образом, у нас появляется возможность равномерно распределить вершины WF по этой синтетической поверхности, в том числе покрыв участки, которые ещё не попали в область видимости сенсора глубины (например, спина или затылок). Более того, раз нам изначально известно телосложение сканируемого человека, то мы можем добавить в функционал энергии ещё и четвёртое слагаемое. Оно будет выступать в роли штрафа для случаев, когда минимуму энергии соответствует движение, явно противоречащее анатомии человека [33].

Схема проведения подобного динамического сканирования представлена на рис. 2. Для нас важно построить параметрическую модель корректно и очень точно, поэтому человек всё время должен находится в области видимости сенсора. Как следствие, мы используем только одну, но более качественную камеру глубины Kinect Azure. С другой стороны, это сильно упрощает процесс сканирования, так как не требуется какое-либо дополнительное оборудование, достаточно только ноутбука и сенсора глубины.

IV. Подготовка и 3D-печать

Чтобы подготовить трёхмерную модель человека к печати, в общем случае необходимо выполнить три шага: обеспечить её герметичность (watertight), добавить подиум для устойчивости, а в целях экономии материала сделать полости внутри фигурки. В работе [16] для этого переработали формулы заполнения вокселей TSDF, что в пять раз уменьшило затраты на расходуемый материал. Вместо подиума они использовали алгоритмически построенный цилиндр, который располагался с учётом центра масс будущей статуэтки. В статье [20] основной проблемой оказалась герметичность фигурок, что потребовало создание отдельной болванки с помощью серии эвристик. Далее эта болванка позволяла заполнить пустые участки поверхности, изначально не попавшие в область видимости сенсора Kinect.

В нашем случае последние две задачи (добавление подиума и внутренних полостей, а также создание временных подложек) могут быть выполнены силами стороннего ПО, поставляемого в комплекте с 3D-принтером (рис. 3, правый столбец). Мы считаем такой подход более правильным, так как алгоритмы, используемые для создания этих полостей, должны учитывать тип и характеристики 3D-принтера. Оставшуюся проблему заполнения пустых участков поверхности удалось решить комбинацией двух факторов. Во-первых,

используемые нами сенсоры глубины и схема их расположения уже позволяют либо полностью избежать появления таких областей (статичный случай), либо существенно их минимизировать (динамический случай), что продемонстрировано в левом столбце рис. 3. Во-вторых, мы вычисляем интерполированные нормали для полигональной сетки, что позволяет рассматривать её вершины как облако направленных точек, после чего применить преобразование Пуассона [24] и, тем самым, заполнить пустующие области. Опытным путём было установлено, что предпочтительным вариантом является хранение таких направленных точек в дереве из восьми уровней, так как это обеспечивает хороший баланс между сглаживанием артефактов, заполнением пустот и сохранением детализации. Однако, если в исходной сканограмме изначально не наблюдались существенные артефакты, то имеет смысл использовать дерево из девяти уровней, так как с ним удаётся чуть лучше сохранить мелкие детали (пуговицы, застёжки, ремень и подобное).

Представленная в качестве примера фигурка (см. рис. 3, её высота 18 см, а вес около 110 г) напечатана на 3D-принтере Raise3D Pro2 примерно за 14 часов, в роли расходного материала выступал полилактид серого цвета с плотностью 1.25 г/см³. Единственной механической обработкой, которая была осуществлена, стало удаление временных колонн-поддержек. Какие-либо другие манипуляция, например, шлифовка, устранение заусенцев или погружение в ацетон не требовались, так как поверхность имела достаточно хорошее качество.

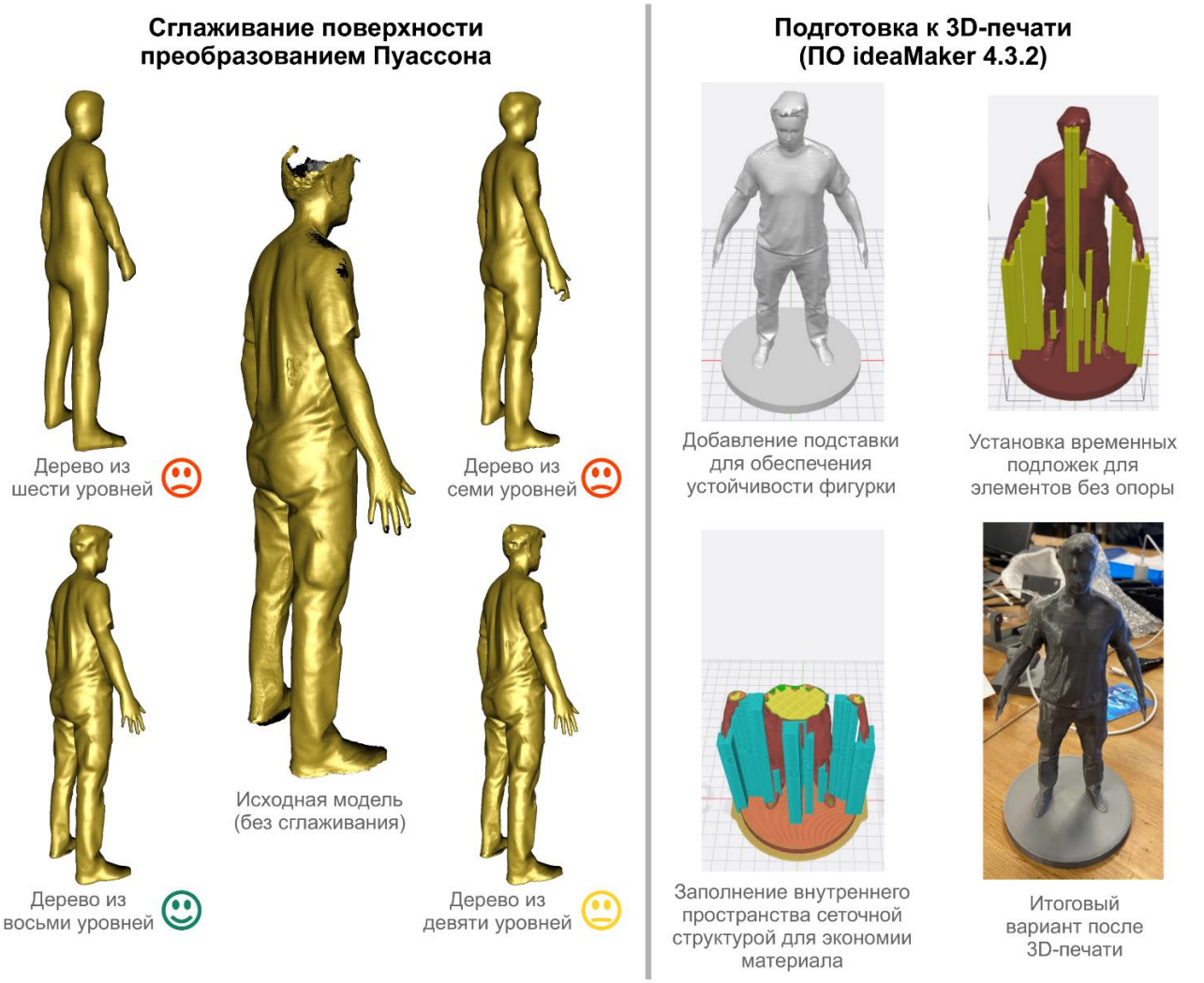


Рис. 3. Процесс изготовления фигурки человека по его реконструированной сканограмме. Сначала применяется преобразование Пуассона, благодаря которому сглаживаются артефакты поверхности и заполняются пустоты (часть нашего алгоритма). Далее с помощью ideaMaker формируется подиум-подставка, временные колонны для слоев без опоры, а также внутренний каркас (стороннее ПО).

V. Качество статуэток

В первую очередь стоит отметить, что более важным является выбор 3D-принтера, чем сенсоров глубины или схемы сканирования. На рис. 4 хорошо заметно, что ни FlashForge Dreamer, ни Raise3D Pro2 не способны в полной мере воспроизвести черты лица, причёску, тип футболки, а также передать прочие небольшие детали, которые делают статуэтку действительно уникальной и узнаваемой. Отчасти эти проблемы могут быть решены за счёт дополнительной обработки уже готовой фигурки. Например, чтобы сделать её похожей на бронзовую статуэтку, достаточно погрузить заготовку в ацетон и, тем самым, сгладить заусенцы и характерные горизонтальные полоски, после чего в несколько слоёв покрыть бронзовой краской разных оттенков (рис. 4, центральная фигурка). С другой стороны, более простым способом может оказаться переход к профессиональным цветным 3D-принтерам. В частности, на рис. 4 также представлена аналогичная гипсополимерная фигурка, распечатанная с помощью 3D Systems ProJet. Визуально она воспринимается как более детализированная и весьма близкая к результатам из [20], хотя при сканировании использовался точно такой же статический подход.

Рис. 4. Сравнение статуэток, сканограммы для которых получены статическим и динамическим методами сканирования. Как можно заметить, основным ограничением для передачи мелких деталей является разрешение 3D-печати, а не точность используемого метода. В качестве иллюстрации в центре размещена аналогичная цветная статуэтка, напечатанная с использованием более качественного 3D-принтера 3D Systems ProJet. Рядом с ней расположена схожая фигурка, но вручную покрытая бронзовой краской.

Ещё одной возможностью для существенного повышения качества статуэток является предварительная доработка соответствующих им сканограмм в 3D-редакторе. В таком случае уже не столь важно, какое изначальное качество было у трёхмерной модели человека и какой подход к сканированию был применён. Полученная сканограмма выступает в роли начального приближения, а далее производится существенное улучшение её геометрической формы и цвета. Подобный процесс ручной доработки оказывается весьма трудоёмким и требует дизайнерских навыков, однако, при наличии профессиональных 3D-принтеров уровня MCor Iris, позволяет достичь высокого уровня правдоподобности. Некоторые из таких примеров приведены на рис. 5.

Если сравнивать между собой качество поверхностей, получаемых в рамках статического и динамического подходов, то в каждой из них можно выделить свои плюсы и минусы. Человеку достаточно сложно оставаться неподвижным, обычно это выражается в небольшом колебании ладоней, которое при статическом сканировании приводит к непропорциональному утолщению и «слипанию» пальцев. Динамический подход искажает пропорции телосложения, а также не всегда корректно обрабатывает небольшие деформации (например, складки свободной одежды), в результате чего образуются своеобразные швы и вмятины. Другим существенным их отличием является уровень детализации. В первом случае каждая часть тела равноудалена от сенсоров глубины, поэтому качество поверхности оказывается равномерным. Во втором сценарии бёдра и талия расположены ближе всего к сенсору, в то время как голова и ступни отдалены. Проявление этой особенности хорошо заметно, если сравнивать, например, детализацию лица и застёжки пояса.

Однако после 3D-печати вклад этих недостатков в общее восприятие стоит признать несущественным, что достаточно хорошо заметно по рис. 4, где соответствующие области с артефактами отмечены красными стрелочками. Например, небольшие «вмятины» на майке и плече, которые остаются даже после преобразования

Пуассона (рис. 3), на серых фигурках практически незаметны. Тоже самое можно сказать и о непропорционально больших пальцах белых статуэток.

Рис. 5. Примеры более качественных цветных фигурок, которые можно получить на базе статического метода сканирования людей. В отличие от рис. 4, здесь было проведено предварительное ретуширование сканограмм в 3D-редакторе, а сами статуэтки дополнительно покрыты лаком.

VI. Выводы

Как показали проведённые сравнения, оба рассмотренных подхода (а точнее, наши собственные реализации алгоритмов KinectFusion и DynamicFusion) одинаково хорошо подходят для быстрого создания недорогих статуэток. В обоих случаях качество поверхности оказывается достаточно близким и в разы превосходит возможности бюджетных 3D-принтеров по воспроизведению мелких деталей. Поэтому выбор конкретного метода сканирования лучше осуществлять, исходя из ограничений к их применению (Табл. I). В качестве основных факторов можно назвать потребность в цветных сканограммах, наличие открытого исходного кода, общую стоимость подобной установки или одежду, в которой допускается быть сканируемым людям.

ТАБЛИЦА I
СОПОСТАВЛЕНИЕ ПОДХОДОВ ПРИ СТАТИЧЕСКОМ
И ДИНАМИЧЕСКОМ СКАНИРОВАНИИ ЛЮДЕЙ

Особенность	**Статический подход** (KinectFusion, 4 x Xtion Pro)	**Динамический подход** (DynamicFusion, 1 x Azure Kinect)
Схема работы	Человек стоит неподвижно, его вращает платформа	Человек, двигаясь свободно, совершает один оборот
Ограничения	Любая поза и одежда, но человек должен быть строго неподвижным	Рекомендуется оставаться в одной позе, одежда обтягивающая
Детализация поверхности	Средняя или высокая, равномерная	Низкая или средняя, зависит от области тела
Характерные артефакты	Утолщение рук и пальцев, «отсутствие» частей тела	Искажение пропорций тела, «вмятины» и «швы» на поверхности
Цвет	Среднего или высокого качества	Отсутствует или смазанный
Время сканирования одного человека	1-3 минуты	20-40 секунд
Открытый	Да	Нет

исходный код	(например, в PCL)	
Объём вычислений	Небольшой (GeForce 3060: 120+ кадров/сек.)	Существенный (GeForce 4090: 30-35 кадров/сек.)
Стоимость сканера	Высокая или очень высокая	Низкая или средняя

Если необходимо изготовить реалистичные фигурки очень высокого качества, то этот вопрос становится скорее инженерным, чем научным. Наибольший эффект обеспечивается за счёт предварительного ретуширования сканограммы в 3D-редакторе, механической обработки статуэтки (погружение в ацетон и покраска, либо шлифовка и покрытие лаком), а также использования профессионального цветного 3D-принтера. Наши наблюдения хорошо соотносятся с демонстрационными материалами, предоставляемыми изготовителями подобных принтеров. Например, ставшая канонической фигурка женщины с татуировками, распечатанная с помощью 3D-принтера MCor Iris и, независимо, его более современной версии MCor Iris HD, оказывается существенно разного качества. Хотя в обоих случаях производитель этих принтеров использовал одну и ту же трёхмерная модель, которая по степени детализации близка к представленным в данной работе сканограммам.

Стоит отметить, что имеются примеры успешного сканирования людей без применения сенсоров Kinect. Рассмотрение таких альтернативных подходов выходит за рамки настоящей работы, однако некоторые их них всё же стоит упомянуть. Например, можно достичь приемлемого уровня детализации, заменив карты глубины на контуры человека, полученные из цветных кадров с обычной RGB-камеры [36]. В смартфонах Apple iPhone имеется собственные датчики глубины LiDAR и TrueDepth, однако их разрешение недостаточно для стабильной работы алгоритмов семейства DynamicFusion. Эта проблема может быть решена за счёт обучения нейросети, которая повышает качество таких карт глубины до уровня сенсора Azure Kinect [37]. Кроме того, современные нейросетевые алгоритмы уже способны создать цветную сканограмму даже по одной фотографии [38]. Нельзя не вспомнить и про классическое лазерное сканирование, которое с рядом ограничений, но всё же применено для неподвижных людей и манекенов [19].

Источник финансирования. Благодарности

Часть работ была выполнена в рамках гранта Российского фонда развития информационных технологий, грантовое соглашение № 2022-550-16 от 23.12.2022.

Список литературы

1. Richardt C. et al. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos //Computer graphics forum. – Oxford, UK : Blackwell Publishing Ltd, 2012. – Vol. 31. – №. 2pt1. – P. 247-256.

2. Newcombe R. A. et al. Kinectfusion: Real-time dense surface mapping and tracking //2011 10th IEEE international symposium on mixed and augmented reality. – IEEE, 2011. – P. 127-136.

3. Zeng M. et al. A memory-efficient kinectfusion using octree //Computational Visual Media: First International Conference, CVM 2012, Beijing, China, November 8-10, 2012. Proceedings. – Springer Berlin Heidelberg, 2012. – P. 234-241.

4. Nießner M. et al. Real-time 3D reconstruction at scale using voxel hashing //ACM Transactions on Graphics (ToG). – 2013. – Vol. 32. – №. 6. – P. 1-11.

5. Whelan T. et al. Kintinuous: Spatially extended kinectfusion. Technical Report – 2012. – P. 1-10.

6. Bondarev E. et al. On photo-realistic 3D reconstruction of large-scale and arbitrary-shaped environments //2013 IEEE 10th Consumer Communications and Networking Conference (CCNC). – IEEE, 2013. – P. 621-624.

7. Roth H., Vona M. Moving Volume KinectFusion //BMVC. – 2012. – Vol. 20. – №. 2. – P. 1-11.

8. Bylow E. et al. Real-time camera tracking and 3D reconstruction using signed distance functions //Robotics: Science and systems (RSS) conference 2013. – Robotics: Science and Systems, 2013. – P. 1-9.

9. Lee S. O. et al. RGB-D fusion: Real-time robust tracking and dense mapping with RGB-D data fusion //2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. – IEEE, 2014. – P. 2749-2754.

10. Giancola S. et al. Integration of absolute orientation measurements in the kinectfusion reconstruction pipeline //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. – 2018. – P. 1486-1495.

11. Newcombe R. A., Fox D., Seitz S. M. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. – P. 343-352.

12. Innmann M. et al. Volumedeform: Real-time volumetric non-rigid reconstruction //Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. – Springer International Publishing, 2016. – P. 362-379.

13. Guo K. et al. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera //ACM Transactions on Graphics (ToG). – 2017. – Vol. 36. – №. 4. – P. 32:1-32:13.

14. Slavcheva M., Baust M., Ilic S. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – P. 2646-2655.

15. Guo K. et al. Twinfusion: High framerate non-rigid fusion through fast correspondence tracking //2018 International Conference on 3D Vision (3DV). – IEEE, 2018. – P. 596-605.

16. Sturm J. et al. CopyMe3D: Scanning and printing persons in 3D //German Conference on Pattern Recognition. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2013. – P. 405-414.

17. Tong J. et al. Scanning 3d full human bodies using kinects //IEEE transactions on visualization and computer graphics. – 2012. – Vol. 18. – №. 4. – P. 643-650.

18. Chen Y. et al. Fast capture of personalized avatar using two Kinects //Journal of Manufacturing systems. – 2014. – Vol. 33. – №. 1. – P. 233-240.

19. Lin S. et al. Fast capture of textured full-body avatar with rgb-d cameras //The Visual Computer. – 2016. – Vol. 32. – P. 681-691.

20. Li H. et al. 3D self-portraits //ACM Transactions on Graphics (TOG). – 2013. – Vol. 32. – №. 6. – P. 1-9.

21. Zhang Q. et al. Quality dynamic human body modeling using a single low-cost depth camera //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – P. 676-683.

22. Mao A. et al. Easy and Fast Reconstruction of a 3D Avatar with an RGB-D Sensor //Sensors. – 2017. – Vol. 17. – №. 5. – P. 1113.

23. Cui Y. et al. Kinectavatar: fully automatic body capture using a single kinect //Computer Vision-ACCV 2012 Workshops: ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part II 11. – Springer Berlin Heidelberg, 2013. – P. 133-147.

24. Kazhdan M., Bolitho M., Hoppe H. Poisson surface reconstruction //Proceedings of the fourth Eurographics symposium on Geometry processing. – 2006. – Vol. 7. – №. 4. – P. 1-10.

25. Tölgyessy M. et al. Evaluation of the azure kinect and its comparison to kinect v1 and kinect v2 //Sensors. – 2021. – Vol. 21. – №. 2. – P. 413.

26. Mejia I. C. P., Zell A. 3d reconstructions with kinfu using different rgbd sensors //2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS). – IEEE, 2018. – P. 13-18.

27. Curless B., Levoy M. A volumetric method for building complex models from range images //Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. – 1996. – P. 303-312.

28. Lorensen W. E., Cline H. E. Marching cubes: A high resolution 3D surface construction algorithm //Seminal graphics: pioneering efforts that shaped the field. – 1998. – P. 347-353.

29. Parker S. et al. Interactive ray tracing for isosurface rendering //Proceedings Visualization'98 (Cat. No. 98CB36276). – IEEE, 1998. – P. 233-238.

30. Besl P. J., McKay N. D. Method for registration of 3-D shapes //Sensor fusion IV: control paradigms and data structures. – Spie, 1992. – Vol. 1611. – P. 586-606.

31. Xie L. et al. SkeletonFusion: Reconstruction and tracking of human body in real-time //Optics and Lasers in Engineering. – 2018. – Vol. 110. – P. 80-88.

32. Zollhöfer M. et al. Real-time non-rigid reconstruction using an RGB-D camera //ACM Transactions on Graphics (ToG). – 2014. – Vol. 33. – №. 4. – P. 1-12.

33. Yu T. et al. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – P. 7287-7296.

34. Weber D. et al. Efficient GPU data structures and methods to solve sparse linear systems in dynamics applications //Computer graphics forum. – Oxford, UK : Blackwell Publishing Ltd, 2013. – Vol. 32. – №. 1. – P. 16-26.

35. Loper M. et al. SMPL: A skinned multi-person linear model //Seminal Graphics Papers: Pushing the Boundaries, Vol. 2 – 2023. – P. 851-866.

36. Alldieck T. et al. Video based reconstruction of 3d people models //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2018. – P. 8387-8397.

37. Shabanov A. et al. Self-supervised depth denoising using lower-and higher-quality RGB-d sensors //2020 International Conference on 3D Vision (3DV). – IEEE, 2020. – P. 743-752.

38. Saito S. et al. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization //Proceedings of the IEEE/CVF international conference on computer vision. – 2019. – P. 2304-2314.

УДК 004.352.243
DOI: 10.25206/978-5-8149-3873-2-2024-519-524

Распознавание рукописных текстов отчетов губернаторов Енисейской губернии 19 века
Handwritten text recognition of the reports of governors of the Yenisei province in the 19th century

П. А. Шерстнев, К. Д. Кожин, А. В. Пятаева
Центр искусственного интеллекта Сибирского федерального университета, Красноярск, Российская Федерация

P. A. Sherstnev, K. D. Kozhin, A.V. Pyataeva
Artificial Intelligence Center, Siberian Federal University, Krasnoyarsk, Russian Federation

Аннотация.
Текст отчетов губернаторов Енисейской губернии имеют значительные особенности, такие как дореформенный язык отчетов, устаревшие орфографические нормы, старые буквы и символы. Кроме того, фотографии отчетов могут содержать различные артефакты, связанные с особенностью письма и сканирования этих документов. Стандартные методы распознавания для распознавания текстов отчетов являются малоэффективными, поэтому авторами разработан специализированный алгоритм для распознавания текста на таких изображениях. Алгоритм включает в себя два этапа: детекция строк на изображении с помощью модели CRAFT и последующее распознавание текста с помощью модели TPS-ResNet-BiLSTM-Attn. Для обучения и тестирования моделей использованы самостоятельно созданные комбинированные наборы данных. Экспериментальные исследования подтверждают эффективность предложенных методов.

***Ключевые слова*:** распознавание текста, оцифровка документов, дореформенный русский язык, машинное обучение, исторические документы.

Аннотация. *The text of the reports of the governors of the Yenisei Governorate have significant peculiarities, such as pre-reform language of the reports, outdated spelling norms, old letters and symbols. In addition, photos of the reports may contain various artifacts related to the peculiarities of writing and scanning of these documents. Standard recognition methods for recognizing report texts are inefficient, so the authors have developed a specialized algorithm for recognizing text in such images. The algorithm includes two stages: line detection in the image using the CRAFT model and subsequent text recognition using the TPS-ResNet-BiLSTM-Attn model. Self-created combined datasets were used for training and testing the models. Experimental studies confirm the effectiveness of the proposed methods.*

***Ключевые слова*:** *text recognition, document digitization, pre-reform Russian language, machine learning, historical documents.*

I. Введение

Официальные отчеты губернаторов Енисейской губернии XVIII и XIX века представляют собой ценный исторический источник, содержащий важную информацию о городах, организациях и людях, живших в регионе. Эти документы охватывают период с момента основания губернии в 1822 году и до революции, предлагая представление о социальной, экономической и политической жизни того времени.

В Сибирском федеральном университете (СФУ) реализуется уникальный для России проект по сбору и оцифровке этих отчетов. В проекте задействованы сотрудники вуза, студенты и волонтеры, которые вручную сканируют и распознают тексты, делая их доступными для исследователей и широкой общественности. Чтобы автоматизировать и ускорить процесс расшифровки отчетов, было принято решение создать сервис для распознавания рукописного текста, ориентированный в первую очередь на тексты Енисейской губернии.

Разработанный сервис предполагает создание рабочей среды для историков, в которой они смогут не только распознавать и редактировать документы, но и создавать коллекции, делиться результатами и дополнительно обучать модели для улучшения качества распознавания. Для реализации такого сервиса потребовалось разработать модуль распознавания рукописных текстов, технические аспекты которого описаны в данной статье. Также будут рассмотрены результаты экспериментов и дальнейшие планы по развитию сервиса.

II. Официальные отчеты губернаторов Енисейской губернии

1. Характеристика отчетов

Официальные отчеты губернаторов Енисейской губернии служат важным историческим источником, охватывающим период с момента образования губернии в 1822 году до революции. Эти документы дают всестороннее представление о различных аспектах жизни региона, в том числе об административных, экономических, социальных и культурных аспектах в мельчайших подробностях [1].

Исследователи считают эти губернаторские отчеты бесценными, поскольку они дают полное представление о развитии Енисейской губернии на протяжении почти столетия. Отчеты содержат широкий спектр информации, такой как:

- демографические данные: данные о населении, его составе, религиозных убеждениях и социальном статусе;
- экономическая информация: информация о сельском хозяйстве, промышленности, торговле и государственных финансах;
- инфраструктура: состояние дорог, мостов, паромов и общественных зданий, включая этапы и тюрьмы;
- социальная сфера: статистика здравоохранения, образования, благотворительных организаций и общественного благосостояния;
- правоохранительная деятельность: статистика преступности, меры пожарной безопасности и общественная безопасность.

На рис. 1 представлен пример нескольких изображений страниц из отчетов губернаторов Енисейской губернии.

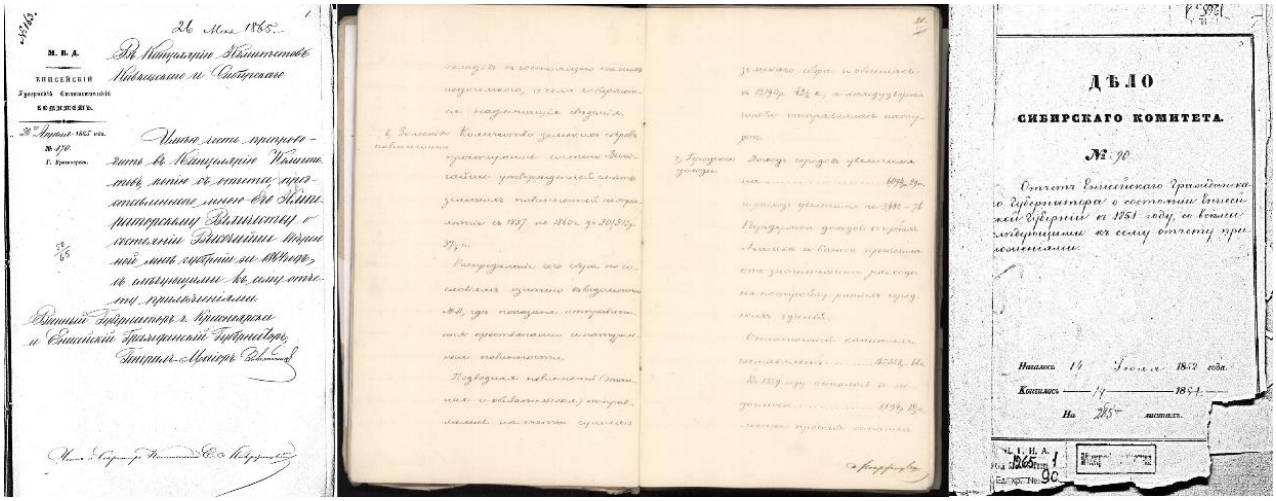


Рис. 1. Примеры страниц отчетов за разные годы

Каждый отчет, как правило, состоит из нескольких разделов, в которых подробно описывается состояние провинции за отчетный период. К основным разделам, содержащимся в этих документах, относятся:

-общая информация: подробные сведения о территории провинции, природных ресурсах и климатических условиях;

-население: статистические данные о численности населения, этническом составе, религиозной принадлежности и демографических изменениях;

-экономическое развитие: отчеты о состоянии сельского хозяйства, промышленности, ремесел и торговли, включая данные об урожайности сельскохозяйственных культур, поголовье скота и его состоянии, а также о развитии различных ремесел и коммерции;

-инфраструктура и транспорт: описание состояния дорог, мостов, почтовых маршрутов и других транспортных сетей;

-социальная сфера: информация о здравоохранении, доступности больниц и аптек, развитии образования, наличии школ и учебных заведений;

-закон и порядок: данные о преступлениях, пожарах, стихийных бедствиях и мерах, принимаемых для поддержания общественного порядка и безопасности.

Отчеты губернаторов Енисейской губернии физически хранятся в различных архивах, что создает трудности для всестороннего и систематического изучения. Большинство из этих документов не переведены в цифровую форму и доступны только в бумажном виде, что существенно затрудняет их использование современными

исследователями. Оцифровка этих отчетов, позволит сделать их доступными для более широкой аудитории и сохранить для будущих поколений.

2. *Особенности текста отчетов губернаторов Енисейской губернии*

Уникальные характеристики текста в отчетах губернаторов Енисейской губернии накладывают определенные ограничения на использование стандартных решений для распознавания рукописного текста. Во-первых, текст написан на дореформенном русском языке, который существенно отличается от современного языка [2]. Это проявляется в использовании устаревших орфографических норм, старых букв и символов, которые больше не используются в современном письме. Например, буквы "Ѣ" и "Ѳ" использовались до орфографической реформы 1918 года. Во-вторых, отчеты содержат множество специализированных терминов и фраз, характерных для административной и бюрократической лексики 19 века. Эти термины могут быть незнакомы современным системам распознавания текста, что затрудняет автоматическое понимание и обработку таких документов. Кроме того, почерк и каллиграфия той эпохи могли сильно отличаться от современных стандартов, что создавало дополнительные трудности для распознавания текста. Чернила, использовавшиеся в то время, могли растекаться, оставляя разводы на бумаге, что также затрудняло распознавание.

Учитывая эти факторы, стандартные решения для распознавания рукописного текста могут оказаться недостаточно эффективными. Они могут иметь проблемы с дореформенным русским языком и специфическими стилями почерка, что приводит к низкой точности распознавания и необходимости внесения существенных исправлений вручную. Поэтому для эффективного и точного распознавания отчетов губернаторов Енисейской губернии был разработан сервис распознавания рукописных текстов со специализированным алгоритмом, который учитывает особенности дореформенного русского языка, специализированную терминологию и почерк того периода, что значительно повышает точность и скорость распознавания.

III. Алгоритм распознавания рукописного текста

Все современные архитектуры для распознавания рукописного текста, а также доступные наборы данных, работают на уровне слов или строк [3]. Это означает, что для решения задачи недостаточно иметь только модель распознавания текста; необходим также инструмент для нахождения и выделения текста из всего изображения. В связи с этим, для успешного выполнения задачи требуется использование двух последовательных моделей машинного обучения: модели детекции строк текста и модели распознавания текста.

Для детекции строк на изображении была выбрана модель CRAFT (Character Region Awareness for Text detection) [4]. CRAFT - одна из самых современных и точных архитектур для детекции текста, позволяющая находить текстовые области на изображении, включая строки текста. Кроме того, CRAFT имеет доступные реализации в пакетах EasyOCR [5] и CRAFT-pytorch [6] и доступные предобученные модели.

Последовательность шагов алгоритма распознавания рукописного текста показана на рис. 2. Для распознавания текста выбрана модель TPS-ResNet-BiLSTM-Attn, которая сочетает четыре основные компонента:

1. TPS: это метод нормализации изображений, который помогает исправить геометрические искажения текста, такие как искривления и повороты;
2. ResNet: сверточная нейронная сеть, извлекающая высокоуровневые признаки из изображения;
3. BiLSTM: двунаправленная рекуррентная нейронная сеть, учитывающая как предшествующий, так и последующий контексты в последовательности данных;
4. Attention: механизм, позволяющий модели сосредоточиться на наиболее важных частях последовательности при генерации текста.

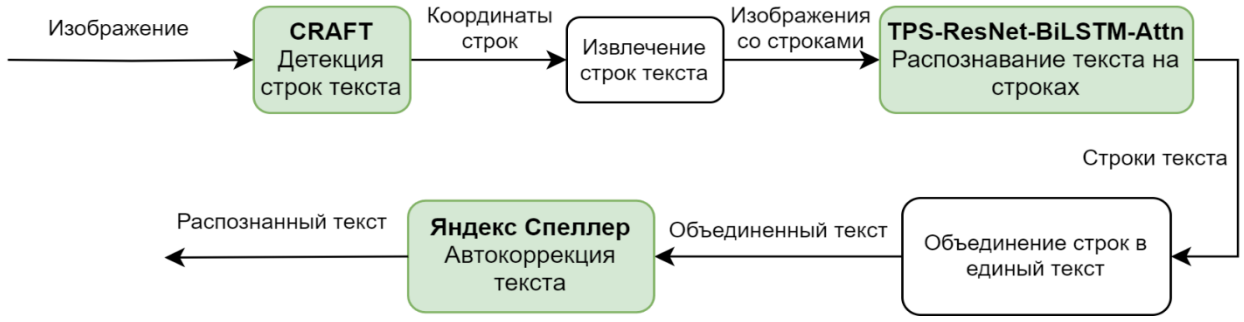


Рис. 2. Последовательность шагов алгоритма распознавания рукописного текста

Эта архитектура, впервые представленная в [7], зарекомендовала себя как одна из самых быстрых и точных для задач распознавания текста и стандартизировала подход к построению моделей. Так, в TPS-ResNet-BiLSTM-Attn за нормализацию отвечает TPS, исправляя геометрические искажения текста, а за извлечение признаков — ResNet. В None-VGG-BiLSTM-Attn за признаки отвечает сверточная нейронная сеть VGG, а нормализация отсутствует. Такой подход позволяет гибко адаптировать архитектуру под конкретные задачи и требования. Текст, полученный на выходе из модели, не всегда идеален и может содержать опечатки, разрывы слов и другие ошибки. Поэтому дополнительным, третьим, этапом является поиск и исправление этих ошибок с помощью систем автоматической коррекции текста. Наиболее эффективным решением является создание специализированной модели коррекции, обученной на большом корпусе текстов. Однако, требует огромного объема данных и значительных вычислительных ресурсов, сравнимых с затратами на само распознавание изображений. Оптимальным вариантом оказалось использование автокорректоров, основанных на словарях и простых эвристических правилах. Такие методы не только эффективны, но и обеспечивают высокую скорость работы. В данной реализации используется Яндекс Спеллер [8], который является бесплатным и обеспечивает высокую точность автоматической коррекции текста на русском языке.

IV. Наборы данных и метрики

1. Наборы данных для распознавания текста

Так как отчеты губернаторов Енисейской губернии содержат текст, который сложно распознать стандартными методами, то необходимо обучать модели распознавания текста с нуля, либо использовать дообучение существующей модели на своих данных. Поэтому в настоящей работе решена задача построения собственного набора данных. Было расшифровано вручную более 500 страниц отчетов, благодаря чему удалось собрать набор объемом в 64 тысячи предложений, слов и словосочетаний, что сравнимо с размерами других русскоязычных корпусов. В рамках данной статьи этот набор данных будет называться «Reports».

Однако, обучение модели распознавания текста требует большего количества разнообразных данных. Поэтому было принято решение включить другие открытые наборы, такие как:

- **Cyrillic Handwriting Dataset (Cyrillic)**: содержит 73830 сегментов рукописных текстов на русском языке. Датасет предоставлен SHIFT Lab CFT [9];
- **Handwritten Kazakh and Russian Database (HKR)**: включает около 95% русских и 5% казахских слов/предложений. Содержит 64943 предложений и 715699 символов, написанных около 200 авторами [10];
- **School Notebooks Dataset (School)**: содержит 864 изображения и 161662 предложения. Взят из открытых источников [11].

Суммарный объем данных составил 364805 предложений. Дополнительно использованы методы искусственного повышения объема и разнообразности данных, такие как синтез рукописного текста (Synthetic) [12] и метод StackMix [13]. Эти методы позволяют увеличивать разнообразие данных за счет генерации новых рукописных текстов и смешивания различных единиц текста для создания новых синтетических примеров, что значительно улучшает обучение модели распознавания [13]. В Табл. 1 представлена статистическая информация о полученном в результате комплексном наборе данных.

ТАБЛИЦА 1
СТАТИСТИКА ИТОГОВОГО НАБОРА ДАННЫХ

№	Набор данных	Количество предложений
1	Reports	64370
2	Cyrillic	73830
3	HKR	64943
4	School	161662
5	Reports (StackMix)	2310265
6	Cyrillic (StackMix)	3700269
7	School (StackMix)	3000000
8	Cyrillic (Synthetic)	3700269
9	HKR (Synthetic)	2634473

Общий объем итогового набора данных для решения задачи распознавания текста составил 15710081 предложений, слов и словосочетаний.

2. Метрики для оценки качества распознавания текста

Для оценки качества моделей были выбраны две основные метрики. Первая метрика — это точность (Accuracy). Она показывает долю правильно распознанных символов относительно общего числа символов в выборке [14], эта метрика напрямую отражает эффективность модели в распознавании каждого символа, предоставляя ясное представление о проценте правильных предсказаний. Второй метрикой было выбрано нормализованное редакционное расстояние (Normalized Edit Distance, NED). Эта метрика измеряет, насколько предсказанный текст отличается от эталонного текста, учитывая длину текста. Нормализованное редакционное расстояние включает количество операций (вставка, удаление, замена), необходимых для преобразования распознанного текста в эталонный [15]. Сочетание этих двух метрик использовано для комплексной оценки производительности модели.

V. Эксперименты и результаты

Для выбора архитектуры нейронной сети при решении задачи распознавания рукописного текста отчетов губернаторов Енисейской губернии проведен ряд экспериментальный исследований, выполненных по принципу от простого к сложному, путем поэтапного расширения набора данных и усложнения архитектуры моделей. Результаты экспериментальных исследований по выбору модели представлены в Табл. 2.

ТАБЛИЦА 2
РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Архитектура модели	**№ набора данных**	**Accyracy**	NED
None-VGG-BiLSTM-CTC	1	40,28	0,73
None-VGG-BiLSTM-CTC	1, 5	58,08	0,84
None-ResNet-BiLSTM-Attn	1, 2, 5, 6	70,44	0,91
TPS-ResNet-BiLSTM-Attn	1 - 9	90,22	0,97

Базовая модель None-VGG-BiLSTM-CTC, которая использовала набор данных, полученных отчетов губернаторов, продемонстрировала начальный уровень точности и NED. Однако, внедрение генерации новых слов с помощью метода StackMix заметно улучшило качество модели. Этот подход не только увеличил объем данных, но и расширил словарный запас, что в итоге привело к более высоким показателям качества. Использование метода StackMix показало, как синтетическое расширение данных может значительно повысить возможности модели, делая ее более адаптивной и точной в распознавании текста.

Так как в задаче распознавания рукописного текста основная сложность заключается в разнообразии почерков и стилей написания, поэтому добавление данных из различных наборов улучшает обобщающую способность модели. Исходя из этого, на следующем этапе решения архитектура механизма извлечения признаков и генерации выходов модели была усложнена, а обучающий набор дополнен данными из других источников.

Самая сложная архитектура TPS-ResNet-BiLSTM-Attn, которая использует обширный набор данных, включая синтетические, продемонстрировала наилучшие результаты по метрикам. Это подтверждает эффективность комплексного подхода, который сочетает сложные архитектурные решения с разнообразием данных. Использование синтетических данных в сочетании с передовыми методами обработки информации позволяет этой архитектуре достичь выдающихся показателей.

VI. Выводы и заключение

Распознавание отчетов губернаторов Енисейской губернии с использованием технологий компьютерного зрения представляет значительный шаг вперед в автоматизации и улучшении качества обработки исторических документов. Важным аспектом работы стало создание специализированных моделей для достижения наилучших результатов распознавания рукописных документов отчетов. Результаты экспериментальных исследований подтверждают, что комбинирование передовых архитектур и разнообразных наборов данных является ключевым фактором в повышении точности и эффективности моделей распознавания рукописного текста на исторических документах.

Таким образом, проект Сибирского федерального университета по оцифровке и распознаванию отчетов губернаторов Енисейской губернии вносит значительный вклад в сохранение и доступность исторических документов, обеспечивая качество, скорость и точность их обработки для исследователей и широкой общественности. Дальнейшим развитием настоящей работы является создание Сервиса с веб-интерфейсом, включающего внедрение моделей для распознавания таблиц и других форматов данных, а также дальнейшее расширение и улучшение набора данных, что позволит дополнительно увеличить качество и эффективность распознавания исторических текстов.

Список литературы

1. Чернышев В. В. Документы волостных правлений как источники изучения истории Енисейской губернии // Енисейская губерния – Красноярский край: 190 лет истории: VII краеведческие чтения, Красноярск, декабрь 2012 г. / М-во культуры Краснояр. края, Арх. агентство Краснояр. края, Гос. универс. науч. б-ка Краснояр. края; [науч. ред. В. И. Федорова ; ред. : Н. Л. Казачкина, С. А. Кудряшова]. – Красноярск : ГУНБ, 2012. – С. 61–63;

2. Губернаторские отчеты: об этом проекте вы еще не раз услышите // СФУ | Мозговой Шторм | Дзен [Электронный ресурс]. URL: https://dzen.ru/a/ZKkVs6CwchYo-AkM (дата обращения: 23.05.2024);

3. Retsinas G., Sfikas G., Gatos B., Nikou C. Best Practices for a Handwritten Text Recognition System // In: Uchida S., Barney E., Eglin V. (eds) Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science, vol 13237. Springer, Cham, 2022. URL: https://doi.org/10.1007/978-3-031-06555-2_17;

4. Baek Y., Lee B., Han D., Yun S., Lee H. Character Region Awareness for Text Detection // Computer Vision and Pattern Recognition. 2019. 12 pages, 11 figures. arXiv:1904.01941 [cs.CV]. URL: https://doi.org/10.48550/arXiv.1904.01941;

5. EasyOCR Documentation // URL: https://jaided.ai/easyocr/documentation/ (дата обращения: 23.05.2024);

6. CRAFT-pytorch // URL: https://github.com/clovaai/CRAFT-pytorch (дата обращения: 23.05.2024);

7. Baek J., Kim G., Lee J., Park S., Han D., Yun S., Oh S. J., Lee H. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis // Computer Vision and Pattern Recognition. Oral paper at ICCV'19. 2019. arXiv:1904.01906 [cs.CV]. URL: https://doi.org/10.48550/arXiv.1904.01906 (дата обращения: 23.05.2024);

8. Яндекс Спеллер // URL: https://yandex.ru/dev/speller/ (дата обращения: 23.05.2024);

9. Cyrillic Handwriting Dataset // URL: https://www.kaggle.com/datasets/constantinwerner/cyrillic-handwriting-dataset (дата обращения: 23.05.2024);

10. Nurseitov D., Bostanbekov K., Kurmankhojayev D., Alimova A., Abdallah A., Tolegenov R. Handwritten Kazakh and Russian (HKR) database for text recognition // Multimedia Tools and Applications. Springer, 2021. 1-23;

11. AI-Forever School Notebooks RU // URL: https://huggingface.co/datasets/ai-forever/school_notebooks_RU (дата обращения: 23.05.2024);

12. Balreira D., Walter M. Handwriting Synthesis from Public Fonts // 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 246-253. 2017. URL: https://doi.org/10.1109/SIBGRAPI.2017.39;

13. Shonenkov A., Karachev D., Novopoltsev M., Potanin M., Dimitrov D. StackMix and Blot Augmentations for Handwritten Text Recognition // arXiv preprint arXiv:2108.11667. 2021. URL: https://doi.org/10.48550/arXiv.2108.11667;

14. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Journal. 2001. Vol. 14;

15. Marzal A., Vidal E. Computation of normalized edit distance and applications // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1993. Vol. 15, No. 9. P. 926–932. DOI: 10.1109/34.232078.

УДК 681.7.069.32
DOI: 10.25206/978-5-8149-3873-2-2024-525-531

Вейвлет-фильтрация зашумленных изображений с использованием их аппроксимаций в субполосах деталей
Wavelet-based filtering of noisy images using their approximations in subbands of details

Ю. С. Бехтин, В. Т. Чинь
ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В. Ф. Уткина», Рязань, Россия

Yu. S. Bekhtin, V. T. Chin
Ryazan State Radio Engineering University named after V. F. Utkin, Ryazan, Russia

Аннотация. Искажения изображений могут возникать как под действием внешних факторов, так и в процессе их формирования из-за физической природы различных оптико-электронных систем. Для подавления широкополосных шумов разработано множество методов, среди которых фильтрация на основе вейвлет-преобразования занимает особое место благодаря возможности эффективно подавлять не только аддитивные, но и мультипликативные шумы. Широко применяемые на практике методы вейвлет-фильтрации базируются на использовании пороговой обработки вейвлет-коэффициентов деталей, то есть обработки высокочастотных субполос зашумленного изображения. При этом низкочастотная составляющая изображения – аппроксимация – как правило, не подвергается пороговой обработке. В связи с этим, подбор эффективных правил пороговой обработки и значений порогов не является тривиальной задачей, поскольку часто пороговая обработка деталей приводит к появлению артефактов из-за искажений в низкочастотной области. В работе приводятся результаты исследования модифицированного метода вейвлет-фильтрации зашумленных изображений, который при пороговой обработке использует аппроксимацию в субполосах деталей, что не только повышает отношение сигнал-шум, но и уменьшает количество артефактов.

***Ключевые слова*:** цифровая обработка изображений, вейвлет-преобразование, вейвлет-фильтрация, пороговая обработка.

Abstract. Image distortions can occur both under the influence of external factors and in the process of image formation due to the physical nature of various optoelectronic systems. To suppress broadband noise, many methods have been developed, among which filtering based on the wavelet transform occupies a special place due to the ability to effectively suppress both additive noise and also multiplicative noise. The methods of wavelet filtering widely used in practice are based on thresholding of wavelet coefficients which are belong in high frequency subbands named details. At the same time, the low frequency component of the image named approximation is not mostly subjected to threshold processing. In this regard, the choice of effective thresholding rule and threshold values is not a trivial task. It is because thresholding of details leads to the appearance of artifacts due to distortions in the low frequency image components. The paper presents the results of a study of a modified wavelet filtering method for noisy images which uses approximation in details. The results have shown that the suggested method not only increases the signal-to-noise ratio, but also reduces the number of artifacts.

***Keywords*:** digital image processing, wavelet transform, wavelet-based filtering, thresholding.

I. Введение

Несмотря на повышение разрешающей способности видеокамер и датчиков в последнее время, изображения на их выходе часто формируются с наличием помех и шумов. И если от искажений, вызванных внешними факторами, можно избавиться, если найти источник этих искажений и его устранить, то от искажений, обусловленных воздействием химико-физических процессов первичных датчиков в ходе формирования изображений, практически избавиться не удается. Вопрос заключается лишь в технологии производства самих первичных датчиков, которая может обеспечивать или не обеспечивать минимальный теоретически достижимый уровень внутренних искажающих воздействий. Таким образом, подавление остаточных помех и шумов, вызванных физической природой формирования изображений в самих видеодатчиках, всегда остается актуальной задачей вторичной обработки изображений. Характерными примерами являются изображения, сформированные датчиками оптического диапазона с некалиброванными ПЗС матрицами, многоэлементными фотоприемниками с неоднородностью фотоэлементов, ультразвуковые и радиолокационные изображения с сильно выраженным спекл-шумом и т.п. [1-6].

Как правило, проблема подавления внутренних шумов носит специфический характер, поскольку создать универсальный фильтр для всех случаев невозможно. Поэтому на производстве для каждого типа датчиков разрабатывают свои собственные методы фильтрации. К сожалению, эффективность подобных, как правило, технических решений не всегда высокая, поэтому при решении различных задач вторичной обработки изображений неизбежно возникает вопрос об их предварительной обработке, включая фильтрацию шумов.

Выбор метода фильтрации, очевидно, зависит, в том числе, и от того, какую модель шума использует специалист по обработке изображений. Общепризнанным является разделение широкополосных шумов на две группы: аддитивные и мультипликативные шумы [1, 2, 8].

В первом случае в качестве модели аддитивного шума используют гауссовский белый шум с нулевым средним [7-9]. Известно большое количество методов шумоподавления, которые получены путем строгого математического вывода на основе данной модели шума. Поскольку белого шума не существует в природе, то, очевидно, подобные методы не всегда являются эффективными. Тем не менее, сегодня подавление аддитивного высокочастотного шума не вызывает больших проблем в смысле устойчивости соответствующих алгоритмов обработки [3, 10].

Гораздо хуже ситуация обстоит в отношении мультипликативных шумов [8-12]. Подавление мультипликативного шума сводится к отделению полезной компоненты от шумового мультипликатора, что трудно или практически невозможно сделать без предварительных преобразований данных изображения, чтобы перейти к аддитивной модели. Одним из подходов является применение логарифмирования, которое можно отнести к методам гомоморфной обработки изображений [3, 7]. К сожалению, логарифмирование приводит к существенному снижению отношения сигнал-шум и искажению оригинала изображения. В результате оценка изображения (то есть восстановленное после фильтрации и взятия обратного логарифма изображение) оказывается смещенной, появляются артефакты, сглаживание краев объектов и т.п.

Другим подходом является применение преобразований перед собственно фильтрацией, чтобы перейти в другое пространственное представление изображений, где подавление мультипликативного шума можно рассматривать с точки зрения фильтрации аддитивного шума. Одним из таких преобразований является вейвлет-преобразование [13-17], которое, в отличие от преобразования Фурье, позволяет сохранить пространственные особенности изображений в каждой из частотных субполос. Показано [14], что полученное после быстрого вейвлет-преобразования (БВП) [13] трансформированное изображение можно рассматривать как совокупность массивов вейвлет-коэффициентов, представляющих масштабированные копии исходного изображения в соответствующих субполосах. В результате принято называть деталями (или высокочастотными компонентами) все субполосы, кроме одной субполосы, которая называется аппроксимацией (или низкочастотной составляющей) и куда «перетекает основная часть энергии» изображения (рис. 1).

Для фильтрации и/или компрессии изображения используется пороговая обработка вейвлет-коэффициентов (thresholding). Существует много правил пороговой обработки [7, 13-20], сущность которой заключается в сравнении величины вейвлет-коэффициента с заранее вычисленным или эмпирически найденным пороговым значением, после чего вейвлет-коэффициенты, оказавшиеся ниже порога, обнуляются, а оставшиеся вейвлет-коэффициенты либо остаются без изменений (hard thresholding), либо модифицируются по какому-либо закону (например, soft thresholding). Поскольку обнуляются те вейвлет-коэффициенты, которые соответствуют деталям и которые больше искажены из-за действия шума, то таким образом происходит фильтрация зашумленных изображений после обратного вейвлет-преобразования. Как правило, вейвлет-коэффициенты аппроксимации не подвергаются пороговой обработке, поскольку считается, что отношение сигнал-шум в данной субполосе относительно высокое. Таким образом, сохраняется информативный низкочастотный компонент.

Из вышесказанного следует, что выбор правила пороговой обработки и величины порога играет существенную роль в процессе фильтрации зашумленного изображения. Известно много методов расчета порогового значения [13, 16]. В частности, можно рассчитывать глобальный порог для обработки всех деталей, а можно искать значения порогов для каждой субполосы. Но в любом случае возможны ситуации, когда обнуление некоторых вейвлет-коэффициентов выражается в значительном сглаживании изображения и/или появлении артефактов (последнее проявляется как размытие краев объектов). Предугадать, какие коэффициенты, оказавшиеся ниже порога, не нужно обнулять, теоретически и практически очень сложно, и авторам неизвестны подобные решения. Очевидно, что причиной обнуления значимых вейвлет-коэффициентов является низкое отношение сигнал-шум в субполосах из-за действия шума. В связи с этим возникает задача повысить отношение сигнал-шум за счет добавления низкочастотных составляющих в субполосы. Идея предлагаемого метода вейвлет-фильтрации заключается в использовании для этой цели аппроксимацию, поскольку она практически не содержит искажений из-за действия шума. Альтернативой является добавление предварительно обработанного любым фильтром изображения к исходному искаженному изображению, однако данный подход не эффективен из-за непредсказуемых искажений после предварительной фильтрации.

II. Постановка задачи

Будем считать, что в процессе формирования изображений происходит их искажение мультипликативным шумом, в результате чего математическая модель наблюдаемого сигнала Y имеет вид:

$$Y = XZ, \quad (1)$$

где X – оригинальное (неискаженное) изображение, Z – мультипликативный шум с единичным средним. В зависимости от типа видеодатчика распределение шума не всегда является гауссовским; например, для спекл-шума (характерно для радиолокационных и ультразвуковых изображений) принимается экспоненциальное распределение с единичным средним, для шумов ПЗС – логнормальное и т.п.

Для получения оценки изображения $\hat{X}$ будем применять трехуровневое вейвлет-преобразование (рис. 1) с использованием последующей пороговой обработки вейвлет-коэффициентов по т.н. «грубому» правилу (hard thresholding). На рис 1 буквами A, D, H и V с индексами обозначены массивы вейвлет-коэффициентов аппроксимации, диагональных, горизонтальных и вертикальных субполос соответственно. При обработке вейвлет-коэффициентов по «грубому» правилу те коэффициенты, которые оказались ниже порога, обнуляются, а оставшиеся коэффициенты остаются без изменений. Отметим, что вейвлет-коэффициенты аппроксимации (сегмент $A3$ в верхнем левом углу на рис. 1) не подвергаются пороговой обработке.

Тогда модификация метода пороговой вейвлет-фильтрации зашумленного изображения будет заключаться в использовании информации об аппроксимации в субполосах деталей путем восстановления «псевдоизображений» по отдельным субполосам с добавлением низкочастотного компонента, полученного из аппроксимации.

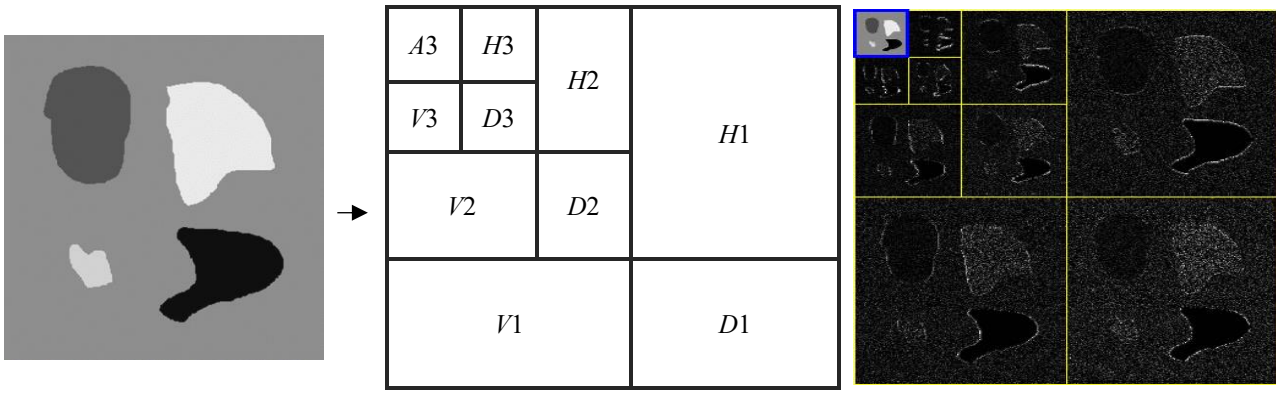


Рис. 1. Схема трехуровневого быстрого вейвлет-преобразования

III. Использование аппроксимации при пороговой обработке деталей

Основной принцип работы предложенного метода и схема проведения экспериментальных исследований для выявления его эффективности иллюстрируется на рис. 2. Описать работу метода удобно в форме алгоритма, причем в целях упрощения записи не рассматривается, как происходят вычисления в пределах субполос.

1. На первом шаге (этапе) вначале выполняется трехуровневое двумерное вейвлет-преобразование, как показано на рис. 1, и формируются соответствующие массивы вейвлет-коэффициентов.
2. Теперь необходимо получить псевдоизображение по аппроксимирующим коэффициентам, для чего, например, восстанавливаем изображение Y_A3 из коэффициентов аппроксимации $A3$, применяя процедуру обратного вейвлет-преобразования. При этом детализирующие коэффициенты берутся нулевыми.
3. На зашумленное изображение добавляется изображение, восстановленное по аппроксимирующим коэффициентам:

$$\tilde{Y} = Y + Y\_A3. \quad (2)$$

4. Снова выполняется трехуровневое вейвлет-преобразование над модифицированным изображением $\tilde{Y}$. Формируются массивы вейвлет-коэффициентов деталей.
5. Применяется «грубая» пороговая обработка вейвлет-коэффициентов деталей.
6. С помощью обратного вейвлет-преобразования получаем «грубую» оценку оригинала изображения $\hat{Y}$.
7. Для получения оценки оригинала $\hat{X}$ необходимо вычесть из «грубой» оценки $\hat{Y}$ псевдоизображение, полученное по коэффициентам аппроксимации Y_A3, то есть выполнить расчеты вида:

$$\hat{X} = \hat{Y} - Y\_A3. \quad (3)$$

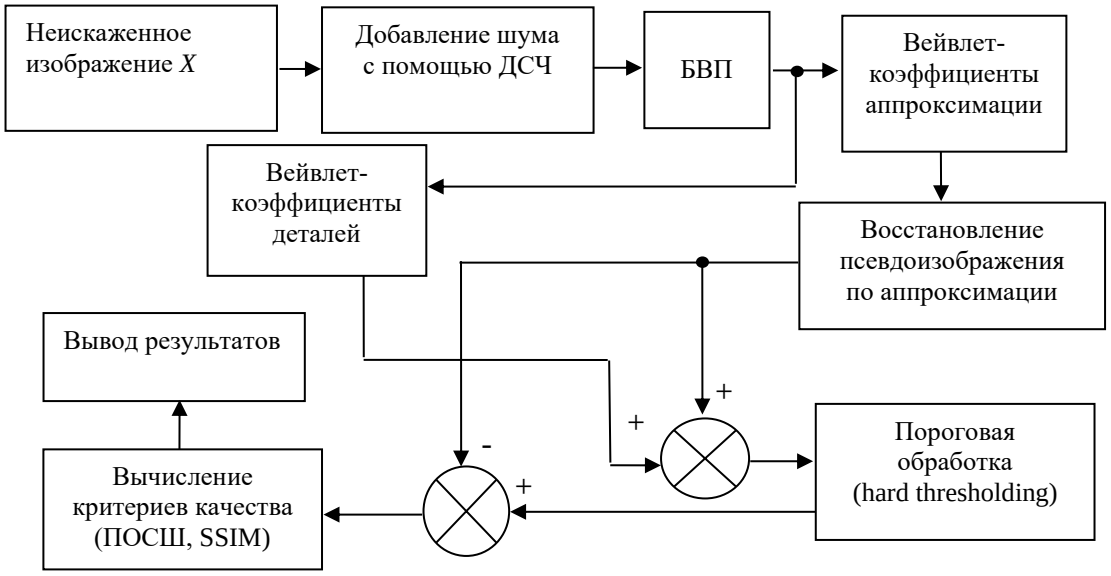


Рис. 2. Принцип работы модифицированного метода вейвлет-фильтрации и проведения экспериментальных исследований

IV. Результаты экспериментов

Эффективность предложенного метода вейвлет-фильтрации проверялась путем статистического моделирования на ЭВМ в среде программирования Matlab. Кроме использования известных функций из Wavelet Toolbox дополнительно были разработаны коды на языке C++, которые после компиляции в виде динамических библиотек применялись для выполнения таких рутинных операций, как приведение промежуточных изображений к исходному яркостному динамическому диапазону, согласование размеров субполос на разных уровнях вейвлет-преобразования при формировании псевдоизображений (зависит от типа вейвлет-базиса), учет краевых эффектов и т.п.

Работоспособность предложенного метода проверялась с помощью библиотеки изображений TID2013, которая содержит изображения различных классов, а также зашумленные изображения. Тем не менее, для оценки эффективности метода использовался подход, отраженный на рис. 2. Для генерации мультипликативного шума использовался датчик равномерно распределенных случайных чисел (ДСЧ), имеющийся в среде Matlab. Полученные с его помощью псевдослучайные числа подвергались обработке с целью получения массивов шумовых отсчетов, распределенных по экспоненциальному закону с единичным средним. Дисперсия шума изменялась в ходе экспериментов. Таким образом, моделировался «тяжелый» случай мультипликативного шума, который встречается в тепловизионных, ультразвуковых и радиолокационных изображениях.

В Табл. 1 приводятся результаты обработки эталонного изображения в градациях серого, с разрешением 8 бит/пиксель, размером 512×512 элементов, содержащим четыре объекта с произвольными формами и яркостями на темном фоне. Величина глобального порога для «грубого» правила находилась через дисперсию массива вейвлет-коэффициентов субполосы $H1$. В качестве критериев эффективности использовались среднеквадратическая ошибка оценивания (СКО), которая вычислялась через разностное изображение $X - \hat{X}$ (рис. 2) и на основе которой затем вычислялись значения пикового отношения сигнал-шум (ПОСШ), а также индекс структурного сходства SSIM (Structure Similarity Index Measure) [3, 7]. Для сравнения в Табл. 1 помещены также данные по СКО, ПОСШ и SSIM для традиционного метода вейвлет-фильтрации на основе правила «грубой» пороговой обработки. В ходе исследования предложенного метода использовались различные типы вейвлетов; в табл. 1 отражены результаты, полученные при использовании «классического» для шумоподавления вейвлет-базиса db7 из семейства вейвлетов Добеши [15].

Для наглядности полученные данные статистической обработки для СКО и ПОСШ показаны также в виде соответствующих зависимостей на рис. 3 и рис. 4 соответственно. На рис. 5, а приведено неискаженное тестовое изображение, на рис. 5, б представлено зашумленное тестовое изображение, а рис. 5, в и 5, г отражают результаты обработки зашумленного тестового изображения известным и предложенным методом вейвлет-фильтрации соответственно.

ТАБЛИЦА 1
РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

СКО шума	Зашумлённое изображение			Восстановленное изображение					
	СКО	ПОСШ (дБ)	SSIM	СКО		ПОСШ (дБ)		SSIM	
				(*)	(**)	(*)	(**)	(*)	(**)
0.016	4.29	41.91	0.95	4.29	2.86	41.91	43.49	0.95	0.96
0.021	7.42	39.42	0.91	7.42	4.78	39.42	41.32	0.91	0.94
0.026	11.97	37.53	0.87	11.97	6.83	37.53	39.93	0.87	0.92
0.032	17.06	35.63	0.82	17.06	8.76	35.63	38.65	0.82	0.91
0.038	29.24	33.58	0.77	29.24	15.05	33.58	36.39	0.77	0.90
0.045	50.30	31.18	0.73	50.30	29.24	31.18	33.53	0.73	0.90
0.051	82.74	28.97	0.68	82.74	53.33	28.97	30.86	0.68	0.91
0.059	125.61	27.15	0.63	101.32	85.93	28.08	28.79	0.76	0.93
0.066	179.05	25.62	0.59	137.46	130.63	26.75	26.96	0.83	0.94
0.075	226.45	24.56	0.54	167.00	167.92	25.90	25.87	0.90	0.94

(*) – традиционный метод вейвлет-фильтрации, (**) – предложенный метод вейвлет-фильтрации.

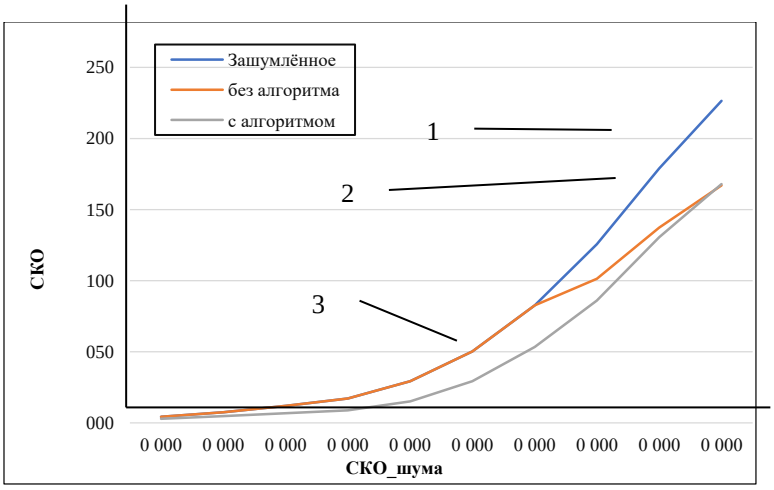


Рис. 3. Зависимость СКО оценки *X* от СКО шума: 1 – СКО для зашумленного изображения; 2 – СКО для традиционного метода; 3 – СКО для предложенного метода

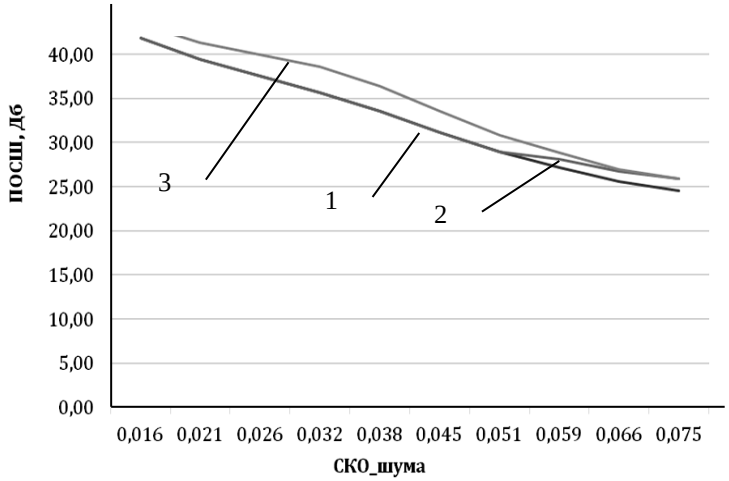
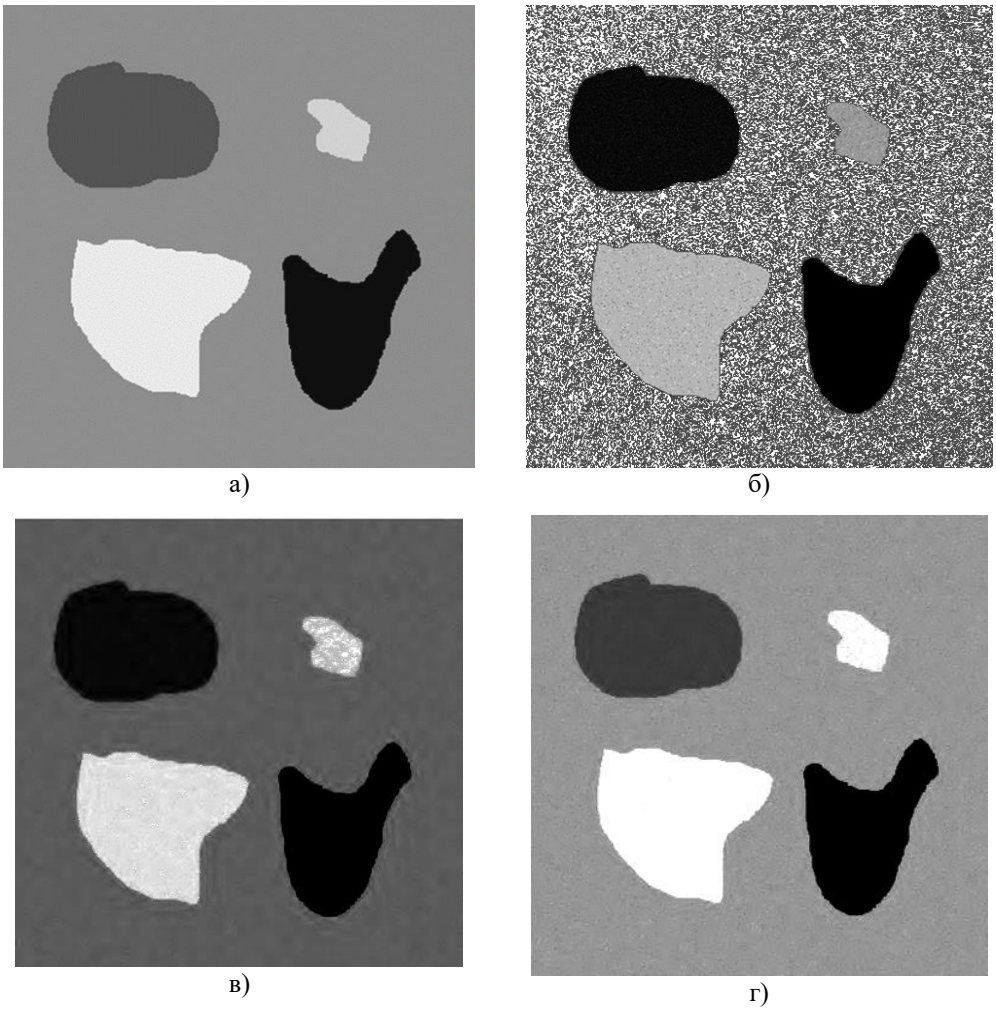


Рис. 4. Зависимость ПОСШ от СКО шума: 1 – ПОСШ для зашумленного изображения; 2 – ПОСШ для традиционного метода; 3 – ПОСШ для предложенного метода



а) б) в) г)

Рис.5. Результаты обработки: а) – неискаженное тестовое изображение; б) – зашумленное изображение; в) – результат пороговой обработки; г) – результат обработки предложенным методом вейвлет-фильтрации

V. Обсуждение результатов

Как видно из рис. 3 и рис. 4, а также Табл. 1, предложенный метод выигрывает у традиционной пороговой обработки зашумленного изображения при относительно небольших значениях дисперсии шума по критериям СКО, ПОСШ и SSIM. Тем не менее, оба метода показывают примерно одинаковое увеличение ПОСШ (или уменьшение СКО оценки оригинала) при увеличении дисперсии шума до таких значений, когда практически объекты на изображении становятся неразличимы визуально. Данный факт можно объяснить тем, что при относительно больших шумах добавление аппроксимации в детали вносит дополнительные искажения вместо повышения отношения сигнал-шум, поскольку аппроксимация модифицированного изображения, полученного через псевдоизображения с добавлением аппроксимации исходного изображения, будет соизмерима по величине с последней. Таким образом, предложенный метод следует применять в тех случаях, когда удаление шумов не должно приводить к появлению новых артефактов. Данное свойство предложенного метода иллюстрируется рис. 5, из которого видно, что в данном конкретном примере традиционный метод вейвлет-фильтрации вносит свои артефакты, причем они появляются не только по краям объектов, но также и в изображениях самих объектов (рис. 5, в). Как видно из рис. 5, г, количество артефактов, вносимое предложенным методом, значительно ниже, и они заметны только по краям некоторых объектов. Это подтверждается и численными значениями индекса структурного сходства SSIM, которые больше у предложенного метода (Табл. 1). Необходимо также отметить, что предложенный метод вейвлет-фильтрации с использованием аппроксимации в субполосах практически восстанавливает оригинальное распределение яркостей объектов и фона (рис. 5, а и рис. 5, г), в то время, как традиционный подход может установить ошибочную яркость объектов и фона (рис. 5, в).

VI. Выводы и заключение

Результатом данной работы является проверка работоспособности идеи о добавлении в зашумленное изображение его низкочастотных компонентов с целью повышения исходного отношения сигнал-шум перед фильтрацией. Реализация идеи осуществлена в рамках вейвлет-фильтрации изображений, искаженных мультипликативным шумом с единичным средним. Результаты статистического моделирования на ЭВМ показали работоспособность такого подхода в целом, что указывает на необходимость дальнейшего проведения исследований и модификаций разработанного метода. Перспективным можно считать сочетание локальной пороговой обработки по различным правилам с предварительным извлечением и последующим добавлением низкочастотных компонентов (аппроксимаций) для каждой субполосы вейвлет-преобразования. Также открытыми пока остаются вопросы о выборе вейвлет-базиса, эффективности подхода при обработке и комплексировании мультиспектральных изображений и др.

Список литературы

1. Katsuhiko Ogata, Modern Control Engineering. Fifth Edition, Prentice Hall, 2010. 905 с.
2. Бесекерский, В. А. Радиоавтоматика. М.: ВШ, 1985. 271 с.
3. Morris, A., Langari, R. Measurement and Instrumentation: Theory and Application, Second Edition, Academic Press, 2016. 640 с.
4. Norton, P., Campbell and S. Horn. Third generation of infrared imagers // Proceeding SPIE, vol. 4130, 2000, pp. 226-235.
5. P. Singh, and R. Mukundan. Synthetic Models of Ultrasound image formation for Speckle Noise Simulation and Analysis // International Conference on Signals and Systems (ICSigSys), 2017, pp. 278-284.
6. Edelman, K. Understanding Ultrasound Physics. E.S.P. Ultrasound, 4th edition. 2012. 567 с.
7. Гонсалес, Р., Вудс, Р. Цифровая обработка изображений. М.: Техносфера, 2012. 1104 с.
8. Vaseghi, S. Advanced digital signal processing and noise reduction. John Wiley & Sons, 2000. 493 с.
9. Hayes, M. H. Statistical digital signal processing and modeling. John Wiley & Sons, 2009. 608 с.
10. Moon, Todd K., Stirling, Wynn C. Mathematical Methods and Algorithms for Signal Processing, 2000. 937 с.
11. Lee, J.S. Refined filtering of image noise using local statistics. Computer Vision, Graphics and Image Processing. 1998, P. 380-389.
12. D. T. Kuan, A. A. Sawchuk, T. C. Strand and P. Chavel, "Adaptive Noise Smoothing Filter for Images with Signal-Dependent Noise," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-7, no. 2, pp. 165-177, March 1985.
13. Vidakovic, B. Statistical Modeling by Wavelets. John Wiley & Sons, 1999. 416 с.
14. Воробьев, В.И., Грибунин, В.Г. Теория и практика вейвлет-преобразования. С.-Петербург, ВУС, 1999. 206 с.
15. Малла, С. Вейвлеты в обработке сигналов. М.: Мир, 2005. 671 с.
16. Бехтин, Ю. С., Емельянов С.Г., Титов Д.В. Теоретические основы цифровой обработки изображений встраиваемых оптико-электронных систем. Монография, изд-во Аргамак-Медиа, Москва, 2016 г. 296 с.
17. Бехтин, Ю.С. Обработка зашумленных изображений на основе вейвлет-преобразований // Вестник РГРТУ. 2006. № 18. С.141-144.
18. Bruce, A. and Gao, H. -Y. WaveShrink: Shrinkage Functions and Thresholds. Proc. SPIE, San Diego, CA, 1995.
19. Stein, C. Estimation of the mean of a multivariate normal distribution // Ann. Statistics, No. 9, 1981, pp. 1135–1151.
20. Donoho, D.L. De-noising by Soft-Thresholding // IEEE Trans. on Information Theory, Vol.41, No.3, 1995, pp. 613-627.

УДК 621.319
DOI: 10.25206/978-5-8149-3873-2-2024-532-537

Подход к автономному сопровождению объектов на базе оптических и радио-систем видения
The concept of autonomous object tracking based on optical and radio- vision systems

В. К. Клочко
Рязанский государственный радиотехнический университет имени В.Ф. Уткина, Гагарина,59/1, Рязань, 390005, Россия

V. K. Klochko
Ryazan State Radio Engineering University named after V.F. Utkin, Gagarin, 59/1, Ryazan, 390005, Russia

Аннотация. Предложена концепция построения автономной навигационной системы при сопровождении БПЛА. Концепция основана на классификации особых точек эталонного и текущего изображения местности для коррекции БПЛА и отличается наличием критерия классификации по минимуму квадрата нормы вектора ошибок сопряжения пар особых точек. Целью работы является развитие метода анализа опорных точек, полученных на базе оптических и радио- изображений местности в автономной навигационной системе. Предложен алгоритм коррекции движения БПЛА, позволяющий использовать вместо корреляционной функции быстрые матричные операции решения линейных алгебраических уравнений, учитывать пространственную ориентацию БПЛА, определять вектор его направления движения к расчетному положению в текущие моменты времени по экстраполированной траектории и совмещать работу оптической и радио- систем.

Ключевые слова: Автономная навигация, беспилотный летательный аппарат, эталонное и текущее изображения, опорные точки, оптическая и радио системы.

Abstract. The concept of building an autonomous navigation system when accompanied by a UAV is proposed. The concept is based on the classification of singular points of the reference and current terrain images for UAV correction and is distinguished by the presence of a classification criterion for the minimum square of the norm of the error vector of conjugation pairs of singular points. The aim of the work is to develop a method for analyzing singular points obtained on the basis of optical and radio- images of the terrain in an autonomous navigation system. An algorithm for correcting the movement of a UAV is proposed, which allows using fast matrix operations for solving linear algebraic equations instead of the correlation function, taking into account the spatial orientation of the UAV, determining the vector of its direction of movement to the calculated position at current moments along an extrapolated trajectory and combining the operation of an optical and radio- systems.

Keywords: Autonomous navigation, unmanned aerial vehicle, reference and current images, singular points, optical and radio systems.

Введение

При разработке современных навигационных систем для беспилотных летательных аппаратов (БПЛА, в дальнейшем ЛА) стоит задача совершенствования существующих методов навигации [1, 2] для их применения в условиях автономного движения без поддержки спутниковых и радионавигационных данных. Применение в таких условиях инерциальных навигационных систем ограничено, так как они имеют свойство накапливать погрешности. Существует множество научных и практических работ, посвященных системам автономной навигации. Наибольшее распространение на практике получили корреляционно-экстремальные системы навигации (КЭСН), основанные на обработке оптических, тепловых и радио- изображений [3]. Известно большое количество методов решения задач КЭСН: методы анализа оптического потока; методы морфологического анализа изображений; методы, основанные на сопоставлении изображений в целом; методы, основанные на сопоставлении элементов (особых точек) и локальных участков изображений. Преимущество последних методов (анализ особых точек) состоит в использовании инвариантного описания изображений, позволяющего сопровождать объекты в условиях взаимных геометрических преобразований текущего и эталонного изображений, например [4 − 7]. Направлению анализа особых точек для коррекции траектории движения ЛА в автономных КЭНС придерживается данная работа.

Целью работы является развитие метода анализа опорных точек, полученных на базе оптических и радио-изображений местности в автономной навигационной системе. Предлагается концепция, которая заключается в коррекции траектории движения ЛА на основе сопоставления опорных точек эталонного изображения (ЭИ) и текущего изображения (ТИ) по определенному критерию сопряжения.

Анализ известного решения

Рассмотрим известное решение построения автономной КЭСН, основанное на сравнении текущего и эталонного изображений местности и вычислении поправок горизонтальных смещений ЛА при наличии измерений высоты полета, например, [1]. Эти поправки учитываются при коррекции траектории движения ЛА.

Алгоритмически решение сводится к следующему.

1. По электронной карте местности, представляющей ЭИ в прямоугольной наземной системе координат $O_1X_1Y_1Z_1$ рассчитывается траектория движения ЛА в виде зависимости координат $x_1(t), y_1(t), z_1(t)$ точки положения ЛА $M_1(x_1, y_1, z_1)$ от времени *t*, где $z_1(t) = h_1(t)$ – высота полета.

2. Формируется плоское (двумерное) оптическое или радио- ТИ участка местности с помощью с помощью оптической станции или бортовой РЛС в текущий момент времени *t* в системе координат ЛА $O_2X_2Y_2$, где точка O_2 совпадает с точкой $M_2(x_2, y_2, z_2)$ положения ЛА в момент времени *t* в наземной системе координат $O_1X_1Y_1Z_1$.

3. Измеряется высота полета ЛА $h_2(t)$ с помощью бортовой аппаратуры и с учетом $\Delta h = h_2(t) - h_1(t)$ ТИ масштабируется по отношению к ЭИ.

4. Изображения ЭИ и ТИ совмещаются и вычисляется корреляционная функция $f(\Delta x, \Delta y)$, зависящая от горизонтальных смещений $\Delta x(t) = x_2(t) - x_1(t)$ и $\Delta y(t) = y_2(t) - y_1(t)$ ТИ относительно ЭИ в системе координат $O_1X_1Y_1Z_1$.

5. Находятся экстремальные значения $\Delta x^*(t)$ и $\Delta y^*(t)$, на которых достигается наибольшее значение функции $f_{\text{наиб}} = f(\Delta x^*, \Delta y^*)$.

6. Осуществляется коррекция движения ЛА путем изменения высоты $h_2(t)$ во времени *t* до расчетного значения $h_1(t)$ и изменения положения ЛА от точки $M_2(x_2, y_2, z_2)$ до расчетной точки $M_1(x_1, y_1, z_1)$ в системе координат $O_1X_1Y_1Z_1$, где $x_1(t) = x_2(t) - \Delta x^*(t)$ и $y_1(t) = y_2(t) - \Delta y^*(t)$.

Данное решение имеет следующие недостатки.

1. Вычисление корреляционной функции на основе изображений местности требует больших вычислительных затрат.

2. Пространственная ориентация ЛА в системе координат $O_1X_1Y_1Z_1$ не учитывается, что не позволяет правильно совмещать ЭИ и ТИ с учетом угловых отклонений системы координат $O_2X_2Y_2Z_2$ относительно $O_1X_1Y_1Z_1$.

3. Отсутствует информация о направлении движения ЛА по результатам коррекции положения ЛА.

4. В случае использования РЛС работа станции ведется непрерывно в активном режиме излучения зондирующего сигнала и подвержена внешним радиопомехам.

Предлагаемое решение

Для устранения указанных недостатков предлагается следующее решение.

1. Рассчитывается траектория движения ЛА по электронной карте местности (ЭИ) в виде зависимости координат $x_1(t), y_1(t), z_1(t)$ положения ЛА $M_1(x_1, y_1, z_1)$ во времени *t* в наземной системе координат $O_1X_1Y_1Z_1$.

2. Включается периодически бортовая РЛС, формируется радио ТИ и после этого РЛС временно выключается.

3. По определенному правилу [4 – 7] находятся опорные точки $M_1(i) = M(x_1(i), y_1(i), z_1(i))$, $i = \overline{1, n_1}$, в количестве n_1 в трехмерной системе координат $O_1X_1Y_1Z_1$ ЭИ и по такому же правилу находят опорные точки $M_2(j) = M_2(x_2(j), y_2(j), z_2(j))$, $j = \overline{1, n_2}$, в количестве n_2 в трехмерной системе координат $O_2X_2Y_2Z_2$ ТИ.

4. Осуществляется перебор *k*-х вариантов соединения точек $M_1(i)$, $i = \overline{1, n_1}$, и $M_2(j)$, $j = \overline{1, n_2}$, в m_k непересекающихся сопряженных пар $M_1(i)$ и $M_2(j_i)$, где номер j_i поставлен в соответствие номеру *i*, $i = \overline{1, m_k}$, $m_k \le \min(n_1, n_2)$, $m_k > 12$.

5. Для каждого *k*-го варианта соединения сопрягаемых пар координаты точек $M_1(i)$, $i = \overline{1, m_k}$, помещаются в $3m_k \times 1$ - матрицу Y_k, а координаты соответствующих им точек $M_2(j_i)$ помещаются в $3m_k \times 12$ -матрицу X_k:

$$Y_k = \begin{bmatrix} M_1(1) \\ \dots \\ M_1(m_k) \end{bmatrix}, \quad M_1(i) = \begin{bmatrix} x_1(i) \\ y_1(i) \\ z_1(i) \end{bmatrix}, \quad X_k = \begin{bmatrix} X_{k1} \\ \dots \\ X_{k,m_k} \end{bmatrix},$$

$$X_{ki} = \begin{bmatrix} x_2(j_i) & y_2(j_i) & z_2(j_i) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_2(j_i) & y_2(j_i) & z_2(j_i) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_2(j_i) & y_2(j_i) & z_2(j_i) \end{bmatrix},$$

где дополнительно в составе 3×12 -матрицы X_{ki} расположены еще 3 последних столбца $[1 \quad 0 \quad 0]^T$, $[0 \quad 1 \quad 0]^T$, $[0 \quad 0 \quad 1]^T$, и вычисляется 12×1-вектор P_k параметров ориентации ТИ относительно ЭИ по формуле

$$P_k = (X_k^T X_k)^{-1} X_k^T Y_k, \tag{1}$$

где в составе $P_k = (p_1, p_2, \dots, p_9, b_x, b_y, b_z)^T$ находятся 9 параметров матрицы H_k поворота осей системы $O_2X_2Y_2Z_2$ относительно $O_1X_1Y_1Z_1$:

$$H_k = \begin{bmatrix} p_1 & p_2 & p_3 \\ p_4 & p_5 & p_6 \\ p_7 & p_8 & p_9 \end{bmatrix}$$

и 3 координаты вектора $b_k = (b_x, b_y, b_z)^T$, соединяющего точки O_1 и O_2.

6. В каждом *k*-м варианте соединения опорных точек $M_1(i)$ и $M_2(j_i)$ в *i*-й паре вычисляется показатель сопряжения

$$J_{ki} = (M_1(i) - H_k M_2(j_i) - b_k)^T (M_1(i) - H_k M_2(j_i) - b_k) \tag{2}$$

и суммарный показатель $J_k = \sum_{i=1}^{m_k} J_{ki}$.

7. Выбирается k^*-й вариант соединения опорных точек с наименьшим значением показателя J_{k^*} при условии $J_{k^*} \le \gamma$, где γ – заданный порог, и запоминаются матрица H_{k^*} и вектор b_{k^*}.

8. Включается оптическая станция и формируется оптическое ТИ. Находятся по определенному правилу на ЭИ опорные точки в трехмерной системе координат ЭИ, а также опорные точки в двумерной системе координат ТИ. Двумерные координаты опорных точек ТИ переводятся в трехмерные координаты ортов векторов направлений в системе координат $O_2X_2Y_2Z_2$.

9. Осуществляется перебор *k*-х вариантов соединения опорных точек ЭИ и ортов векторов направлений ТИ в m_k сопрягаемых пар ($m_k > 12$), при этом в каждом *k*-м варианте для каждой сопрягаемой пары вычисляются трехмерные координаты опорной точки ТИ по формуле

$$M_{\text{ТИ}} = \hat{r}a, \ \hat{r} = a^T H_{k^*}^T \cdot (M_{\text{ЭИ}} - b_{k^*}),$$

где $M_{\text{ЭИ}}$ – опорная точка ЭИ, a – 3×1-орт вектора направления ТИ.

10. Помещаются координаты опорных точек ЭИ в $3m_k \times 1$-вектор Y_k, а координаты опорных точек ТИ в $3m_k \times 12$-матрицу X_k, и вычисляется 12×1-вектор P_k параметров ориентации ТИ относительно ЭИ по формуле (1).

11. Выбирается k^*-й вариант соединения опорных точек с наименьшим значением показателя сопряжения J_{k^*}, после чего последние три элемента вектора P_{k^*} помещаются в базовый вектор b_{k^*} и движение ЛА направляется по вектору $-b_{k^*}$. Затем по истечение заданного периода времени вновь включается работа РЛС и все операции пп. 3 − 11 периодически повторяются.

12. Если условия наблюдения не позволяют использовать оптическую станцию, то работу продолжает РЛС с периодическим отключением, за время которого движение ЛА осуществляется по экстраполированной траектории в направлении вектора $-b_{k^*}$.

Расчетная часть

1. При использовании РЛС, работающей в режиме синтезирования апертуры [8], на земной поверхности формируется сетка с линиями уровня дальности, угла места и азимута. Сферические координаты центров ячеек сетки пересчитываются к прямоугольные координаты системы $O_2X_2Y_2Z_2$ ТИ. На ЭИ также определяются по определенному правилу опорные точки, которые сопрягаются с опорными точками ТИ в системе координат $O_1X_1Y_1Z_1$.

Пусть в *k*-м варианте сопряжения вектор $M_2(j_i)=(x_2(j_i),y_2(j_i),z_2(j_i))^T$, определенный в системе координат $O_2X_2Y_2Z_2$, поставлен в соответствие вектору $M_1(i)=(x_1(i),y_1(i),z_1(i))^T$, определенному в системе координат $O_1X_1Y_1Z_1$. Запишем условие линейной зависимости векторов или связи их координат с точностью до случайного вектора ошибок сопряжения $e_{ki}=(e_x,e_y,e_z)^T$ в системе координат $O_1X_1Y_1Z_1$:

$$M_1(i)-H_kM_2(j_i)-b_k=e_{ki}\ ,\ i=\overline{1,m_k}\,. \tag{3}$$

Критерием правильного сопряжения *i*-й пары векторов в *k*-м варианте возьмем ограничение на квадрат евклидовой нормы вектора ошибок e_{ki} в формуле (2):

$$J_{ki}=\|e_{ki}\|^2=(M_1(i)-H_kM_2(j_i)-b_k)^T(M_1(i)-H_kM_2(j_i)-b_k)\le\gamma\,.$$

Критерием сопряжения m_k пар векторов будет суммарный показатель

$$J_k=\sum_{i=1}^{m_k}J_{ki}\le\gamma_\Sigma\,,$$

где пороги γ и γ_Σ можно выбрать статистически по критерию хи-квадрат.

В показателе J_{ki} присутствуют 9 + 3 = 12 неизвестных параметров ориентации в составе H_k и b_k. Чтобы найти оценки параметров алгебраическим способом, систему уравнений (3) запишем в виде

$$M_1(i)=X_{ki}P_k+e_{ki}\,,\quad i=\overline{1,m_k}\,,\quad m_k>12\,,$$

или в блочно-матричной форме

$$Y_k=X_kP_k+E_k\quad\Leftrightarrow\quad\begin{bmatrix}M_1(1)\\ \dots\\ M_1(m_k)\end{bmatrix}=\begin{bmatrix}X_{k1}\\ \dots\\ X_{k,m_k}\end{bmatrix}\cdot P_k+\begin{bmatrix}e_{k1}\\ \dots\\ e_{k,m_k}\end{bmatrix},$$

где 3×12-матрицы X_{ki} определены выше.

Оценка $\hat{P}_k$ неизвестного 12×1-вектора P_k находится из условия минимума квадрата евклидовой нормы блочного вектора ошибок E_k

$$J_k=\|E\|^2=(Y_k-X_kP_k)^T(Y_k-X_kP_k)\,.$$

Из необходимого условия существования экстремума J_k по P_k получается оценка $\hat{P}_k$ в (1):

$$\frac{dJ_k}{dP_k}=2(Y_k-X_kP_k)^T(-X_k)=O^T\quad\Rightarrow\quad X_k^T(Y_k-X_kP_k)=O\,,\qquad\hat{P}_k=(X_k^TX_k)^{-1}X_k^TY_k\,.$$

Невырожденность матрицы $X_k^T X_k$ обеспечивается алгоритмом выбора пар сопряженных опорных точек, в котором предусматривается линейная независимость строк матрицы X_k при выборе этих точек.

2. При использовании оптической станции в оптическом приемнике пространственное положение точки *M* в системе координат $O_2X_2Y_2Z_2$, совмещенной по центру с центром оптической линзы, отображается через центр линзы в плоскости oxy в виде точки *m* с координатами *x*, *y*, которые находятся известными методами обнаружения и оценивания положения точечных объектов или центра тяжести протяженных объектов в видеокадрах [4].

Из геометрических соображений координаты орта a направления на точку *M* в системе координат $O_2X_2Y_2Z_2$ с учетом фокусного расстояния f линзы находятся как

$$\bar{a} = \overrightarrow{oM} / |\overrightarrow{oM}| = \overrightarrow{oM} / r = (a_x, a_y, a_z), \quad a_x = -x / |\overrightarrow{om}| = -x / c,$$

$$a_y = -y / |\overrightarrow{om}| = -y / c, \qquad a_z = \sqrt{1 - a_x^2 - a_y^2} = f / c, \qquad c = \sqrt{x^2 + y^2 + f^2},$$

или в векторно-матричной форме $a = (a_x, a_y, a_z)^T = (-x, -y, f)^T / \sqrt{x^2 + y^2 + f^2}$.

Для *i*-й пары опорных точек ЭИ и ТИ: $M_1(i)$ и $M_2(j_i)$, поставленных в соответствие друг другу в *k*-м варианте соединения, уравнение связи координат (сопряжения) запишется как

$$M_1(i) - r_2 H_k a_2(j_i) - b_k = e_{ki}, \qquad r_2 a_2(j_i) = M_2(j_i).$$

Оценка дальности $\hat{r}_2$ находится из условия минимума квадрата нормы вектора ошибок сопряжения e_{ki}

$$J_{ki} = \| e_{ki} \|^2 = (M_1(i) - r_2 H_k a_2(j_i) - b_k)^T \, (M_1(i) - r_2 H_k a_2(j_i) - b_k), \tag{4}$$

а именно:

$$\frac{dJ_k}{dr_2} = 2 \, (M_1(i) - r_2 H_k a_2(j_i) - b_k)^T (-H_k a_2(j_i)) = 0,$$

$$a_2^T(j_i) H_k^T (M_1(i) - r_2 H_k a_2(j_i) - b_k) = 0,$$

где учитывается, что для ортогональной матрицы H_k: $H_k^T H_k = I$, I – единичная матрица, и для орта $a_2^T(j_i) a_2(j_i) = 1$. Получается следующая оценка дальности:

$$\hat{r}_2 = a_2^T(j_i) H_k^T (M_1(i) - b_k). \tag{5}$$

Оценка (5) подставляется вместо r_2 в выражение показателя (4), который участвует в сопряжении пар векторов.

Заключение

Предложена концепция построения автономной навигационной системы, отличающаяся коррекцией движения ЛА на основе классификации пар опорных точек ЭИ и ТИ по определенному критерию сопряжения и позволяющая использовать вместо корреляционной функции быстрые матричные операции решения линейных алгебраических уравнений, учитывать пространственную ориентацию ЛА, определять вектор направления движения ЛА к расчетному положению в текущие моменты времени по экстраполированной траектории и совмещать работу бортовой РЛС с работой оптической станции при периодическом отключении РЛС, что обеспечивает ее помехозащищенность.

Дальнейшее исследование направлено на разработку программно-алгоритмического обеспечения для реализации предложенного решения в условиях моделирования автономной навигационной системы.

Решение может найти применение в навигационных системах сопровождения движущихся объектов разной физической природы: летательных, наземных, водных и подводных беспилотных аппаратов.

Список источников

[1] Бакулев П.А., Сосновский А.А. Радионавигационные системы. Учебник для вузов. М.: Радиотехника, 2005. 224 с.

[2] Авиационная радионавигация: Справочник /А.А. Сосновский, И.А. Хаймович, Э.А. Лутин, И. Б. Максимов; Под ред. А. А. Сосновского. М.: Транспорт, 1990. 264 с.

[3] Баклицкий В.К., Бочкарёв А.М. Методы фильтрации сигналов в корреляционно-экстремальных системах навигации. М.: Радио и связь, 1986. 216 с.

[4] Методы автоматического обнаружения и сопровождения объектов. Обработка изображений и управление / Б.А. Алпатов, П.В. Бабаян, О.Е. Балашов, А.И. Степашкин. М.: Радиотехника, 2008. 176 с.

[5] Zhou S.B., Li Y. The SIFT image feature matching based on the Plural Differential // Advanced Materials Research (Volumes 268 – 270). 2011 P. 2172 − 2184.

[6] Bastanlar Y., Temizel A., Yardımcı Y. Improved SIFT Matching for Image Pairs with a Scale Difference // IET Electronics. – Volume 46, Issue 5, 2010. P. 346 − 352.

[7] Wu Z., Wang J., Li C., Yan Y., Chu C. Bidirectional Matching Algorithm for Target Tracking Based on SIFT // Advanced Research on Computer Education, Simulation and Modeling Communications in Computer and Information Science. Volume 175, 2011. P. 253 − 258.

[8] Кондратенков Г.С., Фролов А.Ю. Радиовидение. Радиолокационные системы дистанционного зондирования Земли: учеб. пособие для вузов / под ред. Г.С. Кондратенкова. М.: Радиотехника, 2005. 368 с.

УДК 004.932.4
DOI: 10.25206/978-5-8149-3873-2-2024-538-543

Исследование влияния неравносторонних форм апертуры пятикаскадного рекурсивно-сепарабельного цифрового фильтра на отношение сигнал/шум и быстродействие

Study of the influence of non-equilateral aperture shapes of a five-stage recursive-separable digital filter on the signal-to-noise ratio and performance

А. В. Каменский1, А. Поднебесный1, М. И. Курячий1, К. С. Куприянова1
1 ТУСУР, ул. Ленина 40, г. Томск, 634050, Россия

A.V. Kamenskiy1, A. Podnebesnyy1, M.I. Kuryachiy1, K.S. Kupriyanova1
[1]TUSUR, 40 Lenina Prospect, Tomsk, 634050, Russia

Аннотация. Разрабатывая новые алгоритмы цифровой обработки изображений важно проверять способы их использования и реализации. Разработанный авторами пятикаскадный рекурсивно-сепарабельный цифровой фильтр, является быстродействующим алгоритмом цифровой обработки, за счет использования свойств рекурсии и сепарабельности. В представленном исследовании был выполнена оценка его влияния на отношение сигнал/шум, при изменении формы апертуры обработки за счет корректировки количества строк и столбцов в апертуре обработки. Оценено быстродействие исследуемого фильтра и было показано закономерное изменение времени его работы. Изменение времени работы фильтра варьировалось в диапазоне от 1,063 до 1,373 раз, при росте числа элементов в строке или столбце апертуры обработки. В итоге наблюдался рост ОСШ с 17 дБ до 28 дБ. Полученные результаты позволят заложить основы разрабатываемой методики применения рекурсивно-сепарабельных цифровых фильтров.

***Ключевые слова*:** цифровая обработка изображений, рекурсия, сепарабельность, отношение сигнал/шум, быстродействующие алгоритмы.

Аннотация. When developing new algorithms for digital image processing, it is important to check the methods of their use and implementation. The five-cascade recursive-separable digital filter developed by the authors is a high-speed digital processing algorithm due to the use of recursion and separability properties. In the presented study, its effect on the signal-to-noise ratio was assessed when changing the shape of the processing aperture by adjusting the number of rows and columns in the processing aperture. The performance of the filter under study was assessed and a regular change in its operating time was shown. The change in the filter operating time varied in the range from 1.063 to 1.373 times, with an increase in the number of elements in a row or column of the processing aperture. As a result, an increase in the SNR from 17 dB to 28 dB was observed. The results obtained will allow laying the foundations for the developed methodology for using recursive-separable digital filters.

***Ключевые слова*:** digital image processing, recursion, separability, signal-to-noise ratio, high-speed algorithms.

I. Введение

С развитием оптических технологий возрастают требования к качеству и быстродействию обработки информации. Рост размеров цифровых изображений (ЦИ) приводит к увеличению вычислительных затрат, связанных с их обработкой. В условиях постоянно возрастающих объемов данных, особенно в таких областях как медицина, спутниковая съемка, мониторинг окружающей среды и безопасности, становится критически важным улучшить эффективность обработки ЦИ и минимизировать количество необходимых вычислительных операций.

Трудностью, с которой сталкиваются при решении задач восстановления изображений, является проблема их размывания и смазывания [1, 2], она возникает при взаимном относительном движении оптического сенсора и объекта во время съемки. Размывание (аналог расфокусировки изображения), возникает, когда сцена и регистрирующее устройство находятся в покое по отношению друг к другу, и его можно смоделировать низкочастотными фильтрами в пространственной или частотной областях. Другая важная модель искажения изображений смазыванием соответствует равномерному прямолинейному перемещению сцены относительно регистрирующих устройств и датчиков в процессе фиксирующей съемки [3-5].

Одним из перспективных решений данной задачи является разработка алгоритма, который оптимизирует процесс путем создания неравносторонних апертур для реализации цифровых фильтров. Этот подход позволяет

значительно снизить вычислительные затраты за счет более рационального использования ресурсов и повышения точности обработки изображений. В результате можно добиться улучшения качества конечных изображений и ускорения процессов их обработки, что особенно важно для систем реального времени. Коллективом автором разработаны различные виды рекурсивно-сепарабельных цифровых фильтров, которые различаются количеством каскадов (связок строчных и кадровых рециркуляторов), а также формой итоговой апертуры обработки. В работе выполненное исследование одного из этих фильтров, а именно пятикаскадного рекурсивно-сепарабельного (ПРС) цифрового фильтра.

II. Постановка задачи

Работа посвящена разработке основ методики обработки цифровых изображений пятикаскадным рекурсивно-сепарабельным цифровым фильтром. Разрабатываемая методика должна учитывать в себе способность выбранной апертуры повысить отношение сигнал/шум (ОСШ) обрабатываемого изображения, при этом неравносторонняя форма апертуры должна сокращать время их обработки.

Для достижения поставленной цели необходимо решить ряд задач.

1. Определить оптимальные коэффициенты фильтрации для апертур 7x7, 9x9 и 11x11 элементов.
2. Выяснить как изменится ОСШ при фиксированном оптимальном коэффициенте фильтрации и изменении формы апертуры обработки ПРС цифрового фильтра.
3. Оценить быстродействие ПРС фильтра.

III. Исходные данные

1. Цифровой ПРС фильтр

Цифровой ПРС фильтр классифицируется как матричный фильтр. Благодаря использованию свойства рекурсии в процессе построении, этот фильтр использует результаты предыдущих вычислений для расчета последующих шагов, что значительно уменьшает необходимое количество математических операций для создания итоговой маски. Такой метод позволяет достичь более высокого быстродействия по сравнению с аналогичными фильтрами, которые не применяют рекурсивный подход [6].

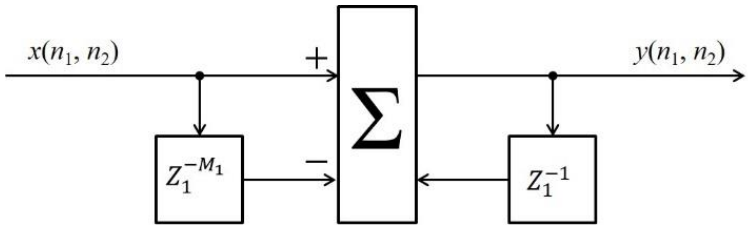
При разработке рекурсивно-сепарабельного алгоритма используются рециркуляторы (строчные и кадровые). На рис. 1 а) представлена образующая рекурсивная ячейка по n_1-строке (строчный рециркулятор – СР), а на рис. 1 б) образующая рекурсивная ячейка по n_2-кадру (кадровый рециркулятор – КР), реализующие соответствующие ортогональные направления обработки “скользящее среднее”. При этом СР – осуществляет обработку по строке матрицы входных данных, а КР – по столбцу данной матрицы [7], где: $x(n_1,n_2)$ – входные данные, $y(n_1,n_2)$ – выходные данные, $n_{1,2}$ – текущий номер отсчета сигнала, включая нулевой отсчет ($n_{1,2}$ = 0, 1, ...), M_1 – направление по строкам, M_2 – направление по кадру, Z – элементы задержки, $\sum$ - суммирование.

За счет использования блоков СР и КР рециркуляторов происходит формирование апертуры обработки ПРС фильтра, структурная схема которого представлена на рис. 2, где: L – коэффициент для свертки дополнительной маски, a – коэффициент сдвига дополнительной маски, $m_{1,2,3,4,5}$ – коэффициенты СР и КР для дополнительной маски, $n_{1,2,3,4,5}$ – коэффициенты СР и КР для основной маски, c – коэффициент сдвига для маски увеличения центрального элемента, A_1 – коэффициент подъема центральной апертуры, A_2 – коэффициент увеличения центрального элемента.

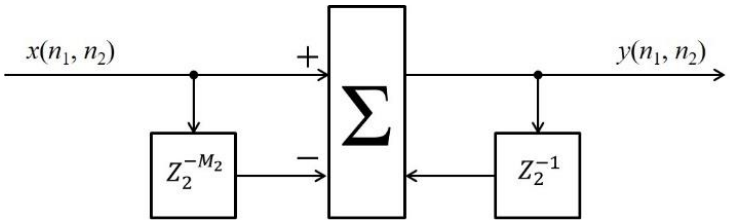
Первая ветвь фильтра формирует вторую маску той же размерности, что и внутренняя маска с нижней ветви. При этом имеет свой определенный коэффициент L, он будет подаваться на эту ветвь и с ним будет происходить свертка строчными и кадровыми рециркуляторами.

Во второй ветви формируется маска заданной размерности и соответствующих коэффициентов строчных и кадровых рециркуляторов. Данная ветвь основная и на сумматор подается с отрицательным знаком.

Полученная в первой ветви маска суммируется с маской со второй ветви, но результат не будет иметь сумму элементов равной нулю. Для этого необходима третья ветвь, задача у данной ветви аналогичная что и у первой ветви исходного фильтра, данная ветвь формирует однородную маску размером 3x3, при этом маска заполняется определенным числом. Эти значения также должны быть рассчитаны в зависимости от недостающей разницы до суммы равной нулю. Когда заданный размер маски предполагает увеличение центральной маски, происходит и увеличение коэффициента формируемой в верхней ветви маски [8, 9].



а)



б)

Рис. 1. Строчный (а) и кадровый рециркуляторы (б)

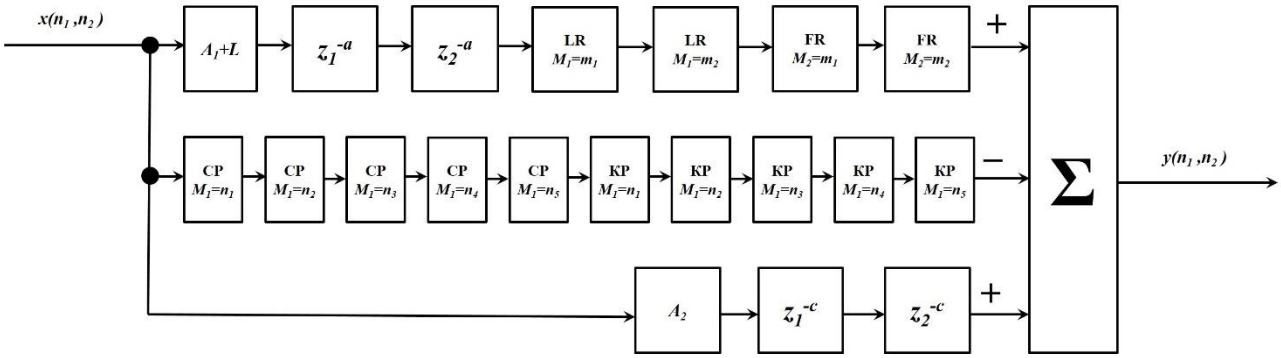


Рис. 2. Структурная схема цифрового ПРС фильтра

2. Отношение сигнал/шум

Оценка ОСШ проводилась при помощи программного обеспечения (ПО) ImageJ. ImageJ – открытое ПО для обработки и анализа цифровых изображений. В ПО ImageJ на тестовом изображении выбиралась область, по которой строилась гистограмма. По полученной гистограмме ПО ImageJ определяло данные о среднем значении яркости и среднеквадратическом отклонении.

$$\text{ОСШ} = 20\,lg\frac{M}{\sigma},$$

где M – средняя яркость изображения,

σ – среднеквадратическое отклонение.

3. Тестовое изображение

В качестве тестового изображения (рис. 3) для проведения исследования было использовано изображение (1032x772 элементов), сформированное активно-импульсной телевизионно-измерительной системой [10]. Измерение ОСШ проходил на белой области тестового объекта.

Рис. 3. Тестовое изображение

4. Параметры вычислительной системы

Для выполнения эксперимента использовалась вычислительная платформа, со следующими характеристиками: процессор – Intel(R) Core(TM) i7-10750H CPU 2,59 ГГц; оперативная память 16 ГБ; видеокарта GTX1650; операционная система – Windows 11Pro; количество ядер – 6; тип системы - 64-разрядная операционная система.

IV. Результаты экспериментальных исследований

1. Определение оптимальных коэффициентов фильтрации

В рамках определения оптимального коэффициента (A_1) фильтрации ПРС фильтра, будут использоваться 3 размера (7x7, 9x9, 11x11 элементов) базовой апертуры обработки и при обработке будет изменяться значение коэффициента фильтрации A_1. В Табл. 1 представлены результаты измерения ОСШ для 3-х апертур обработки ПРС фильтра. ОСШ исходного изображения равно 17 дБ.

ТАБЛИЦА 1
РЕЗУЛЬТАТЫ ИЗМЕРЕНИЯ ОСШ НА ТЕСТОВОМ ИЗОБРАЖЕНИИ

Апертура 7x7 элементов		**Апертура 9x9 элементов**		**Апертура 11x11 элементов**	
Коэффициент фильтрации A_1	ОСШ, дБ	Коэффициент фильтрации A_1	ОСШ, дБ	Коэффициент фильтрации A_1	ОСШ, дБ
1	16	1	3	1	5
5	19	5	13	5	14
10	21	10	11	10	14
15	**28**	15	13	15	15
20	25	20	18	20	16
25	23	25	16	25	17
49	24	49	22	49	15
75	13	**75**	**24**	**75**	**20**
99	15	99	23	99	17

2. Выявление изменений в значениях ОСШ при изменении размеров и формы апертуры обработки

В эксперименте производилось изменение размеров апертуры обработки по строкам и столбцам, то есть относительно базовых размеров (7x7, 9x9 11x11 элементов) и происходило уменьшение числа элементов либо в строке, либо в столбце матрицы значений фильтра. Определенный ранее оптимальный коэффициент фильтрации оставался неизменным и в ходе эксперимента было оценено ОСШ при изменении размеров и формы апертуры обработки. Полученные результаты представлены в Табл. 2.

ТАБЛИЦА 2
РЕЗУЛЬТАТЫ ИЗМЕРЕНИЯ ОСШ ПРИ ИЗМЕНЕНИИ РАЗМЕРОВ И ФОРМЫ АПЕРТУРЫ ОБРАБОТКИ

Размер апертуры обработки	ОСШ, дБ	Размер апертуры обработки	ОСШ, дБ	Размер апертуры обработки	ОСШ, дБ
17x7	2	19x9	2	19x11	2
15x7	7	17x9	6	17x11	4
13x7	16	15x9	12	15x11	8
11x7	22	13x9	18	13x11	11
9x7	26	11x9	21	**11x11**	**20**
7x7	**28**	**9x9**	**24**	9x11	15
5x7	27	7x9	22	7x11	15
7x5	26	9x7	22	11x7	17
7x7	**28**	**9x9**	**24**	11x9	18
7x9	26	9x11	21	**11x11**	**20**
7x11	21	9x13	17	11x13	13
7x13	15	9x15	12	11x15	10
7x15	7	9x17	6	11x17	6
7x17	2	9x19	2	11x19	2

3. Оценка быстродействия ПРС фильтра

В исследовании быстродействия использовались апертуры, которыми осуществлялась обработка тестового изображения. Для каждого размера апертуры обработки проводилось 10 измерений времени работы фильтра, после чего вычислялось среднее время работы фильтра. Полученные результаты оценки быстродействия ПРС фильтра представлены в Табл. 3.

ТАБЛИЦА 3
РЕЗУЛЬТАТЫ ИЗМЕРЕНИЯ ВРЕМЕНИ РАБОТЫ ПРС ФИЛЬТРА

Размер апертуры обработки	Время обработки, с	Размер апертуры обработки	Время обработки, с	Размер апертуры обработки	Время обработки, с
17x7	0,8686	19x9	1,1250	19x11	1,4798
15x7	0,7579	17x9	0,9906	17x11	1,3259
13x7	0,6763	15x9	0,8737	15x11	1,1755
11x7	0,6365	13x9	0,7697	13x11	1,0325
9x7	0,5215	11x9	0,6862	11x11	0,8870
7x7	0,4273	9x9	0,5670	9x11	0,7035
5x7	0,3113	7x9	0,4557	7x11	0,5632
7x5	0,3424	9x7	0,4426	11x7	0,5421
7x7	0,4273	9x9	0,5670	11x9	0,6526
7x9	0,4942	9x11	0,6497	11x11	0,8870
7x11	0,5537	9x13	0,7640	11x13	1,0212
7x13	0,6151	9x15	0,8512	11x15	1,1799
7x15	0,7139	9x17	0,9697	11x17	1,3282
7x17	0,7909	9x19	1,0510	11x19	1,5051

V. Обсуждение результатов

Анализируя результаты, представленные в Табл. 1, можно сделать вывод о том, что применение ПРС фильтра для повышения ОСШ целесообразно, поскольку удается добиться его повышения, однако рост ОСШ зависит от размеров апертуры обработки. Лучше всего себя показала апертура 7x7 элементов (ОСШ=9,68 дБ), также данный результат был получен при наименьшем значении коэффициента фильтрации (A_1=15). В то время как два других (9x9 и 11x11 элементов) вида базовой апертуры показали худшие результаты по повышению ОСШ, и они были достигнуты при больших значениях коэффициента фильтрации (A_1=75).

Из Табл. 2 следует, что при изменении размеров апертуры обработки за счет корректировки количества элементов в столбце и строке матрицы апертуры обработки, при сохранении определенного ранее оптимального коэффициента фильтрации приводит к уменьшению значений ОСШ от значений полученных при использовании базовых размеров апертур обработки. Уменьшение ОСШ свидетельствует о том, что ближайшие по размерности апертуры обработки к базовой (равносторонней) апертуре показывают незначительное снижение ОСШ. В тоже время при увеличении числа элементов в строке или столбце матрицы апертуры обработки, уменьшается ОСШ и это означает, что необходимо сделать выбор оптимального коэффициента обработки.

Результаты оценки времени обработки различными апертурами ПРС фильтром представленные в Табл. 3 показывают закономерные результаты. С ростом числа элементов в апертуре обработки время работы фильтра увеличивается и наоборот с уменьшением числа элементов время работы уменьшается.

VI. Выводы и заключение

Представленные результаты позволяют заложить основы разрабатываемой методики обработки изображения рекурсивно-сепарабельными цифровыми фильтрами. Использование ПРС фильтра для повышения ОСШ целесообразно и показывает хорошие результаты. Незначительно снижение значений ОСШ при уменьшении размеров апертуры позволяет использовать для обработки апертуры меньшей размерности, тем самым сокращая время обработки, что подтверждают и результаты оценки времени работы ПРС фильтра.

Список литературы

1. Коробейников А. Г., Федосовский М. Е., Алексанин С. А. Разработка автоматизированной процедуры для решения задачи восстановления смазанных цифровых изображений // Кибернетика и программирование. 2016. № 1. С. 115–119.

2. Sun M. J., Edgar M. P., Phillips D. B., Gibson G. M. & Padgett M. J. Improving the signal-to-noise ratio of single-pixel imaging using digital microscanning // Optics express. – 2016. – Vol. 24. – No. 10. – P. 10476-10485.

3. Трушин А. А. Повышение качества цифровых изображений, искаженных в процессе фотосъемки // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. – 2020. – №. 3. – С. 111-122.

4. Волков В. Ю. Различение объектов на цифровых изображениях // Международная конференция по мягким вычислениям и измерениям Учредители: Санкт-Петербургский государственный электротехнический университет" ЛЭТИ" им. ВИ Ульянова (Ленина). – 2022. – Т. 1. – С. 218-221.

5. Chambolle A., De Vore R. A., Lee N. Y., & Lucier B. J. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage // IEEE Transactions on image processing. – 1998. – Vol. 7. – No. 3. – P. 319-335.

6. Burger W., Burge M. J. Digital image processing: An algorithmic introduction. – Springer Nature, – 2022. – P. 944.

7. Kamenskiy A. V. High-speed recursive-separable image processing filters // Computer Optics. – 2022. – Vol. 46. – No. 4. – P. 659-665. DOI: 10.18287/2412-6179-CO-1063.

8. Комарницкий М. А. Разработка двумерного рекурсивно-сепарабельного пятикаскадного фильтра повышения четкости изображений // Материалы Международной научно-технической конференции студентов, аспирантов и молодых ученых «Научная сессия ТУСУР–2022» – Томск, 2022. – Ч. 1. – С. 96–99.

9. Поднебесный А. В. Оценка быстродействия равносторонней и неравносторонней формы построения цифрового пятикаскадного рекурсивно-сепарабельного фильтра обработки изображений // Электронные средства и системы управления: материалы докладов XIX Международной научнопрактической конференции (15–17 ноября 2023 г.): в 2 ч. – Ч. 2. – Томск: В-Спектр (ИП Бочкарева В.М.), – 2023. – 366 с.

10. Kapustin V. V., Zahlebin A. S., Movchan A. K., Kuriyachiy M. I., Kruticov M. Experimental assessment of the distance measurement accuracy using the active-pulse television measuring system and a digital terrain model // Computer Optics. – 2022. – Vol. 46. – No. 6. – P. 948-954. DOI: 10.18287/2412-6179-CO-1114.

УДК 004.93
DOI: 10.25206/978-5-8149-3873-2-2024-544-550

Разработка системы для распознавания людей по изображениям земной поверхности
Development of a system for recognizing people from images of the earth's surface

А.А. Залищук[1], В.А. Ненашев[2]
[1-2]Санкт-Петербургский государственный университет аэрокосмического приборостроения
Санкт-Петербург, Россия

A. A. Zalishchuk, V.A. Nenashev
Saint Petersburg State University of Aerospace Instrumentation Saint Petersburg, Russia

Аннотация. В задачах распознавания и выделения объектов на изображениях земной поверхности признаки имеют высокую вариативность ракурсов и оттенков, что затрудняет применение явных алгоритмов. Нейронные сети обладают свойством контекстуальной обработки информации с помощью обучения на примерах (case-based) и реализуют один из подходов к решению неформализуемых или трудноформализуемых задач, связанных с обработкой изображений, в связи с чем тема работы имеет высокий уровень актуальности. Целью работы является разработка системы распознавания людей на изображениях земной поверхности с использованием технологий машинного обучения и технического зрения. На основе сравнительного анализа вариантов реализации осуществлен выбор и обоснование применяемой модели нейронной сети. При реализации и обучении выбранной модели нейронной сети «YOLOv8» выполнена оптимизация с помощью метода срезов «SAHI», разработанного специально для обнаружения небольших объектов и точной настройки.

***Ключевые слова*:** нейронные сети, машинное обучение, распознавание объектов, обработка изображений, техническое зрение

Annotation. *In tasks of recognizing and identifying objects in images of the earth's surface, features have a high variability of angles and shades, which makes it difficult to use explicit algorithms. Neural networks have the property of contextual information processing using case-based learning and implement one of the approaches to solving unformalizable or difficult-to-formalize problems associated with image processing, and therefore the topic of the work has a high level of relevance. The goal of the work is to develop a system for recognizing people in images of the earth's surface using machine learning and technical vision technologies. Based on a comparative analysis of implementation options, the selection and justification of the neural network model used was made. When implementing and training the selected "YOLOv8" neural network model, optimization was performed using the "SAHI" slicing method, developed specifically for small object detection and fine tuning.*

***Keywords*:** *neural networks, machine learning, object recognition, image processing, computer vision.*

I. Введение

Технологии распознавания образов являются одним из актуальных направлений исследований при разработке нейросетевых методов обработки данных. Такие технологии применяются в ряде областей и сферы их использования охватывают большой спектр, в том числе: обнаружение возгораний по спутниковым снимкам, обнаружение дефектов железнодорожных путей, обнаружение заболеваний по медицинским изображениям, определения состояния металла по металлографическим изображениям, распознавание лиц в системах контроля и наблюдения, определение видов заканчивающихся товаров в магазине и многое другое [1].

Для обучения машины видению и пониманию объекта на изображении используются либо методы технического зрения, либо технологии машинного обучения. При использовании нейронных сетей требуется большое количество вычислительных ресурсов для выделения информации об объектах, а также наличие значительного объема визуальных предварительных данных об объекте для проверки ранее выделенной информации о том же объекте.

Классический подход к решению задачи распознавания образов состоит в выделении формализованных признаков в явном виде и представляет из себя ограниченный круг алгоритмов и методов. В задачах распознавания и выделения объектов на изображениях земной поверхности признаки имеют высокую вариативность ракурсов и оттенков, что затрудняет применение явных алгоритмов.

С другой стороны, задача распознавания образов становится более ориентированной под конкретные практические решения. В сфере видеонаблюдения и обработки изображений распознавание выполняется на

основе анализа объектов интереса. Затем искусственный интеллект с помощью машинного обучения распознает действия и классифицирует их.

Целью работы является разработка системы распознавания людей на изображениях земной поверхности с использованием технологий машинного обучения и технического зрения. Работа посвящена исследованию алгоритмов, применяющих сверточные нейронные сети. На основе этого исследования требуется выбрать модель нейронной сети, а также методы оптимизации для автоматического обнаружения и выделения людей на оптических изображениях земной поверхности.

II. Постановка задачи

Основная цель данной работы является разработка данной системы, которая будет эффективным инструментом для автоматизации процесса распознавания объектов на изображениях земной поверхности [2, 3]. Система должна будет обеспечивать высокую точность распознавания и возможность работы с большим объемом информации.

Для достижения этой цели необходимо решить ряд задач.

На первом этапе следует выбрать оптимальные и современные методы и алгоритмы для разрабатываемой системы, ведь новейшие модели лишены ошибок старых разработок и эффективнее выделяя информацию об объектах интереса.

Вторым этапом будет являться создание базы данных признаков, которая будет использоваться для обучения и тестирования выбранной модели машинного обучения.

Третьим этапом является обучение и тестирование нейронной сети, выполняющей задачу распознавания объектов (людей) на изображениях. Также следует интегрировать систему в пользовательский интерфейс для удобства использования и обеспечить безопасность хранения и передачи данных пользователей.

И четвертым, заключительным этапом, является тестирование работы разработанной системы на отобранных (валидационных) данных.

III. Теория

3. Выбор модели нейронной сети

Архитектура «YOLO» является одной из наиболее популярных и быстроразвивающихся для реализации одностадийных детекторов для анализа изображений. В работе для реализации системы была выбрана модель архитектуры «YOLOv8» как одна из последних в семействе [4].

Архитектура «YOLOv8» оптимизирована под быстрое и точное решение задач компьютерного зрения таких как обнаружение и классификация объектов интереса на изображениях и видео. Модели этой архитектуры способны работать на различных устройствах, включая мобильные телефоны и компьютеры.

Важной особенностью «YOLOv8» является его способность эффективно учиться на больших объемах данных, сохраняя при этом высокую точность обнаружения объектов. По сравнению с предыдущими версиями «YOLOv5» и «YOLOv7» в выбранной архитектуре увеличено количество внутренних слоев (274 против 224 и 255 соответственно), а также расширено количество обучаемых параметров (11,2 млн. против 7,2 млн. и 9,6 млн. соответственно). Это делает «YOLOv8» отличным инструментом для обработки огромного количества информации в различных областях, включая транспорт, медицину, науку и многие другие сферы.

На рисунке 1 показано сравнение работы четырех разных версий архитектуры «YOLO» на датасете «MS COCO». По графику зависимости точности распознавания от размера используемой модели на левой части рисунка видно, что модели архитектуры восьмой версии определяют объекты в датасете точнее, чем модели предыдущих архитектур (показатель средней точности) при меньшем размере самих моделей. В соответствии к левой части рисунка приводится график скорости обработки изображений (миллисекунда/картинка) в зависимости от размера используемой модели на правой части, который показывает преимущество восьмой версии над предыдущими.

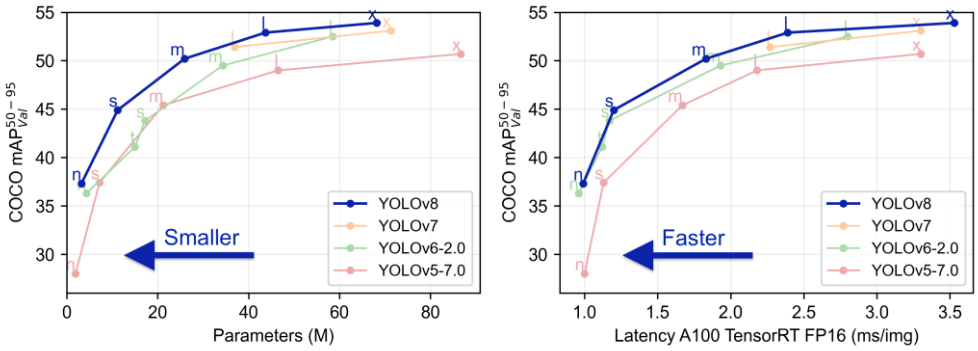


Рис. 1. Сравнение «YOLOv8» с предыдущими моделями

Из этого следует вывод, что модель архитектуры «YOLOv8» является эталонной («SOTA» – state of the art) [5] для обнаружения объектов и в дальнейшем модель архитектуры этой версии будет использоваться в системе.

4. Формирование датасета

Так как применяемый метод распознавания объектов относится к категории методов обучения с учителем, требуется предоставить модели данные о местоположении ограничивающих рамок людей на изображении. Нужно собрать обширный набор изображений, которые будут использоваться для обучения и проверки модели.

Были подготовлены изображения для обучения модели, т.е. были приведены к одному размеру и масштабированы. Для этого необходимо провести аннотирование изображений. Для выполнения этой задачи использовалась платформа «Roboflow» [6], которая позволяет не только размечать изображения, но также позволяет экспортировать аннотированные наборы данных в различных форматах.

Кроме того, «Roboflow» предлагает широкий выбор различных датасетов для каждой задачи в области компьютерного зрения. Поэтому было решено найти и объединить несколько уже аннотированных датасетов с оптическими снимками земной поверхности с людьми.

Первым из выбранных датасетов является «LADD dataset» [7], который содержит аннотированные изображения из когда-то ранее собранной информации для проекта «LACMUS», который занимался разработкой программы для поиска потерявшихся людей в лесу на фотографиях с воздуха с помощью нейросети. Данный датасет был выбран за большого количество изображений с углом обзора вида «птичьего полета» над человеком. Пример из датасета представлен на рис.2.

Рис.2. Пример аннотированного изображения из датасета «LADD dataset»

Вторым стал датасет под именованием «ROGER-VISION» [8], который был найден на платформе при переборе популярных датасетов и очень схож по формату с ранее найденным «LADD dataset». За счет найденного датасета происходит увеличение разнообразия обучающих изображений аннотациями с другими масштабами, увеличивая точность нахождения человека на земной поверхности как под другими ракурсами, так и по другим размерам ограничительных окон. Пример аннотированного изображения представлен на рис. 3.

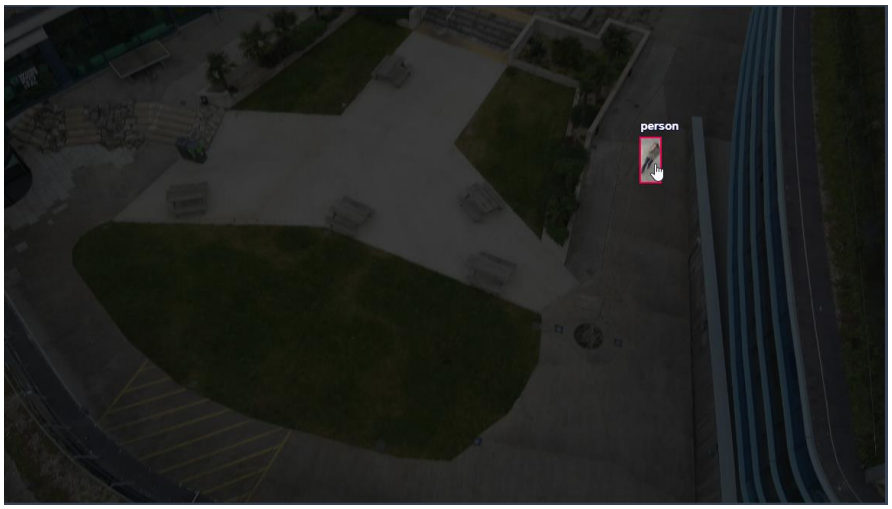


Рис. 3. Пример аннотированного изображения из датасета «ROGER-VISION»

И третьим, последним и дополняющим датасетом, был выбран «internal-flight» [9]. По анализу изображений с датасета (рис. 4) можно говорить о том, что он схож с предыдущими двумя и был сделан с дрона. Этот датасет стал частью финальной выборки за счет наличия изображений с множеством объектов интереса («людей») в одном кадре.

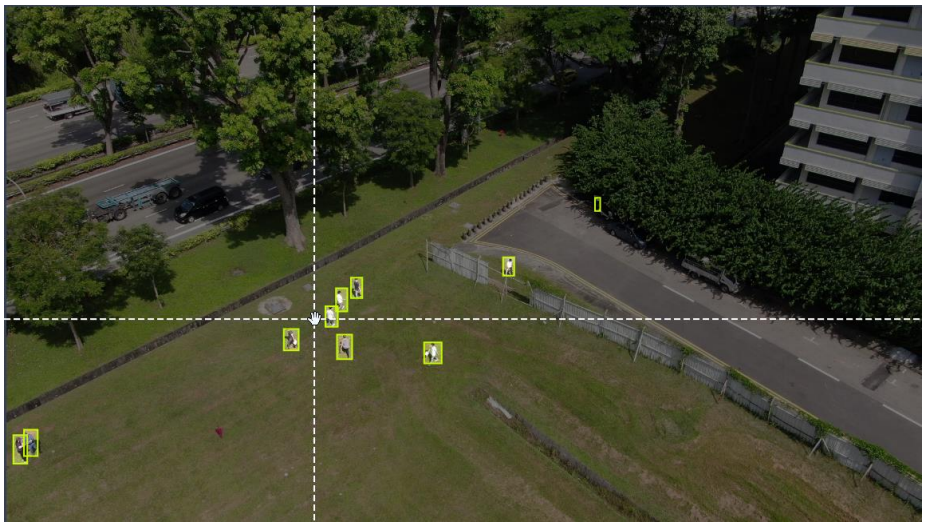


Рис. 4. Пример аннотированного изображения из датасета «internal-flight»

Ранее созданная разметка необходима для предоставления информации об изображении нейронной сети. Процесс обучения и оценка качества обученной модели зависят от того, насколько хорошо была выполнена разметка, сколько данных имеется и как они разделены на «тренировочную» (train), «валидационную» (valid) и «тестовую» (test) выборки.

5. Обучение и тестирование нейронной сети

Для достижения наилучшего баланса классов и предотвращения переобучения модели данные были разделены на тренировочный, валидационный и тестовый наборы. Тренировочный набор данных (80%), включающий 5916 изображений, использовался для обучения модели. Тестовый набор данных (15%), состоящий из 696 изображений, применялся для оценки производительности модели в процессе обучения. Валидационный

набор данных (5%), включающий 348 изображений, использовался для окончательной оценки производительности модели после завершения обучения [10].

Обучение модели «YOLOv8» проводилось в течение 50 эпох с использованием входных изображений фиксированного размера 640x640 пикселей. Для обучения использовался оптимизатор «Adam» с параметрами «learning_rate» = 0,001 (скорость обучения) и коэффициентом регуляризации 0,1.

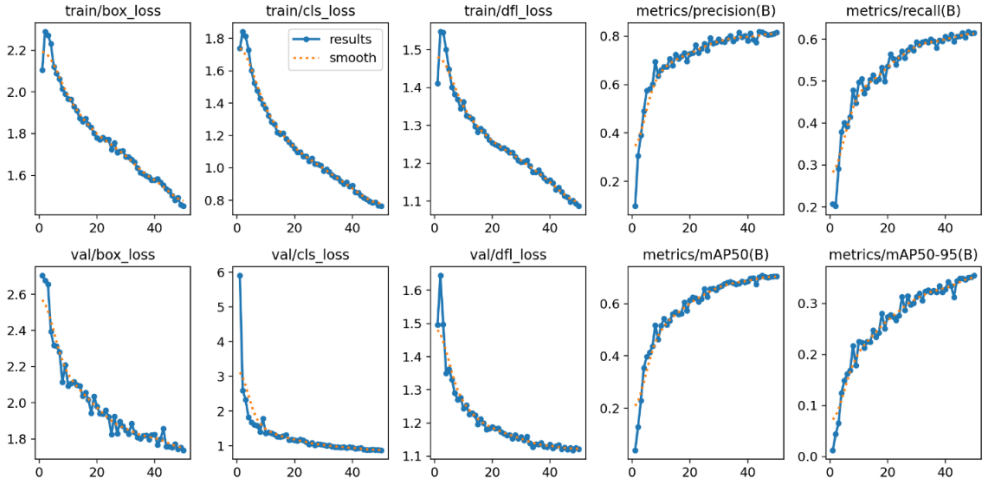


Рис. 5. Графики обучения системы с определением её точности

Анализ кривых «box_loss», «cls_loss», «precision», «recall», «mAP50» и «mAP50-95» позволил определить оптимальное количество эпох с учетом, поданных тестовых и валидационных данных, при которых модели достигают наилучшего баланса между точностью и полнотой.

IV. Результаты экспериментов

По результатам тестирования системы было выделено, что обученная модель распознает людей с вероятностью от 50 до 68 процентов, что является весьма хорошим результатом для выбранной модели «YOLOv8».

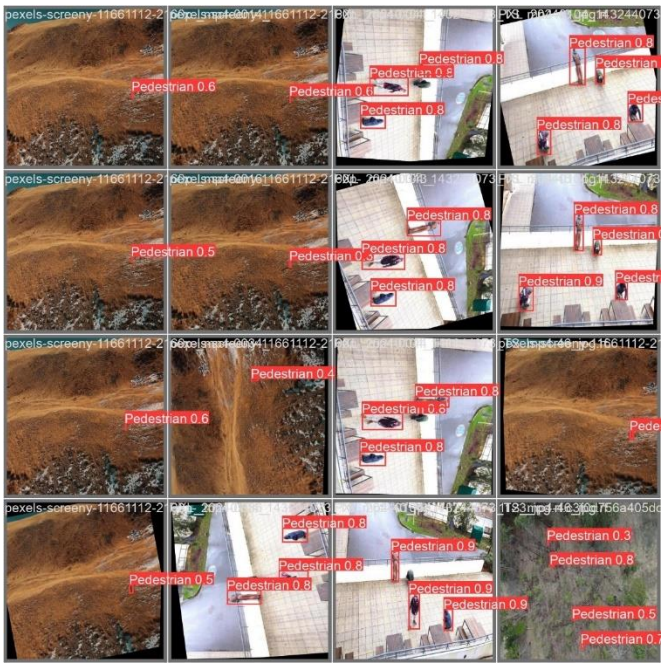


Рис. 7. Предсказания нейронной сети на изображениях валидационной выборки

Для получения результатов эксперимента была создана система с выделением ограничительных рамок вокруг объектов интереса (людей) с точностью классификации, которая обученная модель могла передать системе. Однако у системы осталась проблема с определением объектов небольшого размера, которая характерна для обученной модели архитектуры «YOLOv8». Для решения проблемы семейства моделей «YOLO» [11] с

определением малых объектов интереса на оптических изображениях в работе было решено улучшить систему при помощи метода «SAHI» [12]. «SAHI (Slicing Aided Hyper Inference)» - конвейер, разработанной специально для обнаружения небольших объектов. «SAHI» использует возможности методов вывода и точной настройки с помощью срезов.

а) Ultralytics *б) SAHI*

Рис. 8. Работа распознавания объектов на изображении через запуск в «Ultralytics» (а) и «SAHI» (б)

Следует отметить, что с запуском через «Ultralytics» теряется часть объектов из-за малого размера (рисунок 8а). Поэтому стоит обратить внимание на аналогичный запуск через «SAHI» - улучшается не только точность определения, но и появляются объекты, которые раньше не находились (рисунок 8б). Однако, из-за ограничений обученной модели в 68% вероятности для обнаружения объектов, система может не находить человека в кадре, что и видно на рисунке 8а и 8б.

V. Обсуждение результатов

В статье была показана обработка изображений при помощи системы, способной распознавать людей на изображениях земной поверхности с вероятностью 68% (рисунок 7 и рисунок 8). Полученные результаты системы могут быть полезными в различных сферах, включая безопасность, мониторинг территорий, поиск и спасение людей. К примеру, данная система может быть использована для автоматического обнаружения людей в районах бедствий или для контроля доступа на охраняемых объектах.

Также за счет введения в систему метода «SAHI» была улучшена возможность обнаружения объектов интереса при малом масштабе. Но следует уточнить, что эта возможность ограничена рамками обученной модели, которые выражены в вероятности 68% обнаружения.

VI. Выводы и заключение

Результатом работы стало создание системы для обработки видео и изображений, предназначенной для обнаружения и классификации людей на оптических снимках земной поверхности, полученных с борта малого летательного аппарата. Полученная система является эффективным инструментом для автоматизации процесса распознавания объектов на изображениях земной поверхности.

Созданная система способна распознавать людей на аэрофотоснимках, обеспечивая высокую точность распознавания и возможность работы с большим объемом информации. Система обеспечивает определение человека на различных типах земной поверхности как на видео, так и на аэрофотоснимках.

Источник финансирования

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации, соглашение No FSRF-2023-0003, «Фундаментальные основы построения помехозащищенных систем космической и спутниковой связи, относительной навигации, технического зрения и аэрокосмического мониторинга».

Список литературы

1. Система компьютерного зрения движущегося воздушного объекта: [Электронный ресурс] // Обработка изображений, распознавание образов. Титов И.О., Емельянов Г.М. URL: https://cyberleninka.ru/article/n/sistema-kompyuternogo-zreniya-dvizhuschegosya-vozdushnogo-obekta/viewer (дата обращения: 25.03.2024).

2. Nenashev V. A., Shepeta A.P., Kryachko A.F. Fusion radar and optical information in multiposition on-board location systems // 2020 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2020, Saint-Petersburg, 01–05 июня 2020 года. – Saint-Petersburg, 2020. – P. 9131451.

3. Nenashev V.A., Khanykov I.G. Formation of fused images of the land surface from radar and optical images in spatially distributed on-board operational monitoring systems // Journal of Imaging. – 2021. – Vol. 7, No. 12. – DOI 10.3390/jimaging7120251

4. Репозиторий модели YOLOv8 // Github URL: https://github.com/ultralytics/ultralytics (дата обращения: 12.03.2024).

5. Дьяченко А.А., Гущина О.М. Анализ этапов развития одноступенчатых детекторов объектов на основе YOLO // Информационные технологии в моделировании и управлении: подходы, методы, решения. 2022. С. 427-435.

6. Платформа Roboflow // Roboflow URL: https://roboflow.com/ (дата обращения: 01.02.2024).

7. Датасет LADD // Roboflow URL: https://universe.roboflow.com/danil-cxed4/ladd-dataset (дата обращения: 10.03.2024).

8. Датасет ROGER-VISION // Roboflow URL: https://universe.roboflow.com/detection-kzvbv/roger-vision (дата обращения: 15.03.2024).

9. Датасет internal-flight // Roboflow URL: https://universe.roboflow.com/werkzeug-simu-nav9w/internal-flight (дата обращения: 17.03.2024).

10. Шестопалов Р.П. О выборе предпочтительного алгоритма подстройки весов при обучении нейросетевых алгоритмов обработки информации // Информационно-экономические аспекты стандартизации и технического регулирования. 2021. № 2 (60). С. 90-101.

11. Redmon J. You Only Look Once: Unified, Real-Time Object Detection / J. Redmon, S. Divvala, R. Girshick et al. // arXiv:1506.02640v1. – 2015. – 10 p.

12. Репозиторий метода SAHI // Github URL: https://github.com/obss/sahi (дата обращения: 20.03.2024).

Секция 4

СИСТЕМЫ АВТОМАТИЗАЦИИ ПРОЕКТИРОВАНИЯ — РАЗРАБОТКА, ВНЕДРЕНИЕ И ПРИОРИТЕТЫ РАЗВИТИЯ

Design Automation Systems — Development, Implementation and Development Priorities

GraphiCon 2024
17 − 19 сентября 2024
Омск, Россия

УДК 004.896
DOI: 10.25206/978-5-8149-3873-2-2024-552-563

Технология адаптивного пользовательского интерфейса САПР
CAD adaptive user interface technology

В.И. Ерофеев[1], Д.Д. Куликов[2]
[1] Институт проблем машиностроения РАН – филиал ФГБНУ «Федеральный исследовательский центр Институт прикладной физики им. А.В. Гапонова – Грехова Российской академии наук», Нижний Новгород, Российская Федерация
[2] Институт проблем машиностроения РАН – филиал ФГБНУ «Федеральный исследовательский центр Институт прикладной физики им. А.В. Гапонова – Грехова Российской академии наук», Нижний Новгород, Российская Федерация

V.I. Erofeev[1], D.D. Kulikov[2]
[1] Mechanical Engineering Research Institute of the RAS – Branch of "Federal Research Center A. V. Gaponov-Grekhov Institute of Applied Physics of the Russian Academy of Sciences", Nizhny Novgorod, Russian Federation
[2] Mechanical Engineering Research Institute of the RAS – Branch of "Federal Research Center A. V. Gaponov-Grekhov Institute of Applied Physics of the Russian Academy of Sciences", Nizhny Novgorod, Russian Federation

Аннотация. Эффективность работы проектировщиков в САПР во многом определяется удобством графического пользовательского интерфейса (GUI). Современные САПР предоставляют доступ к тысячам команд. Часть команд доступна в GUI в виде графических элементов. Часть редко используемых команд остаётся доступной только в режиме ручного ввода для опытных пользователей, а количество отображаемых элементов измеряется сотнями. Целью работы является реализация технологии формирования списка команд, максимально соответствующих текущему контексту и накопленному опыту работы пользователя САПР. В разработанном решении использованы современные методы машинного обучения (ML). Создана библиотека программного обеспечения, реализующая технологию. Библиотека протестирована в интеграции с несколькими современными САПР. Разработанная технология увеличивает эффективность работы проектировщиков как за счет их собственного опыта, так и за счет трансфера знаний в виде моделей машинного обучения от более опытных инженеров к менее опытным. В рамках разработки технологии был собран набор данных, имеющий перспективы использования при решении других актуальных проблем САПР с помощью методов ML. Технология имеет перспективы использования в других типах приложений.

***Ключевые слова*:** адаптивный интерфейс пользователя, предиктивный интерфейс пользователя, контекстно-информированный интерфейс пользователя, интерфейс пользователя САПР, искусственный интеллект, машинное обучение, нейронная сеть.

Аннотация. The efficiency of CAD designers is largely determined by the convenience of the graphical user interface (GUI). Modern CAD systems provide thousands of commands. Some of the commands are available in the GUI in the form of graphical elements. Some of the rarely used commands remain available only in manual input mode for experienced users, and the number of displayed elements is measured in hundreds. The purpose of the work is to implement the technology of forming a list of commands that best match the current context and accumulated experience of a CAD user. The developed solution uses modern machine learning algorithms (ML). A software library has been created that implements the technology. The library has been tested in integration with several modern CAD systems. The developed technology increases the efficiency of designers' work both through their own experience and through the transfer of knowledge in the form of ML models from more experienced engineers to less experienced ones. As part of the technology development, a data set was collected that has prospects for use in solving other CAD problems using ML methods. The technology has prospects for use in other types of applications.

***Ключевые слова*:** adaptive user interface, predictive user interface, context-aware user interface, CAD user interface, artificial intelligence, machine learning, neural network.

I. Введение

Постепенное развитие современных САПР по принципу увеличения количества поддерживаемых функций в совокупности с заметным дефицитом высококвалифицированных инженеров-проектировщиков выводит задачу поиска более эффективных методов взаимодействия с пользователями в разряд наиболее актуальных проблем развития современных САПР.

Цель данной работы – продолжить развитие технологии GUI, появившейся в 80-х годах прошлого века с выходом на массовый рынок операционных систем для персональных компьютеров, в специфичном для САПР контексте. Одновременно с нарастанием сложности GUI разнообразных приложений и операционных систем появляется ряд методов, позволяющих улучшить GUI на этапе его проектирования, например, Мавандади и соавт., 2006 [1].

Скачок в развитии алгоритмов ML, происшедший в последние годы, открывает широкие возможности для улучшения GUI не только на этапе проектирования, но также и на этапе использования программных средств пользователями. Методы ML улучшают пользовательский опыт, решая широкий спектр задач: персонализация, голосовые и текстовые помощники, предиктивная аналитика, а также адаптация приложений для нужд пользователей с ограниченными возможностями. Одной из наиболее актуальных задач в данной области является создание персонифицированных помощников для решения рутинных задач, выполняемых пользователями ПК. Например, широко известным стало применение компанией Microsoft большой языковой модели (LLM) GPT-4 в поисковом сервисе с целью суммаризации и объяснения результатов поисковых запросов [2]. С целью улучшения пользовательского опыта с помощью динамической адаптации GUI находят применения не только алгоритмы ML, но и алгоритмы обучения с подкреплением, например, Тоди и соавт., 2021 [3]. Подобно принципу переноса знаний в LLM, Намоуном и соавт., 2024 [4] была предложена большая модель пользовательского интерфейса (LUIM), позволяющая предиктивно генерировать GUI в зависимости от текущего контекста работы в приложении и нашедшая свое применение пока только в версиях приложений для мобильных платформ.

Несмотря на существование методов предиктивного адаптивного пользовательского интерфейса общего назначения (например, Карбун и соавт., 2019 [5]), качество их предсказаний в применении к САПР и многим другим типам приложений остаётся невысоким.

Практическое применение идей адаптивного пользовательского интерфейса к созданию GUI нового поколения для приложений любого назначения – по-прежнему чрезвычайно сложная задача с большим количеством неопределённых входных данных. Поэтому в данной работе мы ограничиваемся созданием технологии динамического контекстного (адаптивного) GUI для САПР. Все приложения данного типа соответствуют двум существенным для решения задачи критериям:

1. Единообразие команд редактирования. Любой современный САПР предоставляет своим пользователям в высокой степени стандартный набор команд рисования, редактирования, визуализации, аннотации, импорта, экспорта и т.д.
2. Единообразие содержательной структуры чертежа, состоящей из примитивов (точек, прямых линий, дуг, 3D примитивов и т.д.), находящихся между собой в отношениях.

Общие признаки САПР позволяют определить единый формат исходных данных, который можно подавать на вход моделям ML для получения необходимых предсказаний. В случае появления новых артефактов (например, неизвестной команды) модель будет накапливать пользовательские данные и дообучаться.

Общепринятым в настоящее время стандартом GUI в САПР является статический контекстный интерфейс, основной особенностью которого является статичное расположение всех элементов управления, которые могут потребоваться пользователю в тот или иной момент времени. При этом пользователь имеет возможность вручную переключать контекст, например выбрав вкладку риббона или переключив текущее рабочее окружение с “2D Моделирования” на “3D Моделирование”. Кроме того, современные САПР, как правило, поддерживают некоторые элементы динамического контекстного GUI, например модифицируют интерфейс, добавляя статическое множество команд в зависимости от типа выбранного объекта на чертеже.

Статический контекстный GUI в САПР, как правило, перегружен чрезмерным количеством команд, которое нередко достигает нескольких сотен, что снижает экранную площадь, доступную для отображения чертежа и увеличивает время поиска необходимых элементов управления, вызывая снижение общей эффективности работы пользователей.

Особенно актуально проблемы статических контекстных GUI стоят перед версиями САПР для мобильных операционных систем, которые получили развитие в последнее время. Меньшие размеры экранного пространства и ограничение на минимальный размер элементов управления делают наличие развитого динамического контекстного GUI необходимым требованием для мобильных решений.

Результаты нашей работы позволяют производителям САПР улучшать эффективность работы пользователей с помощью элементов адаптивного GUI, существенно сокращая элементы статического интерфейса в пользу динамического.

Ключевым результатом, обеспечивающим успешное применение программного продукта, разработанного в рамках данной работы, является создание модели ML на высоких уровнях качества. Создаваемые в данной работе алгоритмы искусственного интеллекта являются инновационным фундаментом всего решения и обеспечивают его высокую эффективность в решении задачи создания адаптивного пользовательского интерфейса САПР.

II. Постановка задачи

Общей целью выполнения работы является создание библиотеки прикладного программного обеспечения, реализующей технологию адаптивного пользовательского интерфейса для САПР. Продуктовое использование создаваемой библиотеки подразумевает интеграцию с существующими на рынке САПР решениями с целью накопления пользовательского опыта, использования накопленных данных для улучшения качества прогнозирующих моделей, а также прогнозирования следующей введённой пользователем команды в зависимости от его предыдущего опыта, опыта других пользователей САПР и текущего контекста использования САПР.

Данная цель подразумевает последовательное выполнение следующих задач: разработка модели пользовательского опыта САПР и модели текущего контекста САПР, разработка финальной модели, объединяющей результаты моделей пользовательского опыта и текущего контекста, а также проверка моделей на тестовом наборе данных, разработка библиотеки прикладного программного обеспечения, разработка технической документации библиотеки, разработка интеграции библиотеки с САПР, тестирование прототипа программной библиотеки адаптивного пользовательского интерфейса САПР.

Все модели ML, созданные в рамках данной работы, обучались на наборе данных WebCAD, собранном нами из открытых источников в сети Интернет.

III. Теория

1. Разработка модели пользовательского опыта

Модель пользовательского опыта решает задачу прогнозирования вероятности ввода команды, исходя из данных о ранее введённых командах. Исходными данными для модели являются:

1. Набор данных “Последовательность команд” WebCAD, содержащий 3863430 команд в общей последовательности.
2. Набор данных “Описания команд” WebCAD, содержащий короткие имена команд, подробные описания и расположения команд в GUI САПР.
3. Набор данных “Текстовый корпус САПР” WebCAD, включающий 2582 текстовых документа общим объёмом 30 Мб по тематике САПР: справки популярных современных САПР, тексты статей тематических форумов и пр.

В рамках данной работы нами было опробовано несколько типов моделей ML, применяемых в настоящее время для решения похожих задач [6, 7, 8, 9] и выбран подход, показавший наивысшие метрики качества на тестовой выборке.

Основной метрикой качества, в соответствии с которой мы будем сравнивать различные методы, является Top 10 Accuracy. В соответствии с данной метрикой предсказание модели считается верным, если введённая команда содержится в списке из 10 наиболее вероятных команд. Целевым значением данной метрики, к достижению которого мы стремимся в рамках работы, является 0.95, что обеспечило бы целесообразность применения метода на практике. Во всех описанных ниже методах размер тестовой выборки принимается равным 30% от размера всего набора данных.

1.1. Метод предсказания следующего предложения

Идея метода состоит в замене команды в последовательности команд её текстовым описанием, построении языковой модели BERT [6] и решении задачи предсказания следующего предложения, где в качестве первого предложения выступает свёртка описаний предыдущих команд, а в качестве предсказываемого предложения – описания из списка всех команд. Далее мы приводим последовательность действий для построения модели пользовательского опыта данным методом.

1. Дообучение языковой модели BERT (“bert-base-uncased”) [6] для предсказания следующего предложения на наборе данных “Текстовый корпус САПР” адаптивным методом обратного распространения ошибки Adam. Количество эпох обучения: 10, темп обучения: 0.0005. Длина входной последовательности ограничена 60 токенами.
2. Создание текстовой свёртки истории команд.
3. Вычисление предсказания модели для предложения А = текстовая свёртка истории команд, предложение Б = описание команды.
4. Итоговое предсказание – список команд, отсортированный в порядке убывания значений вероятностей соответствующих предложений.

Top 10 Accuracy данного метода на тестовой выборке составляет 0.71.

1.2. Метод схожести компактных представлений описаний команд

Данный метод является развитием предыдущего метода, в котором задача предсказания следующего предложения заменена попарным сравнением компактных представлений истории выполнения и списка всех команд. Далее мы приводим последовательность действий для построения модели пользовательского опыта данным методом.

1. Дообучение языковой модели Sentence Transformer (“paraphrase-mpnet-base-v2”) [8] на наборе данных “Текстовый корпус САПР” адаптивным методом обратного распространения ошибки Adam. Размер батча: 64, количество итераций сэмплирования пар предложений: 5, количество эпох обучения: 10, количество предварительных шагов: 10, функция потерь: CosineSimilarityLoss.
2. Создание с помощью дообученной модели компактных представлений (энкодингов) описаний команд из набора данных “Описания команд”.
3. Вычисление энкодинга истории выполнения команд глубины n по формуле затухающей суммы (1).

$$H_n = \sum_{k=-1}^{-n} \mathrm{m} * E_k \tag{1}$$

где n – это количество команд в предшествующей последовательности,

m – коэффициент затухания, который служит для того, чтобы более поздние команды в сценарии имели больший вклад в прогноз,

k – индекс команды (0 – прогнозируемая команда, -1 – предыдущая команда и т.д.),

E_k – энкодинг команды.

4. Вычисление вероятности команд по формуле косинусной схожести (2).

$$S(E_k, H_n) = \frac{E_k \cdot H_n}{\|E_k\| * \|H_n\|} \tag{2}$$

В рамках данной работы мы использовали значение n = 5 и m = 1.2. Top 10 Accuracy на тестовой выборке составляет 0.78.

1.3. Метод схожести комплексных представлений команд

Заметим, что метод схожести компактных представлений описаний команд показывает более высокое качество на тестовой выборке. Существенным источником улучшения качества данного метода может послужить использование данных о расположении команд в GUI, которые представляют собой источник информации о взаимной схожести команд. Заметим, что использование энкодингов описаний команд, полученных из языковой модели Sentence Transformer, является достаточно сложным методом, который хорошо работает для текстов любого происхождения и сложности, но может проиграть сравнительно простому методу TF-IDF [10] для текстов относительно короткого словаря и коротких предложений.

Метод схожести комплексных представлений команд является развитием предыдущего метода согласно данным замечаниям. Далее мы приводим последовательность действий для построения модели пользовательского опыта данным методом.

1. Создание энкодинга расположения команд в GUI методом One Hot Encoding. Созданный таким образом энкодинг представляет собой вектор размером 156, состоящий из 0 и 1.
2. Создание энкодинга методом TF-IDF для предварительно токенизированного и лемматизированного текста описания команды. Энкодинг представляет собой вектор размером 552.
3. Нормализация и объединение энкодингов. Размер комплексного энкодинга – 708.
4. Вычисление комплексного энкодинга истории выполнения команд по формуле (1).
5. Вычисление вероятности команд по формуле косинусной схожести (2) для комплексных энкодингов.

Top 10 Accuracy данного метода на тестовой выборке составляет 0.81.

1.4. Рекуррентная нейронная сеть со слоем эмбедингов из комплексных представлений команд

Данный метод является развитием предыдущего метода, в котором пункты 4 и 5 заменены предсказанием рекуррентной нейронной сети. Мотивация данного изменения заключается в том, что формулы (1) и (2) представляют собой достаточно простые функции, которые плохо выучивают сложные или долгосрочные

зависимости, имеющие место в последовательности команд, например циклы. В настоящее время стандартом решения данных задач являются рекуррентные нейронные сети с ячейками долгосрочной памяти [9].

Общая схема слоёв модели, показавшей наибольшую эффективность в вычислительных экспериментах, представлена на рис. 1.

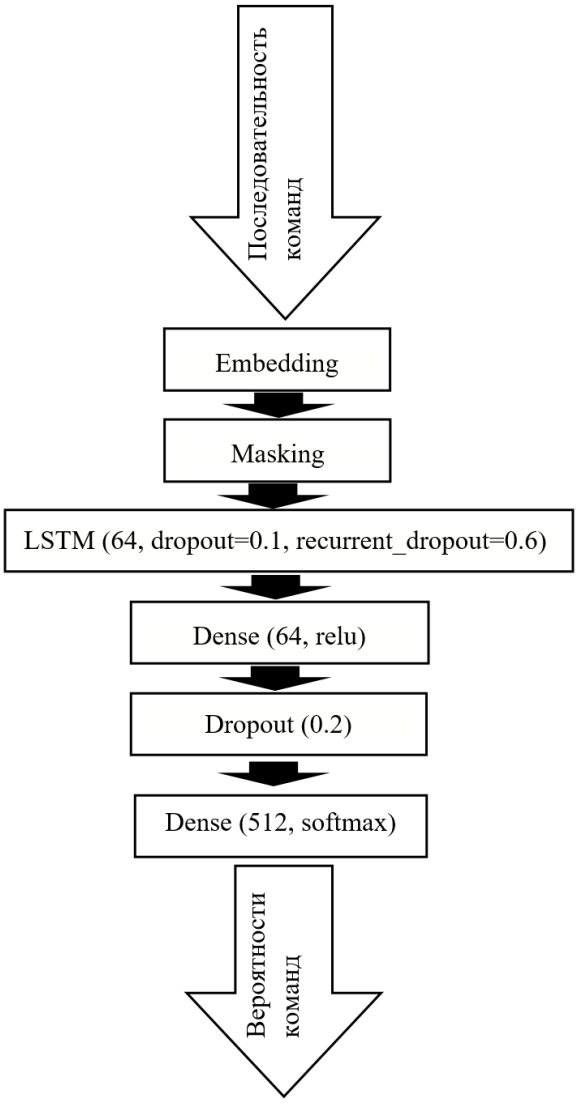


Рис. 1. Модель пользовательского опыта

Нейронная сеть обучалась адаптивным методом обратного распространения ошибки Adam. Функция потерь Categorical Crossentropy, размер батча: 1024, количество эпох обучения: 200, предусмотрена ранняя остановка обучения в случае отсутствий улучшения метрик качества на валидационной выборке в течение минимум 15 эпох. Слой Embedding является обучаемым и инициализирован значениями комплексных представлений команд.

Top 10 Accuracy на тестовой выборке составляет 0.94, что значительно превосходит качество предыдущих методов. Далее в рамках данной работы мы используем рекуррентную нейронную сеть для построения модели пользовательского опыта.

2. Разработка модели текущего контекста

Модель текущего контекста САПР строится на основе набора данных “Текущий контекст” WebCAD для предсказания следующей команды.

Набор данных “Текущий контекст” – это таблица, каждая строка в которой представляет собой независимое состояние, характеризующее текущий контекст работы пользователя. Столбцами таблицы являются параметры, описывающие текущее состояние:

1. Количество объектов определённого типа, содержащихся в чертеже.
2. Количество объектов определённого типа, содержащихся во множестве выбранных объектов.
3. Значения системных переменных САПР.

Строки таблицы содержат значения следующей команды, которую ввёл пользователь из данного состояния (столбец NextCommand). Данный столбец является целевой переменной для дальнейшего обучения модели.

Набор данных “Текущий контекст” – это структурированные, табличные данные, общим стандартом для моделирования которых в настоящее время являются алгоритмы градиентного бустинга на деревьях решений. В вычислительных экспериментах мы использовали несколько наиболее эффективных алгоритмов градиентного бустинга: XGBoost, LightGBM, CatBoost. В качестве алгоритма обучения модели мы выбрали XGBoost [11], который показал наивысшие значения метрик качества на тестовой выборке.

Разработанная модель XGBoost текущего контекста представляет собой ансамбль деревьев решений, построенных по методу градиентного бустинга, в котором каждое следующее дерево решений строится по принципу уменьшения совокупной ошибки предыдущих деревьев, а итоговые вероятности команд определяются как сумма выводов всех входящих в ансамбль деревьев решений. Общий принцип функционирования модели текущего контекста изображён на рис. 2.

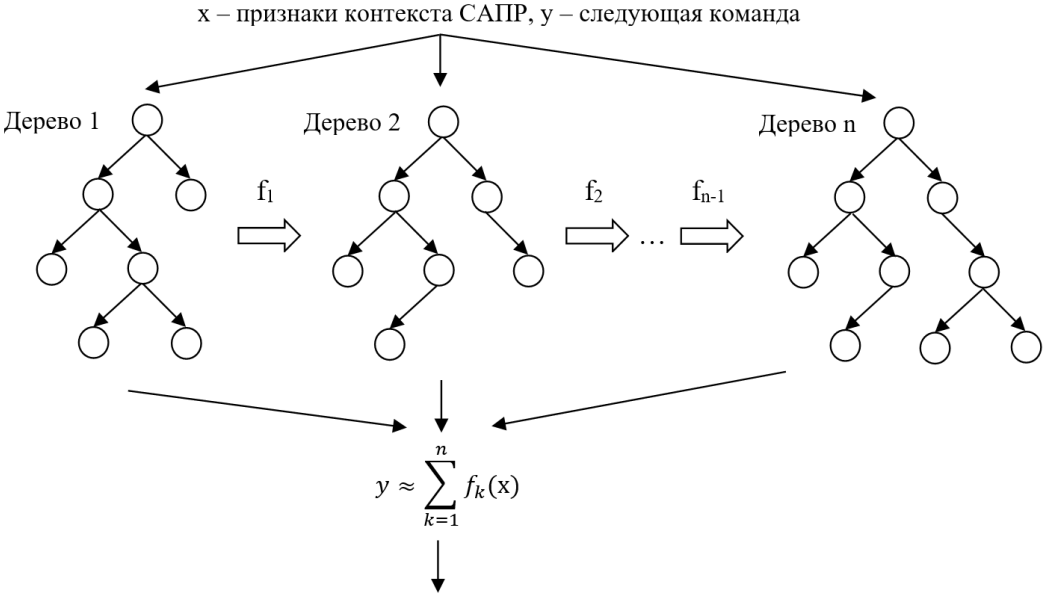


Рис. 2. Модель текущего контекста

Процесс создания модели включал три основных этапа:

1. Предобработка данных. Данный этап включал следующие процедуры: Labels Encoding для столбца NextCommand, исключение сильно коррелирующих признаков, создание новых и замена существующих признаков функцией (логарифмирование признаков с сильным разбросом и добавление линейных комбинаций признаков), стандартизация числовых признаков.
2. Обучение модели. Данный этап начинался с подбора значений гипер-параметров модели с целью определения значений, максимизирующих площадь под ROC-кривой. Поиск оптимальных значений гипер-параметров проходил в два шага: RandomizedSearch на сетке гипер-параметров в широком диапазоне значений, затем GridSearch в окрестности найденных на предыдущем шаге значений. После определения оптимальных значений гипер-параметров модель обучалась на части данных, предназначенной для обучения. Был определен оптимальный размер обучающей выборки с помощью метода построения кривой обучения.
3. Тестирование модели. Этап включал определение метрик качества модели (F1, точность и полнота в зависимости от порога), построение графиков кривой ROC, кривой Precision-Recall, значимости параметров.

Цикл построения модели был повторен несколько раз для достижения наивысшего значения Top 10 Accuracy.

Итоговая модель текущего контекста, созданная в рамках данной работы, обладает следующими значениями метрик качества:

1. Взвешенная усреднённая точность: 0.65.
2. Взвешенная усреднённая полнота: 0.65.
3. Top 10 Accuracy: 0.98.

В таблице 1 приведены значения параметров обучения, которые продемонстрировали наивысшие показатели качества модели текущего контекста.

ТАБЛИЦА 1
ЗНАЧЕНИЯ ПАРАМЕТРОВ ОБУЧЕНИЯ МОДЕЛИ ТЕКУЩЕГО КОНТЕКСТА

№	Параметр	Значение
1	Objective (функция потерь)	multi:softprob
2	booster (тип модели)	gbtree (дерево решений)
3	tree_method (метод построения дерева)	gpu_hist (гистограмма для GPU)
4	learning_rate (темп обучения)	0.07
5	gamma (минимальное изменение функции потерь для разделения)	0 (нет ограничений)
6	max_depth (максимальная глубина дерева)	7
7	min_child_weight (минимальная сумма весов)	1
8	max_delta_step (максимальное изменение веса)	0 (нет ограничений)
9	lambda (коэффициент L2 регуляризации)	10
10	alpha (коэффициент L1 регуляризации)	2
11	subsample (часть тренировочных данных для выборки на каждой итерации бустинга)	0.7
12	Размер обучающей выборки	25478

3. Разработка объединения моделей

Пользовательский опыт представляет собой сложный информационный объект, имеющий несколько независимых источников происхождения, что в свою очередь, вызывает необходимость создания нескольких моделей машинного обучения, которые независимо описывают различные части данного объекта. Общая схема функционирования метода представлена на рис. 3. Таким образом для получения итоговых предсказаний нашим методом требуется объединить предсказания всех независимых моделей с помощью модели-дискриминатора так, чтобы качество итогового прогноза было выше качества отдельных моделей.

Исходными данными для работы дискриминатора являются вероятности команд, предсказанных моделью пользовательского опыта и моделью текущего контекста. Дискриминатор прогнозирует итоговые вероятности команд.

Обучающая и тестовая выборка для обучения модели состояла из объединения выводов обоих моделей.

В качестве архитектуры модели-дискриминатора использовалась мультиномиальная логистическая регрессия [12]. Выбор данной архитектуры обусловлен весьма простой структурой входных данных: признаки – вероятности команд с выхода обоих моделей, подавляющее большинство которых близки к нулевым значениям и целевая переменная – итоговые вероятности команд. Принципиальная схема функционирования дискриминатора изображена на рис. 4: W – матрица весовых коэффициентов модели, N – общее количество прогнозируемых команд.

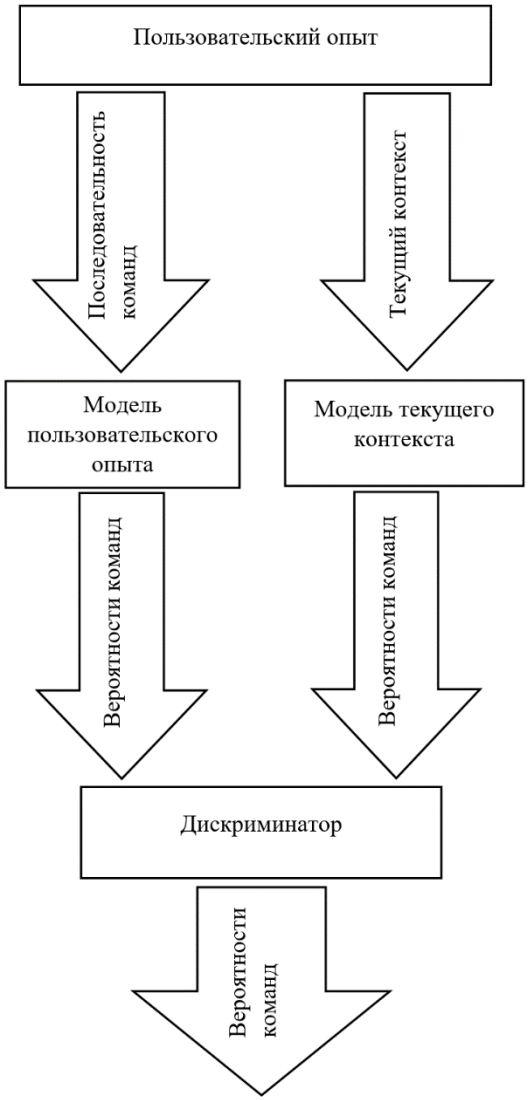


Рис. 3. Общая схема функционирования метода адаптивного пользовательского интерфейса САПР

Процесс создания модели включал три основных этапа:

1. Предобработка данных: исключение сильно коррелирующих признаков, создание новых и замена существующих признаков функцией (логарифмирование признаков с сильным разбросом и добавление линейных комбинаций признаков).
2. Обучение модели. Данный этап начинался с подбора параметров регуляризации, максимизирующих площадь под ROC-кривой, после чего модель обучалась на части данных, предназначенной для обучения, с помощью метода градиентного спуска. Был определен оптимальный размер обучающей выборки с помощью метода построения кривой обучения.
3. Тестирование модели. Этап включал определение метрик качества модели (F1, точность и полнота в зависимости от порога), построение графиков кривой ROC, кривой Precision-Recall.

Лучшая из обученных нами моделей в рамках данной работы продемонстрировала следующие значения метрик качества:

1. Взвешенная усреднённая точность: 0.68.
2. Взвешенная усреднённая полнота: 0.66.
3. Top 10 Accuracy: 0.99.

В таблице 2 приведены значения параметров обучения, которые продемонстрировали наивысшие показатели качества модели.

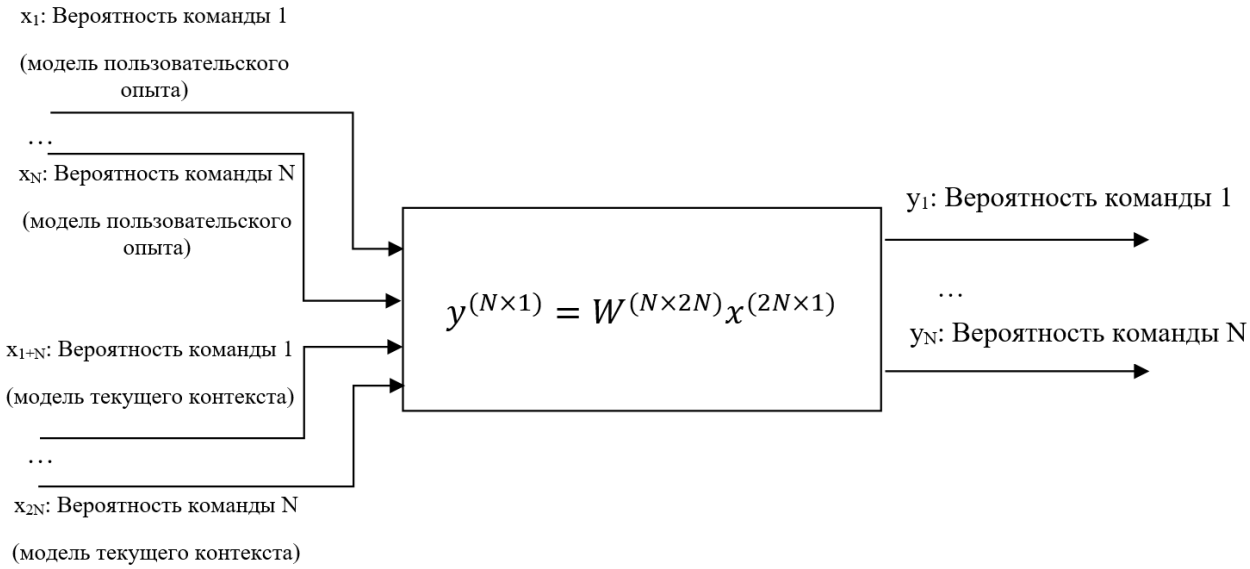


Рис. 4. Модель-дискриминатор

ТАБЛИЦА 2
ЗНАЧЕНИЯ ПАРАМЕТРОВ ОБУЧЕНИЯ МОДЕЛИ-ДИСКРИМИНАТОРА

№	Параметр	Значение
1	solver (алгоритм оптимизации)	lbfgs
2	penalty (способ регуляризации)	L2
3	C (коэффициент регуляризации)	0.7
4	Размер обучающей выборки	25478

IV. Результаты экспериментов

Вычислительные эксперименты в рамках данной работы заключались в тестировании итоговой цепочки моделей с целью определения значений ключевых метрик качества дискриминатора на тестовой выборке. Для получения сравнимых результатов в каждом из экспериментов тестирования моделей тестовая выборка была зафиксирована и содержала 10919 записей пользовательского опыта.

На рис. 5 представлено изображение матрицы неточностей, визуализирующей структуру ошибок модели-дискриминатора на тестовой выборке для нескольких наиболее популярных команд САПР. Строки матрицы соответствуют верным значениям следующей команды, столбцы – предсказанным командам, таким образом верные случаи предсказаний соответствуют главной диагонали матрицы. Высокие значения вне главной диагонали матрицы соответствуют устойчивым ошибкам и являются предметом дальнейших исследований с целью достижения более высокого качества модели.

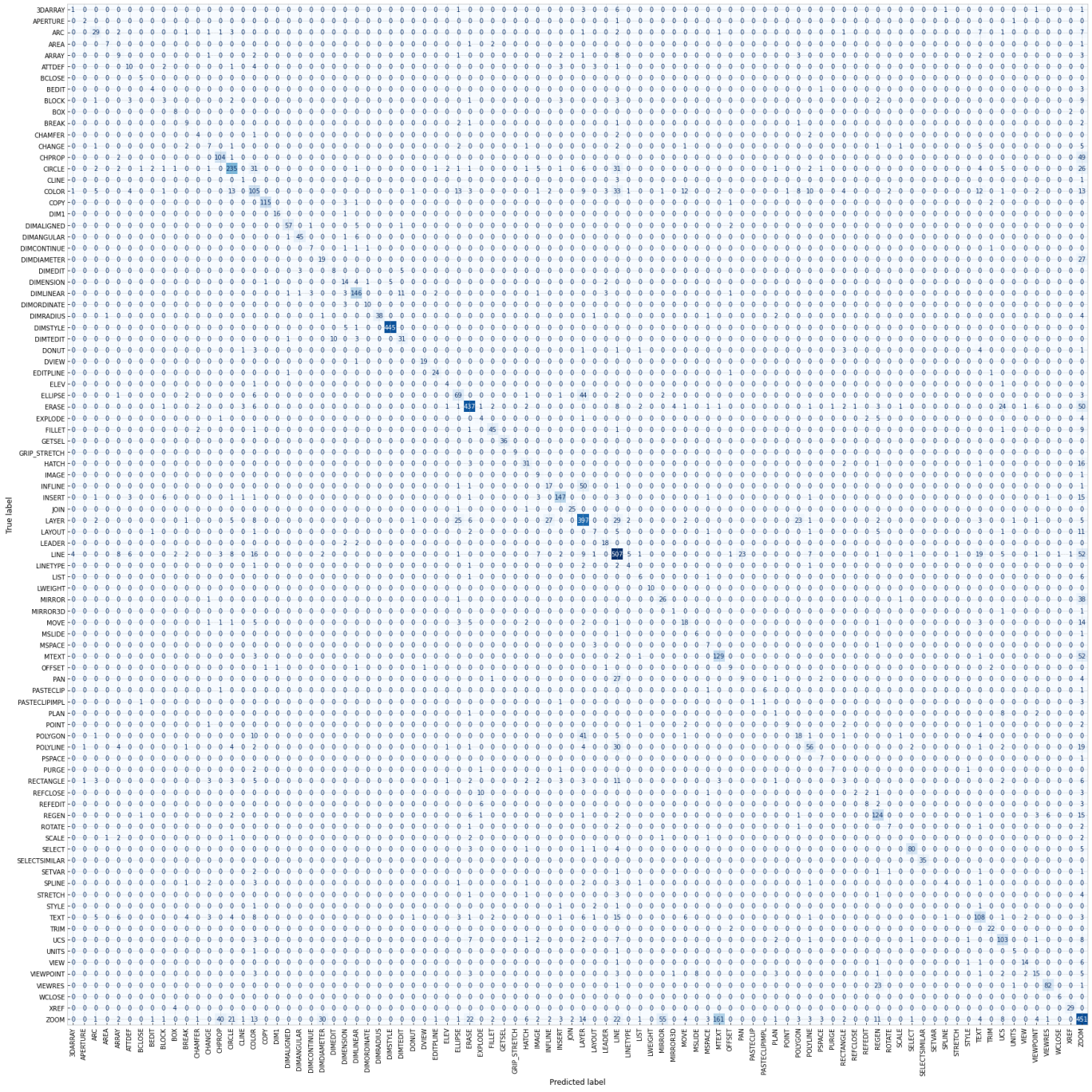


Рис. 5. Матрица неточностей тестирования модели

В таблица 3 представлены усреднённые по всем командам значения точности, полноты и F1, полученные моделью на тестовой выборке.

ТАБЛИЦА 3
УСРЕДНЁННЫЕ МЕТРИКИ

Значение	Точность	Полнота	F1-score
Macro avg	0.56	0.55	0.54
Weighted avg	0.65	0.65	0.64

1. Accuracy: 0.65.
2. Top 5 accuracy: 0.94.
3. Top 10 accuracy: 0.99.

4. Micro-averaged One-vs-Rest ROC AUC score: 0.99.

Данные результаты тестирования свидетельствуют о высоком качестве разработанных моделей.

V. Обсуждение результатов

Достигнутое в рамках данной работы значение Top 10 accuracy на уровне 0.99 существенно превосходит необходимый для практического применения уровень 0.95. Значение метрики Top 5 accuracy на высоком уровне 0.94 позволяет на практике использовать более короткие списки рекомендованных команд. Значение метрики Accuracy на уровне 0.65 свидетельствует о том, что разработанная технология не позволяет в точности предсказывать следующую команду, что могло бы в перспективе быть использовано для автоматизации сценариев проектирования. Однако Accuracy имеет относительно высокое значение, что может свидетельствовать о существовании некоторых частных сценариев проектирования, поддающихся достаточно надёжной автоматизации. Целью будущей работы может служить определение таких сценариев.

Мы рассматриваем в качестве основного источника улучшения качества работы модели изучение и устранение причин систематических ошибок, наблюдаемых в матрице неточностей, что также может рассматриваться в качестве темы будущего исследования.

Ещё одним источником улучшения технологии может служить использование альтернативных архитектур моделей, например нейронных сетей. Мы рассматриваем в качестве перспективной идею о замене модели текущего контекста, модели пользовательского опыта и модели-дискриминатора единой нейронной сетью, которая позволила бы обучиться на всём наборе данных в едином контексте. Единая модель потребовала бы существенно больше данных, чем мы располагаем в данный момент в наборе данных WebCAD, что в сою очередь свидетельствует о сохраняющейся актуальности задачи продолжения работ над пополнением набора данных WebCAD. Заметим, что предложенная в данной работе технология базируется на весьма вычислительно эффективных алгоритмах логистической регрессии, градиентного бустинга и рекуррентных нейронных сетях. Замена данных алгоритмов единой нейросетевой моделью, вероятно, приведёт к снижению вычислительной эффективности. Таким образом предметом будущих экспериментов может служить отыскание необходимого баланса между увеличением точности прогнозирования за счет использования нейросетевой архитектуры модели и эффективностью использования вычислительных ресурсов.

Подчеркнем, что вне рамок данной работы осталась задача предсказания параметров команд. Большинство команд САПР имеют параметры числового, векторного, перечисляемого или текстового типа. Мы считаем, что задача предсказания значений параметров может быть решена с высоким качеством при помощи предложенной технологии по крайней мере для случая перечисляемого типа.

VI. Выводы и заключение

В рамках данной работы была создана модель пользовательского опыта САПР, модель текущего контекста САПР и объединение моделей в виде модели-дискриминатора для получения итогового предсказания относительно следующей команды САПР. Работа моделей в решении задачи предсказания следующей команды протестирована на тестовом наборе данных. Достигнуто качество работы модели, существенно превосходящее целевой уровень, необходимый для практического применения технологии.

Непосредственные результаты выполнения работы – разработанная технология предсказания следующей команды САПР – может быть использована в качестве дополнения к функциональности существующих на рынке решений САПР. Это позволит существенно повысить эффективность труда инженеров-проектировщиков, снижая себестоимость и сроки работ по проектированию в САПР, что в свою очередь будет иметь высокий долгосрочный технико-экономический эффект в разных отраслях экономики.

Источник финансирования. Благодарности

Основные результаты работы получены в рамках гранта №171ГС1ИИС12-D7/80645 от 12.12.2022 на проведение НИОКР от ГБУ «Фонд содействия развитию малых форм предприятий в научно-технической сфере».

Работа была продолжена в рамках государственного задания на проведение фундаментальных научных исследований на 2024-2026 гг. (FFUF -2024-0031, № НИОКТР 1023032800130-3-2.3.2).

Список литературы

1. S. Mavandadi, P. Aarabi, A. Khaleghi and R. Appel, "Predictive Dynamic User Interfaces for Interactive Visual Search," 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 2006, pp. 381-384, doi: 10.1109/ICME.2006.262516.
2. OpenAI: GPT-4 Technical Report, arXiv:2303.08774v1 [cs.CL], 15 Mar 2023.
3. Kashyap Todi, Gilles Bailly, Luis A. Leiva, Antti Oulasvirta “Adapting User Interfaces with Model-based Reinforcement Learning”, arXiv:2103.06807 [cs.HC], 11 Mar 2021.
4. Abdallah Namoun, Ahmed Alrehaili, Zaib Un Nisa, Hani Almoamari, Ali Tufail, “Predicting the usability of mobile applications using AI tools: the rise of large user interface models, opportunities, and challenges”, arXiv:2405.03716 [cs.HC], 5 May 2024.
5. Carbune, Victor; Keysers, Daniel; and Deselaers, Thomas, "Visual User Interface (UI) Interaction Prediction", Technical Disclosure Commons, 26 February, 2019).
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
7. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.
8. Spyros Zoupanos, Stratis Kolovos, Athanasios Kanavos, Orestis Papadimitriou, Manolis Maragoudakis. Efficient comparison of sentence embeddings. 18 Apr 2022, arXiv:2204.00820.
9. Alex Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. 9 Aug 2018, arXiv:1808.03314.
10. Amir Jalilifard, Vinicius F. Caridá, Alex F. Mansano, Rogers S. Cristo, Felipe Penhorate C. da Fonseca. Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. 2021, arXiv:2001.09896.
11. Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. 2016, arXiv:1603.02754.
12. Felix Abramovich, Vadim Grinshtein, Tomer Levy. Multiclass classification by sparse multinomial logistic regression. 2020, arXiv:2003.01951.

УДК 004.925.8:658.5
DOI: 10.25206/978-5-8149-3873-2-2024-564-572

Реинжиниринг параметрической модели просечно-вытяжного листа с использованием оператора скиннинга поверхности

Parametric reverse engineering of expanded sheet metal parts based on surface skinning operator

С. Е. Сляднев, А. М. Воеводин
ООО «Студия Квавар», г. Нижний Новгород, Россия

S.E. Slyadnev, A.M. Voevodin
Quaoar Studio LLC, Nizhny Novgorod, Russia

Аннотация. Предложен новый подход к параметризации геометрических моделей просечно-вытяжных металлических листов (ПВЛ). Получаемая геометрическая модель представлена в виде решетки, состоящей из «минимально повторимых» (minimal repeatable) шаблонов. Шаблон реконструируется оператором скиннинга поверхности по данным лазерного сканирования, представленным производителем ПВЛ. Предложенный подход обеспечивает «гибкую» параметризацию модели с высокой детализацией, приближенной к настоящей форме изделия. Наличие параметрической модели позволяет ставить задачи оптимизации. Продемонстрирован пример постановки такой задачи, где целевой функцией является геометрическая невязка модели по отношению к данным сканирования. Предложенный подход основан на реконструкции семейства сплайн-кривых невысоких степеней (3) и протягивании сплайн-поверхности вдоль опорных кривых для каждой грани шаблона. Параметрами проектирования являются толщина листа, свойства инструмента (длина и угол «зуба»), подъем шаблона над плоскостью и т.д. Заявленные параметры имеют инженерный смысл и могут использоваться для подбора конфигурации ПВЛ производителем под нужды заказчика. Результаты исследования были использованы в коммерческом программном обеспечении одного из ведущих мировых поставщиков ПВЛ.

***Ключевые слова*:** просечно-вытяжной лист (ПВЛ), реинжиниринг, скиннинг B-сплайн поверхности.

Abstract. This study explores a parameterization approach for precise modeling of expanded sheet metal parts. The obtained geometric model is represented as a rectangular grid of "minimal repeatable" patterns. Each pattern is reconstructed using surface skinning techniques based on the laser scan data that is provided by a manufacturing shop. The proposed method enables the generation of accurate and visually realistic geometric models of "minimal repeatable" patterns for various punch tools. Consequently, the shape's parametric model can be subjected to optimization, where the objective function measures the deviation of a laser scan from the nominal shape. The described approach involves creating a number of curve networks and approximating each network using B-spline surfaces. The design parameters exposed by the constructed parametric model possess significant geometric meaning, such as the thickness of the material, the length of the tool, the width of the tool, and so on. Therefore, the exposed variables can be used as meaningful "handles" for design optimization. The results of this study have been incorporated into a commercial software package that generates a variety of expanded metal grids for a range of user-defined parameters.

***Keywords*:** expanded sheet metal, reverse engineering, surface modeling, B-spline skinning.

I. Introduction

Expanded sheet metal parts are commonly used in AEC (Architecture, Engineering, and Construction) applications due to their remarkable functional and aesthetic properties, as well as their simple fabrication method. Nevertheless, despite the ease of manufacturing (punching a sheet metal blank with progressive longitudinal and transversal offset), the resulting form may exhibit various shapes of distortion and deformation, which can be difficult to comprehend.

The precise modeling of expanded metals does not seem to be an often-addressed geometric problem. In CAD, these shapes typically serve as decorative elements within a larger information model. In such a context, it is not reasonable to create smaller-scale models of expanded metals, since their geometric features will not be perceptible to the human eye at the commonly employed level of detail. At the same time, complexity calls for higher processing power in order to cope with a vast number of geometric features. Consequently, a search for better compactness and more efficient visualization typically leads to the sacrifice of accuracy in ornamental parts, such as expanded metals.

The situation changes when expanded sheet metals are analyzed at their own (smaller) scale. Frequently, manual selection of the desired shape by a human operator is made by referring to existing catalogs, where the only way of visually evaluating the expanded metal product is by examining a picture of it.

II. Problem statement

This paper discusses a geometric modeling technique that enables the representation of expanded metals in three dimensions, similarly to ordinary sheet metal parts, whose modeling in 3D has been a common practice for decades. The availability of a precise geometric model of an expanded sheet metal allows for further research into the functioning of these parts, such as fluid flow separation (CFD), thermal or structural analysis.

The main contribution of this study is the proposed parameterization technique, which offers a sufficient number of "sensitives" to accurately convey a wide range of expanded metal parts. It may come as a surprise that apparently simple expanded metals call for sophisticated surface modeling techniques that are almost as complex as the modeling operators employed in, for example, turbomachinery design. While it is possible to simplify the precise modeling of expanded metals for some specific cases, our real-life experience has shown that a general approach necessitates the use of techniques such as curve and surface approximation.

The goal of this study is to construct a parametric model $S(\mathbf{a})$, where $\mathbf{a} \in \mathbb{R}^N$ is a vector of design variables for representing a single "minimal repeatable" pattern of the expanded sheet metal (ESM) grid (here N is the number of design variables). The requirements imposed on such a model are twofold: a) the obtained geometric model should be accurate enough to communicate the shape of the ESM grid to human consumers; b) the enabled design parameters should be flexible enough to allow for deforming the parametric model to fit into different shapes of punches.

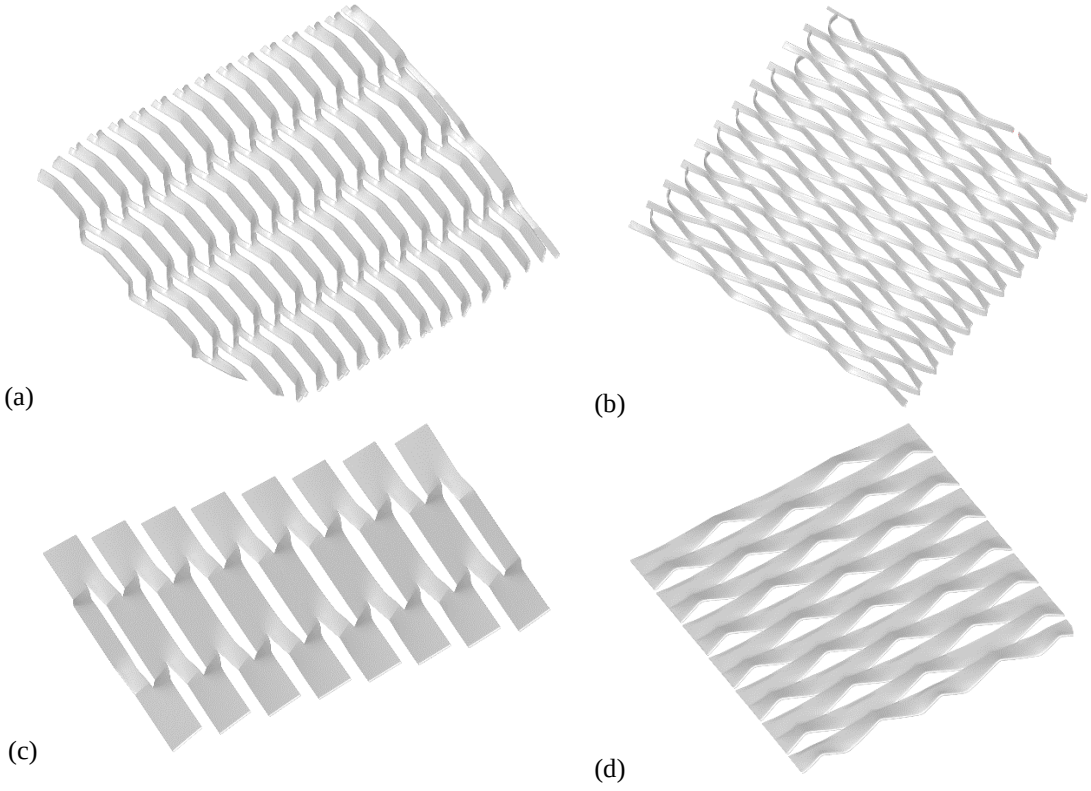


Fig. 1. Laser scan data for expanded sheet metal grids of different types (provided by manufacturing shop): (a) "armadillo", (b) "wave", (c) "prism", (d) "rhomb".

III. Theory

The modeling process starts with manual calibration of the key feature points based on the provided laser scans (fig. 1). The scan data captures the real shape of the final product and is used as a reference for constructing parametric models. For each type of expanded sheet metal, the corresponding laser scan is prepared as a surface mesh stored in the corresponding STL file. For the sake of simplification, we extract a "minimal repeatable" pattern M as a portion of surface

triangulation whose replication in the modeling space with different transformations would end up in the desired grid (fig. 2).

Fig. 2. The "minimal repeatable" pattern extracted from the laser scan of a "prism" grid.

We aim at reconstructing a precise parametric model of the minimal repeatable portion M that provides "close to reality" accuracy (fig. 3).

Fig. 3. The "minimal repeatable" pattern's scan (rendered as a point cloud) matched with the reconstructed parametric model (shown in blue). The point that yields the maximum geometric deviation is depicted with a red ball marker (~1.5 mm in this case).

6. Modeling stages

The modeling process starts with the preparation of a regular point grid $\{\mathbf{R}_i\}$ located at the imaginary midsurface of the pattern (fig. 4). At this stage, the minimal repeatable is constructed as a single strip, which would later undergo cutting and replication operations to generate the final state of the model. By varying the slopes of the transversal segments (denoted with s in fig. 4), we can achieve the desired twists of the strip shape and better capture the initial scan.

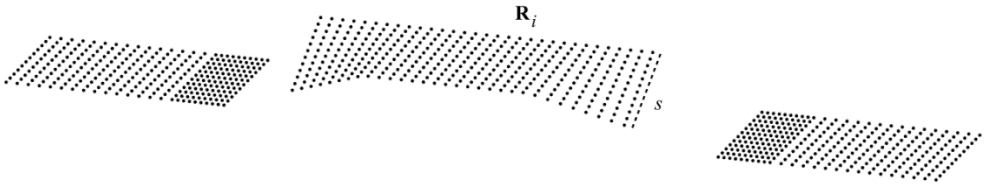


Fig. 4. The initial regular point cloud at the midsurface of the pattern.

The midsurface point set $\{\mathbf{R}_i\}$ (fig. 6a) is used to construct the longitudinal interpolant curves (fig. 6b). Once these curves are defined, they are discretized with a uniform parametric step for the sake of better covering the otherwise undefined strands. It is the shape of the interpolant curves $\mathbf{c}_j(u)$ that largely determines the shape of the strands (see fig. 5 which illustrates what is a "strand"). The points sampled over the longitudinal curves $\mathbf{c}_j(u)$ are shifted to the upper and inner directions for modeling the material thickness (*sw* parameter in Table 1). As a result, two point sets are derived: $\{\mathbf{R}_i^u\}$ and $\{\mathbf{R}_i^l\}$ (fig. 5, fig. 6c).

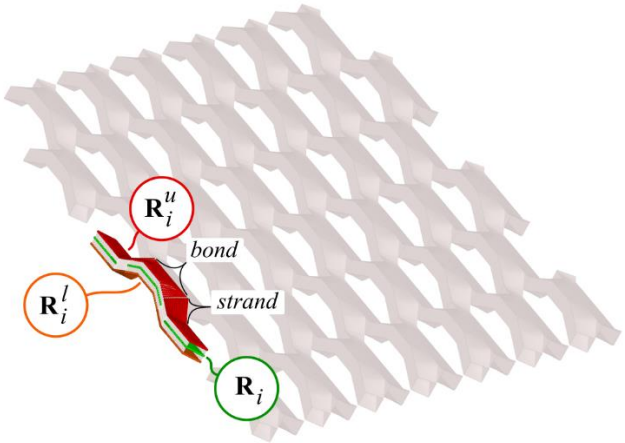


Fig. 5. Final ESM grid with minimal repeatable highlighted.

Using the obtained point-set layers, the cross-sectional profiles $\mathbf{p}_k$ for a volumetric strip are defined (fig. 6d). These profiles are obtained using the same synchronous skinning procedure as applied for constructing the longitudinal curves $\mathbf{c}_j(u)$. Spatial spacing of profiles follows the discretization of the curves $\mathbf{c}_j(u)$. Because of dense sampling, the profiles can be reliably skinned into a B-spline surface without unexpected oscillations (Fig. 6e).

At the last stage, the constructed volumetric strip is cut at the mid-strands, and the obtained pieces are moved side-by-side with the initial strip to form a "cross" pattern (Fig. 6f).

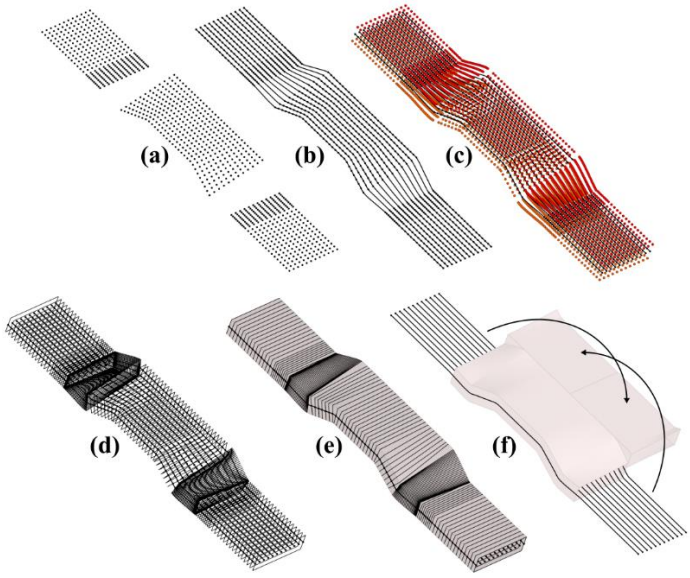


Fig. 6. ESM modelling stages.

The presented approach to modeling proved to be practical and computationally quick for expanded sheet metal pieces of locally symmetric shape (otherwise, the split-replication technique depicted in fig. 6f is not valid). It should be noted that expanded sheet metal models are not necessarily symmetric (e.g., the "wave" pattern in fig. 1b exposes geometric anisotropy between odd and even rows). Still, as the core of the method consists of surface skinning, it can be reused for non-symmetrical patterns with some amelioration in the employed parameterization.

7. Curve interpolation

The central stage of the modeling procedure is applying the skinning operator (fig. 6e) to approximated B-spline curves. In our study, we explored two ways of curve approximation: the "variational" approach that optimizes the differential properties of a curve [9] and plain interpolation with centripetal parameterization.

The variational setup finds the approximant curve **c** defined in the range $[a, b]$ as the minimizer of a quadratic criterion supplemented with fairing functionals as denoted in Eq. (1). This method, among others, has been used by S. Auriemma in the work related to gradient-based shape optimization [3].

$$\mathrm{J} = \mathrm{J}_0 + \omega_1 \mathrm{J}_1 + \omega_2 \mathrm{J}_2 + \omega_3 \mathrm{J}_3 \tag{1}$$

Here J_0 is defined with Eq. (2) and serves to minimize the distance between the feature points $\mathbf{P}_i$ and the approximant curve.

$$\mathrm{J}_0[\mathbf{c}] = \sum_{i=0}^{m} \|\mathbf{c}(t_i) - \mathbf{P}_i\|^2 \tag{2}$$

The supplementary functionals $\mathrm{J}_{1,2,3}$ are used in the "variational" optimizer [9] to minimize the curve length (Eq. 3), curvature (Eq. 4), and sudden variations in torsion (Eq. 5).

$$\mathrm{J}_1[\mathbf{c}] = \int_a^b \|\dot{\mathbf{c}}(t)\| dt \tag{3}$$

$$\mathrm{J}_2[\mathbf{c}] = \int_a^b \rho^2(t) dt \tag{4}$$

$$\mathrm{J}_3[\mathbf{c}] = \int_a^b \dot{\rho}^2(t) dt \tag{5}$$

Here ρ is the curvature computed as follows:

$$\rho(t) = \frac{\|\dot{\mathbf{c}}(t) \times \ddot{\mathbf{c}}(t)\|}{\|\dot{\mathbf{c}}(t)\|^3}$$

Curve optimization assumes that there are a number of constraints imposed on the behavior of the minimizer. Also, the objective functional J is not linear by **c** while the functionals $\mathrm{J}_{1,2,3}$ are generally not quadratic.

The technical report by Matra Datavision [9] gives a profound description of the numerical method to solve the problem $\mathrm{J} \to min$. This report is remarkable in the sense that it corresponds to *AppDef_Variational* algorithm, being the legacy of STRIM 100 modeling core [6] (developed by CISIGRAPH and acquired by Matra Datavision in 1994). This algorithm is available in open source [2] and can be plugged into the computation with minimal effort. At the same time, the computational burden posed by this algorithm seems to outweigh the benefits of smoothing out curve defects, especially because the feature points $\mathbf{P}_i$ are free of noise and allow for exact interpolation.

To speed up computations while keeping the curve shape satisfactory, we employ the simplest method of interpolation with centripetal parameterization, as proposed by E. Lee [8].

$$u_0 = 0$$
$$u_i = u_{i-1} + \frac{\sqrt{\|\mathbf{P}_i - \mathbf{P}_{i-1}\|}}{\sum_{j=1}^{n} \sqrt{\|\mathbf{P}_j - \mathbf{P}_{j-1}\|}}$$
$$u_n = 1$$

The control point coordinates $\mathbf{Q}_i$ are derived from the B-spline curve formula and interpolation conditions (Eq. 6).

$$\mathbf{c}(u_k) = \sum_{i=0}^{n} N_{i,p}(u_k)\mathbf{Q}_i = \mathbf{P}_k \tag{6}$$

The knot values are chosen by the averaging technique, and the spline degree is fixed at $p = 3$. Therefore, the only unknowns remain the coordinates of the curve control points, and the whole problem is reduced to solving a linear system.

8. Design variables

Although the most flexible parameterization is clearly achieved by declaring each point from the set $\{\mathbf{R}_i\}$ as a design variable, such an approach does not appear to be practical as it may not preserve the "design intent" of one or another type of pattern. To keep the design space observable and to expose the familiar parameters (e.g., see p.6 in the paper by Larsson and Stahlbrand [7]) to human engineers, we prepared different parameter sets for various shapes of a pattern.

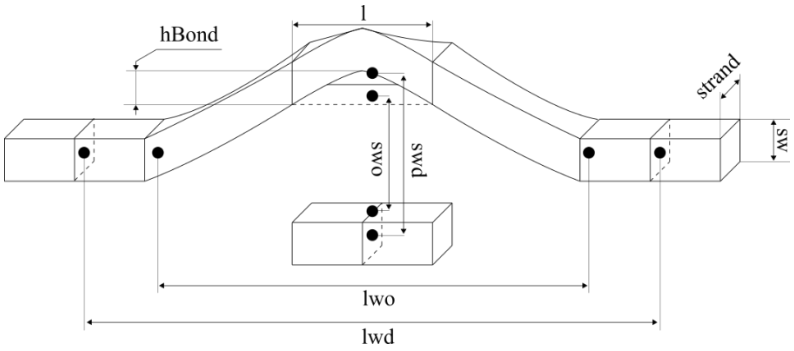


Fig. 7. ESM modeling parameters for the "rhomb" type of the pattern.

Design variables for the "rhomb" and "prism" types of the model are depicted in fig. 7 and fig. 8 respectively. Their meanings are briefly exposed in Tables 1 and Table 2 respectively.

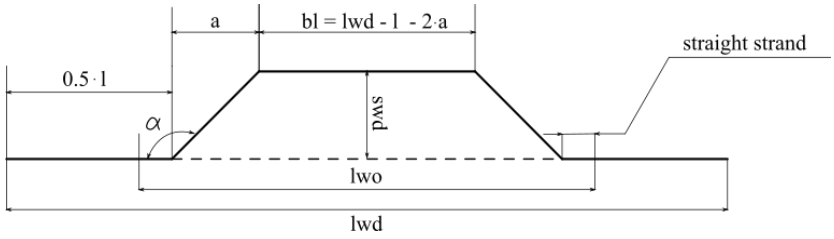


Fig. 8. ESM modeling parameters for the “prism” type of the pattern.

TABLE 1
"RHOMB" DESIGN VARIABLES

Parameter	Meaning	Unit
lwd	long-way design width	mm
swd	short-way design width	mm
sw	material thickness	mm
$strand$	strip width	mm
$hBond$	bond deformation	mm
l	length of the punch tip	mm

9. Shape optimization

Optimizing the pattern shape over the set of the exposed design variables $\mathbf{a} \in \mathbb{R}^N$ is the empirical way to prove that the constructed parametric model is flexible enough, i.e., it can capture the real form of a part. The typical use of the optimizer employs the following steps. First, the engineer prepares a representative laser scan model M of a minimal repeatable pattern (fig. 2). The system also expects that the "initial guess" model is provided, i.e., the human operator selects the components of $\mathbf{a}_0 \in \mathbb{R}^N$ in the parameter space. The task of the optimizer is to find another set $\mathbf{a}^*$ by incrementally refining the initial guess for the sake of minimization of the objective function being the sum of squared distances from scan points to the model $\mathrm{S}(\mathbf{a})$ (Eq. 7).

$$\mathrm{F} = \sum_i d_i^2 \to min \tag{7}$$

It should be noted that for successful optimization, the parametric model $\mathrm{S}(\mathbf{a})$ should deform *predictably* with respect to the scan data M. Assuming that both the scan data and the parametric model have their local coordinate systems associated with the centers of bonds, we algorithmically ensure that our parameterization keeps the origin point of the local axis frame intact.

We employ the Particle Swarm Optimization (PSO) technique [4,5] for the selection of $\mathbf{a}^*$. The choice of the optimization algorithm is driven by our previous positive experience with PSO in a somewhat similar "shape / point cloud" registration problem where it gave practically valuable results.

For the sake of simplicity, we limit the dimension of the search space with $N = 4$. We declare the parameters named $lwd, swd, hBond$ and l to be the design variables and define all other parameters as design constants. Table 2 summarizes the design parameters for a "prism" type of ESM. All parameters exposed to the optimizer as design variables are marked with asterisk.

TABLE 2
"PRISM" DESIGN VARIABLES

Parameter	**Meaning**	**Unit**
$lwd*$	long-way design width	mm
$swd*$	short-way design width	mm
sw	material thickness	mm
$strand$	strip width	mm
$hBond*$	bond deformation	mm
$l*$	length of the punch tip	mm
α	angle of the punch tip	degrees

For different types of expanded metals, various design variables can be locked or unlocked in the search space, depending on the a priori defined heuristics. For example, it might make no sense to vary the material thickness if this property of the sheet is fixed in the manufacturing shop.

For each particle in the swarm, the velocity $\mathbf{V}_{i+1}$ at the $(i+1)$-th step is defined according to the Eq. 8, where ξ_1, ξ_2 are random numbers and ω, c_1 and c_2 are empirically defined constants.

$$\mathbf{V}_{i+1} = \omega \mathbf{V}_i + \left(\mathbf{P}_{min}^{self} - \mathbf{P}_i\right) c_1 \xi_1 + \left(\mathbf{P}_{min}^{global} - \mathbf{P}_i\right) c_2 \xi_2 \tag{8}$$

The position of a particle in the space of parameters is obtained according to the Eq. 9.

$$\mathbf{a}_{i+1} = \mathbf{a}_i + \mathbf{V}_{i+1} \tag{9}$$

IV. Results of experiments

All geometric computations and optimization techniques presented in this paper were developed in the ecosystem of Analysis Situs open source CAD environment [10]. Curve interpolation methods have been reused from the open source Mobius SDK [1].

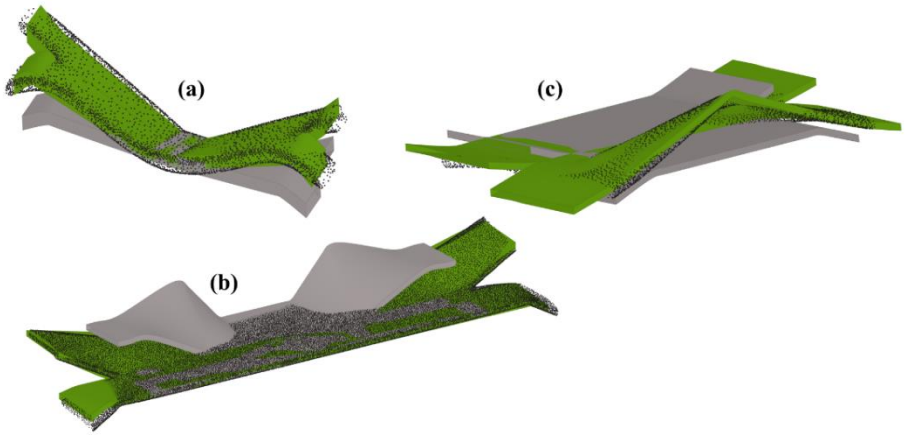


Fig. 9. Optimization results (green) for "armadillo" (a), "prism" (b) and "rhomb" (c) pattern types. The "initial guess" models are shown in grey. The scan data is depicted with black points. The centre point of a bond is a fixed as an anchor point for a local coordinate system: this point stays invariant over the optimization process.

Some typical optimization results are shown in Fig. 9 for "armadillo" and "prism" types of expanded sheet metals. It appears that reasonably good solutions are located quickly in about 5 steps (fig. 10), and the remaining steps do not improve the fitness value significantly. The fitness values can be computed either on the entire point cloud (fig. 10b,c) or on a randomly selected point set (fig. 10a) to speed up computations.

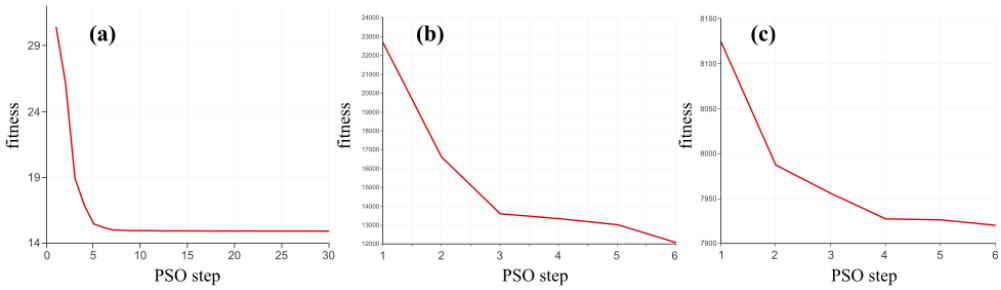


Fig. 10. Fitness values for each PSO step for the optimization results shown in fig. 9.

V. Discussion

The commonly used simplistic approaches to ESM modeling employ sketching the pattern shapes as series of lines and arcs, with subsequent extrusion. We presented a more accurate method of parameterization for expanded sheet metal parts. Our approach gives better flexibility in terms of the final shape thanks to the employed surface skinning apparatus. The obtained parametric models were matched with the provided laser scans of real products both manually and with the help of a global optimization loop. The achieved shapes look realistic enough, as it was confirmed by the commercial machining shop that industrialized the prototyped technology.

After many hours of experimenting with the parameters, we concluded that there is still a room for further automation, especially when it comes to the selection of parameters based on the initial scans. The computations conducted in the parameters space of the reduced dimension ($N = 4$) reveal that better accuracy calls for more parameters to be unlocked as design variables (e.g., α and swo, which stands for "short-way opening" distance and is not exposed as a design parameter). At the same time, exploring more design variables in an optimization loop may easily become computationally prohibitive, even with heuristic search methods such as PSO. Consequently, the directions for further research include working on computational efficiency.

Setting up the rectangular domain in the search space $\mathbb{R}^N$ and calibrating the PSO coefficients, such as ω, c_1 and c_2 is another problem that required human intervention for obtaining adequate optimization results. The effects of changing these constants on the convergence of the global optimizer are yet to be explored. It also appears that each type of sheet pattern ("rhomb", "prism", etc.) requires specific rules for setting up the initial guess and sensible optimization parameters.

VI. Conclusions

Modeling expanded sheet metal parts has received little attention in the CAD community, mostly because such parts are rarely considered functionally critical and are so idealized with elementary geometric models. In this study, we presented an accurate geometric modelling technique which is capable of constructing realistic shapes which better reflect tool properties and design parameters specified by an engineer. The automated workflow which was built on top of the presented method is as follows:

1. A set of design parameters is chosen by the user from a catalog. Each parameter has its “validity range” which is dependent of material properties and “know-how” accumulated in the machining shop.
2. The specified design parameters are passed to an executable, which constructs the “minimal repeatable” pattern according to the logic presented in this paper.
3. The constructed “minimal repeatable” is saved as a glTF file to be replicated as many times as needed to form the ultimate expanded sheet metal mesh.

The "modeler" executable, which runs at the second stage of the workflow, generates a "nominal" shape of the pattern in boundary representation. The created CAD model can be triangulated with differed mesh generation parameters to achieve the appropriate level of detail from this representation. This way, the geometric complexity (in the sense of "triangles budget") is kept under control by the target system. The PSO-based optimizer proved that the developed parameterization is "flexible" enough to represent the realistic geometry of a part. One takeaway from this research is that an optimization loop can only be set up for predictable parameterization, which means that all design variables must be independent, properly scaled, and convey intuitive "sensitives" in terms of their effect on final shape. Exploring the multidimensional parameter space (with $N = 4$ in the simplest setup) calls for high efficiency of the modeler. The use of B-spline interpolation functions results in decent performance because they reduce the entire problem to a series of linear systems to solve. As a result, the users of the modeler can quickly iterate over a number of design variants and choose the most appropriate ESM parameters based on visual feedback from the system.

References

1. Mobius API documentation. www.openmobius.org, accessed: 2024-07-04
2. Open CASCADE geometric modeling kernel. https://dev.opencascade.org/, accessed: 2024-06-21
3. Auriemma, S.: Development of parametric CAD models for gradient-based aerodynamic shape optimisation. Ph.D. thesis, Queen Mary University of London (2021)
4. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science. pp. 39–43 (1995). https://doi.org/10.1109/MHS.1995.494215
5. Eberhart, R., Shi, Y.: Comparison between genetic algorithms and particle swarm optimization. vol. 1447, pp. 611–616 (03 1998). https://doi.org/10.1007/BFb0040812
6. Lang, J., Massabo, A., Pyzak, D.: Practice of Computer Aided Geometric Design Cisigraph's Presentation, pp. 79–106. Vieweg+Teubner Verlag, Wiesbaden (1991)
7. Larsson, L., Stahlbrand, M.: Development of a framework for the design of expanded metal facades – Using artificial intelligence to streamline pre-production work. Master's thesis, Linkoping University (2022)
8. Lee, E.: Choosing nodes in parametric curve interpolation. Computer-Aided Design 21(6), 363–370 (1989)
9. Mangin, P.: Curve by smoothing N points (in French). Tech. rep., Matra Datavision (1995), https://quaoar.su/files/papers/occ/appdef variational pp2-35.pdf, accessed: 2024-06-21
10. Slyadnev, S., Malyshev, A., Turlapov, V.: CAD model inspection utility and prototyping framework based on OpenCascade. In: GraphiCon 2017. pp. 323–327. Perm, Russia (2017)

УДК 528.92:630
DOI: 10.25206/978-5-8149-3873-2-2024-573-581

Опыт внедрения и адаптации ИС AISPOL и ГИС в лесоустройстве
Experience of implementation and adaptation of AISPOL and GIS in forest management

Л.В. Черных, Д.В. Черных
Поволжский государственный технологический университет, пл. Ленина, д. 3, г. Йошкар-Ола, Россия

L.V. Chernykh, D.V. Chernykh
[1]Volga State University of Technology, Yoshkar-Ola, Russian Federation

Аннотация. Приведён обзор информационных систем и комплексов программ обработки лесоустроительной информации. Рассмотрена автоматизированная информационная система AISPOL, в которой интеграция подсистем с ГИС достигается на основе организационной, функциональной, технической, программной и информационно-лингвистической совместимости. Цель исследования заключается в совершенствовании автоматизации технологических процессов, связанных со сбором и представлением данных о лесах в цифровом виде при лесоустроительном проектировании с использованием AISPOL. Разработаны основные математические модели, составляющие ядро контроля измеряемых и исчисления расчетных показателей по лесоустроительной базе данных. Предложен алгоритм автоматического построения тематических лесных карт на лесничество или лесной участок по материалам совмещенной с геоинформацией повыдельной базой данных. Установлено, что производительность труда специалиста по созданию лесных тематических карт с использованием AISPOL в сравнении с автоматизированным способом повышается в 25…30 раз.

***Ключевые слова*:** автоматизированная обработка данных, лесоустройство, геоинформационные системы, тематические лесные карты, цифровизация лесного комплекса.

Annotation. The review of information systems and program complexes of forest inventory information processing is given. The automated information system AISPOL is considered, in which the integration of subsystems with GIS is achieved on the basis of organizational, functional, technical, software and information-linguistic compatibility. The aim of the research is to improve the automation of technological processes related to the collection and presentation of forest data in digital form in forest planning using AISPOL. The basic mathematical models that constitute the core of the control of measured and calculation of calculated indicators on the forest inventory database are developed. The algorithm of automatic construction of thematic forest maps for a forestry or a forest area based on the materials of the database combined with geoinformation has been proposed. It is established that the labor productivity of a specialist on creation of forest thematic maps using AISPOL in comparison with the automated method increases by 25...30 times.

Keywords: automated data processing, forest inventory, geographic information systems, thematic forest maps, digitalization of the forest complex.

I. Введение

Для решения задач по цифровизации лесного комплекса первоочередными направлениями являются сбор, структурирование, обработка, представление и хранение данных о лесах, а также оценка их ресурсного потенциала [1]. В настоящее время работы по сбору количественных и качественных данных о лесах проводятся в рамках лесоустройства и государственной инвентаризации лесов, лесопатологического мониторинга, отчетов по сохранению лесов с использованием различных отраслевых автоматизированных информационных систем и геоинформационных систем уже более 30 лет. За этот период было создано большое количество информационных систем, отраслевых геоинформационных систем, программно-аппаратных комплексов для сбора и обработки данных о лесах. К таким программным средствам можно отнести АРМ-Таксатора; ПК СОЛИ, ПК СОЛИ-2, СОЛИ-3 («Рослесинфорг»); ГИС «Лесфонд» (ООО «Лабмастер»); ГИС TOPOL-L (ООО "ЛесИС"); ПК «ЕСАУЛ» («Рослесинфорг»–«Севзаплеспроект»); PLP 2015 и Forest Book (ООО «Леспроект»); ПК «ЛесГИС» («Рослесинфорг» «Запсиблеспроект»); АИС ГЛР и др. [2-4]. Многие системы обработки лесоустроительной информации уже закончили жизненный цикл или находятся на последнем его этапе по причине отсутствия функций обработки, редактирования и представления пространственных данных о лесах, а также старения системного программного обеспечения и технических средств.

В настоящее время современные информационные системы используют независимые системы управления базами данных и геоинформационные системы для создания баз данных о лесах. Одной из таких систем является AISPOL [5, 6]. Данная система широко применяется для создания, актуализации и ведения лесоустроительных

баз данных, а также для проектирования мероприятий по использованию и сохранению лесов в рамках проектной документации различных уровней − от лесного участка до лесного плана субъекта Российской Федерации.

Автоматизированная система AISPOL состоит из трех подсистем[12] (рис. 1):

• АП-КТ – автоматизированная подсистема «Карточки таксации», используемая для ввода, контроля, корректировки лесоустроительной информации;

• АП-ДОЛП – автоматизированная подсистема «Документационное обеспечение лесоустроительного проектирования» используется для выполнения расчетов и печати первичных документов по запросам пользователя, по материалам таксации лесов и проектирования лесохозяйственных мероприятий по охране, защите и воспроизводству лесов;

• АП-КАРТЫ – автоматизированная подсистема «Картографическое обеспечение лесоустроительного проектирования», предназначенная для создания и печати лесных карт в среде ГИС Аксиома, MapInfo (или другой ГИС) по запросам пользователя по материалам таксации лесов и проектирования лесохозяйственных мероприятий по охране, защите и воспроизводству лесов.

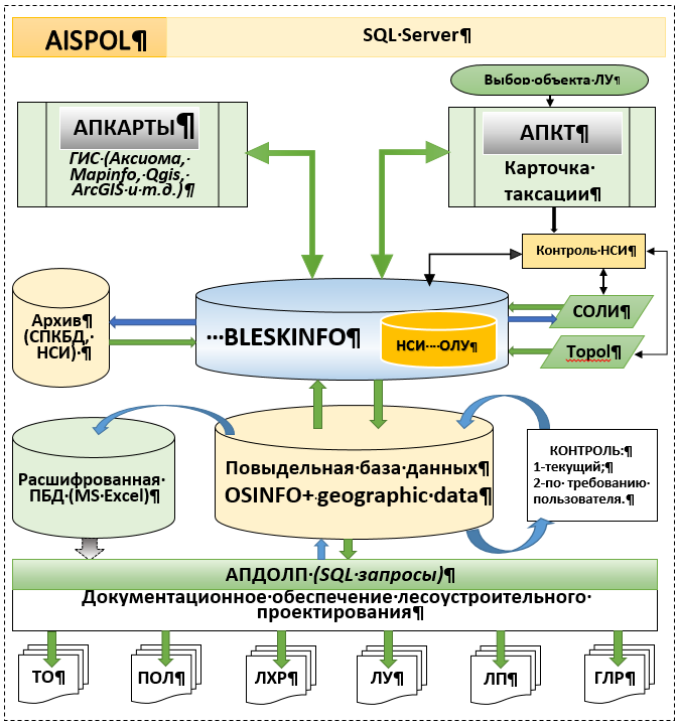


Рис.1. Схема взаимодействия подсистем в среде AISPOL [4]

Интеграция указанных подсистем и их органичное объединение в единую систему достигается на основе организационной, функциональной, технической, программной и информационно-лингвистической совместимости.

По данным Рослесхоза, на территории России действует более 75 тысяч договоров аренды лесных участков на площади в 237,7 млн га. На этой территории осуществляются мероприятия по сохранению и использованию лесов в соответствии с проектной документацией, которая приводится в проектах освоения лесов арендных участков и лесохозяйственных регламентах лесничеств[13]. Отметим, что проектная документация по освоению, охране, защите и восстановлению лесов обновляется один раз в 10 лет после проведения мероприятий по лесоустройству. Такое требование устанавливается лесным законодательством в части назначения мероприятий по сохранению лесов[14]. В составе лесоустроительной документации и проектов освоения на различные виды использования лесов имеются тематические карты, большинство из которых создаётся на основе атрибутивной повыдельной информации о лесах встроенными инструментами геоинформационных систем (ГИС), которые

[12] Автоматизированная обработка лесоустроительной информации AISPOL 2.5 Версия БД 3.1. Руководство пользователя. // Йошкар-Ола: ООО «Марлеспроект», 2021. 128 с. URL: https://disk.yandex.ru/i/6PC-LMmDMYEytg

[13] Приказ Минприроды России от 16.11.2021 № 864 (ред. от 07.08.2023) «Об утверждении Состава проекта освоения лесов, порядка его разработки и внесения в него изменений, требований к формату проекта освоения лесов в форме электронного документа» (зарегистрировано в Минюсте России 30.12.2021 N 66777).

[14] Приказ Минприроды России от 05.08.2022 № 510 «Об утверждении Лесоустроительной инструкции» (зарегистрировано в Минюсте России 30.09.2022 N 70328).

требуют ручной настройки на каждый объект лесоустройства исходя из уникального набора тематических данных, хранящихся в совмещенной атрибутивной и картографической базе данных.

Для создания картографических материалов и проектной документации при производстве лесоустроительных работ и проектировании сохранения и использования лесов также необходимы расчетные показатели, назначение лесохозяйственных мероприятий в рамках лесного законодательства по закрепленным в них формальным признакам[15], что, в свою очередь, требует высокой квалификации специалистов, значительных трудозатрат и отдельных мероприятий по контролю данного проектирования со стороны органов государственной власти.

II. Постановка задачи

Цель исследования − автоматизация технологических процессов, связанных со сбором и представлением данных о лесах в цифровом виде в ИС AISPOL.

Задачи:

- выполнить анализ нормативно-правовой базы, технологического процесса лесоустройства и лесного проектирования;
- создать алгоритм расчета необходимых показателей для проектирования по лесоустроительной базе данных;
- создать алгоритм назначения лесохозяйственных мероприятий по сохранению и использованию лесов;
- автоматизировать проектирование и контроль лесоустроительной базы данных;
- автоматизировать создание тематических лесных карт (ТЛК) в геоинформационных системах.

III. Теория

Назначение лесохозяйственных мероприятий и создание тематических лесных карт при лесоустроительном проектировании требует как определяемых при таксации лесов фактических показателей, так и расчетных показателей.

Для получения расчетных показателей используются:

а) сортиментные и товарные таблицы древостоев по субъектам Федерации и лесорастительным зонам: выход ликвидной и деловой древесины;

б) стандартные таблицы сумм площадей сечений для древостоев с относительной полнотой 1.0: относительная полнота древостоя, состав древостоя, запас древостоя;

в) нормативный возраст рубки лесообразующих пород в лесных районах субъектов Федерации[16], в том числе по бонитетам насаждений и целевому назначению лесов; группа возраста;

г) шкала классов бонитетов;

д) лесная типология: класс природной пожарной опасности[17].

Назначение лесохозяйственных мероприятий в части сохранения лесов проводится по формальным признакам соответствующих нормативно-правовых актов. Так, для назначения мероприятий по лесовосстановлению и уходу за несомкнувшимися лесными культурами руководствуются Правилами лесовосстановления[18]. Мероприятия, направленные на лесовосстановление, зависят от нормативного количества подроста ценных древесных пород и характера их пространственного распределения по площади. Данная информация, получаемая по итогам проведения мероприятий по таксации лесов или составлению проекта лесовосстановления, имеется в Государственном лесном реестре. Уходы за несомкнувшимися лесными культурами проектируются в соответствии с лесохозяйственными регламентами лесничеств и проводятся, как правила, в первые 3-5 лет. Заключительным мероприятием по уходу является «лесоводственный уход», который проводится в год перевода в покрытую лесом площадь лесных земель.

Мероприятия по уходу за лесами регламентируются Правилами ухода за лесом[19], где основанием для назначения рубок ухода являются такие таксационные показатели, как состав древостоя, группа типов леса, относительная полнота и возраст древостоя.

[15] «Лесной кодекс Российской Федерации» от 04.12.2006 N 200-ФЗ (ред. от 04.08.2023) (с изм. и доп., вступ. в силу с 01.01.2024)
[16] Приказ Рослесхоза от 09.04.2015 № 105 (ред. от 02.07.2015) «Об установлении возрастов рубок».
[17] Приказ Рослесхоза от 05.07.2011 № 287 «Об утверждении классификации природной пожарной опасности лесов и классификации пожарной опасности в лесах в зависимости от условий погоды» (зарегистрировано в Минюсте РФ 17.08.2011 N 21649.
[18] Приказ Минприроды России от 04.12.2020 № 1014 «Об утверждении Правил лесовосстановления, состава проекта лесовосстановления, порядка разработки проекта лесовосстановления и внесения в него изменений» (зарегистрировано в Минюсте России 18.12.2020 N 61556).
[19] Приказ Минприроды России от 30.07.2020 № 534 «Об утверждении Правил ухода за лесами» (зарегистрировано в Минюсте России 18.12.2020 N 61555).

Общий объем мероприятий по охране лесов от пожаров проектируется исходя из размера лесного участка или лесничества по целевому назначению и с учетом по площади защитных и эксплуатационных лесов[20]. Данные мероприятия проектируются точечно для каждого лесного участка, исходя из субъективных факторов, наличия и распределения лесной инфраструктуры, что на данном этапе развития информационных систем исключает их полную автоматизацию.

Санитарно-оздоровительные мероприятия назначаются по итогам лесопатологических обследований, относящихся к уровню ежегодных наземных обследований, что также отображается в проектной документации разных уровней.

Мероприятия по использованию лесов в части заготовки древесины проводятся в спелых и перестойных лесных насаждениях согласно Правилам заготовки древесины с учетом целевого назначения лесов и относительной полноты древостоев[21].

Тематические лесные карты и планы лесонасаждений формируются исходя из значений соответствующих атрибутивных данных о лесотаксационном выделе в соответствии с инструкцией о порядке создания и размножения лесных карт [7]. В 2017 году в ФГБУ «Рослесинфорг» были разработаны Временные технические указания по созданию и оформлению лесных карт, утвержденные приказом директора ФГБУ «Рослесинфорг» от 20.07.2017 № 320-П[22].

Перечень лесных тематических карт и необходимых атрибутов для их построения на объект исследования приведен в таблице 1.

ТАБЛИЦА 1
НЕОБХОДИМЫЕ АТРИБУТЫ ТЕМАТИЧЕСКИХ ЛЕСНЫХ КАРТ И ПЛАНОВ ЛЕСОНАСАЖДЕНИЙ ЛЕСНИЧЕСТВА, ЛЕСНОГО УЧАСТКА

Название лесной карты	Количество возможных вариаций тематики*	Необходимые атрибуты БД	
		фактические	расчетные
План лесонасаждений	34674	Преобладающая порода, категория земель	Группа возраста
ТК классов природной пожарной опасности	5	Тип леса	Класс пожарной опасности
ТК целевого назначения лесов	22	Целевое назначение лесов	٧
ТК особо защитных участков леса	18	Категория особо защитного участка	٧
ТК охраны объектов животного и растительного мира	54	Целевое назначение лесов и категория особо защитного участка	٧
ТК мероприятий по воспроизводству и уходу за лесами	122	нет	Лесохозяйственное мероприятие
ТК наличия очагов вредных организмов, загрязнений	160	Наличие вредных организмов, поврежденных насаждений	
ТК план рубок	45	нет	Лесохозяйственное мероприятие
Косметическая тематика	140	Древесные породы второго яруса, единичных деревьев, несомкнувшихся лесных культур под пологом	

** Количество возможных вариаций лесной тематики может изменяться*

Отметим, что самое большое число вариаций тематики имеет план лесонасаждений, а минимальное значение − ТЛК классов природной пожарной опасности.

[20] Приказ Рослесхоза от 27.04.2012 № 174 «Об утверждении Нормативов противопожарного обустройства лесов» (зарегистрировано в Минюсте России 07.06.2012 № 24488)
[21] Приказ Минприроды России от 01.12.2020 № 993 (ред. от 17.10.2022) «Об утверждении Правил заготовки древесины и особенностей заготовки древесины в лесничествах, указанных в статье 23 Лесного кодекса Российской Федерации» (зарегистрировано в Минюсте России 18.12.2020 № 61553)
[22] Временные технические указания по созданию и оформлению лесных карт. М.: ФГБУ «Рослесинфорг», 2017. 75 с.

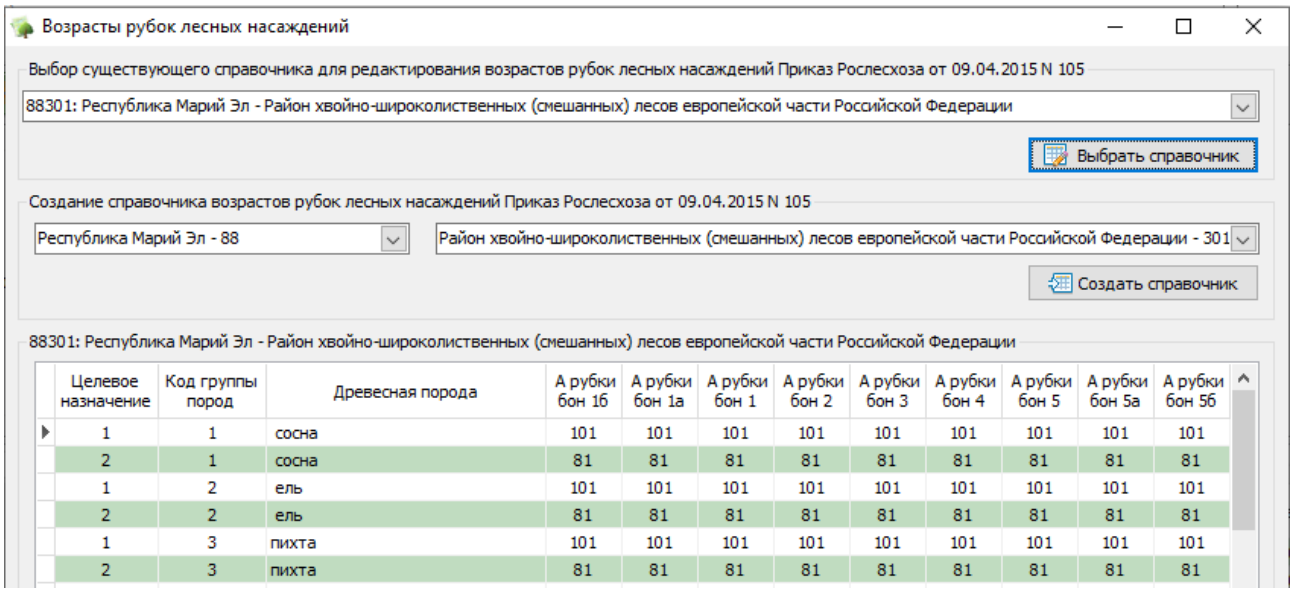
Для автоматического построения тематических лесных карт на лесничество или лесной участок нами реализован следующий алгоритм:

1. Выбор лесничества или лесного участка;
2. Логический контроль на корректность и адекватность характеристик таксационных выделов.
3. Выбор тематической лесной карты;
4. Установление ключевых атрибутов ТЛК объекта картографирования (таксационного выдела), i=1…n;
5. Выбор таблицы шаблонов ТЛК;
6. Установление фильтра значений по таблице шаблонов ТЛК;
7. Выбор и установление вида изображения для лесотаксационного выдела ТЛК (знаки, цветовая схема и т.д.) ;
8. Проверка условия по количеству объектов картографирования. Если i < n возвращаемся к п 4.
9. Переход в ГИС и обновление ТЛК;
10. Оформление ТЛК в среде ГИС и запись в файл или печать.

Вышеприведенный алгоритм создания тематических лесных карт реализован в автоматизированной информационной системе AISPOL [4].

IV. Результаты экспериментов

Для получения всех расчетных показателей производятся настройка и заполнение нормативно справочной информации (НСИ) базы данных AISPOL в виде программного кода и заполнения соответствующих справочников системы. Для примера приведем справочник «Возрасты рубок лесных насаждений» (рис. 2).



Целевое назначение	Код группы пород	Древесная порода	А рубки бон 1б	А рубки бон 1а	А рубки бон 1	А рубки бон 2	А рубки бон 3	А рубки бон 4	А рубки бон 5	А рубки бон 5а	А рубки бон 5б
1	1	сосна	101	101	101	101	101	101	101	101	101
2	1	сосна	81	81	81	81	81	81	81	81	81
1	2	ель	101	101	101	101	101	101	101	101	101
2	2	ель	81	81	81	81	81	81	81	81	81
1	3	пихта	101	101	101	101	101	101	101	101	101
2	3	пихта	81	81	81	81	81	81	81	81	81

Рис. 2. Пример справочника возрастов рубок

К главной НСИ относят стандартные таблицы сумм площадей сечений и запасов при полноте 1,0; сортиментные и товарные таблицы; шкалы классов бонитетов; типы леса, типы лесорастительных условий. НСИ по причине большого массива данных, их дискретности и имеющих строгие ограничения по максимальным значениям аргументов в AISPOL реализованы в виде математических моделей по всем лесорастительным зонам всех субъектов Российской Федерации по основным лесообразующим древесным породам.

Для примера приведем основные математические модели, составляющие ядро контроля измеряемых и исчисляемых расчетных показателей по лесоустроительной базе данных.

1) Класс возраста – возрастной интервал, устанавливаемый по возрасту преобладающей древесной породы основного яруса. Величина возрастного интервала (Δ_a) зависит от биологических особенностей древесной породы и составляет:

5 лет – для быстрорастущих древесных пород;

10 лет – для мягколиственных и твердолиственных порослевых;

20 лет – для хвойных и твердолиственных семенных;

40 лет – для кедра, бука и других долгорастущих древесных пород.

Класс возраста вычисляется по формуле

$$KA = \frac{A+(\Delta_a-1)}{\Delta_a}, \quad (1)$$

где A – возраст преобладающей древесной породы, лет;

Δ_a – возрастной интервал, лет;

KA – класс возраста, целая часть числового значения, ед.

2) Класс бонитета характеризует фактическую производительность насаждения. Величина класса бонитета (КБ) устанавливается с учетом происхождения по среднему возрасту и средней высоте преобладающей древесной породы основного яруса

$$KB = f(A, H, PR), \quad (2)$$

где A – средний возраст, лет;

H – средняя высота, м;

PR – происхождение и скорость роста групп древесных пород: 1 – семенное; 2 – порослевое; 3 – быстрого роста.

Значения высот для конкретного класса бонитета (индекс класса бонитета) моделируется по формуле

$$H_i = H_{max} \cdot [1 - \exp(-c_1 \cdot A)]^{c_2}, \quad (3)$$

где c_1 – параметр роста в высоту;

c_2 – параметр формы кривой;

A – возраст, лет;

H_{max} – асимптотическое значение высоты, м;

H_i – значение высоты i класса бонитета в возрасте A, м.

Предлагаемая функция для моделирования БШ является трехпараметрической и имеет наименьшую ошибку аппроксимации в сравнении с другими аналогичными функциями [8]. Модель корректна и адекватна экспериментальным данным. Коэффициент корреляции моделирования высоты по классам бонитетов выше 0,955.

3) Определение запаса древостоя элемента леса, яруса. При глазомерно-измерительном способе таксации лесов запас древесины определяется по сумме площадей поперечного сечения стволов деревьев элемента леса (древесной породы), определенной на реласкопических площадках или круговых площадках постоянного радиуса, и видовой высоте, вычисленной по математическим моделям или таблицам, входящим в состав нормативно-справочной информации, предусмотренной главой IV Лесоустроительной инструкции[23], по следующим формулам:

$$M_{эл} = \sum g_{эл} \cdot \text{Н}F_{эл}, \quad (4)$$

$$M_{я} = \sum M_{эл}, \quad (5)$$

где $M_{эл}$ – запас древесины по элементу леса на 1 га, м^3;

$M_{я}$ – запас древесины по ярусу на 1 га, м^3;

$\sum g_{эл}$ – среднее арифметическое по элементу леса значение суммы площадей поперечного сечения стволов деревьев на 1 га по данным измерений на реласкопических площадках или перечетных площадках постоянного радиуса (количество площадок указано в пункте 202 лесоустроительной инструкции), м^2.

Сумма площадей сечений по элементу леса при полноте 1,0 рассчитывается по формуле

$$\Sigma g = a_0 \cdot [1 - \exp(-a_1 \cdot h)]^{a_2}, \quad (6)$$

где h – средняя высота элемента леса, м;

a_0 и a_1 – коэффициенты уравнения, ед.

Н$F_{эл}$ – видовая высота элемента леса древостоя яруса лесного насаждения, значение которой вычисляется по математическим моделям или по таблицам, входящим в состав нормативно-справочной информации, предусмотренной главой IV лесоустроительной инструкции, м.

Видовая высота по элементу леса рассчитывается по формуле

$$HF = a_0 + a_1 \cdot h, \quad (7)$$

где h – средняя высота элемента леса, м;

a_0 и a_1 – коэффициенты уравнения, ед.

[23] Приказ Минприроды России от 05.08.2022 № 510 «Об утверждении Лесоустроительной инструкции» (зарегистрировано в Минюсте России 30.09.2022 № 70328)

После вычисления запасов элементов леса формула состава яруса определяется в соответствии с пунктом 252 Лесоустроительной инструкции.

4) Товарные таблицы представлены в виде электронного справочника «Сортиментные и товарные таблицы для таксации лесов России» [9].

5) Тип лесорастительных условий (ТЛУ) устанавливается в полевых условиях.

ТЛУ – лесоводственная классификационная единица, объединяющая сходные лесорастительные условия лесных земель, которые обеспечивают произрастание лесной растительности определенного состава и производительности (ОСТ 56-108-98) [10].

6) Тип леса (ТЛ) устанавливается в полевых условиях.

ТЛ − лесоводственная классификационная единица, объединяющая леса с однородными лесорастительными условиями определенного типа с соответствующим им породным составом древостоя, другой растительностью и фауной (ОСТ 56-108-98) [10].

7) Тип ландшафта (ТЛФ) устанавливается в полевых условиях в соответствии со шкалой групп и типов ландшафтов.

8) Тип почв (ТП) – основная классификационная единица, характеризуемая общностью свойств, обусловленных режимами и процессами почвообразования и единой системой основных генетических горизонтов (ГОСТ 27593-88) [11].

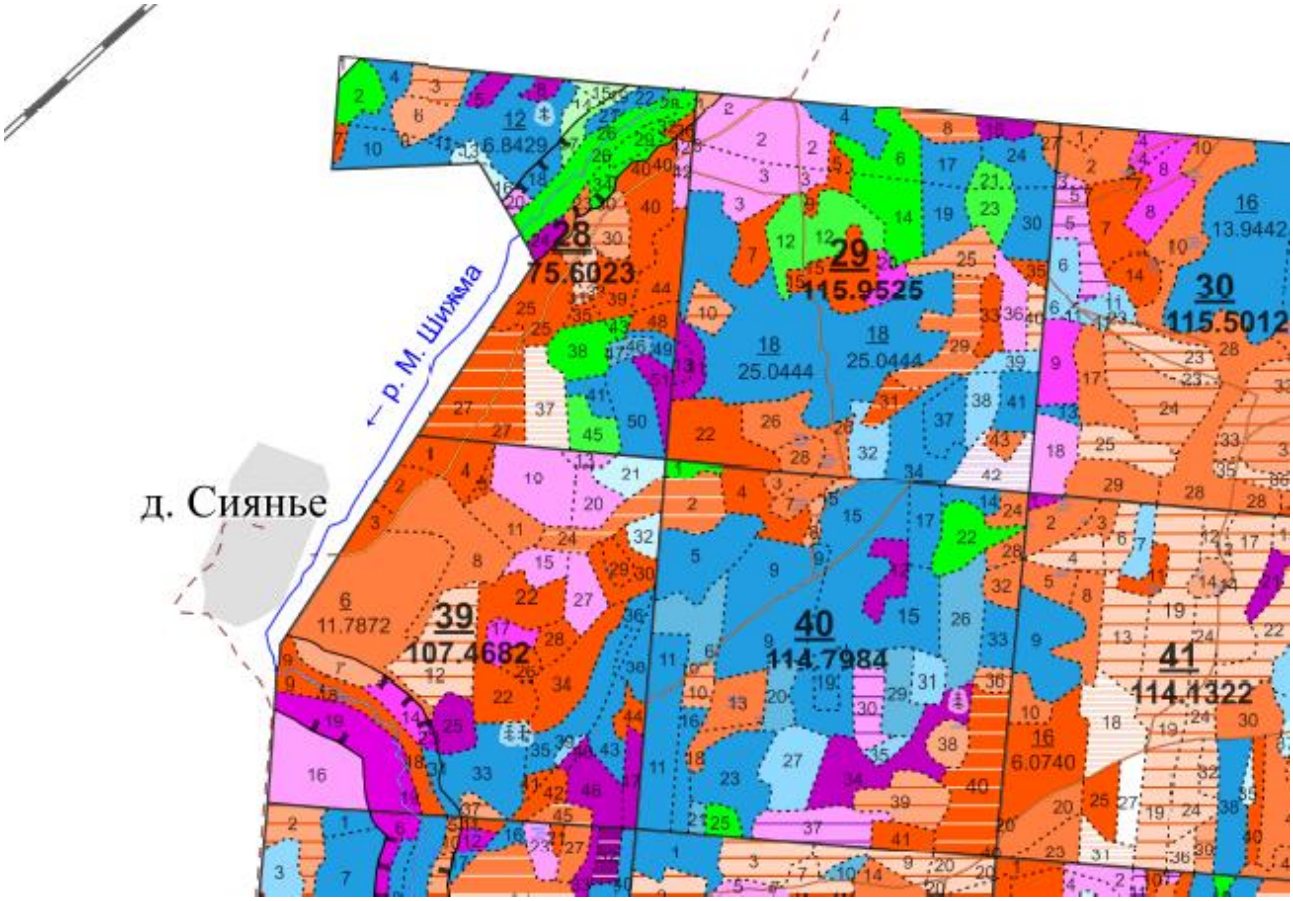


Рис. 3. Фрагмент плана лесонасаждений

С использованием таких таблиц и моделей автоматизирован контроль и расчет таксационных показателей древостоев объекта исследований и выполняется автоматическое построение тематических лесных карт (рис. 3).

V. Обсуждение результатов

Автоматизация технологических процессов при лесоустройстве и лесном проектировании в системе AISPOL потребовало структурирование большого количества нормативно-справочной информации, разработки и реализации в программном коде множества алгоритмов расчета и заполнения баз данных, что в конечном итоге

оправдано повышением производительности труда и увеличением достоверности и оперативности представления информации о лесных экосистемах и их ресурсном потенциале.

В среднем на лесном участке имеется около 1400 возможных вариаций сочетания различных атрибутов по всем тематическим лесным картам. На ручную настройку тематики по одному атрибуту в среде ГИС специалист затрачивает около одной минуты рабочего времени. Таким образом, для создания тематических лесных карт на один лесной участок высококвалифицированному специалисту необходимо затратить три рабочих дня. Такие расчеты включают в себя только создание соответствующей тематики в геоинформационной системе. По нашему мнению, подобная оценка являются консервативной. На практике, согласно приказу ФГБУ «Рослесинфорг» «Об установлении цены нормативного человеко-дня при выполнении работ на 2023 год», полное оформление одной карты с зарамочным текстом и легендой в среднем может занимать до 0,9 рабочего дня. Отметим, что создание одной лесной карты с использованием AISPOL специалисту необходимо затратить до 15 минут. Следовательно, производительность труда специалиста по созданию лесных тематических карт с использованием AISPOL в сравнении с автоматизированным способом повышается в 25…30 раз.

Результаты проведенной работы дают представление о дальнейшем развитии информационных систем и цифровизации лесного комплекса. Ограничения автоматизированного проектирования и представления информации на данном этапе заключаются в особенностях лесного законодательства, несовершенстве методов получения информации о лесах и закрытости имеющейся информации в Государственном лесном реестре на повыдельном уровне.

Хотя вся автоматизация и проводилась согласно лесному законодательству, атрибуты и критерии для принятия решений по назначению и проведению лесохозяйственных мероприятий в части сохранения лесов основаны на многолетнем опыте лесоводов и действуют на всей территории Российской Федерации. Нормативное назначение мероприятий по сохранению лесов может давать и отрицательный эффект для лесных экосистем, но данная особенность заключается в несовершенстве лесного законодательства и малой изученности динамики лесных экосистем в условиях антропогенного воздействия, так, например, оборот рубки древостоев может достигать целого столетия.

Дальнейшее развитие САПР AISPOL с совершенствованием геоинформационных систем и повышением вычислительных мощностей аппаратных средств дает возможность создания таких аналитических технологий как цифровые двойники лесного фонда с моделированием динамики количественных и качественных показателей лесного фонда и его ресурсного потенциала.

VI. Выводы и заключение

Автоматизация вычисления расчетных таксационных показателей, назначение всего комплекса лесохозяйственных мероприятий по нормативному признаку, создание тематики для лесных карт, использование функций САПР AISPOL и геоинформационных систем способно повысить качество проектной лесоустроительной документации и увеличить производительность труда высококвалифицированных специалистов в области лесоустройства и лесного проектирования.

AISPOL также позволяет производить контроль, импорт и экспорт лесоустроительных баз данных с геоинформацией, что может использоваться как инструментарий для наполнения Федеральной государственной информационной системы лесного комплекса (ФГИС ЛК).

Список литературы

1. Носков В.А. Цифровизация лесного комплекса, учитывающая качество и характер использования древесины // Россия: тенденции и перспективы развития. 2021. № 16-2. С. 233-237.
2. Программное обеспечение в лесном хозяйстве и при лесозаготовках /А.Н. Заикин, В.В. Сиваков, В.В. Никитин, А.А. Брионес // Лесной вестник / Forestry Bulletin. 2023. Т. 27, № 4. С. 172–184. DOI: 10.18698/2542-1468-2023-4-172-184.
3. Северский С.И., Малышева Н.В., Пьянков И.Б. Проблемы внедрения информационно- коммуникационных технологий в лесное хозяйство и комплексные решения // Вестник МГУЛ – Лесной вестник. 2015. № 2. С. 144-152.
4. Дышлюк С.С., Николаева О.Н., Ромашова Л.А. Об использовании экологических карт в создании экологической компоненты инфраструктуры пространственных данных // Геодезия и картография. 2016. № 4. С. 18–25. Doi: 10.22389/0016-7126-2016-910-4-18-25.
5. Свидетельство о государственной регистрации программы для ЭВМ 2020612490 Российская Федерация. Автоматизированная обработка лесоустроительной информации AISPOL / Д.В. Черных, Л.В. Черных, В.Л.

Черных; заявитель и правообладатель Общество с ограниченной ответственностью «Марлеспроект». № 2019667371; заявл. 19.12.2019; опубл. 25.02.2020. 1 с.
6. Лесовосстановление: монография / Л.В. Черных, Д.В. Черных, В.Л. Черных, С.А. Денисов. Москва; Вологда, Инфра-Инженерия, 2023 г. 216 с. ISBN: 978-5-9729-1154-7
7. Инструкция о порядке создания и размножения лесных карт / Государственный комитет СССР по лесному хозяйству //Москва: ЦБНТИ, 1987. 80 с.
8. Черных, В. Л. Автоматизированная актуализация лесотаксационных нормативов (на примере лесов Урала). дис. д-ра сельскохоз. наук: специальность: 06.03.02. Йошкар-Ола, 2002. 401 с.
9. Электронный справочник «Сортиментные и товарные таблицы для таксации лесов России» / В.Л. Черных, Д.М. Ворожцов, Чернов А.П., Баранов А.Ф., Щевцов Б.П., Войнов Г.С., Шевелев С.Л. // Международное сотрудничество в лесном секторе: баланс образования, науки и производства: материалы международной конференции. Йошкар-Ола: МарГТУ, 2009. С. 233-237.
10. ОСТ 56-108-98. Лесоводство. Термины и определения. М.: ВНИИЦлесресурс, 1998. 56 с.
11. ГОСТ 27593-88 (СТ СЭВ 5298-85). Государственный стандарт Союза ССР. «Почвы. Термины и определения» (утв. и введен в действие Постановлением Госстандарта СССР от 23.02.1988 № 326). – 9 с.

Источник финансирования.

Работа выполнена при финансовой поддержке ООО «Марлеспроект».
Научный руководитель: Черных Валерий Леонидович д-р с.-х. наук, проф. каф. Лесоводства и лесоустройства, Поволжский государственный технологический университет.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-582-589

Оптимизация процесса коммуникации разработчиков 3D-моделей
Optimizing the communication process of 3D model developers

К.А. Асямов
ПАО «ОНХП», г. Омск, Россия

K.A. Asyamov
PJSC 'ONHP', Omsk, Russia

Аннотация. В статье рассматривается проблема коммуникации разработчиков 3D-моделей. Для решения проблемы был предложен способ оптимизации коммуникации, сокращающий временные и ресурсные затраты. Была предложена замена форматов, передающих только данные о геометрии 3D-моделей, на формат передачи веб-приложений, что освобождает пользователей от использования дополнительного ПО при визуализации 3D-графики. Дано обоснование применению HTML как портативного формата для передачи данных о 3D-моделях. Произведено расширение структуры HTML-документа путем добавления новых элементов, позволяющих описывать передаваемые данные о 3D-моделях.

***Ключевые слова*:** 3D-моделирование, HTML-формат, веб-приложения, коммуникация разработчиков, передача геометрии, оптимизация взаимодействия

Аннотация. The paper deals with the problem of communication of 3D model developers. To solve the problem, a way to optimize communication that reduces time and resource costs was proposed. It was proposed to replace formats that transmit only data about 3D model geometry with the web application transmission format, which frees users from using additional software when visualizing 3D graphics. A justification is given for the use of HTML as a portable format for 3D model data transmission. The structure of HTML-document is extended by adding new elements that allow describing the transferred data about 3D-models.

***Ключевые слова*:** 3D modelling, HTML format, web applications, developer communication, geometry transfer, interaction optimization

I. Введение

В связи с тем, что разработка моделей геометрических объектов в большинстве случаев является комплексным процессом, требующим взаимодействия многих специалистов, для компаний крайне важно внедрение актуальных программных средства внутренней и внешней коммуникации. Из-за ориентированности решений на области, не связанные с геометрическим моделированием, разработчики 3D-моделей сталкиваются с трудностями в объяснении своих концепций и получении обратной связи.

Проблема коммуникации разработчиков 3D-моделей заключается в невозможности одновременно и эффективного формирования отчета со стороны разработчика 3D-модели, и удобного, с обратной связью, просмотра геометрии заказчиком.

Цель работы заключается в предложении решения данной проблемы.

Согласно гипотезе исследования, HTML-формат может быть эффективно применен для передачи данных 3D-моделей. В данном случае может быть реализована отправка полноценного веб-приложения, не требующего дополнительного ПО (программное обеспечение) для воспроизведения файлов с 3D, что способствует повышению эффективности обмена информацией разработчиков 3D-моделей.

Результат работы представляет собой библиотеку для приложения КОМПАС-3D, реализующую автоматизированное формирование отчетов, включающих в себя 3D-модель. Кроме возможности просмотра 3D-моделей, создаваемые HTML-отчеты предоставляют пользователям возможности, дополнительно упрощающие процесс коммуникации, такие как добавление текстовых и графических комментариев, а также передача данных отчета без использования внешних приложений.

Актуальность исследования обусловлена динамичным развитием технологий и возрастающей конкуренцией в различных отраслях, где геометрическое моделирование является ключевым элементом.

Решением проблемы занимались сообщества разработки ПО с открытым исходным кодом [1, 2] а также компании, такие как Slack Technologies, Microsoft Corporation, Zoom Video Communication, Google. Так, были разработаны различные инструменты для взаимодействия, стандарты обмена данными, были улучшены

интерфейсы существующих ПП (программный продукт), а также была улучшена их интеграция. Решения проблемы коммуникации разработчиков 3D-моделей касаются различные исследования [3, 4]. При этом она в полной мере не решена в отношении существующих CAD-систем и иного вида ПО.

В настоящее время процесс просмотра, комментирования и отправки отчетов производится в различных приложениях, большая часть программных продуктов, способных быть примененными как альтернативы HTML, не доступны в России, для пользователей, получающих файлы с геометрией, необходим поиск воспроизводящих 3D программ, приобретение их и, соответственно, дополнительная трата личного времени, невозможен нативный просмотр геометрии в существующих 3D-форматах с мобильных устройств, универсальное решение обозначенной проблемы отсутствует.

Предложенное решение выгодно отличают интерактивные возможности HTML5. Передача данных в виде веб-приложения делает передаваемые данные доступными для чтения, а также способными быть представленными в предельно удобном для потребителя информации виде.

Методологической основой работы являлись исследования в областях проектирования программных систем [5, 6] и компьютерной графики [7, 8, 9]. В работе применялись спецификации и документации технологий HTML5 [10], WebGL 2.0 [11], языков программирования c# [12], JavaScript [13].

II. Постановка задачи

Решение проблемы коммуникации участников разработки 3D-моделей заключается в предоставлении им возможности передачи геометрии в формате, доставляющим минимум неудобств, являющимся наиболее автономным, не требующим от участников коммуникации использования специализированного ПО для работы с геометрией. Вместо стандартных для области компьютерного моделирования STEP, M3D, OBJ и так далее, предлагается использовать HTML как формат, обладающий наибольшей совместимостью с различными платформами, предоставляющий возможность воспроизведения графического содержания отчета на мобильном устройстве, встраивания в отчет форм для ввода комментариев и дачи обратной связи при проверке отчета.

Цель работы – оптимизировать процесс коммуникации разработчиков 3D-моделей. Для достижения цели работы требуются решения следующих задач: выполнить анализ недостатков существующих решений для коммуникации, применяемых в области 3D-моделирования; выполнить анализ технологий, потенциально применимых для решения проблемы коммуникации разработчиков 3D-моделей; произвести оценку востребованности, эффективности и применимости предлагаемого решения проблемы; выполнить описание его архитектуры, определение технологий и подходов, которые должны быть использованы для разработки решения; разработать прототип решения проблемы коммуникации разработчиков 3D-моделей.

III. Теория

Существует два основных варианта решения задачи по генерации HTML-документов с геометрией – это разработка либо внешнего конвертера, либо аддона для существующего приложения по 3D-моделированию. В первом случае на создание документа тратится гораздо больше времени, чем при использовании решения, интегрированного в ПП, с которым работает разработчик 3D-модели. В результате целесообразной оказывается доработка существующего ПО. В качестве такового был выбран КОМПАС-3D, отечественный продукт, применяемый для моделирования технических объектов в различных областях промышленности, предоставляющий все необходимые средства для программирования приложений.

Модуль Kompas-3D Invisible представляет собой документированное API, набор интерфейсов и программных модулей, позволяющих реализовывать создание dll и rtw ActiveX-библиотек для КОМПАС-3D, разработанных на языках Delphi, Visual Basic, C++ или C#. Выбор языка программирования в данном случае не играет существенной роли, тем не менее наиболее перспективным и современным из представленных является C#, в данном случае его выгодно отличает богатая экосистема сторонних библиотек, а также достаточно лаконичный синтаксис.

Решение задачи рендеринга графики в браузере требует определения API для обращения к видеочипу устройств, в которых будет воспроизводиться контент HTML-отчетов. Подобными веб-интерфейсами являются WebGL и WebGPU. Последний является более производительным решением, предоставляет возможности Vulkan, Direct3D 12 и Metal, однако при этом в отношении поддержки браузерами уступает конкуренту. WebGPU по умолчанию отключен почти во всех браузерах. WebGL 2.0 является стандартом, способным запускать трёхмерную графику в вебе на всех платформах без исключения. Так, в качестве API для обращения к графическому процессору вычислительного устройства был выбран WebGL.

Обоснование применения HTML-формата для сохранения геометрии представлено в таблице 1. Для расчета сжатия использовалась модель без текстур, содержащая 16919 полигонов и 8666 вершин.

ТАБЛИЦА 1
СРАВНЕНИЕ ФОРМАТОВ ПРЕДСТАВЛЕНИЯ ГЕОМЕТРИИ

Формат	Области применения	Способ формирования	Сжатие модели	Поддержка	Сохраняемые данные
M3D	Машиностроение, проектирование, 3D-печать	Генерация в ПО для моделирования	13642 KB	Низкая	Геометрия модели, текстуры, цвет
OBJ	Компьютерное моделирование	Генерация в ПО для моделирования	2819 KB	Высокая	Геометрия модели, текстуры, цвет
STEP	Машиностроение, проектирование	Генерация в ПО для моделирования	24557 KB	Высокая	Геометрия модели, текстуры, цвет
STL	Проектирование, 3D-печать	Генерация в ПО для моделирования	3461 KB	Высокая	Только геометрия модели
PDF	Обмен информацией	Внешняя конвертация, платные аддоны для ПО	739 KB	Adobe Acrobat	Геометрия модели, текстуры, цвет
HTML	Обмен информацией	Внешняя конвертация, предлагаемый аддон для ПО	3945 KB	Chrome, Firefox, Safari, Edge, Opera и т. д.	Модели, иерархия объектов, параметры сцены, материалы, текстуры
glTF	Обмен информацией	Внешняя конвертация	1720 KB	Высокая	Модели, иерархия объектов, параметры сцены, материалы, текстуры
DWG	Машиностроение, проектирование	Генерация в ПО для моделирования, внешняя конвертация	9723 KB	AutoCAD, CorelCAD, BricsCAD, nanoCAD, Caddie	Сборки моделей, текстовые аннотации, чертежи, спецификации
DXF	Обмен информацией, дизайн	Генерация в ПО для моделирования, внешняя конвертация	13423 KB	Большинство существующих CAD-приложений	Чертежи, модели, текстура, цвет
FBX	Обмен информацией, дизайн, проектирование	Генерация в ПО для моделирования, внешняя конвертация	3207 KB	Продукты компании Autodesk	Модели, иерархия объектов, параметры сцены, материалы, текстуры, анимация

Сравнение форматов представления геометрии и ПО для работы с ними показывает, что в настоящее время имеется как область неудовлетворенных потребностей разработчиков 3D-моделей, так и возможность для решения имеющихся проблем. Реализация генерации HTML-отчетов может позволить участникам разработки более тесно взаимодействовать, зависеть при этом от меньшего числа внешних факторов и быть в большей степени сосредоточенными на выполнении своих рабочих задач.

В результате формат HTML является наиболее перспективным средством решения проблемы коммуникации разработчиков 3D-моделей. Подавляющее число форматов, в которых могут быть сохранены данные о 3D-моделях, отличаются тем, что для их воспроизведения пользователь обязан обладать специфическим для него программным обеспечением, что вынуждает его приобретать дорогостоящие лицензии, тратить время на понимание специфики работы в выбранных приложениях. Если от участников коммуникации требуется лишь ознакомление с материалами проекта и дача обратной связи, то применение стандартных форматов передачи 3D-моделей доставляет участникам взаимодействия неудобства.

Все вышеперечисленное говорит о целесообразности использования выбранных технологий для решения проблемы коммуникации разработчиков 3D-моделей.

Решение поставленной задачи осуществляется путем разработки системы (рис. 1), которую образуют компоненты для извлечения параметров 3D-моделей и генерирования интерактивных отчетов. Для формируемых HTML-документов далее описаны необходимые функциональные модули, делающие файлы с геометрией пригодными для эффективной коммуникации разработчиков 3D-моделей.

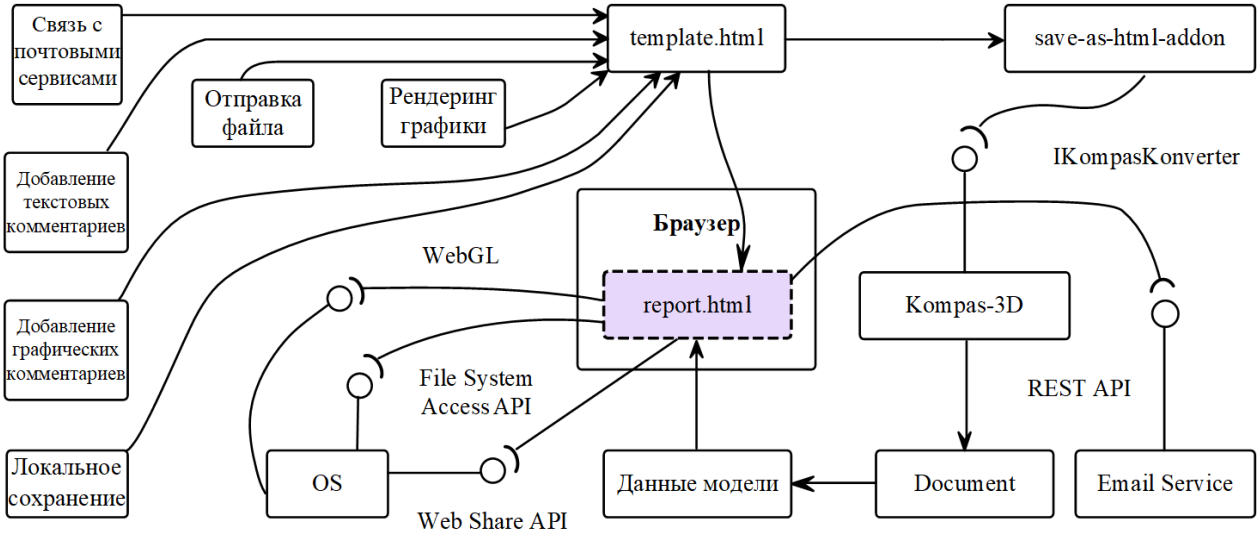


Рис. 1. Схема компонентов системы

Проектируемая библиотека, элемент «save-as-html-addon», подключается к приложению, элемент «Kompas-3D», так как она реализует необходимый интерфейс – IKompasKonverter [14], применяемый для добавления новых форматов сохранения чертежей и геометрии в КОМПАС-3D. Разработанная библиотека (см. рисунок 1) содержит шаблон HTML-документа, «template.html», для которого реализуется форма для работы с электронной почтой, отправки и изменения файла, возможность добавления текстовых и графических комментариев, а также рендеринг 3D-модели. Приложение, элемент «Kompas-3D», включает в себя создаваемую пользователем 3D-сцену, элемент «Document», и содержит данные о геометрии разрабатываемой модели. При объединении шаблона страницы и данных о геометрии происходит формирование HTML-документа, «report.html», который далее может быть открыт пользователем как веб-приложение.

Полученный текстовый файл сохраняется с расширением «HTML». Воспроизведение браузером отчета происходит при интерпретации данных, распределенных в HTML-файле так, как показано на рисунке 2:

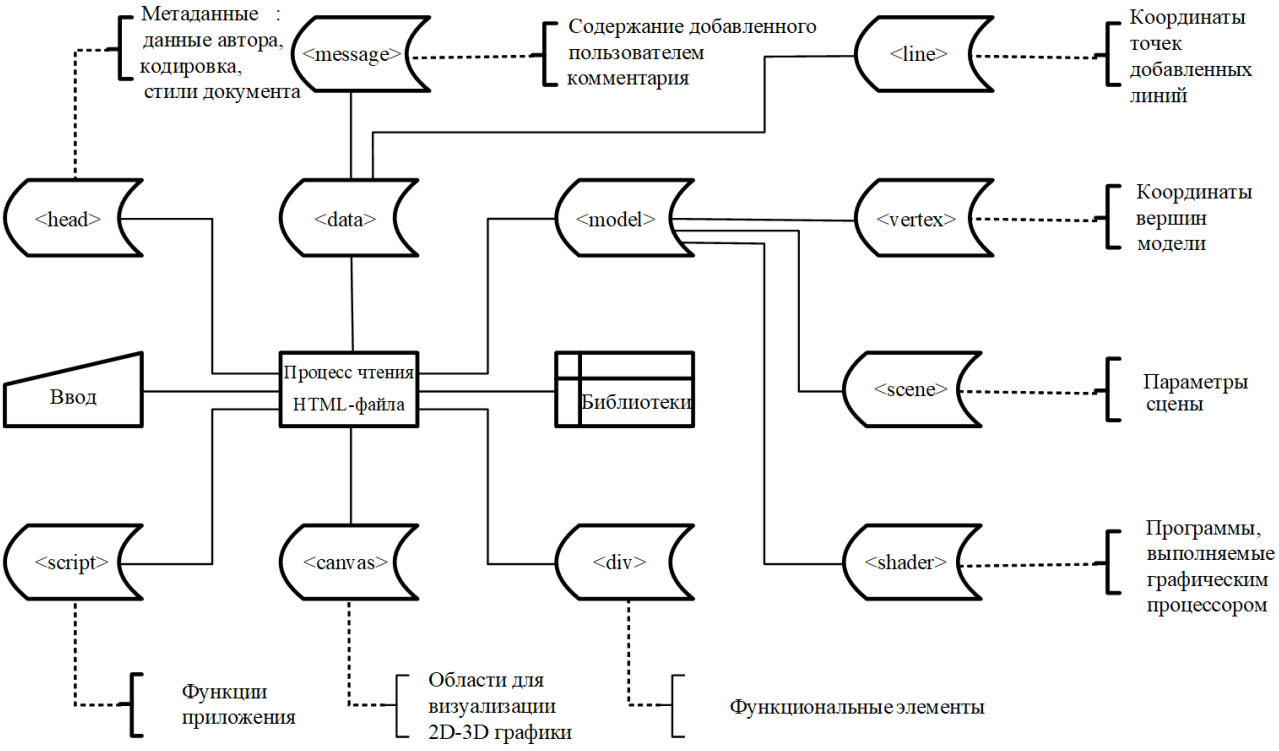


Рис. 2. Структура HTML-документа

После чтения данных (рис. 2) в процессе воспроизведения файла отчета запускается конвейер WebGL, определяющий последовательность вершинных и пиксельных шейдеров и промежуточных процедур для рендеринга кадров.

Важным аспектом реализации проекта (рис. 1) является интеграция решения с внешними сервисами. Они позволяют связать различные компоненты системы воедино, обеспечивая беспрепятственный обмен данными и повышая эффективность работы. В качестве сервиса для отправки писем по электронной почте был выбран emailjs [15]. Для обращения к операционной системе воспроизводящего устройства были использованы доступные для веб-приложений Web Share API, File System Access API [10].

IV. Результаты экспериментов

Результатом выполнения поставленных задач является аддон для приложения КОМПАС-3D, который реализует сохранение 3D-моделей в HTML-формате. Использование аддона конструктором позволяет сократить время на формирование отчета, а отправка HTML-файла сокращает время на просмотр и дачу обратной связи заказчиком.

Применение аддона позволяет конструктору и заказчику обмениваться информацией о проекте большим числом способов, чем им было доступно ранее. Так достигается удобство коммуникации, сокращаются временные издержки. Способы взаимодействия проиллюстрированы на рисунке 3:

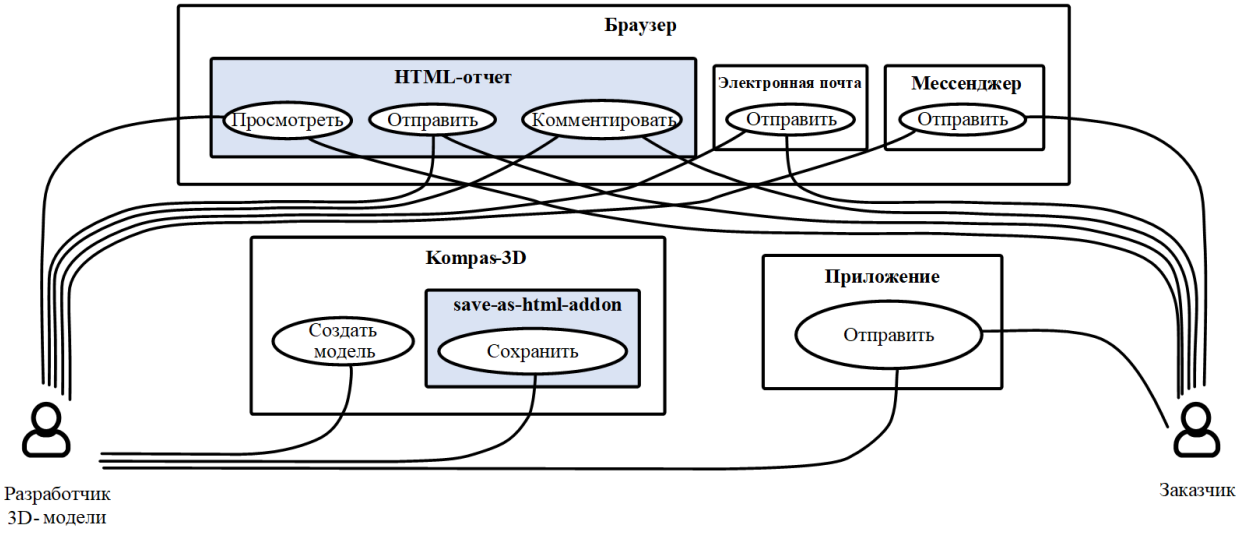


Рис. 3. Схема способов использования решения

Процесс использования библиотеки начинается с разработки геометрической модели в приложении КОМПАС-3D, после чего пользователь вызывает окно сохранения проекта, где выбирает добавляемую аддоном опцию формирования HTML-отчета, см. рис. 4:

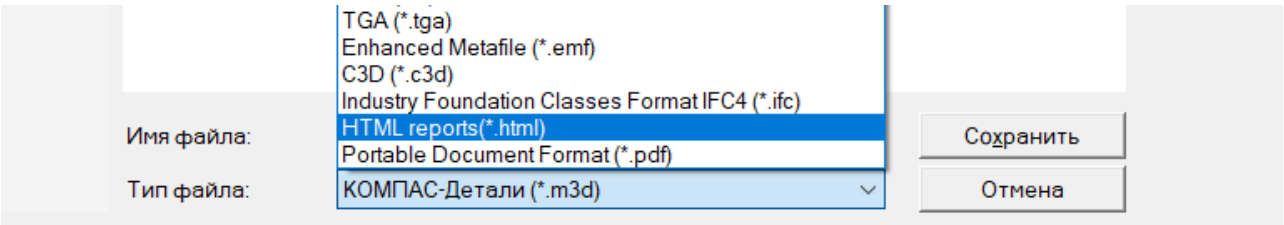


Рис. 4. Панель сохранения модели в КОМПАС-3D

Далее разработчик может отправить отчет как обычный файл через мессенджер или приложение электронной почты либо воспользоваться встроенной в отчет функцией передачи данных. Благодаря HTML-формату получатель отчета приобретает возможность его чтения на мобильном устройстве, а также возможность дачи обратной связи без использования дополнительных приложений.

Апробация результатов исследования производилась в среде моделирования инженерами элементов трубопроводных систем в ПАО «ОНХП». Для передачи использовались экспортированные 3D-модели, для которых разрабатываются семейства в ПП Autodesk Revit. Применение аддона подтвердило возможность,

благодаря использованию HTML-формата, упростить процесс взаимодействия коллег отдела, избавить их от необходимости дополнительно перемещаться либо подключаться к чужому компьютеру при обсуждении вопросов, касающихся избыточности геометрии, ее соответствия стандартам и требованиям ведения каталога оборудования изделий и материалов компании. Вместе с этим на практике была подтверждена функциональность всех реализованных модулей и производительность рендеринга 3D-моделей при Full HD и 2K-разрешении экрана. Передаваемые 3D-модели, такие как заглушки, отводы, отступы, переходы, тройники, крестовины и подобные им элементы трубопроводных систем, из-за специфики своего применения, отличаются предельной простотой геометрии. В рассматриваемом случае (рис. 5) 3D-модель представляет собой турбореактивный двигатель в разрезе, она разрабатывалась в целях получения качественных изображений модели, содержащей 236629 вершин, 491651 полигонов. При открытии отчета в браузере получатель видит следующий интерфейс:

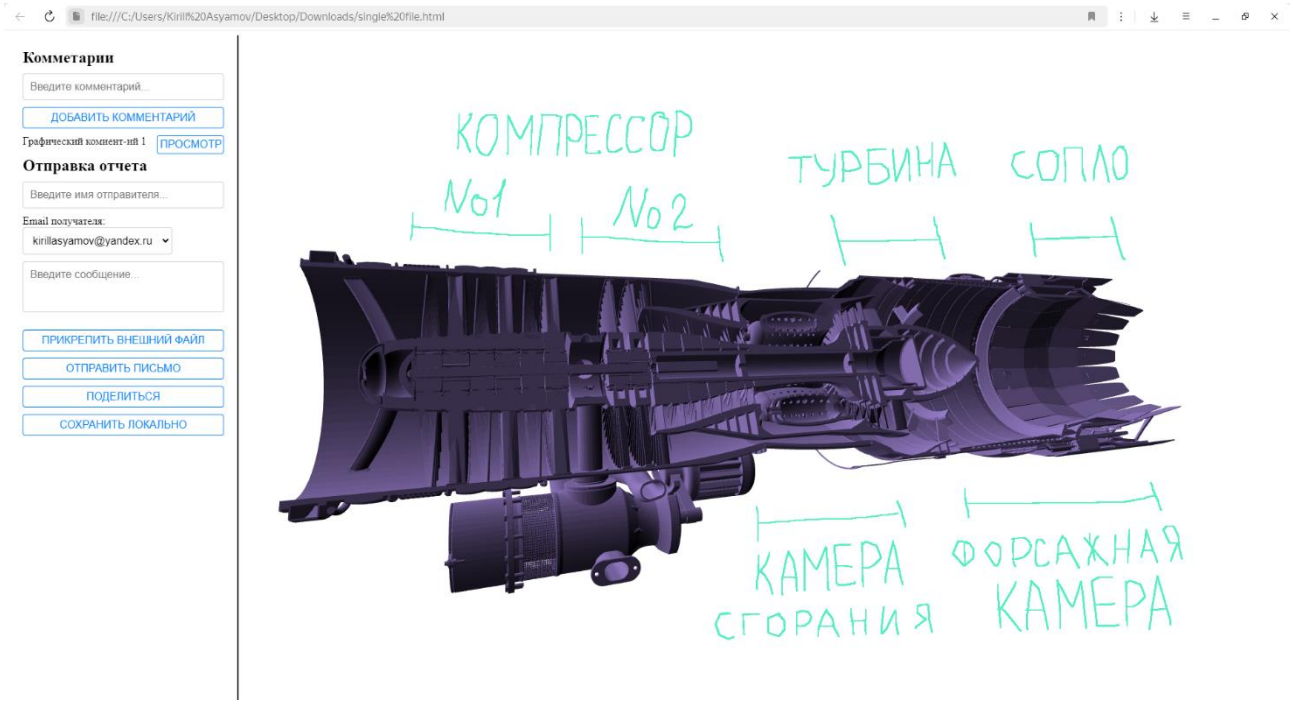


Рис. 5. HTML-отчет в браузере

В данном примере разработчиком геометрии были добавлены графические пояснения к модели (рис. 5). При ее вращении комментарии будут автоматически скрыты, их отображение производится только в плоскости, где они были изначально созданы. При формировании отчета была добавлена электронная почта автора для дачи обратной связи (рис. 5), так получатель файла лишается необходимости ввода почты отправителя, а также своей благодаря тому, что для отправки электронных писем используется соответствующий сервис.

В данном случае от участника коммуникации необходим только ввод письма и список дополнительно передаваемых файлов, возможно добавление текстовых и графических комментариев в отчет.

Примеры использования создаваемого аддона включают его применение в учебном процессе в университетах, в процессе проектирования, при предложении товаров и обсуждении технического задания с потребителем, при публикации отчетов на форумах, а также при использовании сгенерированных HTML-файлов в личных целях.

Результаты применения решения демонстрируют его работоспособность, функциональность примененных веб-технологий и возможность практического применения решения.

V. Обсуждение результатов

Разработанный прототип библиотеки для приложения КОМПАС-3D демонстрирует возможность сохранения и передачи геометрии в HTML-формате, возможность воспроизведения 3D-графики с частотой 60 кадров в секунду без снижения разрешения экрана воспроизводящих устройств. Проведенное исследование показало работоспособность и применимость выбранных веб-технологий в различных областях.

Применение HTML-отчетов с геометрией способно в наибольшей степени оптимизировать процесс коммуникации там, где данный процесс отличают высокие темпы, где имеет место привлечение новых клиентов, не обладающих навыками чтения чертежей и потенциально не владеющими ПО для работы с геометрией, также там, где отсутствует жесткий регламент взаимодействия участников разработки 3D-моделей.

Во-первых, существует множество фриланс-бирж, где размещаются заказы по моделированию разного рода технических объектов. При этом важным является поддержание связи с заказчиком и предоставление промежуточных результатов.

Во-вторых, представители конструкторских бюро также заинтересованы в поддержании активной связи с заказчиками, постоянной отправке отчетов, согласовании результатов, демонстрировании деталей перед их изготовлением, например, перед печатью на 3D-принтере.

В-третьих, возможность передачи HTML-отчетов может быть востребована в учебном процессе в университетах, где имеет место постоянная необходимость в проверке и обсуждении студенческих работ, а также, где относительно часто применяется КОМПАС-3D, для которого и создается данная библиотека.

Кроме того, возможность хранения геометрии в HTML-формате позволяет пользователям КОМПАС-3D распоряжаться плодами своей профессиональной деятельности недоступными ранее способами. HTML-формат предоставляет больше удобств в аспекте хранения, анализа и демонстрации моделей, например, позволяет прикреплять геометрию на форумах, посвященных разработке геометрии моделей, где ранее было необходимо прикреплять PDF-документы либо скриншоты, формирование которых отнимало много времени.

VI. Выводы и заключение

В работе был рассмотрен способ решения проблемы коммуникации участников разработки геометрических моделей. В связи с отсутствием эффективного способа реализации данного процесса было предложено передавать HTML-файлы с геометрией, генерируемые в приложении КОМПАС-3D.

Была рассмотрена архитектура библиотеки для реализации генерации HTML-файлов в приложении КОМПАС-3D, которая может являться основой для разработки аналогичных расширений для ПП по 3D-моделированию. Был разработан прототип решения проблемы коммуникации разработчиков 3D-моделей, была проверена его применимость, были проведены тесты производительности. В результате была подтверждена работоспособность всех реализованных функциональных модулей HTML-отчета и показано их положительное влияние на процесс коммуникации разработчиков 3D-моделей.

Разработка и внедрение библиотеки представленной архитектуры должны оказать положительное влияние на подлежащий трансформации процесс коммуникации участников разработки геометрических моделей. Переход от передачи как обычных изображений, так и STL-, STEP-, OBJ- или M3D-файлов, к передаче HTML-документов является шагом в направлении совершенствования важных для области геометрического моделирования бизнес-процессов, повышения адаптивности современных компаний, применяющих КОМПАС-3D.

Выбранный формат выгодно отличает как наличие уникальных интерактивных возможностей HTML5, так и доступность для чтения, что позволяет 3D-моделированию быть удобным одновременно и для изготовителя, и для потребителей.

Таким образом, разработка и внедрение библиотеки должны быть выгодны как пользователям приложения, так и работодателям, внедряющим КОМПАС-3D и заинтересованным в повышении производительности производства, а также компании, разрабатывающей само приложение КОМПАС-3D, заинтересованной в расширении возможностей продукта и повышении его конкурентоспособности.

Рассмотренный способ оптимизации взаимодействия разработчиков 3D-моделей может быть рекомендован для его реализации в различных ПП для 3D-моделирования не только в КОМПАС-3D. HTML-формат делает просмотр генерируемых отчетов доступным более широкому кругу пользователей, чем форматы представления 3D-моделей, используемые в настоящее время. Веб-технологии позволяют сделать визуализацию 3D-графики и технической документации интегрированной в отчет вместе элементами интерфейса, предоставляющими пользователям функциональность, упрощающую взаимодействие разработчиков 3D-моделей.

Источник финансирования. Благодарности

Научный руководитель – Панчук Константин Леонидович, Омский Государственный Технический Университет.

Список литературы

1. Репозиторий проекта WebGL [Электронный ресурс]. – 2024 – Режим доступа: https://github.com/KhronosGroup/WebGL, – дата обращения: 10.06.2024.

2. Репозиторий проекта WebGPU [Электронный ресурс]. – 2024 – Режим доступа: https://github.com/webgpu, – дата обращения: 10.06.2024.

3. Чувиков Д.А., Феоктистов В.П. СРАВНИТЕЛЬНЫЙ АНАЛИЗ 3D ФОРМАТОВ ХРАНЕНИЯ ДАННЫХ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ И СИСТЕМАХ ВИРТУАЛЬНОЙ РЕАЛЬНОСТИ // Автоматизация и управление в технических системах. – 2014. – № 4; URL: auts.esrae.ru/12-222 (дата обращения: 15.07.2024).

4. Веретёхин А.В. Актуальные цели и практические результаты управления цифровыми трансформациями коммерческих организаций / А. В. Веретёхин, А. V. Veretyokhin // Экономика. Информатика. – 2023. – № 3. – С. 539-551. — ISSN 2687-0932. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/journal/issue/346904 (дата обращения: 30.07.2024). – Режим доступа: для авториз. пользователей.

5. Гамма Э., Хелм Р., Джонсон Р., Влиссидес Дж. Приемы объектно-ориентированного проектирования. Паттерны проектирования. Пер. с англ. Москва: Вильямс, 2001. 368 с.

6. Басс Л., Клементс П., Кацман Р. Архитектура программного обеспечения на практике. — СПб.: Символ-Плюс, 2004. — 560 с.

7. Никулин, Е. А. Компьютерная геометрия и алгоритмы машинной графики / Е. А. Никулин. — Санкт-Петербург: Издательство «БХВ-Петербург», 2003. — 552 с.

8. Эйнджел Э. Интерактивная компьютерная графика. Вводный курс на базе OpenGL, 2-е изд. — М.: Издательский дом «Вильямс», 2001. — 592 с.

9. Боресков А.В. Программирование компьютерной графики. Основные понятия и алгоритмы. Метод трассировки лучей. Современный OpenGL. – Москва: ДМК Пресс, 2019. – 480 с.

10. Описание стандарта HTML [Электронный ресурс]. – 2024 – Режим доступа: https://html.spec.whatwg.org/, – дата обращения: 2.05.2024.

11. Спецификация WebGL [Электронный ресурс]. – 2024 – Режим доступа: https://registry.khronos.org/webgl/specs/latest/2.0/, – дата обращения: 2.05.2024.

12. Документация по языку программирования C# [Электронный ресурс]. – 2024 – Режим доступа: https://learn.microsoft.com/ru-ru/dotnet/csharp/, – дата обращения: 2.05.2024.

13. Спецификация стандарта ECMAScript 262 [Электронный ресурс]. – 2024 – Режим доступа: https://262.ecma-international.org/14.0/index.html?_gl=1*5b43p7*_ga*MTIwODc2ODg3Ny4xNzIyMjg4MDg4*_ga_TDC K4DWEPP*MTcyMjI4ODA4OC4xLjAuMTcyMjI4ODA4OC4wLjAuMA..#sec-intro , – дата обращения: 02.05.2024.

14. Официальная страница модуля KOMPAS-Invisible – основная информация, руководство пользователя [Электронный ресурс] : URL: https://kompas.ru/solutions/developers/kompas-invisible/ (дата обращения 20.12.2023).

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-590-599

Разработка и внедрение интерактивного электронного технического руководства с интеграцией технологий дополненной реальности для оптимизации производственных процессов

Design and Deployment of interactive electronic technical manuals with Augmented Reality Integration for Enhanced Production Efficiency

Н. М. Ермолаева
CompSkills, Нижний Новгород, Россия

N. M. Ermolaeva
CompSkills, Nizhny Novgorod, Russia

Аннотация. Несмотря на значительную цифровизацию и автоматизацию производственных линий, уникальные знания операторов продолжают играть ключевую роль в цехах. Операторы обеспечивают гибкость и оперативность в своих действиях, а также способны автономно реагировать на основе своих технических знаний и опыта. Преодоление разрыва между современными цифровыми системами создания и управления информацией и работниками на производстве является актуальной задачей как для научного сообщества, так и для инженеров-технологов. Разработка и внедрение системы автоматизации проектирования с применением интерактивного электронного технического руководства (ИЭТР) могут помочь в решении данной задачи. Одним из современных способов доставки ИЭТР операторам на рабочем месте является использование дополненной реальности (ДР), которая позволяет визуализировать инструкции в поле зрения оператора с помощью цифровых изображений, анимации и текстовых указаний. В рамках настоящей работы предложен комплекс инструментальных средств, ориентированных на решение задачи подготовки ИЭТР и использовании его на производстве. В частности, разработан оригинальный инструмент, который позволяет определить оптимальную последовательность разборки изделия и включить в себя вспомогательную информацию о выполняемых действиях. В дополнение к нему разработан второй инструмент — приложение ДР, предназначенное для отображения подготовленных инструкций на мобильном устройстве в контексте окружающей реальной среды, например, в условиях производственного процесса. Результаты могут быть представлены как в виде ИЭТР на экране, так и в виде приложения ДР на мобильном устройстве. Предложенный подход продемонстрирован в прототипе процесса сборки станка. Ожидается, что представленные решения и подходы способствуют повышению эффективности и точности производственного процесса.

***Ключевые слова*:** индустрия 4.0, 3D-моделирование, дополненная реальность, адаптивность ИЭТР, промышленная эксплуатация, система автоматизированного проектирования, сборка

Аннотация. Despite significant digitalization and automation of production lines, unique operator knowledge continues to play a crucial role in workshops. Operators provide flexibility and responsiveness in their actions, as well as the ability to autonomously react based on their technical knowledge and experience. Overcoming the gap between modern digital systems for creating and managing information and production workers is a pressing task for both the scientific community and engineers. Developing and implementing a design automation system using interactive electronic technical guidance (IETG) can help solve this problem. One of the modern ways to deliver IETG to operators at the workplace is through the use of augmented reality (AR), which allows for visualizing instructions in the operator's field of vision using digital images, animations, and text prompts. Within the framework of this study, a complex of instrumental tools has been proposed to address the task of preparing IETG and its use in production. Specifically, an original tool has been developed that allows for determining the optimal disassembly sequence of a product and including auxiliary information about the actions performed. In addition, a second tool - an AR application - has been developed to display prepared instructions on a mobile device in the context of the surrounding real environment, for example, in the conditions of a production process. The results can be presented both as IETG on a screen and as an AR application developed for mobile devices. The proposed approach was demonstrated in a prototype of a machine assembly process. It is expected that the presented solutions and approaches will contribute to improving the efficiency and accuracy of the production process.

***Ключевые слова*:** industry 4.0, 3D modeling, augmented reality, adaptiveness, interactive electronic technical manuals, industrial operation, computer-aided design, assembly.

I. Введение

Как описано авторами работы [1], что поскольку цифровизация влияет на большинство методов современного производства, то же самое происходит и с передачей технических инструкций. Предпринимаются попытки перейти от традиционных бумажных руководств, в которых отсутствует взаимодействие и механизмы обратной связи, к цифровым подходам. Современные приложения с 3D визуализацией инструкций, поддерживающие взаимодействие, воспринимаются легче, тем самым оператор не получает инструкций, для понимания которых требуется сильная техническая подготовка, как это было в случае с традиционными методами.

ДР — это пример новой технологии, которая способствует улучшению доставки информации оператору и своевременному выполнению поставленных задач. Этот новый подход к визуализации инструкций нашел благодатную почву в производстве в различных приложениях. В работе [2] отмечается, что цель ДР не ограничивается добавлением объектов в реальную среду, но также включает в себя скрытие объектов из нее в режиме реального времени при этом предоставляя пользователю возможность взаимодействовать с виртуальным миром. Виртуальная реальность (ВР) и ДР находят широкое применение в различных секторах экономики, включая образование, здравоохранение, промышленность, индустрии развлечений (игровую, кинематографическую, спортивные трансляции и шоу-бизнес), социальные сети, маркетинговые кампании, торговлю, недвижимость, а также оборонно-промышленный комплекс.

В области образования, как было продемонстрировано в работе [3], ДР проявляет себя как мощный инструмент поддержки обучения в контексте процедур промышленного обслуживания. Благодаря возможности непосредственной связи инструкций по выполнению задач обслуживания с деталями машины, подлежащими обработке, ДР-технологии могут быть использованы для создания эффективных систем обучения навыкам технического обслуживания и сборки. Это, в свою очередь, позволяет ускорить процесс обучения технических специалистов и освоение новых процедур технического обслуживания.

В области здравоохранения, в работе [4] описывается применение хирургической системы смешанной реальности, показывающей трехмерные модели анатомических структур в виде голограмм, связанных с объектами реального мира, через очки ДР. Для указания расположений объектов реального мира используются искусственные маркеры, закрепляемые на теле пациента. Авторы подчеркивают, что ключевые компоненты систем смешанной реальности должны уметь определять положение реальных объектов, принимать и обрабатывать полученную информацию, а также визуализировать объекты ДР в комплексе с реальными объектами.

В промышленности технология ДР успешно применяется для оптимизации производственных процессов, таких как сборка, и предотвращения ошибок. Как отмечено в работе [2], приложения ДР, разработанные для производства, помогают моделировать и улучшать производственные операции, а также предлагать решения в случае возникновения ошибок.

Сборка является одной из ключевых областей исследований в производстве приложений ДР, где внедрение этой технологии привело к значительным улучшениям. Сравнительный анализ использования традиционных бумажных инструкций по сборке и экрана с ДР показал, что время завершения, уровень ошибок и основная рабочая нагрузка сборщиков были существенно улучшены.

Кроме того, исследования подтверждают, что эффект от внедрения ДР в сборке тем более значим, чем сложнее производственный процесс. Пошаговая визуализация инструкций по техническому обслуживанию, ранее представленных в бумажном виде, также является перспективной областью исследований, демонстрируя улучшения во времени выполнения задач и уменьшение количества ошибок.

В работе [5] отмечается, что сложные ручные операции характеризуются низкой эффективностью и высокой вариабельностью качества, что делает их узким местом в интеллектуальном производстве. Поэтому разработка эффективных и надежных вспомогательных методов для операторов на месте является обязательным условием для снижения требований к опыту и улучшения эффективности работы, а также для стабилизации качества.

Одним из таких методов является ИЭТР, которое представляет собой цифровой ресурс, состоящий из текстов, изображений, динамических эффектов, 3d-моделей и другой информации, доступной на планшетах, промышленных компьютерах и других устройствах отображения. По сравнению с традиционными бумажными инструкциями, ИЭТР является более интуитивным, интерактивным, информативным и понятным, что позволяет снизить временные требования и усилия операций. Широкое применение ИЭТР наблюдается в таких отраслях, как авиация, автомобилестроение, тяжелое оборудование, электроника и другие.

Однако ИЭТР имеет ограничения, связанные с необходимостью переключения внимания оператора между виртуальной информацией и реальной средой. В связи с этим, технология ДР была признана многообещающей для производства, поскольку она может синхронизировать информационное пространство с физическим

пространством и отображать информацию, соответствующую задаче, в режиме реального времени, что позволяет снизить умственную нагрузку операторов и когнитивную дистанцию.

Кроме того, было показано, что помощь в операциях на основе ДР может сократить время выполнения и значительно снизить частоту ошибок. Однако, несмотря на свои преимущества, ДР в промышленности еще не достигла своего полного потенциала. Одна из ключевых технических проблем такого руководства - это плохая адаптивность, известная в литературе как возможность повторного использования, применимость, масштабируемость и гибкость. В идеальном случае использование ДР в промышленности должно легко настраиваться при изменениях в составе персонала, продуктах, технологических процессах и условиях труда. Кроме того, подобного рода приложения должны быть спроектированы с учетом возможности удобной доставки разработчиками и развертывания конечными пользователями без необходимости в перепрограммировании. Отличаясь от электронных и традиционных бумажных рабочих инструкций, здесь предлагается новый формат рабочих инструкций, организующий контент ДР и обеспечивающий поддержку и руководство для работников на производственном объекте.

Адаптивность процедуры разработки приложения ДР в промышленности, по мнению авторов [5], играет решающую роль в обеспечении масштабируемости системы. С этой целью предлагается наполнить ИЭТР в ДР локальными объектами и изменяющимся контекстом, которые должны динамически обновляться в процессе выполнения. Кроме того, приведены результаты исследований применения ИЭТР для типовых промышленных операций разного типа и сложности, которые показывают, что затраты времени и рабочая нагрузка приемлемы для инженеров и операторов. Такое приложение должно иметь основу для быстрого проектирования и развертывания. При этом самое сложное — спланировать, какие локальные объекты нужны и где их разместить. Также в работе приведены результаты исследования общности структуры. Подтверждено, что инженеры и операторы проектировали различные инструкции к работе в ДР без участия программиста в разумные сроки и с разумной рабочей нагрузкой. Кроме того, технические и управленческие требования промышленных операций могут быть удовлетворены путем контроля нормальной работы и проверки качества. Но некоторые ограничения все же существуют. Во-первых, невозможно осуществить проектирование инструкций к работе в ДР без испытаний на производстве, т.к. необходимо наполнение локальных объектов в реальных условиях. Во-вторых, живое создание индикаторов и отслеживание изображений по заполнителям действительно занимает некоторое время в ходе опытной эксплуатации. В-третьих, если требуется создание динамической 3d-модели, инженеру все равно придется выполнить эту работу.

Кроме того, немало работ, в которых описано применение на практике ИЭТР.

К примеру, в работе [6] описано применение разработанного ИЭТР для ремонта турбинных лопаток. В ходе эксперимента ручной станок для шлифовальных работ был оснащен дополнительным оборудованием, включающим: сенсорный монитор, монитор для отображения содержимого ДР, двумя камерами и сканером. На основе отсканированного серийного номера турбинной лопатки цифровые рабочие инструкции отображались на сенсорном экране и на AR мониторе. Авторы провели сравнительное исследование времени выполнения задачи и удобства использования ИЭТР с бумажным рабочим процессом, опираясь на опыт десяти опытных слесарей.. Результаты показали, что применение ИЭТР устраняет необходимость в печати, поиске, хранении и извлечении бумажных документов и избавляет исполнителя от необходимости переключения между двумерными рабочими инструкциями и выполнением задачи на трехмерном физическом объекте.

В работе [7] идет речь об использовании цифрового двойника для конкретных объектов, который использует веб-приложение для ИЭТР и технологию ДР для представления пространственной информации. Цифровой двойник технически обслуживаемого изделия может получить доступ к данным из существующих информационных систем в цехе, таких как система управления производством, планирование ресурсов предприятия и управление жизненным циклом продукции. Доступность цифровой информации об объекте обеспечивает новые вспомогательные функции, такие как динамические рабочие инструкции и визуализация дефектов с помощью ДР, доступ к которым осуществляется через человеко-машинный интерфейс.

В работе [8] приведена методология рационализации опытно-промышленных процессов подтвержденная успешной реализацией десяти реальных продуктов. Авторы показали, что существует зависимость между выбранными для создания инструкций средствами визуализации в условиях технологической подготовки производства и характером производственных процессов.

В связи с тем, что в индустрии 4.0 наблюдается сильная тенденция к автоматизации, что также сообщается как в работе [8], так и в работе [9], авторами предлагается помощь инженерам-технологам, т.е. лицам, ответственным за создание рабочих инструкций на заводе, в выборе адекватной последовательности сборки и быстром создании инструкций по сборке для операторов.

Несмотря на динамичное развитие исследований, научных разработок и реализаций авторских инструментов в России (ИТОРУМ, IGA SYSTEMS, SeaProject и т.д.) и за рубежом (Dassault Systèmes, SAP, VISCOPIC Steps,

VKS, Cortona3D и др.) многие компании предпочитают разрабатывать собственное программное обеспечение с целью создания ИЭТР, и улучшения эффективности производственного процесса.

Согласно приведенным работам, в контексте САПР создание и внедрение комплексов программных средств для поддержки производственных процессов приобретает особую значимость. Анализ результатов внедрения и адаптации САПР в различных отраслях промышленности свидетельствует о высокой эффективности таких систем в оптимизации производственных процессов, уменьшении ошибок и улучшении качества продукции, что подтверждает актуальность разработки и внедрения таких систем в современной промышленности.

II. Постановка задачи

Целью работы является автоматизация технологического проектирования процесса создания сборочных единиц на производстве, что влечет за собой уменьшение времени выполнения сборочных операций, снижение количества ошибок и улучшение эффективности производственного процесса.

Задачи исследования:

1. Разработать приложение для подготовки ИЭТР, описывающее действия, выполняемые на производстве.
2. Разработать приложение для отображения созданных ИЭТР на мобильном устройстве.
3. Оценить функциональность и возможность реализации данных приложений и их работу в комплексе в плане передачи данных и отображения визуальной информации.
4. Выбрать подходящие инструменты реализации и построить законченное решение, которое можно протестировать в реальной работе.
5. Проверить работоспособность комплекса приложений в реальных условиях эксплуатации.

III. Теория

3. Разработка ИЭТР с ДР-интеграцией

3.1 Принципы построения

В качестве описания общей конструкции системы подготовки инструкций для работы в ДР мы будем использовать принципы, приведенные в работе [5]:

- многоразовое использование. Приложение должно иметь возможность динамической загрузки для различных систем ДР, другими словами, оно не должно зависеть от какой-либо системы ДР и не должно быть жестко запрограммировано. Приложение должно динамически загружаться системой ДР как нейтральное. В данной работе целью являются мобильные устройства с установленной системой Android.

- двунаправленное взаимодействие в реальном времени. Возможности приложения должны быть пригодными для использования в любой реальной среде, поддерживать сбор данных о реальных операциях, а также информацию обратной связи через функцию отслеживания и наложение виртуальной информации с помощью контекстно-зависимых методов. В приведенной работе мы отображаем реальную среду на экране устройства с наложением дополнительной информации из созданных рабочих инструкций.

- использование не предполагает программирования. Содержимое ИЭТР для работы в ДР должно быть спроектировано с учетом обеспечения доступности для конечных пользователей без необходимости в программировании. Одновременно, структура и содержимое должны обладать адаптивными свойствами, позволяющими реагировать на частые изменения производственных сценариев. Для реализации этого принципа, приложение для создания и редактирования рабочих инструкций должно быть оснащено обширным набором вспомогательных инструментов, обеспечивающих гибкость и возможность настройки системы для удовлетворения потребностей различных пользователей.

3.2. Жизненный цикл приложения ИЭТР для работы в ДР

В данном разделе мы будем использовать проектирование рабочего процесса, предложенное в [5]. Оно включает следующие этапы разработки:

1. Планирование. В рамках этого этапа инженер-технолог определяет общую структуру работ, учитывая требования проектирования рабочего процесса. Он создает ИЭТР, которое содержит технические параметры, изображения и описание действий. Кроме того, он выбирает оптимальный ракурс сцены и необходимые модели деталей. В результате формируется контент, не зависящий от технологии ДР. На этом этапе приложение может быть рассмотрено как инициализированные рабочие инструкции, содержащие зарезервированные пустые блоки, предназначенные для наполнения при использовании в приложении ДР.

2. Обогащение. В рамках этого этапа частичное содержимое ИЭТР для работы в ДР загружается в приложение ДР, что позволяет оператору на месте провести пробную эксплуатацию и оценить эффективность подготовленного скелета ИЭТР. Он собирает дополнительную информацию о реальной физической среде в режиме реального времени, выявляя места, требующие коррекции. Полученная информация передается инженеру-технологу для выполнения необходимых изменений и дополнений к предустановленному контенту. После постобработки нового содержимого ИЭТР, созданная инструкция становится полноценным описанием рабочего процесса, интегрированным с технологией ДР и учитывающим реальные условия физической среды.

3. Развертывание. В рамках данного этапа происходит интеграция содержимого ИЭТР в систему ДР, которая затем переводится в режим реального времени. Оператор производства выполняет свои задачи под руководством системы, а информация о его деятельности может быть зарегистрирована в активной системе ИЭТР для работы в ДР для целей последующей персонализированной оценки и оптимизации производственного процесса.

4. Ревизия. Результаты рабочего процесса подвергаются тщательному анализу, и, если необходимо, инициируется новый цикл итерации, начиная с этапа планирования, для обеспечения оптимальной эффективности производственного процесса.

В рассматриваемом цикле задействованы два типа специалистов: инженер-технолог и оператор (высококвалифицированный и исполнитель), что согласуется с принципами адаптивности, изложенными в работе [5]. Согласно этим принципам, для создания контента ИЭТР для работы в ДР требуются только инженеры и операторы, работающие на месте, без необходимости привлечения дополнительных ресурсов, таких как разработчики.

3.3 Взаимодействие приложений разрабатываемого комплекса

Для создания и использования ИЭТР была разработана интеллектуальная пилотная система промышленной эксплуатации на основе ДР. Она состоит из двух пользовательских приложений, разделяющих одно ядро работы с данными. На вход приложению по созданию рабочих инструкций идут 3d-модели и изображения. Выходом является документ, содержащий созданные инструкции, который служит входными данными приложению, работающем на устройствах с ДР. Данная схема взаимодействия предоставляет возможность редактировать созданные инструкции и направлять откорректированный документ приложению ДР.

Архитектура предлагаемого комплекса приложений разделена на пять уровней как показано на рис. 1.

На приведенном рисунке внутренняя имплементация разрабатываемых приложений разделена на логические блоки. При этом такие блоки, как уровень данных, уровень визуализации и уровень логики являются общими для разрабатываемых приложений, что позволяет работу с одним документом и визуальными элементами. Функциональный уровень и уровень приложения разрабатываются независимо и различаются.

На уровне данных реализована модель данных, которая направлена на сохранение создаваемых инструкций и их элементов. Она способна связываться и сохранять в себе 3d-модель и ссылки на загружаемые изображения. Модель данных содержит информацию о геометрии, топологии и атрибутах трехмерных объектов, полученную из файлов формата STEP. Для сохранения и трансляции модели данных мы используем бинарный формат документа, что позволяет снизить размер передаваемых данных. Все необходимые файлы с данными сохраняются в файловой системе, поддерживающей Windows и Android.

Уровень визуализации предоставляет доступ к настройке положения камеры сцены, освещению. Он предоставляет набор презентаций и команды по управлению ими.

На логическом уровне представлены менеджеры для взаимодействия с моделью данных приложения и модулем визуализации.

На функциональном уровне представлены сущности и управляющие элементы, которые обеспечивают реализацию взаимодействия с пользователем приложения. Имплементация перечисленных элементов различается между разрабатываемыми приложениями. Это связано с различием целей разрабатываемых приложений, а также аппаратных систем.

На уровне приложения инженер либо оператор используют разработанный интерфейс приложений.

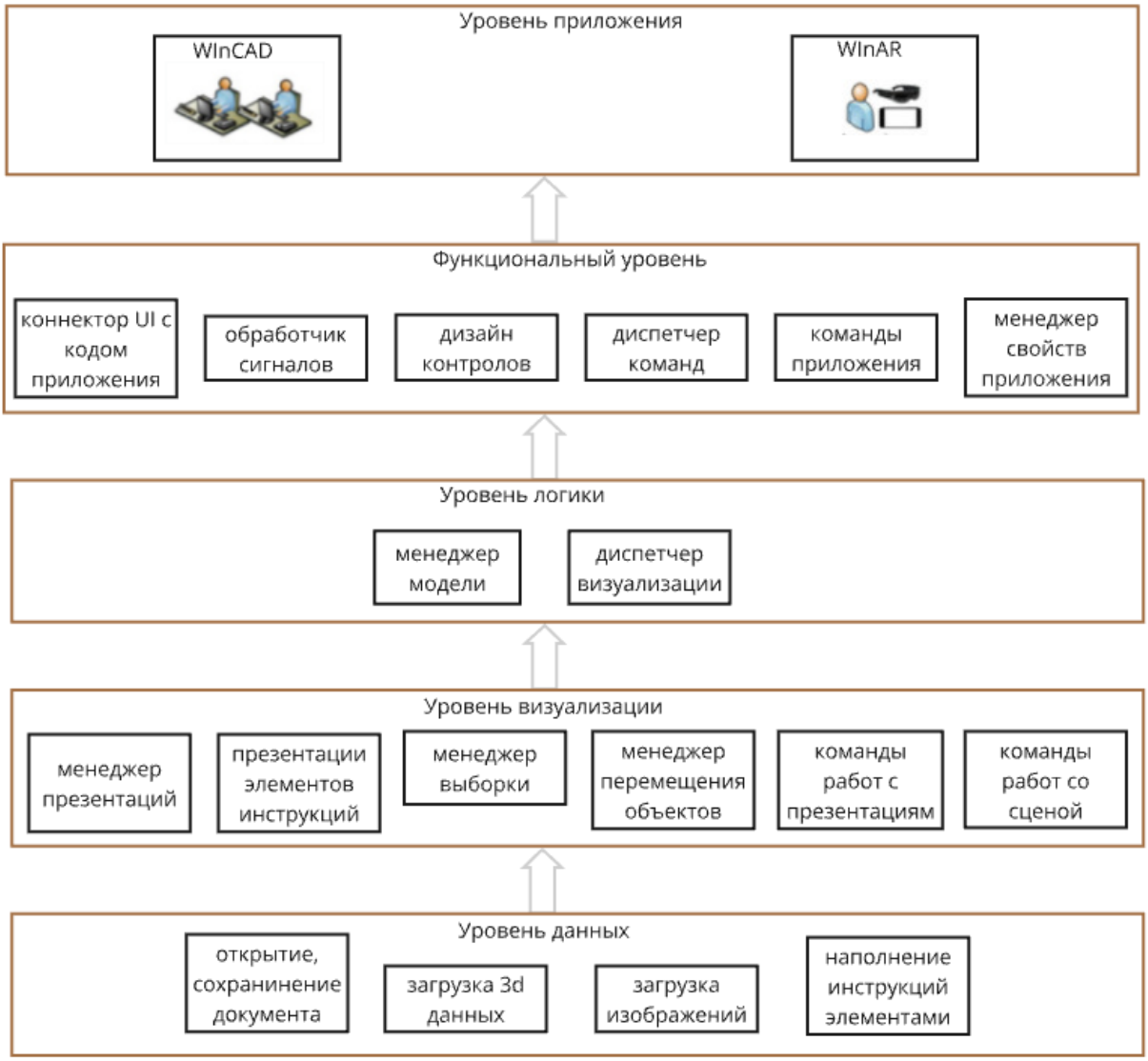


Рис. 1. Архитектура разрабатываемого комплекса приложений

3.4 Приложение для подготовки ИЭТР

Данное приложение разработано для использования на ПК и совместимо с операционными системами Windows и Linux. Инструмент предназначен для описания последовательности действий ИЭТР, которые в дальнейшем могут быть визуализированы с помощью ДР.

Разработка приложения была осуществлена с учетом требований национальных [10-13] и международных [14-15] стандартов.

Основные функциональные возможности приложения включают в себя:

- импорт трехмерных моделей с возможностью максимального извлечения данных из файлов описания моделей;
- интерактивное взаимодействие с трехмерной моделью сборки;
- навигацию по элементам модели, обеспечивающую выборку отдельных частей с возможностью фильтрации интересующих компонентов;
- создание и навигацию по рабочим картам;
- добавление текстовой информации о выбранном элементе, включая его тип, размеры и дополнительные комментарии;
- загрузку цифровых изображений, полученных в производственных условиях, для иллюстрации производственных процессов.

Инженер-технолог подготавливает ИЭТР, которое включает в себя набор рабочих карт. Каждая рабочая карта представляет собой окружающую среду и обеспечивает руководство к действиям. Данный элемент содержит

заголовок, набор объектов разметки, цифровых изображений и элементы 3d-модели. Он также определяет текущее положение камеры вида. Объект разметки, или маркер, представлен в виде аннотаций, стрелок, текста и изображений. Кроме того, рабочая карта может быть связана с визуальными объектами, такими как секущие плоскости и ассоциативные виды.

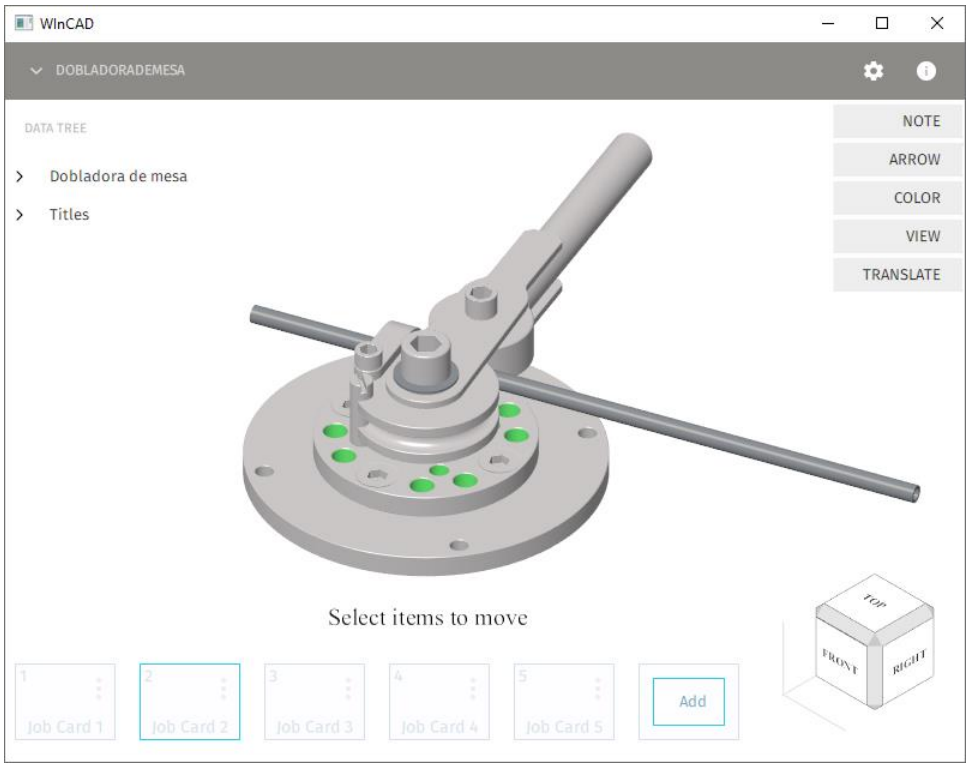


Рис. 2. Приложение по созданию ИЭТР

3.5 Приложение для использования ИЭТР в ДР

Это приложение обеспечивает функциональность на мобильных устройствах под управлением операционной системы Android.

Приложение позволяет загрузить ИЭТР, подготовленное инженером-технологом и передвигаться по его рабочим картам, отображая те же элементы визуализации, которые были подготовлены при создании карты. Эти элементы накладываются поверх изображения окружающего мира на экране.

Кроме того, при необходимости работы с камерой вида, предоставляется возможность ее позиционирования. На рис. 3 представлены ключевые особенности приложения ДР и то, как изображены различные ручные операции. Предоставление оператору такого уровня детализации снижает вероятность недопонимания (из-за отсутствия опыта или неправильного толкования) и уменьшает количество ошибок, допущенных на производстве.

3.6 Технологии, используемые для реализации комплекса приложений

В основе разработки ядра приложения лежит язык программирования C++ и библиотека 3D-моделирования OpenCascade [16], что обеспечивает совместимость с различными форматами данных САПР и позволяет организовывать модель данных, визуализировать их и выполнять над ними геометрические операции.

Пользовательский интерфейс приложения для рабочей станции разработан с использованием технологии QML библиотеки Qt, что обеспечивает гибкость и кроссплатформенность. Сборка приложения осуществляется средствами Visual Studio под Windows и системой Makefiles под Linux.

Для разработки пользовательского интерфейса приложения для ДР использован язык Java, а готовое приложение собирается средствами Android Studio, что позволяет подготовить его для различных конфигураций мобильных устройств.

В процессе разработки были рассмотрены различные инструменты и технологии, включая Unity и OpenXR, но в итоге было принято решение использовать вышеуказанные технологии ввиду их гибкости и кроссплатформенности.

Рис. 3. Приложение по использованию ИЭТР в ДР

IV. Результаты экспериментов

В качестве эксперимента мы выбрали подготовку ИЭТР по сборке ручного гибочного станка. Для данного примера будем использовать 3d-модель сборки, полученную в [17].

Сборка является обратным процессом к разборке. Поэтому, имея в доступе готовую модель сборки, мы подготавливаем инструкцию по разборке данной модели и будем использовать ее в дальнейшем в обратном порядке.

Разработанная инструкция для приложения на ПК представлена на рис. 4

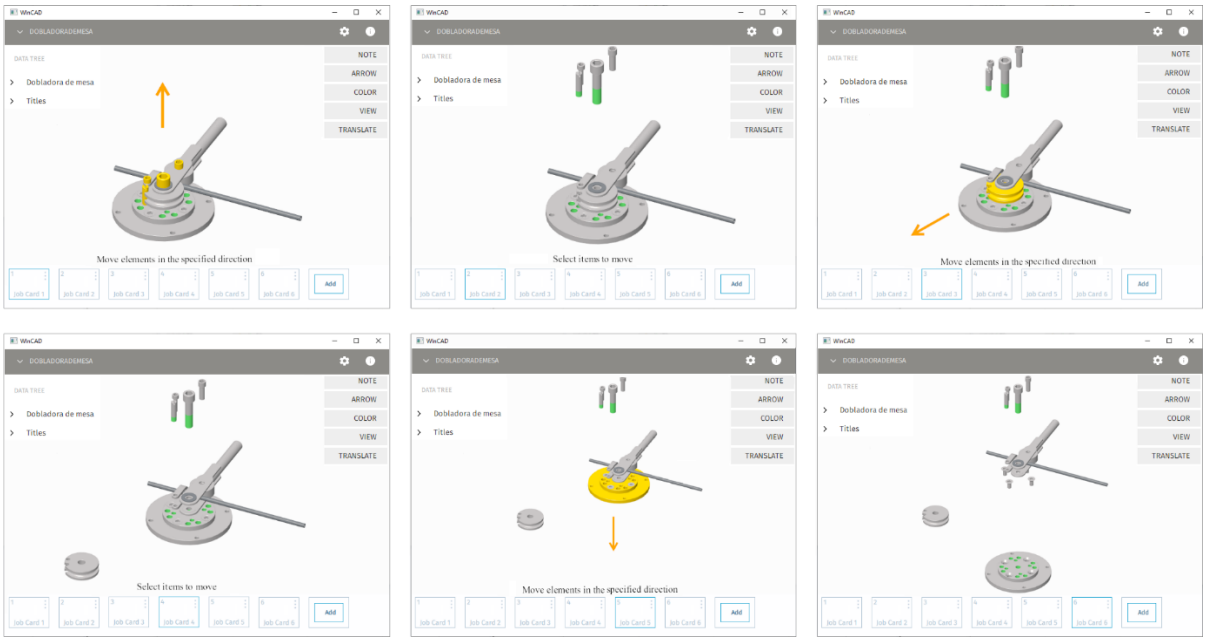


Рис. 4. Рабочие инструкции по сборке ручного гибочного станка

При создании инструкций происходит следующий процесс: инженер-технолог загружает 3d-модель и создает рабочие карты. Во время создания первой рабочей карты он выбирает элемент модели, который хочет переместить, и добавляет стрелку, указывающую направление движения. Затем он создает новую рабочую карту и сдвигает выбранный элемент в указанном направлении. Далее он создает следующую рабочую карту и так до тех пор, пока не разберет всю модель. При этом он может изменять положение камеры сцены, и новая созданная карта будет использовать предыдущее положение. Во время навигации по картам в сцене будет восстанавливаться положение камеры, использованное при наполнении данной рабочей карты. Каждая рабочая карта имеет заголовок, в котором указывается рекомендация оператору о том, что он должен сделать на текущем шаге исполнения.

Рис. 3 иллюстрирует пример представления инструкции на мобильном устройстве, где рабочая область отображает окружающую среду, на которую наложены элементы инструкции. Интерфейс также включает в себя функции навигации по рабочим картам инструкции и интерактивный элемент для изменения положения камеры.

В ходе эксперимента была реализована интеграция приложений для рабочей станции и мобильного устройства. В частности, было подготовлено ИЭТР на персональном компьютере, после чего оно было перенесено на мобильное устройство и отображено в производственном помещении.

V. Обсуждение результатов

Результат разработанных приложений может быть представлен в виде ИЭТР на экране или в виде приложения ДР, предназначенного для мобильных устройств.

Важной частью данной работы является анализ использования комплекса приложений для подготовки ИЭТР и отображения его в ДР для промышленности. Приложения предлагается писать на базовых инструментах программирования, что дает широкие возможности для расширения функционала, не ограниченные коммерческими лицензиями или узостью интерфейса.

VI. Выводы и заключение

В приведенной работе представлен комплекс приложений для автоматизации технологического проектирования процесса создания сборочных единиц на производстве с использованием средств ДР: описан жизненный цикл работ и особенности работы с данной системой. Подробно раскрыты архитектура и примененные технологии. Кроме того, описаны способы применения данного комплекса на реальном примере. Оговорены ограничения применения данных приложений в текущем состоянии.

Будущая работа над данной системой будет сосредоточена на расширении возможностей комплекса приложений с целью включения алгоритмов отслеживания элементов окружающей среды, например, основанное на методах распознавания объектов. Более того предполагается адаптировать приложение ДР на другие устройства ДР (умные очки и др.) с целью освободить руки оператора и повысить эффективность выполнения операций. В дальнейшем планируется реализовать автоматическое создание инструкций для случая сборки деталей основываясь на 3d-моделях сборок. Внедрение подобного алгоритма сократит время, необходимое для генерации инструкций. Наконец, будущая работа над сгенерированными инструкциями ДР может позволить адаптировать контент к опыту и уровню компетенции оператора.

Список литературы

1. Vasilios Zogopoulosa, Eva Geurtsb, Dorothy Gorsa, Steven Kauffmann Authoring Tool for Automatic Generation of Augmented Reality Instruction Sequence for Manual Operations, 2022/03 Procedia CIRP. 9th CIRP Conference on Assembly Technology and Systems.

2. G. Plakasa, S. T. Ponisa, K. Agalianosa, E. Aretoulakia, S. P. Gayialis Augmented Reality in Manufacturing and Logistics: Lessons Learnt from a Real-Life Industrial Application, 2020/04, FAIM2021.

3. Sabine Webel, Uli Bockholt, Timo Engelke Nirit Gavish, Manuel Olbrich, Carsten Preusche An augmented reality training platform for assembly and maintenance skills, 2013, Robotics and Autonomous Systems, Pages 398-403.

4. А. Мадалиев, В.М. Иванов Методика позиционирования хирургической системы смешанной реальности, 2023, GraphiCon.

5. Junhao Geng, Xinyu Song, Yuntao Pan, Jianjun Tang, Yu Liu, Dongping Zhao, Yongsheng Ma A systematic design method of adaptive augmented reality work instruction for complex industrial operations, 2020, Computers in Industry, Volume 119.

6. Leon Eversberg, Jens Lambrecht Evaluating digital work instructions with augmented reality versus paper-based documents for manual, object-specific repair tasks in a case study with experienced workers, 2023, The International Journal of Advanced Manufacturing Technology 127(3-4):1-13.

7. Leon Eversberg, Puya Ebrahimi, Martin Pape, Jens Lambrecht A cognitive assistance system with augmented reality for manual repair tasks with high variability based on the digital twin, 2022, Manufacturing Letters, Volume 34, Pages 49-52.

8. Konstantin Novikov, Petr Horejsi, Jan Kubr, Matej Dvorak, Miroslav Bednar , David Krakora, Matej Krnoul Michal Simon, Methodology for Rationalization of Pre-Production Processes, Using Virtual Reality Based Manufacturing Instructions, 2023, Machines.

9. Dorothy Gors, Jeroen Put, Bram Vanherle, Maarten Witters, Kris Luyten Semi-automatic extraction of digital work instructions from CAD models, 2021, Procedia CIRP, Volume 97, Pages 39-44.

10. ГОСТ Р 54088-2017. Интегрированная логистическая поддержка. Эксплуатационная и ремонтная документация в форме интерактивных электронных технических руководств. Основные положения и общие требования.

11. ГОСТ 2.051-2013. Единая система конструкторской документации (ЕСКД). Электронные документы. Общие положения.

12. ГОСТ Р 2.601-2019. Единая система конструкторской документации. Эксплуатационные документы.

13. ГОСТ Р 59278-2020. Информационная поддержка жизненного цикла изделий. Интерактивные электронные технические руководства с применением технологий искусственного интеллекта и дополненной реальности. Общие требования.

14. MIL-DTL-87269D. Detail specification database, revisable interactive electronic technical manuals, for the support of.

15. MIL-DTL-87268D. Interactive electronic technical manuals general content, style, format, and user-interaction requirements.

16. Open CASCADE Technology, https://dev.opencascade.org.

17. Juan Arroyave Workbench bender, https://grabcad.com/library/workbench-bender-1.

УДК: 004.942
DOI: 10.25206/978-5-8149-3873-2-2024-600-607

Сквозные 3D-технологии в составе комплекса программ в защищенном исполнении «Система полного жизненного цикла «Цифровое предприятие» (СПЖЦ САРУС)
End-to-end 3D technologies as part of a secure software package "Digital Enterprise Full Life Cycle System" (SARUS)

М.А. Глазова, А.И. Жукаев, И.В. Калинников, Д.А. Кирпичев, С.А. Зорина, С.Н. Ганчук, А.В. Трищенков
ФГУП «РФЯЦ-ВНИИЭФ», г. Саров, Нижегородской обл., Россия

M.A. Glazova, A.I. Zhukaev, I.V. Kalinnikov, D.A. Kirpichev, S.A. Zorina, S.N. Ganchuk, A.V. Trishchenkov
FSUE "RFNC-VNIIEF", Sarov, Nizhny Novgorod region, Russia

Аннотация. Основная задача современной промышленности – создание глобально конкурентоспособной и востребованной продукции нового поколения в кратчайшие сроки. Для успешного решения этой задачи необходимы постоянная генерация, применение, накопление и трансфер новых знаний, создание и развитие наукоемких технологий с последующим их объединением в технологические цепочки нового поколения, разработка наукоемких инноваций и создание современных цифровых производств.

В 2016-2023 гг. Институтом Цифровых технологий ФГУП «РФЯЦ-ВНИИЭФ» разработан уникальный отечественный импортонезависимый комплекс программ в защищенном исполнении СПЖЦ САРУС, ориентированный как на предприятия оборонно-промышленного комплекса (ОПК), так и гражданских отраслей промышленности.

В СПЖЦ САРУС реализован принцип сквозной 3D-технологии, работа которой построена на взаимодействии независимых проектов PDM, CAD, CAE, CAM в рамках программного модуля «Технологическая платформа», представляющей собой набор программных компонентов, предназначенных для построения, функционирования и интеграции на основе единой объектно-ориентированной структуры данных и типовых протоколов взаимодействия.

Ключевые слова: система полного жизненного цикла, цифровое предприятие, сквозные 3D-технологии.

Annotation. The main task of modern industry is to create globally competitive and in-demand products of a new generation in the shortest possible time. To successfully solve this problem, it is necessary to constantly generate, apply, accumulate and transfer new knowledge, create and develop high-tech technologies with their subsequent integration into new generation technological chains, develop high-tech innovations and create modern digital industries.

In 2016-2023 The Institute of Digital Technologies of the Federal State Unitary Enterprise "RFNC-VNIIEF" has developed a unique domestic import-independent set of programs in a secure version - SRLTS SARUS intended for both enterprises of the military-industrial complex (DIC) and civilian industries.

The SRLTS SARUS implements the principle of the end-to-end 3D technology, which is based on the interaction of independent PDM, CAD, CAE projects and CAM within the framework of the “Technological Platform” software module, which is a set of software components designed for construction, operation and integration based on an integrated object-oriented data structure and standard interaction protocols.

Keywords: full life cycle system, digital enterprise, end-to-end 3D technologies.

I. Введение

Ранее в России отсутствовала комплексная импортонезависимая технологическая платформа, объединяющая операционные системы, системы управления базами данных, системы управления предприятием и сквозные технологии управления полным жизненным циклом изделия. Использование существующих импортных платформ в сфере ОПК ограничено ввиду несоответствия требованиям по информационной безопасности, эмбарго, санкций и связанными с этим рисками.

В соответствии с Указом Президента Российской Федерации от 30 марта 2022 «С 1 января 2025 г. органам государственной власти, заказчикам было запрещено использовать иностранное программное обеспечение на принадлежащих им значимых объектах критической информационной инфраструктуры».

На данный момент реализацию данного Указа Президента Российской Федерации обеспечивает разработанный в ФГУП «РФЯЦ-ВНИИЭФ» импортонезависимый, защищенный отечественный комплекс СПЖЦ САРУС.

II. Постановка задачи

Основной задачей разработки комплекса стала необходимость создание и внедрение сквозной 3D-технологии в отечественный комплекс программ в защищенном исполнении СПЖЦ САРУС в соответствии с государственными программами.

Команде разработчиков смежных проектов предстояло решить задачи по:

— реализации импортонезависимости, защищенности модулей и обеспечению их сквозного взаимодействия между собой;

— завершению проекта по подготовке к внедрению в следующих организациях: АО «ОДК», АО «ЮМАТЕКС», АО «ИВК», АО «РЭП-Холдинг»;

— проведению комплексного функционального тестирования СПЖЦ САРУС на предприятиях – технических заказчиков: ПАО «ОАК», АО «ОСК», АО «Вертолеты России», АО «ХК «Уралвагонзавод», ФГУП «КГНЦ»;

— апробации отраслевых решений СПЖЦ САРУС и его внедрение в организации ОПК.

III. Принцип работы

Внедрения сквозных 3D-технологий в СПЖЦ САРУС обусловлено необходимостью взаимодействия модулей внутри комплекса для выполнения задач конструкторского проектирования, включая предварительные инженерные расчеты, подготовку производства, учет и хранение и сопровождение данных о жизненном цикле изделия. Принцип работы сквозной технологии построен на взаимодействии независимых проектов PDM, CAD, CAE, CAM в рамках программного модуля «Технологическая платформа», которая представляет собой набор программных компонентов, предназначенных для построения, функционирования и интеграции на основе единой объектно-ориентированной структуры данных и типовых протоколов взаимодействия. Технологическая платформа обеспечивает коллективную работу, управление объектной моделью прикладных инженерных систем, функционирующих в едином информационном пространстве с общими базами данных и единым файловым хранилищем. Общая архитектура комплекса СПЖЦ САРУС наглядно отображена на рис. 1.

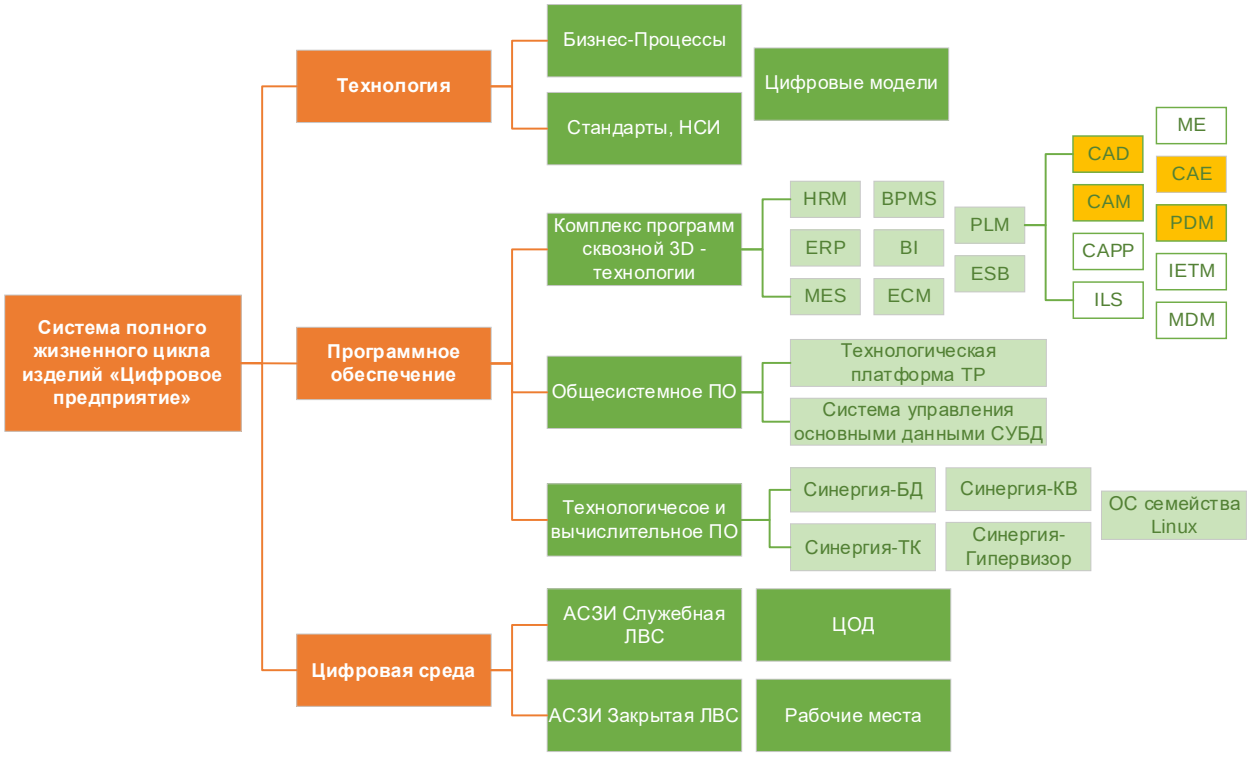


Рис. 1. Архитектура комплекса СПЖЦ САРУС

С целью разработки архитектуры независимых программных модулей, обеспечивающих систему сквозного автоматизированного проектирования, была создана схема информационных потоков (рис. 2), позволяющая выявить общие точки взаимодействия данных модулей между собой.

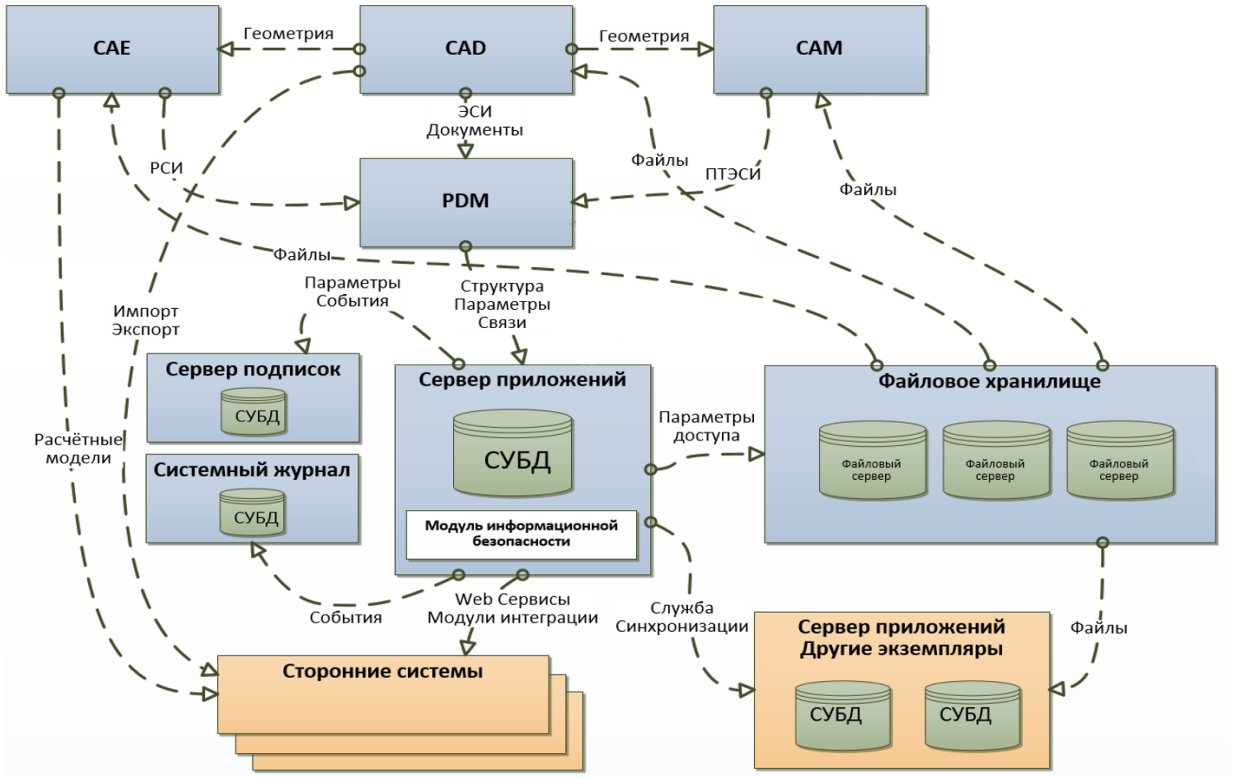


Рис. 2. Схема информационных потоков

В общем случае для обеспечения процессов проектирования, функциональные возможности сквозной 3D-технологии начинаются с программного проекта PDM, который предназначен для организации коллективной работы в единой цифровой среде PLM-системы (рис. 3) и выполняет:

обмен данных между смежными модулями;

формирование графа требований к изделию;

разработку бизнес-процессов и присвоение ролей каждому участнику проекта;

создание и редактирование ЭСИ изделия;

управление архивом технической документации предприятия;

планирование и мониторинг конструкторского и технологического проектирования;

согласование ЭСИ и документации на него;

управление изменениями ЭСИ и документации на нее;

использование простейшей электронной подписи для утверждения ЭКД.

Сформированные в PDM файлы записываются в файловое хранилище и являются основанием для начала работы остальных модулей. Взятие объектов из общей базы данных, на любом этапе жизненного цикла изделия, не приводит к конфликтам коллективной работы в едином информационном пространстве.

После формирования плана разработки КД, подготовки и выдачи заданий начинается этап непосредственного конструкторского проектирования.

Рис. 3. Функциональные возможности программного модуля PDM

В программном проекте CAD осуществляется разработка новых и редактирование ранее созданных геометрических моделей объектов проектирования и производства, включая 3D и 2D-представления, оформление конструкторской документации, визуализация объектов проектирования, геометрический анализ цифровых моделей (рис. 4). Так же в CAD осуществляется создание ассоциативных аннотированных чертежей, создание сопроводительной документации к разрабатываемым изделиям –спецификации на основе сборочных чертежей выпускаемого изделия, извещений об изменениях, и различных текстовых документов, при условии соблюдения норм ЕСКД по оформлению данного вида документации.

С помощью инструментов модуля пользователь создает аннотированную твердотельную модель будущего изделия в соответствии с ЭСИ, содержащую в себе всю необходимую атрибутивной информацией, массо-центровочные характеристики и данные необходимые для изготовления и контроля.

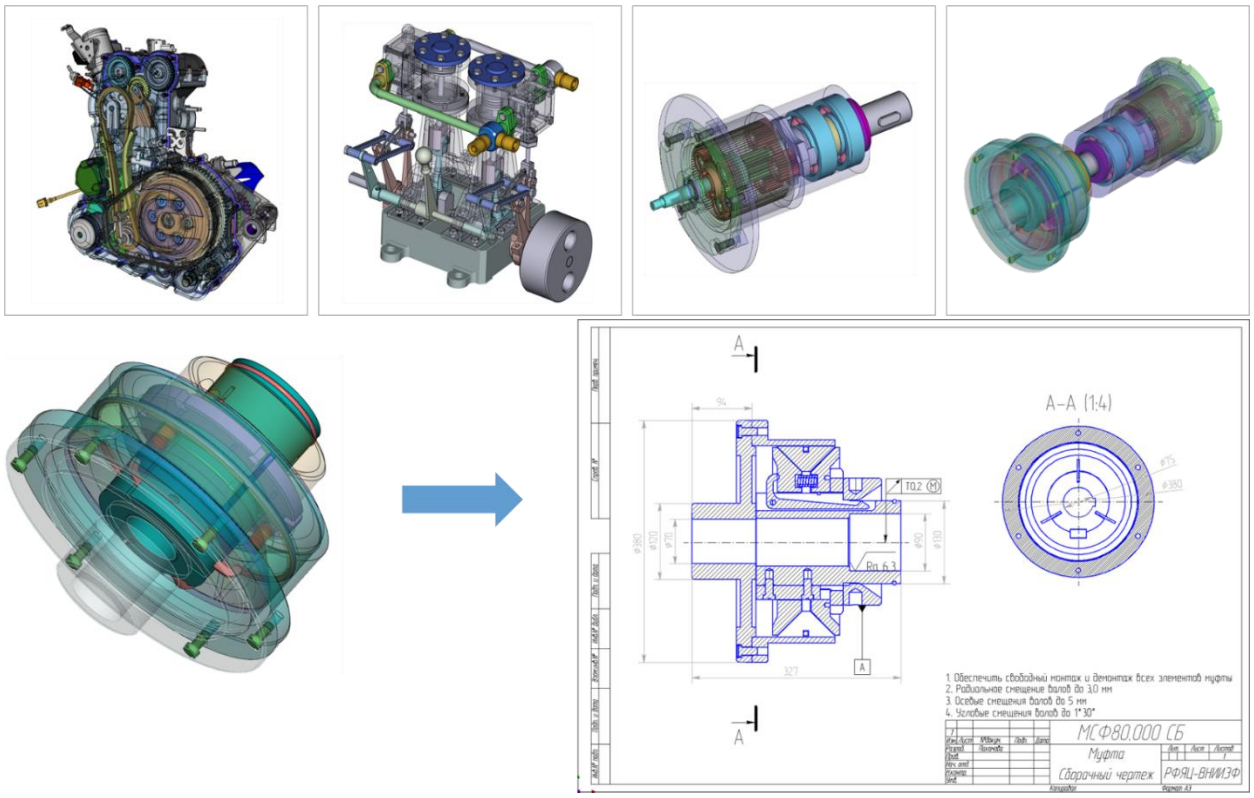


Рис. 4. Функциональные возможности программного модуля CAD

Спроектированная модель или сборка изделия сохраняется в ЭСИ проекта PDM, в котором так же может быть сформированная спецификация, на основании атрибутов загруженных файлов (рис. 5).

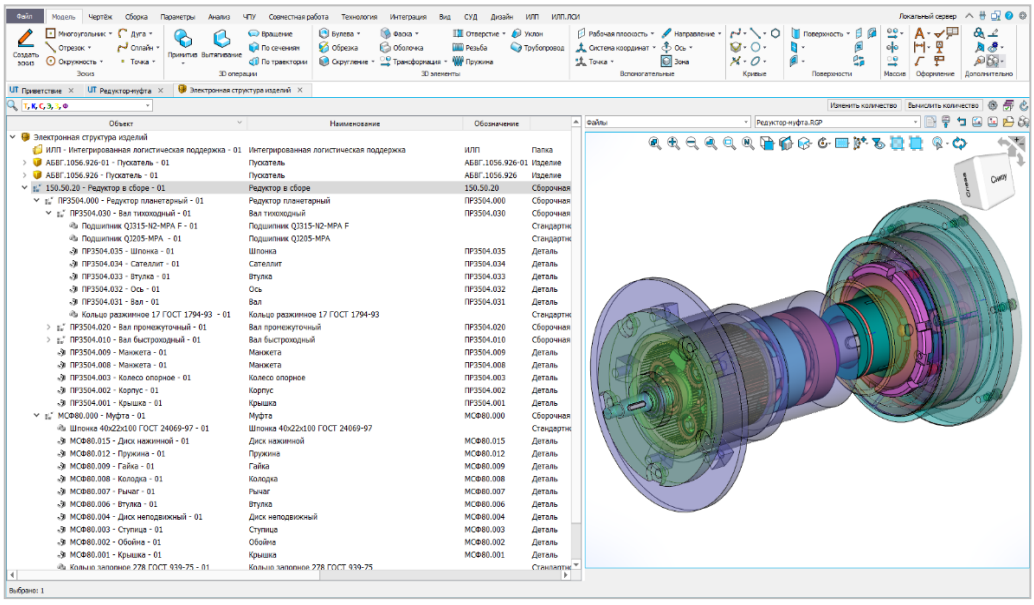


Рис. 5. Загрузка модели в ЭСИ

Электронная модель является основой для дальнейшей работы таких проектов как CAE и CAM.

В процессе конструкторского проектирования в проекте CAE выполняются инженерные расчеты стандартных элементов и типовых соединений, применяемых в конструкции изделия, например: пружины, контактные пары, тонкостенные оболочки, подшипники, зубчатые передачи, сварные, клеевые, шпоночные, шлицевые и резьбовые соединения.

Также выполняются прочностные и тепловые экспресс-расчеты составных частей и деталей конструкции изделия с применением внешних подключаемых решателей «ЛОГОС Прочность» и «ЛОГОС Тепло»: статический анализ, анализ устойчивости, динамический анализ, модальный анализ, гармонический анализ, анализ на воздействие широкополосной случайной вибрации, тепловой анализ (рис. 6 а, б).

Вся работа модуля может быть разделена на 3 этапа:

– препроцессинг – создание расчетной задачи для вычислений МКЭ;

– решение задачи – решение поставленной задачи с применением внешних подключаемых решателей «ЛОГОС»;

– постпроцессинг – анализ результатов расчета.

Из CAE в PDM посредством проекта CAD передается информация о расчетной модели и результатах расчета. Документы и, соответственно, расчетные модели проекта CAE хранятся в файлах. Соответственно, вся функциональность проекта PDM обеспечивает коллективную работу с документами, включающими, в том числе, данные КЭ-анализа моделей и расчетов стандартных элементов и типовых соединений.

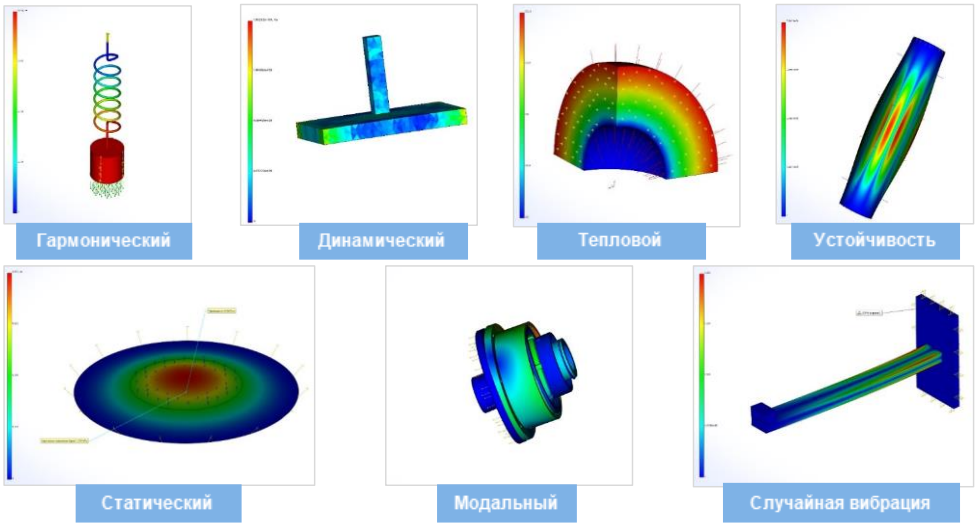


Рис. 6 а. Экспресс анализ в модуле CAE

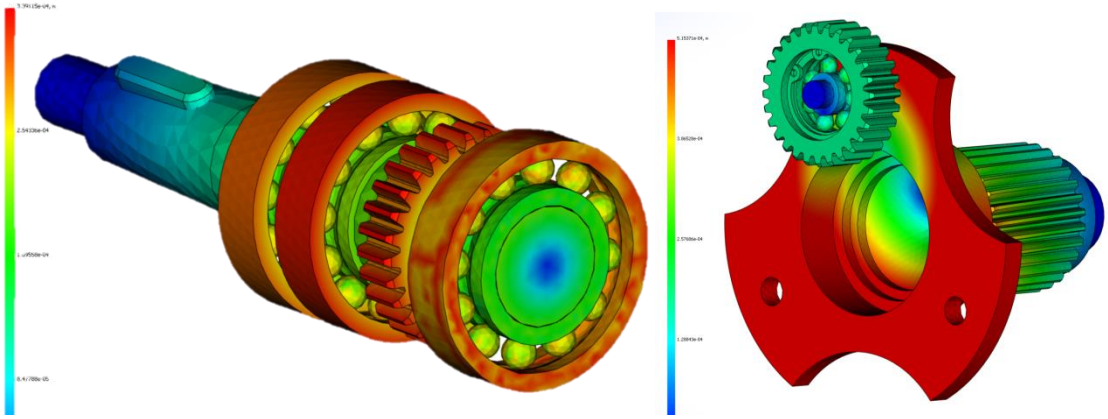


Рис.6 б. Примеры применения программного модуля CAE

После формирования работ проектов КПП, наполнения ЭСИ комплектом КД (техническими требованиями, моделями, расчетами, пояснительными записками и т.п.) наступает этап согласования и утверждения ЭКД. Согласование ЭКД осуществляется в соответствии с бизнес-процессом и ролями, разработанными на начальном этапе проектирования в PDM.

По окончанию этапа согласования выдается задание работнику архива на регистрацию и учет комплекта ЭКД.

По сформированному комплекту ЭКД начинается следующий этап подготовки производства к изготовлению. По готовности конструкции изделия к изготовлению, производится ее подготовка к ЧПУ изготовлению: подготовка оснасток, выбор режущих инструментов, написание управляющих программ и др. в программном проекте CAM. (рис. 7).

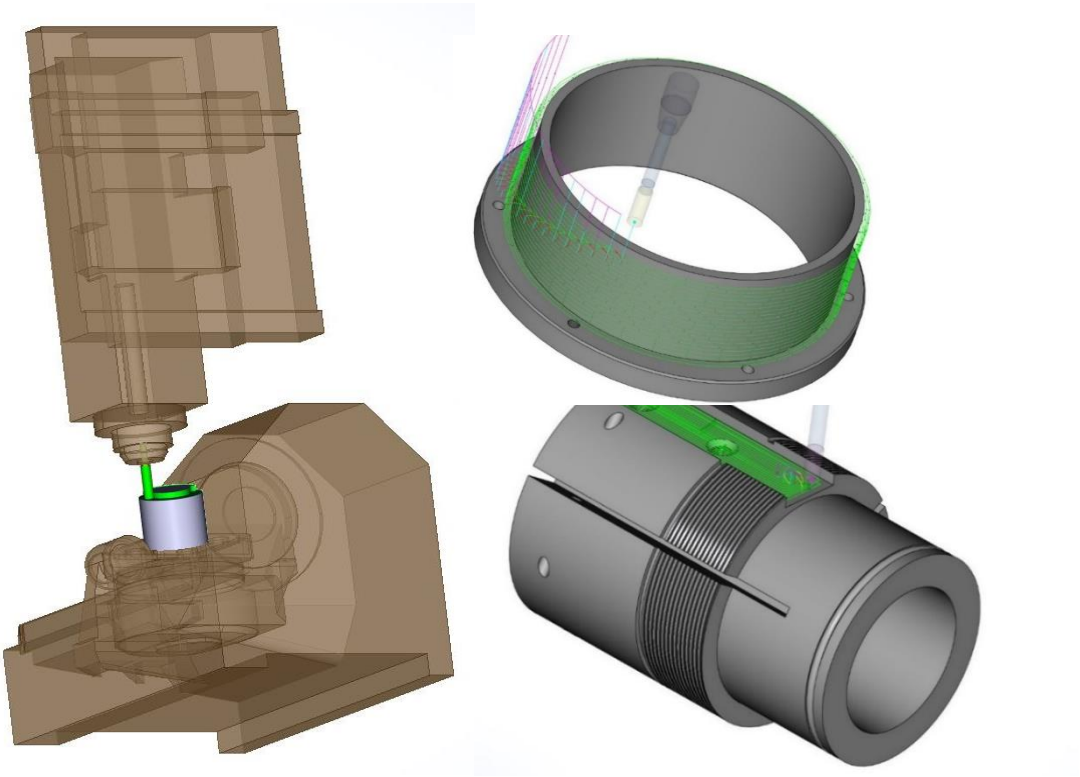


Рис.7. Пример применения программного модуля CAM

Проект CAM взаимодействует в рамках единой системы автоматизации производства с проектом PDM. PDM контролирует выполнение задач и отслеживает изменения в составе и структуре изделий. CAM предназначен для подготовки управляющих программ для станков с ЧПУ, которые интегрированы в проект PDM. За обеспечение процесса моделирования в CAM отвечает CAD-система, где обмен между модулями осуществляется через программный модуль ТП.

Проект CAM разделен на следующие функциональные блоки:

– по функциональному назначению: среда разработки, верификатор, высокоскоростная обработка.
– по виду обработки: токарная обработка, фрезерная обработка, обработка отверстий, резка, электроэрозионная обработка, аддитивное производство.

На основе различных исходных данных и событий, формируется план работ КПП, формируются работы проектов КПП. Информация рассылается исполнителям. Исполнители обеспечивают накопление «базы знаний» о структуре изделия. Инженерами обеспечивается наполнение ЭСИ комплектом КД (техническими требованиями, моделями, расчетами, пояснительными записками и т.п.) с его последующим согласованием (рис. 8).

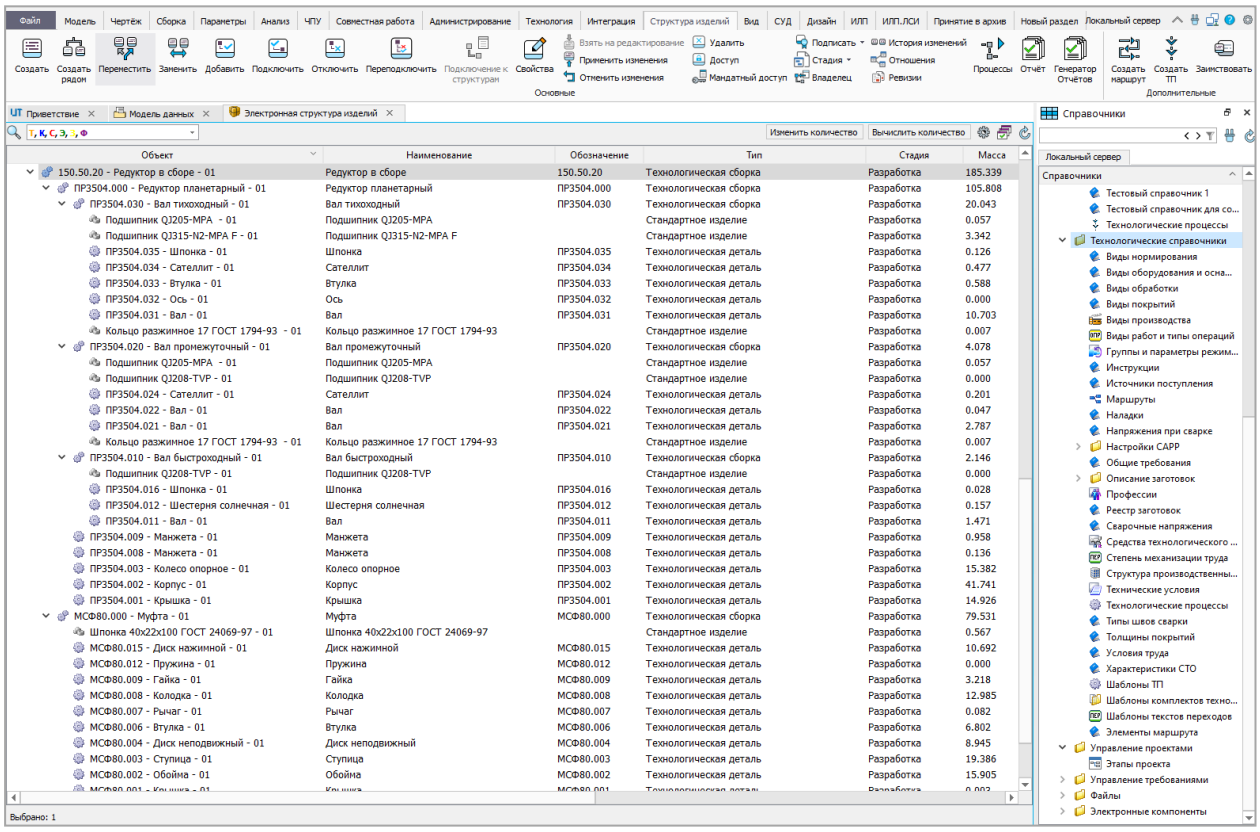


Рис.8. Сформированная производственно-технологическая ЭСИ

Интеграция между проектами, работающими в единой базе данных, происходит путем использования единых справочных данных без дополнительных преобразований. Обмен с помощью файлов обменных форматов предоставляет возможность формирования на стороне системы-отправителя структурированного файла, содержащего данные об объектах системы, а на стороне системы-получателя возможность чтения такого файла и внесения изменений, описание которых содержится в обменном файле, в объекты системы-получателя.

VI. Выводы и заключение

В период 2019-2023 гг. в соответствии с требованиями ТЗ разработано программное импортонезависимое отечественное решение по конструкторскому проектированию в составе СПЖЦ САРУС с внедренной сквозной 3D-технологией, которое:

- обеспечивает решение широкого спектра конструкторских и производственных задач специалистами конструкторских бюро и технологами производства, включая управление и контроль над изделиями на всем жизненном цикле изделия;

- прошло уникальное для Российской Федерации комплексное функциональное тестирование в организациях – технических заказчиках: ПАО «ОАК», АО «ОСК», АО «Вертолеты России», АО «ХК «Уралвагонзавод», ФГУП «КГНЦ», в процессе которого создано более 150 АРМ, обучено более 200 специалистов, разработаны изделия основного производства;

- имеет свидетельства о государственной регистрации программы для ЭВМ, включено в реестр российского программного обеспечения;

- успешно сдано межведомственной комиссии и проектно-архитектурному комитету.

В 2023 г. разработанное программное решение:

- введено в опытную эксплуатацию во ФГУП «РФЯЦ-ВНИИЭФ» по результатам работ, в рамках которых было создано 1366 рабочих мест, обучено 450 специалистов, разрабатывалось 11 изделий по тематическому плану;

- подготовлено к пилотному внедрению в организациях: АО «ОДК», АО «ЮМАТЕКС», АО «ИВК», АО «РЭП-Холдинг», в которых создано более 50 АРМ и обучено более 50 специалистов.

Список литературы

1. Божко А.Н., Волосатова Т.М., Грошев С.В. и др. Основы автоматизированного проектирования, 2021. 329 с.

2. Гражданская продукция организаций оборонно-промышленного комплекса: каталог / РОСАТОМ. 281 с. URL: https://rosatom.ru/upload/iblock/a86/a869c529d553ce1bf7e665c26ab3885b.pdf (дата обращения: 09.09.2024).

3. Инфраструктура и ресурсное обеспечение цифровой экономики / Д. С. Бурцев, Е. С. Гаврилюк, А. Г. Изотова [и др.]. СПб: Университет ИТМО, 2021. 190 с.

4. Компьютерный инжиниринг: учеб. пособие / А. И. Боровков [и др.]. СПб.: Изд-во Политехн. ун-та,2012. 93 с.

5. Наумов М. А., Губин А. Л., Пучкова С. А. Комплекс программ в защищенном исполнении «Система полного жизненного цикла изделий» «Цифровое предприятие». Система конструкторского проектирования // Высокие технологии атомной отрасли. Молодежь в инновационном процессе: сб. материалов XV научно-технической конференции молодых специалистов Росатома (Нижний Новгород, 15–16 сент. 2021 г.) . Саров: Изд-во Всерос. науч.-исслед. ин-та экспериментальной физики, 2021. С. 207–211.

6. Суртаева О. С. Драйверы цифрового развития промышленного производства в России: Монография, 2021. 126 с.

Секция 5

ОБРАБОТКА И АНАЛИЗ БИОМЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Biomedical Image Processing and Analysis

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-609-617

Прогноз и визуализация трендовых тем в обработке биомедицинских изображений на основе машинного обучения
Prediction and visualization of trending topics in biomedical image processing based on machine learning

М. М. Шарнин1, Н. В. Сомин1
Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия

M. Charnine1, N. V. Somin1
[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

Аннотация. В работе представлен библиометрический и визуальный анализ трендовых тем в коллекции PubMed в области обработки изображений на основе машинного обучения. Анализ выполнен с использованием коллекции научных статей PubMed, из которой были отобраны 9875 статей, имеющие в заголовках слова “learning” (обучение) и “image” (изображение). Выявлен значительный рост (в 6.29 раза за 5 лет) ежегодно публикуемых подобных статей. Рассчитан и представлен рейтинг характерных/релевантных ключевых слов в отобранных статьях. Среди релевантных ключевых слов выявлены трендовые ключевые слова со значительными частотами и растущими трендами в отобранных статьях, а также с прогнозируемым долгосрочным ростом трендов в коллекции PubMed. Долгосрочный прогноз трендов рассчитан с помощью пакета машинного обучения CatBoost на основе данных об употреблении ключевых слов в статьях PubMed с учетом ссылок цитирования. Представлена семантическая карта трендовых ключевых слов, содержащая информацию о новизне и долгосрочности трендов. Семантическая карта рассчитана с помощью нейронных сетей Word2Vec и алгоритма t-SNE. Группы из близких по семантике трендовых ключевых слов образуют трендовые темы. Близкие трендовые темы входят/объединяются в кластеры на семантической карте. В результате визуального анализа семантической карты выявлены одиннадцать трендовых тем: (1) виды глубокого обучения, (2) самоконтролируемое обучение и обучение с небольшим количеством примеров (self-supervised and few-shot learning), (3) мета-обучение и доменная адаптация (meta-learning and domain adaptation), (4) автокодировщик и объединение признаков (autoencoder and feature fusion), (5) обнаружение объектов и аномалий (object and anomaly detection), (6) семантическая сегментация (semantic segmentation), (7) механизм внимания (attention mechanism), (8) трендовые болезни, (9) методы получения изображений, (10) персонализация и оптимизация, (11) сравнительный анализ и обзоры. Дан обзор литературы по наиболее перспективным трендовым темам с указанием дополнительных ключевых слов и ссылок на публикации.

***Ключевые слова*:** библиометрический анализ, визуальная аналитика, трендовые темы, обработка изображений на основе машинного обучения, трендовые ключевые слова, долгосрочные прогнозы трендов, коллекция PUBMED

Аннотация. The paper presents a bibliometric and visual analysis of trending topics in the PubMed collection in the field of image processing based on machine learning. The analysis was performed using the PubMed collection of scientific articles, from which 9875 articles were selected with the words “learning” and “image” in the titles. A significant increase (6.29 times over 5 years) in annually published similar articles was revealed. The rating of relevant keywords in the selected articles was calculated and presented. Among the relevant keywords, trending keywords were identified with significant frequencies and increasing trends in the selected articles, as well as with a predicted long-term increase in trends in the PubMed collection. Long-term trend forecast was calculated using the CatBoost machine learning package based on keyword usage in PubMed articles, taking into account citation links. A semantic map of trending keywords is presented, containing information about the novelty and long-term nature of trends. The semantic map was calculated using Word2Vec neural networks and the t-SNE algorithm. Groups of trending keywords that are similar in semantics form trending topics. Similar trending topics are included into clusters on a semantic map. As a result of a visual analysis of the semantic map, eleven trending topics were identified: (1) types of deep learning, (2) self-supervised and few-shot learning, (3) meta-learning and domain adaptation, (4) autoencoder and feature fusion, (5) object and anomaly detection, (6) semantic segmentation, (7) attention mechanism, (8) trending diseases, (9) imaging techniques, (10) personalization and optimization, (11) comparative analysis and reviews. A review of the literature on the most promising trending topics is provided, indicating additional keywords and links to publications.

***Ключевые слова*:** bibliometric analysis, visual analytics, trending topics, image processing based on machine learning, trending keywords, long-term trend forecasts, PUBMED collection.

I. Введение

Область обработки изображений на основе машинного обучения стремительно развивается. За последние годы количество статей по обработке биомедицинских изображений на основе машинного обучения выросло в разы и превысило две тысячи работ ежегодно. Опубликованные обзоры научной литературы не дают полную картину тенденций в этой области. Существующие обзоры затрагивают только отдельные узкие темы в этой области и не используют возможностей долгосрочного прогноза. Например, в обзоре [1] рассматривается только мультимодальное глубокое обучение для анализа биомедицинских изображений. В обзоре [2] обсуждается использование глубокого обучения в медицинской визуализации для улучшения ухода за позвоночником. В обзоре [3] анализируется машинное обучение в визуализации сердечно-сосудистой системы.

Это приводит к следующим трем основным целям исследования: (1) из имеющихся в литературе работ по обработке изображений на основе машинного обучения выявить трендовые ключевые слова, имеющие наиболее долговременные растущие тренды по количеству статей и цитирований; (2) выявить и визуализировать трендовые темы из близких по семантике трендовых ключевых слов; (3) дать краткий обзор научной литературы в области обработки изображений на основе машинного обучения с точки зрения наиболее перспективных/трендовых тем.

Статья имеет следующую структуру. Сначала описаны обрабатываемые библиографические данные и применяемый метод библиометрического анализа и визуальной аналитики. Далее приводятся результаты применения метода и обзор наиболее перспективных трендовых тем в области обработки изображений на основе машинного обучения, включая краткий обзор наиболее интересных статей по этой тематике. В конце дается заключение.

II. Постановка задачи

В качестве предметной области в работе была выбрана медицина. Область исследования данной статьи - это обработка биомедицинских изображений на основе машинного обучения. Статья посвящена прогнозу и визуализации трендовых тем в исследуемой области.

Более точно ставится следующая задача: дать прогноз с известной точностью на 3 и более лет вперед для трендовых тем в области исследования на базе коллекции PubMed, а также разработать средства визуализации трендовых тем.

Исходный текстовый материал

Для решения поставленной задачи необходимо полностью проанализировать текстовую коллекцию PubMed, которая по состоянию на начало 2024 года содержала более 36 миллионов статей (36 417 711) по медицине, биологии и связанным наукам. Для поиска в PubMed работ, относящихся к исследуемой области (т.е. к обработке изображений на основе машинного обучения), использовался поисковый запрос learn* AND imag*, где звездочка (*) означает любую последовательность букв. При этом учитывалось, что чаще всего в статьях PubMed анализируются биомедицинские изображения. С помощью этого запроса из PubMed было выделено 9875 статей, содержащие в заголовках слова “learning” (обучение) и “image” (изображение) в различных формах. Данные 9875 статей мы называем в дальнейшем Выборочной коллекцией.

Актуальность

Поставленная задача является актуальной, т.к. количество статей в исследуемой области растет.

По данным Выборочной коллекции были рассчитаны следующие графики роста количества статей по обработке изображений на основе машинного обучения за последние годы (см. рис. 1).

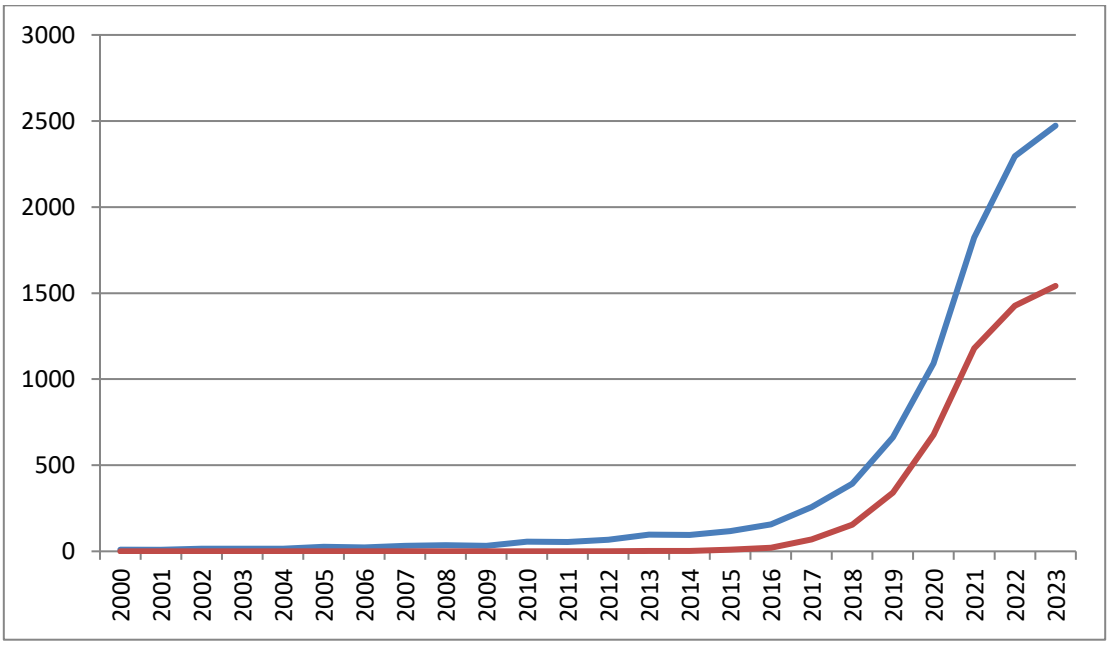


Рис. 1. Верхний график - количество научных статей в Выборочной коллекции в различные годы. Нижний график – количество статей из Выборочной коллекции со словом “deep” (глубокое).

Из рисунка 1 видно, что наблюдается значительный рост количества ежегодно публикуемых научных статей по обработке изображений на основе машинного обучения, индексируемых в PubMed (верхний график). Причем эти статьи все чаще касаются глубокого обучения и имеют слово “deep” в заголовке (нижний график). Так за 5 лет с 2018 по 2023 количество ежегодно публикуемых статей в исследуемой области выросло в 2473/393=6.29 раза. За тот же период количество статей, касающихся глубокого обучения (со словом “deep”), выросло в 1542/155=9.95 раза. После 2000 года общая доля статей со словом “deep” составляет 5438/9875=55%.

III. Теория

Этапы анализа

Для выявления и визуализации трендовых тем используется библиометрический анализ, включающий в себя применение математических и статистических методов к изучению библиографических коллекций книг, периодических изданий и прочих публикаций. Библиометрический анализ исследуемой области содержит следующие этапы:

1) определение поисковых запросов, с помощью которых можно найти статьи из PubMed, относящиеся к исследуемой области,

2) построение Выборочной коллекции из заголовков статей в PubMed, удовлетворяющих поисковым запросам, и анализ динамики роста количества статей в Выборочной коллекции за последние годы,

3) выявление релевантных, а затем и трендовых ключевых слов в Выборочной коллекции,

4) построение семантической карты и выявление трендовых тем с помощью визуальной аналитики,

5) обзор наиболее перспективных трендовых тем.

Основные методы исследования и применяемые программные средства

Трендовые ключевые слова определяются с учетом долгосрочных прогнозов их трендов в PubMed, рассчитанным с помощью пакета машинного обучения CatBoost [4]. CatBoost хорошо подходит для задач прогноза, работает быстрее чем аналогичные нейросетевые алгоритмы при незначительным снижением качества прогноза. Прогнозы трендов для характерных ключевых слов рассчитываются, используя различные показатели статей, содержащих эти ключевые слова. Наиболее важным показателем тренда ключевого слова является количество цитирований ключевого слова (КЦКС). Для расчета этого показателя мы сначала находим все статьи с этим ключевым словом в определенном году (год1), а затем подсчитываем все цитирующие ссылки на эти статьи (за период от год2 до год3). Значение КЦКС определяется по формуле 1:

$$\text{КЦКС}(\text{слово}, \text{год1}, \text{год2}, \text{год3}) = \sum_i \text{Ссылок}(\text{Статья}_i(\text{слово}, \text{год1}), \text{год2}, \text{год3}) \quad (1)$$

Долгосрочность тренда равна количеству лет непрерывного роста усредненного показателя КЦКС. Такая долгосрочность тренда является целью для алгоритма машинного обучения CatBoost. Регрессионная модель CatBoost была обучена для 20 параметров тренда ключевого слова в текущем и предыдущих годах, включая: КЦКС, количество статей с ключевым словом, количество лет роста предыдущего тренда, общий рост цитируемости тренда, количество лет от начала тренда. Более подробно данная методика описана в работах [5,6]. Точность долговременного прогноза на 3 года вперед составляет более 60%. При этом 59% трендовых слов, выявленных в 2020 году, остались трендовыми в 2023 году.

Результаты прогноза трендов визуализируются с помощью алгоритмов Word2Vec и t-SNE на семантической карте. Word2Vec – это разновидность нейронных сетей, отличительной особенностью которых является высокая скорость обработки больших массивов текстовых данных с целью выявления семантических связей между ключевыми словами. Алгоритм t-SNE хорошо подходит для визуализации данных высокой размерности в двухмерном пространстве, так что похожие объекты (ключевые слова) моделируются близко расположенными точками, а непохожие объекты моделируются с большой вероятностью точками, далеко друг от друга отстоящими.

С помощью визуальной аналитики выявляются трендовые темы, входящие в кластеры трендовых слов на семантической карте. Трендовая тема состоит из одного или нескольких близких по смыслу трендовых ключевых слов. Наиболее перспективные трендовые темы рассмотрены в приведенном далее обзоре научной литературы.

Таким образом материалом для исследования является вся коллекция PubMed, в которой особое внимание уделяется ее части под названием Выборочная коллекция. Параметры ключевых слов Выборочной коллекции исследуются как на локальном уровне в Выборочной коллекции, так и на глобальном уровне во всей коллекции PubMed. На семантической карте отображаются как локальные параметры ключевых слов (новизна трендов), так и глобальные параметры – семантическая близость и долгосрочность прогноза трендов. Визуальный анализ семантической карты помогает определить трендовые темы, оценить их динамику и увидеть картину в целом, включая перспективные направления из наиболее перспективных тем.

IV. Результаты экспериментов

Рейтинг наиболее характерных/релевантных ключевых слов

Характерными/релевантными в исследуемой области являются термины, которых относительно много в Выборочной коллекции по сравнению с коллекцией PubMed.

Список/рейтинг наиболее характерных ключевых слов в Выборочной коллекции: learning (обучение), deep learning (глубокое обучение), learning image (обучающее изображение), image classification (классификация изображений), image (изображение), images (изображения), deep (глубокий), dictionary learning (обучение по словарю), image retrieval (поиск изображения), hyperspectral image (гиперспектральное изображение), transfer learning (трансферное обучение), dictionary (словарь), learning-based (на основе обучения), deep learning-based (на основе глубокого обучения), image super (изображение супер), image segmentation (сегментация изображения), images deep (глубокие изображения), machine learning (машина обучение), machine (машина), segmentation (Сегментация), learning hyperspectral (Гиперспектральное обучение), feature learning (Функциональное обучение), hyperspectral (Гиперспектральный), supervised learning (контролируемое обучение), supervised (Контролируемый), image super-resolution (Сверхразрешение изображения), image annotation (Аннотация к изображению), metric learning (Метрическое обучение), medical image (Медицинский образ), learning approach (Подход к обучению), single image (Одно изображение), imagery (Образность), instance learning (Экземплярное обучение), classification (Классификация), learning-based image (Образ на основе обучения), image recognition (распознавание изображений), active learning (активное обучение), sensing images (распознавание изображений), learning framework (структура обучения), semi-supervised (полуконтролируемое обучение), imaging machine (машина формирования изображения), x-ray images (рентгеновские изображения), instance (экземпляр), semi (полу), sensing image (распознаваемое изображение), learning medical (медицинское обучение), extreme learning (экстремальное обучение), semi-supervised learning (полуконтролируемое обучение), ct images (изображения компьютерной томографии), manifold learning (многообразное обучение) и т.д.

Рейтинг трендовых ключевых слов

Трендовые ключевые слова были отобраны среди релевантных/характерных ключевых слов с учетом их локальных параметров/показателей в Выборочной коллекции и глобальных параметров в коллекции PubMed.

С помощью алгоритма машинного обучения CatBoost по методике, описанной в предыдущем разделе, были рассчитаны прогнозы долгосрочности роста трендов для характерных слов и выявлены трендовые ключевые слова. В результате в Выборочной коллекции было выявлено 3788 характерных терминов, у которых прогноз роста был более чем на один год, а количество упоминаний более двух. Из этих терминов было отобрано 716 трендовых терминов, у которых количество упоминаний более 8, а прогноз роста более 4 лет (т.е. после 2027 года).

Отобранные ключевые слова были упорядочены по совокупности показателей и в результате получился рейтинг трендовых ключевых слов. Рейтинг составлялся с учетом следующих показателей: положение в рейтинге характерных слов, данные прогноза, частота в коллекции PubMed и в Выборочной коллекции, а также по тенденциям роста в Выборочной коллекции.

Верхние позиции в рейтинге трендовых ключевых слов занимают следующие термины: self-supervised (самоконтролируемое), meta-learning (метаобучение), cross-domain (междоменный), leveraging (использование), hybrid deep (гибридный глубокий), autoencoder (автоэнкодер), domain adaptation (адаптация предметной области), deep-learning model (модель глубокого обучения), multimodal deep (мультимодальный глубокий), personalized (персонализированный), object detection (обнаружение объектов), attention mechanism (механизм внимания), deep reinforcement (глубокое подкрепление), combined machine (комбинированная машина), performance evaluation (оценка производительности), classification model (модель классификации), contrastive (контрастный), prototype (прототип), anomaly detection (обнаружение аномалий), chest computed (компьютерный анализ грудной клетки), automated machine (автомат), network image (сетевое изображение), ischemic stroke (ишемический инсульт), autoencoders (автоэнкодеры), disentangled (распутывание), osteoarthritis (остеоартрит), optimizing (оптимизация), quality improvement (улучшение качества), knee osteoarthritis (остеоартрит коленного сустава), deep ensemble (глубокий ансамбль), review machine (обзорная машина), benchmarking (бенчмаркинг), review deep (глубокий анализ), classification framework (структура классификации), intraoperative (интраоперационно), rectal cancer (ректальный рак), can improve (можно улучшить), segmentation network (сеть сегментации), scoping review (обзорный обзор), feature fusion (объединение признаков), cervical cancer (рак шейки матки), colon cancer (рак толстой кишки), robust deep (робастная глубокая), semantic segmentation (семантическая сегментация), skin cancer (рак кожи), phantom study (фантомное исследование), cone-beam computed (конусно-лучевые вычисления), data deep (глубокие данные), dual-energy ct (двухэнергетическая компьютерная томография) и т.д.

Построение семантической карты и визуальный анализ

Вообще нужно отметить, что построение семантической карты и ее визуализация – важный и даже необходимый этап всего алгоритма. Визуализация очень полезна, когда нужно срочно понять общую ситуацию, выявить главное и принять, причем коллективно, обоснованное решение.

Для целей визуального анализа с помощью нейросети Word2Vec была рассчитана мера семантического подобия (semantic similarity) трендовых ключевых слов в коллекции PubMed. По этим данным построена семантическая карта с помощью алгоритма t-SNE (см. рисунок 2). Чем выше мера подобия – тем меньше расстояние между ключевыми словами на семантической карте.

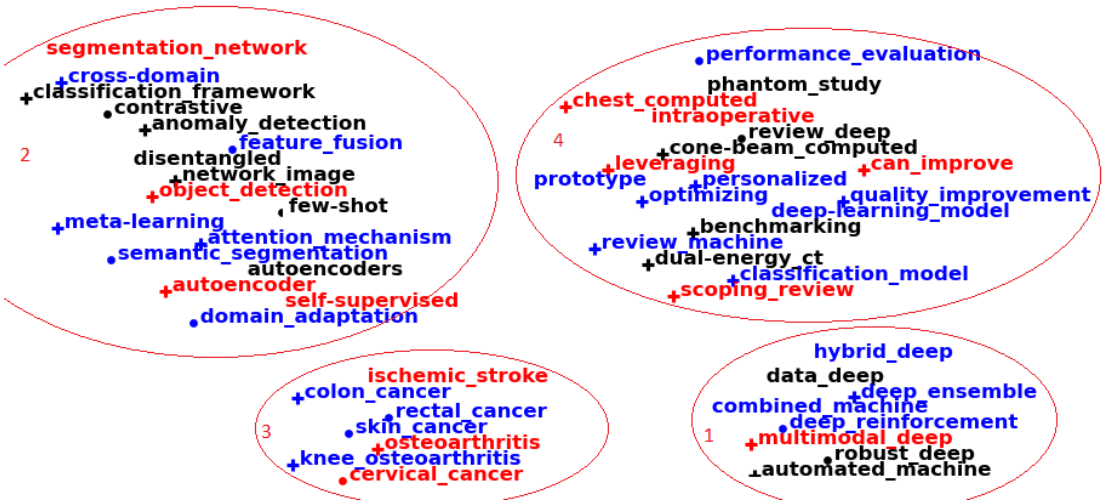


Рис. 2. Семантическая карта трендовых ключевых слов из предметной области обработки изображений на основе машинного обучения. Чем выше семантическое сходство ключевых слов, тем ближе они расположены

на рисунке. Красным цветом выделены наиболее перспективные/трендовые термины, имеющие долгосрочные тренды в коллекции PubMed, синим – средние тренды, черным – минимальные тренды. Плюсом (+) выделены ключевые слова, имеющие самые новые/свежие тренды, а точкой – средние по новизне тренды.

На рисунке 2 красным и синим цветами представлены трендовые термины PubMed. Красным цветом выделены наиболее перспективные термины, имеющие самые долгосрочные тренды в коллекции PubMed, синим – средние тренды, черным – минимальные тренды. Плюсом (+) выделены ключевые слова, имеющие самые новые/свежие тренды, а точкой – средние по новизне тренды.

На рисунке 2 также видны четыре кластера, которые содержат близкие по семантике трендовые термины, которые можно разбить на темы.

На основе визуального анализа выявлены 11 трендовых тем, распределенных между четырьмя кластерами: (1) виды глубокого обучения (types of deep learning), (2) самоконтролируемое обучение и обучение с небольшим количеством примеров (self-supervised and few-shot learning), (3) мета-обучение и доменная адаптация (meta-learning and domain adaptation), (4) автокодировщик и объединение признаков (autoencoder and feature fusion), (5) обнаружение объектов и аномалий (object and anomaly detection), (6) семантическая сегментация (semantic segmentation), (7) механизм внимания (attention mechanism), (8) трендовые болезни, (9) методы получения изображений, (10) персонализация и оптимизация, (11) сравнительный анализ и обзоры.

Тема 1 входит в первый одноименный кластер “виды глубокого обучения”, к которому относятся: гибридное глубокое обучение, мультимодальное глубокое обучение, глубокое обучение с подкреплением, глубокоге ансамблевое обучение.

Темы 2,3,4,5,6,7 входят во второй кластер, который можно назвать “алгоритмы обработки изображений”.

Тема 8 входит в третий одноименный кластер “трендовые болезни”, который включает в себя различные виды рака (cancer) и остеоартрита (osteoarthritis), а также ишемический инсульт (ischemic stroke).

Темы 9,10,11 входят в четвертый кластер, который можно назвать “другие связанные задачи”, включая (9) методы получения изображений (panoramic, chest computed, cone-beam computed, intraoperative, ultrasonic), (10) персонализация и оптимизация, (11) сравнительный анализ и обзоры.

V. Обсуждение результатов

Визуальная аналитика помогает разделить тредовые ключевые слова на кластеры, причем некоторые из этих кластеров представляют одну тему, как например, кластер 3 “трендовые болезни”. Другие кластеры могут содержать несколько различных тем. В этом случае близость ключевых слов внутри кластера помогает определить содержание тем, причем близкие позиции по горизонтали имеют большое значение. Например, темы 2 (self-supervised, few-shot) и 5 (object detection, anomaly detection) имеют близкие позиции по горизонтали у первых букв их ключевых слов. Ключевые слова темы 4 (autoencoder, autoencoders, feature fusion) имеют близкие позиции по вертикали для слов “autoencoder” и “autoencoders”, а также близкие позиции по горизонтали для слов “autoencoder” и “feature fusion”.

Визуальный анализ семантической карты помогает определить трендовые темы, оценить их динамику и увидеть картину в целом, включая перспективные направления из наиболее перспективных тем. Например, близость на семантической карте красных и синих слов (2) self-supervised, (3) “domain adaptation” и (4) “autoencoder” показывает, что соответствующие темы 2,3,4 из кластера “алгоритмы обработки изображений” являются перспективным направлением среди алгоритмов обработки изображений.

Особенности данной визуализации в том, что она состоит только из трендовых ключевых слов, в ней цветом отражены прогнозы долгосрочности трендов, а также новизна этих трендов с помощью символов точка (.) и плюс (+) в начале слов. Такая визуализация помогает быстро понять ситуацию и выявить главные и зарождающиеся тренды/тенденции.

Обзор трендовых тем

Рассмотрим наиболее перспективные трендовые темы в области обработки изображений на основе машинного обучения, выявленные в процессе библиометрического анализа.

Первая трендовая тема “типы глубокого обучения” (types of deep learning) содержит статьи, использующие различные виды глубокого обучения, включая: мультимодальное глубокое обучение, гибридное глубокое обучение, глубокое обучение с подкреплением, глубокое ансамблевое обучение. Эта тема имеет следующие

ключевые слова: multimodal deep (мультимодальное глубокое), hybrid deep (гибридное глубокое), deep reinforcement (глубокое обучение с подкреплением), deep ensemble (глубокое ансамблевое).

Мультимодальное глубокое обучение (Multimodal deep learning) объединяет информацию из нескольких модальностей, таких как текст, изображение, аудио и видео, для получения более точных и полных прогнозов.

Подход гибридного глубокого обучения сначала обучает глубокую сеть на обучающих данных, извлекает функции из глубокой сети, а затем использует эти функции для повторного выражения данных для ввода в метод неглубокого обучения, который обучен выполнять окончательную классификацию.

Глубокое обучение с подкреплением (Deep reinforcement learning) сочетает в себе обучение с подкреплением и глубокое обучение. Модели глубокого ансамблевого обучения (Deep ensemble learning models) сочетают в себе преимущества как моделей глубокого обучения, так и ансамблевого обучения, так что конечная модель имеет лучшую производительность.

Рассмотрим ряд статей из этой темы. В работе [1] дается обзор мультимодального глубокого обучения в биомедицинских изображениях и текстах, с помощью которого решаются такие задачи как создание отчетов, ответы на визуальные вопросы, кросс-модальный поиск, компьютерная диагностика и семантическая сегментация. Статья [7] описывает удаление шума на ультразвуковых изображениях молочной железы на основе гибридного метода глубокого обучения. Работа [8] обсуждает модель подписи к изображению, основанную на механизме внимания и глубоком обучении с подкреплением. Статья [9] посвящена анализу неаннотированных гистопатологических изображений почек с использованием глубокого ансамблевого обучения.

Ко второй теме “самоконтролируемое обучение и обучение с небольшим количеством примеров” (self-supervised and few-shot learning) относятся статьи, в которых обсуждаются вопросы, связанные с обучением алгоритмических моделей на необработанных, неразмеченных данных, а также обучение, в котором модель надо преднастроить на тренировочных данных так, чтобы она хорошо обучалась на ограниченном количестве новых размеченных примеров, что помогает преодолевать недостаток размеченных данных. Эта тема имеет следующие ключевые слова: self-supervised, few-shot, contrastive. Контрастное обучение (contrastive learning) использует не только принцип близости, но и принцип различия, т.е. фокусируется на извлечении значимых представлений путем сопоставления положительных и отрицательных пар примеров. Рассмотрим некоторые статьи из этой темы. В статье [10] обсуждается самоконтролируемое контрастивное обучение для классификации изображений по нескольким примерам. В работе [11] предлагается новый взгляд на сегментацию медицинских изображений по нескольким примерам. Статья [12] посвящена декодированию изображений моторики мозга из нескольких ЭЭГ с использованием извлечения связанных признаков и обучения по нескольким примерам.

Третья трендовая тема “мета-обучение и доменная адаптация” (meta-learning and domain adaptation” содержит статьи, в которых изучаются методы определения наиболее подходящего алгоритма вместе с параметрами к нему для быстрой адаптации к новым задачам, а также методы обучения модели на данных из домена-источника (source domain) так, чтобы она показывала сравнимое качество на целевом домене (target domain). Эта тема имеет следующие ключевые слова: meta-learning, domain adaptation, cross-domain. Приведем примеры статей. В работе [13] рассматривается глубокое метаобучение для классификации медицинских изображений по нескольким примерам. Статья [14] описывает неконтролируемую адаптацию предметной области для классификации изображений и обнаружения объектов с использованием подхода управляемого переноса обучения и дивергенции JS. Работа [15] посвящена минимизации стоимости аннотаций для сегментации ультразвуковых изображений с использованием междоменного переноса обучения.

Четвертая трендовая тема “автокодировщик и объединение признаков” (autoencoder and feature fusion) содержит статьи, в которых изучаются методы поиска и объединения значимых признаков. Автокодировщик – это специальная архитектура искусственных нейронных сетей, позволяющая применять обучение без учителя, при котором на выходном слое нейросети получается отклик, наиболее близкий ко входному. Автокодировщики выделяют из входных данных общие признаки, которые кодируются в значениях весов искусственной нейронной сети. Объединения различных признаков, извлеченных из разных баз данных способствует улучшению обнаружения объектов. Из-за разнообразия внешнего вида объектов объединение их разнородных визуальных особенностей/признаков необходимо для высокоэффективного обнаружения объектов. Эта тема имеет следующие ключевые слова: autoencoder, autoencoders, feature fusion. Приведем примеры статей. Работа [16] посвящена неконтролируемому обучению признакам и кластеризации частиц, изображенных в необработанных голограммах, с использованием автоэнкодера. В статье [17] описывается масштабное объединение признаков на основе обучения для анализа выживаемости по целым изображениям слайдов.

VI. Выводы и заключение

В последние десятилетия наблюдается значительный рост количества англоязычных работ по обработке изображений на основе машинного обучения. Стремительно развиваются такие методы и темы/направления обработки изображений, как использование глубокого обучения и технологии нейросетей.

В результате прогнозного библиометрического анализа на основе коллекции PubMed выявлены и рассмотрены следующие одиннадцать трендовых тем в англоязычной научной литературе по обработке изображений на основе машинного обучения: (1) виды глубокого обучения, (2) самоконтролируемое обучение и обучение с небольшим количеством примеров (self-supervised and few-shot learning), (3) мета-обучение и доменная адаптация (meta-learning and domain adaptation), (4) автокодировщик и объединение признаков (autoencoder and feature fusion), (5) обнаружение объектов и аномалий (object and anomaly detection), (6) семантическая сегментация (semantic segmentation), (7) механизм внимания (attention mechanism), (8) трендовые болезни, (9) методы получения изображений, (10) персонализация и оптимизация, (11) сравнительный анализ и обзоры.

Однако полученные результаты позволяют сделать вывод, что предлагаемая методика библиометрического анализа может быть применена не только к медицине или биологии, но и к другим предметным областям, в которых число публикаций быстро растет, а выявлять тенденции развития и оценивать наиболее перспективные тренды нужно по возможности качественно и в тоже время быстро. Поэтому данную статью можно рассматривать как отработку на основе большой и сложной предметной области общей методики библиометрического анализа, состоящей из следующих этапов: (1) получение Выборочной коллекции, (2) вычисление ключевых слов, (3) выявление трендов, (4) получение и визуализация семантической карты (5) более пристальный анализ основных трендов.

Список литературы

1. Sun, Z., Lin, M., Zhu, Q., Xie, Q., Wang, F., Lu, Z., & Peng, Y. (2023). A scoping review on multimodal deep learning in biomedical images and texts. Journal of Biomedical Informatics, 104482.

2. Constant, C., Aubin, C. E., Kremers, H. M., Garcia, D. V. V., Wyles, C. C., Rouzrokh, P., & Larson, A. N. (2023). The use of deep learning in medical imaging to improve spine care: A scoping review of current literature and clinical applications. North American Spine Society Journal (NASSJ), 15, 100236.

3. [Rouzrokh, P., Khosravi, B., Vahdati, S., Moassefi, M., Faghani, S., Mahmoudi, E., ... & Erickson, B. J. (2023). Machine learning in cardiovascular imaging: a scoping review of published literature. Current Radiology Reports, 11(2), 34-45.

4. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in neural information processing systems. 2018. V. 31.

5. Charnine M., Klokov A., Kochiev L., Tishchenko A. Research trending topic prediction as cognitive enhancement. 2021 international conference on cyberworlds (CW). IEEE. 2021. P. 217–220. doi: 10.1109/CW52790.2021.00044.

6. Charnine M., Tishchenko A., Kochiev L. Visualization of Research Trending Topic Prediction: Intelligent Method for Data Analysis. Proceedings of the 31th International Conference on Computer Graphics and Vision. 2021. V. 2. P. 1028–1037.

7. Vimala, B. B., Srinivasan, S., Mathivanan, S. K., Muthukumaran, V., Babu, J. C., Herencsar, N., & Vilcekova, L. (2023). Image noise removal in ultrasound breast images based on hybrid deep learning technique. Sensors, 23(3), 1167.

8. Bai, T., Zhou, S., Pang, Y., Luo, J., Wang, H., & Du, Y. (2023). An image caption model based on attention mechanism and deep reinforcement learning. Frontiers in Neuroscience, 17, 1270850.

9. Koo, J. C., Ke, Q., Hum, Y. C., Goh, C. H., Lai, K. W., Yap, W. S., & Tee, Y. K. (2023). Non-annotated renal histopathological image analysis with deep ensemble learning. Quantitative Imaging in Medicine and Surgery, 13(9), 5902.

10. Lim, J. Y., Lim, K. M., Lee, C. P., & Tan, Y. X. (2023). SCL: Self-supervised contrastive learning for few-shot image classification. Neural Networks, 165, 19-30.

11. Feng, Y., Wang, Y., Li, H., Qu, M., & Yang, J. (2023). Learning what and where to segment: A new perspective on medical image few-shot segmentation. Medical Image Analysis, 87, 102834.

12. Zhu, L., Liu, Y., Liu, R., Peng, Y., Cao, J., Li, J., & Kong, W. (2023). Decoding Multi-Brain Motor Imagery From EEG Using Coupling Feature Extraction and Few-Shot Learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 4683-4692.

13. Jiang, H., Gao, M., Li, H., Jin, R., Miao, H., & Liu, J. (2022). Multi-learner based deep meta-learning for few-shot medical image classification. IEEE Journal of Biomedical and Health Informatics, 27(1), 17-28.

14. [Goel, P., & Ganatra, A. (2023). Unsupervised domain adaptation for image classification and object detection using guided transfer learning approach and JS divergence. Sensors, 23(9), 4436.

15. Monkam, P., Jin, S., & Lu, W. (2023). Annotation cost minimization for ultrasound image segmentation using cross-domain transfer learning. IEEE Journal of Biomedical and Health Informatics, 27(4), 2015-2025.

16. Liu, Z., Thevar, T., Takahashi, T., Burns, N., Yamada, T., Sangekar, M., ... & Thornton, B. (2021). Unsupervised feature learning and clustering of particles imaged in raw holograms using an autoencoder. JOSA A, 38(10), 1570-1580.

17. Li, L., Liang, Y., Shao, M., Lu, S., & Ouyang, D. (2023). Self-supervised learning-based Multi-Scale feature Fusion Network for survival analysis from whole slide images. Computers in Biology and Medicine, 153, 106482.

УДК: 519.6+004.891.3+616-018
DOI: 10.25206/978-5-8149-3873-2-2024-618-628

Нейросетевой метод поиска по содержимому для гистологических изображений
Neural network method for content-based histological image retrieval

А. С. Вешкин[1], А. В. Хвостиков[1]
[1] *Факультет вычислительной математики и кибернетики, Московский государственный университет имени М. В. Ломоносова, Москва, Россия*

Аннотация. В данной работе исследуется влияние различных подходов извлечения признаков из гистологических изображений на эффективность методов поиска по содержимому. Особое внимание уделяется сравнительному анализу трёх подходов к извлечению признаков: базирующемся на статистических призна- ках изображения, использовании автокодировщика и применении гиперболического экстрактора. В работе рассматриваются теоретические аспекты каждого подхода, проводится их адаптация для задач анализа гистологических изображений и экспериментальная оценка их производительности на наборах гистологических изображений PATH-DT-MSU WSS1, WSS2. Для оценки качества поиска используются метрики Accuracy@10 и MAP@10. Наилучшие результаты показал метод, использующий гиперболический экстрактор признаков: 73.45% по Accuracy@10. Описанный в работе подход является универсальным и может быть применен на дру- гих наборах данных для задачи поиска изображений по содержимому.

Ключевые слова: глубокое обучение, компьютерное зрение, гистология, полнослайдовые гистологические изображения, поиск по содержимому.

Аннотация. In this paper, we study the impact of different approaches to feature extraction from histological images on the performance of content-based retrieval methods. Particular attention is paid to the comparative analysis of three approaches to feature extraction: based on statistical image features, using an autoencoder, and using a hyperbolic extractor. The paper considers theoretical aspects of each approach, adapts them to the tasks of histological image analysis, and experimentally evaluates their performance on the PATH-DT-MSU WSS1, WSS2 histological datasets. The Accuracy@10 and MAP@10 metrics are used to evaluate the search quality. The best results were shown by the method using the hyperbolic feature extractor: 73.45% by Accuracy@10. The approach described in the paper is universal and can be applied to other datasets for the task of content-based retrieval of images.

Ключевые слова: deep learning, computer vision, histology, whole slide histological images, content based image retrival.

I. Введение

Гистологические изображения являются важным инструментом для диагностики различных заболеваний и исследований в области медицины. Они содержат ценную информацию о структуре тканей и клеток, кото- рая может быть использована для определения патологических изменений и оценки эффективности лечения. Однако, интерпретация гистологических изображений требует большого объема времени и экспертизы со сто- роны специалистов. Кроме того, существует проблема, связанная с субъективностью восприятия и различными степенями опыта разных специалистов.

Современные гистологические сканеры позволяют получать полнослайдовые изображения (англ. Whole Slide Image, WSI), типичное разрешение которых крайне велико — порядка 100 тыс. 100 тыс. пикселей, но может достигать и больших значений. Это одновременно и хорошо, и привносит трудности, так как врачи получают много полезной информации, но ручной анализ такого большого объема данных крайне затруднителен.

В свете этих проблем возникает необходимость разработки метода поиска по содержимому для гистологических изображений. Это метод, который для указанной на снимке области способен осуществлять поиск похожих

на нее областей по структуре и содержимому, что поможет автоматизировать процесс анализа и классификации тканей и клеток. Это позволит улучшить точность, скорость и надежность интерпретации гистологических препаратов и значительно облегчит работу врачей-гистологов.

Любой метод поиска по содержимому упирается в качество извлекаемых признаков. Поэтому многие исследования в данной теме посвящены именно процессу извлечения признаков. В статье [1] авторы экспериментируют с использованием в качестве экстрактора признаков вариационного автокодировщика[2], а также подхода на основе контрастивного обучения и генеративных сетей. Последние работы, например, [3] используют в качестве нейросетевых экстракторов признаков архитектуру трансформера. Для изображений эта архитектура была адаптирована и получила название Vision Transformer (ViT) [4]. Стоит однако отметить, что в области анализа полнослайдовых гистологических изображений методы поиска по содержимому развиты еще не достаточно сильно. В том числе из-за малого количества общедоступных наборов данных для валидации данных методов. В текущей работе описывается процесс разработки метода поиска по содержимому для гистологических изображений, включая полнослайдовые. Основной упор делается на обучение нейросетевых экстракторов признаков: автокодировщика и экстрактора, использующего гиперболические расстояния. Также исследует- ся эффективность нейросетевых экстракторов по сравнению с подходом, использующим в качестве признаков статистические характеристики изображения.

II. Использованные наборы данных

Тестирование и сравнение разработанных методов велось на двух наборах данных. Основной и целевой набор данных – это гистологические снимки из PATH-DT-MSU (WSS1, WSS2)[5, 6, 7]. Этот набор данных со- держит полнослайдовые гистологические снимки различных зон желудочно-кишечного тракта. Снимки сделаны в высоком разрешении с помощью микроскопа Leica DM4000B/DFC495 и сканера Leica SCN400.

PATH-DT-MSU содержит разметку по 5 классам тканей. Для некоторых участков снимков определен тип ткани на этой области. В работе было принято решено использовать данную разметку только при оценке точности метода, а при самой разработке метода разметку никак не задействовать. Это мотивировано желанием получить универсальный метод, работающий с любыми медицинскими изображениями без дополнительной информации.

Из PATH-DT-MSU был сформирован тестовый набор данных, содержащий 20 тыс. патчей размера 256 256 пикселей, взятых с масштаба 20. Патчи, содержащие много заднего фона были отброшены.

Также, в ходе разработки и отладки методов поиска по содержимому были использованы изображения из набора данных Common Objects in Context (COCO)[8]. Этот набор данных содержит 330 тысяч изображений, 200 тысяч из которых размечены.

Набор данных COCO был адаптирован под задачу поиска по содержимому. Из изображений были вырезаны области, размеченные как объекты какого-то класса. Области, в которых объекты разных классов наслаива- лись друг на друга были отброшены. В итоге получился набор изображений с одним конкретным объектом и соответствующая этому объекту метка класса. Изображения имели размер 256 256 пикселей.

Часть классов набора данных COCO являются довольно общими. Поэтому примеры изображений каждого из классов были проанализированы и среди всех классов была выделена часть, имеющая четкую структуру. Фрагменты этих классов были оставлены и в дальнейшем использовались для тестирования разработанных методов поиска по содержимому.

Отметим, что COCO не является гистологическим набором данных и нужен для базовой разработки и настройки обучения нейросетевых моделей, общей оценки пригодности их для задачи поиска по содержимому. Детальная настройка методов и финальная оценка точности производится на гистологическом наборе данных PATH-DT-MSU.

III. Оценка точности работы метода

Для оценки точности поиска похожих изображений можно предположить, что метод работает хорошо, если результаты поиска похожи на изображение-запрос и плохо, если не похожи. Эту простую идею можно заложить в метрики, которые будут использоваться для оценки точности работы метода. Для этого можно использовать разметку объектов на классы. Как в COCO, так и в PATH-DT-MSU можно по разметке определить класс для фрагмента изображения. Правда стоит учитывать, что размеченных областей в PATH-DT-MSU на порядок меньше чем в COCO.

Положим, что изображения разных классов не похожи друг на друга, а изображения одного класса – похожи друг на друга. Тогда можно подавать в метод поиска фрагменты определенного класса. И если метод возвращает изображения других классов как похожие, то он ошибается. А если класс возвращаемых изображений совпадает с классом запроса, то метод отработал корректно. Это именно то, что требуется от метода при применении в задачах гистологии.

Самая простой способ превратить описанную выше идею в меру точности – считать долю результатов того же класса, что и изображение-запрос среди самых похожих на запрос изображений по версии метода:

Пусть classtarget – класс изображения, поданного в метод поиска, classn – класс изображения, находящегося на позиции с номером n в выдаче метода, k – число отранжированных кандидатов (в данной работе равно 10), [classtarget = classn] – индикатор того, что класс изображения-запроса совпадает с классом изображения-кандидата на позиции n в выдаче метода.

Тогда описанная выше мера точности поиска по содержимому будет выглядеть следующим образом:

$$Accuracy@k = \frac{\sum_{n=1}^{k}[class_{target} = class_n]}{k}.$$

Для учитывания порядка выдачи можно перейти к MAP@k – Mean Average Precision для top k [9]:

$$MAP@k = \frac{\sum_{n=1}^{k}\frac{1}{n} * [class_{target} = class_n]}{\sum_{n=1}^{k}\frac{1}{n}},$$

где $class_{target}$ – класс изображения, поданного в метод, $class_n$ – класс изображения, находящегося на n-томместе в выдаче метода, k – чисто отранжированных кандидатов.

Данная функция учитывает соответствие класса запроса классу результата на n-том месте с весом 1/n. Таким образом, ошибки на первых результатах выдачи сильнее сказываются на итоговом значении функции, чем ошибки на последних результатах. Также значение функции приводится к диапазону [0, 1] путем деления на $\sum_{n=1}^{k}\frac{1}{n}$.

Обе меры точности хорошо подходят для оценки исследуемых методов и будут использованы далее.

IV. Статистический экстрактор признаков

В качестве первого экстрактора признаков использовались статистические характеристики изображения [10, 11], вычисляемые как для изображения-запроса, так и для всех фрагментов-кандидатов. Используется 2 группы статических признаков. Первая – обычные статистические характеристики: среднее значение, среднеквадратичное отклонение, вариация, асимметрия, эксцесс и энтропия. Вторая – ряд характеристик для матрицы смежности (англ. Grey Level Co-occurrence Matrix, GLCM). Матрица смежности строится по фиксированному вектору $\bar{d} = (dr, dc)$, задающему смещение между пикселом яркостью i и пикселом яркостью j:

$$C_{\bar{d}}[i,j] = |\{(r,c)|I[r,c] = i, I[r+dr, c+dc] = j\}|.$$

Для компактного описания GLCM используются числовые признаки: энергия, энтропия, контрастность, однородность, различие и корреляция.

В работе использовалась матрица GLCM со смещением в 5 пикселей в направлении 0 градусов. Итого получалось 6 признаков для исходного изображения и 6 признаков для матрицы смежности. Таким образом вычисляется вектор признаков длины 12.

В качестве функции расстояния между векторами признаков использовалось косинусное расстояние [12]. Из-за того, что описанные выше признаки имеют сильно отличающийся масштаб, к ним была применена медианная нормализация [13, 14].

В таблице 1 приведены замеры точности статистического экстрактора на наборах данных COCO и PATH- DT-MSU с разбивкой по классам, а также усредненные по классам значения Accuracy@10 и MAP@10.

Точность статистических признаков на гистологических данных на порядок превосходит точность на сложноструктурированных объектах из набора данных COCO. Это говорит о том, что гистологические снимки имеют более выраженные текстурные особенности.

Также были проведены эксперименты по оценке влияния различных групп статистических признаков на точность поиска. Сначала для поиска по содержимому были использованы только показатели MEAN и STD.

ТАБЛИЦА 1
ТОЧНОСТЬ СТАТИСТИЧЕСКОГО ЭКСТРАКТОРА

COCO

Класс	*Accuracy@10*	*MAP@10*
Среднее значение	**12.41%**	**0.13**
zebra	**53.33%**	**0.64**
hot dog	26.67%	0.33
broccoli	20.00%	0.2
umbrella	20.00%	0.18
bear	16.67%	0.23
donut	16.67%	0.14
giraffe	16.67%	0.12
sheep	16.67%	0.094
banana	13.33%	0.24
bus	6.67%	0.027
elephant	6.67%	0.034
cat	3.33%	0.019
motorcycle	3.33%	0.019
person	3.33%	0.028
cow	0.00%	0.0
dog	0.00%	0.0
horse	0.00%	0.0
potted plant	0.00%	0.0

PATH-DT-MSU

Класс	*Accuracy@10*	*MAP@10*
Среднее значение	**65.00%**	**0.66**
AT	**96.20%**	**0.97**
LP	85.80%	0.88
TUM	57.40%	0.57
MM	20.60%	0.21

После к ним были добавлены остальные признаки, не использующие характеристики матрицы смежности. В конце была замерена точность на всех признаках включая GLCM. В таблице 2 представлены результаты измерений точности различных комбинаций статистических признаков.

ТАБЛИЦА 2
ТОЧНОСТЬ СТАТИСТИЧЕСКОГО ЭКСТРАКТОРА НА НАБОРЕ PATH-DT-MSU ДЛЯ РАЗНЫХ НАБОРОВ ПРИЗНАКОВ

Набор признаков	*Accuracy@10*	*MAP@10*
Только MEAN и STD	50.40%	0.5
Статистики без GLCM	59.35%	0.61
Все признаки	**65.00%**	**0.66**

V. Экстрактор на основе автокодировщика

Для повышения точности извлекаемых из изображений признаков была предпринята попытка использовать сверточный автокодировщик [15]. Автокодировщик – это нейросетевая модель, состоящая из двух основных частей: кодировщика и декодировщика. Кодировщик в ходе обучения учится компактно описывать входное изображение, а декодировщик – по этому описанию восстанавливать исходное изображение. На выходе кодировщика (сжимающей части автокодировщика) обучается векторное представление, которое в дальнейшем будет использоваться в методе поиска по содержимому в качестве вектора признаков.

При обучении автокодировщика в качестве функции потерь используется L2 расстояние между входным и выходным изображениями.

Также в данной работе была использована идея маскирования части регионов входного изображения, описанная в работе [16]. Часть регионов входного изображения маскируются, и от модели требуется восстановить их. Это делается для того, чтобы научить модель лучше понимать структуру входных изображений. Пример процесса обучения автокодировщика с маскированием представлен на Рис. 1.

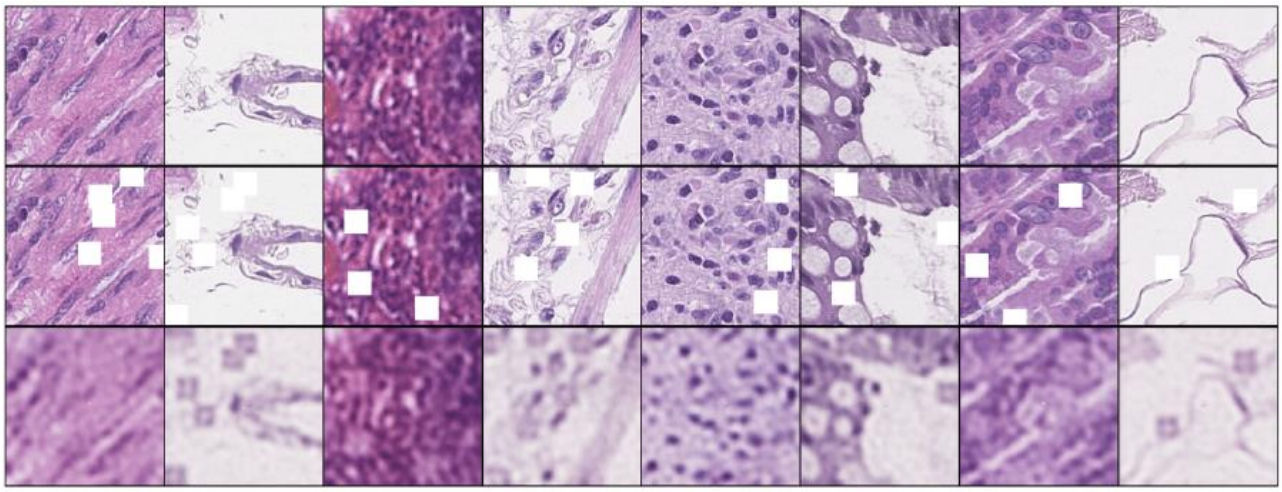


Рис. 1. Пример обучения автокодировщика с маскированием на наборе данных PATH-DT-MSU. Верхний ряд - исходные изображения. Средний ряд - изображения с замаскированными регионами, подаваемые на вход автокодировщику. Нижний ряд - восстановленные автокодировщиком изображения.

В качестве функции расстояния между векторами признаков использовалось косинусное расстояние [12].

В таблице 3 приведены замеры точности экстрактора на основе автокодировщика на наборах данных COCO и PATH-DT-MSU с разбивкой по классам, а также усредненные по классам значения Accuracy@10 и MAP@10.

ТАБЛИЦА 3
ТОЧНОСТЬ ЭКСТРАКТОРА НА ОСНОВЕ АВТОКОДИРОВЩИКА

COCO

Класс	*Accuracy@10*	*MAP@10*
Среднее значение	**14.44%**	**0.17**
zebra	3.33%	0.016
hot dog	13.33%	0.24
broccoli	40.00%	0.52
umbrella	20.00%	0.15
bear	16.67%	0.18
donut	20.00%	0.15
giraffe	0.00%	0.0
sheep	10.00%	0.063
banana	**63.33%**	**0.64**
bus	16.67%	0.29
elephant	13.33%	0.3
cat	13.33%	0.24
motorcycle	0.00%	0.0
person	0.00%	0.0
cow	6.67%	0.0
dog	10.00%	0.11
horse	13.33%	0.03
potted plant	6.67%	0.13

PATH-DT-MSU

Класс	*Accuracy@10*	*MAP@10*
Среднее значение	**44.80%**	**0.46**
AT	**100.00%**	**1.0**
LP	31.00%	0.33
TUM	28.80%	0.29
MM	19.40%	0.22

Средние показатели точности на COCO слегка превосходят показатели статистического экстрактора, но точность поиска все еще неудовлетворительная.

На наборе данных PATH-DT-MSU для класса AT экстрактор показывает максимальные показатели точности, но при этом точность на остальных классах оказывается заметно хуже. Из-за этого средние показатели точности оказываются хуже чем у статистического экстрактора. Это можно объяснить тем, что выход кодировщика, который в итоге используется в качестве экстрактора признаков, никак напрямую не участвует в функции потерь. Из-за этого на итоговые вектора признаков напрямую не накладываются никакие требования, и признаки не имеют хороших метрических свойств.

VI. Экстрактор, использующий гиперболические расстояния

Идея следующего экстрактора признаков была взята из работы [17]. В этой статье авторы описывают схему обучения экстрактора признаков, показывающего лучшие на момент написания работы результаты на 3 наборах данных для поиска похожих изображений: CARS196 [18], CUB-200-2011 [19] и Stanford Online Products [20].

Авторы предлагают использовать при обучении экстрактора гиперболические расстояния и неевклидову геометрию. Поэтому для краткости иногда будем называть описываемый в этом пункте экстрактор "гиперболическим".

На Рис. 2 изображена схема обучения гиперболического экстрактора признаков. Обучение экстрактора состоит из двух частей. Сначала целевая модель применяется к неразмеченному набору изображений. Полученные вектора признаков с помощью специального слоя переводятся в гиперболическое пространство. Далеек этим векторам признаков применяется несколько уровней иерархической кластеризации [21]. Изображениям присваиваются метки кластеров, в которые они вошли. Затем эти метки используются при оптимизации целевой модели. Происходит новая итерация обучения: новой версией модели извлекаются признаки, кластеризуются, модель оптимизируется. Рассмотрим детальнее все этапы и особенности этой схемы обучения.

1. Иерархическая кластеризация

Иерархическая кластеризация – это один из алгоритмов кластеризации, цель которого – объединить близкиепо некоторому критерию объекты в группы. На входе алгоритма множество векторов признаков, на выходе – отображение этих векторов в метки кластеров.

Данный алгоритм производит итеративное слияние разных кластеров друг с другом по определенному критерию до достижения какого-то определенного условия. В данной работе условие остановки – достижение определенного числа кластеров, например когда количество кластеров станет равно 100.

Критерий слияния устроен следующим образом. Пусть изначально каждый вектор признаков представляетиз себя кластер размера один. Посчитаем для каждой пары кластеров a и b межкластерное расстояние d_{ab} по следующей формуле:

$$d_{ab} = \frac{1}{n_a n_b} \sum_{\mathbf{z}_i^a \in C_a, \mathbf{z}_j^b \in C_b} \left\| \mathbf{z}_i^a - \mathbf{z}_j^b \right\|,$$

где C_a и C_b – множество векторов кластеров a и b соответственно, $\mathbf{z}^a$ и $\mathbf{z}^b$ – векторы из кластеров a и b, n_a и n_b – размеры кластеров a и b.

Объединим друг с другом кластеры с наименьшим межкластерным расстоянием. Повторим процедуру до достижения условия остановки (например, что после очередного слияния число кластеров стало равно 100).

Описанная выше схема обучения целевой модели пытается с помощью иерархической кластеризации испра- вить недостатки стандартного подхода Metric Learning [22]. В классической постановке Metric Learning берется якорный объект (Anchor). Для него берутся позитивный (Positive sample) и негативный (Negative sample) при-меры. Далее с помощью специальной функции потерь целевая модель оптимизируется так, чтобы позитивный пример приблизился к якорному объекту, а негативный отдалился от него.

В этом подходе присутствует недостаток, который проиллюстрирован на примере Рис. 3.

Пусть позитивный пример – объект того же класса, что и якорный объект, а негативный пример – объект другого класса. Однако такое бинарное отношение не учитывает степень близости объектов. Например, для якорного объекта класса "Black Footed Albatross"и "Sooty Albatross и "Florida Jay"будут одинаково отдалены,хотя "Sooty Albatross"ближе, так как оба принадлежат к классу "Albatross".

Для учета этой близости можно ввести иерархию классов и на её основе определять степень близости объектов, сближая или отдаляя их соответственно.

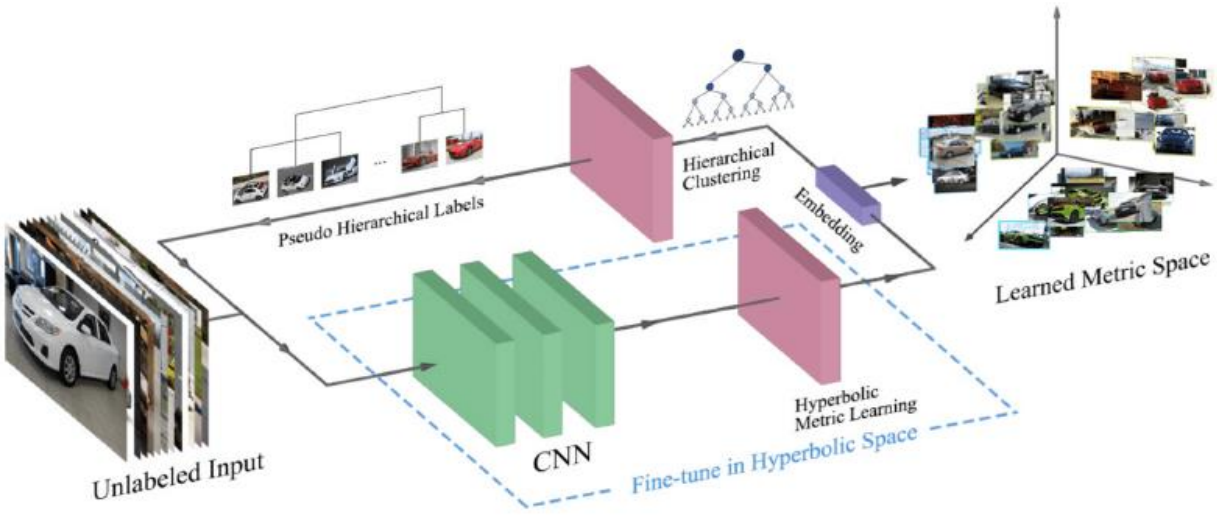


Рис. 2. Схема обучения гиперболического экстрактора признаков.

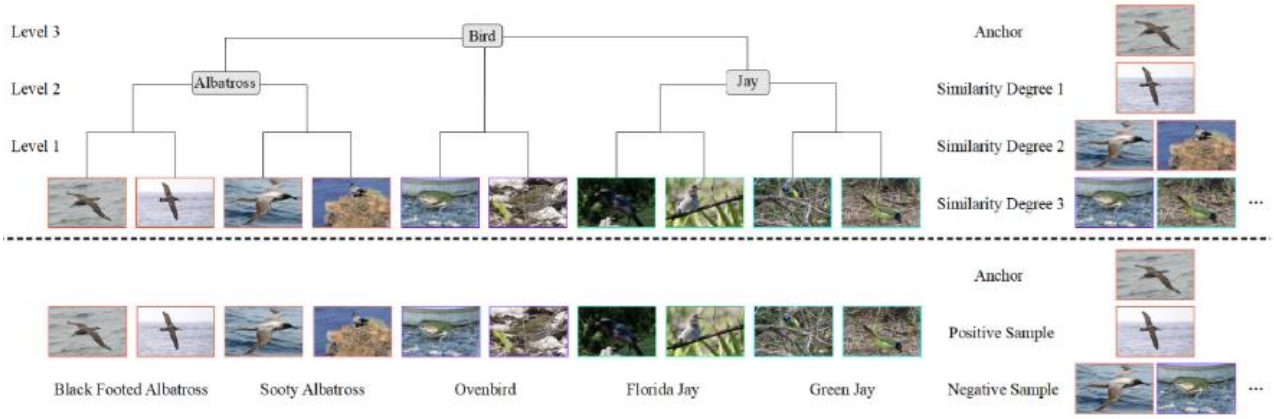


Рис. 3. Проблема неучета разного уровня близости объектов в методах Metric Learning.

В данной задаче разметка на классы не используется, поэтому предлагается создать её с помощью иерархической кластеризации. Применим этот алгоритм несколько раз, постепенно уменьшая число кластеров, например: 100 →50 →10 кластеров. Сначала объекты делятся на 100 небольших кластеров, затем они объединяются в 50, а потом в 10 более крупных кластеров. Кластеры получаются вложенными, что позволяет каждому объектуполучить три метки вложенных кластеров, которые будут использованы в процессе обучения.

2. Переход к гиперболическим расстояниям

В конце этапа извлечения признаков целевой моделью, перед стадией кластеризации к векторам признаков применяется преобразование, переводящее эти вектора в гиперболическое пространство:

$$\mathbf{z} = \exp^{\tau}(\mathbf{x}) := \tanh(\sqrt{\tau}\|\mathbf{x}\|)\frac{\mathbf{x}}{\sqrt{\tau}\|\mathbf{x}\|},$$

где $\mathbf{x}$ – исходный вектор признаков, а τ – это кривизна шара Пуанкаре.

Расстояние между двумя векторами $\mathbf{z}_i$ и $\mathbf{z}_j$ в гиперболическом пространстве определяется следующим образом:

$$d_D(\mathbf{z}_i, \mathbf{z}_j) = \cosh^{-1}\left(1 + 2\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\left(1 - \|\mathbf{z}_i\|^2\right)\left(1 - \|\mathbf{z}_j\|^2\right)}\right).$$

После этого преобразования к переведенным в гиперболическое пространство векторам применяется этап иерархической кластеризации, описанный выше.

Перевод векторов признаков в гиперболическое пространство обусловлен более удобной работой с иерархическими структурами и отношениями. А именно это и происходит на этапе построения последовательности вложенных кластеров.

3. Функция потерь

Выше был описан процесс получения векторов признаков в гиперболическом пространстве и присвоениеим последовательности вложенных меток кластеров. Для обучения модели используется схема, подобная упомянутой выше Anchor-Positive-Negative схеме в Metric Learning [22]. Будем рассматривать тройки объектов $\{\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_l\} \in \mathscr{S}$. Для них определим функцию потерь следующим образом:

$$\mathscr{L}(i,j,l) = \left(\log \frac{\|\mathbf{z}_i - \mathbf{z}_j\|}{\|\mathbf{z}_i - \mathbf{z}_l\|} - \log \Omega^{s_{ij} - s_{il}} \right)^2,$$

где $s_{ij} \in \{1,2,\dots,K\}$ – степень близости объектов по меткам, полученным на этапе кластеризации. Она принимает значение уровня иерархии кластеров, на котором объекты начинают принадлежать к одинаковомукластеру. $\Omega > 0$ – это гиперпараметр метода, определяющий вклад разницы s_{ij} и s_{il} в функцию потерь.

Функция потерь состоит из двух компонентов. Первая – логарифм отношения расстояний между двумя парами объектов, вторая отвечает за степень близости объектов в дереве классов, образованных кластеризацией.В такой конструкции сила сближения/отдаления объектов будет зависеть от того, на каком уровне иерархии кластеров объекты принадлежат одному кластеру.

В таблице 4 приведены замеры точности гиперболического экстрактора на наборах данных COCO и PATH-DT-MSU с разбивкой по классам, а также усредненные по классам значения Accuracy@10 и MAP@10.

ТАБЛИЦА 4
ТОЧНОСТЬ ГИПЕРБОЛИЧЕСКОГО ЭКСТРАКТОРА

COCO

Класс	*Accuracy@10*	*MAP@10*
Среднее значение	**66.48%**	**0.66**
zebra	**100.00%**	**1.0**
hot dog	60.00%	0.59
broccoli	93.33%	0.93
umbrella	86.67%	0.89
bear	50.00%	0.43
donut	40.00%	0.4
giraffe	**100.00%**	**1.0**
sheep	46.67%	0.47
banana	**100.00%**	**1.0**
bus	**100.00%**	**1.0**
elephant	73.33%	0.72
cat	56.67%	0.51
motorcycle	**100.00%**	**1.0**
person	83.33%	0.86
cow	30.00%	0.34
dog	10.00%	0.07
horse	43.33%	0.4
potted plant	23.33%	0.25

PATH-DT-MSU

Класс	*Accuracy@10*	*MAP@10*
Среднее значение	**73.45%**	**0.74**
AT	75.40%	0.75
LP	**82.20%**	**0.84**
TUM	71.60%	0.71
MM	64.60%	0.66

Также были проведены эксперименты по добавлению в обучающее множество фрагментов не только из размеченных областей полнослайдовых изображений, но и просто фрагментов из случайных областей. В описанных выше подходах на гистологических изображениях с масштаба ×20 брались выделенные области определенного класса, из них случайно вырезались фрагменты размера 256 × 256 пикселей. На этих фрагментах происходило обучение. То, что фрагменты, на которых проводилось обучение, брались только из определенных областей потенциально могло бы вносить смещение в модель и ее итоговая обобщающая способность бы падала.

Чтобы уменьшить этот эффект в последующих экспериментах в обучающее множество также добавлялись случайные фрагменты из любой области изображения. Также был проведен эксперимент, где обучение происходило только на фрагментах из случайных областей.

В итоге было рассмотрено три конфигурации обучающих данных: фрагменты только из размеченных областей, фрагменты из размеченных областей + фрагменты из неразмеченных областей и фрагменты только из неразмеченных областей (таблица 5).

ТАБЛИЦА 5
ТОЧНОСТЬ ГИПЕРБОЛИЧЕСКОГО ЭКСТРАКТОРА ПРИ РАЗНЫХ СПОСОБАХ ФОРМИРОВАНИЯ ОБУЧАЮЩЕГО МНОЖЕСТВА ИЗ НАБОРА ДАННЫХ PATH-DT-MSU

Источники фрагментов	*Accuracy@10*	*MAP@10*
Только из размеченных областей	**73.45%**	**0.74**
Размеченные области + неразмеченные области	64.60%	0.67
Только из неразмеченных областей	70.35%	0.7

Из таблицы 5 видно, что лучше показатели точности у метода, который учился только на фрагментах из размеченных областей. Это довольно предсказуемо и объяснимо тем, что множество фрагментов, на котором происходит оценка метода, тоже взято из этих областей. Точность модели, обученной только на случайных фрагментах тоже выглядит хорошей, так как эта модель обучалась на менее смещенных данных, возможно именно ее стоит использовать на практике.

VII. Итоговое сравнение методов

В таблице 6 приведено сравнение всех описанных выше методов и экспериментов с ними.

Видно, что метод поиска по содержимому с использованием гиперболического экстрактора показывает наилучшие результаты как на наборе данных COCO, так и на PATH-DT-MSU. В качестве финального тестирования данный метод был применен к случайным областям из PATH-DT-MSU. Эти условия максимально близки к тем, в которых метод будет использоваться в реальных задачах врачами-гистологами. Пример такого поиска можно увидеть на Рис. 4.

Результаты поиска выглядят похожими на изображения-запросы, из чего можно сделать вывод, что данный метод может быть потенциально использован врачами в качестве вспомогательного инструмента.

VIII. Заключение

В рамках данной работы были разработаны и исследованы различные подходы к извлечению структурных признаков из изображений и изучено их влияние на работу метода поиска по содержимому как для изображений общего типа, так и для полнослайдовых гистологических изображений желудочно-кишечного тракта.

В итоге был реализован метод поиска по содержимому, показывающий высокую точность поиска на наборах данных COCO и PATH-DT-MSU. Наилучший вариант метода достигает 73.45% совпадений класса изображения-запроса с 10 лучшими разультатами поиска. Также описанная в работе схема метода поиска по содержимому универсальна и может быть перенесена на другие источники данных.

Из приведенных результатов видно, что разработанный метод действительно ищет похожие фрагменты изображений и потенциально может быть использован врачами-гистологами в качестве вспомогательного инструмента при проведении диагностических мероприятий.

Метод программно реализован на языке Python 3 и выложен в открытый доступ: https://github.com/ArtemVeshkin/CBIR-v2.

ТАБЛИЦА 6
СРАВНЕНИЕ ТОЧНОСТИ ВСЕХ ОПИСАННЫХ ЭКСТРАКТОРОВ ПРИЗНАКОВ

Метод извлечения признаков	*Набор данных*	*Accuracy@10*	*MAP@10*
Статистический экстрактор	COCO	12.41%	0.13
	PATH-DT-MSU	65.00%	0.66
Статистический экстрактор (только MEAN и STD)	PATH-DT-MSU	50.40%	0.5
Статистический экстрактор (без GLCM)	PATH-DT-MSU	59.35%	0.61
Метод на основе автокодировщика	COCO	14.44%	0.17
	PATH-DT-MSU	44.80%	0.46
Гиперболический экстрактор	COCO	**66.48%**	**0.66**
	PATH-DT-MSU	**73.45%**	**0.74**
Гиперболический экстрактор (с добавлением случайных фрагментов)	PATH-DT-MSU	64.60%	0.67
Гиперболический экстрактор (только на случайных фрагментах)	PATH-DT-MSU	70.35%	0.7

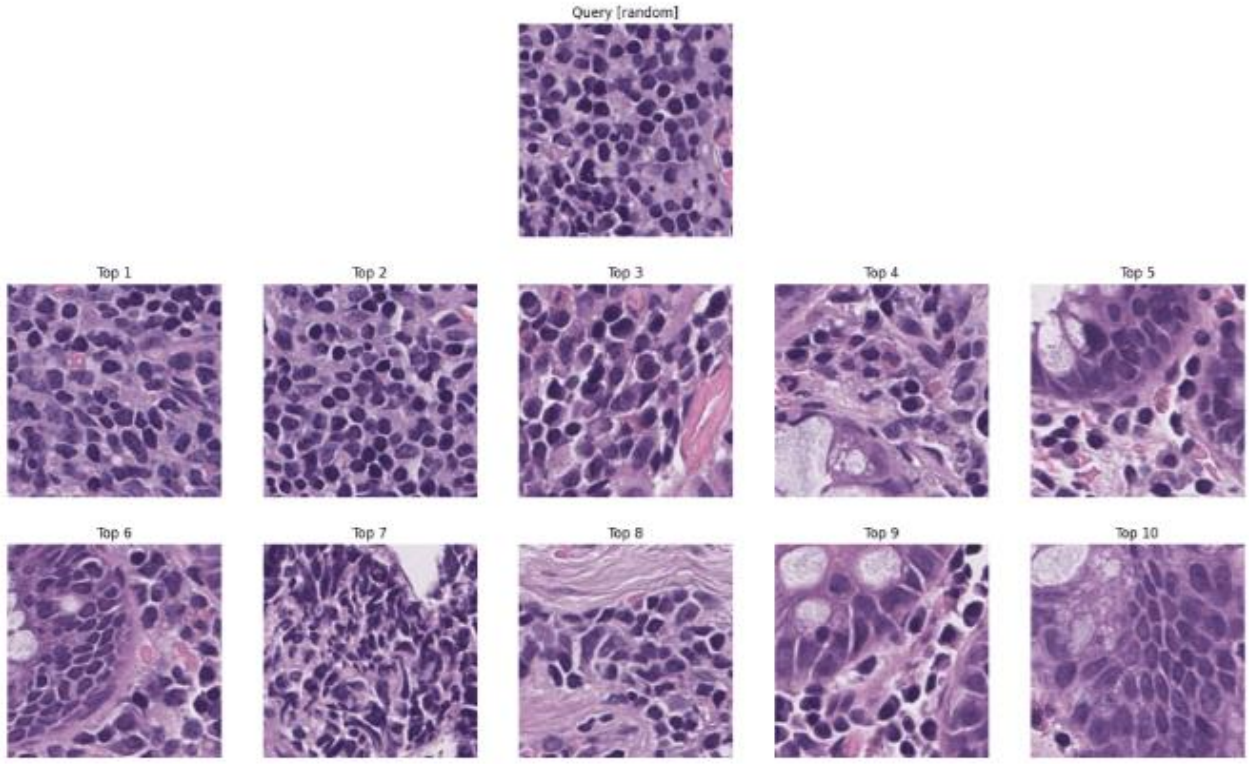


Рис. 4. Пример поиска похожих изображений для случайной области набора данных PATH-DT-MSU с помощью разработанного гиперболического экстрактора. Сверху - изображение-запрос.

Благодарности

Работа выполнена при поддержке гранта Российского Научного Фонда 22-41-02002.

Список литературы

[1] Neural Image Compression for Gigapixel Histopathology Image Analysis / David Tellez, Geert Litjens, Jeroen van der Laak, Francesco Ciompi // IEEE transactions on pattern analysis and machine intelligence.
— 2019. — Vol. 43, no. 2. — Pp. 567–578.

[2] Kingma Diederik P, Welling Max. Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. — 2013.

[3] Gkelios Socratis, Boutalis Yiannis, Chatzichristofis Savvas A. Investigating the vision transformer model for image retrieval tasks // 2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS) / IEEE. — 2021. — Pp. 367–373.

[4] A survey on vision transformer / Kai Han, Yunhe Wang, Hanting Chen et al. // IEEE transactions on pattern analysis and machine intelligence. — 2022. — Vol. 45, no. 1. — Pp. 87–110.

[5] Tissue type recognition in whole slide histological images / Alexander Khvostikov, Andrey Krylov, Ilya Mikhailov et al. // CEUR Workshop Proc. 3027. — Vol. 50. — 2021.

[6] Development of CNN-based algorithm for automatic recognition of the layers of the wall of the stomach and colon / I Mikhailov, A Khvostikov, A Krylov et al. // VIRCHOWS ARCHIV / SPRINGER ONE NEW YORK PLAZA, SUITE 4600, NEW YORK, NY, UNITED STATES. — Vol. 479. — 2021. — Pp. S36–S37.

[7] Veshkin AS, Khvostikov AV. Multiscale Content-Based Image Retrieval for Whole-Slide Histological Images //Computational Mathematics and Modeling. — 2022. — Vol. 33, no. 2. — Pp. 244–254.

[8] Microsoft COCO: Common objects in context / Tsung-Yi Lin, Michael Maire, Serge Belongie et al. // Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 / Springer. — 2014. — Pp. 740–755.

[9] Learning with average precision: Training image retrieval with a listwise loss / Jerome Revaud, Jon Almaza´n, Rafael S Rezende, Cesar Roberto de Souza // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2019. — Pp. 5107–5116.

[10] Kvostikov AV, Krylov Andrey S, Kamalov UR. Ultrasound image texture analysis for liver fibrosis stage diagnostics // Programming and Computer Software. — 2015. — Vol. 41. — Pp. 273–278.

[11] Ultrasound despeckling by anisotropic diffusion and total variation methods for liver fibrosis diagnostics / Alexander Khvostikov, Andrey Krylov, Julius Kamalov, Alina Megroyan // Signal Processing: Image Communication. — 2017. — Vol. 59. — Pp. 3–11.

[12] Similarity between Euclidean and cosine angle distance for nearest neighbor queries / Gang Qian, Shamik Sural, Yuelong Gu, Sakti Pramanik // Proceedings of the 2004 ACM symposium on Applied computing. — 2004. — Pp. 1232–1237.

[13] Singh Dalwinder, Singh Birmohan. Investigating the impact of data normalization on classification performance //Applied Soft Computing. — 2020. — Vol. 97. — P. 105524.

[14] A comparison of normalization techniques for microRNA microarray data / Youlan Rao, Yoonkyung Lee, David Jarjoura et al. // Statistical applications in genetics and molecular biology. — 2008. — Vol. 7, no. 1.

[15] Autoencoder and its various variants / Junhai Zhai, Sufang Zhang, Junfen Chen, Qiang He // 2018 IEEE international conference on systems, man, and cybernetics (SMC) / IEEE. — 2018. — Pp. 415–419.

[16] Context encoders: Feature learning by inpainting / Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — Pp. 2536–2544.

[17] Unsupervised hyperbolic metric learning / Jiexi Yan, Lei Luo, Cheng Deng, Heng Huang // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2021. — Pp. 12465–12474.

[18] 3d object representations for fine-grained categorization / Jonathan Krause, Michael Stark, Jia Deng, Li Fei-Fei // Proceedings of the IEEE international conference on computer vision workshops. — 2013. — Pp. 554–561.

[19] The caltech-ucsd birds-200-2011 dataset / Catherine Wah, Steve Branson, Peter Welinder et al. — 2011.

[20] Deep metric learning via lifted structured feature embedding / Hyun Oh Song, Yu Xiang, Stefanie Jegelka, Silvio Savarese // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — Pp. 4004–4012.

[21] Murtagh Fionn, Contreras Pedro. Algorithms for hierarchical clustering: an overview // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. — 2012. — Vol. 2, no. 1. — Pp. 86–97.

[22] Kaya Mahmut, Bilge Hasan S¸akir. Deep metric learning: A survey // Symmetry. — 2019. — Vol. 11, no. 9. — P. 1066.

УДК 004.8
DOI: 10.25206/978-5-8149-3873-2-2024-629-635

Исследование эффективности современных архитектур сверточных нейронных сетей в задаче распознавания кожных новообразований
Study of the effectiveness of modern convolutional neural network architectures in the task of recognizing skin neoplasms

А.С. Михалев1,2, В.А. Жуковская1, М.А. Кобелев2, А. В. Пятаева1, А.Е. Гец2
[1]Центр искусственного интеллекта СФУ, г. Красноярск
[2]Каф «Программная инженерия» ИКИТ СФУ, г. Красноярск

A.S. Mikhalev1,2, V.A. Zhukovskaya1, M.A. Kobelev2, A.V. Pyataeva1, A.E. Getz2
[1]Center of Artificial Intelligence SFU, Krasnoyarsk
[2]Department of "Software Engineering", Institute of Computer Science and Information Technologies, SFU, Krasnoyarsk

Аннотация. В статье рассматривается проблема роста заболеваемости раком кожи, особенно меланомой, которая является наиболее агрессивной и опасной формой этого заболевания. Основное внимание уделено инновационным технологиям диагностики кожных новообразований, в частности, применению искусственного интеллекта. В ходе исследования анализируется эффективность различных архитектур сверточных нейронных сетей для распознавания кожных новообразований на основе цифровых снимков. Исходные данные включают 25331 изображение различных типов кожных поражений из набора данных ISIC 2019. Для обучения и тестирования использованы методы аугментации данных для борьбы с дисбалансом классов. Было исследовано 20 моделей сверточных нейронных сетей. Основные метрики для оценки эффективности моделей включают ROC-AUC, accuracy, precision, recall и F-мера. Результаты экспериментов показали, что наилучшие показатели продемонстрировала модель EfficientNet_v2_s. Подбор гиперпараметров для этой модели позволил достичь улучшения метрик: ROC-AUC 0.923, accuracy 0.917, precision 0.833, recall 0.786 и F-мера 0.796.Таким образом, использование современных архитектур сверточных нейронных сетей и методов машинного обучения позволяет существенно повысить точность и надежность диагностики кожных новообразований, что имеет важное значение для раннего выявления и лечения рака кожи.

***Ключевые слова*:** рак кожи, онкология, меланома, визуальные данные, глубокое обучение, свёрточные нейронные сети

Аннотация. *The article deals with the problem of increasing incidence of skin cancer, especially melanoma, which is the most aggressive and dangerous form of this disease. The main attention is paid to innovative technologies for diagnosing skin neoplasms, in particular, the use of artificial intelligence. The study analyzes the effectiveness of different convolutional neural network architectures for recognizing skin neoplasms based on digital images. The raw data includes 25331 images of different types of skin lesions from the ISIC 2019 dataset. For training and testing, data augmentation techniques were used to deal with class imbalance. Twenty convolutional neural network models were investigated. The main metrics to evaluate the performance of the models include ROC-AUC, accuracy, precision, recall and F-measure. Experimental results showed that the EfficientNet_v2_s model demonstrated the best performance. Selection of hyperparameters for this model allowed to achieve improvement of metrics: ROC-AUC 0.923, accuracy 0.917, precision 0.833, recall 0.786 and F-measure 0.796. Thus, the use of modern convolutional neural network architectures and machine learning methods can significantly improve the accuracy and reliability of skin neoplasms diagnosis, which is important for early detection and treatment of skin cancer.*

***Ключевые слова*:** *skin cancer, oncology, melanoma, visual data, deep learning, convolutional neural networks*

I. Введение

Рак кожи [1] – это одно из самых частых онкологических заболеваний, встречающихся у людей всех возрастных групп. В последние десятилетия наблюдается значительный рост заболеваемости, что связано с увеличением продолжительности жизни, изменением климатических условий и повышением интенсивности ультрафиолетового излучения. Основные виды рака кожи включают базально-клеточную карциному, плоскоклеточную карциному и меланому, которая является наиболее агрессивной и опасной формой этого заболевания. Факторы риска развития рака кожи разнообразны и включают генетическую предрасположенность,

длительное пребывание на солнце без защиты, частое посещение соляриев, а также наличие предраковых состояний кожи. Важную роль в профилактике и ранней диагностике рака кожи играют регулярные осмотры дерматологом и самообследование кожи. Когда меланома диагностируется, оставаясь всё ещё на внешних слоях кожи, простое удаление обычно ведёт к излечению пациента, а относительная выживаемость в течение 5 лет составляет приблизительно 98%. К сожалению, несмотря на склонность меланомы к ранней диагностике посредством простого визуального осмотра, у многих пациентов по-прежнему диагностируется более прогрессирующее заболевание. В связи с этим актуальной становится задача разработки новых технологий для диагностики кожных новообразований. Одной из таких инновационных технологий является применение искусственного интеллекта, который показывает большие перспективы в области дерматологии.

Технологии искусственного интеллекта могут существенно улучшить процессы выявления и анализа кожных новообразований благодаря своей способности обрабатывать и интерпретировать огромные объемы данных с высокой скоростью и точностью. Современные алгоритмы машинного обучения, обученные на базе тысяч и миллионов изображений различных типов кожных опухолей, способны распознавать малейшие признаки злокачественных изменений, которые могут быть не видны невооруженным глазом. Целью данной статьи является исследование и сравнение эффективности различных архитектур сверточных нейронных сетей в задаче распознавания кожных новообразований на основании анализа цифрового снимка и выявление наиболее эффективных из них.

II. Постановка задачи

Заболеваемость кожной меланомой с 1979 года возрастает каждый год. Несмотря успешную диагностику меланомы на ранних стадиях посредством простого визуального осмотра, к сожалению, у многих пациентов меланома по-прежнему диагностируется как активно прогрессирующее заболевание. В результате, по статистике, более 100 тыс. людей умирают от меланомы каждый год. Хотя поражения кожи видны невооруженным глазом, меланомы на ранних стадиях может быть трудно отличить от доброкачественных поражений кожи с похожими проявлениями.

Первостепенным методом оценки подозрительных новообразований является как дерматоскопия [2]. Затем, если врач может производить выполнение анализа на онкомаркеры или биопсию. Дерматоскопия – метод исследования поверхностных поражений кожи с помощью специального прибора – дерматоскопа. Это инструмент, похожий на камеру, позволяющий осматривать кожные поражения без отражений от поверхности кожи. Дерматоскоп состоит из лупы, источника света (поляризованного или неполяризованного), прозрачной пластины и иногда жидкой среды между инструментом и кожей. Современные цифровые дерматоскопы способны обрабатывать информацию и записывать ее в цифровом виде. Это неинвазивный метод, который позволяет увидеть структуру кожи с увеличением в 10–20 раз. Врач может детально рассмотреть эпидермис, его цвет, дермо-эпидермальные соединения и сосочковый слой дермы. Ценность данной методики заключается в высокой диагностической точности, возможности обеспечить раннюю диагностику злокачественных новообразований кожи, что имеет чрезвычайное значение в связи с ростом заболеваемости меланомой и раком кожи.

Изображения с дерматоскопа могут быть классифицированы на доброкачественные и злокачественные путем анализа цифрового снимка. Это позволит распознать тип кожных образований, обнаруживая подозрительные образования на ранних стациях, когда лечение наиболее эффективно. Имея достаточно объёмный набор цифровых снимков кожных новообразований, путем применения методов компьютерного зрения и машинного обучения можно предсказывать тип кожного новообразования.

III. Литературный обзор

В последние годы использование сверточных нейронных сетей для распознавания кожных новообразований стало одной из наиболее активно развивающихся областей в медицинской диагностике. В данном разделе рассмотрены полученные достижения и исследования, посвященные применению различных архитектур для распознавания кожных новообразований.

В работе [3] авторы предварительно обучили модель сверточной нейронной сети на наборе дерматоскопических изображений ISIC 2017. Далее они применили технику переноса обучения с помощью тонкой настройки. Модель сверточной нейронной сети обучили на наборе данных из 206 оптических изображений, составленных из наборов данных DermQuest и Dermatology Information System. В работе [4] для задачи распознавания новообразований также использовалась техника переноса обучения. В качестве модели были исследованы такие архитектуры сверточных нейронных сетей как VGG16, ResNet18 и ResNet50. В результате авторами был сделан вывод, что техника переноса обучения с большого набора дерматоскопических

данных на небольшой оптический набор данных, позволяет получить достаточно высокие значения метрик - 97% правильно обнаруженных новообразований. В статье [5] основная идея исследования заключается в использовании подхода Image Super Resolution (ISR): способе улучшения изображения путем масштабирования, выделения краев и устранения неровности. Подход ISR повышает начальную точность моделей на 15,59 % для VGG16, 13,85 % для ResNet и 7,78 % для Inception V3.

В сравнительной статье [6] также для решения задачи классификации рака кожи использовались предварительно обученные нейронные сети AlexNet, GoogleNet, VGG16, VGG19, ResNet18, ResNet50, ResNet101, ShuffleNet, MobileNet и DenseNet201 в связке с такими классификаторами как метод опорных векторов (SVM), алгоритм k-ближайших соседей и деревья решений. Для обучения и тестирования использовался набор данных ISBI 2016. Наилучшую точность показала 88,02 %.гибридная модель с использованием DenseNet-201 и MobileNet с классификатором SVM. В статье [7] авторы используют предобученную нейронную сеть ResNet50. Для тренировки использовался набор данных International Skin Imaging Collaboration, включающий в себя 4204 изображения из открытых источников и проверенные биопсией. В результате исследования чувствительность и специфичность для классификации дерматологами составили 67,2 % и 62,2 % соответственно. Обученная модель достигла значений чувствительности 82,3 % и специфичности 77,9 %. В исследовании [8] для распознавания рака кожи использовалась предобученная модель MobileNetV2, в которую после сверточных слоев добавлены несколько слоев для классификации новообразований. В результате экспериментальных исследований на наборе данных ISIC-Archive построенная нейронная сеть может обеспечить точность до 85 %. Однако, предложенная архитектура дает низкую чувствительность для наборов данных ISBI 2016 и PH2, показывая, что предложенный метод не смог правильно классифицировать злокачественные новообразования. Это связано с тем, что эти два набора данных имеют высокий дисбаланс, в котором количество злокачественных данных намного меньше, чем количество доброкачественных данных.

Результаты изучения литературных источников показывают, что использование современных архитектур сверточных нейронных сетей, а также техники переноса обучения значительно повышает точность и чувствительность моделей в задачах распознавания кожных новообразований.

III. Описание набора данных кожных новообразований

Для классификации дерматоскопических изображений на предмет определения типа новообразование требуется подготовить набор данных, в котором будут требуемые классы будут сбалансированы между собой. Изображения для работы нейронной сети в настоящей статье получены из открытого набора данных ISIC 2019 Challenge Dataset [9]. Этот набор данных содержит 25 331 различных образований кожи, формат файлов *.jpeg и включает следующие классы: меланома, меланоцитарный невус, базальноклеточная карцинома, старческий кератоз, доброкачественный кератоз, дерматофиброма, сосудистое поражение, плоскоклеточная карцинома. В наборе данных ISIC 2019 Challenge Dataset изображения имеют различные разрешения, минимальное разрешение изображений 640×480 пикселей, максимальное – 6 000×4 000. Пример изображений различных классов из набора данных приведён на рис. 1.

Рис. 1. Категории кожных новообразований

На рис. 2 представлено распределение классов в исходных данных в виде гистограммы. Классы новообразований имеют следующие условные обозначения MEL - меланома, NV - меланоцитарный невус, BBC - базальноклеточная карцинома, AK - старческий кератоз, BKL - доброкачественный кератоз, DF - дерматофиброма, VASC - сосудистое поражение, SCC - плоскоклеточная карцинома.

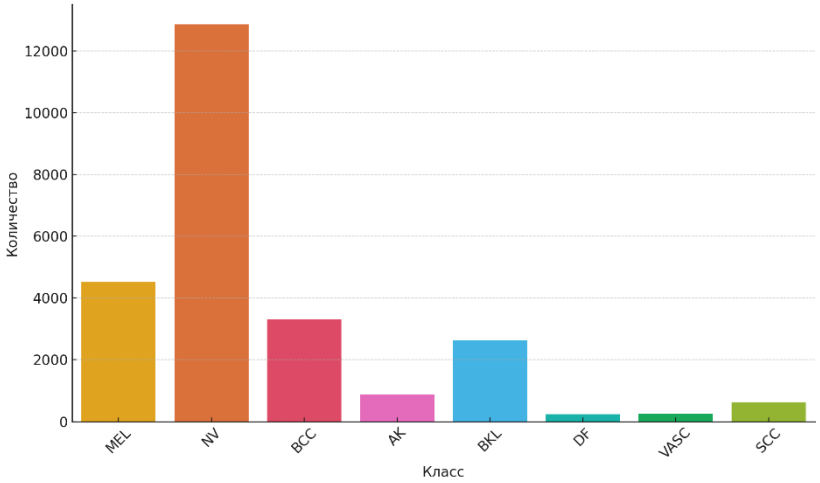


Рис. 2. Распределение классов в исходных данных

Как можно увидеть данные в этом наборе существенно разбалансированы. Качественная и достоверная работа нейронных сетей возможна только при использовании сбалансированного набора данных (рис. 3.)

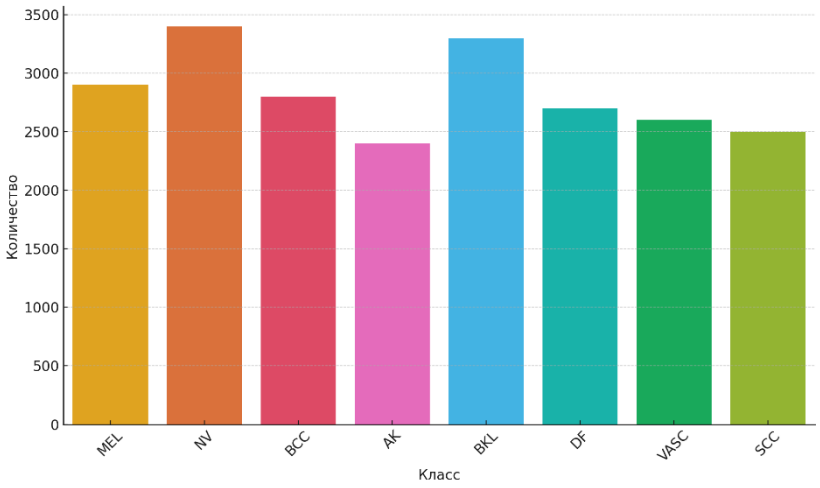


Рис. 3. Распределение классов после преобразований

Для балансировки набора данных случайным способом удалено 1 400 изображений для класса «меланома», 9 000 изображений для класса «невус», 900 изображений для класса «базальноклеточная карцинома» и 900 изображений для класса «доброкачественный кератоз». Полученные изображения объединены в новый набор данных, который составил 11 131 изображение. Затем для увеличения количества представителей изображений других классов была использована аугментация данных: поворот на случайный угол, отражение по горизонтали и по вертикали, что позволило добавить в набор еще 10 200 изображений.

IV. Использование сверточных нейронных сетей для распознавания новообразований

В настоящей работе для исследования эффективности применения сверточных нейронных сетей для классификации дерматоскопических изображений использованы следующие модели нейронных сетей: Regnet_y_3_2gf, Resmlp_12_224, ResNeSt50, ResNet18, ResNet50, Resnext_50_32x4d, Swinv2_cr_tiny_ns_224, VGG11, VGG19, Vit_b_16, Convnext_tiny, DenseNet121, DenseNet161, Efficientnet_v2_s, Efficientnet_v2_m, Inception V3, Levit 256, MnasNet_1_0, MobileNet V2, Nfnet_l0. Наименование моделей и количество их параметров приведены в Табл. 1.

ТАБЛИЦА 1
МОДЕЛИ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

Модель	**Количество параметров**
Regnet_y_3_2gf	19M
Resmlp_12_224	15M
ResNeSt50	27M
ResNet18	12M
ResNet50	26M
Resnext_50_32x4d	25M
Swinv2_cr_tiny_ns_224	28M
VGG11	133M
VGG19	144M
Vit_b_16	86M
Convnext_tiny	28M
DenseNet121	8M
DenseNet161	29M
Efficientnet_v2_s	24M
Efficientnet_v2_m	55M
Inception V3	24M
Levit 256	19M
MnasNet_1_0	4M
MobileNet V2	3.5M
Nfnet_l0	32.8M

Все перечисленные модели были заранее предобучены, использовался оптимизатор Adam со скоростью обучения 0.001, в качестве функции потерь выбрана категориальная кросс-энтропия. Модели обучались максимально в течении ста эпох, а для борьбы с переобучением использовалась техника ранней остановки путем мониторинга производительности модели на валидационном наборе данных во время обучения в течении десяти эпох. В ходе экспериментов было опробовано 17 различных архитектур свёрточных нейронных сетей и обучено 21 модель на одних и тех же данных. В Табл. 2 показаны результаты тестирования обученных моделей на тестовых данных.

ТАБЛИЦА 2
СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МОДЕЛЕЙ

Модель	**Количество эпох**	**ROC-AUC**	**accuracy**	**precision**	**recall**	**F**
Regnet_y_3_2gf	35	0.837	0.81	0.672	0.655	0.64
Resmlp_12_224	15	0.696	0.5	0.28	0.29	0.28
Resnest50	39	**0.859**	**0.868**	**0.71**	**0.71**	**0.69**
Resnet18	28	0.8	0.722	0.56	0.56	0.53
Resnet50	18	0.81	0.735	0.57	0.571	0.55
Resnext_50_32x4d	26	0.808	0.704	0.57	0.575	0.54
Swinv2_cr_tiny_ns_224	21	0.777	0.662	0.485	0.5	0.466
Vgg11	25	0.728	0.613	0.404	0.405	0.367
Vgg19	13	0.667	0.5	0.233	0.257	0.226
Vit_b_16	72	0.788	0.719	0.54	0.544	0.52
Convnext_tiny	52	0.8	0.545	0.584	0.572	0.553
Densenet121	31	0.83	0.8	0.65	0.63	0.615
Densenet161	34	0.827	0.783	0.623	0.636	0.607
Efficientnet_v2_s	42	**0.856**	**0.873**	**0.74**	**0.748**	**0.726**
Efficientnet_v2_m	13	0.838	0.806	0.676	0.654	0.647

Inception V3	41	**0.837**	**0.816**	**0.667**	**0.656**	**0.637**
Levit 256	28	0.83	0.773	0.626	0.619	0.6
MnasNet_1_0	13	0.654	0.427	0.15	0.18	0.11
MobileNet V2	31	0.836	0.79	0.655	0.675	0.64
Nfnet_l0	16	0.838	0.79	0.626	0.627	0.6

В качестве метрик для оценки качестве работы моделей использованы классические метрики, традиционно используемые в задачах компьютерного зрения: ROC-AUC, accuracy, precision, recall и F-мера, Метрики рассчитаны для каждого класса и затем из этих данных построено усредненное значение метрик по всем классам, По данным тестирования моделей на тестовых данных лучшие результаты для классификации новообразований показали модели ResNest50, Efficientnet_v2_s, Inception V3, при этом лучшее качество работы показала модель EfficientNet_v2_s, Для повышения значений точности классификации и значения F-меры в работе выполнен подбор гиперпараметров модели EfficientNet_v2_s согласно Табл, 3,

ТАБЛИЦА 3
ГИПЕРПАРАМЕТРЫ И ДИАПАЗОН ИХ ВАРЬИРОВАНИЯ

Гиперпараметр	**Варьируемые значения**
Тип регуляризации	L1, L2
Количество нейронов на плотных слоях	32, 64, 96, 128, 160, 192
Функция на слое субдискретизации	Среднее значение, максимальное значение
Коэффициент дропаута	0, 0,1, 0,2, 0,4
Коэффициент обучения	10-2, 10-3, 10-4, 10-5, 10-6
Оптимизатор	SGDM, RMSProp, Adam
Размер пакетов при обучении	32, 64, 128, 256, 512

Для каждого набора гиперпараметров создана и обучена отдельная модель сверточной нейронной сети, В результате тестирования, для каждой модели рассчитаны метрики качества, В результате сравнения метрик полученных модификаций моделей выявлен наиболее эффективный набор значений гиперпараметров модели, а именно:

- L1 регуляризатор;
- по 64 нейрона на «плотных» слоях;
- функция взятия максимального значения на слоях субдискретизации;
- 0,2 дропаут;
- коэффициент обучения – 10-4;
- Adam оптимизатор;
- 32 семпла в каждом пакете при обучении,

В Табл, 4 приведены результаты тестирования модели на тестовых данных с подобранными гиперпараметрами,

ТАБЛИЦА 4
МЕТРИКИ РАБОТЫ МОДЕЛИ EFFICIENTNET_V2_S

Модель	**ROC-AUC**	**accuracy**	**precision**	**recall**	**F**
Efficientnet_v2_s	0,923	0,917	0,833	0,786	0,796

Таким образом, проведя расширенный подбор гиперпараметров, удалось существенно повысить значения метрик относительно базовых настроек, так значение accuracy выросло на 0,060, значение precision увеличилось с 0,740 до 0,833, метрики полноса и F-мера также показали рост. Использование модели EfficientNet_v2_s с

выбранными гиперпараметрами позволяет эффективно классифицировать новообразования кожи, и обнаруживать злокачественные заболевания кожи на ранних стадиях.

V. Заключение

В данной статье проведено сравнение эффективности ряда популярных архитектур сверточных нейронных сетей в задаче распознавания кожных новообразований на основании анализа цифрового снимка. Проведены экспериментальные исследования, наглядно демонстрирующие качество работы моделей. Установлено, что лучшая эффективность классификации поражений кожи достигается при применении модели Efficientnet_v2_s, для этой модели сверточной нейронной сети выполнен подбор гирерпараметров. Полученные гиперпараметры значительно увеличивают качество работы модели для задачи классификации новообразований кожи по данным дерматоскопии.

Список литературы

1. Литвиненко Владимир Михайлович Рак кожи причины возникновения, классификация, способы лечения и профилактики // Наука, техника и образование. 2014. №1 (1). [Электронный ресурс], URL: – https://cyberleninka.ru/article/n/rak-kozhi-prichiny-vozniknoveniya-klassifikatsiya-sposoby-lecheniya-i-profilaktiki.
2. Штиршнайдер Ю, Ю, и др, Современные неинвазивные технологии визуализации в дерматологии //Вестник дерматологии и венерологии, – 2011, – №, 5, – С, 41-52
3. American Academy of Dermatology Association, How to prevent skin cancer Association [Электронный ресурс], URL: – https://www.aad.org/public/diseases/skin-cancer/prevent/how
4. *Abbes, W, Deep* Neural Networks for Melanoma Detection from Optical Standard Images using Transfer Learning */ W, Abbes, D, Sellami //* Procedia Computer Science, – 2021, – p,1304–1312
5. Enhancement in Skin Cancer Detection using Image Super Resolution and Convolutional Neural Network */ A, Lembhe, P, Motarwar, R, Patil, S, Elias* // Procedia Computer Science, – 2023,– p,164–173
6. Hybrid convolutional neural networks with SVM classifier for classification of skin cancer / D, Keerthana, V, Venugopal, M,K, Nath, M, Mishra // Biomedical Engineering Advances, – 2023
7. Brinker, T,J, Deep neural networks are superior to dermatologists in melanoma image classification / T,J, Brinker, A, Hekler, A,H, Enk // European journal of cancer, – 2019, – p,11–17
8. Indraswari, R, Melanoma image classification based on MobileNetV2 network / R, Indraswari, R, Rokhana, W, Herulambang // Procedia Computer Science, – 2022, – T,197, – C,198–207
9. ISIC Challenge Datasets // ISICChallenge [Электронный ресурс], URL: – https://challenge.isic-archive.com/data/#2019

УДК 004.89, 611.611
DOI: 10.25206/978-5-8149-3873-2-2024-636-640

Исследование влияния операций предварительной обработки изображений на точность сегментации с помощью сверточных нейронных сетей
Study of the Influence of Image Pre-processing Operations on Segmentation Accuracy Using Convolutional Neural Networks

К. О. Иванов1, А. А. Роженцов1, В. Н. Дубровин2
1Поволжский государственный технологический университет, Йошкар-Ола, Россия
2Республиканская клиническая больница, Йошкар-Ола, Россия

K. O. Ivanov1, A. A. Rozhencov1, V. N. Dubrovin2
[1]Volga State University of Technology, Yoshkar-Ola, Russia
[2]Republican Clinical Hospital, Yoshkar-Ola, Russia

Аннотация. Проведено исследование влияния различных методов предварительной обработки изображений на точность сегментации почечной ткани на медицинских изображениях, полученных с помощью компьютерного томографа. Для этого поставлен ряд экспериментов по обучению сверточной нейронной сети U-Net сегментации ткани почки, которые отличались друг от друга комбинацией операций предварительной обработки изображений, среди которых можно отметить двухуровневое ограничение яркости пикселов, рандомизацию значений яркости фоновых пикселов, спецификацию гистограммы и изменение динамического диапазона изображения с помощью минимаксной нормализации. Установлена квазиоптимальная последовательность операций предварительной обработки изображений, позволяющая повысить точность сегментации изображений почки с помощью сверточной нейронной сети U-Net и стабильность её обучения.

***Ключевые слова*:** глубокое обучение, компьютерная томография, сверточная нейронная сеть, сегментация изображений

Аннотация. A study was conducted of the influence of various image pre-processing methods on the accuracy of renal tissue segmentation in medical images obtained using a computed tomography. To do this, a number of experiments were carried out to train the U-Net convolutional neural network for kidney tissue segmentation, which differed from each other in the combination of image pre-processing operations, among which we can note a two-level limitation of pixel brightness, randomization of background pixel brightness values, histogram specification and changing the dynamic range of the image by minimax normalization. A quasi-optimal sequence of image preprocessing operations has been established to improve the accuracy of kidney image segmentation using the U-Net convolutional neural network and the stability of its training.

***Ключевые слова*:** deep learning, computed tomography, convolutional neural network, image segmentation.

I. Введение

Точная сегментация тканей почки имеет важное значение для построения их трехмерных моделей в системах интраоперационной навигации. Данные системы позволяют совмещать виртуальные трехмерные модели органов с телом пациента. Это повышает информационное обеспечение врача при проведении малоинвазивных операций, делает малоинвазивную операцию более безопасной. Кроме того, данные системы позволяют выбирать на теле пациента точку прокола, которая расположена как можно ближе к области хирургического воздействия. Это позволяет сократить период послеоперационного восстановления [1, 2].

Еще одним важным практическим приложением, в котором могут использоваться результаты точной сегментации тканей почки, является ранняя диагностика онкологических новообразований. Детализированная трехмерная модель почки с ее анатомическими структурами позволит избавить врача от рутинного просмотра каждого КТ-снимка забрюшинного пространства. Кроме того, использование методов сегментации тканей почки в перспективе позволит автоматизировать процесс поиска онкологических новообразований [3].

Однако задача сегментации тканей почки является сложно формализуемой из-за вариативности взаимного расположения и структуры органов брюшной полости, малой контрастности тканей и слабой дифференциации злокачественных и здоровых тканей. В подавляющем большинстве существующих медицинских информационных систем трехмерные модели анатомических структур почки формируются в полуавтоматическом режиме с использованием ручной разметки тканей. Перспективным подходом для решения

научной проблемы сегментации тканей почки является использование методов глубокого обучения, поскольку в настоящий момент именно сверточные нейронные сети показывают наилучшие результаты сегментации объектов на изображении по сравнению с другими методами машинного обучения. При этом требуется решить ряд задач, связанных с повышением точности сегментации, которая зависит от выбранной архитектуры нейронной сети и ее параметров, а также от операций предварительной обработки КТ-изображений [4, 5].

II. Постановка задачи

Целью данной работы является исследование зависимости точности сегментации тканей почки с помощью сверточной нейронной сети от различных комбинаций операций предварительной обработки изображений. При этом в качестве анализируемых изображений выступают послойный снимки, полученные с помощью компьютерного томографа, а среди многообразия архитектур сверточных нейронных сетей выбрана U-Net. Данная нейронная сеть является популярной архитектурой для сегментации медицинских изображений, благодаря своей способности точно определять границы объектов на изображениях [6]. В качестве обучающих данных использовались КТ-изображения базы данных KiTS (Kidney Tumor Segmentation), расположенной в открытом доступе в сети интернет [7]. Данная база данных включает около 3000 томографических изображений с разрешением 512 на 512 пикселов. Эти изображения получены от трехсот пациентов с раком почки во время артериальной фазы КТ-обследования. При этом каждое КТ-изображение сопровождается изображением маски. Области почечной ткани отмечены на маске пикселами со значением яркости 128, областям злокачественного образования соответствуют пикселы со значением яркости 255, остальным тканям – пикселы, значения яркости которых равны нулю (рис. 1). Для проведения исследований из данной базы данных нами было выбрано 800 уникальных изображений, содержащих области тканей почки. Шестьсот из них использовались в качестве обучающих, 100 – в качестве проверочных и еще 100 в качестве тестовых данных.

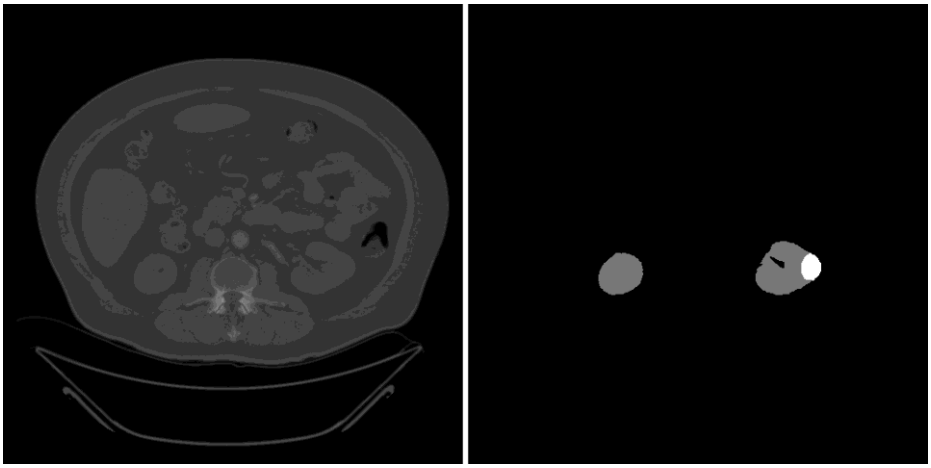


Рис. 1. КТ-изображение и его бинарная маска, взятые из набора данных KiTS (Kidney Tumor Segmentation)

В условиях ограниченного количества обучающих данных нами не наблюдался процесс обучения сверточной нейронной сети U-Net сегментации тканей почки. Однако в случае, когда каждое изображение, поступающее на вход сверточной нейронной сети, подвергалось двухуровневому ограничению яркости с применением операции маскирования, рандомизации величин яркости фоновых пикселов, спецификации гистограммы и минимаксной нормализации, происходил процесс обучения. Переобучение нейронной сети наблюдалось после 80 эпох обучения. При этом удалось достигнуть приемлемого качества сегментации тканей почки без использования топовых версий современных тензорных процессоров (в нашем случае для проведения экспериментов использовалась видеокарта NVidia RTX 4060).

III. Теория

Изображения, полученные с помощью медицинского компьютерного томографа, характеризуются диапазоном разброса значений яркости пикселов в пределах от -2047 до 2048. В качестве значений яркости пикселов выступают относительные величины плотностей тканей по шкале Хаунсфилда, определяющей степень поглощения рентгеновского излучения исследуемым объектом. Использование относительных величин позволяет устранить различия в характеристиках изображений, полученных от различных КТ-томографов. При

этом тканям почки на КТ-изображениям соответствуют только те пикселы, значения яркости которых лежат в интервале от -200 до 500. Эти же пикселы могут соответствовать другим тканям забрюшинного пространства [4].

Данное обстоятельство затрудняет процесс обучения сверточной нейронной сети сегментации тканей почки, поскольку большая часть поступающей на ее вход информации соответствует тканям других органов забрюшинного пространства. Поэтому логично использовать двухуровневое ограничение по яркости пикселов, чтобы исключить значения за пределами диапазона [–200, 500] на этапе предварительной обработки изображений. Кроме того, при пороговой обработке можно использовать операцию маскирования. Данная операция позволяет исключить из последующего анализа те пикселы, значения яркостей которых не соответствуют тканям почки. Значения яркостей этих пикселов могут быть заменены на равномерно распределенные случайные величины, взятые из интервала [-1, 1]. Подобный подход использовался в работе [8] для деперсонализации медицинских изображений и позволяет улучшить качество обучения нейронной сети.

Области тканей почки на различных КТ-изображениях характеризуются разным видом распределения яркостей пикселов, что также затрудняет разработку алгоритмов их сегментации. Чтобы привести эти распределения к одному виду может использоваться процедура стандартизации гистограммы [9]. Для этого из обучающей выборки необходимо случайным образом выбрать одно референсное изображение и построить для него гистограмму яркостей. Далее на остальных изображениях изменить значения яркостей пикселов ткани почки с использованием процедуры стандартизации гистограммы и ранее построенной гистограммы яркости референсного изображения.

На завершающем этапе, в независимости от используемых операций предварительной обработки изображений, динамический диапазон данных, которые поступают на вход нейронной сети, должны быть приведен к интервалу [-1, 1] для исключения проблемы «взрывающихся» градиентов [10]. Для этих целей используется процедура минимаксной нормализации.

IV. Результаты экспериментов

Для исследования зависимости точности сегментации тканей почки с помощью сверточной нейронной сети U-Net от различных комбинаций операций предварительной обработки изображений было проведено пять экспериментов. При этом использовалась стандартная архитектура сверточной нейронной сети U-Net, подробное описание которой представлено в работе [6]. На рис. 2. представлены графики обучения сверточной нейронной сети при использовании различных комбинаций операций предварительной обработки изображений.

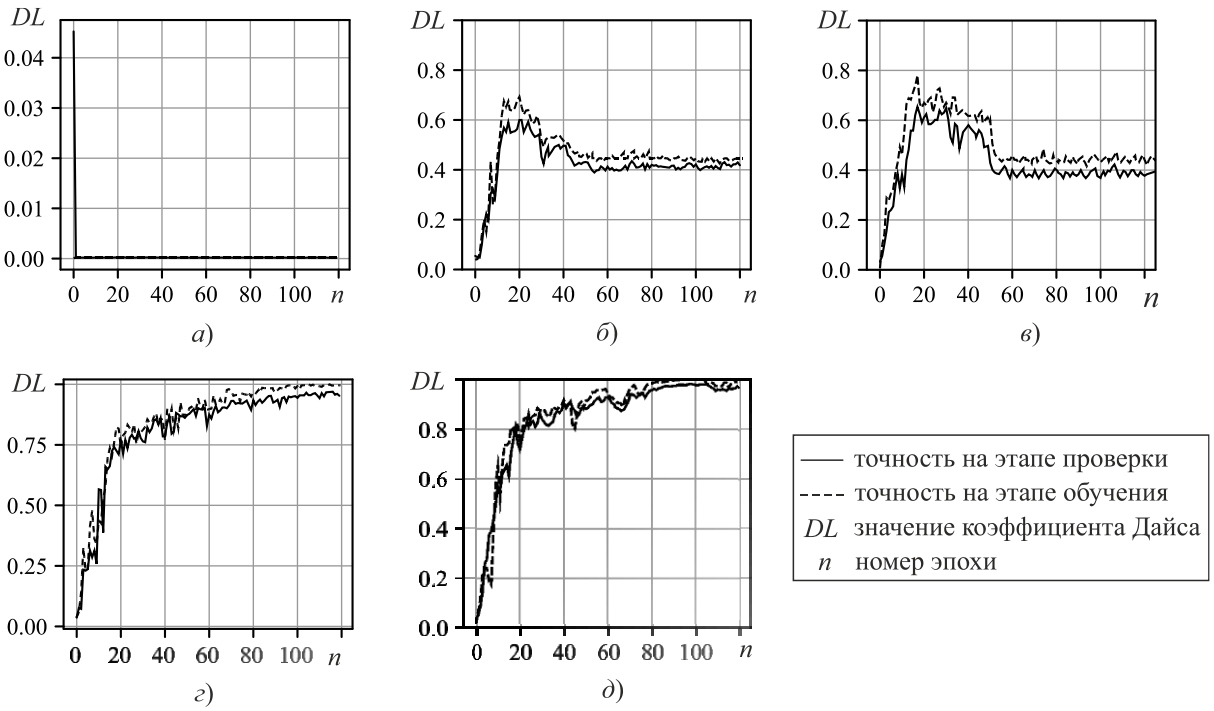


Рис. 2. Графики обучения сверточной нейронной сети в зависимости от комбинаций операций предварительной обработки КТ-изображений: *а*) первый эксперимент; *б*) второй эксперимент; *в*) третий эксперимент; *г*) четвертый эксперимент; *д*) пятый эксперимент

К изображениям, подаваемым на вход нейронной сети в первом эксперименте, применялось лишь изменение значений величин яркостей пикселов к интервалу [-1, 1] при помощи минимаксной нормализации. При этом не наблюдался процесс обучения сверточной нейронной сети, что видно из рисунка 2*а*.

Изображения, подаваемые на вход сверточной нейронной сети во втором эксперименте, предварительно подвергались двухуровневому ограничению яркости и приведению величин яркостей пикселов к интервалу [-1, 1]. В данном случае уже наблюдался процесс обучения сверточной нейронной сети сегментации тканей почки, однако точность сегментации, выраженная в значениях коэффициента Дайса, достигла лишь 60 % на проверочных данных. Переобучение нейронной сети наступало после 20 эпох.

Изображения, подаваемые на вход сверточной нейронной сети в третьем эксперименте, предварительно подвергались двухуровневому ограничению яркости, спецификации гистограммы, приведению значений величин яркостей пикселов к интервалу [-1, 1] при помощи минимаксной нормализации. Точность сегментации тканей почки в данном случае достигла 65 %, при этом переобучение сверточной нейронной сети также наступало после 20 эпох обучения.

К изображениям, подаваемым на вход сверточной нейронной сети в четвертом эксперименте, применялись операции предварительной обработки, аналогичные третьему эксперименту. Однако в отличие от предыдущего эксперимента использовалось двухуровневое ограничение яркости с маскированием. Это означает, что пикселы, значения яркостей которых не принадлежат интервалу [-200, 500], маскируются и не участвуют в процедуре стандартизации гистограммы. Точность сегментации тканей почки при этом возросла до 90 %.

Отличие пятого эксперимента от четвертого заключалось в том, что после двухуровневой пороговой обработки немаскированным пикселям присваивались случайные значения яркостей, равномерно распределенные в диапазоне от -1 до 1. При этом были получены аналогичные результаты точности сегментации тканей почки, как и в предыдущем эксперименте. Тем не менее в результате многократного запуска циклов обучения было установлено, что рандомизация фоновых пикселов в данном случае способствует стабильности обучения сверточной нейронной сети. Так, в четвертом эксперименте успешное обучение сверточной нейронной сети наблюдалось лишь в 50 % случаев, а в пятом уже в 90 % случаев. Таким образом, общая схема алгоритма сегментации тканей почки на КТ-изображениях с помощью сверточной нейронной сети U-Net при ограниченном количестве обучающих данных может иметь структуру, представленную на рис. 3.

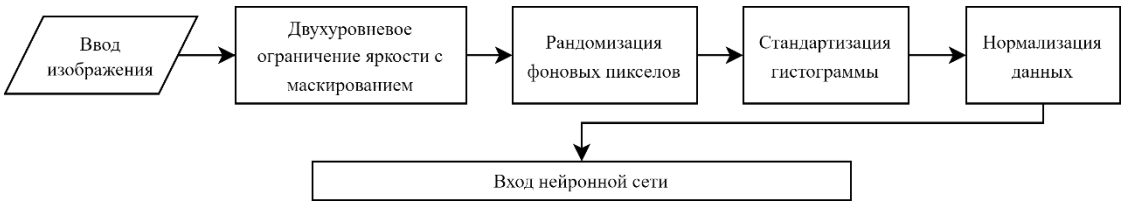


Рис. 3. Структура алгоритма сегментации тканей почки на КТ-изображениях при помощи сверточной нейронной сети U-Net, включающего операции предварительной обработки изображений

V. Обсуждение результатов

В представленной работе выполнено исследование влияния различных методов предварительной обработки изображений на точность сегментации почечной ткани на медицинских изображениях, полученных с помощью компьютерного томографа. Результаты показали, что предварительная обработка изображений играет ключевую роль в успешном обучении и точности сегментации. Важным аспектом работы стало использование операции маскирования областей изображения и присваивание случайных значений пикселам, выходящим за пределы интересующего диапазона яркости. Этот метод, аналогичный использованному для деперсонализации медицинских изображений при объемном рендеринге, повысил стабильность обучения нейронной сети. Маскирование и случайные значения помогли снизить влияние нерелевантной информации, способствуя лучшей концентрации модели на сегментации почечной ткани.

VI. Выводы и заключение

Проведенные эксперименты показали, что правильная предварительная обработка изображений значительно влияет на качество сегментации почечной ткани с использованием сверточной нейронной сети архитектуры U-Net. Среди рассмотренных комбинаций операций предварительной обработки изображений наиболее эффективным подходом оказалось последовательное использование двухуровневого ограничения по яркости с маскированием, рандомизации фоновых пикселов, спецификации гистограммы и минимаксной нормализации. Этот комплексный метод позволяет нейронной сети успешно обучаться и точно сегментировать почечную ткань,

минимизируя влияние нерелевантных данных. Дальнейшие исследования могут быть направлены на оптимизацию данных операций и исследование их применения к другим типам медицинских изображений.

Источник финансирования. Благодарности

Работа выполнена при финансовой поддержке Российского научного фонда, грант «Методы и интеллектуальная система поддержки динамической устойчивости операторов эргатических систем» № 23-19-00568.

Список литературы

1. The application of extended reality technology-assisted intraoperative navigation in orthopedic surgery / B. Dongxian, L. Zhipeng, L. Hao, Z. Qunjie, W. Kaifeng, T. Xiaodong, Z. Jie // Frontiers in Surgery. – 2024. – Vol. 11. – P. 1-15.

2. Optoelectronic navigation system for a flexible surgical instrument based on inertial microelectromechanical sensors / A.A. Rozhentsov, A.A. Baev, M. Halimov, N.N. Mitrakova // BULLETIN OF THE RUSSIAN ACADEMY OF SCIENCES: PHYSICS. – 2021. – Vol. 85. – No. 12. – P. 1434-1438.

3. S.M. Gujarathi, R. Shrivastava, S. K. Angadi. Survey of Kidney Cancer Analysis Using Machine Learning and Deep Learning Algorithms // Journals of Electrical Systems. – 2024. – Vol. 20. – No. 6s. – P. 2491-2501.

4. Kidney segmentation from computed tomography images using deep neural network / L.B. da Cruz, J.D. Lima Araújo, J. L. Ferreira, J. O. Bandeira Diniz, A. C. Silva, J. D. Sousa de Almeida, A. C. de Paiva, M. Gattass // Computers in Biology and Medicine. – 2020. – Vol. 123. – P. 103906.

5. Алгоритм сегментации тканей почки на КТ-изображениях на базе сверточной нейронной сети U-Net / К.О. Иванов, А.В. Казаринов, А.А. Роженцов, В.Н. Дубровин, А.А. Баев, А.О. Евдокимов // Медицинская техника. – №6. – С. 37-39. – 2022 г.

6. Ronneberger, O., Fischer, P., Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation // 18th International Conference Medical Image Computing and Computer-Assisted Intervention. MICCAI. 2015. Munich, Germany. P. 234-241.

7. The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes / https://github.com/neheller/kits19. Retrieved January 25, 2021.

8. Вимкер Р., Буелов Т., Клиндер Т., Бергтольдт М., Вахтер-Штеле И.. Визуальная деперсонализация массивов медицинских данных для защиты при объемном 3d-рендеринге. Патент RU2017102244A: заявл. 22.06.2015: опубликовано 26.07.2018. Бюл. № 21.

9. Гонсалес, Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М.: Техносфера, 2006. – 1072 с.

10. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвиль. – М.: ДМК-Пресс, 2018. – 652 с.

11. Continuous dice coefficient: a method for evaluating probabilistic segmentations / R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, N. Harel // Computer Vision and Pattern Recognition. – 2019. arXiv preprint arXiv:1906.11031.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-641-655

Цветовые преобразования видеоэндоскопических изображений
Color transformations of videoendoscopic images

Н. А. Обухова, А. А. Мотыко, А.А. Поздеев, О.Р. Савельева, С. Ян
СПбГУ, СПбГЭТУ «ЛЭТИ», Санкт-Петербург, Россия

N. A. Obukhova, A. A. Motyko, A.A. Pozdeev, O.R. Savelieva, Xin Yang
SPbU, ETU "LETI", Saint Petersburg, Russia

Аннотация. В работе рассмотрены основные процедуры цифровой обработки эндоскопических изображений, направленные на повышение их диагностической ценности. Показано, что существующие подходы, решая частные задачи, приводят к общей несогласованности цветовых преобразований в эндоскопической видеосистеме. Предложен новый векторный критерий качества для синтеза цветовых преобразований, учитывающий как точность цветопередачи, так и передачу локального цветового контраста. Рассмотрены примеры синтеза цветовых преобразований для целевых задач обработки эндоскопических изображений на основе предложенного критерия.

***Ключевые слова*:** обработка эндоскопических изображений, цветокоррекция, цветовой контраст, цветовые преобразования

Аннотация. The paper considers the main procedures of digital processing of endoscopic images aimed at increasing their diagnostic value. It is shown that the existing approaches, solving private problems, lead to the general inconsistency of color transformations in the endoscopic video system. A new multi-objective quality criterion for the synthesis of color transformations, taking into account both the accuracy of color reproduction and the transmission of local color contrast, is proposed. Examples of synthesis of color transformations for target tasks of endoscopic image processing using the proposed criterion are considered.

***Ключевые слова*:** endoscopic image processing, color corrections, color contrast, color transformations

I. Введение

Медицинские видеосистемы - широкий класс прикладных телевизионных систем, применяемых в различных областях здравоохранения. К ним относятся системы для видеоэндоскопии, телемедицины и медицинской визуализации. Эндоскопия во всем многообразии является сегодня одним из основных инструментов диагностики внутренних органов человека. Новые возможности прикладного телевидения по формированию, анализу и обработке изображений, а также востребованность в системе здравоохранения обуславливают интенсивное развитие видеоэндоскопических систем, которые сегодня включают в себя также элементы телемедицины и медицинской визуализации.

Видеоэндоскопические комплексы в контексте формирования и обработки изображений представляют собой классические прикладные телевизионные системы. Вне зависимости от области применения (бронхоскопия, гастроскопия, колоноскопия) телевизионная часть эндоскопа включает камеру и осветитель, закрепленные на дистальном конце и видеопроцессор с устройством отображения.

Видеоэндоскопия сопряжена с достаточно сложными условиями получения данных, в числе которых: необходимость использования широкоугольной оптики и точечного источника освещения, приводящая к сильным геометрическим искажениям и неравномерности освещения по полю кадра; маленький размер сенсора, характеризующийся высоким уровнем шума; наличие на изображении одновременно очень ярких (вследствие бликов) и темных участков. Для борьбы с указанными проблемами применяют специальные методы и алгоритмы шумоподавления [1], компенсации неравномерности яркости [2] и локального повышения контраста [3].

Основная цель обработки видеоданных в эндоскопических системах – повышение диагностической ценности снимков, обеспечение наилучшей возможности для врача проводить точную диагностику на базе анализа видеоряда. Это достигается в основном за счет решения следующих задач:

- реализации точной цветопередачи;
- повышения локального контраста, обеспечивающего выраженность особенностей тканей;
- увеличения спектрального разрешения за счет использования мультиспектральной съемки;

- формирования эргономичных изображений удобных для визуального анализа (использования карт патологий, тепловых карт и других инструментов визуализации в псевдоцветах).

Разработке общего подхода к решению указанных задач и примерам реализации посвящена данная работа.

II. Постановка задачи

Большинство видеоэндоскопов позволяют получить изображение в белом свете (рис. 1).

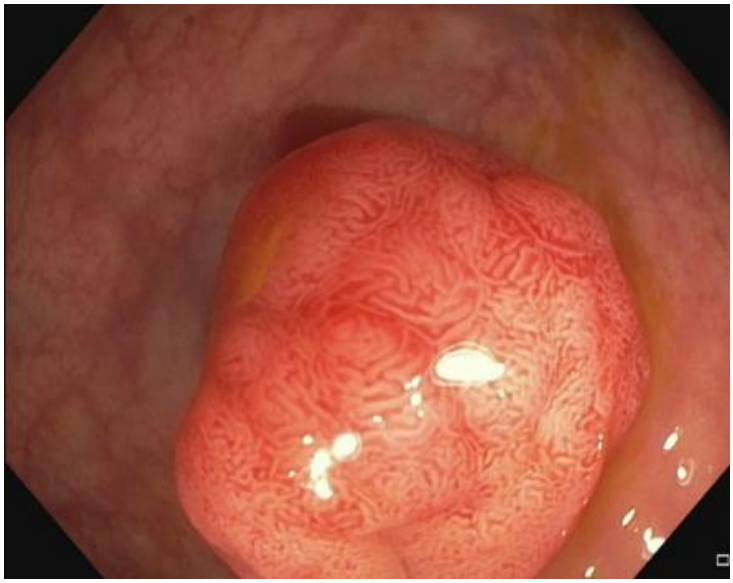


Рис. 1. Эндоскопическое изображение в белом свете

Эндоскопия в белом свете (WLI - white light imaging) имеет свои недостатки при проведении обследования, главный из которых связан с низким спектральным разрешением формируемых изображений и как следствие низким локальном цветовым контрастом, который не позволяет проявить на снимке важные для врача детали. В результате у эндоскопии в белом свете отсутствует ряд диагностических возможностей. В частности, нельзя получить информацию об особенностях, залегающих даже на небольшой глубине в тканях (например, сосудах). Данные с более высокой информативностью позволяют формировать мультимодальные эндоскопы и соответствующие методы обработки изображений. Два самых популярных подхода на современном этапе — это бимодальные эндоскопы и технология хромоэндоскопии.

Технологии цветового контрастирования в эндоскопии относятся к так называемой хромоэндоскопии [4-5]. Оригинальная хромоэндоскопия — это метод исследования, основанный на использовании специальных красящих веществ для улучшения контрастности и визуализации изменений на внутренней поверхности органов при эндоскопическом исследовании. Он позволяет выявлять и оценивать скрытые изменения в тканях, такие как опухоли, полипы, воспаление или дисплазия. Во время хромоэндоскопии красящее вещество наносится на поверхность органа, а затем адсорбируется в определенных тканях, что помогает выделить их и улучшить видимость при проведении обследования. Этот метод позволяет детектировать даже мелкие изменения, которые могут быть незаметны при обычной эндоскопии. Несмотря на свою эффективность, хромоэндоскопия не лишена существенных недостатков: процедура требует использования специальных красящих веществ, которые не всегда доступны; пациенты могут проявлять аллергическую реакцию на используемые красящие вещества; ограниченная глубина проникновения. Как следствие, развитие получили технологии виртуальной хромоэндоскопии, в которой повышение локального цветового контраста достигается без использования красителей. Виртуальную хромоэндоскопию подразделяют на программную и аппаратную. Аппаратная виртуальная хромоэндоскопия (NBI [6], BLI [7], LCI [8]) подразумевает достижение целевых характеристик за счет изменения оптической схемы, манипуляций со спектральным составом освещения, использованием узкоспектральных диапазонов. Программная сопряжена с цифровой обработкой снимков, полученных в белом свете. Иллюстрация хромоэндоскопии и виртуального аналога приведена на рисунке 2.

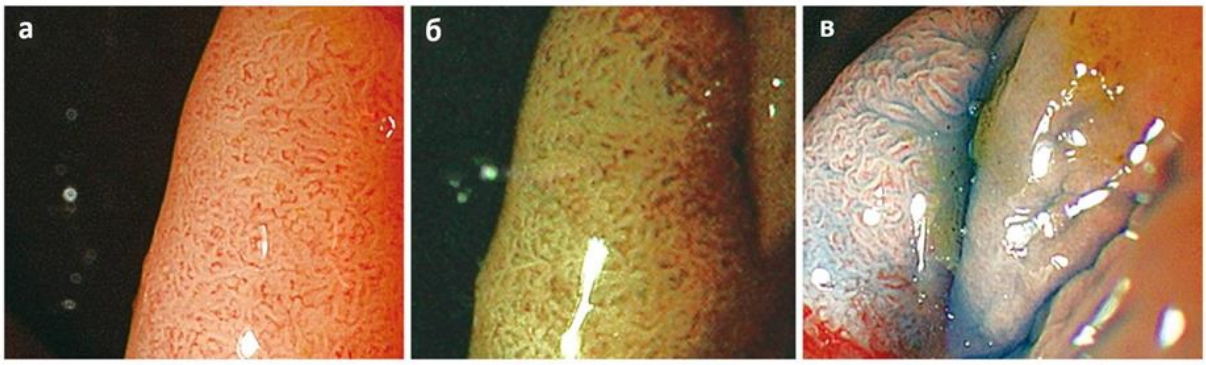


Рис. 2. Аденоматозный полип, а - увеличенное изображение с высоким разрешением в белом свете; б - узкоспектральная визуализация, демонстрирующая структуру капиллярной сети; в - хромоэндоскопия, демонстрирующая структуру капиллярной сети

Исследования применения аппаратной виртуальной хромоэндоскопии подтверждают практическую пользу и важность данной технологии. Однако ее внедрение затруднено существенным усложнением и удорожанием конструктивной части эндоскопа, проблемами, связанными с габаритными размерами системы. Поэтому развитие получают программные методы виртуальной хромоэндоскопии, опирающиеся на обработку и анализ цветовой информации снимков, получаемых от обычных WLI видеоэндоскопов. На сегодня все основные производители эндоскопов оснащают свои продукты функцией программной виртуальной хромоэндоскопии. На рис. 3 приведена иллюстрация работы различных режимов контрастирования видеоэндоскопа Pentax EPK-3000 (Olympus)

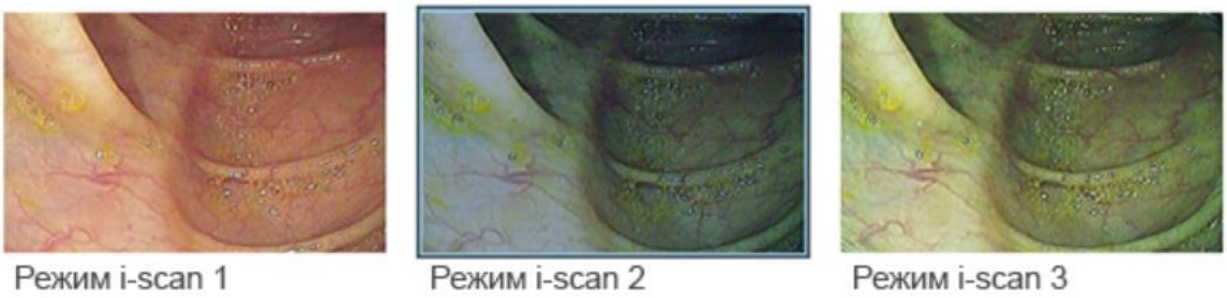


Рис. 3. Режимы цветового контрастирования видеоэндоскопа Pentax EPK-3000 (виртуальная хромоэндоскопия)

К методам повышения спектрального разрешения можно отнести бимодальную эндоскопию, которая, являясь частным случаем мультимодальной, связана с задачами медицинской визуализации, так как формируемые изображения содержат искусственные цвета, практически не связанные с реальным прототипом в сцене. Типовым примером является система, в которой сенсор регистрирует флуоресцентные отклики от освещения различного спектра [9]. В подобных системах повышение диагностической ценности снимков достигается за счет анализа изображений, полученных в различных спектральных диапазонах. Основной задачей в данном случае является организация эффективного комбинирования данных из разных диапазонов и визуализация за счет цветовых преобразований результирующего изображения. Известна реализация бимодальной эндоскопии с применением индоцианина зеленого [10]. В данных эндоскопических видеосистемах помимо цветной видеокамеры присутствует инфракрасный сенсор для регистрации флуоресценции специального фотосенсибилизатора (индоцианина зеленого) в тканях. За счет специальной визуализации, реализующей комбинирование данных из цветного и ИК каналов осуществляется локализация патологии с помощью выделения цветом, см. рис. 4.

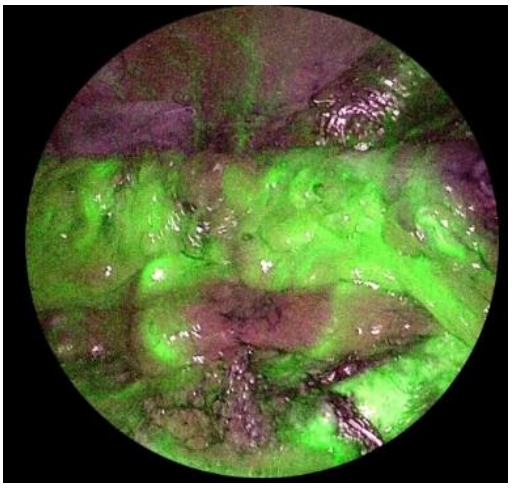


Рис. 4. Изображение, характерное для ICG эндоскопии

Задача специальной визуализации в медицинских видеосистемах также может быть связана с формированием изображений с использованием псевдоцветов (условных цветов). Применение синтетической палитры - один из классических методов визуализации, широко применяемый при отображении снимков, полученных сенсорами в частотном диапазоне, отличным от видимого. Помимо этого псевдоцвета удобны для представления невизуальной информации поверх демонстрируемого изображения. Так, например, их используют в интеллектуальных медицинских видеосистемах при отображении карт патологии (результатов автоматического анализа, рис. 5).

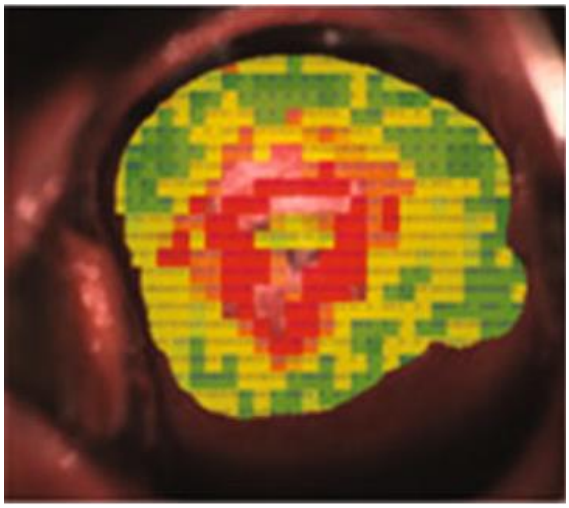


Рис. 5. Карта патологии в псевдоцветах для кольпоскопического изображения

По существу, задача процедур специальной визуализации — это отображение в кадрах видеоряда или статичных снимках объектов интереса или дополнительной информации наилучшим образом с точки зрения восприятия человеком. Очевидный путь для этого - повышение цветового контраста предмета отображения относительно фона - базового контекста изображения. Развитие телемедицины в ряде приложений хирургии формирует повышенные требования к качеству цветопередачи. Врач, проводя операцию дистанционно, должен видеть сцену с минимальными искажениями цветовой и контрастной информации. Это актуально и для видеоэндоскопических систем. В этой связи большое внимание в настоящий момент уделяют непосредственно репродукции цвета.

Таким образом, в видеоэндоскопии целевыми задачами обработки изображений, направленными на повышение диагностической ценности снимков являются:

- цветопередача с сохранением или повышением локального цветового контраста;
- синтез изображения – результата совмещения нескольких спектральных каналов;
- эргономичная визуализация результатов автоматического анализа;

Указанные задачи решаются с помощью цветовых преобразований, обработки и анализа цветовой информации, содержащейся в видеоданных.

Очевидно, что все цветовые преобразования, реализуемые над видеорядом, влияют на цветопередачу, локальный контраст, эргономику и прочие свойства результирующего изображения совместно. При этом в существующих реализациях обработки эндоскопических изображений необходимого совместного учета указанных факторов не производится. Каждая из процедур решает только свою локальную задачу. Известные методы аппаратной и программной эндоскопии (FICE [11], I-Scan [12], Tri-Scan [13]) нацелены на повышение локального цветового контраста, не учитывая при этом точность цветопередачи. В соответствующих открытых работах [13] в расчет принимаются только оценки изменения контраста с помощью различных метрик. Например, Focus value, которая основана на отношении трансформант дискретно-косинусного преобразования (ДКП), соответствующих энергии постоянной составляющей (DC) и энергии остальных составляющих изображения (AC):

$$F_v = \frac{E_{AC}}{E_{DC}},$$
$$E_{DC} = F(0,0)^2,$$
$$E_{AC} = \sum_{x=0}^{M}\sum_{y=0}^{N}(F_{AC}(u,v) - \overline{F_{AC}(u,v)})^2,$$

где *F(u,v)* - трансформанты ДКП *x, y* – координаты на изображении, *M,N* – размер изображения; или метрика для локального контраста из работы [14]:

$$c^{TT} = \frac{1}{MN}\sum_{x=0}^{M}\sum_{y=0}^{N}\frac{R_c(x,y) - R_s(x,y)}{R_c(x,y) + R_s(x,y)},$$

где $R_c(x,y)$, $R_s(x,y)$ - результаты свертки изображения с фильтром Гаусса с различным размером апертуры. Способом оценки контраста основанном не на частотной фильтрации является мера, предложенная Тремо [15], которая учитывает информацию о цвете. Для оценки локального цветового контраста известны метрики на базе структурного тензора Ди Зенцо [16, 17]. Для каждой точки цветного изображения определяется квадратичная форма:

$$\mathbf{C}(x,y) = \mathbf{J}^{\mathrm{T}}\mathbf{G}\mathbf{J},$$
$$\mathbf{J} = \begin{pmatrix} \frac{\partial R}{\partial x} & \frac{\partial R}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \\ \frac{\partial B}{\partial x} & \frac{\partial B}{\partial y} \end{pmatrix},$$

где **J** - матрица частных производных сигналов R, G и B, **G** - отражает зависимость между R, G, B, которая может быть определена, если известны спектральные характеристики камеры, сформировавшей изображение (в общем случае **G** - единичная матрица). Таким образом,

$$\mathbf{C}(x,y) = \begin{pmatrix} C_{xx}(x,y) & C_{xy}(x,y) \\ C_{xy}(x,y) & C_{yy}(x,y) \end{pmatrix},$$

а элементы определяются как:

$$C_{xx}(x,y) = \left|\frac{\partial R}{\partial x}\right|^2 + \left|\frac{\partial G}{\partial x}\right|^2 + \left|\frac{\partial B}{\partial x}\right|^2,$$
$$C_{xy}(x,y) = \frac{\partial R}{\partial x}\frac{\partial R}{\partial y} + \frac{\partial G}{\partial x}\frac{\partial G}{\partial y} + \frac{\partial B}{\partial x}\frac{\partial B}{\partial y},$$
$$C_{yy}(x,y) = \left|\frac{\partial R}{\partial y}\right|^2 + \left|\frac{\partial G}{\partial y}\right|^2 + \left|\frac{\partial B}{\partial y}\right|^2.$$

Собственный вектор

$$\mathbf{v}(x,y) = (v_x(x,y), v_y(x,y))^{\mathrm{T}}$$

соответствующий наибольшему собственному числу $\mu_{max}(x,y)$ матрицы $\mathbf{C}(x,y)$ интерпретируется как направление наибольшего изменения цвета в точке *(x,y)* изображения или, иначе говоря, направлением наибольшего изменения контраста, а само собственное значение отражает выраженность этого изменения.

Построенное таким образом поле векторов для изображения называют псевдоградиентным полем, поскольку в отличии от градиента в данном случае невозможно определить знак вектора - неясно в каком направлении функция возрастает, а в каком убывает. Решению задачи выбора знаков посвящен ряд работ [18, 19]. Псевдоградиентное поле характеризует локальный контраст изображения. Построив два таких поля для сравниваемых изображений (до и после обработки), можно определять изменение контраста в каждой точке *ΔEc(x,y)*, а также, получить некоторую интегральную оценку изменения контраста для изображения:

$$E_{contrast} = \sum_{x=0}^{M}\sum_{y=0}^{N}\Delta E_c(x,y).$$

Таким образом можно оценивать изменение (или наоборот сохранение, в зависимости от задачи) контраста. Применяют и другие глобальные и локальные оценки, не учитывая при этом иные показатели, помимо контраста. В результате получают изображения, подобные приведенным на рис.6.

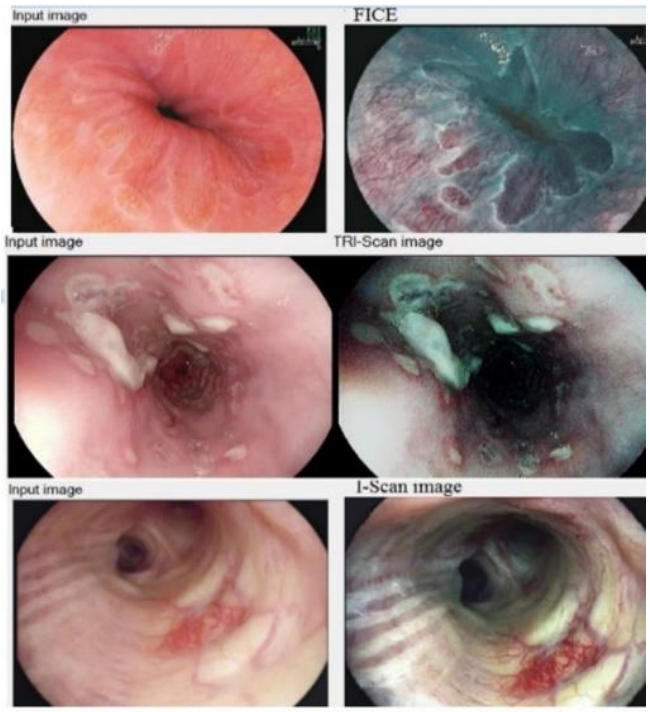


Рис. 6. Результаты обработки известными методами хромоэндоскопии (правый столбец) исходных изображений (левый столбец)

На рис. 6 видно, что несмотря на успешное повышение локального цветового контраста, полученные изображения имеют недостатки, приводящие к снижению эргономичности: значительное смещение цветовой палитры, неравномерность яркости по полю кадра. Причина появления указанных недостатков понятна – данные факторы не были учтены при синтезе цветового преобразования.

Аналогичная проблема наблюдается и у известных алгоритмов медицинской визуализации. Типовым подходом к синтезу объединенного изображения в бимодальной эндоскопии является альфа смешивание, при котором информация из ИК канала отображается зелеными полутонами или голубыми с некоторой прозрачностью. В результате формируется изображение подобное приведенному на рис. 7 [20]

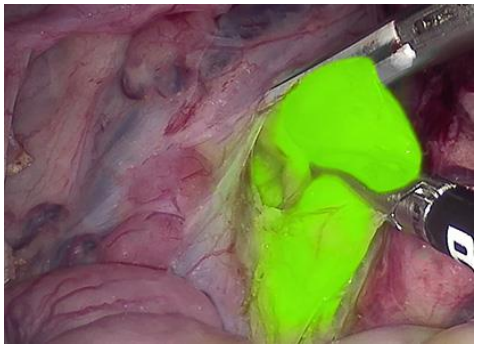


Рис. 7. Результат визуализации одной из систем ICG эндоскопии

Логика выбора цветового тона базируется на максимизации цветовой разницы между подложкой – изображением в белом цвете и информацией из ИК канала. Поставленная цель достигается, однако при этом теряется информация о локальном контрасте подложки, пропадают детали ткани, находящиеся под аппликируемой маской. Такая же проблема возникает и при визуализации карт патологии – результатов анализа изображений: синтез цветового преобразования осуществляется либо с опорой исключительно на цветовую разницу, либо вовсе без какого-либо критерия, на основании эмпирики. Так для процедуры оценки параметров некоторого цветового преобразования в качестве целевого показателя используют среднюю разницу между целевыми значениями цветов и получившимися в результате отображения:

$$E_{color} = \frac{1}{C}\sum_{c=1}^{C} M(l_c - \mathbf{A}(k_c)),\ l_c \in L,\ k_c \in K, \quad (1)$$

где C - число соответствующих друг другу цветов заданной палитры в пространствах прообраза K и образа L, $\mathbf{A}$ – параметры цветового преобразования, M - метрика цветового различия.

Видно, что существующий подход не учитывает локальный контраст, не может адекватно учесть важность тех, или иных цветов изображения. Между тем в силу специфики медицинских снимков различные участки цветовой палитры имеют неодинаковую важность. Так для эндоскопических изображений необходимо точно передавать оттенки красного - цвета тканей, при этом, например, цвета зеленого спектра там не встречаются вовсе - передавать их точно не имеет практического смысла. В существующем подходе, наследуемом от систем общего назначения, можно попытаться учесть важность цветов за счет весов - но возникает трудноразрешимая задача установки их численных значений. В целом, можно сделать вывод, что при синтезе видеоэндоскопической системы видимого диапазона с точки зрения обеспечения точной цветопередачи используют типовой подход для цветной телевизионной системы общего назначения, не учитывающий специфику медицинских систем.

На основании обобщения сформулированных проблем можно сделать вывод, что для синтеза цветовых преобразований в эндоскопических телевизионных системах нужен новый критерий качества, обладающий следующими свойствами.

1. Критерий должен учитывать фактор точности цветопередачи.
2. Критерий должен учитывать фактор передачи локального цветового контраста.
3. Критерий должен иметь возможность учета цветового контекста - отсутствие или наличие цветов, важность отдельных цветов.

III. Теория

1. Векторный критерий качества цветовых преобразований

Для учета точности цветопередачи и передачи локального цветового контраста необходимо сформулировать векторный критерий качества цветовых преобразований:

$$\mathbf{E}_{sum} = (\mathbf{E}_{color}(\mathbf{x}), E_{contrast}(\mathbf{x})),$$

где $\mathbf{x}=(x_1,x_2,\ldots,x_n)$, вектор параметров цветового преобразования; $E_{contrast}(\mathbf{x})$ - целевая функций, характеризующая передачу локального цветового контраста; $\mathbf{E}_{color}(\mathbf{x})$ - векторная целевая функция, характеризующая цветопередачу:

$$\mathbf{E}_{color}(\mathbf{x}) = \{ f_{c1}(\mathbf{x}),\ f_{c2}(\mathbf{x}), \ldots,\ f_{cn}(\mathbf{x})\},$$

где $f_{c1}(\mathbf{x})$, $f_{c2}(\mathbf{x})$, …, $f_{cn}(\mathbf{x})$ - целевые функции, определяющие точность передачи отдельных цветов. В данной формализации терм, отвечающий за передачу локального цветового контраста, представляет собой скалярную функцию, характеризующую общую оценку локального контраста для изображения, как принято в современной практике.

Введенный векторный критерий дает возможность найти параметры цветового преобразования путем решения оптимизационной задачи. Значение критерия в точке оптимума при оптимизации по вектору параметров цветового преобразования соответствует:

$$E^*_{sum} = \min_{\mathbf{x}}\{\mathbf{E}_{color}(\mathbf{x}), E_{contrast}(\mathbf{x})\},$$

а значение аргумента (искомого вектора параметров цветового преобразования) в точке оптимума:

$$\mathbf{x}_{opt} = \arg\min_{\mathbf{x}}\{\mathbf{E}_{sum}\}$$

Решение задачи векторной оптимизации может быть сведено к поиску Парето-оптимальных решений, но для задачи синтеза цветовых преобразований указанный подход малопригоден, так как сопряжен с трудноразрешимой задачей выбора одного решения из набора и избыточно усложняет расчет. В технических задачах [21], подобных рассматриваемой, векторная оптимизация проходит процедуру скаляризации, когда

множество показателей качества (критериев оптимальности) сводятся в один. В простейшем случае в скаляризация заключается в представлении целевых функций в виде взвешенной суммы. Таким образом устанавливают:

$$\mathbf{E}_{color}(\mathbf{x}) \sim E_{color}(\mathbf{x}) = \sum_{c=1}^{C} w_c f_c(\mathbf{x}),$$

где C – цвета, входящие в палитру при расчете точности цветопередачи, w_c – вес, характеризующий важность соответствующего цвета в общей оценке целевой функции. Учет контекста таким образом реализуется через выбор набора цветов C, по которым рассчитывается значение цветовой ошибки, и возможности установки различных весов цветов, используемых в расчете. Веса могут быть в общем случае установлены одинаковыми или оценены, исходя из анализа контекста, из статистически оцененной встречаемости на изображениях, из априорно установленной важности участков палитры.

Провести таким же образом скаляризацию для общей задачи оптимизации, определяющей оценку введенного критерия, возможно только при эмпирическом подборе весов w_{color}, $w_{contrast}$, который не всегда очевиден:

$$E^*_{opt} = \min_{\mathbf{x}}\{w_{color}E_{color}(\mathbf{x}) + w_{contrast}E_{color}(\mathbf{x})\}.$$

Для решения этой задачи предлагается применить метод изменения ограничений, при котором одну из целевых функций оставляют в качестве целевой, а остальные превращают в ограничения. То есть задачу можно переформулировать в следующем виде:

$$E^*_{opt} = \min_{\mathbf{x}}\{E_{contrast}(\mathbf{x}) \big| E_{color}(\mathbf{x}) \le \varepsilon\},$$

где ε – некоторый допустимый уровень для соответствующей целевой функции. В случае скаляризованной $E_{color}(\mathbf{x})$ такой уровень определить достаточно просто. В этом качестве можно использовать известные пороги восприятия или приемлемости [22] для некоторой усредненной оценки цветопередачи. Например, при расчете ошибки цветопередачи в единицах CIEDE2000, порог восприятия (ниже которого человек не видит разницу) составляет около 1 – 1.3:

$$E^*_{opt} = \min_{\mathbf{x}}\{E_{contrast}(\mathbf{x}) \left| \frac{1}{C}\sum_{c=1}^{C} w_c f_c(\mathbf{x}) \le 1.0\right\}.$$

На практике при оценке точности цветопередачи часто используют максимальную ошибку, или 95-й перцентиль, то есть:

$$E^*_{opt} = \min_{\mathbf{x}}\{E_{contrast}(\mathbf{x}) \big| E_{color}(\mathbf{x})_{\max} \le \varepsilon \in [1..2]\},$$

$$E^*_{opt} = \min_{\mathbf{x}}\{E_{contrast}(\mathbf{x}) \big| E_{color}(\mathbf{x})_{0.95} \le \varepsilon \in [1..2]\}.$$

Для оценки передачи контраста имеет смысл использовать локальные метрики, учитывающие цветовую информацию. Подходящей реализаций в данном случае являются оценки на базе структурного тензора Ди Зенцо. В качестве способа расчета точности цветопередачи рассмотрим представленный в выражении (1) аддитивный вариант скаляризации $\mathbf{E}_{color}(\mathbf{x})$ с одинаковыми весами и вычислением E_{color} как среднего значения всех составляющих $E_{color}(\mathbf{x})$ для некоторого $\mathbf{x}$.

Таким образом, векторный критерий качества синтеза цветовых преобразований полностью выполняет заданные требования. Критерий учитывает как точность цветопередачи, так и передачу локального цветового контраста (в зависимости от типа задачи может быть сохранение или повышение контраста), обладает возможностью учета контекста изображения и особенностей состава цветовой палитры.

На основе сформулированного векторного критерия были решены целевые задачи обработки видеоэндоскопических изображений. Разработаны метод виртуальной хромоэндоскопии и методы синтеза изображения, объединяющего изображение, полученное в белом свете и узкоспектральное изображение для бимодальной эндоскопии.

2. Метод виртуальной хромоэндоскопии

Метод виртуальной хромоэндоскопии, основан на разложении изображений, полученных в белом свете на отдельные RGB-компоненты, их независимой обработке, а затем рекомбинации для создания новых изображений, в которых за счет повышения локального контраста подчеркнут рисунок сосудов и изменения слизистой оболочки. Структурная схема алгоритма приведена на рис. 8.

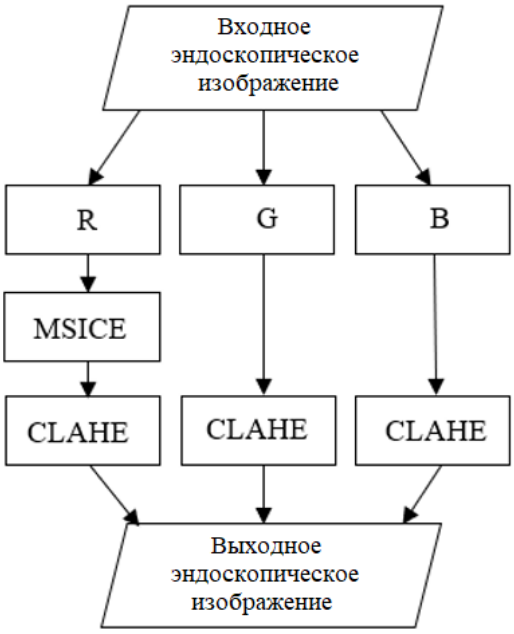


Рис. 8. Структурная схема предложенного алгоритма виртуальной хромоэндоскопии

На первом этапе для улучшения визуализации кровеносных сосудов при обработке красного канала использован нелинейный метод локального повышения контраста – MSICE [23]. На втором этапе производится тоновая коррекция за счет раздельной обработки каждого цветового канала (R, G, B) с помощью алгоритма адаптивной гистограммной эквализации с ограничением контраста (CLAHE) [24].

В рассматриваемой схеме метода большое влияние на результат оказывают параметры эквализации в каждом канале. При установке высоких значений, предлагаемый метод позволяет получить значительное увеличение локального цветового контраста. В то же время контроль над цветопередачей отсутствует. В результате палитра изображения существенно отличается от естественной, что негативно сказывается на эргономике. По этой причине принципиально важным является обоснованный выбор параметров эквализации CLAHE. Необходимо сохранить палитру изображения близкой к исходной, и при этом максимизировать локальный контраст.

Сформулированное требование означает решение следующей оптимизационной задачи на основе введенного векторного критерия. Найти для алгоритма CLAHE набор параметров ограничения контраста (*clip limits*) $\mathbf{l} = (lR, lG, lB)$ для каналов R, G, B, который обеспечил бы максимальное приближение по уровню локального цветового контраста к обработанной части набора данных, сохранив при этом приемлемую разницу *thr* между палитрой и исходной частью набора данных. Для оценки цветового различия использовалось среднее значение меры E_{color} (на базе формулы CIEDE2000 [25]) по всем пикселям изображений. Для расчета расхождения по локальному контрасту $E_{contrast}$ использовалась оценка на основе структурного тензора Ди Зенцо, усредненная по всем пикселям изображений *MN*. Задача была сформулирована как оптимизация с ограничениями:

$$E^*_{opt} = \min_{\mathbf{l}}\{\sum_{MN} E_{contrast} \Big| \sum_{MN} E_{color} < thr\}.$$

Выбор значения *thr* определялся исходя из следующих соображений. Разница по формуле CIEDE2000 меньше *1* находится ниже так называемого порога различимости, при котором большинство людей не замечают разницы между цветами. Поэтому логично провести минимизацию так, чтобы средняя разница цветов была ниже этого порога.

3. Методы визуализации изображений, полученных при бимодальной эндоскопии.

Цель визуализации - сформировать изображение, объединяющее цветное изображение, получаемое в белом свете и монохромный канал, регистрирующий флуоресценцию фотосенсибилизатора. На рис. 9 приведены примеры исходных изображений.

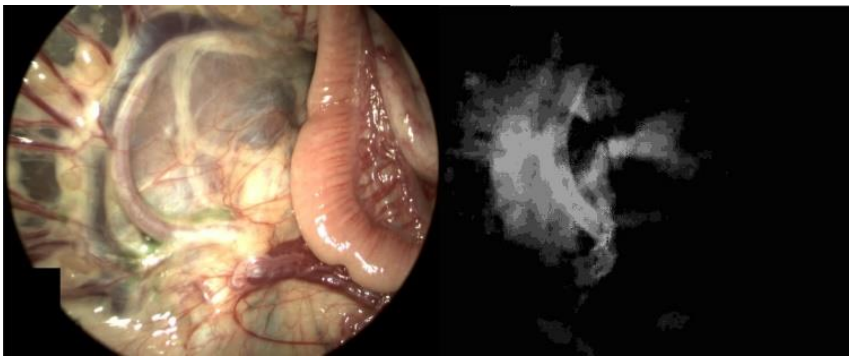


Рис. 9. Изображения, формируемые бимодальным эндоскопом

Предложено два способа формирования результирующего изображения. В первом способе синтезированное изображение получают аппликативным наложением области патологии, отображаемой одним тоном. Во-втором, область патологии отображают псевдоцветным изображением. Особенность сформированного псевдоцветного изображения области поражения следующая: цветовой контраст между областями разного цвета на псевдоцветном изображении соответствует яркостному контрасту областей с разным уровнем флуоресценции (яркости) на изображении, полученном в ИК диапазоне.

Для реализации первого способа визуализации необходимо найти цветовой тон аппликации для отображения монохромного канала, который соответствовал бы максимуму цветового отличия между подложкой (изображением в белом свете) и аппликацией, одновременно с этим сохраняя цветовой контраст подложки. Основой метода является анализ по предложенному векторному критерию зависимостей результирующих характеристик формируемого изображения от цветового тона, используемого для отображения монохромного канала при альфа-смешивании. Анализируют среднее цветовое различие и оценку сохранения локального контраста подложки (изображение в белом свете) и апплицируемых данных (флуоресцентный канал). Для расчета цветового различия использовалась формула CIEDE2000, оценка сохранения локального контраста выполнялась на базе структурного тензора Ди Зенцо.

Во втором способе при формировании псевдоцветного изображения должны быть выполнены следующие требования: учтены свойства человеческого зрения и исключены потери информации о яркости флуоресценции. При замене области флуоресценции, представленной на узкоспектральном изображении в градациях серого, на псевдоцветное изображение необходимо, чтобы визуальный эффект от яркостного перепада соответствовал визуальному эффекту от заменяющего его цветового перепада на псевдоцветном изображении.

Формирования псевдоцветного изображения области поражения состоит в следующем.

Сегментируют область флуоресценции на узкоспектральном изображении. Признаком для сегментации является яркость. Сегментированную область делят на подобласти равной яркости – выполняют кластеризацию изображения. Для каждого кластера выбирался такой цвет, чтобы цветовое расстояние между областями i и j с цветами $color_i = (R_i, G_i, B_i)$, и $color_j = (R_j, G_j, B_j)$, рассчитанное по метрике CIEDE2000 dC_{ij}, соответствовало вычисленному значению dL_{ij} между этими областями в монохромном изображении. Это предполагает решение оптимизационной задачи, в результате которой находится набор цветов для областей **color**, сформулированной следующим образом:

$$E^*_{opt} = \min_{\mathbf{color}}\{\sum_{i=1}^{K}\sum_{j=1}^{K}\left(dL_{ij} - dC_{ij}\right)^2\}, i \neq j$$

Следует отметить, что в случае больших значений dL_{ij} для определения dC_{ij} использовалась оценка цветового расстояния CIE76.

IV. Результаты экспериментов и обсуждение

1. Экспериментальное исследование метода виртуальной хромоэндоскопии

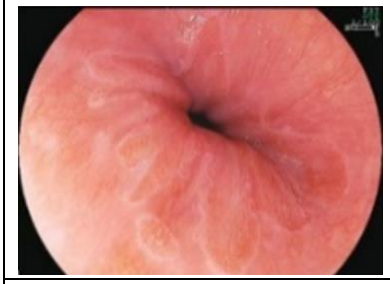
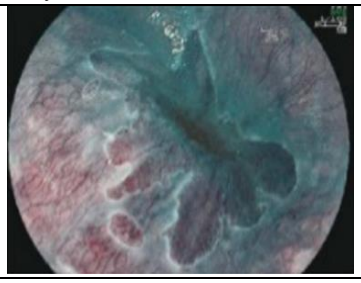
В экспериментах использовалась открытая база данных KVASIR [26], состоящая из 4000 эндоскопических изображений с разным разрешением - от 720x576 до 1920x1072 пикселей. Разнообразие данных в этом наборе позволяет тщательно оценить предложенный алгоритм в различных условиях.

Для определения параметров метода в соответствии с процедурой, описанной в пункте III.2, была создана обучающая выборка изображений, состоящая из двух частей. Набор исходных эндоскопических изображений и набор изображений, полученных на основе предложенного алгоритма с максимизированными параметрами контрастирования. Выборка из сформированных наборов представлена на рис. 10.

Рис. 10. Образцы из обучающего набора данных. Исходные изображения (2 левых столбца) и контрастные изображения (2 правых столбца)

В ходе решения задачи оптимизации были получены следующие оценки для искомых параметров *lR = 2.2, lG = 2.2, lB = 1.2*. Средняя оценка E_{color} по тестовой выборке, состоящей из 100 изображений, составила *1.01*. Это означает, что полученные изображения в среднем практически не выходят за рамки исходной палитры. При этом аналогичная оценка, полученная для этих же изображений при максимизации контрастирования, составила *3.74*, что соответствует легко различимому, выраженному отличию палитры от исходной.

В среднем для изображений с максимизированным контрастом оценка *Fv* оказалась в *2.6* раза выше, чем среднее значение *Fv* для набора оригинальных изображений. На рисунке 9 в правых столбцах заметно явное переконтрастирование. В свою очередь для изображений с рассчитанными параметрами CLAHE средняя оценка *Fv* оказалась в *1.8* раза выше, чем для оригиналов. Как видно на рис. 10, этот результат также соответствует выраженному усилению контраста и практически эквивалентен оценке для известных прототипов *1.6-2.2*. В частности, для рассматриваемой выборки изображений средняя оценка *Fv* после применения TRI-Scan (с рекомендованными авторами метода параметрами) оказалась в 1.65 раз выше, чем для оригиналов, а для подвыборки изображений с известным результатом обработки методом I-Scan увеличение *Fv* достигло 2.2 раза. На рис. 11 показаны примеры исходных изображений (слева), результаты применения конкурирующих методов (в центре) и предлагаемого метода (справа). Нужно отметить, что в рамках рисунка 11 не было возможности показать результат применения всех рассматриваемых методов для каждого исходного изображения. Это связано с тем, что FICE и I-Scan – закрытые технологии, и моделировать их работу вне реального эндоскопического обследования при подаче на вход цифрового изображения невозможно. В данном случае можно лишь воспользоваться опубликованными атласами изображений для данных методов, демонстрирующих пары снимков – исходный и результат применения соответствующей технологии.

<table>
<tr><th>Исходное изображение</th><th>Конкурирующий метод</th><th>Результат предложенного метода</th></tr>
<tr><td rowspan="2">  </td><td>Результат FICE</td><td rowspan="2"> </td></tr>
<tr><td>  </td></tr>
<tr><td></td><td>Результат TRI-Scan</td><td></td></tr>
</table>

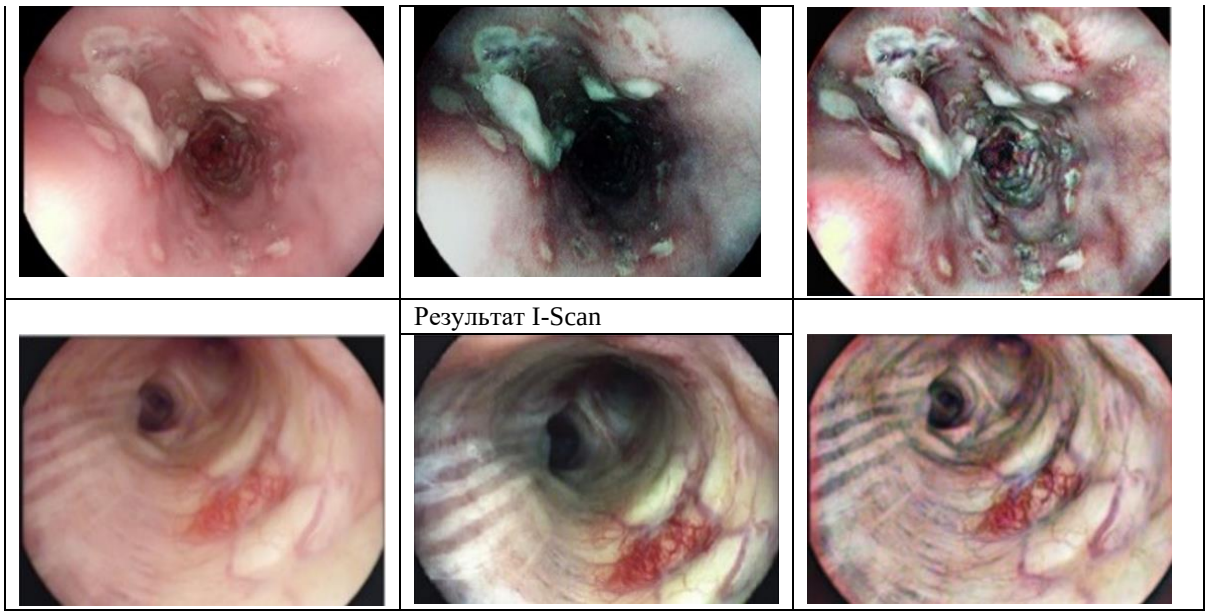


Рис. 11. Результаты предложенного метода в сравнении с конкурентами. Левый столбец – исходные изображения, центральный столбец – результаты конкурирующих методов (FICE, TRI-Scan, I-Scan), правый столбец – результат предложенного метода.

2. Экспериментальное исследование методов визуализации изображений, полученных при бимодальной эндоскопии

Для первого способа визуализации на выборке из 10 типовых изображений был проведен анализ зависимостей результирующих характеристик формируемого изображения от цветового тона, используемого для отображения монохромного канала при альфа-смешивании. На рис. 12 приведены графики E_{color} и $E_{contrast}$ в зависимости от цветового тона в пространстве HSV. Как было указано выше (пункт) для расчета цветового различия использовалась формула CIEDE2000, оценка сохранения локального контраста выполнялась на базе структурного тензора Ди Зенцо.

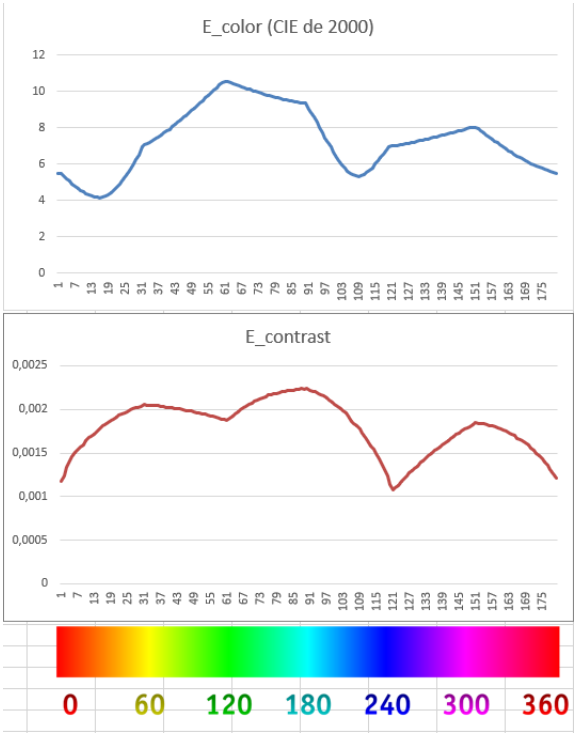


Рис. 12. Графики оценок E_{color} и $E_{contrast}$ в зависимости от цветового тона в пространстве HSV

Исходя из анализа графиков видно, что при выборе любого цветового тона аппликация будет заметной, так как средняя оценка цветового различия относительно подложки достаточно велика и превышает $\Delta E_{00} = 4$, что

соответствует значимой разнице. С точки зрения выбора цветового тона существует два перспективных варианта с координатами *H = 120* и *H = 242*. Третья интересная точка на графиках *H = 0* соответствующая красному цвету, уступает точке *H = 242* как с точки зрения сохранения контраста, так и максимизации цветового различия. Если для целевого назначения синтеза единого изображения (определяемого врачом-эндоскопистом) важнее максимизация цветового расстояния, то имеет смысл выбирать зеленый цвет при альфа-смешивании. Он в наибольшей степени отличается от подложки. Действительно, зеленый цвет далек от желто-красной палитры тканей и от сине-фиолетового цвета венозных сосудов. Если для целевого назначения важно сохранять информацию, содержащуюся в подложке, то для цвета аппликации необходимо выбрать *H = 242*, представляющий оптимальный по предложенному критерию компромисс. Отношение оценок сохранения контраста подложки для *H = 120* и *H = 242* составляет *1.8*. То есть контраст в случае использования цветового тона *H = 242* сохраняется почти в два раза лучше согласно оценке на базе структурного тензора Ди Зенцо. При этом среднее цветовое различие результирующего изображения и снимка в белом свете выше для *H = 120* в *1.5* раза относительно *H = 242*. На рис. 13 приведены результаты формирования единого изображения с использованием указанных значений цветового тона. В областях максимальной яркости апплицируемого канала заметно лучшее сохранение локального контраста для *H = 242* по сравнению с *H = 120*.

Рис. 13. Результаты формирования единого изображения с использованием рассмотренных значений *H*

Результаты визуализации бимодальных изображений вторым способ приведены на рисунке 14. В экспериментальном исследовании второго способом визуализации были использованы реальные лапароскопические изображения и проведена их оценка экспертами – восьмью врачами эндоскопистами. В проведенных экспериментах эксперты сравнивали одновременно предъявленные флуоресцентные изображения и изображения, полученные в белом свете, и синтезированные изображения. Использована шкала [-3,3] для оценки качества изображения и точности локализации поражения. Оценка 3 означает, что синтезированное изображение лучше, а -3 - лучше исходные изображения. Выполненное исследование по методу попарного сравнения показало, что большинство врачей-эндоскопистов предпочли изображение, синтезированное предложенным методом, одновременной визуализации изображений в белом свете и свете флуоресценции. Средняя оценка предпочтения составляет 2,6.

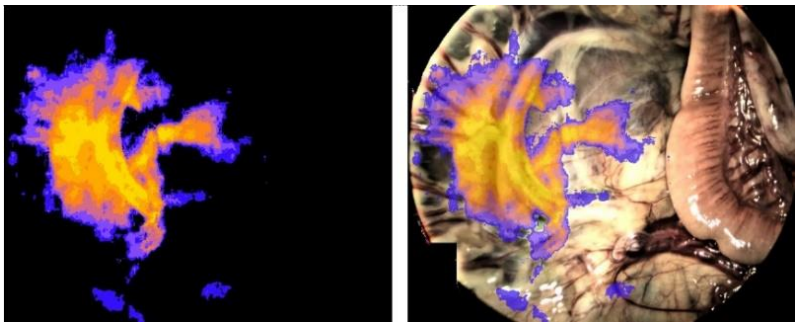


Рис. 14. Изображение, сформированное предложенным алгоритмом, и результат его аппликации

V. Выводы и заключение

В данной работе был представлен векторный критерий для синтеза цветовых преобразований эндоскопических изображений. Критерий учитывает как точность цветопередачи, так и передачу локального цветового контраста (в зависимости от типа задачи может быть сохранение или повышение контраста), обладает возможностью учета контекста изображения и особенностей состава цветовой палитры. Критерий позволяет осуществлять системный подход к синтезу цветовых преобразований эндоскопических изображений, учитывая во взаимосвязи все основные целевые показатели качества целевой обработки изображений, важные для

эндоскопической системы, обеспечивающие получение диагностически денных снимков, а следовательно, повышающие точность диагностики.

На основе предложенного критерия был выполнен синтез алгоритмов обработки эндоскопических снимков. Разработанный алгоритм виртуальной хромоэндоскопии позволяет получить изображения, уровень локального цветового контраста в которых соответствует известным аналогам, при этом алгоритм обеспечивает сохранение исходной цветовой палитры изображения, что делает снимок более удобным для восприятия и анализа.

Разработанный алгоритм синтеза единого изображения для бимодальной эндоскопической системы позволяет получать результат в соответствии с задачей медицинского обследования и назначением системы. Он обеспечивает возможность получения решения с максимизированным цветовым различием аппликации и подложки или обоснованного компромиссного решения с сохранением локального контраста подложки и приемлемым цветовым различием.

Источник финансирования. Благодарности

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках реализации научного проекта по соглашению № 075-15-2024-631

Список литературы

1. N. A. Obukhova, A.A. Motyko, A.A. Pozdeev, B.S. Timofeev, Review of Noise Reduction Methods and Estimation of their Effectiveness for Medical Endoscopic Images Processing, Proceedings of the XXth Conference of Open Innovations Association FRUCT, 2018, т.426, с.204-210.

2. N. A. Obukhova, A.A. Motyko, A.A. Pozdeev, The Endoscopic Images Visualization in Clinical Decision Support Systems, 22th International Conference on Digital Signal Processing and its Applications (DSPA), Moscow, Russia, 2020, pp. 1-4

3. N. A. Obukhova, A.A. Motyko, A.A. Pozdeev, Methods of Endoscopic Images Enhancement and Analysis in CDSS, Intelligent Systems Reference Library, vol. 175, «Computer Vision in Control Systems -5 », Advanced Decisions in Technical and Medical Applications Springer, 2020, p. 225-264

4. East J.E. и др. Advanced endoscopic imaging: European Society of Gastrointestinal Endoscopy (ESGE) Technology Review, Endoscopy. Georg Thieme Verlag, 2016. Т. 48, № 11. с. 1029–1045

5. Al-Kawas F.H. Chromoendoscopy, Gastroenterol. Hepatol. (N. Y). Millenium Medical Publishing, 2006. Т. 2, № 2. с. 101–103

6. Gono K. и др., Appearance of enhanced tissue features in narrow-band endoscopic imaging, J. Biomed. Opt. J Biomed Opt, 2004. т. 9, № 3. с. 568.

7. Togashi K. и др., Blue laser imaging endoscopy system for the early detection and characterization of colorectal lesions: a guide for the endoscopist, Therap. Adv. Gastroenterol. SAGE Publications, 2016. т. 9, № 1. с. 50

8. Shinozaki S. и др., Linked color imaging for the detection of early gastrointestinal neoplasms, Therap. Adv. Gastroenterol. SAGE Publications, 2019. т. 12.

9. J. Kim, H. Al Faruque, S. Kim, E. Kim, J. Hwang, Multimodal endoscopic system based on multispectral and photometric stereo imaging and analysis, Biomed. Opt. Express, 2019, vol.10, pp.2289-2302

10. Web: https://www.karlstorz.com/cps/rde/xbcr/karlstorz_assets/ASSETS/3470896.pdf, дата обращения 13.07.2024

11. FICE Atlas of Spectral Endoscopic images, 2008.
Web: https://www.fujifilmla.com/wp-content/uploads/2021/01/fice-atlas-esp.pdf, дата обращения 13.07.2024

12. PENTAX Medical i-scan Mini- Atlas for Gastroenterology, 2015.
Web: https://www.i-scanimaging.com/fileadmin/user_upload/PENTAX_i-scan_Mini-Atlas.pdf, дата обращения 13.07.2024

13. M.S. Imtiaz, S.K. Mohammed et al. “Tri-Scan: A Three Stage Color Enhancement Tool for Endoscopic Images”, Journal of Medical Systems, vol. 41(6), Jun. 2017, pp. 1-16.

14. Tadmor Y, Tolhurst DJ. Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. Vision Res. 2000, vol. 40(22), pp. 3145-57

15. A. Tremeau, Color contrast parameters for analysing image differences, Colour, Institute, I. (Eds.), Proceedings of the University of Derby Color Image Science, Derby, 2000, pp. 11–23

16. Chatoux H. и др., Gradient in spectral and color images: from the Di Zenzo initial construction to a generic proposition, J. Opt. Soc. Am. A, 2019, v. 36, № 11, p. 154

17. Di Zenzo S. A note on the gradient of a multi-image, Comput. Vision, Graph. Image Process, 1986. v. 33, № 1. pp. 116–125.

18. Socolinsky D. A., Wolff L. B. A, New Visualization Paradigm for Multispectral Imagery And Data Fusion, In proc. of CVPR, 1999, pp. 1319-41.

19. Nikolaev D., Karpenko S., Color-To-Grayscale Image Transformation Preserving The Gradient Structure, In proc. of 20th ECMS, 2006, pp. 427–430

20. Web: https://www.richard-wolf.com/en/disciplines/surgery/system-green, дата обращения 13.07.2024

21. Брахман Т. Р. Многокритериальность и выбор альтернативы в технике. — М.: Радио и связь, 1984. — 287 с.

22. Kim A-Ri, Kim Hong-Suk, Park Seung-ok, Measuring of the Perceptibility and Acceptability in Various Color Quality Measures, Journal of the Optical Society of Korea, 2011, v.15.

23. V. Vonikakis, I. Andreadis, “Multi-Scale Image Contrast Enhancement”, in Proc. 10th IntI. Conf. on Control Automation Robotics and Vision, Dec. 2008, pp. 856-861

24. G. Yadav, S. Maheshwari, A. Agarwal, “Contrast limited adaptive histogram equalization based enhancement for real time video system”, in Proc. ICACCI-2014, Sep. 2014, pp. 2392-2397

25. G. Sharma, W. Wu, and E. N. Dalal, «The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations», Color Res. Appl., y. 30, vol. 1, pp. 21–30, 2005

26. A Multi-Class Image-Dataset for Computer Aided Gastrointestinal Disease Detection. Web: http :// datasets.simula.no, дата обращения 13.07.2024

УДК: 519.6; 004.932.2
DOI: 10.25206/978-5-8149-3873-2-2024-656-668

Методы классификации гистологических изображений на основе сверточных нейронных сетей с фильтром Габора
Histological Image Classification Methods Based on Convolutional Neural Networks with Gabor Filters

Е. А. Мурин[1] , Д. В. Сорокин[1], А. С. Крылов[1]
[1] *Лаборатория математических методов обработки изображений, факультет вычислительной математики и кибернетики, МГУ имени М. В. Ломоносова, Москва, Россия*

Egor A. Murin[1], Dmitry V. Sorokin[1], Andrey S. Krylov[1]
[1] *Laboratory of Mathematical Methods of Image Processing, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia*

Аннотация. В статье представлены методы классификации гистологических изображений, объединяющие использование фильтров Габора и сверточных нейронных сетей (Convolutional Neural Networks, CNN). Филь-тры Габора применяются для предварительной обработки изображений, что позволяет извлекать частотные характеристики и ориентированные текстурные признаки, улучшая эффективность и точность последующей обработки с помощью CNN. Данный подход направлен на уменьшение количества обучаемых параметров, увеличение скорости обучения и повышение адаптивных способностей моделей, что особенно важно при ограниченном объеме обучающих данных. Исследуются различные конфигурации гибридных сетей с фильтрами Габора различных размеров и их влияние на производительность модели. Результаты экспериментов на наборе данных NCT-CRC-HE-100K демонстрируют, что предложенные методы позволяют достичь высокой точности классификации, превосходя традиционные CNN по количеству параметров и скорости обучения.

Ключевые слова: фильтры Габора; сверточные нейронные сети; гистологические изображения; классификация изображений.

Аннотация. The article presents histological image classification methods that combine the use of Gabor filters and convolutional neural networks (CNNs). Gabor filters are used to pre-process images to extract frequency characteristics and oriented texture features, improving the efficiency and accuracy of subsequent CNN processing. This approach is aimed at reducing the number of trainable parameters, increasing the learning speed and increasing the adaptive abilities of models, which is especially important when the amount of training data is limited. Various configurations of hybrid networks with Gabor filters of various sizes and their impact on model performance are investigated. Experimental results on the NCT-CRC-HE-100K dataset demonstrate that the proposed methods can achieve high classification accuracy, surpassing traditional CNNs in terms of the number of parameters and learning speed.

Ключевые слова: Gabor filters; convolutional neural networks; histological images; image classification.

I. Введение

Методы анизотропной фильтрации широко используются для получения надежного представления изображения. В частности, фильтр Габора, см. Рис. 1, широко используется в приложениях классификации текстур, поскольку он может извлекать частотное содержимое, зависящее от ориентации фильтра. Также ядро филь тра Габора можно легко изменять и масштабировать, что позволяет строить его с высокой эффективностью с точки зрения вычислений. При обработке изображений документов функции Габора идеально подходят для идентификации написания слова в многоязычном документе [1]. Фильтры Габора с разными частотами и ориентацией в разных направлениях использовались для локализации и извлечения областей, содержащих только текст, из сложных изображений документов (как серых, так и цветных), поскольку текст богат высокочастотными компонентами, тогда как изображения имеют относительно гладкий характер. Он также применялся для распознавания выражения лица [2], а также в задачах анализа медицинских изображений, например, изучения распределения направленности внутри пористой губчатой трабекулярной кости в позвоночнике [3].

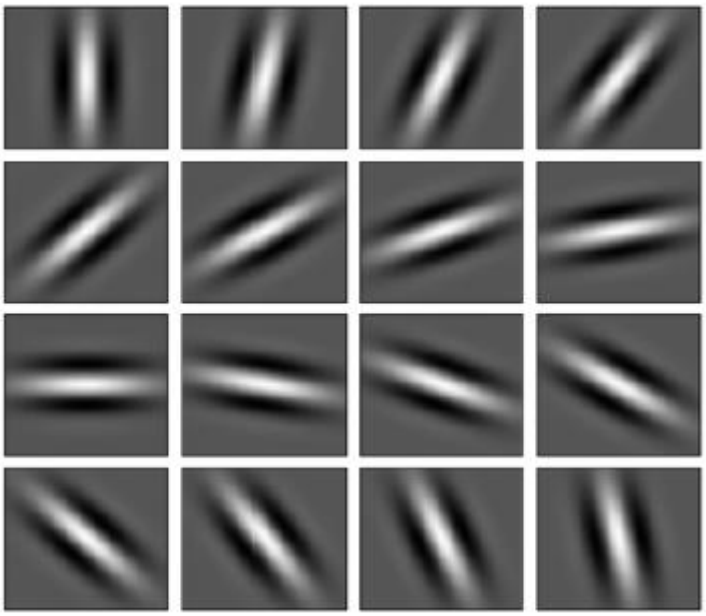


Рис. 1. Блок из 16 вещественных частей фильтров Габора, ориентированных под углами с шагом 11,25°.

Глубокие сверточные нейронные сети (Deep Convolution Neural Networks, далее просто CNN), основанные на сверточных фильтрах, широко используются в области компьютерного зрения, в том числе и в анализе медицинских снимков (например [4]). Их особенностью является автоматическая подстройка ядер сверток под анализируемые данные в процессе обучения нейронной сети. В отличие от созданных вручную фильтров без участия процесса обучения, анализ объектов на основе CNN — это метод, который позволяет анализировать пиксельное представление объекта напрямую. Однако такой метод имеет свои минусы: затраты на вычислительно сложное обучение и большое количество параметров модели. Например, типичная CNN имеет миллионы параметров, которые необходимо оптимизировать, и выполняет более миллиарда высокоточных операций для классификации одного изображения. Поэтому, если обучающих данных недостаточно, CNN не справляются с неизвестными преобразованиями, которые необходимо произвести над изображением, что часто встречается в анализе медицинских изображений. Избежать нехватки данных можно, добавляя в нейронные сети более предсказуемые и проверенные математические методы работы с изображениями. Одним из таких являются фильтры Габора.

Существуют модели с предварительной обработкой входных изображений фильтрами или детекторами признаков для повышения точности. Например, в [5] предлагается метод обнаружения черт лица с использованием комбинации фильтра Габора и нейронной сети. Фильтры Габора сначала применяются к изображениям, чтобы извлечь внутренние черты лица. Затем результирующие изображения признаков, полученных с помощью фильтров Габора, используются в качестве входных данных для нейронных сетей, чтобы извлечь новые признаки. Подобная идея также исследуется в [6] для распознавания рукописных цифр. Фильтры Габора используются исключительно как не меняющийся шаг предварительной обработки для оригинальных изображений. Помимо использования фильтра Габора в шаге предварительной обработки изображений, существует несколько работ, использующих фильтры Габора в сверточном слое нейронной сети. Например, в [7] фильтры Габора заменяют случайные ядра свертки в первом слое сети. В этом случае количество обучаемых параметров уменьшается благодаря введению фильтров Габора в качестве первого слоя. Та же стратегия используется в [8], где несколько слоев заменяются фильтрами Габора для улучшения эффективности модели.

В статьях [9, 10, 11] представлены интересные методы использования фильтра Габора в CNN в том числе и для медицинских задач [12]. Основной идеей является использование небольшого количества фильтров Габора для уменьшения количества обучаемых параметров нейронной сети, за счет встраивания этих фильтров внутрь слоя свертки, чтобы увеличить количество сворачиваемых фильтров этого слоя. Эти реализации применены к разным наборам данных и помимо уменьшения размеров CNN сохраняют точность предсказаний. Подробное описание данных подходов приведено в разделах 2 и 3.

Одно из применений подобной технологии есть в статье [13]. В ней рассматривается задача распознава- ния вен пальцев. Авторы предлагают модель RGCN, которая использует обучаемый слой свертки Габора для увеличения четкости рисунка вен. Модель использует способность фильтра Габора выделять ориентированные элементы на разных масштабах для извлечения информации вен пальцев.

В статье [14] была рассмотрена идея увеличения количества каналов при помощи фильтров Габора, где в

текстурной ветви сложной нейронной сети использовались адаптивные фильтры Габора. Согласно этой статье, такая модификация позволяет увеличить точность детекции или сегментации. Согласно утверждению авторов, при малом увеличении количества параметров (менее 1%) данная модификация в ветви, отвечающей за извле- чение признаков из региона интереса, увеличивает точность предсказаний всей детектирующей сети на 4.9% (с 85.9% до 90.8%).

В отличие от описанных выше работ в данной работе рассматриваются некоторые из этих методов с использованием фильтров Габора бо´льшего размера, чем (5 × 5), в применении к задаче классификации гистологических изображений. Мы проанализировали возможность встраивать эти фильтры в архитектуры для получения преимуществ по сравнению с существующими моделями. Также было исследовано, какую информацию нейронная сеть из фильтра Габора считает значимой и продемонстрированы фильтры Габора, которые получаются в результате обучения сети. Рассмотрены их преимущества и недостатки в сравнении с аналогичными сетями без модификаций.

В одной из предложенных модификаций сети AGResNet5 2 с помощью фильтров Габора получилось уменьшить количество обучаемых параметров на 30% по сравнению с сетью без фильтров Габора, что увеличило адаптивную способность сети, при сохранении точности. Другая предложенная модель FL-G 3 стала лучше не модифицированной модели в любой момент ее обучения и достигает лучшей максимальной точности за счет незначительного прироста количества параметров.

II. Постановка задачи

Пусть X – множество описаний объектов, Y – множество номеров или наименований классов. Существует неизвестная целевая зависимость – отображение $y' : X \to Y$, значения которой известны только на объектах конечной обучающей выборке $X^m = \{(x_1, y_1), ..., (x_m, y_m)\}$. Требуется найти отображение $y' : X \to Y$, определяющее класс произвольного объекта $x \in X$.

В данной работе необходимо классифицировать гистологические изображения, используя сверточную нейронную сеть. Финальным результатом классификации будет значение, соответствующее порядковому номеру класса, которому принадлежит данное изображение (в нашем случае значения от 0 до 8, где 0 - соответствует фону, 1-8 - соответствует определенному типу ткани).

III. Существующие нейронные сети с Фильтром Габора

1. Фильтр Габора

Фильтры Габора, см. Рис. 1, названные в честь Денниса Габора, представляют собой линейные фильтры, часто используемые для анализа текстур, потому что с помощью них можно анализировать наличие на изображении областей, обладающих определенными частотными характкристиками в определенных направлениях, что хорошо подходит для представления текстур и распознавания их на изображениях. В пространственной области двумерный фильтр Габора представляет собой функцию Гаусса, модулируемую синусоидальной плоской волной. Двумерный непрерывный фильтр Габора задается следующим образом:

$$g(x, y; \sigma, \theta, w, \gamma) = exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) exp(i(2\pi x' w)), \quad (1)$$

где $x' = x \cos\theta + y \sin\theta$, $y' = -x \sin\theta + y \cos\theta$. Фильтр Габора имеет следующие параметры: σ — стандартное отклонение Гауссового ядра, θ — ориентация фильтра, w — частота в комлексной экспоненте, и γ — коэффициент сжатия, характеризующий эллиптичность функции Габора.

2. Сверточная нейронная сеть с фильтрами Габора (Gabor Convolutional Network, GCN)

Чтобы понять, что извлекает CNN из данных изображения, визуализируем сверточные фильтры, которым обучилась AlexNet [15] на наборе данных ImageNet (Рис. 2). Можно заметить, что многие фильтры похожи на фильтры Габора. Например, эти фильтры с разной ориен- тацией способны выделять границы объекта в разных направлениях. На основе этих наблюдений в [9] авторы предложили CNN с банком ориентированных фильтров (БОФ) называемую Gabor Convolution Network [9, 16] (далее GCN). Было предложено взять небольшой набор фильтров, которые затем модулируются БОФ Габора с различной ориентацией и масштабами. БОФ реализованы с помощью базового элемента CNN — операции свертки, и, таким образом, могут быть легко интегрированы в

глубокую сверточную нейронную сеть любой архитектуры. Кроме того, поскольку БОФ генерируется на основе небольшого набора обучаемых фильтров свертки, GCN является более компактной моделью с меньшим количеством обучаемых параметров, чем исходная CNN. Однако, есть недостатки — введение дополнительных сверток приводит к увеличению количества операций свертки, по сравнениию с немодифицированными слоями.

Фильтры Габора имеют U направлений и V масштабов, за них отвечают параметры θ и σ. В GCN в качестве U берется равномерная сетка из полуинтервала $[0, \pi)$, а в V варьируется в зависимости от глубины слоя. Благодаря информации об ориентации и масштабе, захваченной фильтрами Габора в БОФ, соответствующие сверточные признаки улучшаются. Перед тем, как быть модулированными фильтрами Габора, сверточные фильтры в стандартных CNN обучаются алгоритмом обратного распространения ошибки. Пусть обученный фильтр имеет размер $N \times W \times W$, где $W \times W$ - размер фильтра, а N - число каналов. Если размерность весов на слой в традиционных CNN выражается как $C_{out} \times C_{in} \times W \times W$, то в GCN он будет представлен как $C_{out} \times C_{in} \times N \times W \times W$, где C_{out} и C_{in} представляют канал карты признаков вывода и ввода соответственно. Чтобы сохранить количество каналов карты признаков во время процесса прямого сверточного преобразования, N выбирается равным U, что является количеством ориентаций фильтров Габора, которые будут использоваться для модуляции данного обученного фильтра. БОФ получается на основе процесса модуляции с использованием U фильтров Габора на обученных фильтрах для заданного масштаба V (Рис. 3).

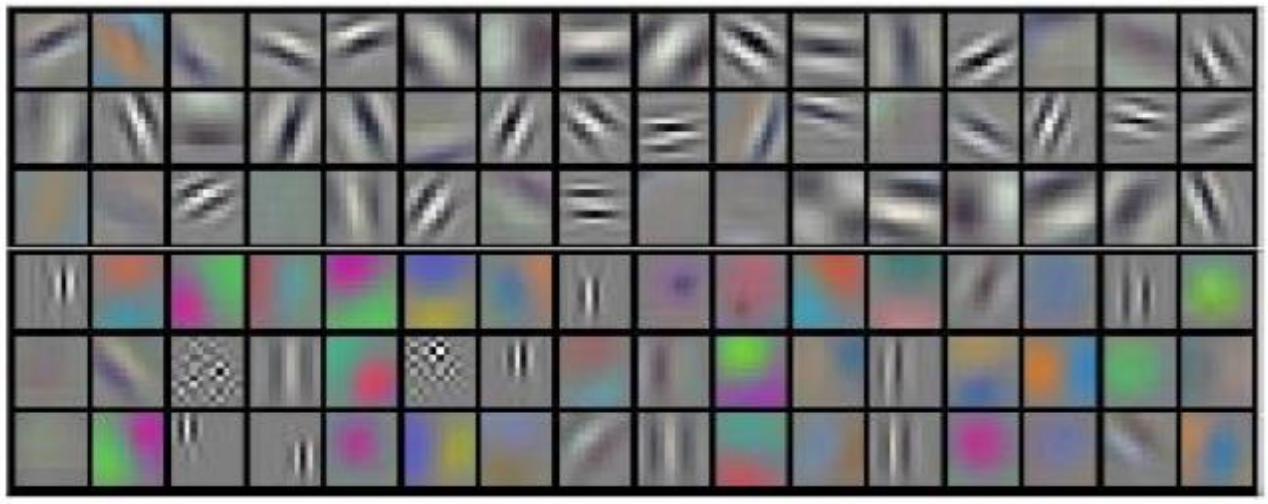


Рис. 2. Фильтры AlexNet обученной на наборе данных ImageNet.

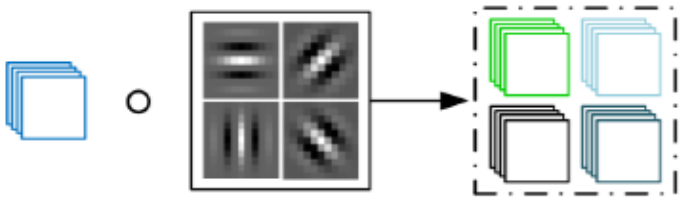


Рис. 3. Схема сверточного слоя с фильтром Габора.

Получить новые фильтры можно по следующей формуле

$$C_{i,u} = C_i \circledast G_u . \quad (2)$$

3. Адаптивный фильтр Габора (Adaptive Gabor Convolutional Network, AGCN)

Также существует модификация GCN с адаптивным фильтром Габора [10]. Идея заключается в том, чтобы обучать параметры фильтра Габора. Таким образом за счет небольшого увеличения количества параметров сети сильно увеличивается ее гибкость, так как при отсутствии необходимости использования фильтров Габора они станут вырожденными. При использовании K фиксированных направлений, т.е. фиксированных параметров θ, обучаться будут σ, w, γ. Итого, для одного фильтра будет $3K$ дополнительных обучаемых параметров (где K — количество фильтров). Визуализация количества параметров для одного обучаемого фильтра приведена на рис. 4.

0,0	0,1	0,2	0,3	0,4
1,0	1,1	1,2	1,3	1,4
2,0	2,1	2,2	2,3	2,4
3,0	3,1	3,2	3,3	3,4
4,0	4,1	4,2	4,3	4,4

$+ K \times$

σ
λ
γ

Рис. 4. Визуализация обучаемого фильтра 5×5, представленной в виде матрицы с пронумерованными элементами, и параметрами адаптивных фильтров Габора.

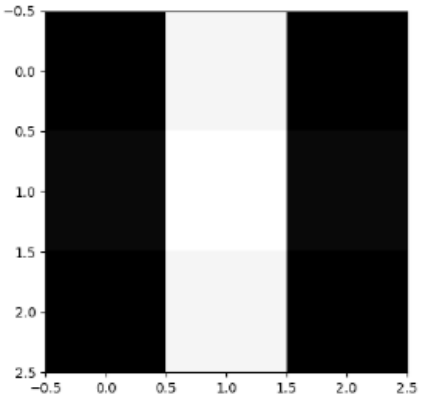


Рис. 5. Фильтр Габора 3x3 $\sigma = 1, \theta = 0, w = 1, \gamma = 0.2$.

IV. Предложенные нейронные сети с Фильтром Габора

1. Модифицированная GCN с большим размером ядра

В работе [9] авторы используют фильтры размера 3 × 3 в небольшой сети из блоков, состоящих из двух последовательных сверток (Рис. 6). Такая сеть применялась к задаче распознавания цифр на наборе данных MNIST [17], где размер изображений достаточно малый, а сама задача является несложной. В данной работе мы будем рассматривать модели с большими размерами фильтров.

Для того, чтобы сохранить количество параметров сопоставимым для сетей с разным размером фильтров, будем использовать различное количество блоков. Для сетей с фильтрами 3 × 3 и 5 × 5 будем использовать 5 последовательных блоков, приведенных ниже, для 7 × 7 и 9 × 9 — 4, а для 11 × 11 — 3. Во всех моделях после сверточных блоков в сети используется один полносвязный слой.

Основной целью использования БОФ остается уменьшение количества обучаемых параметров модели, с возможностью сохранения точности. Поэтому для сравнения параметры сетей подбирались так, чтобы сохранить это значение в пределах погрешности. Таким образом были построены 5 нейронных сетей с использованием фильтров Габора. В каждой сети использовались блоки со свертками определенного размера: (3 × 3), (5 × 5), (7 × 7), (9 × 9), (11 × 11). В качестве гиперпараметров использовались значения U = 4 и V = 1. Здесь и далее,

чтобы избежать переобучения, был добавлен слой Dropout с вероятностью выброса 0.2 перед слоем классификации. Также во всех последующих экспериментах использован размер пакета 64 и шаг обучения 0.0001. Данные сети были обучены на наборе данных NCT-CRC-HE-100K. Все сети имеют одинаковое количество параметров (примерно 3.5 млн обучаемых параметров) и обучаются по 100 эпох. В качестве функции потерь использовалась перекрестная энтропия (Cross-entropy loss):

$$L(y, \hat{y}) = -\sum_{i=1}^{K} 1_{[y=i]} \log \hat{y} \tag{3}$$

$$\hat{y} = \mathrm{Softmax}(x) = \left\{ \frac{\exp(x^i)}{\sum_{j=1}^{K} \exp(x^j)} \right\}_{i=1}^{K}, \quad K - \text{число классов.} \tag{4}$$

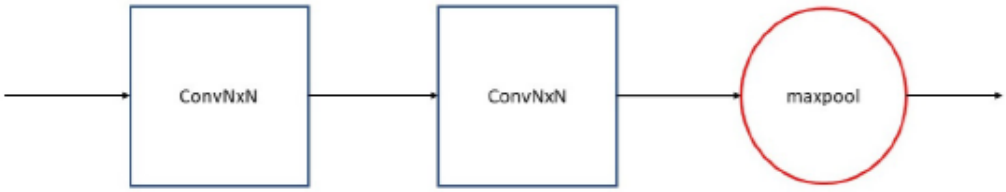


Рис. 6. Блок нейронной сети.

Ниже приведен алгоритм обучения сверточной нейронной сети с фильтрами Габора.

Algorithm 1 Алгоритм обучения сети.

Инициализация гиперпараметров и параметров структуры сети;
Задание значений U и V (U и V - количество ориентаций и масштабов фильтров Габора соответственно);
Начало обучения:
repeat
Ввод мини-пакета изображений для обучения;
Получение БОФ путем использования фильтров Габора на обученных фильтрах;
Прямое сверточное преобразование GCN на основе входных карт признаков и БОФ;
Расчет функции потерь кросс-энтропии, а затем обратное распространение ошибки;
Обновление обученных фильтров;
until Достигнуто максимальное количество эпох

Результаты обучения предложенных сетей изображены на Рис. 11.

Далее мы объединили обученные сети в суммирующую сеть: предсказания каждой сети подавалось в новый полносвязный слой, который на основе предсказаний делал свои. Полученные веса этого слоя приведены в таблице 1.

Сеть	Вес
GCN3	0.0064
GCN5	0.3751
GCN7	0.4126
CN9	0.4571
GCN11	0.4745

Таблица 1. Веса суммирующей сети.

В таблице 1 видно, что чем больше размер ядра свертки, тем больший вес имеет данная сеть. Это означает, что суммирующая сеть считает предсказания сети с большими фильтрами более надежными, несмотря на ее относительно низкую точность.

2. Свертка с фильтром Габора в ResNet

Результаты, полученные для модифицированной GCN, демонстрируют, что идея использования больших фильтров повышает точность классификации. Поэтому было предложено применить подобный подход для сетей типа ResNet. Мы использовали сеть ResNet50 [18], в которой были заменены слои сверток (3 × 3) на слои с фильтрами Габора разных размеров, а так же их адаптивные аналоги. Результаты обучения данных моделей изображены на Рис. 12 - Рис. 18.

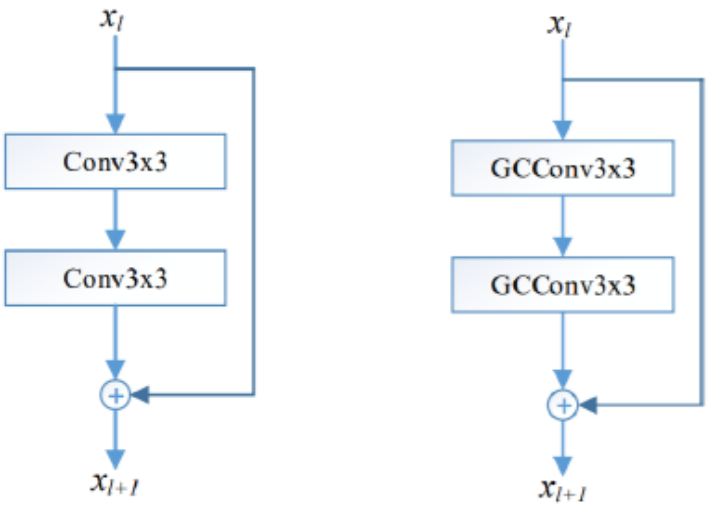


Рис. 7. Блок ResNet и GResNet.

Преимущества подхода заключается в перебрасывании уже полученных признаков мимо сверточных слоев. Таким образом использование неудачно обученных фильтров, не будет портить результат предыдущих слоев. В случае использования фильтров Габора данное свойство особенно актуально, т.к. для некоторых данных и некоторых параметров фильтров Габора соответствующие фильтры будут низкоинформативными и иметь достаточно низкий отклик.

3. FL-G (First layer Gabor)

У предыдущего подхода есть недостаток — свертка производится с фильтрами, которые оптимизирует сеть. Относительно изображения эти фильтры имеют малый размер. Если использовать фильтр Габора большего размера, то придется увеличивать и ядро свертки. Во-первых, это противоречит идее современных сверточных нейронных сетей – множество последовательных слоев свертки с малыми ядрами, обычно не превосходящими 7 × 7. Во-вторых, увеличение размера ядра увеличивает количество параметров сети, а основной идеей использование фильтров Габора является их уменьшение.

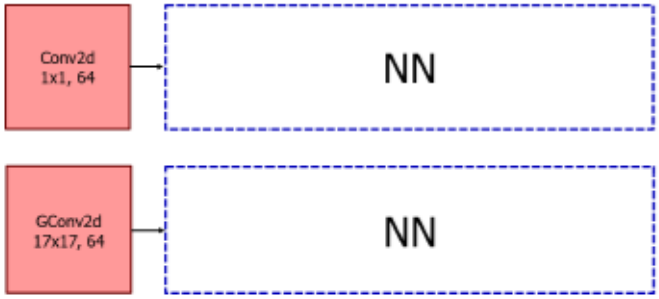
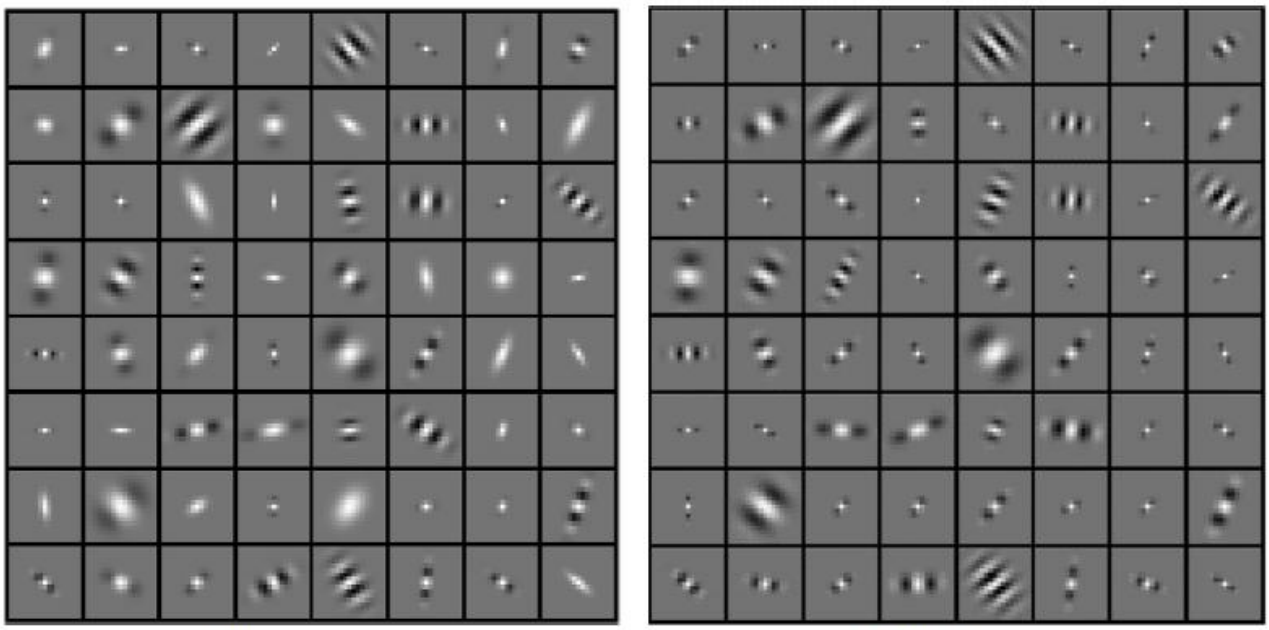


Рис. 8. Архитектура сетей.

Поэтому используем адаптивные фильтры Габора в качестве первого слоя нейронной сети, чтобы увеличить количество каналов, которые будут подаваться в оставшуюся часть сети. Дополнительно введем ограничения на параметры фильтров: $\sigma > 1$, $\gamma > 1$, $w \in (0, 5)$, чтобы в процессе оптимизации они находились в пределах значений, которые могут поместиться в фильтр размера 17 × 17, а локальные максимумы функции не попадали на каждую точку сетки, т.е. фильтр Габора не вырождался в Гауссовский фильтр.

Для эксперимента построим две одинаковые сети, отличающиеся лишь первым слоем. В одной, в качестве этого слоя используем свертку с адаптивными фильтрами Габора. В другой — это будет слой со сверткой 1 × 1. Таким образом количество каналов, которые будут поданы дальше, будут совпадать в обеих сетях.

Дополнительно приведем фильтры Габора, которые получаются в процессе обучения (Рис. 9). Заметные изменения происходят на первых 25 эпохах, потом фильтры почти не изменяются, что соответствует графикам обучения модели (Рис. 19), где видно, что уже на 20 эпохе модель выходит на плато и дальнейшие изменения ошибки несут уточняющий характер.



(a) При инициализации (b) На 100 эпохе

Рис. 9. Примеры вещественной части фильтров Габора, полученных при обучении.

V. Результаты

1. *Набор данных*

Данные для обучения берутся из открытого набора данных NCT-CRC-HE-100K [19], см. Рис. 10. Эти изображения были вручную извлечены из окрашенных слайдов раковой ткани человека. Количество изображений отображено в таблице 2. Каждый Для экспериментов использовалось 22000 случайно выбранных изображений из всего набора и разбиты на две не пересекающихся группы: 20000 для обучения, 2000 для тестов.

2. *Результаты экспериментов*

На Рис. 11 приведены результаты обучения для блочных сетей из раздела 1. Заметно выбиваются сети с фильтром свертки (3 × 3) с фильтром Габором и без. Они сильно колеблются и не выходят на плато, что говорит о их медленном обучении и нестабильности. Дополнительно приведем средние и максимальное значение ошибки на тестовой выборке для представленных в таблице 3 сетей. Чем больше размер фильтра, тем быстрее сеть обучается, что повторяет результаты графиков, где сеть с фильтрами (11 × 11) сразу выходит на значения > 60%. Однако, ее максимальное значение проигрывает сети с фильтрами (5 × 5).

На Рис. 12 представлены графики обучения для сетей ResNet50 и ее модификациями с фильтрами Габора. В Таблице 4 также приведены средняя и максимальная точности сетей. В этом эксперименте ситуация продемонстрирована обратная ситуация, в сравнении с первым экспериментом. Чем больше размер фильтра, тем ниже способность к обучению и ниже максимальная точность. Добавление сверток с фильтрами Габора согласно методу, представленному в статье [9], в сеть ResNet с фильтрами малого размера (GResNet3) дает хороший результат. При этом количество обучаемых параметров уменьшается почти в 4 раза, значение метрики почти не падает, а скорость обучения растет. Большой фиксированный фильтр в сетях GResNet7 - GResNet11 может накладывать ограничения, которые сказываются на точности, поэтому далее будут приведены результаты для адаптивных фильтров Габора, которые будут обучаться совместно с сетью.

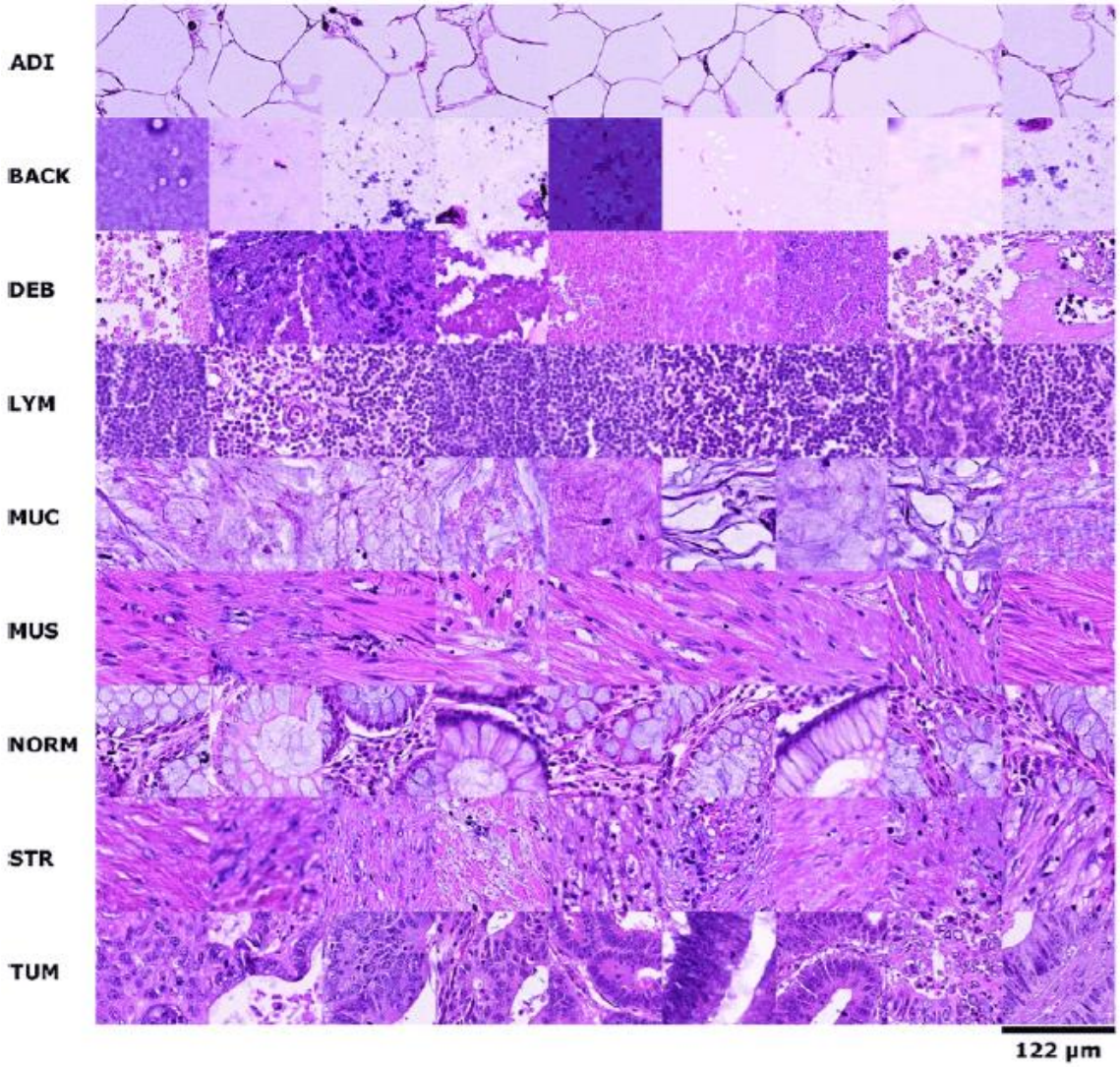


Рис. 10. Пример гистологических изображений из NCT-CRC-HE-100K.

Класс	Примерное число изображений в тыс. (±100 изображений)
ADI	10,4
BACK	10,6
DEB	11,5
LYM	11,6
MUC	8,9
MUS	13,5
NORM	8,8
STR	10,4
TUM	14,3

Таблица 2. Распределение изображений по классам в наборе данных NCT-CRC-HE-100

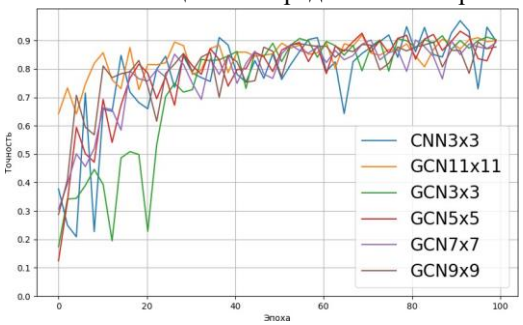


Рис. 11. График обучения сетей из раздела 1.

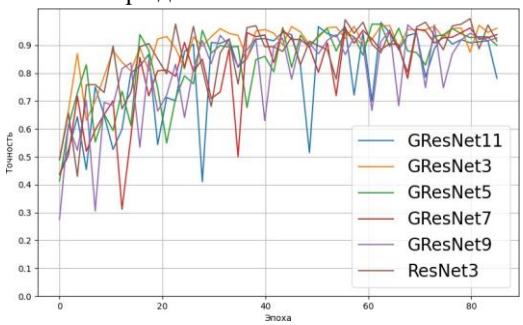


Рис. 12. Сравнение GResNet50 и ResNet50.

Сеть	Среднее значение точности	Максимальная точность
GCN3	74,6	91.8
GCN5	78,3	93.9
GCN7	78.7	90.0
GCN9	80.7	89.6
GCN11	84.1	91.7
CNN3	78.4	96.5

Таблица 3. Среднее и максимальное значение точности сетей из раздела 1. Здесь CNN3 - аналогичная сеть с фильтрами размера (3×3) без модификаций.

Также в ходе экспериментов были обучены сети ResNet50, у которых часть блоков была заменена на свертки с фильтром Габора, а остальные оставлены без изменений. На Рис. 13 - 17 представлены графики их обучения. Эти сети во многом оказались хуже, чем свои аналоги с полной заменой на свертки с фильтром Габора или же без этой модификации вовсе.

Сеть	Среднее значение точности	Максимальная точность
GResNet3	89.5	97.1
GresNet5	84.4	96.9
GResNet7	82.4	96.3
GResNEt9	81.3	95.6
GResNet11	81.4	95.3
ResNet3	87.0	97.7

Таблица 4. Среднее и максимальное значение точности сетей ResNet и GResNet. Здесь ResNet3 - ResNet50 с фильтрами размера (3×3) без модификаций.

На Рис. 18 представлены графики обучения для сети ResNet50 и ее модификациями с адаптивными фильтрами Габора. Также в Таблице 5 представлены средняя и максимальная точности сетей. Как видно из таблиц, большие фильтры не дают преимущества в более сложных архитектурах сетей, в то время как фильтры малого размера дают выигрыш в скорости обучения и почти не проигрывают в точности, а местами даже превосходят ее. Самый значимый результат у AGResNet3 помимо увеличения точности это уменьшение количества обучаемых параметров на 30%.

Приведем результаты применения сети FL-G, см. Рис. 19. В отличие от других предложенных моделей, данная модель не имеет преимущества в количестве параметров, т.к. здесь в качестве первого слоя добавлен адаптивный фильтр Габора, что увеличивает количество обучаемых параметров на 192. Однако, это очень малое значение, потому что даже небольшие сети содержат тысячи и миллионы обучаемых параметров. В то же время, дополнительный слой с фильтром Габора дает прирост и к точности (с 94.6% до 95.6%), и к скорости обучения (Рис. 19). Сеть быстро обучается и выходит на плато, в то время как ее аналог без фильтров Габора все еще сильно колеблется. Графики обучения сети FL-G все время находится выше, чем график сети без модификации. Таким образом, на любом этапе обучения сети, она дает лучший результат.

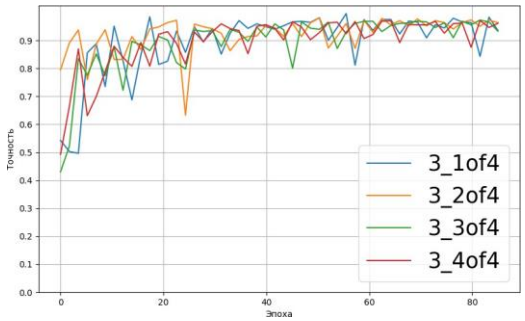


Рис. 13. График обучения ResNet50 и частичных GResNet со сверткой (3×3).

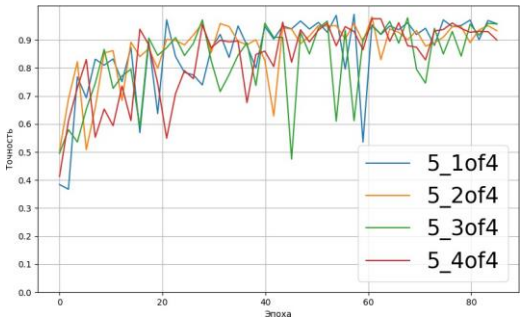


Рис. 14. График обучения ResNet50 и частичных GResNet со сверткой (5×5).

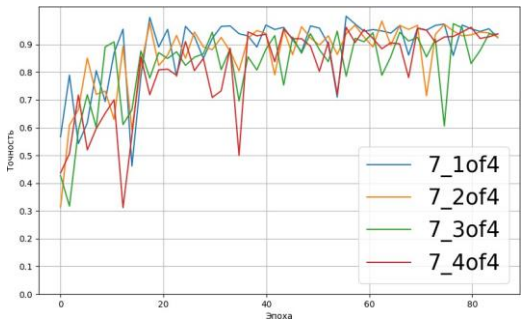


Рис. 15. График обучения ResNet50 и частичных GResNet со сверткой (7×7)

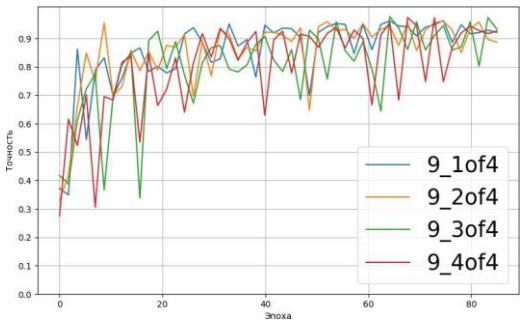


Рис. 16. График обучения ResNet50 и частичных GResNet со сверткой (9×9).

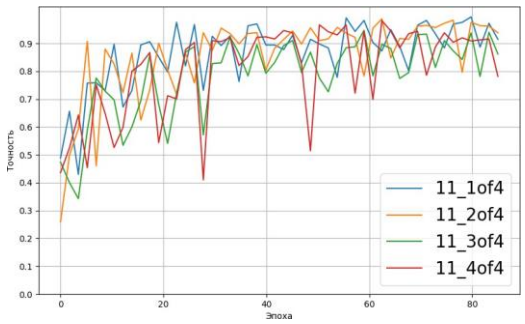


Рис. 17. График обучения ResNet50 и частичных GResNet со сверткой (11×11).

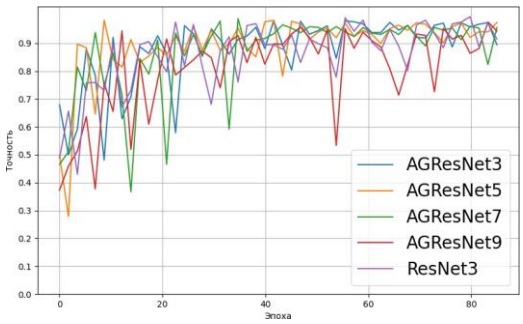


Рис. 18. Графики обучения AGResNet50 с разными размерами и ResNet50.

VI. Заключение

В данной работе были предложены методы классификации гистологических изображений с использованием сверточных нейронных сетей с фильтрами Габора. Предложенные модели были протестированы на наборе данных NCT-CRC-HE-100K. Модели с адаптивными фильтрами Габора, встроенными в архитектуру ResNet50 (AGResNet) показали хорошую точность (>97%). При интеграции фильтров Габора в эту сеть удалось снизить количество обучаемых параметров на 30% и при этом получить схожую точность 97,5% по сравнению с моделью, не использующей фильтры Габора (98%). Модель с добавленным первым слоем для увеличения количества входных каналов при помощи фильтров Габора (FL-G) за счет незначительного прироста параметров позволила увеличить и скорость обучения, и точность с 94.6% до 95.6%.

Сеть	Среднее значение точности	Максимальная точность
AGResNet3	88.6	97.9
AGresNet5	89.8	97.5
AGResNet7	88.6	97.6
AGResNEt9	83.1	97.2
AGResNet11	82.9	97.0
ResNet3	87.0	97.7

Таблица 5. Среднее и максимальное значение точности сетей ResNet50 и GResNet50.

Источник финансирования.

Исследование выполнено за счет гранта Российского научного фонда №22-41-02002.

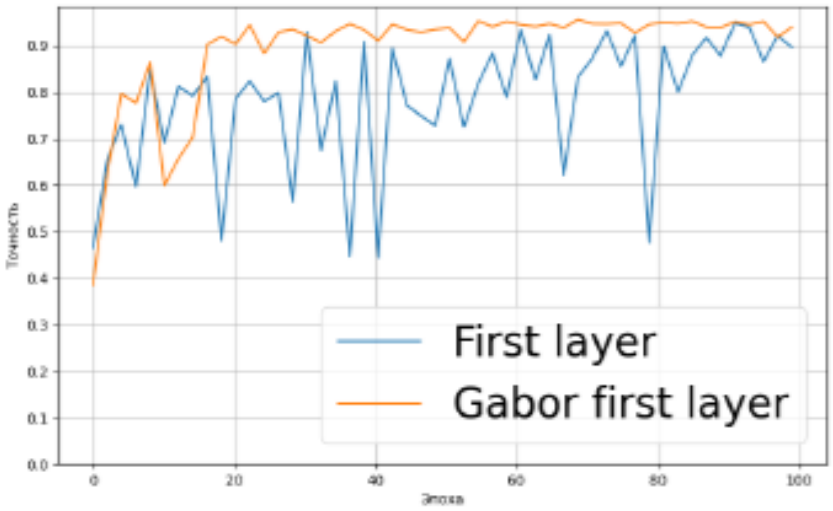


Рис. 19. Сравнение G-FL и FL и ResNet50.

Список литературы

[1] Fogel Itzhak, Sagi Dov. Gabor filters as texture discriminator // Biological cybernetics. — 1989. — Vol. 61, no. 2. — Pp. 103–113.

[2] Coding facial expressions with gabor wavelets / Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba // Proceedings Third IEEE international conference on automatic face and gesture recognition / IEEE. — 1998. — Pp. 200–205.

[3] On estimating the directionality distribution in pedicle trabecular bone from micro-CT images / Catherine M Gdyczynski, Amir Manbachi, SayedMasoud Hashemi et al. // Physiological measurement. — 2014. — Vol. 35, no. 12. — P. 2415.

[4] Classification of breast cancer histology images using convolutional neural networks / Teresa Arau´jo, Guilherme Aresta, Eduardo Castro et al. // PloS one. — 2017. — Vol. 12, no. 6. — P. e0177544.

[5] Kwolek Bogdan. Face detection using convolutional neural networks and Gabor filters // Artificial Neural Networks: Biological Inspirations–ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part I 15 / Springer. — 2005. — Pp. 551–556.

[6] Calderon Andres, Roa Sergio, Victorino Jorge. Handwritten digit recognition using convolutional neural networks and gabor filters // Proc. Int. Congr. Comput. Intell. — 2003. — Pp. 1–9.

[7] Chang Shuo-Yiin, Morgan Nelson. Robust CNN-based speech recognition with Gabor filter kernels // Fifteenth annual conference of the international speech communication association. — 2014. — Pp. 905–909.

[8] Sarwar Syed Shakib, Panda Priyadarshini, Roy Kaushik. Gabor filter assisted energy efficient fast learning convolutional neural networks // 2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED) / IEEE. — 2017. — Pp. 1–6.

[9] Gabor convolutional networks / Shangzhen Luan, Chen Chen, Baochang Zhang et al. // IEEE Transactions on Image Processing. — 2018. — Vol. 27, no. 9. — Pp. 4357–4366.

[10] Adaptive Gabor convolutional networks / Ye Yuan, Li-Na Wang, Guoqiang Zhong et al. // Pattern Recognition. — 2022. — Vol. 124. — P. 108495.

[11] Learnable Gabor Convolutional Networks / Guoqiang Zhong, Wei Gao, Wencong Jiao et al. // International Conference on Neural Information Processing / Springer. — 2019. — Pp. 305–313.

[12] Deformable gabor feature networks for biomedical image classification / Xuan Gong, Xin Xia, Wentao Zhu et al. // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. — 2021. — Pp. 4004–4012.

[13] Residual Gabor convolutional network and FV-Mix exponential level data augmentation strategy for finger vein recognition / Yifan Wang, Huimin Lu, Xiwen Qin, Jianwei Guo // Expert Systems with Applications. — 2023. — Vol. 223. — P. 119874.

[14] Learning gabor texture features for fine-grained recognition / Lanyun Zhu, Tianrun Chen, Jianxiong Yin et al. // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2023. — Pp. 1621–1631.

[15] Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey E. Imagenet classification with deep convolutional neural networks // Communications of the ACM. — 2017. — Vol. 60, no. 6. — Pp. 84–90.

[16] Fast and efficient facial expression recognition using a gabor convolutional network / Ping Jiang, Bo Wan, Quan Wang, Jiang Wu // IEEE Signal Processing Letters. — 2020. — Vol. 27. — Pp. 1954–1958.

[17] Gradient-based learning applied to document recognition / Yann LeCun, Le´on Bottou, Yoshua Bengio, Patrick Haffner // Proceedings of the IEEE. — 1998. — Vol. 86, no. 11. — Pp. 2278–2324.

[18] Deep residual learning for image recognition / Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — Pp. 770–778.

[19] Kather Jakob Nikolas, Halama Niels, Marx Alexander. 100,000 histological images of human colorectal cancer and healthy tissue. — 2018. — URL: https://doi.org/10.5281/zenodo.1214456.

УДК: 167
DOI: 10.25206/978-5-8149-3873-2-2024-669-674

Scale Invariance property of Kolmogorov-Arnold Networks for MRI Gibbs Ringing Suppression

M. A. Penkin[1], A. S. Krylov[1]
[1] *Faculty of computational mathematics and cybernetics, Lomonosov Moscow state university Student of the Faculty of Computational Mathematics and Cybernetics at Lomonosov Moscow State University, Moscow, Russia*

Abstract. Gibbs ringing is an artifact in signal and image processing, particularly in scenarios involving reconstruction from sparse or limited data. This paper presents an approach for suppressing Gibbs ringing through the application of a convolutional Kolmogorov-Arnold Network (KAN), a type of artificial neural network inspired by the theory of functional analysis. Our method leverages KANs' capability of reshaping the neural network learning by the stack of non-linear learnable activation functions, derived from the Kolmogorov-Arnold representation theorem, to effectively mitigate oscillatory artifacts. First, we train KANs over the classical solvers: Perona-Malik and Kellner, – thus proving KANs' potential in rediscovering classical deringing algorithms. Second, we examine KANs' scale invariance over the multi-scale testing synthetic dataset, revealing this crucial property of KANs for real-life medical applications. Experiments show a significant reduction in ringing artifacts while preserving valuable image features and overall signal fidelity. The implications of our findings suggest that Kolmogorov-Arnold Networks represent a promising tool for magnetic resonance imaging, enhancing reconstruction quality to a large extent.

Keywords: Gibbs ringing artifact, Medical imaging, Magnetic resonance imaging, Kolmogorov-Arnold, Anisotropic diffusion.

I. Introduction

Magnetic resonance imaging (MRI) is a biomedical imaging modality that underpins many recent breakthroughs in biology and medicine. Due to the non-ideal nature of MRI reconstruction, sharp discontinuities in an underlying signal lead to oscillatory artifacts, known as Gibbs ringing [1] or truncation artifacts (see Fig. 1). Gibbs ringing detracts from the quality and the interpretability of reconstructed images, either obscuring important details for medical diagnosis or preventing accurate data analysis. So, post-processing of MR images [2] is often incorporated into MRI scanners by manufacturers, and such software plays a critical role in the final image quality for clinical reporting and interpretation.

Traditional methods [3] for Gibbs ringing suppression, such as windowing functions and frequency domain filtering, often fall short in preserving high-frequency details or even introduce their own artifacts, necessitating the exploration of more advanced techniques.

In recent years deep learning approaches have gained prominence as powerful tools for various image processing tasks, capitalizing on their ability to model complex patterns in a data-driven manner. Among these approaches, Convolutional Neural Networks (CNNs) have demonstrated a notable success in image deringing [4, 5], denoising [6, 7], deblurring [8] and inpainting [9], largely due to the hierarchical feature extraction. However, CNNs' training can be challenging, particularly in ensuring their generalization across diverse image types, noise conditions and input scales [10].

This paper considers an alternative framework for MRI reconstruction by utilizing recently proposed convolutional Kolmogorov-Arnold Neural Networks (KANs) [11, 12]. This hybrid architecture integrates the principles of convolutional processing with the theoretical underpinnings of the Kolmogorov-Anold representation theorem [13], providing a new mechanism for non-linear functions approximation. While CNNs have fixed activation functions on nodes ("neurons"), KANs have learnable activation functions on edges ("weights"), thus dispensing from linear weights: every weight parameter is now replaced by a univariate spline function. The advanced generalization is given by the construction: splines are able to smoothly represent an arbitrary activation function that can hardly be found using a ReLU in between convolutions. One might worry that KANs are hopelessly expensive, since each CNN's weight parameter becomes KAN's spline function. Fortunately, KANs usually require much smaller computation graphs in contrast to CNNs [12].

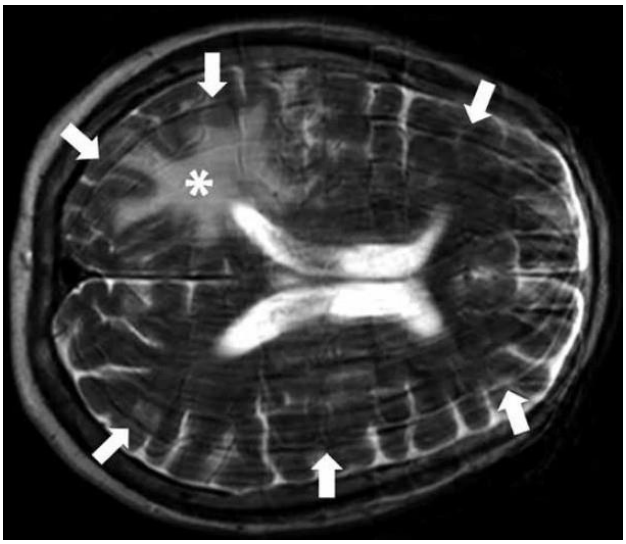


Figure 1. Axial T2-weighted image of the brain with notable truncation artifacts (white errors).

We explore the architecture of KANs, studying how they can be employed to effectively suppress Gibbs ringing while preserving essential image features. Through the experimentation on IXI dataset (http://brain-development.org/ixi-dataset/) we demonstrate that the method can not only rediscover existing classical deringing algorithms (i.e. Perona-Malik [14] and Kellner [2]) but, moreover, possesses a strong scale invariance over the unseen multi-scale testing synthetic dataset, obtaining a principle property for real-world medical applications.

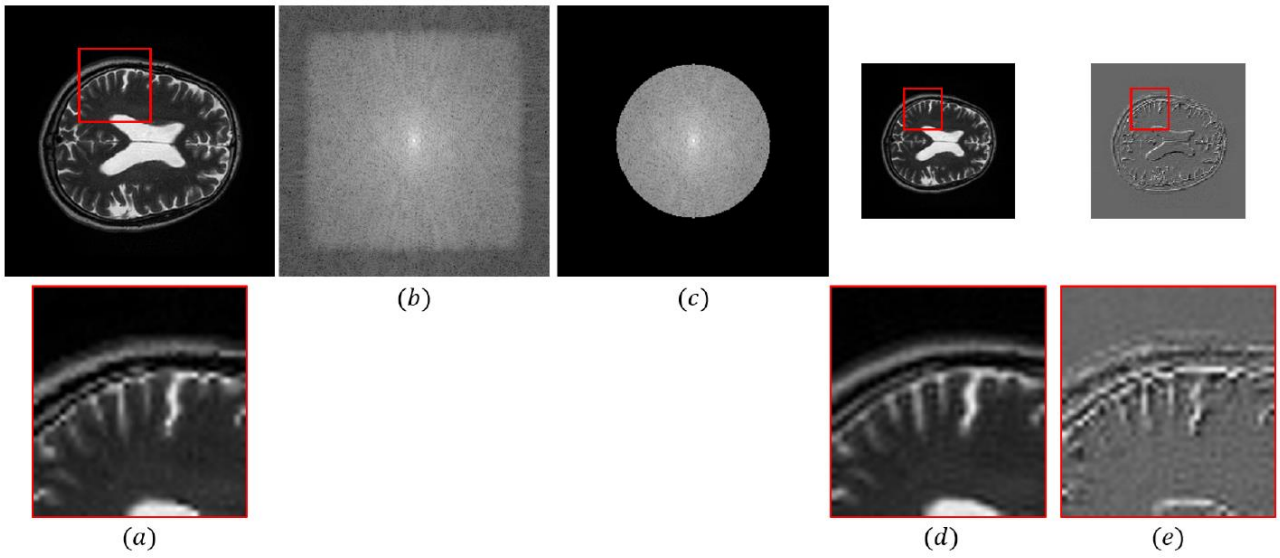


Figure 2. Low-resolution magnetic resonance image data retrieval. Red rectangles point out zoomed areas. (*a*) – high-resolution (HR) ground-truth image of shape (255x255x1), I^1, (*b*) – Fourier spectrum of (*a*), (*c*) – Fourier spectrum truncation, (*d*) – inverseFourier transform of (*c*), yielding low-resolution (LR) Gibbs-corrupted image of shape (145x145x1), I^0, (*e*) – residual image: (*d*) – (*a*), – demonstrating the distortion.

Synthetic data is used in medical imaging data-driven algorithms as a common solution to a variety of problems, i.e. privacy issues. Without access to data, AI scientists and developers cannot make meaningful progress, whereas atthe same time data cannot be shared carelessly. The result is typically a long series of “jumping through hoops” in anattempt to access the necessary data. So, synthetic data acquisition (see Fig. 2) offers a potential solution.The presented findings contribute to the growing body of literature on physics-informed neural networks [15, 16] for image processing.

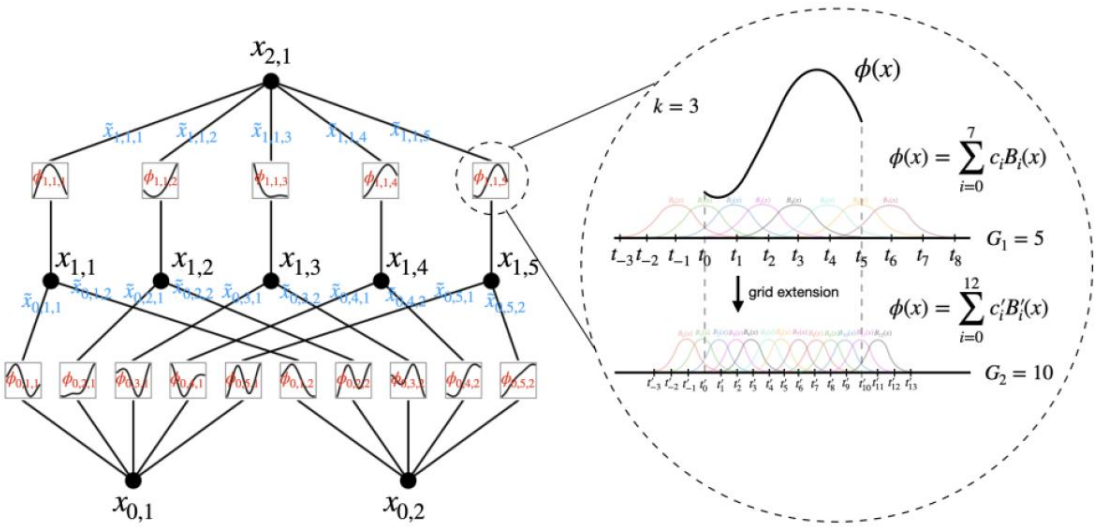


Figure 3. Kolmogorov-Arnold Network.

II. Related work

The Kolmogorov-Arnold representation theorem [13] is a fundamental result in the theory of functions of several variables.

Andrey Kolmogorov and Vladimir Arnold established that if $f : [0, 1]^n \to \mathbb{R}$ is a multivariate continuous function, then f can be written as a finite composition of continuous functions of a single variable and the binary operation of addition. More specifically,

$$f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} \Phi_q \sum_{p=1}^{n} \phi_{q,p}(x_p), \tag{1}$$

where $\varphi_{q,p} : [0, 1] \to \mathbb{R}$ and $\Phi_q : \mathbb{R} \to \mathbb{R}$.

One might naively consider this great news for machine learning: learning a high-dimensional function boils down to learning a polynomial number of 1D functions. However, non-smoothness, or even fractal, of the inner functions has limited the practical use of the promising representation. Although, there is still some debate on this.

There have been various attempts [17, 18, 19] to employ neural networks, called Kolmogorov-Arnold Networks, modeled on the Kolmogorov-Arnold representation instead of the universal approximation theorem [20]. To further encourage usage of the Kolmogorov-Arnold representation theorem the authors [11] generalized the architecture to arbitrary widths and depths, keeping the idea of the representation theorem but not sticking rigorously to the original equation, which has only two-layer non-linearities and a small number of terms (2n + 1) in the hidden layer. Hereby, KANs are usually composed of a series of concatenated Kolmogorov-Arnold layers (see Fig. 3, each containing a set of learnable one-dimensional activation functions. Such architecture has proven to be effective at approximating high-dimensional functions, demonstrating robust performance across various applications [11].

Still, there has been a lack of KANs' incorporation into medical imaging.

III. Method

We propose the application of Kolmogorov-Arnold Networks for MRI Gibbs ringing suppression, verifying KANs' capacity for rediscovering two classical deringing solvers: Perona-Malik (PM) and Kellner.

The training, validating and testing datasets $\{(I_i^0, I_i^*)\}$ for KANs fitting and evaluating, where I_i^0 – is an input corrupted image and I_i^* – is a classical solver output for I_i^0, – were created in a synthetic manner from the high-resolution (HR) MRI IXI dataset (http://brain-development.org/ixi-dataset/), composed of 581 T1, 578 T2 and 578 PD volumes.

10427 training, 2016 validating and 1617 testing image pairs were generated using the following data processing pipeline (see Fig. 2):

- $I_i^1 \leftarrow$ load high-resolution ground-truth image;

- apply Fourier transform;
- crop frequency spectrum: central 25% of frequency domain is kept;
- I_i^0 apply inverse Fourier transform, getting low-resolution Gibbs-corrupted image, additionally augmented by random Gaussian noise with $\sigma = 0.01$;
- I_i^* ← apply classical solver, i.e. PM(I_i^0; T = 3, K = 0.1, $\Delta t = 0.1$) and Kellner(I_i^0; $k_1 = 1$, $k_2 = 3$).

Kolmogorov-Arnold Networks are employed as deep learning operators to build the function space mappings: $\hat{I} : I^0 \rightarrow I^*$, – in a data-driven manner from the provided dataset $\{(I_i^0, I_i^*)\}$, informed by the classical solvers: Perona-Malik and Kellner.

We utilized the following KAN setup for training both the Perona-Malik and the Kellner function space mappings:

- 16 (chosen by cross-validation from 4, 8, 16) KAN layers with 3×3 kernel size, projecting input images into the feature space;
- a single convolutional 3 × 3 restoration layer, reconstructing grayscale MRI images.

Adam [21] optimizer is utilized to fit the algorithm: $lr = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$.

IV. Results

Following the results analysis, Fig. 4 illustrates strong scale invariance of Kolmogorov-Arnold Networks over the multi-scale $\{\times 1.0,\ \times 0.9,\ \times 0.8,\ \times 0.7,\ \times 0.6\}$ testing synthetic dataset. Gibbs ringing suppression effectiveness is evaluated by total variation (TV), coupled with peak signal-to-noise ratio (PSNR) to quantify reconstruction quality.

Indeed, KANs, being trained on a single scale 1, learned the Perona-Malik and the Kellner function space mappings and, what's more, are robust across the testing grid of scales. It is worth noting that, additionally, KANs demonstrate a certain regularization by uplifting total variation of PM towards the ground-truth values and improving PSNR of Kellner. While we expected such behaviour of Fourier Neural Operators [16] due to the inside harmonics truncation module, here KANs reveal the same effect as a nice dividend.

The code is available by the following link: https://github.com/MaksimPenkin/MedicalKAN.

V. Conclusion

The application of Kolmogorov-Arnold Networks for Gibbs ringing suppression in MRI demonstrates effectiveness at enhancing image quality by rediscovering the classical deringing solvers and, furthermore, discloses KANs' scale invariance over the grid of input scales, unseen during training. At present, the primary limitation of KANs is slow training [11]. Luckily, KANs typically need significantly smaller computation graphs than CNNs, so we avoided this issue when working on the stated MRI restoration task. Continued research in this area is likely to further KANs' optimization, leading to more advanced implementations within clinical MRI protocols. As machine learning techniques evolve, the integration (e.g. quantization and compression) [22] of neural networks is developed simultaneously and plays an increasingly important role in pushing deep learning solutions towards the edge devices.

Acknowledgments

The work was supported by Russian Science Foundation Grant 22-41-02002.

References

[1] Gibbs J Willard. Fourier's series. — Nature, 1898.

[2] Gibbs-Ringing Artifact Removal Based on Local Subvoxel Shifts / Elias Kellner, Bibek Dhital, Valerij G Kiselev, Marco Reisert // Magnetic resonance in medicine. — 2016. — Vol. 76, no. 5. — Pp. 1574–1581.

[3] Jerri Abdul J. The Gibbs phenomenon in Fourier analysis, splines and wavelet approximations. — Springer Science & Business Media, 1998. — Vol. 446.

[4] Penkin Maxim A, Krylov Andrey S, Khvostikov Alexander V. Hybrid Method for Gibbs-Ringing Artifact Suppres- sion in Magnetic Resonance Images // Programming and Computer Software. — 2021. — Vol. 47, no. 3. — Pp. 207–214.

[5] Penkin Maxim A, Krylov Andrey S, Khvostikov Alexander V. Attention-based Convolutional Neural Network for MRI Gibbs-ringing Artifact Suppression // CEUR Workshop Proceedings. — 2020.

[6] How to choose adaptively parameters of image denoising methods? / Krylov Andrey, Penkin Maxim, Ma- maev Nikolay, Alexander Khvostikov // 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA) / IEEE. — 2019. — Pp. 1–6.

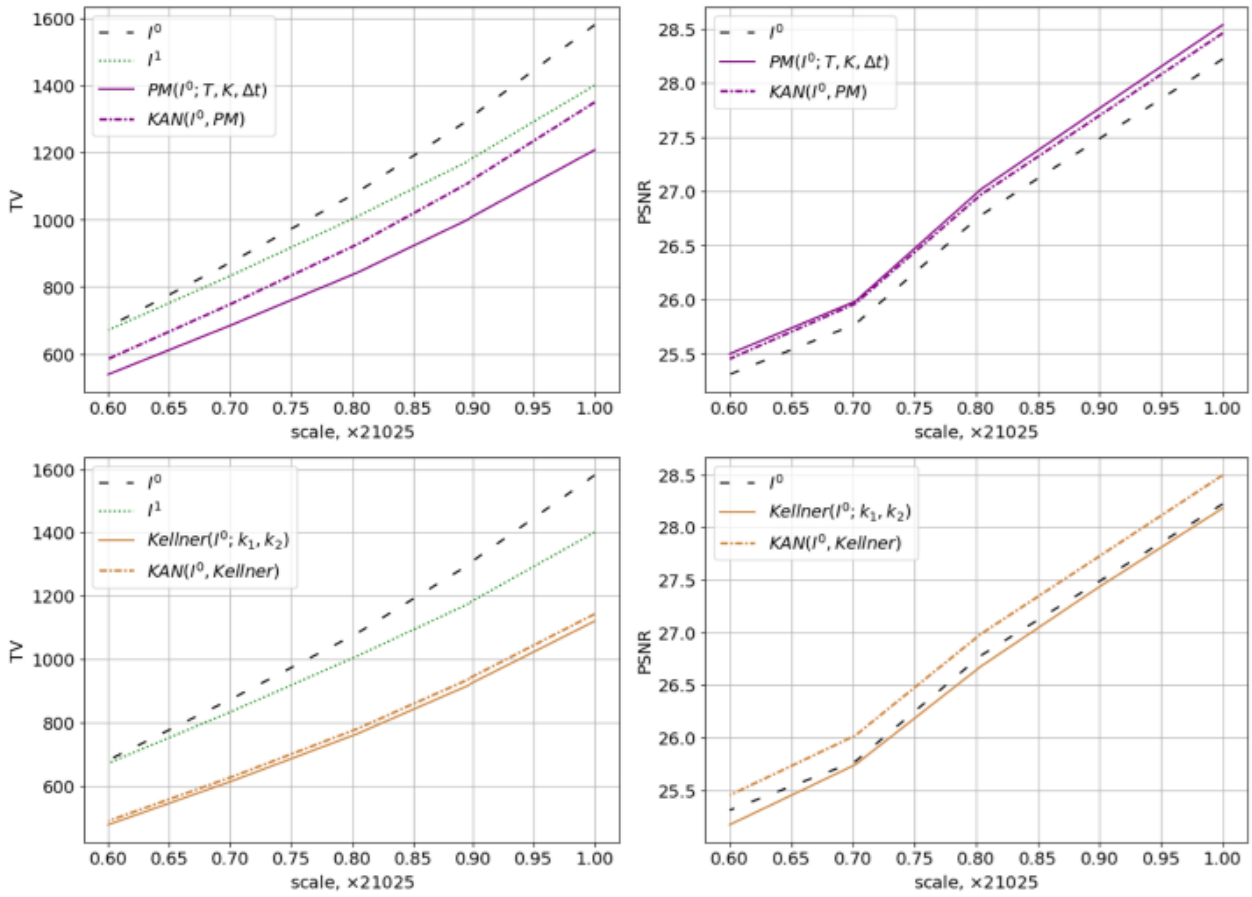


Figure 4. KANs quantitative evaluation results. PSNR is calculated relative to I^1 – high-resolution ground-truth images from IXIdataset, downsampled with nearest-neighbor interpolation on each level of the testing scale.

[7] HCformer: Hybrid CNN-Transformer for LDCT Image Denoising / Jinli Yuan, Feng Zhou, Zhitao Guo et al. // Journal of Digital Imaging. — 2023. — Vol. 36, no. 5. — Pp. 2290–2305.

[8] Blind Image Deblurring Using Multi-Task Convolutional Neural Network and Spatially Variant Recurrent Neural Network / VV Anisimovskiy, MA Penkin, EA Dorokhov et al. // Graphicon-Conference on Computer Graphics and Vision. — Vol. 33. — 2023. — Pp. 455–462.

[9] Rakibe Poonam, Patil Pramod D. Recent Methods on Medical Image Inpainting and Multi-task Learning Using Deep Learning Models // World Conference on Information Systems for Business Management / Springer. — 2023. — Pp. 403–413.

[10] Penkin Maksim A, Khvostikov Alexander V, Krylov Andrey S. Automated Method for Optimum Scale Search when Using Trained Models for Histological Image Analysis // Programming and Computer Software. — 2023. — Vol. 49, no. 3. — Pp. 172–177.

[11] KAN: Kolmogorov-Arnold Networks / Ziming Liu, Yixuan Wang, Sachin Vaidya et al. // arXiv preprint arXiv:2404.19756. — 2024.

[12] Convolutional Kolmogorov-Arnold Networks / Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, Santiago Pourteau // arXiv preprint arXiv:2406.13155. — 2024.

[13] Kolmogorov Andrei Nikolaevich. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. — American Mathematical Society, 1961.

[14] Perona Pietro, Shiota Takahiro, Malik Jitendra. Anisotropic diffusion // Geometry-driven diffusion in computer vision. — 1994. — Pp. 73–92.

[15] Namaki Neda, Eslahchi MR, Salehi Rezvan. The use of physics-informed neural network approach to image restoration via nonlinear PDE tools // Computers & Mathematics with Applications. — 2023. — Vol. 152. — Pp. 355–363.

[16] Penkin Maksim, Krylov Andrey. Medical Image Joint Deringing and Denoising using Fourier Neural Operator // Proceedings of the 2023 8th International Conference on Biomedical Imaging, Signal Processing. — 2023. — Pp. 40–45.
[17] Sprecher David A, Draghici Sorin. Space-filling curves and Kolmogorov superposition-based neural networks // Neural Networks. — 2002. — Vol. 15, no. 1. — Pp. 57–67.
[18] Ko¨ppen Mario. On the Training of a Kolmogorov Network // Artificial Neural Networks—ICANN 2002: International Conference Madrid, Spain, August 28–30, 2002 Proceedings 12 / Springer. — 2002. — Pp. 474–479.
[19] Lai Ming-Jun, Shen Zhaiming. The Kolmogorov Superposition Theorem can Break the Curse of Dimensionality When Approximating High Dimensional Functions // arXiv preprint arXiv:2112.09963. — 2021.
[20] Chen Tianping, Chen Hong. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems // IEEE transactions on neural networks. — 1995. — Vol. 6, no. 4. — Pp. 911–917.
[21] Diederik P Kingma. Adam: a method for stochastic optimization // (No Title). — 2014.
[22] Neural Network Quantization with AI Model Efficiency Toolkit (AIMET) / Sangeetha Siddegowda, Mar- ios Fournarakis, Markus Nagel et al. // arXiv preprint arXiv:2201.08442. — 2022.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-675-689

XAI и слабая аннотация в обеспечении точности и интерпретируемости вложенной классификации опухолей мозга

XAI AND WEAK ANNOTATION TOWARD BETTER ACCURACY AND INTERPRETABILITY IN THE NESTED CLASSIFICATION OF BRAIN TUMORS

Рехамния Валид[1], Турлапов Вадим Евгеньевич[1]
[1]Центр Искусственного Интеллекта, Университет им.Н.И.Лобачевского, Нижний Новгород, Россия

Rehamnia Walid[1], *Turlapov Vadim Evgenievich*[1]
[1]*Center for Artificial Intelligence, N.I.Lobachevsky University, Nizhny Novgorod, Russia*

Аннотация. Опухоли мозга ежегодно уносят тысячи жизней. Их ранняя диагностика, которая обычно опирается на МРТ, имеет здесь решающее значение. В последнее время было предложено много методов диагностики на базе искусственного интеллекта (ИИ), большинство из которых использует глубокое обучение (DL), имеющее и ряд недостатков: ненадежность; ресурсоемкость; необходимость разметки больших биомедицинских датасетов для обучения моделей DL. В статье предлагается алгоритм, который в значительной мере решает проблемы автоматической разметки и объяснимости диагноза, используя слабую аннотацию (WA) и объяснимый ИИ (XAI). Алгоритм включает предобработку изображений аннотированной области, обеспечивающую извлечение признаков опухоли, классификацию ее типа и кластеризацию, определяющую степень развития опухоли. Достигнутая точность бинарной классификации опухолей глиомы и менингиомы с использованием случайного леса (RF) составила 96%, и конкурентна среди лучших результатов публикаций 2023 года, полученных моделями DL. Затем, алгоритм обеспечивает объяснимую кластеризацию глиомы по степени злокачественности: на глиому низкой злокачественности (LGG) и высокой злокачественности (HGG), что важно для ранней диагностики. Алгоритм также предлагается для объяснении результатов DL, опираясь на результат XAI, и будет продолжать исследоваться в этом направлении.

Ключевые слова: здравоохранение, опухоль мозга, диагностика глиомы, МРТ, искусственный интеллект, объяснимый ИИ, слабая аннотация, машинное обучение, глубокое обучение

Abstract. Brain tumors are among the most serious and fatal cancers, claiming thousands of lives annually. Early diagnosis, which typically relies on MRI, is crucial for saving lives. Recently, many Artificial Intelligence (AI) techniques have been proposed to alleviate radiologists' workload and expedite diagnosis. Most of these methods utilize deep learning (DL), which, while powerful, can be less reliable and resource-intensive. A significant challenge is the labeling of large and frequently updated biomedical datasets for training DL models. This paper proposes an algorithm that addresses these challenges by leveraging weak annotation and eXplainable AI (XAI). The algorithm also incorporates various image processing steps, such as feature extraction and clustering, to determine both the type and degree of brain tumors. Our study demonstrates state-of-the-art performance, achieving 96% accuracy in the binary classification of Glioma and Meningioma tumors using Random Forest (RF). Additionally, the algorithm provides transparent clustering of Glioma into Low-Grade Glioma (LGG) and High-Grade Glioma (HGG) after type classification. The algorithm is also designed to offer XAI results for explaining DL outcomes and will continue to be developed in this area.

Keywords: healthcare, brain tumor, glioma diagnosis, MRI, Artificial Intelligence, eXplainable AI, Weak Annotation, Machine Learning, Deep Learning

I. Введение

Опухоли головного мозга являются опасными для жизни заболеваниями, которые ежегодно становятся причиной более 18 000 смертей, согласно исследованию Американского общества клинической онкологии (ASCO) [1]. Эти опухоли различаются в зависимости от клеток, в которых они развиваются, при этом наиболее распространенными являются глиома, менингиома и опухоли гипофиза. Глиомы возникают из глиальных клеток, которые обеспечивают поддержку мозга [2]. Менингиомы, составляющие 15-20% опухолей головного мозга, возникают из оболочек, окружающих мозг [3]. Опухоли гипофиза возникают в гипофизе, который регулирует гормоны, репродуктивный цикл и другие функции организма [4]. Всемирная организация

здравоохранения (WHO) классифицирует опухоли центральной нервной системы на четыре степени злокачественности: степень I и II — это низкозлокачественные или доброкачественные опухоли, а степень III и IV — это высокозлокачественные или злокачественные опухоли [5]. В некоторых исследованиях и клинических контекстах, глиомы в целом классифицируются на глиомы низкой степени злокачественности (LGG), которые включают степени I и II, и глиомы высокой степени злокачественности (HGG), которые включают степени III и IV.

Для анализа и диагностики опухолей головного мозга радиологи используют различные методы медицинской визуализации. Магнитно-резонансная томография (MPT, MRI) является одним из самых популярных и эффективных методов, предоставляя изображения высокого разрешения как в 2D, так и в 3D форматах без необходимости инвазивных процедур [6].

Слабая аннотация (Weak Annotation, WA) - это метод графического или текстового описания области локализации (как правило прямоугольной) патологии, обнаруженной специалистом по данным медицинского исследования. Хотя он менее точен, но его значительно проще и быстрее создавать, и можно автоматически извлечь из текстовых заключений диагностов [7]. Общие задачи диагностики опухолей головного мозга с использованием ИИ обобщены на рис. 1. К этим задачам относятся классификация опухолей по типам (например, глиома, менингиома, гипофизарная или неопухолевая), и их градация (например, LGG против HGG) или сегментация различных областей опухоли на изображении на уровне пикселей или суперпикселей .

Основные вклады этой статьи следующие:

1. Использование объяснимого ИИ (XAI) как методологической основы, избегание глубокого обучения (DL), при этом допуская потенциальную интеграцию результатов DL, которые могут быть объяснены результатами XAI и/или использованы для повышения точности.
2. Использование зоны слабой аннотации в качестве исходной разметки (вместо ручной разметки сегментации градации опухоли), которое позволяет заменить сегментацию на основе глубокого обучения (DL) кластеризацией методом k-средних, то есть машинным обучением (ML) без учителя, без потери точности. При необходимости, - повысить точность кластеризации/сегментации за счет расширения состава признаков высокого уровня.
3. Предварительная обработка данных для извлечения признаков высокого уровня, которые конкурируют с признаками, полученными сверточными нейронными сетями (CNN).

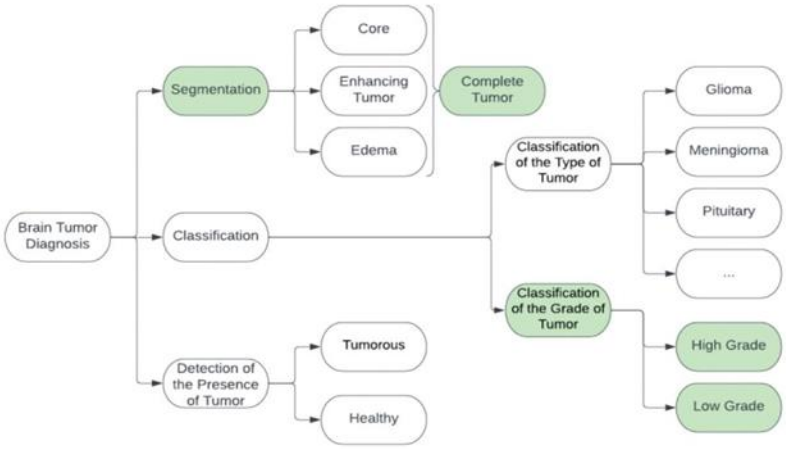


Рис. 1: Различные диагностические задачи, которые может выполнять ИИ [8]

После введения статья организована следующим образом: обзор публикаций; материалы и методы (датасет и алгоритм диагностики на основе WA); результаты и их обсуждение, заключение.

II. Обзор публикаций

1. Диагностика глиомы, используя ИИ

Недавние достижения в области ИИ показали многообещающий потенциал в области диагностики глиомы. Еще 12-13 лет назад исследования ранней диагностики изучали связь между глиомой и другими симптомами: так [9] изучали связь между временем начала и продолжительностью аллергии, а также использованием антигистаминных препаратов в развитии глиомы, в то время как [10] исследовали связь между

преддиагностическими уровнями антител IgE и риском глиомы, предполагая обратную связь между биомаркерами аллергии и риском глиомы.

Сегодня внимание исследований переключилось на решение более сложных задач, в отом числе: с обнаружения опухолей на их классификацию и сегментацию, используя компьютерное зрение и ИИ. Так [11] изучает эффективность методов глубокого обучения (DL) на гистологических изображениях, чтобы классифицировать различные подтипы глиомы. Метод достиг точности определения патча 86,5%, а по пациенту в целом точности 87,5%. Аналогично, авторы [12] разработали DL-модель EfficientNetB0 для классификации изображений МРТ по разным степеням глиомы (низкоуровневая глиома или высокоуровневая глиома). Модель протестирована на наборе данных BraTS-2019, продемонстрировав точность 98,8%, но игнорируя этап балансировки данных. Авторы [13] (2022) сосредоточились на поиске эффективных признаков и построили 102 признака на основе текстурного, морфологического и статистического анализа. Из них 20 были выбраны, путем удаления дубликатов, проверки коррелированности по Пирсону и рекурсивного исключения, и использованы для бинарной и мультиклассификации опухолей головного мозга. По мнению авторов, бинарная классификация глиома/менингиома, использующая модели долговременной краткосрочной памяти (LSTM), дает лучший результат по точности (97,7%).

Команда авторов [14] (2023) разработала быстрый интраоперационный метод молекулярной диагностики глиомы на основе ультразвуковых сигналов диапазона радиочастот.

Интересны также обзоры двух последних лет. Так обзор 2023 года [15] дает сравнительный анализ методов аугментации данных, нейронных сетей, трансферного обучения, и сетей на основе кросс-трансформеров, с целью оптимизации обнаружения и классификации опухолей головного мозга на данных МРТ с использованием ИИ. Рассмотрены классы опухолей: глиома, менингиома и опухоль гипофиза. Наилучший результат в классификации (97,2%) показала модель InceptionResNetV2.

В обзоре 2024 года [16] рассмотрено 70 исследований, в которых для анализа гистопатологических изображений глиом использован ИИ. Большинство из них использовали модели сверточной нейронной сети (CNN) на датасете The Cancer Genome Atlas (TCGA) для различных задач диагностики, таких как классификация, прогнозирование молекулярных маркеров и прогнозирование выживаемости. Исследовались гистопатологические изображения срезов тканей диффузных глиом взрослого типа, окрашенные гематоксилином и эозином.

Авторы [17] (2023) на датасете Brats2020 применили CNN для бинарной классификации глиом в комбинации с методами, не основанными на глубоком обучении, чтобы преодолеть проблему высокой дисперсии. Для этого использован адаптивный порог, гистограмма, морфологическая операция и адаптивный алгоритм слияния признаков. В классификации степени злокачественности глиомы показан подход с использованием CNN и адаптивной сегментации.

Подраздел в целом демонстрирует успехи применения DL в классификации и сегментации опухолей мозга, несмотря на то, что DL-модели являются моделями с высокой дисперсией, и любое изменение в обучающих данных существенно влияет на истинность результата. Тем не менее, разнообразие моделей и их возможности постоянно увеличиваются, а сами модели остаются современными в задаче диагностики опухолей мозга.

2. Диагностика глиомы, используя XAI

Прежде всего остановимся на направлении разработки методов направленных на повышение интерпретируемости результатов моделей глубокого обучения (DL) в диагностике глиомы как для врачей, так и для специалистов по данным. Интерпретируемость очень часто близка к объяснимости, являющейся необходимым условием объяснимого искусственного интеллекта (XAI).

В обзоре [18] (2022) рассматривается более 200 статей, использующих различные методы повышения интерпретируемости результатов DL. Сегодня собран арсенал таких методов, наиболее популярными из которых являются: Grad-CAM, DeepLIFT, LIME, SHARP. Система методов повышения интерпретируемости несомненно полезна для легитимизации применения DL в медицине.

В публикации [19] (2022) проблема объяснимости решается путем введения фреймворка «NeuroXAI», который использует 7 современных методов объяснения, предоставляя визуальные карты. Фреймворк исследован как для классификации, так и для сегментации изображений МРТ, и продемонстрировал потенциал для визуализации внутреннего потока при сегментации средствами DL.

Два года спустя тот же коллектив авторов разработал в [20] (2024) новый гибридный метод под названием «TransXAI», который обеспечивает после локализации опухоли ее объяснение путем создания тепловой карты аналогичной Grad-CAM и понятной хирургам. На датасете Brats 2019 достигнута средняя точность мультимодальной сегментации изображения в режиме T1Gd, по метрике Dice-scores, - 0.88, 0.78, и 0.75 для областей опухоли necrotic, enhancing и edema соответственно. Планируется развитие работы на 3D изображения

с увеличением точности сегментации. Авторы [21] (2023) исследовали 2 CNN-модели, достигнув точности 96% за счет аугментации данных и оптимизации гиперпараметров. Также «под капотом» использованы методы повышения интепретируемости результата DL. Код 3 предыдущих исследований находится в открытом доступе на GitHub.

Авторы [22] (2023) построили свой объясняющий DL-решения фреймворк «Explainable Brain Tumor Detection Framework for MRI Analysis», которая строит сегментацию, используя 3D U-Net, классификацию через VGG и RepOptimizer, а объяснимость через процесс CAM как постфактум-технику. Исследования выполнены на датасете BraTS2018, и показали результаты на уровне state-of-the-art. Похожее решение показано в статье [23], но для сегментации опухоли почки и на данных КТ.

Системный обзор [24] (2021) исследовал проблемы и перспективы новых подходов на основе «чистого» XAI в системах поддержки клинических решений, базированных на классических методах машинного обучения с использованием деревьев решений. Под чистым XAI понимается ИИ, который в медицине должен полностью соответствовать требованиям к XAI со стороны стандартов. Статья [25] (2023) рассматривает следующие 5 рекомендаций соответствия XAI в применении к DL-моделям: понятность, клиническая релевантность, правдивость, информативная корректность и вычислительная эффективность (рис. 2). По данным статьи эти рекомендации не прошли 60 из наиболее популярных тепловых карт (несоответствие 3 и 4 рекомендациям).

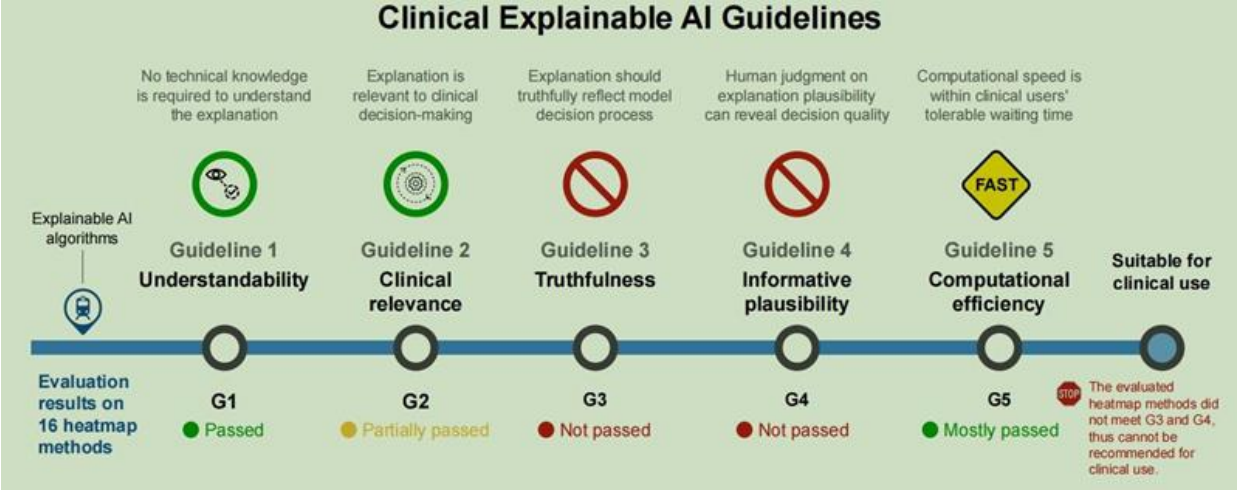


Рис. 2: Рекомендации (GNo), составляющие клиническую объяснимость решений XAI [25]

3. *Диагностика глиомы, используя слабую аннотацию* (*Weak Annotation,WA*)

По итогам предыдущего подраздела представляется очевидным вывод, что дополнение возможностей DL-моделей средствами повышения интерпретируемости продуктивно для диагностики опухолей мозга, но необходимая медицине 100% объяснимость может быть достигнута только применением чистого XAI. Это делает интересным для медицины использование XAI-диагностики в паре с интерпретируемым DL, результаты которого могут получить XAI-объяснение.

Хотя методы машинного и глубокого обучения показывают многообещающие результаты в диагностике глиомы, проблема подготовки данных маркированных в достаточной степени все еще остается сложной, особенно если это точная площадная или контурная маркировка для обучения сегментации. Слабая аннотация/разметка в этом случае является эффективным решением, если после нее получается построить кластеризацию аннотированной области, которая будет давать результаты близкие к сегментации. Так в статье [26] (2020) предложена слабо контролируемая структура глубокого обучения под названием «ADGNET», которая показала превосходные результаты на основе следующего вектора метрик: Kappa, Sen, Spe, Pr, Acc и F1-score.

Авторы [27] (2023), вдохновленные успехом обучения на основе NLP-подсказок (natural language processing, обработки естественного языка), предложили подход на основе подсказок «PL-BTS» для сегментации без учителя. Вместо обучения модели на аннотированных данных, подход стремится ответить на вопросы о патологичности вокселя входного изображения. Т.е., обусловлено ли его изображение опухолеподобной гипер/гипоинтенсивностью, путем проведения обучения на правилах/признаках созданных вручную. Подход продемонстрировал замечательные результаты как на внутренних, так и на общедоступных датасетах. Для имитации сложной структуры человеческого мозга, которая обеспечивает точные результаты с очень небольшим

руководством, в [28] (2022) предложена модель ”vMFNet”, на вероятности vMF для прогнозирования маски сегментации в различных частях изображения, и самое главное, что ядра vMF можно изучить даже из немаркированных данных, реконструируя их снова для формирования исходных данных. Результаты исследования улучшили производительность обобщения на 2 бенчмарках, особенно с WA-данными.

Аналогично, в [29] (2024) приняли смешанный метод надзора «vMFNet» для получения устойчивых композиционных представлений посредством комбинации неконтролируемого обучения и слабого надзора, в то же время включив неполные метки патологии на уровне пикселей. После проведения экспериментов с набором данных Brats автор заявил об эффективности vMFNet в случае большого количества слабо аннотированных данных, сопровождаемых небольшим количеством полностью аннотированных.

Авторы [30] (2022) использовали WA, созданную путем применения методов NLP в радиологических отчетах для обнаружения изменений глиомы. Подчеркнуто ее сочетание с индуктивным переносом обучения. Методология авторов продемонстрировала лучшие результаты и общность модели на изображениях МРТ.

В статье [31] (2024) разработана новая стратегия аннотации (простое перетаскивание), затем проверил ее с помощью нового слабо контролирующего метода на основе алгоритма водораздела. Эта аннотация достигла сопоставимой производительности с подробными попиксельными аннотациями. Авторы этого исследования обнаружили, что в ситуации ограниченных ресурсов использование слабых аннотаций в различных группах пациентов может привести к более надежной производительности модели по сравнению с распределением попиксельных аннотаций для небольшого подмножества изображений.

В 2022 [32] на основе генерации суперпикселей и кластеризации суперпикселей в качестве слабого аннотирования предложена модель использования методов с учителем для задачи сегментации мозга. Получен со средним значением точности сегментации 95% на двухмерных срезах МРТ набора данных Brats 2020.

Чтобы решить проблемы, с которыми сталкиваются врачи, диагностирующие PPGL на КТ-снимках, которые состоят из вариации размера и времени, необходимого для рутинного отслеживания и измерения, статья [33] (2024) представила метод обнаружения «PPGL» с помощью задачи прокси-сегментации, где он обучил NNUnet с полным разрешением на 2D и 3D слабо аннотированных данных и получил точность 70% и чувствительность 64,1% с помощью теста на 53-х КТ-исследованиях.

Статья [34] (2023) провела сравнение между различными стратегиями активного обучения на реальных наборах данных, а затем предложила структуру, которая минимизирует объем необходимой аннотации, сохраняя точность.

Вместе эти исследования иллюстрируют, как включение слабых аннотаций может повысить как эффективность, так и точность моделей, используемых для обнаружения и классификации опухолей головного мозга.

III. Материалы и методы

1. Использованные датасеты

В этом исследовании мы использовали общедоступный датасет [13], состоящий из 2 типов опухолей головного мозга: глиомы и менингиомы. Эти данные размечены в формате YOLO руками эксперта-рентгенолога с 10-летним опытом, где были определены тип и ось ограничивающего прямоугольника каждого изображения опухоли (в отдельных текстовых файлах). Всего имеется 1420 изображений, где:

• Глиома 602: 482 для обучения и 120 для тестирования

• Менингиома 818: 654 для обучения и 164 для тестирования

На рис. 3 ниже показаны примеры, извлеченные из набора данных (до и после применения ограничивающего прямоугольника и метки), где первая строка представляет тип глиомы, а вторая строка представляет тип менингиомы. Как можно видеть, изображения были сделаны под разными углами.

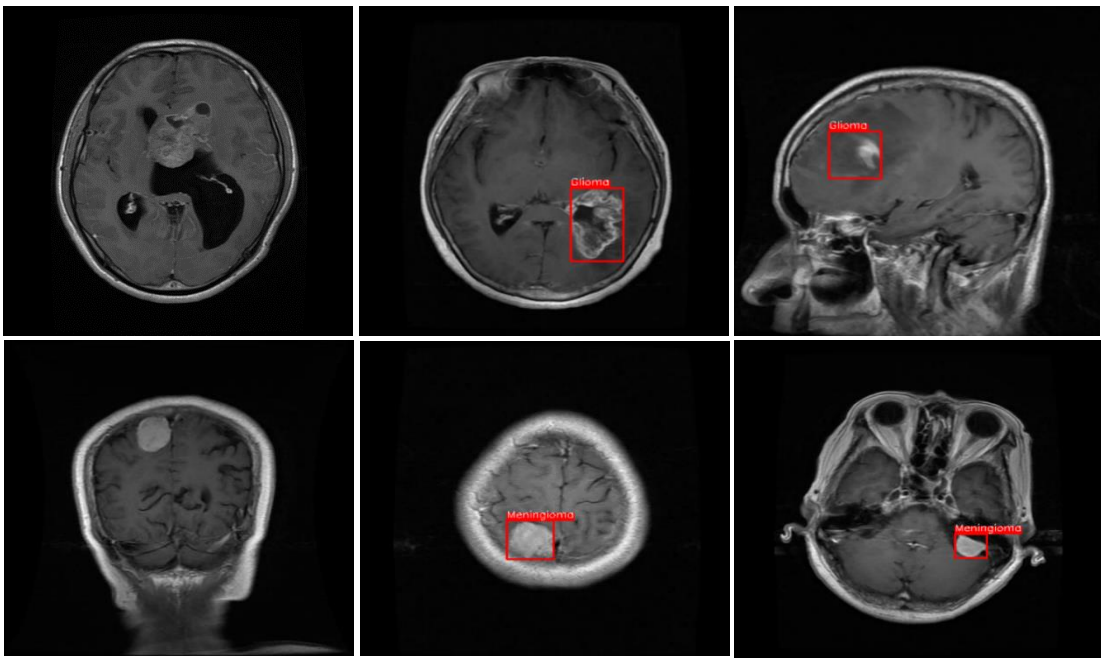


Рис. 3: Пример изображений из датасета (до и после применения ограничивающей рамки)

2. *Алгоритм*

Предлагаемый алгоритм состоит из множества шагов, каждый из которых будет объяснен далее:

Кластеризация k-means. Для кластеризации пикселей области интереса (ROI), которая предопределена маркировкой ограничивающего прямоугольника были применены как k-средние, так и нечеткие-средние. Результаты обоих алгоритмов кластеризации были почти идентичны, поэтому алгоритм использовал k-средних из-за его простоты и производительности, рис. 4 демонстрирует результат кластеризации с различными значениями k.

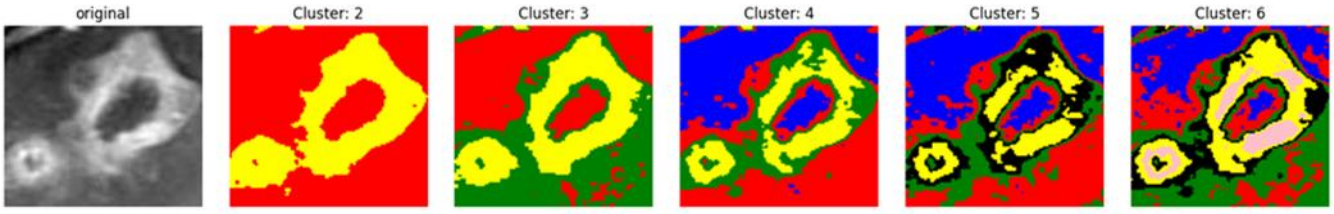


Рис. 3: Результат k-means кластеризации, используя величины k от 2 до 6.

Custom Preprocessing. Включает удаление фона, очистку изображения от шума путем удаления выбросов…, выход этого шага представляет собой сегментированное на уровне пикселей изображение со следующими 3 метками:

- 0: Фон
- 1: Расширенная опухоль (ET)
- 2: Некротическое ядро опухоли (NCR)

Для окончательного построения маски использовались 2 формата масок. Например, на рис. 5 ниже показаны 4 изображения. Они, слева направо соответственно: исходное изображение с ограничивающей рамкой слабой аннотации; маска опухоли (области ET+NCR); маска области NCR опухоли отдельно (формат 1); маска области интереса (ROI) для опухоли после объединения ET+NCR (формат 2). Установлено, что этих 3-х компонентов PCA достаточно, чтобы покрыть более 85% дисперсии датасета.

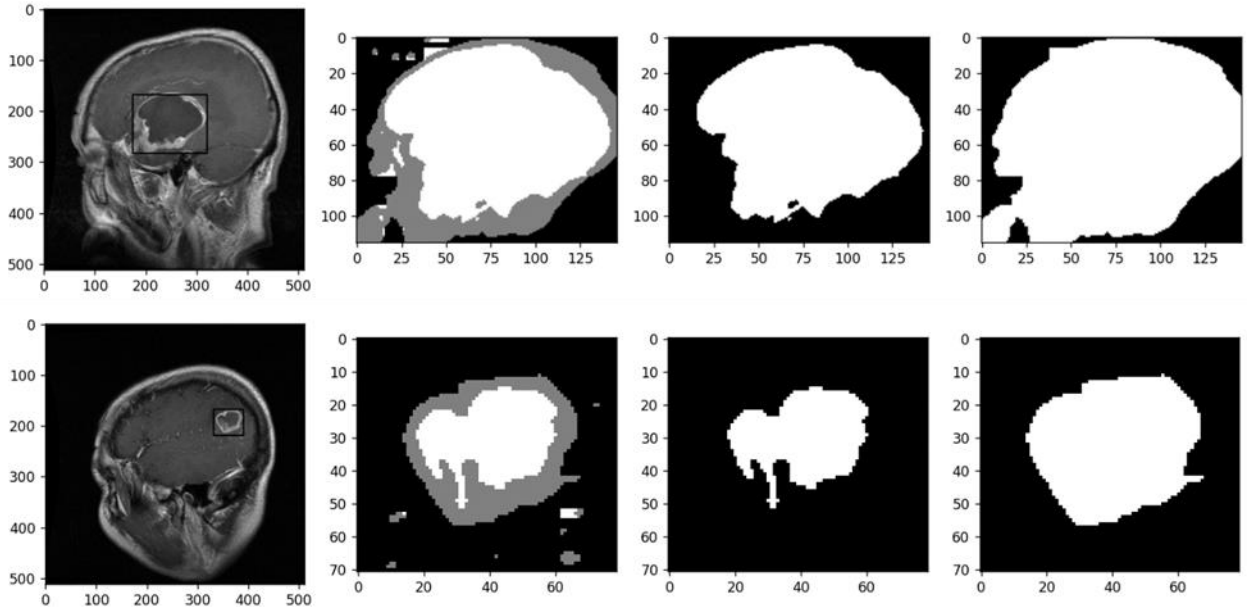


Рис. 4: Рис. 5: Маски слабой аннотации (слева, черный прямоугольник) и их 3 кластеризации (справа)

Feature selection/extraction. Извлечение количественных признаков, связанных с формой, таких как площадь, периметр, и т.д. (эти признаки стали доступны благодаря маске, сгенерированной на этапе слабой аннотации). помимо набора признаков из библиотеки «pyradiomics» Python, которые извлекают статистику первого порядка (среднее значение, дисперсия, …), текстурные признаки, в основном матрица совместного появления уровней серого (GLCM) описывает частоту пар интенсивностей вокселей, встречающихся вместе на определенном расстоянии и в определенной ориентации. Чтобы получить наилучший набор этих признаков, были опробованы различные методы, включая анализ главных компонентов (PCA), быстрый анализ независимых компонентов (ICA), t-распределенное стохастическое встраивание соседей (t-SNE), локально-линейное встраивание, рекурсивное исключение признаков (RFE) … были протестированы, и количество извлеченных признаков для каждого из этих методов было различным.

Classification. Классификация проводилась с использованием простых моделей ML, относящихся к чистому XAI, в том числе: Decision Trees, Random Forests, и Light Gradient boosting (LGB).

Grading clustering. На этом этапе алгоритм использует область интереса (опухоль, выделенную рамкой) для извлечения дискриминантных признаков, которые помогают различать LGG и HGG. Эти признаки являются подмножеством признаков, описанных выше (только форма и признаки GLCM).

eXplainable AI (XAI). Включает объяснение и интерпретацию решений, принятых алгоритмами классификации (для определения типа) и кластеризации (для определения степени злокачественности) с использованием диаграмм визуализации, обсуждения анализа…, см. раздел результатов и обсуждения ниже. Приведенная ниже диаграмма на рис. 6 суммирует различные части нашего алгоритма.

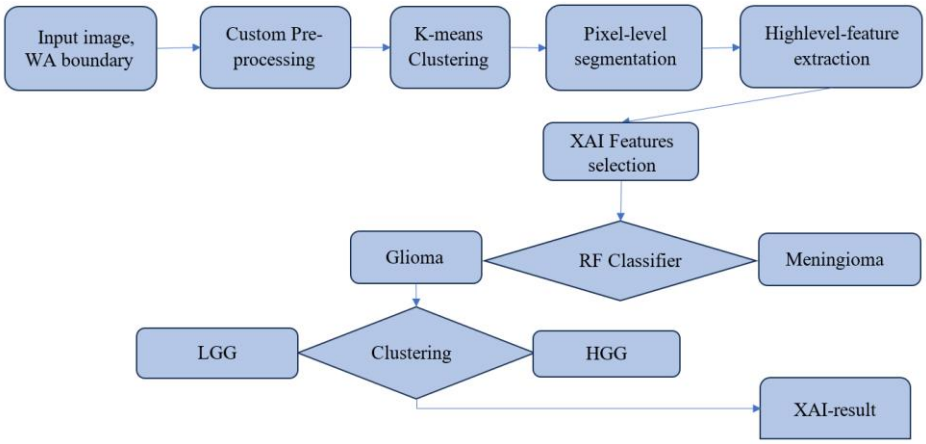


Рис. 6: Схема предложенного алгоритма

IV. Результаты эксперимента и их обсуждение

1. Определение и релевантность характеристик формы

ТАБЛИЦА 2
ОПРЕДЕЛЕНИЕ ПРИЗНАКОВ ФОРМЫ

Feature	Definition
Area	Площадь - общее количество пикселей в области опухоли на ее изображении.
Perimeter	Периметр - длина границы области опухоли: число пикселей, принадлежащих границе опухоли.
Eccentricity	Эксцентриситет - мера того, насколько форма опухоли отклоняется от идеальной окружности. Определяется как отношение расстояния между фокусами эллипса, вписанного в опухоль, к длине ее большой оси. Диапазон изменения: от 0 до 1. • 0: Указывает на идеальную окружность. • 1: Указывает на сильно вытянутую форму, близкую к отрезку прямой
Major Axis Length	Длина большего диаметра эллипса
Minor Axis Length	Длина меньшего диаметра эллипса

Характеристики площади и периметра, из приведенной выше таблицы 1 напрямую связаны с размером и сложностью границ опухоли, при этом HGG обычно имеют большую площадь и более неровные границы, с другой стороны, эксцентриситет помогает понять форму опухоли, при этом более вытянутые или неровные формы потенциально указывают на HGG, и, наконец, длина большой и малой оси описывают форму и ориентацию опухоли, при этом более длинные оси или более сложные формы потенциально указывают на опухоль более высокой степени злокачественности.

2. Эксперимент

Исследование было проведено на компьютере с ОС Windows 11, с 16 GB RAM, графическим процессором NVIDIA GeForce GTX 1060 и процессором Intel(R) Core (TM) i7-8750H. Язык программирования: Python 3.x. Характеристика использованного датасета дана выше. Напомним, что из 602 изображений глиомы для обучения и тестирования использованы соответственно 482 и 120, а из 818 изображений менингиомы - 654 и 164.

В ходе эксперимента определены параметры алгоритма, обеспечившие наилучшую точность классификации 96% (перечислены ниже):

1) Модель классификации: случайный лес со 100 деревьями (для обеспечения надежности модели путем балансировки снижения дисперсии и повышениевычислительной эффективности, обеспечивая стабильный и надежный ансамбль) и максимальной глубиной 16 (для захвата сложных паттернов без переобучения, что делает классификацию эффективной для различения глиомы и менингиомы, сохраняя при этом обобщение на невидимые данные).
2) Для фаз классификации и кластеризации выбрано 10 признаков формы.
3) Начальные центроиды алгоритма k-средних выбраны с использованием метода инициализации k-средних++, который улучшает скорость сходимости и создает более стабильные кластеры путем разнесения начального положения центров кластеров.
4) В фазе кластеризации (k-средних) экспериментально выбран k=2, поскольку для него границы опухоли оказались хорошо очерчены.
5) Первый извлеченный параметр (формат) маски опухоли имел наивысшую точность.
6) С помощью PCA были извлечены 7 признаков.
7) Отбор 6 наиболее эффективных признаков выполнен с помощью метода Recursive Feature Elimination (RFE). Ими оказались: средняя интенсивность; округлость ET; минимальная интенсивность; площадь ET; различие GLCM dissimilarity; GLCM correlation. Рис. 7 показывает их относительную важность.
8) Все признаки задач классификации и кластеризации перед этапом обучения были стандартизированы путем вычитания среднего значения и нормировки до единичной дисперсии
9) Веса глиомы, менингиомы {8, 2}

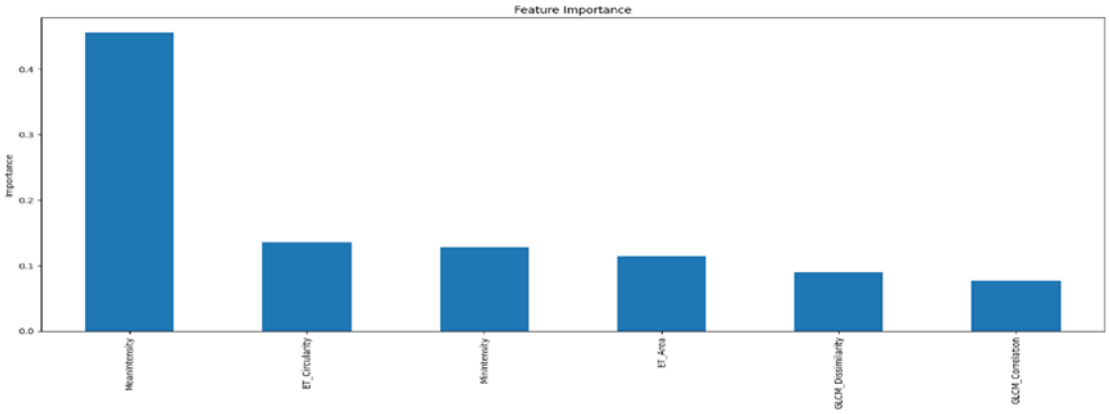


Рис. 7: Важность признаков определенная на основе RFE

Экспериментально оценено, что сокращение площади фона в области интереса (ROI) при переходе от слабой аннотации прямоугольником к ROI, определяемой извлеченнной маской, повышает точность классификации примерно на 5–6 %. После этого, чтобы кластеризовать изображения глиомы на LGG и HGG, алгоритм извлек 10 дискриминантных признаков и применил алгоритм k-средних с k=2.

Чтобы сделать результаты более надежными и достоверными, были предприняты некоторые шаги XAI для объяснения и интерпретации результатов как классификатора, так и кластеризации:

a. Классификация:

Важность признаков: рассчитывается как уменьшение доли узла, взвешенное по вероятности достижения этого узла. Чем выше значение, тем важнее признак. Наиболее важные признаки будут выше в дереве. Один признак может использоваться в разных ветвях дерева, тогда важность признака - это его общий вклад в уменьшение примеси. Согласно рис. 8, классификатор подчеркивает, что средняя интенсивность, округлость ET (Расширенная опухоль) и площадь ET являются наиболее важными признаками для различения глиом и менингиом. Средняя интенсивность и минимальная интенсивность отражают плотность и гетерогенность опухоли, при этом глиомы часто демонстрируют большую изменчивость. Округлость и площадь ET помогают отличать обычно округлые менингиомы от нерегулярных форм глиом. GLCM dissimilarity и GLCM correlation дополнительно подтверждают это, фиксируя вариации текстуры. Эти признаки согласуются с известными клиническими характеристиками, повышая интерпретируемость и надежность модели при классификации опухолей.

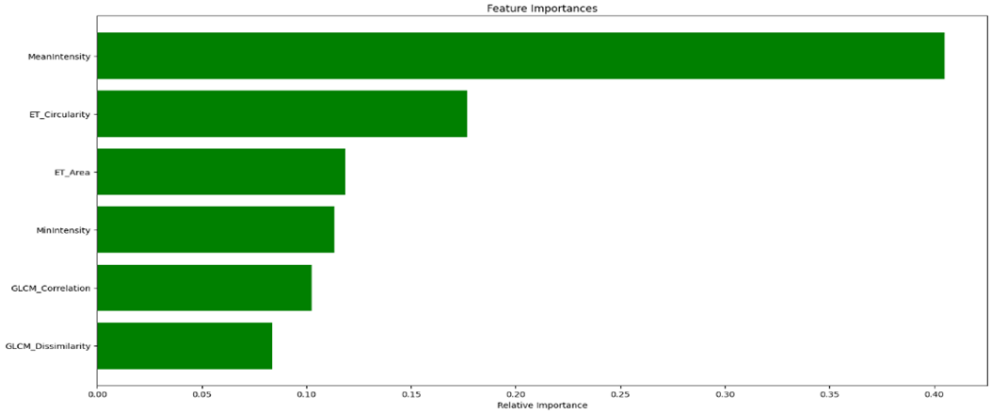


Рис. 5: Важность признаков в алгоритме классификации

Визуализация дерева: Для упрощения визуализации и анализа схема RF подогнана к дереву решений. Визуализация дерева решений на рис. 9 показывает, как классификатор Random Forest различает глиому и менингиому, фокусируясь на таких признаках, как средняя интенсивность, корреляция GLCM, округлость ET (Расширенная опухоль) и площадь ET. Модель сначала разделяется по средней интенсивности, подчеркивая ее важность в дифференциации более плотных менингиом от более неоднородных глиом. Последующие разделения по корреляции GLCM и округлости ET уточняют это различие, причем глиомы обычно демонстрируют большую изменчивость текстуры и нерегулярные формы, в то время как менингиомы кажутся более округлыми и однородными. Структура дерева с четкими путями принятия решений и важностью признаков делает прогнозы модели объяснимыми, и саму модель объяснимой (eXplainable). Это имеет решающее значение для клинического

доверия в медицине. Совмещая фокус модели с заданными клиническими характеристиками опухолей, дерево решений поддерживает необходимую объяснимость решения и модели ИИ. Деревья решений и их визуализация служат одним из важнейших инструментов в Explainable AI (XAI).

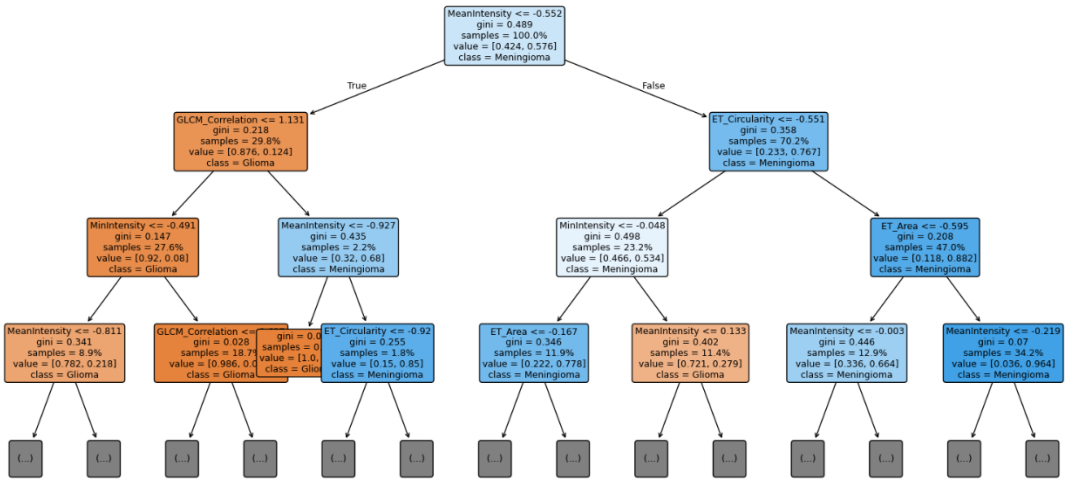


Рисунок 6. Визуализация дерева решений

b. Кластеризация:

Визуализация данных. При визуализации в 2D (рис.10) результата применения PCA видно, что для разделения данных без заметного перекрытия и снижения размерности достаточно только 2 признаков (компонентов PCA). Оказалось, что 2 компонента покрывают более 95% дисперсии данных. Точки кластера 0 (слева, синий цвет) имеют низкую дисперсию и меньшие значения Component1 по сравнению с кластером 1 (справа, зеленый цвет), который имеет более высокую дисперсию и значения Component1. Это означает, что использованные при кластеризации признаки были достаточными (дискриминантными) для разделения классов HGG (кластер1 имеет бо́льшие значения признаков и бо́льшую дисперсию) и LGG (кластер 0 имеет меньшие значения и более однороден).

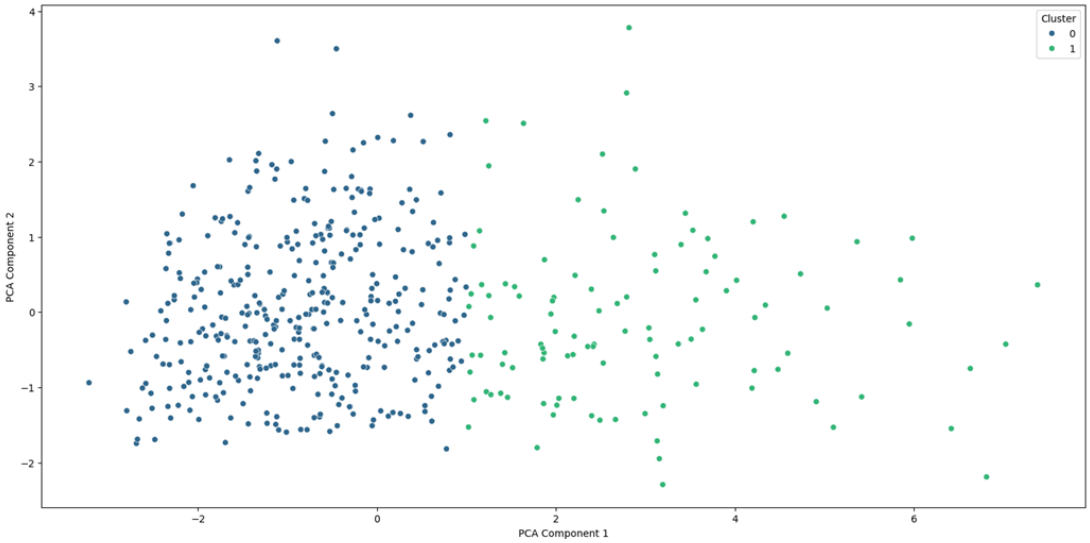


Рис. 10: Точки данных кластеризации оценок с использованием двух компонентов PCA

Агрегация. На рис.11 показаны в виде диаграмм результаты анализа кластеров по каждому из признаков (среднее значение, стандартное отклонение и дисперсию). Из диаграмм видно, что кластер 0 представляет опухоли меньшего размера (меньшая площадь, периметр и длины осей), тогда как кластер 1 представляет опухоль большего размера с менее округлой формой, поэтому наиболее вероятно, что кластер 0 относится к LGG, а кластер 1 — к HGG.

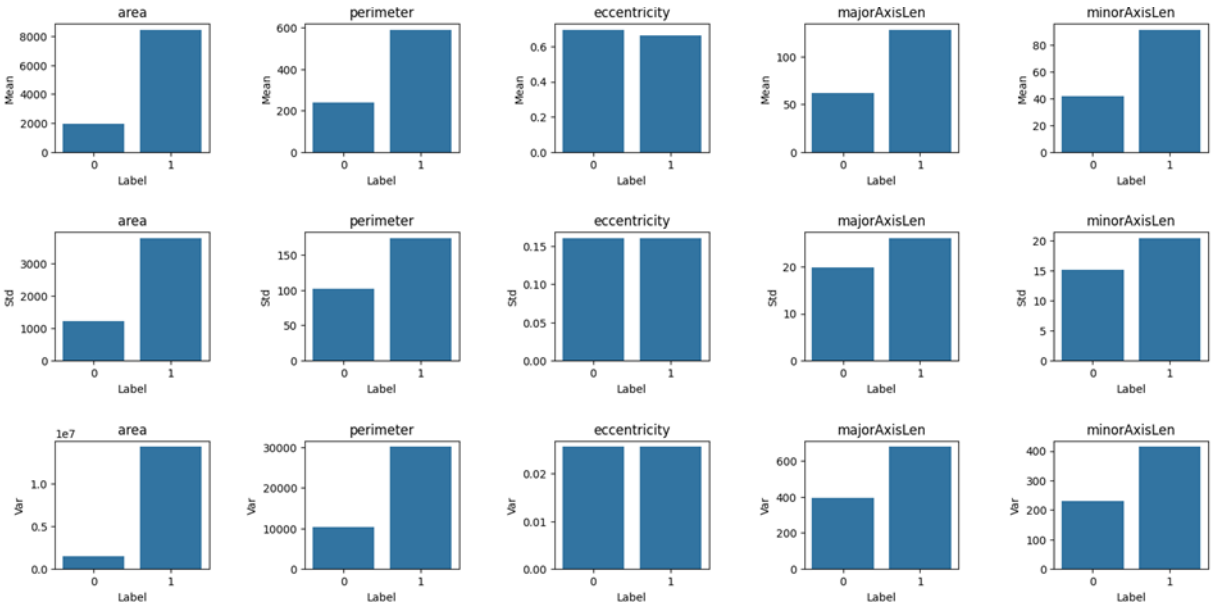


Рис. 11: Диаграммы агрегации признаков кластеризации и классификации

Важность. На этапе кластеризации необходимо определить вклад (важность) признаков. Для этого применен PCA с 3 компонентами, которых оказалось достаточно, чтобы покрыть более 98% дисперсии данных. На основе вклада исходных признаков в эти компоненты построена таблица важности признаков (табл.2). Цель этой таблицы — определить и ранжировать вклад каждого признака в их общую корреляцию и кластеризацию.

ТАБЛИЦА 3
ВАЖНОСТЬ ПРИЗНАКОВ В КЛАСТЕРИЗАЦИИ СТЕПЕНИ ЗЛОКАЧЕСТВЕННОСТИ ОПУХОЛЕЙ

Rank	Feature	Importance
0	Eccentricity	0.92
1	Perimeter	0.88
2	Area	0.57
3	Minor Axis Length	0.31
4	Major Axis Length	0.30

Based on the aforementioned analysis and interpretation, the clustering output become transparent and trustfully where the group 0 represents LGG while group 1 represents HGG.

3. *Обсуждение результатов*

a. ***Задача классификации***:

ТАБЛИЦА 4
STATE-OF-THE-ART ИССЛЕДОВАНИЯ, ИХ МЕТОДЫ И ДОСТИГНУТАЯ ТОЧНОСТЬ

Year/Ref	Dataset	Task	Method	Accuracy
2021 [11]	300 патчей изображений 2048 × 1536 pix при 40x увеличении	Классификация 5 типов Глиомы	Custom DenseNet	87%
2022 [12]	BraTS-2019	Классификация Глиомы: LGG, HGG	EfficientNetB0	98.8%
2022 [13]	3064 T1-weighted images	Классификация: Glioma vs Meningioma	LSTM	97.7%
2023 [14]	3064 T1-weighted images	Классификация: Glioma, Meningioma, Pituitary	InceptionResNetV2	97.2%

2023 [17]	BraTS 2020	Классификация Глиомы: LGG, HGG	CNN + Transfer Learning	99%
2023 [17]	BraTS 2020	Сегментация опухолей	Adaptative thresholding mechanism	70.58%
2023 [21]	Br35H (3000 images) + Классификация опухолей мозга (MRI) (2870 images)	Классификация опухолей мозга	CNN models	96%
2023 [22]	BraTS2018	Классификация: LGG, HGG	VGG + RepOptimizer	95.46%
Данное исследование	1420 2D изображений (WA с прямоуг. рамкой и меткой, Рис.3)	Классификация: Glioma vs Meningioma [13]	Random Forests	96%

Согласно Таблице 3, в которой сравниваются некоторые из вышеупомянутых современных исследований с данным исследованием:

- Хотя исследование [11] использовало глубокое обучение и больший набор данных, они получили точность, которая на 9% меньше по сравнению с представленным исследованием. Эта большая разница может относиться к разнице в задаче (бинарная или мультиклассификация) и природе базы данных.
- При сравнении этого исследования с [12] [13] [14] и [17] (часть классификации) настоящий результат можно считать удовлетворительным и многообещающим, поскольку с использованием простой модели (случайный лес) и слабо аннотированных данных мы достигаем немного меньшей точности примерно на 1-3% по сравнению с теми исследованиями, которые используют преимущества глубокого обучения, трансферного обучения и качества данных (полностью аннотированных). Чтобы увидеть, насколько отсутствие DL может повлиять на точность, см. две строки «2023 [17]», которые представляют низкую точность сегментации по сравнению с ее современным состоянием.
- Наконец, это исследование получает преимущество по сравнению с [21] и [22], которые используют CNN и VGG соответственно на больших и полностью аннотированных данных, но уступают нашему результату.

b. ***Слабая Аннотация (WA)***:

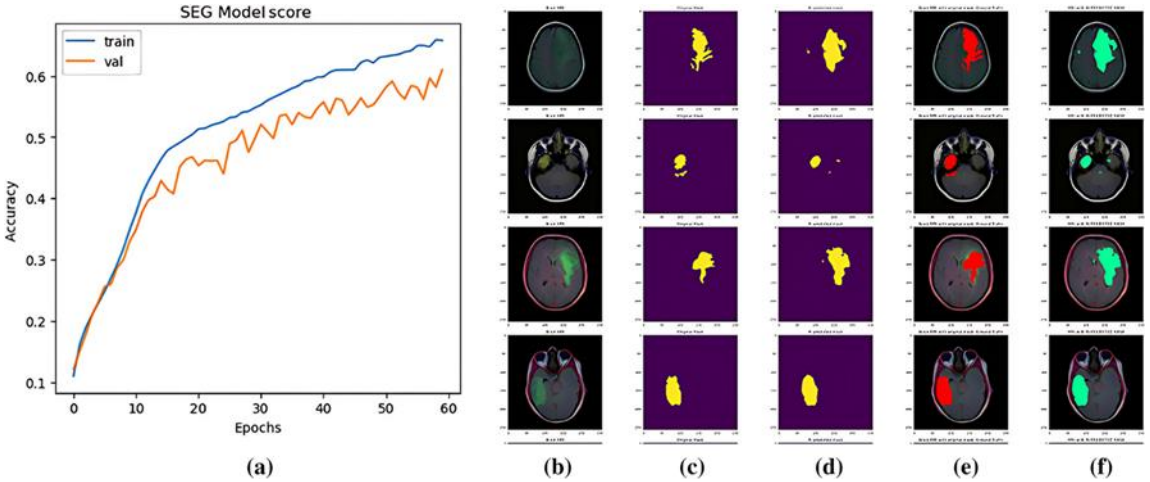


Рис.12: Сегментация Vgg19Unet - график обучения, (b) входные изображения, (c) истинные данные, (d)прогнозируемое значение, (e) маска истинных данных, (f) маска прогнозируемого значения [37]

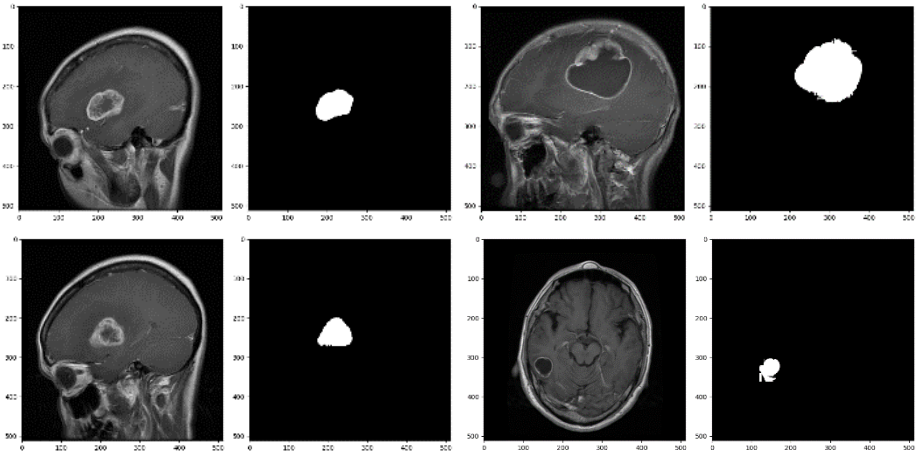


Рис. 13: Результаты после слабой аннотации (WA) в парах: входное изображение (слева); опухоль (справа).

Следует заметить, что датасет, используемый в этом исследовании не содержит разметку сегментации опухолей специалистами. Соответственно нет обучения на такой сегментации и нет тестовой части датасета. Поэтому выходные маски опухолей сравниваются с опухолью на входном изображении визуально.

Сравнивая выходные данные сегментации (на рис.12), которые представляют собой результат недавнего исследования (VGG19Unet [37]) с использованием новой DL-модели на полностью аннотированном наборе данных, с результатом (рис.13) по слабой аннотации, мы обнаруживаем, что слабая аннотация в этом исследовании обеспечивает менее шумные маски, которые в основном идентичны фактическим формам опухоли и по качеству конкурируют с выходными данными новейших алгоритмов глубокого обучения. Но ошибки не исключены и 4-е изображение на рисунке 13 специально показывает пример частичного выхода за зону опухоли.

c. Ограничения:

Предложенный алгоритм классифицирует изображения МРТ по трем категориям: LGG, HGG и менингиома. Понятно, однако, что точность результата существенно зависит от точности разметки области и доли фона в указанной области. В случае только словесного описания зоны опухоли в заключении диагноста, без указания координат ограничивающего прямоугольника, разумно включить предобработку, которая позволит уменьшить долю фона. Альтернативным решением может быть совместное использование данного объяснимого алгоритма XAI и модели(ей) глубокого обучения, который мог бы получать объяснение XAI-результатами и постоянно дообучаться на них в случае успешности прогноза.

V. Заключение

В данном исследовании представлена алгоритм классификации и сегментации глиомы мозга, стартующий со слабой аннотации (WA) и сегментации опухоли, заданной прямоугольной рамкой в какой-либо моде МРТ-изображения. Предложена структура объяснимости, предусматривающая предобработку входных изображений в зоне разметки, которая должна экстрагировать эффективные признаки классификации, кластеризации и сегментации опухоли. Затем обеспечить обучаемость модели классификации и XAI-свойства алгоритма: объяснимость входных данных и классификации опухолей по различным типам (глиома, менингиома и степени злокачественности); объяснимость сегментации на основе понятных признаков кластеризации. В значительной мере это в нашем алгоритме достигнуто, обеспечена сегментация изображения со слабой разметкой изображений МРТ с точностью 96% с использованием случайного леса, конкурентной по сравнению с state-of-the-art результатами.

Сосредоточившись на ключевых характеристиках, таких как меры формы и текстуры, и используя методы XAI, такие как визуализация дерева решений и важность признаков, модель обеспечивает высокую интерпретируемость и согласуется с клиническими характеристиками. В продолжении, планируется расширить методы слабой аннотации, найти еще более эффективные признаки кластеризации, чтобы довести показатель точности до максимальных значений и иссследовать парное применение XAI и DL моделей.

Благодарности

Работа выполнена при поддержке Аналитического центра при Правительстве Российской Федерации (идентификатор соглашения 000000D730324P540002, грант № 70-2023-001320 от 27.12.2023).

Список литературы

[1] A. Biswas and M. S. Islam, "Brain Tumor Types Classification using K-means Clustering and ANN Approach," *International Conference on Robotics, Electrical and Signal Processing Techniques*, pp. 654–658, 2021, doi: 10.1109/ICREST51555.2021.9331115.

[2] D. Wong and S. Yip, "Pathology of Primary Brain Tumors—Gliomas," *Comprehensive Overview of Modern Surgical Approaches to Intrinsic Brain Tumors*, pp. 121–137, Jan. 2019, doi: 10.1016/B978-0-12-811783-5.00006-9.

[3] F. S.-T. N. Dergisi and undefined 2011, "meningiom Sınıflaması ve Histopatolojik özellikleri," *norosirurji.dergisi.orgF SöylEmEzoğluTürk Nöroşirürji Dergisi, 2011•norosirurji.dergisi.org*, Accessed: Jul. 30, 2024. [Online]. Available: http://www.norosirurji.dergisi.org/pdf.php?&id=702

[4] G. Gürsoy, B.Ç.-A.E. ve A.H. Tıp, and undefined 2019, "Hipofiz adenomlarında Radyoterapi," *dergipark.org.trG Gürsoy, BD ÇarmıklıAnkara Eğitim ve Araştırma Hastanesi Tıp Dergisi, 2019•dergipark.org.tr*, Accessed: Jul. 31, 2024. [Online]. Available: https://dergipark.org.tr/en/download/article-file/775671

[5] S. Sowmyayani, "Brain tumor volume estimation of 3d magnetic resonance imaging," *Futuristic Trends in Computing Technologies and Data Sciences Volume 2 Book 18*, pp. 109–117, Nov. 2023, doi: 10.58532/V2BS18P2CH3.

[6] Z. Ullah, M. U. Farooq, S. H. Lee, and D. An, "A hybrid image enhancement based brain MRI images classification technique," *Med Hypotheses*, vol. 143, p. 109922, Oct. 2020, doi: 10.1016/J.MEHY.2020.109922.

[7] Z. H. Zhou, "A brief introduction to weakly supervised learning," *Natl Sci Rev*, vol. 5, no. 1, pp. 44–53, Jan. 2018, doi: 10.1093/NSR/NWX106.

[8] C. Ozkaya and S. Sagiroglu, "Glioma Grade Classification Using CNNs and Segmentation With an Adaptive Approach Using Histogram Features in Brain MRIs," *IEEE Access*, vol. 11, pp. 52275–52287, 2023, doi: 10.1109/ACCESS.2023.3273532.

[9] B. J. McCarthy *et al.*, "Assessment of Type of Allergy and Antihistamine Use in the Development of Glioma," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 20, no. 2, pp. 370–378, Feb. 2011, doi: 10.1158/1055-9965.EPI-10-0948.

[10] J. Schwartzbaum *et al.*, "Association Between Prediagnostic IgE Levels and Risk of Glioma," *JNCI: Journal of the National Cancer Institute*, vol. 104, no. 16, pp. 1251–1259, Aug. 2012, doi: 10.1093/JNCI/DJS315.

[11] L. Jin *et al.*, "Artificial intelligence neuropathologist for glioma classification using deep learning on hematoxylin and eosin stained slide images and molecular markers," *Neuro Oncol*, vol. 23, no. 1, pp. 44–52, Jan. 2021, doi: 10.1093/NEUONC/NOAA163.

[12] Z. Khazaee, M. Langarizadeh, M. Ebrahim, and S. Ahmadabadi, "Developing an artificial intelligence model for tumor grading and classification, based on MRI sequences of human brain gliomas," *brieflands.com*, vol. 15, no. 1, p. 120638, 2022, doi: 10.5812/ijcm.120638.

[13] S. Bhattacharjee, D. Prakash, C. H. Kim, H. C. Kim, and H. K. Choi, "Texture, Morphology, and Statistical Analysis to Differentiate Primary Brain Tumors on Two-Dimensional Magnetic Resonance Imaging Scans Using Artificial Intelligence Techniques," *Healthc Inform Res*, vol. 28, no. 1, pp. 46–57, Jan. 2022, doi: 10.4258/HIR.2022.28.1.46.

[14] A. Anaya-Isaza, L. Mera-Jiménez, L. Verdugo-Alejo, and L. Sarasti, "Optimizing MRI-based brain tumor classification and detection using AI: A comparative analysis of neural networks, transfer learning, data augmentation, and the cross-transformer network," *Eur J Radiol Open*, vol. 10, p. 100484, Jan. 2023, doi: 10.1016/J.EJRO.2023.100484.

[15] X. Xie *et al.*, "Rapid intraoperative multi-molecular diagnosis of glioma with ultrasound radio frequency signals and deep learning," *EBioMedicine*, vol. 98, Dec. 2023, doi: 10.1016/j.ebiom.2023.104899.

[16] J.-P. Redlich *et al.*, "Applications of artificial intelligence in the analysis of histopathology images of gliomas: a review," *npj Imaging 2024 2:1*, vol. 2, no. 1, pp. 1–16, Jul. 2024, doi: 10.1038/s44303-024-00020-8.

[17] C. Ozkaya and S. Sagiroglu, "Glioma Grade Classification Using CNNs and Segmentation With an Adaptive Approach Using Histogram Features in Brain MRIs," *IEEE Access*, vol. 11, pp. 52275–52287, 2023, doi: 10.1109/ACCESS.2023.3273532.

[18] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med Image Anal*, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/J.MEDIA.2022.102470.

[19] R. A. Zeineldin *et al.*, "Explainability of deep neural networks for MRI analysis of brain tumors," *Int J Comput Assist Radiol Surg*, vol. 17, no. 9, pp. 1673–1683, Sep. 2022, doi: 10.1007/S11548-022-02619-X/FIGURES/5.

[20] R. A. Zeineldin *et al.*, "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–14, Feb. 2024, doi: 10.1038/s41598-024-54186-7.

[21] D. E. B. Schiavon, C. D. L. Becker, V. R. Botelho, and T. A. Pianoski, "Interpreting Convolutional Neural Networks for Brain Tumor Classification: An Explainable Artificial Intelligence Approach," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14196 LNAI, pp. 77–91, 2023, doi: 10.1007/978-3-031-45389-2_6/FIGURES/10.

[22] F. Yan, Y. Chen, Y. Xia, Z. Wang, and R. Xiao, "An Explainable Brain Tumor Detection Framework for MRI Analysis," *Applied Sciences 2023, Vol. 13, Page 3438*, vol. 13, no. 6, p. 3438, Mar. 2023, doi: 10.3390/APP13063438.

[23] P. K. Rao *et al.*, “Optimizing Inference Distribution for Efficient Kidney Tumor Segmentation Using a UNet-PWP Deep-Learning Model with XAI on CT Scan Images,” *Diagnostics 2023, Vol. 13, Page 3244*, vol. 13, no. 20, p. 3244, Oct. 2023, doi: 10.3390/DIAGNOSTICS13203244.

[24] A. Markella Antoniadi *et al.*, “Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review,” *mdpi.comAM Antoniadi, Y Du, Y Guendouz, L Wei, C Mazo, BA Becker, C MooneyApplied Sciences, 2021•mdpi.com*, 2021, doi: 10.3390/app11115088.

[25] W. Jin, X. Li, M. Fatehi, G. H.-M. image analysis, and undefined 2023, “Guidelines and evaluation of clinical explainable AI in medical image analysis,” *ElsevierW Jin, X Li, M Fatehi, G HamarnehMedical image analysis, 2023•Elsevier*, Accessed: Aug. 04, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522003127?casa_token=vP5GtDlApE0AAAAA:aysSBLDC-5wMH-PJoEZ200f1NoQdV9I2c1o4EO3Dee_s9kj6zmNwUwdDABopBsSOGbFtShcN

[26] S. Liang and Y. Gu, “Computer-Aided Diagnosis of Alzheimer’s Disease through Weak Supervision Deep Learning Framework with Attention Mechanism,” *Sensors 2021, Vol. 21, Page 220*, vol. 21, no. 1, p. 220, Dec. 2020, doi: 10.3390/S21010220.

[27] X. Zhang, N. Ou, C. Liu, Z. Zhuo, Y. Liu, and C. Ye, “Unsupervised Brain Tumor Segmentation with Image-based Prompts,” p. 1, Apr. 2023, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/2304.01472v1

[28] X. Liu, S. Thermos, P. Sanchez, A. Q. O’Neil, and S. A. Tsaftaris, “vMFNet: Compositionality Meets Domain-Generalised Segmentation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13437 LNCS, pp. 704–714, 2022, doi: 10.1007/978-3-031-16449-1_67/FIGURES/2.

[29] X. Liu, A. Kascenas, H. Watson, S. A. Tsaftaris, and A. Q. O’Neil, “Compositional Representation Learning for Brain Tumour Segmentation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14293 LNCS, pp. 41–51, 2024, doi: 10.1007/978-3-031-45857-6_5/FIGURES/3.

[30] T. Di Noto *et al.*, “Weakly Supervised Learning with Automated Labels from Radiology Reports for Glioma Change Detection,” Oct. 2022, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/2210.09698v2

[31] Y. C. Chou, B. Li, D. P. Fan, A. Yuille, and Z. Zhou, “Acquiring Weak Annotations for Tumor Localization in Temporal and Volumetric Data,” *Machine Intelligence Research*, vol. 21, no. 2, pp. 318–330, Apr. 2024, doi: 10.1007/S11633-023-1380-5.

[32] J. J. Yoo, K. Namdar, and F. Khalvati, “Deep Superpixel Generation and Clustering for Weakly Supervised Segmentation of Brain Tumors in MR Images,” pp. 1–9, Sep. 2022, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/2209.09930v2

[33] D. C. Oluigbo *et al.*, “Weakly supervised detection of pheochromocytomas and paragangliomas in CT,” *https://doi.org/10.1117/12.3008833*, vol. 12927, pp. 175–179, Apr. 2024, doi: 10.1117/12.3008833.

[34] D. D. Kim *et al.*, “Active Learning in Brain Tumor Segmentation with Uncertainty Sampling, Annotation Redundancy Restriction, and Data Initialization,” Feb. 2023, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/2302.10185v1

[35] L. Kapoor and S. Thakur, “A survey on brain tumor detection using image processing techniques,” *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, pp. 582–585, Jun. 2017, doi: 10.1109/CONFLUENCE.2017.7943218.

[36] P.L. Saveri, S. Kumar, and M. Bharti, “Segmentation and Area Calculation of Brain Tumor Images Using K-Means Clustering and Fuzzy C-Means Clustering,” *Lecture Notes in Networks and Systems*, vol. 673 LNNS, pp. 643–653, 2023, doi: 10.1007/978-981-99-1745-7_47/TABLES/1.

[37] B. Vijayakumari, N. Kiruthiga, and C. P. Bushkala, “Automated evaluation and parameter estimation of brain tumor using deep learning techniques,” *Neural Computing and Applications 2024*, pp. 1–17, Aug. 2024, doi: 10.1007/S00521-024-10255-6.

Секция 6

ЦИФРОВАЯ ЗЕМЛЯ И БОЛЬШИЕ ДАННЫЕ

Digital Earth and Big Data

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 008+9112.7
DOI: 10.25206/978-5-8149-3873-2-2024-691-696

Концептуальная модель экологической цивилизации в цифровой среде Антропоцена
A conceptual model of ecological civilization in the digital environment of the Anthropocene

В.Н. Бочарников
Тихоокеанский институт географии ДВО РАН, г. Владивосток, Россия

V.N. Bocharnikov
[1]Pacific Geographical Institute FEBRAS, Vladivostok, Russia

Аннотация. Цивилизация – глобальная структура, определенный исторический этап, человекоразмерная идея комфортного социального обеспечения безопасной жизнедеятельности и антропологического существования на планете Земля. Предполагается, что эффективное геопространственное моделирование может быть осуществлено в рамках междисциплинарных научных исследований, которые декларируются в идеологии проектных решений «Цифровой Земли». В сообщении рассматриваются авторские результаты экоцивилизационной оценки дикой природы России, обсуждается возможность парадигмального вовлечения цифровых технологий в методологию эколого-географических исследований. Эколого-ориентированный цивилизационный сценарий, или экологическая цивилизация, представляется как новый геополитический концепт, получивший известность в качестве кризисного средства решения острых глобальных проблем человечества, связанных с окружающей средой. Первым шагом к подготовке национальной цивилизационной программы России должна стать разработка в цифровой среде комплексной многоуровневой модели, включающей математический, системно-синергетический, организационно-управленческий, этнокультурный, геоэкологический и др. ее аспекты ее создания.

***Ключевые слова*:** цивилизационный подход, междисциплинарные исследования, экологическая цивилизация, Антропоцен, Цифровая Земля, концептуальная модель, философия

Аннотация. Civilization is a global structure, a certain historical stage, a human–sized idea of comfortable social security for safe life and anthropological existence on planet Earth. It is assumed that effective geospatial modeling can be carried out within the framework of interdisciplinary scientific research, which are declared in the ideology of design solutions of the "Digital Earth". The report examines the author's results of the ecocivilization assessment of the wildlife of Russia, discusses the possibility of paradigmatic involvement of digital technologies in the methodology of ecological and geographical research. The eco-oriented civilizational scenario, or ecological civilization, is presented as a new geopolitical concept that has gained alternative fame as a crisis solution to acute global problems of mankind related to the environment. The first step towards the preparation of the national civilizational program of Russia should be the development of a complex multilevel model in the digital environment, including mathematical, system-synergetic, organizational and managerial, ethnocultural, geoecological, etc. its aspects of its creation.

***Ключевые слова*:** civilizational approach, interdisciplinary research, ecological civilization, Anthropocene, Digital Earth, conceptual model, philosophy

I. Введение

Наука как способ познания современности не имеет конкурентов: осмысление ею взаимосвязи и зависимости общества от природы предстают насущной необходимостью постижения истинной реальности. Задача человека-исследователя состоит в том, чтобы раскрыть возможности как объективные значения и воплотить в них личностный смысл. В рамках неоклассической (постнеклассической) науки ориентация делается на связку «знание – инструмент». Результаты познания конструируются в формате словесной деятельности, предполагающей наличие культуры составления и пользования профессиональными дефинициями. Характерной особенностью современной науки является постепенное и неуклонное ослабление логического, понятийного компонента и усиление роли внерационального компонента с его интуитивной составляющей. Здесь мы четко обозначаем, что человек как биосоциальное существо для своей полноценной жизнедеятельности и развития нуждается не только в качественной социальной среде, но и в естественном своем природном окружении определенного качества [1].

Культура человека формирует наглядно облик антропогенного ландшафта, а философы убеждают нас, что мир пребывает в «цивилизационная усталости», и как следствие в мире наблюдается высокий уровень

турбулентности процессов интеграции и дифференциации мирового геопространства. В условиях расширения обитаемой территории планеты возникает настоятельная потребность соотнесения темпов изменения окружающей среды с адаптационными возможностями человека и человеческой популяции, определения допустимых пределов их воздействия на биосферу, исходя из допустимых границ ее изменения, что может быть уверенно объяснено наступлением эпохи Антропоцена [2]. Фундаментальные проблемы взаимодействия человека и природы не теряют своей социальной актуальности в любую эпоху, не является исключением и современный этап существования информационного общества.

Человечество перешло в «эпоху идей», и подтверждение этому мы находим в таких понятиях, как «постиндустриальное общество», «информационное общество», «общество знания». Глобализация как особое явление представляет ныне ведущую тенденцию мирового развития, ее оборотная же сторона, предстает причиной перманентных общемировых кризисов, наиболее остро проявившихся в первых декадах третьего тысячелетия. Технологический порядок связан со стремлением упорядочить, организовать, поставить на службу человеку хаос «естественного». Специфика «времени цифры» состоит в том, что современное общество не просто становится все более восприимчивым к новым технологиям, но и проникается особым технологическим мировосприятием [3].

В современной науке требуется постоянный процесс обеспечения «потока» нового знания - как мониторингового, оперативного, так и фундаментального характера, в которых должно происходить методологическое и теоретическое сближение различных дисциплин. Исследование технонауки зависит от того, как выстроены вопросы, на которые предполагается дать ответы. Знание только тогда приобретает информационный характер, когда оно ставит перед общественной системой проблему выбора из многообразия вариантов тех, что полезны системе, одновременно расширяя многообразие внутреннее [3]. Каждая тема служит как бы набором инструкций, отвечая на которые можно разработать принципы, понятия и постулаты теории. Попытаемся включить новые возможности в полученные традиционным способом карты антропогенной освоенности планеты, разделяя уверенность в том, что синергетическая методология сделала понимание нелинейности мира не только интеллектуальной, но и жизненной необходимостью.

II. Постановка задачи

География традиционно является одной из структур, особым языком для представления картины человеческого мира. В кризисных условиях последнего десятилетия остро встает глобальная проблема сосуществования многих культур и видения цивилизационных ценностей через призму всего человечества, что противоречит прежнему восприятию мира как единого культурно-цивилизационного континуума, имманентно присущего социуму. Новое понимание задает вектор обновления методологии науки, и тогда зададим вполне конкретный вопрос, зачем в обыденности необходимо знание?

Территория – это проекция пространства, отображение конкретного пространственно-временного континуума. В географии главенствует территориальная парадигма, суть которой в том, что территорию в географии рассматривают как арену взаимодействия различных движущих сил и компонентов, развития природы и общества, их территориальную обусловленность и выраженность. Пространственно-временное географическое изучение «следов» человека, эмпирическое раскрытие категорий «природа», «человек», «культура», «общество», семантика комбинаций их многообразного использования составляют огромный информационный пласт, детально рассматривать который не представляется возможным без конкретизации и постановки конкретной научной задачи. В географической науке возможно успешно осуществлять обобщение результатов междисциплинарных исследований по оценке состояния и устойчивости земных пространств разных размерностей, в пределах которых биотические и абиотические компоненты находятся в системной связи друг с другом, взаимодействуя с человеческим обществом в формате определяемой целостности [4].

Давно известно, что, создавая модель объектов территории и взаимосвязей между объектами как единую целостность, стало возможным получать разностороннюю информацию о территории. Обозначим направленность перспективных цивилизационных исследований в контексте раскрытия статичной картины традиционных карт, и сложной динамики геоинформационного моделирования в рамках крупного информационного проекта «Цифровая Земля» [5]. В рамках такого развития требуется пересмотреть традиционнные позиции важности дифференциации любой территории на освоенное и не освоенное человеком пространство, следует найти новый способ рассмотрения геопространства для решения междисциплинарных задач и включить как методологический фундамент активного применения в проекте «Цифровая Земля». Эмпирическим снованием к постановке новой задачи является апробированная нами в процессах многолетних исследований традиционная схема ГИС-обеспечения, которая основывается на выделении и сопоставлении культурно-географических и историко-географических районов, рассматриваемых как антропогенные (культурные) ландшафты, и крупных участков естественной (дикой) природы [6].

III. Теория

В прикладных исследованиях и технонауке процессе моделирования носит, как правило, системно-структурный характер, достаточно строго передаваемый понятиями логики и терминами науки, редко перегружаемый текстовыми описаниями, представляемый нередко графикой причинно-следственных связей, с использованием онтологических и деятельностных мотивов, существенных для целеполагания, а также в процессах принятия управленческих решений [7]. В рамках разработки ГИС нами была сформирована база геоданных, на основе которой впервые сделан расчет освоенной и неосвоенной человеком территории, путем инвертирования выделенной антропогенной инфраструктуры рассчитана территория дикой природы, где определены целостные участки местности с минимальным антропогенным воздействием, было принято, что за минимальную, ненарушенную хозяйственным воздействием территорию берутся участки площадью не менее 50 тыс. га (500 кв. км). Эти участки показывают соотношения населения и природных целостных геосистем (дикой природы) на территории административно-территориальных субъектов России.

Многообразие мира особенно важно обеспечивать особенным пространственно-временным видением, которое дает лишь эффект пребывания в «здесь и сейчас». После нашей работы появилась возможность доказательно отметить как целостные участки дикой природы обозначают границы восприятия природного, и антропогенных районов, культурных ландшафтов, что и оказалось возможным к моделированию на территории Азиатской или "Восточной" России, лишь с небольшими фрагментами природных ландшафтов в пределах "северных территорий" Европейской части Российского Дальнего Востока). Доминирующую на планете ныне западную цивилизацию следует признать техногенной, а по мере ее развития темп природных и социальных изменений, «технизация» ускоряется в такой степени, что ставится амбициозная задача отображения реального состояния биосферы и ландшафтов Земли в оцифрованном виде [8]. Технологии ГИС позволили воссоздать картину визуальных «следов присутствия человека» (дороги, поселения, промышленные объекты), рассматриваемых как антропогенно-измененные ландшафты, с отбором и выделением их противоположностей – территорий с высокой степенью сохранности природной составляющей – «дикой природы», обеспечивающей возможность соотнести степень сохранности дикой природы России с основными ландшафтными группами, представленными на территории России.

Разнородные «блоки» знаний, убеждений, мыслей, чувств, настроений, стремлений, надежд, соединяясь в мировоззрении, не образуют более или менее целостное понимание людьми мира в пространстве политики и самих себя в этом мире. Планетарный предел жизни человечества к концу XXI в. прогнозируется как сценарий катастрофы из-за истощения ресурсов в связи с ростом населения и экономики. Мы стремительно движемся к антропоцентрическому «абсолюту», но для того, чтобы понять чем это нам «грозит», представляется целесообразным обратиться к новым технологиям. В своем существовании каждый человек детерминирует специфический характер восприятия мира, эмоционального реагирования, речи, поведения, деятельности, самоидентификации субъекта, обеспечивает единство и преемственность существования социальной общности, а также стимулирует социальный прогресс посредством продуцирования культурных новаций. В любой статичной картине наблюдается постоянное и выявляемое смешение конструктов и событий вследствие практического использования таких терминов как «переменная», «параметры», «координаты» и т.п., участвующих в конструировании модельного отображения.

Объект исследования: геоэкосистемы (в понимании В.А. Кобылянского), которые оформляются в качестве различных классификационных единиц, в т.ч. ландшафтных групп, речных бассейнов, биомов или экорегионов, культурно-географических областей. Мы основывались на том, что дикая природа – общественная идея, научный конструкт, геоконцепт, особая категория природоохранных территорий, экологический индекс, в то время как свойства культурного ландшафта - это свойства поляризованного ландшафта, у которого, с одной стороны, должны сохраняться функции природного ландшафта (средоформирующие, ресурсовоспроизводящие, хранения генофонда биосферы и др.), с другой - функции обустройства «мест» и «районов» с целью создания комфортных условий жизни человека [9].

IV. Результаты экспериментов

Концептуальная модель – конструирование и отображение упрощенных сущностных характеристик объекта или предмета, системного явления, сложного образования, представленное в определенном виде, схематично (в абстрактном виде) передающего смыслы фундаментальной идеи [10]. Современные интегральные (ландшафтные) геосистемы есть часть глобальной структуры геоверсума, главным образом его социосферы, развивающиеся под влиянием его общих закономерностей и формирующие ландшафтную среду жизни человека со своими специфическими территориальными особенностями. Геоинформационная модель (ГИС-модель) обеспечивает возможность унификации, обработки, анализа и обращения к изоморфной геосистемной структуре

цивилизационных ядер, образовавшимся социально-экономическим анклавам магистральных осей и периферии сосуществования «дикой» и «очеловеченной природы» как функциональной специфике природно-цивилизационных регионов России.

Феномен огромной широтной протяженности территории России, когда с севера на юг выделяются девять природных зон, дополненных секторами высотной поясности, в целом конструируются геомоделью-районированием, состоящим из тринадцати физико-географических стран. К крупным и целостным участкам дикой природы в границах крупных физико-географических объектов на территории Азиатской России относятся: 1) арктические острова - Новая Земля, Северная Земля, Новосибирские о-ва, о. Врангеля; 2) морское побережье Северного Ледовитого океана и прилежащие к нему участки суши - полуострова Ямал, Гыданский, Таймыр, и низменности: Северо-Сибирская, Яно-Индигирская и Колымская; 3) горные территории центральной и окраинной части Азиатской России, а именно: Средне-Сибирское плоскогорье и Северо-Востока, особенно: хребет Черского, Колымское, Чукотское и Корякское нагорья; 4) акватории арктических и тихоокеанских морей.

В проекте визуализируются, демонстрируются и обсуждаются примеры динамичных процессов переходов от этапа «слитности» (синкретичности) человека и природы через обособление человека от природного, формирования культуры и создания «второй (очеловеченной)» природы, завершение натурального обмена и организации производства к прогрессорским природопокорительским стратегиям, строительству городов и организации агломераций и перехода в итоге к проективно-конструкторскому типу хозяйственной деятельности. Взаимодействие различных геосистем создает конкретную географию каждой территории, выражаемую через пространственно-временную парадигму, связанную с определением «генетического кода» каждого ландшафта, природно-хозяйственного района, каждой геосистемы, так как напластование прошлых эпох, тренды развития природной среды, хозяйства и общества, человечества и отдельных этносов - все это причудливым образом, весьма мозаично отражается в ландшафтах Земли. На основе мультидисциплинарных аналитических работ и экспедиционно-полевых обследований изучаются и характеризуются пространственные паттерны локальных протоцивилизаций: «Озерной» (Байкальский регион); «Прибрежно-морской» (Северо-Тихоокеанский регион); «Речной» (Западно-Сибирский регион); «Северной» (Арктический регион); «Горной» (Алтайский регион).

Подтвердилось, что географическое (образное и ситуативное) знание существенно дополнит традиционную логично-рациональную научную рациональность модельного рассмотрения посредством отображения и научного обращения к пространственно-временному континууму эколого-цивилизационных ядер Сибири и Дальнего Востока России. В качестве ключевых факторов цивилизационно-геокультурной дифференциации, определяющих обособление геокультурных территориальных систем разного ранга, выделяются: природный (физикогеографические рубежи – водоразделы, типы ландшафтов, высотные пояса, климатические сектора); этнический (ареалы проживания этнических и субэтнических групп, этноконтактные зоны); хозяйственный (распространение систем природопользования различных типов и видов); религиозный (конфессиональные ареалы, религиозные центры и периферии); политико-административный (границы административных образований, центры административного управления); культурноисторический (исторические рубежи и места, имеющие мемориальное значение); топонимо-лингистический (распространение языков, диалектов и «маркировка» геокультурного пространства географическими названиями, фиксирующими принадлежность определенному этнокультурному сообществу); ментальный (представления местного населения о геокультурном статусе той или иной территории, а также о местоположении геокультурных рубежей).

V. Обсуждение результатов

Цивилизация – «человекоразмерный» концепт, всеохватывающий социальные, культурные, информационные, аксиологические, политические и иные аспекты земного человеческого существования. Российская Федерация является мировым экологическим донором, крупнейшим хранителем пресной воды, крупнейшим в мире государством, в пределах границ которого общая доля сохранности естественных природных геосистем превышает 50% всей территории страны. Цивилизационная идентичность, складывающаяся из менталитета цивилизации, из многих взаимопереходов локалитета в глобалитет и обратно является приоритетной для идеи экологической цивилизации и ее удержания как одной из наиболее привлекательных будущностей России. Конструктивистские воззрения, гласящие, что релевантная реальность активно конструируется нами, должны включать и мысль о том, что конструирование окрашивается тем, что представляется наиболее важным [11].

Философская позиция релятивизма как мировоззренческой установки выражается формулой: «все относительно», но под «относительностью» имеют ввиду обычно изменяемость значения какого-либо свойства, состояния, характеристики в зависимости от определенных обстоятельств, а также понимание самих характеристик объектов. Сущность редукционизма определяется тем, что состояния «высших уровней» можно полностью описать посредством правил и теории «низших уровней». Функционирование сознания человека

предопределяет построение особой нефизической или ментальной реальности либо с помощью рационального (терминологического) рассуждения, который характерен для классического научного дискурса, либо с широким использованием метафоры, занимающей центральное место в религиозных, метафизических и мистических текстах. Гуманитарное пространство России следует рассматривать как синонимичный термин географическому пространству (географической картине мира), и выраженной как конкретно-исторический интегральный результат длительного со-развития и самоорганизации структурированных природных и общественных систем и сопряженных им процессам (производственные, инфраструктурные, бытовые и др.) в контексте соответствующих отношений (экономическими, социальными, культурными, экологическими и др.), ориентированных, в связи с потребностями человека.

Географические объекты оцениваются с точки зрения оценки значимости их в культуре, отдельная составляющая культурного объекта информационная, которая определяет его связь с историей. В противоположность дикой природе, культура играет доминирующую роль в признании человеческих ценностей, рассматривая памятники архитектуры, археологии, этнологии, природные и антропогенные объекты, находя элементы, которые указывают связь объекта со значимыми историческими событиями. Свойства культурного ландшафта - это свойства поляризованного ландшафта, у которого, с одной стороны, должны сохраняться функции природного ландшафта (средоформирующие, ресурсовоспроизводящие, хранения генофонда биосферы и др.), с другой - функции обустройства «мест» и «районов» с целью создания комфортных условий жизни человека, как материальных, так и духовных. Такое понимание цивилизации предметно отображается в формате конструирования геоконцепта, обязательным условием которого является установление не только конкретики пространственно-временного позиционирования идейного содержания, но цивилизационная идентичность, убедительность которой предстоит как один из структурирующих элементов общественной и социокультурной системы всего человечества.

Уверенно прогнозируется, что география будущего скорее будет заниматься не столько реальным, сколько различными виртуальными пространствами, создавая такие «карты», которые будут очень далеки от традиционных, и в то же время, наиболее подходящими для формирования нужных фрагментов человеческого мировоззрения. Геоинформационный подход с использованием современного программного обеспечения позволяет быстро найти и определить цивилизационные зоны на основе сформированной в географическую базу данных информации. проекций явлений, событий, предметов и процессов, происходящих в структуре географической оболочки. Новизна предложенного в проекте цивилизационного подхода заключается в том, что предлагается научно-обоснованная и практичная концепция внешнеполитического представления Российской Федерации в качестве мирового экологического лидера в текущий период финансово-экономической нестабильности и геополитической турбулентности. В методологическом отношении, геосистемный подход наилучшим образом обеспечивает комплексность при оценке состояния окружающей среды и управлении природопользованием, дает систему операционно-территориальных единиц, в которых наблюдается однородность физических параметров среды, учитываются системные связи между природными компонентами. Соответственно, в социальной системе выводится на центральное место «очеловеченные» процессы, в т.ч. и в связи с территориальным освоением человека; техносфера рассматривается в связи с техническим воздействием и технологиями, прямым (отходы, загрязнение и т.п.) и обратным взаимодействием с природой (экологические проблемы здоровья человека и среды).

VI. Выводы и заключение

Наука стремится познать сущность (глубинное внутреннее содержание) процессов, явлений через построение идеальных объектов, разрабатывая в этой связи, применяя для их наглядности и объяснения - концепции, конструкты, модели и алгоритмы. В исследованиях для создания комплексной цивилизационной модели будут использованы установления территориальных продвинутые методы пространственного анализа и математико-статистического моделирования. Освоенную часть географической оболочки советские географы называли «географической средой», понимая ее как отображение множества геобъектов и их отношений. В культурной географии, в гуманитарном, географическом и историческом познании поддерживается идея о неповторимости и уникальности каждого явления и процесса действительности. Культура играет доминирующую роль в признании человеческих ценностей, рассматривая памятники архитектуры, археологии, этнологии, природные и антропогенные объекты, находя элементы, которые указывают связь объекта со значимыми историческими событиями. Понимание – это поиск смысла, ведь понять можно только то, что имеет смысл, и если в обыденное жизни данный процесс происходит в условиях общения, коммуникации и диалога, то в науке осуществляется на базе применения методологических регулятивов. Выявлено, что природные экосистемы, местообитания и виды на территории Российской Федерации испытывают влияние ряда негативных факторов, которые имеют как

природное, так и антропогенное происхождение. В реальности результатом любого познания будет организация, формулирование и усвоение, «схватывание» неких новых смыслов, проверяемых, в той или иной степени.

Очевидно, для обоснования и описания локальных цивилизаций геокультурное районирование должно основываться на общенаучных методологических принципах: объективности, историзма, главного критерия, системности, ныне цивилизационный подход позволяет в анализе исторических процессов повернуться к человеку, смягчить традиционный экономический детерминизм. Важно здесь отметить, что **новое может быть лишь принято лишь тогда**, когда пройдет внутреннюю проверку уже известным, и наше персональное сознание найдет для вновь полученного – «ментальное место».

Источник финансирования. Благодарности

Работа выполнена при поддержке гранта РНФ (проект № 22–17–00186, «Потенциал приморских поселений для долгосрочного устойчивого развития: содержание и методы оценки на примере Тихоокеанской России») и в рамках Соглашения о представлении из федерального бюджета грантов в форме субсидий от 19 сентября 2023 г. № 075-15-2023-584.

Автор благодарит к.г.н. Егидарева Евгения Геннадьевича за многолетний совместный труд по картографированию дикой природы России.

Список литературы

(исключаются учебники, учебно-методические пособия, рукописи (диссертации, авторефераты))

1. Бочарников В. Н. Ценности в дикой природе для современного человека – новый актуальный фокус междисциплинарных исследований // Астраханский вестник экологического образования. 2012. № 4. URL: https://cyberleninka.ru/article/n/tsennosti-v-dikoy-prirode-dlya-sovremennogo-cheloveka-novyy-aktualnyy-fokus-mezhdistsiplinarnyh-issledovaniy (дата обращения: 31.07.2024).
2. Бинчик Э. Эпоха человека: риторика и апатия антропоцена. Пер. с польского Т. Пирусской. М.: Новое литературное обозрение. 2012 392 с.
3. Бочарников В. Н. Глобальный вызов технонауки, география и ГИС // География и водные ресурсы. 2014. № 1. URL: https://cyberleninka.ru/article/n/globalnyy-vyzov-tehnonauki-geografiya-i-gis (дата обращения: 31.07.2024).
4. Кузьмин С. Б. Методология геосистемных исследований в русле диалога естественных и гуманитарных наук // Общество. Среда. Развитие (Terra Humana). 2018. № 1 (46). С. 112-122.
5. Eremchenko E. What is and what is not the digital earth? // CEUR Workshop Proceedings : 30, Saint Petersburg, 22–25 сентября 2020 года. – Saint Petersburg, 2020.
6. Бочарников В.Н. Географическая экология эпохи Антропоцена. Владивосток: ТИГ ДВО РАН, 2022. 274 с.
7. Мохов В. А., Гринченков Д. В., Власова Л. М., Нгуен Тхи Тху, Пидоненко Г. В. Концептуальное моделирование как основа проектирования сложных систем // Известия вузов. Северо-Кавказский регион. Серия: Технические науки. 2018. №2 (198). URL: https://cyberleninka.ru/article/n/kontseptualnoe-modelirovanie-kak-osnova-proektirovaniya-slozhnyh-sistem (дата обращения: 29.07.2024).
8. Дергачева Е. Цифровая Земля: от идей к воссозданию виртуальной копии планеты // GeoContext. Vol. 8, N 1, 2020. С. 5-16.
9. Bocharnikov V. Concepts and measures for maintaining wilderness and landscape biodiversity in the Anthropocene // Current trends in landscape research. L. Mueller, F. Eulenstein (eds). Springer, Cham, Switzerland. 2019.
10. Меньшикова А.А. Методологические подходы и философские основания исследования концептуальной схемы // Манускрипт. Тамбов: Грамота, 2019 Том 12 Выпуск 10 С. 123-128. ISSN 2618-9690. С. 123,124.
11. Бочарников В.Н. Феномен гештальта в парадигме социально-экономической географии // Астраханский вестник экологического образования, № 3. 2012. С. 54 – 63.

УДК 303.4:004
DOI: 10.25206/978-5-8149-3873-2-2024-697-700

Пятна на ясном Солнце или проблемы Цифровой Земли у коллективных субъектов
Spots in the clear Sun or problems of the Digital Earth for collective subjects

В.А. Филимонов
Институт математики СО РАН, Новосибирск, Россия

V.A. Filimonov
Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia

Аннотация. Отмечена фундаментальность задачи избавления от принудительной концептуализации. Обоснована необходимость концептуализации и категоризации в процессах коллективной работы. Рассмотрены проблемы, связанные с процессами коллективной работы. Приведены сведения о некоторых компонентах авторской кросс-технологии ситуационного центра применительно к проекту «Цифровая Земля». Подчёркнута необходимость рефлексивных компонентов.

***Ключевые слова*:** кросс-технологии, ситуационный центр, коллективные субъекты, конфигураторы, концептуализация, когнитивная осведомлённость, когнитивная графика, неоплатонизм, рефлексия

Аннотация. The fundamental nature of the task of getting rid of forced conceptualization is noted. The necessity of conceptualization and categorization in the processes of collective work is substantiated. The problems related to the processes of teamwork are considered. Information is provided on some components of the author's cross-technology of the situation center in relation to the Digital Earth project. The necessity of reflexive components is emphasized.

***Ключевые слова*:** cross-technologies, situational center, collective subjects, configurators, conceptualization, cognitive awareness, cognitive graphics, neoplatonism, reflection

I. Введение

Научные интересы автора связаны с проектированием многодисциплинарных технологий коллективной деятельности [1, 2]. Как правило, эти технологии ориентированы на построение конфигураторов исследуемых и проектируемых систем, при этом формально-логические методы дополнялись правополушарными методами рефлексивного театра ситуационного центра (СЦ).

В работах авторов группы «Неогеография», в частности, в [3], обозначена фундаментальная проблема несводимости (образа) Единого к любому конфигуратору системных представлений. Утверждается, что проект «Цифровая Земля» обеспечивает целостное восприятие обстановки без её принудительной категоризации. Аргументация во многом основана на классической работе А.Ф. Лосева [4].

Существует термин «*ситуационная осведомлённость*» (*situational awareness*), который подразумевает знание обо всех обстоятельствах, критических для определённой деятельности в конкретной ситуации. В [3] утверждается, что «*…в Цифровой Земле за счёт использования беззнаковых средств представления геоконтекста реализован сверхголографический режим представления обстановки, удовлетворяющий условиям обеспечения ситуационной осведомлённости*». Размышления над вопросом Н. Талеба «*Почему колёса к чемоданам догадались приделать спустя 2000 лет после изобретения колеса?*», и вопросом: «*Что мешало изобрести фаустпатрон в СССР в 1941 г.?*», обусловили введение термина «*когнитивная осведомлённость*», под которым понимается рассмотрение ситуации в более широком контексте [5].

В итоге можно констатировать подтверждение гипотезы автора о том, что предлагаемые в неогеографии решения являются прототипами решения более сложных задач [6], а также рассмотреть одну из таких задач.

Дополнительно заметим, что СЦ в РФ не стали эффективным средством коллективной работы из-за их использования только в качестве мультимедиа-аудитории при сохранении старой технологии проведения партийно-хозяйственных мероприятий. По этой причине в 2020 г. в РФ появились Центры Управления Регионом, задача которых состоит в мониторинге обсуждения местных проблем в социальных сетях. Аналогичным образом ограничилось использование учебных СЦ, в том числе созданных в МГИМО в 2008 г. В этой связи представляется важной задача создания СЦ для управления проектом «Пояс и Путь», объявленная на Форуме

BRFonDE 2023 в Пекине **(**2023 Belt and Road Forum on DigitalEarth).

I. Постановка задачи

Общую постановку задачи можно сформулировать следующим образом:

- как в коллективной деятельности найти баланс между созданием конфигуратора и пониманием Единого ;
- или, в терминах автора, как строить систему ситуационной и когнитивной осведомлённости коллективного субъекта в ситуационном центре.

II. Теория

В работе использованы классические результаты А.А. Зиновьева, В.А. Лефевра, В.В. Налимова, А.Ф. Лосева, а также идеи из произведений Дж. Свифта «Путешествие Гулливера в Лапуту» и Ст. Лема «Формула Лимфатера». Для рассматриваемой темы важны положения, описанные ниже.

Логическая «клеточка» А.А. Зиновьева «Субъект + Логика (конструктор) + Онтология (образ ситуации, сформированный Субъектом с помощью его Логики)» [7]. Здесь важно, что без привязки к Субъекту рассыпается любая система утверждений. В связи с этим интересно описание Дж. Свифтом дискуссий в Лапуте с использованием остенсивных определений (вместо слов использовались предметы, имеющиеся у собеседников). Однако у коллективных субъектов Логика и Онтология очень специфичны. В частности, нейросети, обученные на большом количестве текстов разных авторов, имеют эту специфику [8].

Концепция одушевлённого существа в фундаментальной психологии В.А. Лефевра [9, 10]: субъект не занимается классификацией, а реагирует на «давление Мира» в сторону полюса, позитивного либо негативного на основании текущей ситуации и истории ситуаций. Заметим, что такого рода поведение демонстрируют даже модельные организмы, не имеющим мозга (напр. *Physarum polycephalum*). В.А. Лефевр также предложил формулу для описания искажения результатов при категоризации (замене оценок их категориями, напр. длины предметов в заданных единицах измерения на оценку категорий предметов по длине). Считал способность к схематизации одним из признаков интеллекта. Пример использования такого подхода приведён в [11]. В последние годы разрабатывал теорию эйдос-навигатора на основе идей Платона.

Идея континуальности мышления В.В. Налимова вместе с визуализацией коллективного бессознательного [12] расширяет как возможности концептуализации, так и того, что в неогеографии называется «семиотикой без знаков».

Концепция экстатического переживания Единого у А.Ф. Лосева [4] показывает ориентиры движения к пониманию и достижения этого понимания. Заметим, что у Ст. Лема в рассказе «Формула Лимфатера» совершенный разум не обладает пониманием Единого, а конфигурирует знание об окружающем мире, которое распространяется для него со скоростью света от места его нахождения.

III. Результаты экспериментов

Результаты экспериментов «Зловещая долина», «Нетранзитивные предпочтения» и «Рефлексивный коллайдер», проведённых под руководством автора в 2021-2022 гг. со студентами и школьниками в Омске и Владивостоке [13-14] показали, что категориальные оценки формируются как процессы, которые описываются теорией катастроф, а также, что коллективная экспертиза при правильной организации может быть эффективной. Также были проделаны три мысленных эксперимента. В первом из них была представлена сцена объяснения автором [4] способа посетителя покинуть его Сад, для чего, естественно, потребовалась концептуализация. Аналогично концептуализация потребовалась в сюжете «Адам и Ева в саду Эдема» с момента вручения Еве iPhone Apple с интернетом. В третьем эксперименте автор попытался представить «клеточку» Цифровой Планеты Маленького Принца, на которой обитает коллективный субъект: Маленький Принц и Роза. На этой Цифровой Планете, помимо вулканов и баобабов, появились образы субъектов, а также рефлексивный образ самой Цифровой Планеты. Читатель, безусловно, может самостоятельно реализовать эти эксперименты.

IV. Обсуждение результатов

Главным итогом является Парадокс Идеального Исследования, в котором всё более детализированная концептуализация и соответствующие конфигураторы по мере усложнения в конечном счёте исчезают, и трансформируются в экстатическое переживание постижения Единого, что напоминает метаморфоз куколки в бабочку. Более понятным является экстатическое переживание понимания чего-то важного для субъекта. Примером является реакция Архимеда «*Эврика!*» на открытый им закон.

Представляется, однако, что избежать принудительной концептуализации полностью не удастся. Одушевлённые существа (и их аналоги на базе искусственного интеллекта) имеют за счёт раскодирования ДНК программы реакции, на требующие концептуализации (напр., импринтинг). Коллективные субъекты (стая, племя, улей) вырабатывают свои специфические средства взаимодействия (например, воробьи в ареале Средиземного моря используют две разные системы знаков). Далее появляются языки, интеллигибельные истины и т.п. инструменты взаимодействия. Отдельным инструментом является схематизация, которую В.А. Лефевр считал одним из ключевых признаков интеллекта. В частности, он считал, что некоторые наскальные рисунки древних людей являются схемами организации коллективного процесса – охоты. Операции формализации и интерпретации, как и алгебра в математике, позволяют компактно представлять большие массивы разнородной информации. Замечательным примером является формализация В.Я. Проппом русской волшебной сказки, сделанная в 1928 г.

Развитие формально-логических методов привело к тому, что не все сенсорные возможности человека, в том числе возможности подсознания, были использованы в процессах коллективной работы. Для восстановления баланса была предложена когнитивная графика (Лица Чернова и др.), рефлексивный театр ситуационного центра, метафорические когнитивные карты и другие компоненты [5].

Отдельной задачей создания образа Мира, а также образов себя и других в этом Мире, является определение степени адекватности этих образов реальному Миру относительно решаемой задачи. Проблема состоит не столько в наличии ложных образов (напр., надувных танков), сколько в том, что фрагменты правды могут быть сконфигурированы так, что создают ложное впечатление о целом.

V. Выводы и заключение

Стремление авторов группы «Неогеография» подняться над концептуализацией является фундаментальным достижением, которое требует создания методов, адекватных поставленным задачам.

Концептуализация в процессе практической деятельности коллективных субъектов является неизбежной и в этом смысле принудительной. Построение образов ситуации (конфигураторов), адекватных задачам этих субъектов, требует грамотной рефлексии, а также создания их рефлексивных образов.

Техники коллективной медитации, в которых участники не используют концептуализации, а ориентированы на экстатическое переживание иллюзии познания Единого, могут быть использованы как вспомогательные компоненты.

Источник финансирования.

Работа выполнена в рамках государственного задания ИМ СО РАН, проект FWNF-2022-0016.

Список литературы

1. Филимонов В.А. Картосемиотика на глобусе кросс-технологий ситуационного центра // Diskussionsbeiträge zur Kartosemiotik und zur Theorie der Kartographie Intern. Korrespondenz-Seminar, Band 17. Dresden 2014, S. 5-13.
2. Филимонов В.А. Конфигуратор технологий коллективной работы для виртуальных ситуационных центров // Военный инновационный технополис "ЭРА". Анапа, 2020. С. 126-131.
3. Ерёмченко Е. Н. Неоплатонизм, изложенный ясно, как солнце, и Цифровая Земля // Вестник Южно-Российского государственного технического университета. Серия: Социально-экономические науки. 2024. Т. 17, № 2. С. 197–209.
4. Лосев А. Ф. Неоплатонизм, изложенный ясно, как солнце / А. Ф. Лосев, А. А. Тахо-Годи // Контекст: Литературно-теоретические исследования. – 2003. – Т. 2003. – С. 269-288.
5. Филимонов В.А. Когнитивная инфраструктура коллективной деятельности // Когнитивные штудии: когнитивная наука и когнитивная психология : сборник материалов Междун. научно-практ. конф., г. Минск, 25–26 мая 2023 г. / Белорус. гос. пед. ун-т им. М. Танка; Минск : БГПУ, 2023, С. 282-289.
6. Филимонов В. А. Неогеография: постановки задач к имеющемуся решению // Геоконтекст. – 2021. – Т. 9, № 1. – С. 17-23.
7. Зиновьев А.А. Логический интеллект – М.: Изд-во Московского гуманитарного ун-та, 2005. – 284 с.
8. Филимонов В.А. Парадокс субъектности нейросетей // Робототехника и искусственный интеллект: материалы XV Всеросс. научно-техн. конф. с междун. участием (г. Железногорск, 2 декабря 2023 г.). –

Красноярск: ЛИТЕРА-принт, 2023, С. 247-250.

9. Лефевр В.А. Что такое одушевлённость?– М: Когито-Центр, 2017.– 122 с.

10. Лефевр В.А. Общая схема современной психологии. Место рефлексивных исследований в системе наук // Математические структуры и моделирование, № 1 (45), 2018, С. 108-110.

11. Филимонов В.А., Чернявская В.С. Формализация одушевлённости на примере понятия «любовь» // Онтология проектирования. – 2022. – Т. 12, №1(43). - С. 11-24.

12. Налимов В.В., Дрогалина Ж.А. Реальность нереального. Вероятностная модель бессознательного . — М.: Мир идей; АО Акрон, 1995. — 432 с.

13. Burmistrova N., Filimonov V. Experimental game "Rock-Paper-Scissors" for teams of people and artificial intelligence systems // Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB), Novosibirsk, Yekaterinburg, 04–08 июля 2022 года. – Novosibirsk, Yekaterinburg: Institute of Electrical and Electronics Engineers, Inc., 2022. – P. 40-43.

14. Филимонов В.А., Чернявская В.С. Катастрофы субъективных оценок и их модели: к вопросу инструментализации диагностики эффекта "зловещей долины" // Территория новых возможностей. Вестник Владивостокского гос. ун-та экономики и сервиса. – 2021. – Т. 13, № 2. – С. 170-178.

УДК 91:681.3(07)
DOI: 10.25206/978-5-8149-3873-2-2024-701-707

Информационная система биосфероориентрированного эргодизайна среды жизнедеятельности
Information system biosphere of the oriented ergodesign of the environment of vital activity

Е. А. Дергачева[1], А. А. Кузьменко[2] [1-2] ФГБОУ ВО Брянский государственный технический университет, Брянск, Россия

E. A. Dergacheva 1, A. A. Kuzmenko 2
[1-2] Bryansk State Technical University, Bryansk, Russia

Аннотация. В статье описывается разработка информационной системы биосфероориентированного эргодизайна среды жизнедеятельности общества. Рассматриваются вопросы необходимости учета естественных природных систем при предметно-пространственном проектировании. Предлагаемая система опирается на междисциплинарный подход, учитывающий сложные взаимосвязи и взаимовлияния общества, техносферы и биосферы. На данном этапе проектирования разработаны подходы для интеграции и анализа научных данных в области геоморфологии, почвоведения, ботаники, зоологии, географии, техносферной безопасности, метеорологии, эргономики и эргодизайна. На основе используемых подходов строится система хранения, обработки и визуализации данных.

***Ключевые слова*:** биосфероориентированный эргодизайн, эргокаркас (эргоядро) территории, геоинформационные системы, социально-техногенный мир.

Annotation. The article describes the concept of the biosphere information system for oriented ergodesign of the society's living environment. The issues of the need to take into account biosphere cycles and subsystems of the biosphere in the subject-spatial design of the ergoclass of the territory are considered. The proposed system is based on an interdisciplinary approach, taking into account the complex interrelationships and mutual influences of society, the technosphere and the biosphere. At this stage of the design, approaches have been developed for the integration and analysis of scientific data in the fields of geomorphology, soil science, botany, zoology, geography, technosphere safety, meteorology, ergonomics and ergodesign. Based on the approaches used, a data storage, processing and visualization system is being built. The module of expertise and forecasting based on our assessment methodology demonstrates the most effective biosphere-oriented solutions in the field of designing and transforming the subject-spatial environment of society.

***Ключевые слова*:** biosphere-oriented ergodesign, ergocarcas (ergonuclear) territories, geoinformation systems, socio-technogenic world.

I. Введение

В современном мире все чаще встречается понятие «эргодизайн», который объединяет научные исследования по эргономике и дизайну в единую систему и представляет собой синергетический процесс проектирования предметно-пространственной среды жизнедеятельности. Сегодня эргодизайн выходит за рамки традиционных представлений об эргономике и дизайне, предлагая анализ взаимодействия человека с окружающей природной средой (естественной и искусственной) и техносферными элементами, а также их гармонизацию.

В трудах Л.П. Окуловой одним из результатов эргодизайна рассматривается эргономизация. «Эргономизация - это всемирно-исторический процесс развития и улучшения эргономического качества среды обитания людей, нацеленный на повышение продуктивности человеческого мозга, интенсификацию интеллектуальной деятельности человечества, увеличение производительности труда» [7]. Применение данного подхода к гармонизации социотехноприродной среды (данное понятие введено в научный оборот одним из авторов статьи Е.А. Дергачевой) в контексте проектирования позволяет нам переосмыслить эргодизайн как процесс, повышающий безопасность искусственных сред, созданных техногенным обществом не только по отношению к человеку, но и по отношению к биосфере в целом. Таким образом, современный эргодизайн необходимо рассматривать как сложный комплекс преобразований и адаптации систем, изделий, технологий и сред не только для людей, но и для окружающей естественной природы. Однако нам необходимо осознавать, что такой процесс

должен быть направлен на создание комфортных и безопасных условий жизни планеты Земля вне временных и пространственных ограничений. Другими словами, необходима единая система проектирования и управления, которая позволит оценивать влияние техногенной среды на биосферу и общество, а также биосферы на элементы техносферы и общество с позиции пространственно-временных условий и бифуркаций, к которым они могут приводить.

В данном контексте руководящий принцип эргодизайна при создании объектно-пространственной социотехноприродной среды жизнедеятельности можно сформулировать следующим образом: для повышения безопасности и стабильности систем необходимо изменить конструктивные особенности проектирования с учетом биосферно-ориентированного подхода. Предлагаемый подход направлен на повышение уровня жизни общества за счет гармонизации глобальных социально-техногенных и биосферных систем. Для реализации данного подхода процессы проектирования должны основываться на междисциплинарной методологии проектирования систем, изделий и технологий. Эргодизайн должен проводить первичную оценку стабильности сложных систем, изучая их взаимосвязи и взаимное воздействие на биосферную жизнь. В рамках данной статьи описывается информационная система биосферо-ориентированного эргодизайна, основанная на интеграции и анализе научных данных в области геоморфологии, почвоведения, ботаники, зоологии, географии, техносферной безопасности, метеорологии, эргономики и эргодизайна. Данная система строит эргокаркас искусственно созданной среды жизнедеятельности, оценивая функциональный комфорт, безопасность среды, экоэффективность и экопродуктивность проектируемых систем, изделий и технологий, а также особенности их влияния на естественные био-геосистемы.

II. Постановка задачи

В ходе разработки биосфероориентированных принципов эргодизайна [5] была выдвинута задача создания информационной системы, основанной на междисциплинарном подходе к анализу состояния биосферных био-геосистем. Решение данной задачи было найдено в методах исследования естественных природных ландшафтов. В основу разрабатываемой технологии были положены принципы геоботанического районирования, геолого-геоморфологической оценки состояния территории, подвергающейся трансформационным воздействиям, географического районирования, метеорологических условий региона и др.

III. Теория

Размышляя о тонком балансе между урбанизацией и сохранением окружающей среды, В.В. Владимиров выдвинул идею «экологического каркаса» территории [1]. По его мнению, этот каркас состоит из узлов и осей, которые концентрируют значительную экологическую активность, служащую основой экологической стабильности. Б.Б. Родоман ввел теоретическое понятие «поляризованный ландшафт» [8]. Как он указывает, эта идея представляет собой универсальный механизм пространственного разделения урбанизированных и охраняемых природных территорий. Данный ландшафт направлен на обеспечение сохранения биоразнообразия и рекреационных ресурсов в условиях растущего влияния социально-техногенного мира.

Научно-философская основа этих идей заключается в признании внутренней ценности естественной природы и необходимости поддержания экологических сетей биосферной биологической жизни. Однако данная концепция упускает из виду повышение безопасности и комфорта человека и преодоления влияния естественного отбора как основных показателей, к которым стремится общество при проектировании предметно-пространственных территорий. Именно эти особенности не позволяют рационально оценивать пространственно-временные изменения, протекающие в биосфере под влиянием социально-техногенных трансформаций. Человек стремится создать безопасную и комфортную среду здесь и сейчас, не оценивая её влияние на будущие поколения. Поэтому, расширяя концепцию экологического каркаса, следует говорить о существовании каркаса, объединяющего все элементы сложного социотехноприродного мира («социо», «техно», «био»).

В рамках данной работы в качестве такого каркаса мы рассматриваем эргокаркас. Для теоретического обоснования его существования обратимся к работам ведущих отечественных и зарубежных ученых, занимающихся изучением вопросов эргономики и эргодизайна в контексте парадигмы устойчивого развития. Так, А. Виснер в 1985 г. отмечает необходимость формированию знаниевого эргоядра [16]. Н. Морей и Р. Никерсон, поддерживая подход А. Виснера, призывают эргономическое сообщество заняться проблемами изменения окружающей среды и глобальными вызовами, сформировав на основе межпредметных знаний каркас устойчивого развития [13]. На практике эта концепция предполагает стратегическое планирование и эргодизайн, при котором городские пространства продуманно интегрируются с природной средой. Такой подход гарантирует, что развитие не будет происходить в ущерб экологической целостности, а скорее будет способствовать гармоничному сосуществованию общества, техносферы и биосферы. Придерживаясь этой точки

зрения, техногенное общество может стремиться к будущему, в котором города и природа не будут конфликтовать, а будут стремиться к гармоничному сосуществованию.

С момента публикации трудов Н. Морея и Р. Никерсона о необходимости централизации межпредметных знаний в области устойчивого развития появляется большое число трудов, посвященных формированию эргоцентра устойчивости. Так, Г. Гарсия-Акоста в концепции эргоэкологии поднимает вопрос о необходимости формирования синергетического ядра эргоэкологических знаний как феномена устойчивого развития социально-технических систем [12]. А. Тэтчер в парадигме экоэргономического развития также обращает внимание на необходимость создания экоэргономического ядра. Однако, в отличие от Г. Гарсия-Акосты, А. Тэтчер делает упор на устойчивое развитие экосферы [15].

Впервые про ядро устойчивости для обобщения междисциплинарных знаний в своей работе писали П. Хэнкок и С. Друри. Они подчеркивали важность объединения междисциплинарных методов вокруг эргономики с целью формирования устойчивых социально-технических систем, которые смогут способствовать решению глобальных проблем человечества.

Большую работу в области глобальной системной эргономики устойчивого развития социально-эколого-технического мира проделали Пол М. Сэлмон, Невилл А. Стэнтон, Джемма Дж. М. Рид, Гай Х. Уокер, Николас Дж. Стивенс, Питер А. П. Хэнкок. В труде «От системной эргономики к глобальной эргономике: мир как социально-эколого-техническая система» авторы продемонстрировали работу над сложной моделью эргоустойчивости социально-эколого-технической системы [14].

Таким образом, мы видим, что многие исследователи говорили о необходимости создания эргокаркаса или эргоядра, объединяющего общество, техносферу и биосферу и преводящих к устойчивости мира. Поддержку данного подхода мы также находим в трудах С.М. Михайлова, посвящённых эргоцентризму. Ученый обосновывает развитие эргоцентризма и поясняет: «Принцип эргоцентризма, отражающий специфику современного развития общества, стал логическим продолжением постоянно присутствующих в истории развития человечества идей антропоцентризма» [6]. С.М. Михайлов рассматривает эргоцентризм как способ построения гармоничной объектно-пространственной среды. Это свидетельствует о важности выделения эргоцентрического подхода, который способен приводить к гармоничному развитию общества и среды его жизнедеятельности. Обобщив данные рассуждения, считаем необходимым наравне с экологическим каркасом выделить эргокаркас урбанизированных территорий как структуру, позволяющую гармонизировать социотехноприродное взаимодействие в созданной искусственной среде жизнедеятельности. Данный подход способен лечь в основу перехода от социально-техногенного к социально-биосферному миру, предложенного Э.С. Демиденко и Е.А. Дергачевой [2].

Предлагаемая информационная система строит «предметно-пространственный эргокаркас» демонстрируя возможности современных интеллектуальных технологий проектирования среды жизнедеятельности, основанной на принципах эргодизайна. Спроектированные пространства смогут обладать сложной структурой, компоненты которой выполняют строго заданные функции и регулируются биосфероориентированными правилами использования.

IV. Результаты экспериментов

На данный момент мы считаем, что узлами сосредоточения наибольшей социально-техногенной активности эргокаркаса выступают ключевые территории городской, пригородной и сельской застройки. Данные территории имеют значительную площадь и высокий уровень трансфрмационного влияния на биосферу. Урбанизированное пространство соединяет городскую застройку с открытым пространством сельской местности, осуществляя транзитные и коммуникационные функции, благодаря которым производятся вещественно-энергетические связи между элементами социально-техногенной системы. Формирование дорожной системы (системы транзитных «коридоров») играет ведущую роль связи систем среды жизнедеятельности общества. Интеграция ориентированных на человека систем в мир природы требует создания переходных зон (буферных зон). Эти зоны выступают в качестве посредников между созданной искусственной техногенной средой и естественной природной средой. К ним относятся сельскохозяйственные комплексы, санитарно-защитные зоны, а также зоны охраны водных ресурсов и мест добычи полезных ископаемых. Разработка предлагаемого эргокаракаса позволит разделить все искусственно созданные территории по степени их влияния на общество и биосферу, а также дать рекомендации по гармонизации элементов социотехноприродной системы.

Разрабатываемая система на основе методологии гибридной оценки биосферной трансформации способна оценивать возможность и рациональность расширения искусственной среды жизнедеятельности общества. Она также призвана давать рекомендации по восстановительным мероприятиям и применению биосфероподобных технологий в различных отраслях промышленности и сферах жизнедеятельности.

Разрабатываемая система позволит найти сложный баланс между развитием и бережным отношением к окружающей природной среде, подчеркивая научно-философскую сущность биосфероориентированного эргодизайна в контексте проекта Цифровая Земля [3, 10, 11]. Она воплощает в себе целостный подход, который направлен не только на гармонизацию деятельности человека с миром природы, но и на восстановление хрупкого биосферного равновесия, поддерживающего жизнь на Земле. Поскольку мы продолжаем формировать нашу окружающую среду, крайне важно учитывать особенности создания функциональных биосфероориентированных пространств [4].

На данный момент мы находимся на третьем этапе проектирования системы. В качестве языка программирования был выбран Python с использованием фреймворка FastAPI. FastAPI обеспечивает эффективное и быстрое взаимодействие с клиентской частью через RESTful API. Основной функционал включает в себя обработку данных, полученных от пользователей, анализ экологических параметров с помощью статистических методов и машинного обучения, а также визуализацию результатов в виде графиков, которые передаются на клиентскую часть для отображения пользователю.

Для проектирования предлагаемой системы были проанализированы программы для обработки научных данных. В качестве аналогов для построения отдельных модулей системы были выбраны следующие программы: 1) ботанический модуль (Indicator, Canoco, Turboveg), 2) зоологический модуль (ZOOINT), 3) модуль почвоведения (Программа определения почв в соответствии с Мировой реферативной базой почвенных ресурсов) 4) климатологический модуль (SkyHelios), 5) модуль пространственного расширения городов (Atlas of Urban Expansion). Все представленные аналоги являются независимыми. Они не обобщают научные данные в единую систему, что не позволяет делать прогнозы по взаимовлиянию элементов окружающей среды друг на друга. Предлагаемая нами система осуществляет попытку обобщить данные из разных областей науки, чтобы повысить уровень анализа трансформационных изменений естественной природной среды под воздействием техносферы.

На первом этапе работ был подготовлен инструментарий для научных исследований в области геоботаники, зоологии, почвоведения, климатологии. Весь созданный инструментарий позволяет проводить анализ природных территорий, классификацию растительных сообществ, фиксировать биоразнообразие, фиксировать почвенные срезы и производить их классификацию, работать с консортами, анализировать сукцессии, оценивать пространственное расширение техногенной среды и прогнозировать его влияние на трансформацию естественных природных ландшафтов. На втором этапе работы начался сбор данных, которыми наполнялась база данных (БД) проектируемой системы. На третьем этапе осуществляется интеграция данных и их представление в географических картах. Оценивается динамика городской экспансии. Проводится оценка изменения общей протяженности городов и населения на этой протяженности, фрагментация застроенной территории за счет открытых пространств, компактности, географической формы. Для анализа берутся крупные города России, данные спутниковых снимков, а также данные переписи населения. На последующих этапах будет осуществлена попытка формирования эргокаркаса на основе спутниковых предметно-пространственных данных о городской среде. Будут выявлены и классифицированы зоны эргокаркаса по силе их воздействия на биосферу и общество, а также доработан алгоритм рекомендаций по пространственному изменению искусственной среды, прогнозированию возможных результатов этого изменения.

Рассмотрим пример работы с модулем системы геоботанической классификации ученого, использующего разработанную систему. Все ученые, работающие с системой, являются авторизованными пользователями. На первом этапе производится ввод основных характеристик описываемой площадки. Для работы нашей системы наиболее значимыми полями для заполнения, без которых система откажется продолжать работу, выступают: номер площадки, тип площадки, широта, долгота (Рис.1).

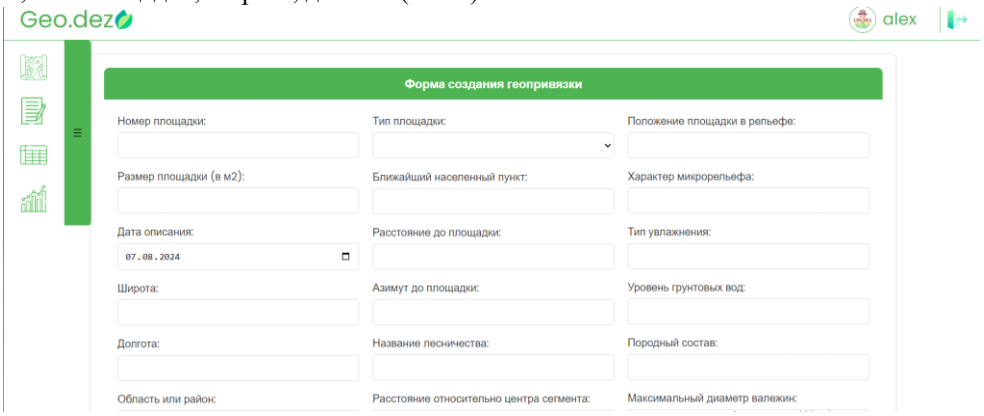


Рис. 1 Добавление в систему данных геопривязки

После добавления геопривязки производится добавление видов растений, обнаруженных на обследуемой территории (Рис. 2). Данный инструмент позволяет исследователям привязать к площадке не только набор видов растений, но и их обилие и ярус. На следующем этапе полученные данные подвергаются классификации.

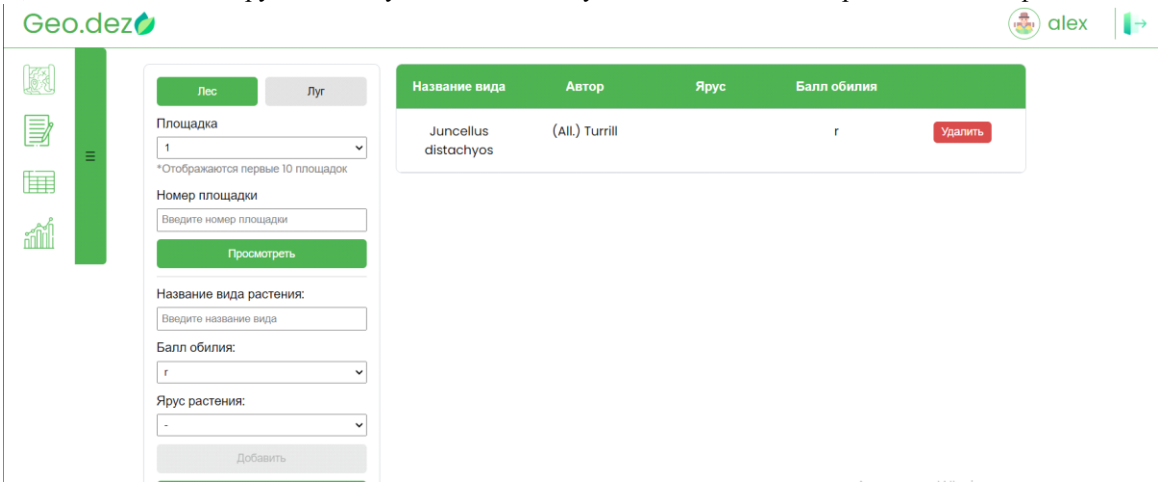


Рис. 2 Модуль работы с геоботаническими описаниями

В системе реализовано два механизма классификации: ручная и автоматическая, основанная на алгоритмах машинного обучения (Рис. 3). Применяется несколько алгоритмов, которые можно задействовать независимо друг от друга (например, алгоритм машинного обучения случайный лес и многослойный персептрон). Классификация, основанная на данных методах, облегчает работу исследователя, автоматически формирует геоботанические таблицы, что оптимизирует временные затраты на их структурирование.

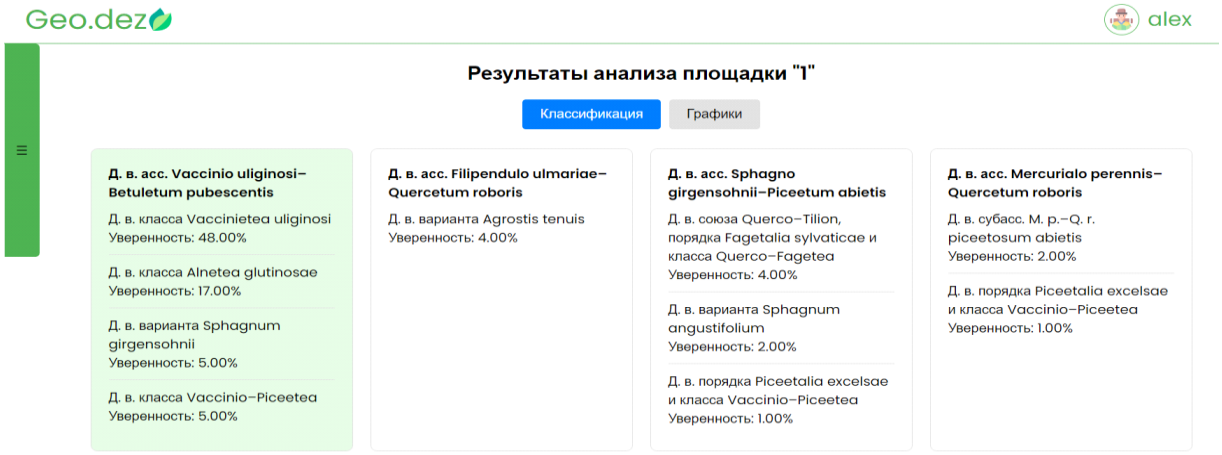


Рис. 3 Результат классификации растительного сообщества основанного на методах ИИ

На следующем этапе автор геоботанического описания при желании может опубликовать полученные данные. На карту можно вывести следующую информацию: название сообщества (обязательное поле), ФИО автора, число видов, фотография сообщества.

Аналогичным образом разработаны инструменты для проведения исследований в области почвоведения, зоологии, климатологии. В рамках данной работы мы не станем демонстрировать все интерфейсы доступные, для научного сообщества.

Неавторизованные пользователи могут видеть обобщенные данные системы. На рис. 3 демонстрируется пользовательский интерфейс.

Интерфейсная часть приложения разработана с использованием Vue.js – современного JavaScript фреймворка для создания пользовательских интерфейсов. Vue.js взаимодействует с backendom через API, используя библиотеку Axios которая позволяет выполнять HTTP-запросы для получения данных.

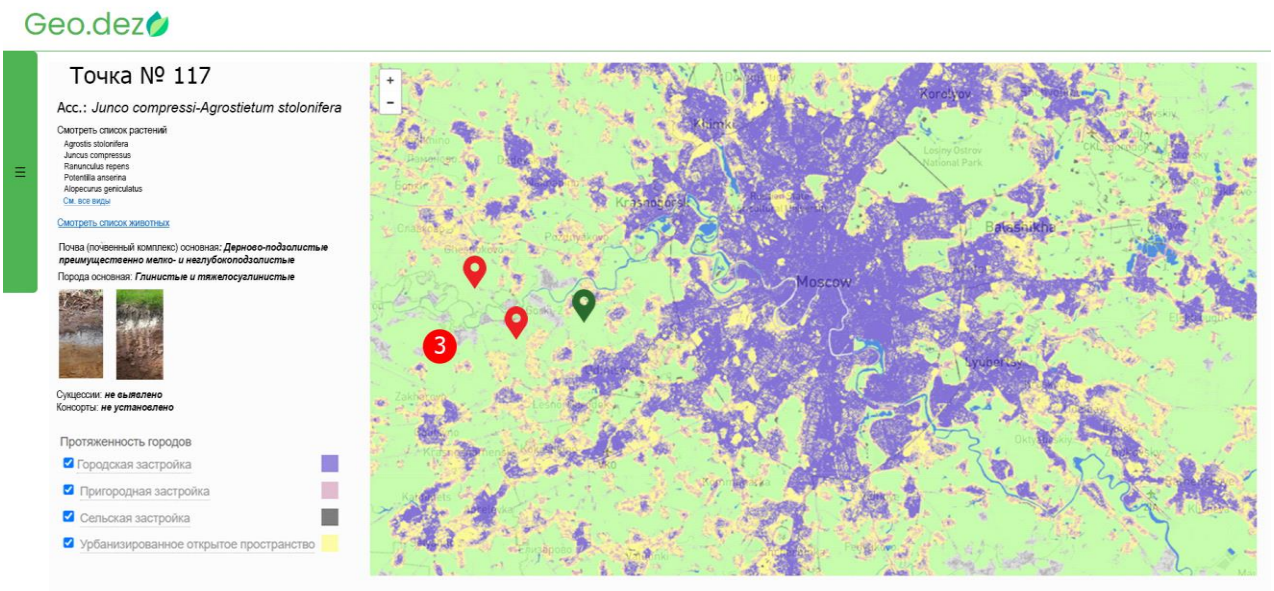


Рис. 4 Интерфейс неавторизованного пользователя

Система автоматически выстраивает связи между научными данными, обобщая их и приводя к единой точке. Так, если геоботаническое описание было сделано в 500 метрах от зоологических исследований, система установит общую точку, обобщит данные ботаников и зоологов и персонализирует каждое исследование.

V. Обсуждение результатов

Полученные результаты позволяют обобщить исследования из разных научных направлений, применив междисциплинарный подход. Применение методологии эргоцентризма и социотехноприродного подхода позволяют создать информационную систему поддержки принятия решений в области проектирования среды жизнедеятельности при помощи расширенной методологии эргодизайна.

На данный момент разрабатываемая система находится на локальном сервере авторов и проходит дополнительные тестирования. Разработанные инструменты для анализа научных данных проходят апробацию в научных центрах и институтах (например, модуль геоботанической классификации направлен на апробацию в гербарий BRSU). По окончанию тестирования научных модулей и наполнения БД данными из открытых источников в разрабатываемую систему будет интегрирован эргокаркас для каждой из нарушенных территорий.

VI. Выводы и заключение

Подводя итог, хочется отметить, что в процессе эргодизайна должно происходить приспособление искусственной среды к возможностям и ограничениям человека, не наносящее вреда естественной окружающей природной среде. Такой подход, оринетированный на социотехнобиосферную модель развития, обеспечивает хорошую основу для применения принципов эргодизайна ко всем элементам техносферы при проектировании биосфероподобных технологий. Формируемый эргокаркас социотехноприродной среды жизнедеятельности позволит найти и внедрить справедливую и сбалансированную взаимосвязь между всеми элементами социотехнобиосферного мира, уменьшив негативное влияние предметно-пространственной среды на биосферу.

В заключение необходимо отметить, что на данном этапе разрабатываемая концепция находится на начальной стадии и требует дальнейшей работы над формулированием подходов, моделей и методов, основанных на принципах биосфероориентированного эргодизайна. Разрабатываемая система позволит достичь высокого уровня гармонизации, что будет способствовать повышению качества жизни и, как результат, формированию безопасных условий жизнедеятельности, позволяющих эффективно управлять ресурсами и проектировать сложные биосфероориентированные системы, изделия, технологии и среды, способствующие переходу от социально-техногенного к социально-биосферному миру.

Список литературы

1. Владимиров В.В. Расселение и окружающая среда. – М.: Стройиздат, 1982. – 228 с
2. Демиденко, Э. С. Буржуазно-техногенное уничтожение биосферной жизни и земного мира: междисциплинарное исследование / Э. С. Демиденко, Е. А. Дергачева. – Москва : Ленанд, 2023. – 276 с.

3. Ерёмченко, Е. Н. (2024). Неоплатонизм, изложенный ясно, как солнце, и Цифровая Земля. Bulletin of the South-Russian State Technical University (NPI) Series Socio-Economic Sciences, 17(2), 197–209. https://doi.org/10.17213/2075-2067-2024-2-197-209

4. Калинин, В. С. (2024). Социально-философская проблематика ноосферной безопасности в техногенном мире. Bulletin of the South-Russian State Technical University (NPI) Series Socio-Economic Sciences, 17(3), 154–165. https://doi.org/10.17213/2075-2067-2024-3-154-165

5. Кузьменко, А. А. Тенденции развития эргономики в контексте парадигмы устойчивого развития социально-техногенного мира: социально-философский анализ: монография / А. А. Кузьменко. – Брянск : БГТУ, 2024. – 159 с. ISBN 978-5-907570-86-3

6. Михайлов, С. М. Принцип эргоцентризма и индуктивный подход в организации предметно-пространственной среды города / С. М. Михайлов, А. С. Михайлова // Innovative Project. – 2016. – Т. 1, № 4(4). – С. 43-50. – DOI 10.17673/IP.2016.1.04.6. – EDN ZWRHIP.

7. Окулова Л.П. Формирование эргономической образовательной среды в системе «школа – вуз»: монография. Ижевск: Удмуртский университет, 2019. 224 с.

8. Родоман Б.Б. Поляризация ландшафта как средство сохранения биосферы и рекреационных ресурсов // Ресурсы, среда, расселение. – М.: Наука, 1974. – С. 150-162.

9. Сиротский, М. С., & Дергачева, Е. А. (2024). Проблемы и противоречия четвертой промышленной революции. Bulletin of the South-Russian State Technical University (NPI) Series Socio-Economic Sciences, 17(3), 219–232. https://doi.org/10.17213/2075-2067-2024-3-219-232

10. Dergacheva E.A.,Demidenko E.S. Visualizing Global Socio-Technogenic Human Transformation: Digital Challenges of Living Earth // CEUR Workshop Proceedings of the 30th International Conference on Computer Graphics and Vision (Graphicon 2020). Proceedings of the 30th International Conference on Computer Graphics and Machine Vision, Vol. 2744, pp.44-1 – 44-14.

11. Dergacheva E. Visualizing Socio-techno-natural Processes: Issues and Challenges // CEUR Workshop Proceedings of the 29th International Conference on Computer Graphics and Vision (Graphicon 2019), Vol. 2485, pp. 168-172.

12. García-Acosta G., Pinilla M.H.S., Larrahondo P.A.R., Morales K.L. (2014). Ergoecology: fundamentals of a new multidisciplinary field // Theoretical Issues in Ergonomics Science, 15(2), 111-133.

13. Nickerson R.S., Moray N. (1995). Environmental change / In R.S. Nickerson (ed.) // Emerging needs and opportunities for human factors research (pp. 158-176). National Academy Press, Washington, DC.

14. Paul M. Salmon N.A., Gemma J.M., Guy H. Nicholas J.A. Hancock. Contemporary Ergonomics and Human Factors. 2019 / eds. Rebecca Charles and David Golightly. CIEHF.

15. Thatcher A., Yeow P.H. (2016). A sustainable system of systems approach: a new HFE paradigm // Ergonomics, 59(2), 167-178.

16. Wisner A. (1985). Ergonomics in industrially developing countries // Ergonomics, 28(8), 1213-1224.

УДК 004.056
DOI: 10.25206/978-5-8149-3873-2-2024-708-714

Новый дескриптор на основе частотного анализа для защиты векторных карт
The development of a novel frequency analysis-based descriptor for the protection of vector maps

О.П. Якимова, В.Г. Горшков
Ярославский государственный университет им. П.Г.Демидова, г. Ярославль, Россия

O.P. Yakimova, V.G. Gorshkov
Yaroslavl Demidov State University, Yaroslavl, Russia

Аннотация. В работе предлагается новый дескриптор векторной карты в качестве так называемого нулевого водяного знака (zero watermark) в задачах защиты авторских прав на цифровые картографические данные. Дескриптор основан на частотном анализе и имеет два варианта: линейный и матричный. Исследуется устойчивость предлагаемого дескриптора к различным искажениям векторной карты (сдвиг, добавление новых объектов, удаление части точек и т.п.). В ходе экспериментальных исследований использовались фрагменты цифровых карт трех масштабов 1:200000, 1:500000, 1:1000000. Для них вычислялись оба варианта дескриптора с различными параметрами, затем производилась атака на векторные данные, после чего оценивалась устойчивость значений дескриптора. Проведённые исследования показали, что оба варианта предлагаемого дескриптора могут быть применены на практике для защиты авторских прав на цифровую карту.

***Ключевые слова*:** нулевое встраивание цифрового водяного знака, дескриптор, частотный анализ, защита векторной карты, ГИС.

Abstract. This paper puts forth a novel vector map descriptor, which is employed as a zero watermark for the purpose ofsafeguarding digital cartographic data from unauthorized reproduction and distribution. The descriptor is based on frequency analysis and comprises two variants: linear and matrix. This study investigates the stability of the proposed descriptor in the presence of various distortions to the vector map, including shifts, the addition of new objects, and the removal of points. In the course of the experimental studies, fragments of digital maps at three different scales (1:200,000, 1:50,000, and 1:100,000) were utilized. The two variants of the descriptor with different parameters were calculated for the aforementioned vector data, which was then subjected to an attack. Following this, the stability of the descriptor values was evaluated. The results of this research demonstrate that both variants of the proposed descriptor can be applied in practice for digital map copyright protection.

Keywords: zero watermarking, descriptor, vector map protection, GIS, frequency analysis

I. Введение

Пространственные данные, то есть сведения, характеризующие местоположение и геометрическое описание объектов в пространстве и относительно друг друга (на местности), играют важнейшую роль при решении как фундаментальных задач научного исследования природных и социально-экономических процессов, так и прикладных задач в области территориального планирования и развития национальной экономики. Как правило, пространственные данные обрабатываются с помощью геоинформационных систем (ГИС) и предоставляются пользователям в виде векторных (цифровых) карт посредством веб-сервисов или веб- порталов, таких как российская геологическая карта от ВСЕГЕИ, 2ГИС, Яндекс-карты, карта Роскадастра и т.п. Несмотря на развитие информационных технологий, создание тематической цифровой карты является трудоемким процессом, занимающим не один месяц работы опытного специалиста. Отсюда возникает задача защиты авторских прав. Как правило, для этой цели используются цифровые водяные знаки (ЦВЗ), которые внедряются в пространственные данные или картографические изображения и содержат информацию об авторе[1-5]. Защита цифровых карт должна вестись по двум направлениям: во-первых, защита пространственных данных от незаконного копирования и распространения, во-вторых, защита от внесения изменений, искажения набора пространственных данных. Исходя из первого направления, ЦВЗ должен быть устойчив ко всем трансформациям векторной карты, таким как сдвиг, поворот, масштабирование, геометрическое упрощение, добавление или

удаление объектов [6,7]. Для оповещения о несанкционированных изменениях ЦВЗ должен быть «хрупким», то есть разрушаться при любых преобразованиях [8,9].

Для защиты карт также используется особый подход «нулевого встраивания цифрового водяного знака» (zero watermarking), который идентичен процедуре хеширования: для защищаемой карты вычисляется некоторая идентифицирующая информация, называемая дескриптором [10-14]. Дескриптор впоследствии хранится в базе данных отдельно. При использовании дескриптор вычисляется заново, после чего сформированный ЦВЗ посредством запроса к базе данных сравнивается с исходным [14]. В работе [10] ЦВЗ строится на основе триангуляции Делоне и вычисления локальных характеристик внутри каждого треугольника, в то время как в [11] вычисляется отношение расстояний между вершинами, после чего для большей секретности объединяется с битами ЦВЗ. Исследования [13,14] посвящены сравнению существующих дескрипторов векторных карт по поводу их устойчивости к искажениям, и, следовательно, применимости на практике, при этом эксперименты проводились над полигональными объектами. Был сделан вывод, что наибольшей стойкостью для слоев полигональных данных обладает дескриптор, который вычисляется как среднее отношение площадей ограничивающих прямоугольников.

В настоящей работе предлагается два варианта нового дескриптора для векторных карт на основе частотного анализа, работающего по сценарию «нулевого встраивания цифрового водяного знака» (zero watermarking), и проводится его экспериментальная проверка.

II. Теория. Алгоритм построения дескриптора

1. Линейный вариант

Исходная векторная карта (набор пространственных данных) разбивается на k частей прямыми перпендикулярными оси X, как показано на рис. 1а. Вопрос оптимального выбора k обсуждается ниже в разделе «Результаты экспериментов». Для каждой из k частей производится подсчет количества точек векторной карты ей принадлежащих. Получаем целочисленный вектор *Watermark* размера k. После этого вектор нормируется. Находим *min* и *max* – минимальное и максимальное значение в массиве *Watermark* и вычисляем его элементы по формуле:

$$Watermark[i] = \frac{Watermark[i] - min}{max - min}, где\ i = \overline{0, k-1.}$$

Получаем линейный дескриптор *Watermark* в виде вектора, каждый элемент которого находится в промежутке от 0 до 1.

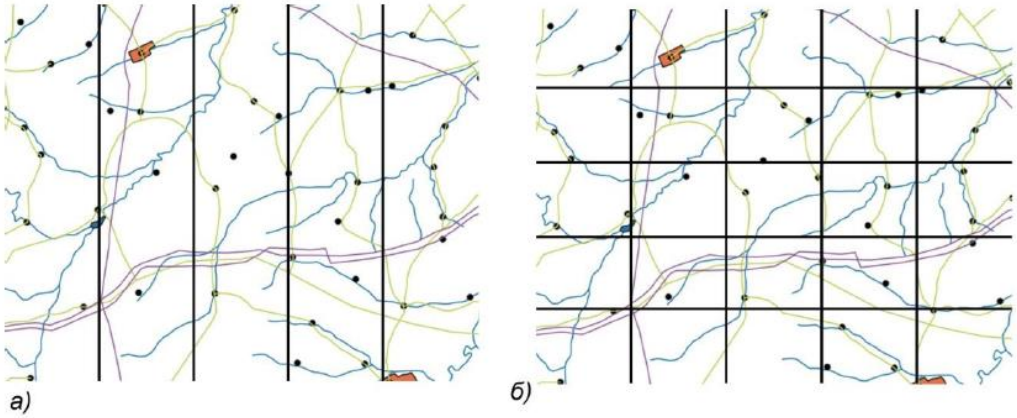


Рис. 1. Разбиение цифровой карты при k=5 для а) линейного варианта дескриптора б) матричного

2. Матричный вариант

На исходную векторную карту набрасывается прямоугольная сетка $k \times k$ (см. рис. 1б). Для каждой ячейки сетки подсчитывается количество принадлежащей ей точек карты. Как и в линейном случае, в полученной матрице находится минимум и максимум, а затем значения нормируются. На выходе получаем дескриптор в матричном виде.

III. Результаты экспериментов

При проведении экспериментов использовался следующий алгоритм проверки подлинности карты:

– вычислялся дескриптор для эталонной карты и для проверяемой карты с одним и тем же параметром k;

– производилось поэлементное сравнение вычисленных дескрипторов в прямом (первый элемент с первым, второй – со вторым и т.д.) и обратном (первый с последним, второй – с предпоследним и т.д.) направлении. Считается, что элементы совпадают, если их разность по модулю меньше установленной ошибки – 5%. Обозначим количество совпавших элементов в прямом направлении через c_1, а количество совпавших элементов в обратном направлении через c_2;

– определяется процент соответствия одной карты по отношению к другой по формуле:

$$\frac{max(c_1,c_2)}{n} \cdot 100\%, \text{ где } n = k \text{ для линейного варианта и } n = k^2 \text{ для матричного.}$$

Сравнение эталонного и проверяемого дескрипторов производится в прямом и обратном направлениях для противодействия искажению координат путем умножения на отрицательное число.

Была проведена серия экспериментов для определения стойкости предлагаемого дескриптора, выбора его оптимальной формы – линейной или матричной, а также для определения параметра k. Дескриптор исходной цифровой карты сравнивался с дескриптором той же карты, подвергнутой какой-либо атаке.

Перечень искажений, применявшихся к цифровым картам:

1. Сдвиг (для каждой точки набора пространственных данных ее координаты изменяются на некоторую константу *b* большую или меньшую нуля).
2. Масштабирование (каждая координата умножается на некоторое число *a, a=2; -2; 0.5;-0.5*).
3. Комплексный сдвиг (измененные координаты точек вычисляются по формулам: $x'= ax+b$, $y'=ay+b$, причем рассматриваются различные сочетания параметров a, b – больше нуля, меньше нуля, больше единицы, от нуля до единицы).
4. Увеличение числа точек, т.е. добавление новых объектов на карту, на 25%, 33%, 50% и 75%
5. Сокращение числа точек путем геометрического упрощения на 25%, 33%, 50% и 75%.

В качестве исходных данных рассматривались фрагменты цифровых карт Республики Башкортостан масштаба 1:200000 и 1:500000, а также севера Краснодарского края масштаба 1:1000000.

И линейная и матричная форма дескриптора показала 100% стойкость при сдвиге, масштабировании или комплексном сдвиге карты (искажения 1 – 3) при любом значении *k*.

Для атаки добавлением объектов к исходному набору данных, представляющему собой реки и каналы, добавлялись дополнительные слои, как показано на рис. 2 для карты масштаба 1:1000000. Слои подбирались таким образом, чтобы обеспечить увеличение исходного количества точек на 25%, 33%, 50% и 75%.

Для проведения экспериментов по оценке стойкости дескриптора параметр k подбирался таким образом, чтобы сторона ячейки была равна 1, 2, 5, 10, 15, 20, 50 мм в масштабе карты. Например, для масштаба 1:200000

это соответствует стороне ячейки в 200 м, 400 м, 1 км, 2 км, 3 км, 4 км, 10 км. Аналогично для других масштабов. Результаты представлены в Табл. 1.

Атака геометрическим упрощением производилась с помощью алгоритма Дугласа – Пейкера [15]. Параметр алгоритма подбирался таким образом, чтобы сократить 25, 33, 50 и 75% точек исходного набора пространственных данных соответственно. На рис. 3 показан фрагмент исходных данных и его последовательное упрощение. Аналогично случаю с добавлением объектов после геометрического упрощения оценивалась стойкость каждого варианта дескриптора. Результаты экспериментов приведены в Табл. 2.

Последний эксперимент был проведен для определения уникальности дескриптора относительно исходных данных. Для двух фрагментов векторных карт масштаба 1:1000000 (см. рис. 4) примерно одинакового размера было произведено сравнение дескрипторов.

IV. Обсуждение результатов

Экспериментальная проверка стойкости предлагаемого дескриптора к различным атакам показала, что и линейный, и матричный варианты дескриптора устойчивы как к трансформациям карты путем сдвига и масштабирования, так и к геометрическому упрощению картографических объектов. Это легко объяснить: сокращение числа точек при генерализации происходит более-менее равномерно и поэтому после нормирования значения элементов дескриптора остаются теми же. При сдвиге, переносе или масштабировании элементы дескриптора не изменяются совсем.

Атака добавлением объектов оказалась самой сложной. Здесь более стойким показал себя матричный вариант частотного дескриптора при разбиении на большее количество ячеек – при выборе k таким, что сторона ячейки равна 1-5 мм в масштабе карты. Разброс полученных результатов можно объяснить тем, что в процессе эксперимента добавлялись реальные слои данных и при этом преследовалась цель добавить примерно нужное количество точек. Некоторые слои пространственных данных содержат объекты равномерно распределенные по

поверхности карты (например, речные системы), другие же, допустим, слой дорог, содержат сгущения около областных центров и редкие объекты на периферии.

Проверка дескриптора на соответствие исходным данным показала, что лучшие результаты показал линейный вариант и параметр k должен соответствовать 1..2 см в масштабе карты.

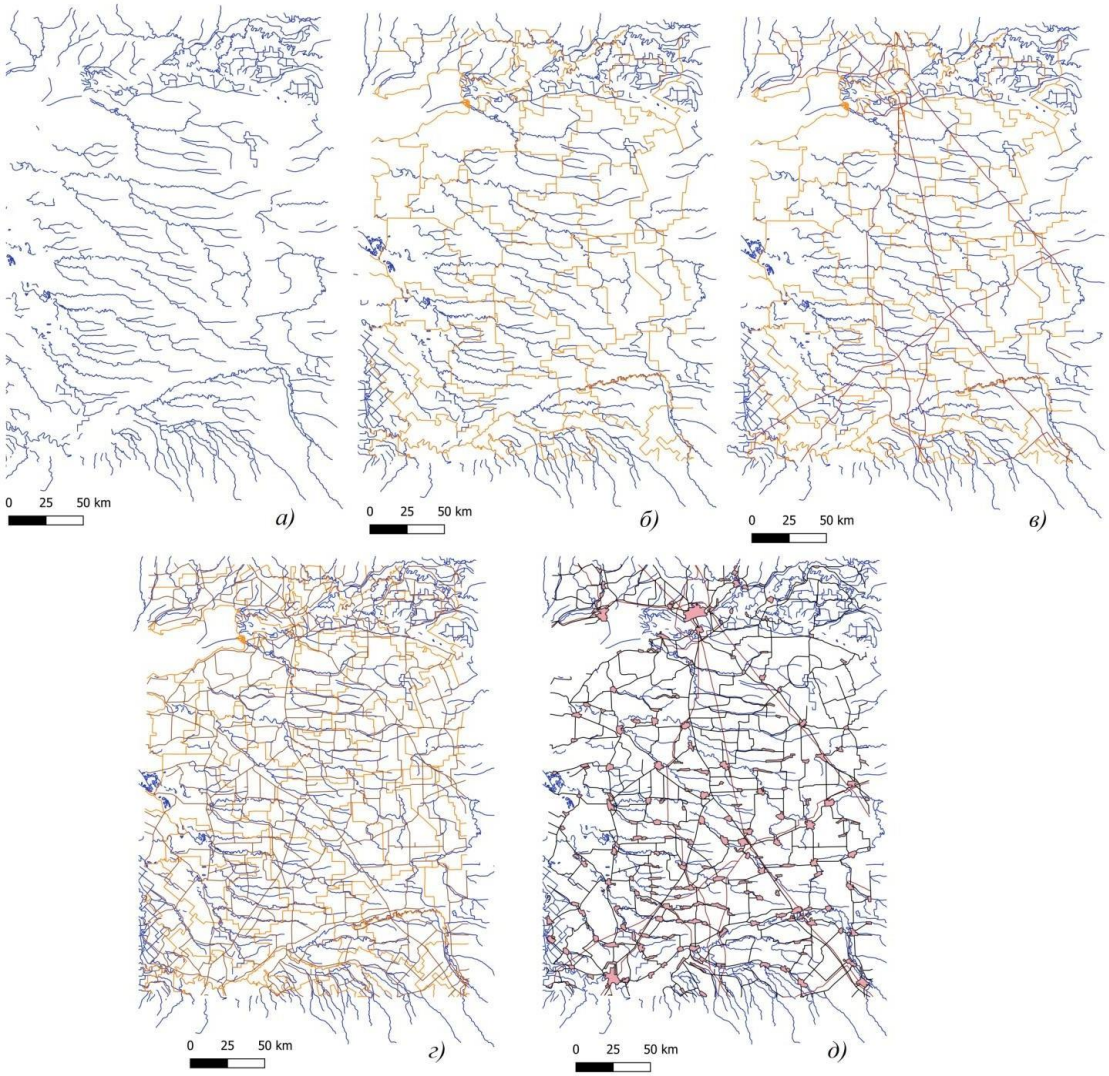


Рис. 2. Фрагмент цифровой карты масштаба 1:1000000 а) исходный набор данных; б) +25% точек; в) +33% точек; г) +50% точек; д) +75%

ТАБЛИЦА 1
РЕЗУЛЬТАТЫ СТОЙКОСТИ ДЕСКРИПТОРА ПОСЛЕ ДОБАВЛЕНИЯ ОБЪЕКТОВ

Масштаб	Дескриптор	Увеличение числа точек на %	Сторона ячейки для вычисления дескриптора(в мм в масштабе карты)						
			1	2	5	10	15	20	50
1:200000	линейный	25	36%	50%	48%	45%	22%	19%	21%
		33	57%	55%	50%	52%	48%	47%	65%
		50	20%	22%	17%	15%	18%	19%	26%
		75	20%	21%	17%	23%	22%	21%	21%
	матричный	25	**98%**	**96%**	**92%**	**85%**	76%	73%	51%
		33	**97%**	**94%**	80%	62%	50%	47%	22%
		50	**90%**	**74%**	57%	39%	30%	28%	17%
		75	**89%**	**72%**	55%	38%	29%	28%	22%
1:500000	линейный	25	25%	31%	30%	30%	24%	21%	45%
		33	45%	37%	33%	37%	40%	46%	45%
		50	27%	23%	24%	26%	21%	35%	18%

		75	20%	17%	16%	23%	16%	21%	18%
	матричный	25	**85%**	**73%**	51%	33%	30%	20%	21%
		33	**88%**	**80%**	72%	39%	31%	19%	17%
		50	**91%**	**91%**	79%	48%	45%	38%	28%
		75	**89%**	**88%**	72%	46%	41%	36%	18%
1:1000000	линейный	25	65%	31%	31%	68%	73%	73%	**83%**
		33	63%	31%	27%	48%	68%	66%	50%
		50	59%	34%	20%	51%	52%	46%	50%
		75	45%	42%	50%	20%	31%	46%	50%
	матричный	25	**93%**	**93%**	**92%**	**85%**	82%	74%	41%
		33	**92%**	**92%**	**90%**	80%	76%	65%	30%
		50	**90%**	**90%**	**84%**	69%	58%	47%	52%
		75	**89%**	**89%**	**80%**	69%	67%	53%	38%

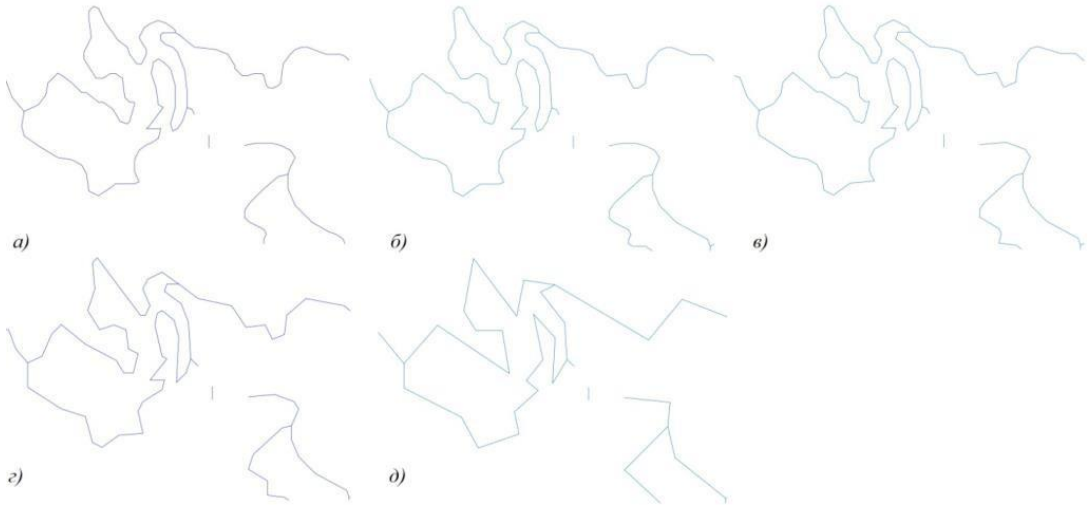


Рис. 3. Геометрическое упрощение линейных объектов а) фрагмент исходных данных; б) сокращение числа точек на 25% ; в) сокращение числа точек на 33%; г) на 50% ; д) на 75%

ТАБЛИЦА 2
РЕЗУЛЬТАТЫ СТОЙКОСТИ ДЕСКРИПТОРА ПОСЛЕ УДАЛЕНИЯ ЧАСТИ ТОЧЕК

Масштаб	Дескриптор	Сокращение числа точек на %	Сторона ячейки для вычисления дескриптора (в мм в масштабе карты)						
			1	2	5	10	15	20	50
1:200000	линейный	25	93%	94%	99%	100%	100%	100%	100%
		33	77%	89%	94%	98%	100%	100%	100%
		50	67%	80%	80%	95%	96%	96%	95%
		75	22%	37%	46%	57%	65%	63%	60%
	матричный	25	98%	97%	95%	97%	96%	96%	100%
		33	97%	95%	92%	94%	94%	90%	98%
		50	95%	91%	87%	88%	82%	83%	91%
		75	94%	87%	76%	63%	45%	53%	58%
1:500000	линейный	25	86%	80%	86%	83%	75%	75%	63%
		33	68%	69%	69%	66%	54%	53%	54%
		50	53%	51%	60%	58%	45%	46%	36%
		75	40%	46%	54%	57%	48%	46%	45%
	матричный	25	96%	96%	95%	86%	90%	83%	71%
		33	95%	93%	85%	71%	80%	68%	55%
		50	93%	91%	84%	69%	71%	65%	45%
		75	91%	89%	80%	69%	66%	63%	44%

1:1000000	линейный	25	82%	94%	100%	100%	100%	100%	100%
		33	85%	88%	98%	100%	100%	100%	100%
		50	65%	73%	50%	89%	47%	93%	83%
		75	26%	30%	31%	41%	26%	33%	33%
	матричный	25	96%	96%	98%	99%	99%	100%	100%
		33	94%	94%	96%	96%	99%	99%	100%
		50	92%	91%	93%	86%	87%	82%	91%
		75	91%	86%	80%	68%	67%	56%	41%

ТАБЛИЦА 3
РЕЗУЛЬТАТЫ СРАВНЕНИЯ ДЕСКРИПТОРОВ РАЗЛИЧНЫХ ФРАГМЕНТОВ КАРТ

Дескриптор	Сторона ячейки для вычисления дескриптора (в мм в масштабе карты)					
	2	5	10	15	20	50
линейный	14%	12%	10%	11%	7%	17%
матричный	45%	32%	26%	24%	24%	5%

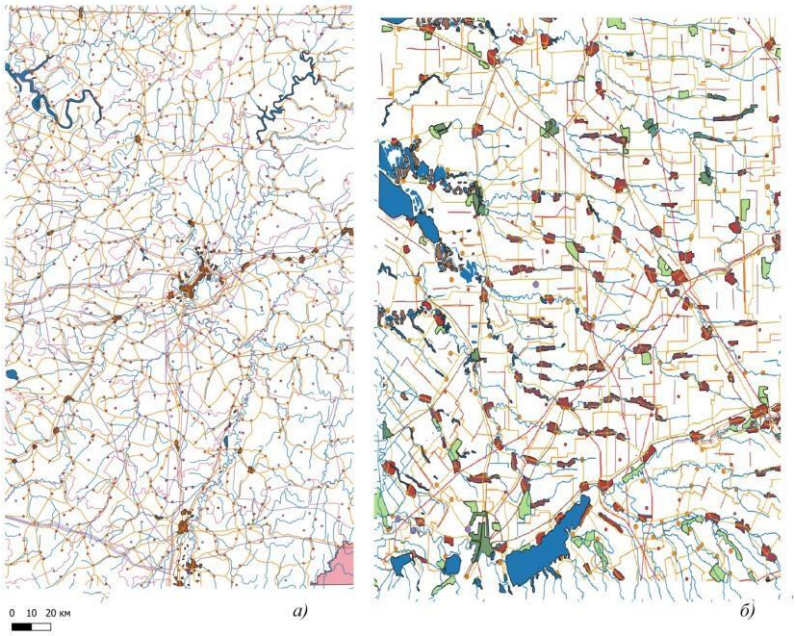


Рис. 4. Фрагменты векторных карт масштаба 1:1000000 а) Башкирии б) севера Краснодарского края

IV. Обсуждение результатов

Экспериментальная проверка стойкости предлагаемого дескриптора к различным атакам показала, что и линейный, и матричный варианты дескриптора устойчивы как к трансформациям карты путем сдвига и масштабирования, так и к геометрическому упрощению картографических объектов. Это легко объяснить: сокращение числа точек при генерализации происходит более-менее равномерно и поэтому после нормирования значения элементов дескриптора остаются теми же. При сдвиге, переносе или масштабировании элементы дескриптора не изменяются совсем.

Атака добавлением объектов оказалась самой сложной. Здесь более стойким показал себя матричный вариант частотного дескриптора при разбиении на большее количество ячеек – при выборе k таким, что сторона ячейки равна 1-5 мм в масштабе карты. Разброс полученных результатов можно объяснить тем, что в процессе эксперимента добавлялись реальные слои данных и при этом преследовалась цель добавить примерно нужное количество точек. Некоторые слои пространственных данных содержат объекты равномерно распределенные по поверхности карты (например, речные системы), другие же, допустим, слой дорог, содержат сгущения около областных центров и редкие объекты на периферии.

Проверка дескриптора на соответствие исходным данным показала, что лучшие результаты показал линейный вариант и параметр k должен соответствовать 1..2 см в масштабе карты.

V. Выводы и заключение

В работе приводится описание нового дескриптора на основе частотного анализа, который предназначен для защиты векторной карты от несанкционированного использования или подделки. Предлагаемый дескриптор работает по сценарию нулевого встраивания водяного знака (zero watermarking) и имеет два варианта: линейный и матричный. Проведена серия экспериментов по исследованию устойчивости предлагаемого дескриптора к различным атакам. Более стойким к атакам добавления объектов оказался матричный вариант дескриптора, к другим атакам (геометрическим упрощением, сдвигом, масштабированием) устойчивы оба варианта дескриптора. Следовательно, предлагаемый дескриптор может быть применен для защиты авторских прав на векторную карту. В дальнейшем планируется провести дополнительные исследования по выбору параметра для расчета дескриптора в зависимости от графической плотности карты.

Список литературы

1. Peng Y., Lan H., Yue M. et al. Multipurpose watermarking for vector map protection and authentication // Multimed Tools Appl, 2018. Vol.77. P. 7239–7259. https://doi.org/10.1007/s11042-017-4631-z

2. Выборнова Ю.Д., Сергеев В.В. Метод защиты векторных карт с использованием изображения ЦВЗ как вторичного контейнера // Компьютерная оптика. 2019. Т. 43, № 3. С. 474-483. DOI: 10.18287/2412-6179- 2019-43-3-474-483.

3. Qiu Y., Gu H, Sun J. Reversible watermarking algorithm of vector maps based on ECC// Multimedia Tools and Applications. 2018. V 77. DOI:10.1007/s11042-018-5680-7.

4. Zhang L., Yan H., Zhu R., Du P. Combinational spatial and frequency domains watermarking for 2D vector maps

// Multimedia Tools and Applications. 2020. V. 79. P.1-13. DOI: 10.1007/s11042-020-09573-3.

5. Wang Y., Yang C., Zhu C. A multiple watermarking algorithm for vector geographic data based on coordinate mapping and domain subdivision// Multimedia Tools and Applications. 2018. V.77. DOI:10.1007/s11042-017- 5358-6.

6. Li Y., Zhang L., Wang X., Zhang X., Zhang Q. A Novel Invariant Based Commutative Encryption and Watermarking Algorithm for Vector Maps// ISPRS Int. J. Geo-Inf. 2021, 10, 718. DOI: 10.3390/ijgi10110718

7. Wang Y., Yang C., Ding, K. Multiple Watermarking Algorithms for Vector Geographic Data Based on Multiple Quantization Index Modulation // Appl. Sci. 2023, 13, 12390. DOI: 10.3390/app132212390

8. Wang N., Kankanhalli M. 2D Vector Map FragileWatermarking with Region Location// ACM Trans. Spatial Algorithms Syst.2018. 4, 4, Article 12 (September 2018), 25 pages. DOI:10.1145/3239163

9. Zhang G., Da Q., Zhang L. et al. A Fragile Watermarking Scheme of Anti-deleting Features for 2D Vector Map// Advanced Hybrid Information Processing. First International Conference, ADHIP 2017, Harbin, China, July 17– 18, 2017, Proceedings, Springer. 2018 DOI: 10.1007/978-3-319-73317-3_42

10. Xi X . et al. Dual Zero-Watermarking Scheme for Two-Dimensional Vector Map Based on Delaunay Triangle Mesh and Singular Value Decomposition // Appl. Sci. Multidisciplinary Digital Publishing Institute, 2019. Vol. 9, No. 4. P. 642.

11. Peng Y., Yue M. A Zero-Watermarking Scheme for Vector Map Based on Feature Vertex Distance Ratio // Journal of Electrical and Computer Engineering. 2015. Vol. 2015 P. 1-6.

12. Zhou Q. et al. Zero Watermarking Algorithm for Vector Geographic Data Based on the Number of Neighboring Features // Symmetry. Multidisciplinary Digital Publishing Institute, 2021. Vol.13, No. 2. P. 208.

13. Асанов А. С., Выборнова Ю. Д., Федосеев В. А. Исследование стойкости компактных дескрипторов векторной карты к трансформациям ее содержимого// Информационные технологии и нанотехнологии (ИТНТ-2022) : сб. тр. по материалам VIII Междунар. конф. и молодеж. шк. (г. Самара, 23 - 27 мая) : в 5 т. / М-во науки и образования Рос. Федерации, Самар. нац. исслед. ун-т им. С. П. Королева (Самар. ун-т), Ин-т систем обраб. изобр. РАН - фил. ФНИЦ "Кристаллография и фотоника" РАН. - Самара : Изд-во Самар. ун- та, 2022Т. 3: Распознавание, обработка и анализ изображений / под ред. В. В. Мясникова. 2022. С. 031343.

14. Асанов А. С., Выборнова Ю. Д., Федосеев В. А. Сравнительный анализ компактных дескрипторов для защиты векторной карты // Proceedings of the 32nd International Conference on Computer Graphics and Vision 2022. С. 796–804. DOI: 10.20948/graphicon-2022-796-804

15. Douglas D.H., Peucker T.K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature // Cartogr. Int. J. Geogr. Inf. Geovisualization. University of Toronto Press, 1973. Vol. 10, No. 2. pp. 112–122.

УДК 004.932.2
DOI: 10.25206/978-5-8149-3873-2-2024-715-724

Анализ и обработка смоделированных изображений морских навигационных карт

Analysis and processing of simulated images of nautical charts

Н. В. Дмитриев1
1ФГБОУ ВО «Уральский государственный университет путей сообщения», Екатеринбург, РФ

N. V. Dmitriev1
[1]Ural State University of Railway Transport, Yekaterinburg, Russia

Аннотация. В работе проведены теоретические и экспериментальные исследования анализа и обработки растровых изображений, обладающих сложной структурой. В качестве представителя таких изображений выбраны изображения морских навигационных карт. Учитывая необходимость наличия нескольких этапов обработки и оценки критериев качества на каждом из них а также сложность получения размеченных данных вручную реальные изображения были заменены правдоподобными моделями, полученными с использованием алгоритма получения псевдопланетарного рельефа diamond-square в соответствии с действующим ГОСТ Р 58251–2018 «Карты навигационные бумажные внутренних водных путей Российской Федерации». Разработана обобщённая методика обработки таких карт для получения информации об отметках глубин морского дна, состоящая из этапов предобработки, многометочной сегментации, локализации, двухкаскадного распознавания и группировки. Во время экспериментальных исследований все разработанные алгоритмы были исследованы для определения лучшего набора фильтрующих операций, типа интеллектуальной модели и их гиперпараметров по критериям полноты, точности и скорости обработки. Итоговая точность составляет 89%, полнота 77%, время работы 4.3 секунды на мегапиксель. Полученные результаты свидетельствуют о возможности применения разработанной методики для других сложноструктурных изображений.

***Ключевые слова*:** обработка изображений, сложноструктурные изображения, навигационные карты, моделирование изображений, сегментация, локализация, распознавание образов, группировка данных, компьютерное зрение, машинное обучение, интеллектуальные системы, искусственные нейронные сети.

Аннотация. The paper presents theoretical and experimental studies of the analysis and processing of raster images with a complex structure. Images of nautical charts have been selected as a representative of such images. Taking into account the need for several stages of processing and evaluating quality criteria on each of them, as well as the difficulty of obtaining manually marked data, real images were replaced with plausible models in accordance with the current GOST R 58251-2018 "Paper navigation maps of inland waterways of the Russian Federation" using the diamond-square algorithm for getting pseudoplanetary relief. A generalized technique of processing of such charts for getting information about the seabed depth marks has been developed, consisting of preprocessing stages, multi-label segmentation, localization, two-stage recognition and data grouping. During the experimental studies, all the developed algorithms were investigated to determine the best set of filtering operations, the type of intelligent model and their hyperparameters according to the criteria of recall, accuracy and processing speed. The final accuracy is 89%, recall is 77%, and the time of processing is 4.3 seconds per megapixel. The obtained results indicate the possibility of using the developed technique for other complex structured images.

***Ключевые слова*:** image processing, complex structural images, nautical charts, simulating of images, segmentation, localization, image recognition, data grouping, computer vision, machine learning, intelligent systems, artificial neural networks.

I. Введение

Анализ и обработка растровых изображений, обладающих сложной структурой, в настоящий момент не является решённой проблемой как с теоретической, так и с практической точки зрения. Объясняется это значительным разнообразием такого класса изображений: к ним можно отнести как естественные проекции (аэрофотоснимки, рентгенограммы, гистологические срезу и т. д.), так и изображения, созданные человеком рукотворно (эскизы, схемы, рукописные документы) или полученные в автоматизированном режиме (чертежи, диаграммы, карты). Особенностями сложноструктурных изображений (ССИ) являются значительная доля пересечений между распределёнными образами [1], наличие неявных правил отображения оригинальных объектов на изображении [2] и высокая информационная избыточность, что приводит к необходимости

применения контекстного или семантического анализа [3], с помощью которого можно было бы выявить ошибки распознавания.

Обработка ССИ должна выполняться в несколько этапов не только потому, что на каждом этапе может происходить преобразование типов обрабатываемых данных, но и потому, что при этом можно достичь лучшего результата за счёт большей формализации алгоритмического и программного обеспечения [4]. Современные исследователи всё чаще отмечают [5, 6], что одно лишь использование больших и сложных интеллектуальных моделей не является целесообразным для изображений, содержащих пространственно-распределённые объекты с возможностью пересечения. Например, свёрточные нейронные сети показывают достаточно хорошее качество при распознавании отдельных образцов, но для распознавания группы образцов потребуется значительное увеличение тренировочных данных, что часто невозможно [7]; сложные интеллектуальные модели являются менее наглядными и интерпретируемыми, поэтому для них сложно понять, как обеспечить увеличение качества их работы; найденные вручную признаки изображений часто более надёжны, чем найденные автоматически [8]. Кроме того, сложные интеллектуальные модели обычно применяются для изображений высокой чёткости, ССИ же могут обладать низким разрешением порядка 100-300 dpi, например, если они получаются при сканировании материалов, напечатанных с помощью старых матричных или струйных принтеров. Выполнение же дальнейшего контекстного и семантического анализа в любом случае требует выполнения формальных операций, которые задействуют графовый подход [9] или использование статистики.

Тем не менее, при разделении обработки на несколько этапов возникает проблема значительного увеличения разметочных данных для реализации обучения и тестирования разработанных программных модулей: если для задачи распознавания в целом нужно лишь найти необходимые образы и вручную классифицировать их, что выполнимо за время, линейно зависящее от количества образов, то для разметки промежуточных этапов в общем потребуется время, линейно зависящее от количества пикселей (площади) изображения.

II. Постановка задачи и материалы исследования

Исходя из написанного выше, исследование ССИ, особенно относящихся к типу полученных автоматизировано, по возможности лучше производить на смоделированных изображениях, которые максимально достоверно отражают реально существующие. Целью данного исследования является анализ и обработка смоделированных изображений морских навигационных карт (МНК) с расширением 200 dpi на основе правил Управления навигации и океанографии Министерства обороны Российской Федерации и ГОСТ Р 58251–2018 «Карты навигационные бумажные внутренних водных путей Российской Федерации» для распознавания отметок глубин морского дна.

Для получения изображений МНК вначале нужно получить цифровую модель рельефа (ЦМР), выраженную в виде матрицы высот, для чего можно воспользоваться итеративным алгоритмом diamond-square [10], основанный на получении значений высот в новых точках с помощью усреднения значений соседних и добавления случайной компоненты, амплитуда которой уменьшается с увеличением номера итерации, в результате чего можно получить псевдопланетарный рельеф. В итоге получаем ЦМР размера 2044×2044.

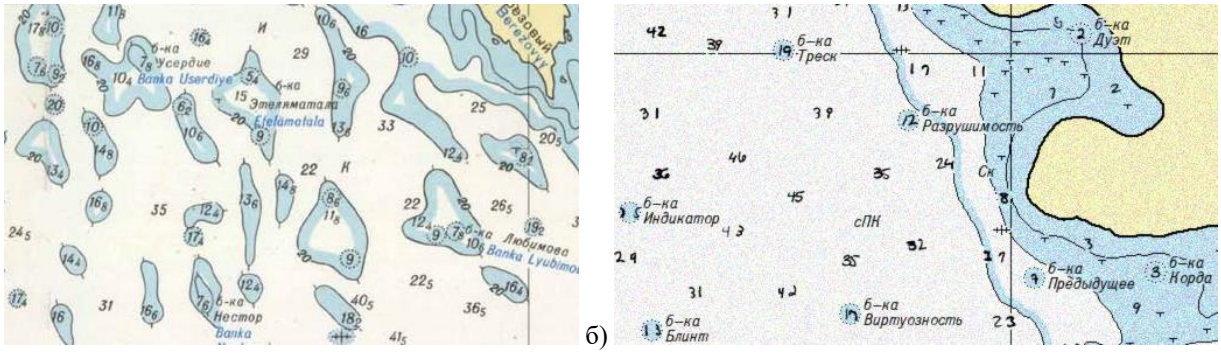
В полученной по стандартному алгоритму ЦМР вероятность возникновения вершин и впадин одинаковая, что не является характерным для рельефа морского дна, поэтому модифицируем алгоритм с помощью добавления выпуклых вниз «впадин», являющихся двумерными гауссианами со случайным расположением центра и значением дисперсии. Далее осуществляем сглаживание возможных резких скачков с помощью медианной фильтрации и нормализуем изображение карты, чтобы вода занимала 90% площади, с максимальным значением глубины в 100 метров.

Координаты отметок глубин выбираются таким образом, чтобы между ними было 50-100 пикселей (всего на одном изображении присутствует порядка 1000 отметок глубин), при этом отметки глубин могут изображаться поверх изобат. Для усложнения распознавания в качестве цифр использовались рукописные из датасета MNIST. Для повышения обобщающей способности алгоритмов и интеллектуальных методов расстояние между центрами цифр варьируется, что может привести как к наслоению образов, так и к их достаточному отдалению. Высота цифр в датасете MNIST составляет порядка 20 пикселей, что соответствует расширению примерно в 300 dpi при стандартном кегле шрифта, равном 4 или 5 пунктам. Для соответствия расширения уровню в 200 dpi, цифры были уменьшены до высоты 13 пикселей.

Кроме отметок глубин также были добавлены отметки банок (мелей) вместе со случайными названиями, скал, затонувших кораблей, обозначение грунтов, координатная сетка под случайным углом, изобаты с шагом в 20 метров с обозначениями (также изобаты с глубиной в 5 и 10 метров). Все данные образы обозначены чёрным цветом, шрифт Д-431, земля – светло-жёлтым, глубины до 10 метров закрашиваются голубым, как и частично изобаты с уровнем 20 метров в сторону берега.

Полученные изображения размывались с использованием фильтра Гаусса со среднеквадратичным отклонением равным 0.5 и добавлялся цветной шум с σ = 0.01. Итоговые изображения сохранялись в формате jpg, что зачастую соответствует формату существующих изображений и вносит дополнительные искажения, характерные соответствующему алгоритму сжатия. Всего было смоделировано 20 изображений, половина из которых использовалось для обучения работы алгоритмов, половина – для тестирования.

Пример реального изображения МНК и смоделированного показан на рис. 1. Можно сделать вывод, что полученные смоделированные изображения достаточно похожи на оригинальные, хоть и не используют весь спектр условных знаков, которые могут быть задействованы на реальных изображениях навигационных карт.



а) б)

Рис. 1. Примеры изображений: а) участок сканированного изображения карты восточной части Финского залива в масштабе 1 : 250 000 (59°40’); б) смоделированное изображение МНК

III. Разработанная методика анализа и обработки смоделированных изображений МНК

Для уменьшения влияния неявных правил отображения ССИ на их понимание, анализ и обработку необходимо прежде всего уменьшить влияние элементов изображения друг на друга, возникающее из-за пересечений образов. Для этого воспользуемся принципом декомпозиции и восходящим подходом к созданию общей методики обработки ССИ: вначале обработку будем производить для отдельных пикселей, затем для их групп, затем выделим образы, потом будем группировать эти образы, пока не дойдём до понимания изображения в целом. В дальнейшем также можно использовать аналогичный подход для объединения нескольких изображений, для карт это будет естественной операцией, так как все они также являются частью глобальной сетки. Предлагаемая методика основана на разработанной ранее для случая анализа и обработки сканированных изображений топографических карт [11] и состоит из следующих этапов:

1. предобработка – выполнение операций фильтрации и улучшений над изображением как единым множеством пикселей, на входе и на выходе этапа – цветные изображения RGB;

2. многометочная сегментация – поточечная классификация отдельных пикселей для выделения сегментов, относящихся к определённым образам с мерой принадлежности, при этом один пиксель может принадлежать сразу нескольким классам сегментов, на выходе этапа – набор полутоновых изображений;

3. локализация – выполнение морфологических и прочих фильтрующих операций над полутоновым сегментированным изображением для получения подклассов, удаления ложно найденных образов, уточнения центров правильно найденных образов, на выходе этапа – набор бинарных изображений;

4. распознавание – выполняемая в два каскада операция:

4.a. распознавание образов скользящим окном исходя из предположения, что их центры лежат в пикселях, полученных на этапе локализации, на выходе подэтапа – индексированное изображение, где каждый пиксель отнесен к определённому подклассу, данное изображение состоит из суперсегментов – связного множества пикселей, которые могут быть отнесены к разным подклассам;

4.b. распознавание образов внутри суперсегментов, при этом в качестве признаков используется доля пикселей для каждого подкласса, на выходе – набор объектов с параметрами координат и значений подкласса;

5. группировка – объединение нескольких точечных объектов в один объект с использованием контекстного и семантического анализа, на выходе этапа – набор объектов с параметрами координат и значений.

В качестве методов и моделей предлагается использовать как формальные (логистическую регрессию LR, линейный дискриминантный анализ LDA, наивный байесовский классификатор NB), так и интеллектуальные (искусственные нейронные сети прямого распространения ANN, метод k ближайших соседей k-NN, Метод опорных векторов SVM, метод решающих деревьев DT). Кроме изменения гиперпараметров также исследуем изменение размера *a* квадратного скользящего окна для сегментации, последовательность морфологических и фильтрующих операций для локализации, набор фильтров для выполнения группировки цифр. Учитывая

небольшой размер распознаваемых образов и сильный уровень зашумления и размытия в использовании моделей глубокого обучения, например, свёрточных нейронных сетей CNN, не имеет смысла.

Общая схема обучения и тестирования набора алгоритмов и моделей для обработки точечных объектов на МНК показана на рис. 2. Преимуществом моделированных изображений является то, что мы можем получить полный набор разметочной информации и таким образом реализовать расчёт критериев качества для каждого из этапов методики.

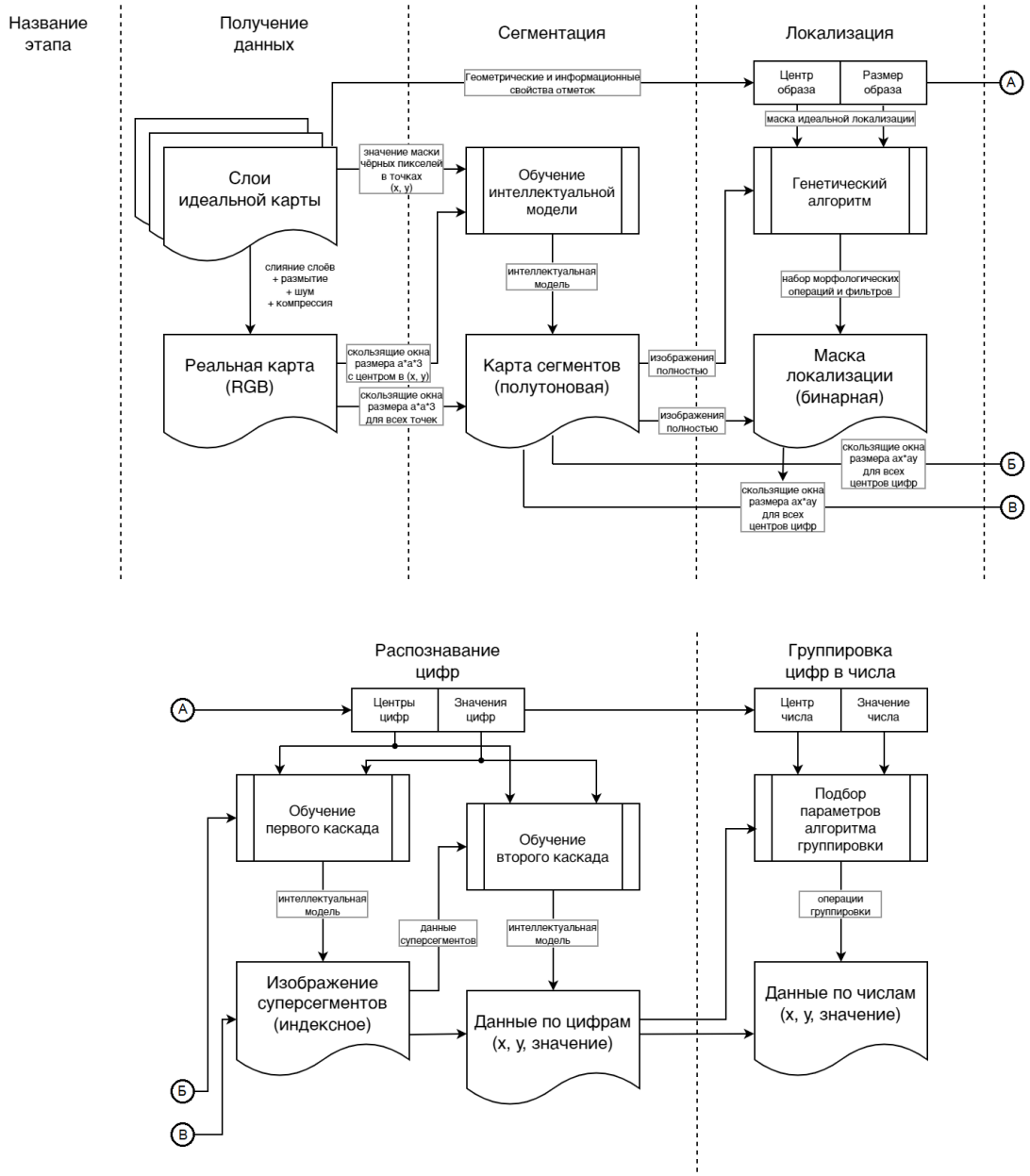


Рис. 2. Схема обучения и тестирования обработки точечных объектов на изображениях МНК

Базовым критерием качества работы алгоритмов является их время работы, так как это напрямую влияет на экономическую эффективность автоматизации. Для учёта различного разрешения считаем время в секундах на мегапиксель. При этом исправление ошибок оператором также занимает некоторое время. Основными типами ошибок, которые могут возникнуть при работе алгоритмов, являются: неверное распознавание (может быть исправлено за условную секунду), ложное распознавание, когда за числовую отметку указывается какой-то другой образ (аналогичное исправление за условную секунду) и пропуск отметки. Ошибки последнего типа являются наиболее трудозатратными для оператора, так как время их исправления зависит не от количества отметок, а от размера МНК, поэтому их количество необходимо минимизировать на этапах сегментации и локализации.

IV. Экспериментальные исследования

Алгоритмы были реализованы в виде программных модулей с помощью пакета прикладных программ MATLAB R2022a под управлением 64-разрядной операционной системы Windows 10 с 8 ГБ оперативной с использованием видеокарты AMD Radeon 530 (2 ГБ).

Так как уровень искажений в смоделированных изображениях МНК уже является достаточно значимым, то этап предобработки в данном случае не выполняем (может быть потеряна информация), и первым этапом исследования является проверка моделей, выполняющих сегментацию. Так как на данном этапе критичен пропуск необходимых пикселей, который может привести к некорректной работе алгоритмов следующих этапов, то в качества критерия качества будем использовать F_β-меру со значением $\beta = 2$.

Для обучения моделей для сегментации было выбрано по 1000 точек на каждом тестовом изображении, относящихся к искомому классу (бинарная классификация). Данные точки выбирались центрами скользящих окон, из которых производилась выборка признаков. Результаты по моделям представлены в таблицах 1–3: если на разных картах результат вычисления критерия качества различался, то в соответствующих ячейках представлены средние значения и среднеквадратическое отклонение, оно в том числе отражает стабильность работы метода. Не показаны модели и методы, работающие медленнее, чем со временем 10 с/Мп.

ТАБЛИЦА 1
РЕЗУЛЬТАТЫ ПРИ ИСПОЛЬЗОВАНИИ ФОРМАЛЬНЫХ МЕТОДОВ, a = 1

Метод	**F_β, %**	**T, с/Мп**
Пороговый: $0 \le R, G, B \le 160$	96.5±0.2	0.16
Пороговый: $0 \le \mathrm{mean}(R, G, B) \le 150$	96.3±0.2	0.17

ТАБЛИЦА 2
РЕЗУЛЬТАТЫ ПРИ ИСПОЛЬЗОВАНИИ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

Размер скользящего окна, пикс.	**Структура сети**	**F_β, %**	**T, с/Мп**
a = 1	3–3–1	96.9±0.2	0.60
	3–6–1	97.1±0.2	0.72
a = 3	27–7–1	97.0±0.2	2.61
	27–13–1	97.1±0.2	2.85
	27–27–1	**97.3**±0.2	3.17
	27–54–1	97.2±0.2	4.56

ТАБЛИЦА 3
РЕЗУЛЬТАТЫ ПРИ ИСПОЛЬЗОВАНИИ ПРОЧИХ МОДЕЛЕЙ, a = 3

Метод	**Параметры**	**F_β, %**	**T, с/Мп**
DT	7 листов	95.7±0.4	0.99
	35 листов	95.1±0.4	1.08
SVM	линейное ядро	94.3±0.5	1.86
LDA	–	96.4±0.3	2.24

Для осуществления этапа локализации необходимо получить последовательность операций, которые максимально уменьшат количество исследуемых пикселей при этом не уменьшат полноту. Данная последовательность операций находилось с помощью разработанного генетического алгоритма (ГА). Для реализации ГА были созданы функции получения изображений по данным расположения чисел; непосредственного выполнения генетического алгоритма; вычисления функции приспособленности; реализации генетических операций мутации и скрещивания; вычисления центромеры хромосомы; удаления пустых генов.

В качестве функции приспособленности использовалось произведение полноты найденных центров отметок на отношение IoU (intersection over union) между полученным бинарным отношением и идеальным, полученным из маски центров отметок с помощью дилатации со структурным элементом (СЭ) «квадрат» размером 7×7.

В качестве хромосомы использовалась изменяемая по длине последовательность операций-генов, каждый из которых обладал следующей структурой: имя операции и параметры операции. В качестве операций использовались морфологические операции (дилатация, эрозия, открытие, закрытие), пороговая обработка и фильтрующие операции (по площади, эксцентриситету, длине осей эквивалентных эллипсов, ориентации) [12]. В качестве параметров операций могли быть названия СЭ или границы для фильтрации.

Параметры генетического алгоритма:

– размер популяции – 100 хромосом;
– максимальное количество итераций – 100;
– доля элитных хромосом, переходящих в следующее поколение – 30%;
– вероятность скрещивания – 10%;
– вероятность мутации – 90%;
– вероятность вставки нового гена – $\frac{5}{N_{\text{генов}}}$ (при $N_{\text{генов}} > 5$, иначе 1);
– вероятность удаления гена – $0{,}05(1 - \frac{1}{N_{\text{генов}}})$ (при $N_{\text{генов}} > 5$, иначе 0);
– вероятность модификации гена – $0{,}95 - \frac{4{,}95}{N_{\text{генов}}}$ (при $N_{\text{генов}} > 5$, иначе 0);
– вероятность взять ген для мутации из левой части – 50%;
– вероятность взять ген для мутации из правой части – 50%.

ГА для получения наилучшей хромосомы запускался на одном изображении из тренировочных данных, так как на них присутствует достаточное количество отметок глубин: 926, всего цифр 1764. Результаты ГА при получении оптимальной хромосомы на этом изображении продемонстрированы на рис. 3.

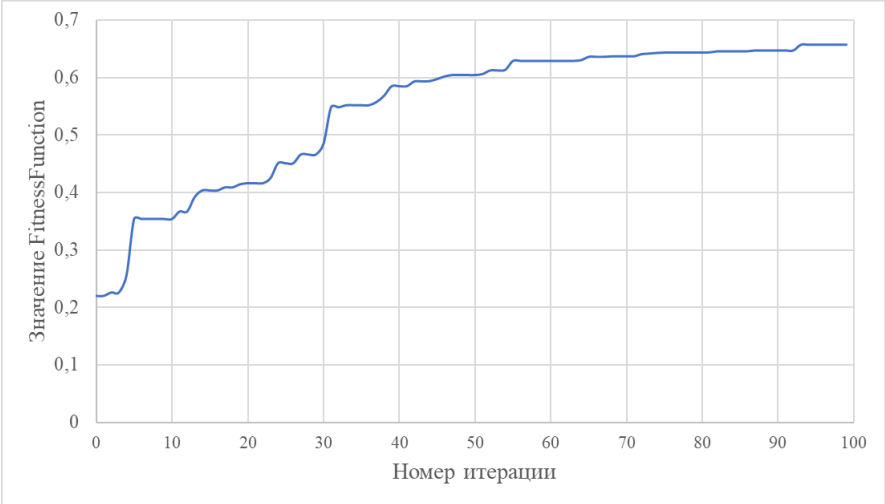


Рис. 3. График функции приспособленности от номера итерации ГА

Итоговая хромосома была запущена на всех 20 изображениях. Время обработки составило 0.16 с/Мп, значение функции приспособленности равно 64.4% ± 1.0%, полнота равна 96.1% ± 0.7%.

Разберём обработку одного изображения. После сегментации в нём содержится ~160 тыс. точек, отнесённых к классу, соответствующему отметкам глубин. При этом центр цифры может не относиться к принадлежащему ему сегменту (например, для цифр 0, 2 или 7), поэтому для таких сегментов часто применяют дилатацию с СЭ «квадрат» размера 5×5, при этом количество исследуемых пикселей увеличивается до 320 тыс. точек при полноте почти равной 100%. Если же применить полученный с помощью ГА список операций, то количество точек сокращается до 94 тыс., что приводит к уменьшению времени распознавания в 3.4 раза; в среднем по всем изображениям датасета это отношение составляет 3.7 ± 0.5. Пример запуска операций, полученных с помощью ГА, показан на рис. 4 в цветовом пространстве CMYK: координата M отвечает за изображение, полученное с этапа сегментации, C – выходное изображение после локализации.

Для этапа распознавания были обучены следующие модели: ИНС с одним скрытым слоем, дерево решений, k-NN классификатор, машина опорных векторов. Для k-NN и SVM количество тренировочных данных было уменьшено в 10 раз для соответствия времени работы другим моделям. Распределение на тренировочный, валидационный и тестовый датасеты происходило в пропорции 2:1:1.

В качестве обучающих данных выступали изображения цифр размера 13×11, полученные с этапа сегментации. В первом случае в качестве центров при обучении использовались только истинные центры цифр, во втором случае к ним добавлялись 4 соседних пикселя по окрестности фон Неймана, в третьем случае – 8 соседних пикселей по окрестности Мура (всего 9). Таким образом производилось исследование влияния аугментации на точность и полноту интеллектуальных моделей.

ИНС обучались с помощью метода сопряжённых градиентов, выходной слой имел функцию softmax. Деревья решений ограничивались в 100, 1000 и 10000 ветвлений. k-NN классификатор запускался для k, равных 1, 3 и 5

соседей. Для SVM использовалось только линейное ядро, так как нелинейные показывали значительное снижение быстродействия.

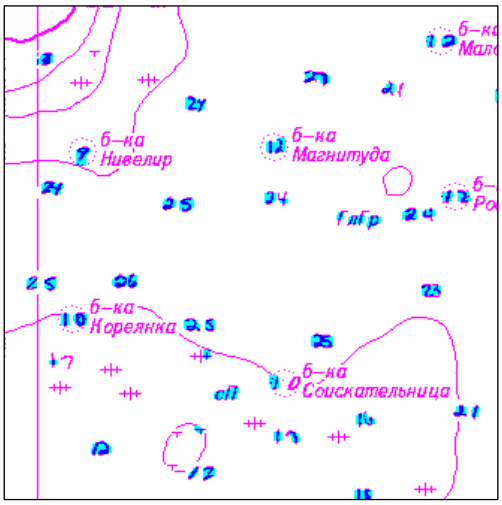


Рис. 4. Результат работы списка операций, полученных в результате ГА

Для обеспечения разделения сегмента, содержащего число, на несколько отдельных сегментов, соответствующих цифрам, дополнительно была обучена ИНС с бинарным надклассом «цифра/не цифра»: позитивными примерами класса выбирались пиксели из квадрата 3×3 относительно центра, негативными – вершины и середины сторон квадрата 7×7 относительно центра. Выход данной нейросети умножался на выход исследуемой интеллектуальной модели.

В таблице 4 представлены критерии качества и быстродействия первого каскада распознавания. Полнота рассчитывалась относительно центрального квадрата размера 5×5: если в данной области присутствовал хоть один пиксель, принадлежащий к правильному классу, то это увеличивало число правильно распознанных объектов. Точность рассчитывалась относительно всех пикселей, указанных к данному классу.

ТАБЛИЦА 4
РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ ПЕРВОГО КАСКАДА РАСПОЗНАВАНИЯ

Метод распознавания	**Recall, %**			**Precision, %**			**T, с/Мп**
	Обучающая выборка			Обучающая выборка			
	1	5	9	1	5	9	
ANN, 143–50–10	77.0	91.3	93.0	18.2	26.3	28.8	0.55
ANN, 143–100–10	78.7	92.0	93.4	18.3	27.0	29.9	0.64
ANN, 143–200–10	79.0	92.7	94.0	**18.8**	27.1	32.8	0.74
ANN, 143–400–10	78.4	93.4	94.2	18.6	**27.2**	**33.4**	0.98
DT, 100	87.7	90.9	91.5	8.9	9.5	10.0	0.15
DT, 1000	**90.7**	93.9	93.6	8.3	10.3	10.6	
DT, 10000	**90.7**	**95.3**	**95.5**	8.3	10.0	11.1	
kNN, 1	90.6	92.1	92.2	9.2	9.7	11.0	1.1
kNN, 3	86.5	86.3	87.6	9.9	13.2	12.9	
kNN, 5	87.2	87.5	85.5	10.5	12.8	14.2	
SVM	81.3	80.0	75.6	10.7	12.0	12.6	1.8

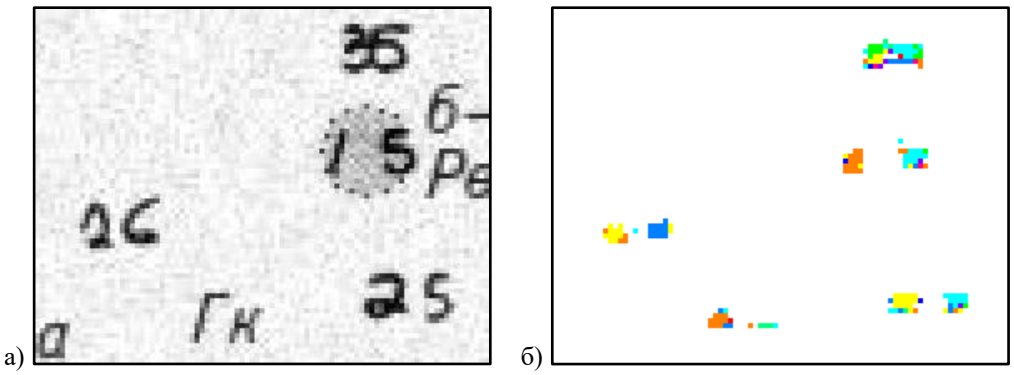
На рис. 5 показан пример работы первого каскада алгоритма для выбранной ИНС, имеющей архитектуру 143–400–10, обученной с использованием аугментации по окрестности Мура. Рис. 5, а) представляет собой сегментированное изображение, а рис. 5, б) изображён в индексном пространстве: белым цветом обозначается фон, пиксели, относящиеся к центру цифры «1» показаны оранжевым, «2» – жёлтым, «5» – циановым, «6» – небесно-голубым и т. д. Видно, что обозначение грунта «Гк» также определилось как цифра «1», что можно объяснить схожестью начертания в скользящем окне 13×11. Так как цифры в числе «35» оказались расположены слишком близко друг к другу, то их сегменты не были разделены на этом этапе, поэтому перед работой второго каскада распознавания были проведены следующие операции над получившимися суперсегментами:

– удаление из них сегментов, состоящих из одного пикселя (в данной операции использовалась окрестность Мура, для остальных – фон Неймана);

– удаление суперсегментов, состоящих меньше, чем из 5 пикселей;

– разбивка сегментов, состоящих больше, чем из 28 пикселей, с помощью разреза посередине.

Обучение интеллектуальных моделей для второго каскада происходило аналогично предыдущему. В качестве обучающих данных были произведены операции первого каскада распознавания над тренировочными изображениями. Для каждого суперсегмента были найдены доли вхождения в него десяти искомых классов.



а) б)

Рис. 5. Пример работы первого каскада распознавания:
а) сегментированное изображение; б) полученные суперсегменты

Критерии качества работы второго каскада распознавания показаны в таблице 5, где кроме интеллектуальных моделей производилась также простая максимизация. Время работы всех моделей оказалось примерно равным 0,2 с/Мп. Отличие критерия качества Common Precision от Identity Precision заключается в том, что Common Precision рассчитывается в том числе для случая, когда распознаваемый образ отсутствует, а Identity Precision рассчитывается только по множеству образов, которые существуют. Таким образом можно оценить уровень всех трёх типов ошибок, упомянутых ранее. Формулы для полноты и двух данных показателей точности могут быть получены с использованием матрицы неточностей (confusion matrix):

$$\text{Recall} = \frac{\sum_{i=1}^{10} \text{cm}(i,i)}{\sum_{i=1}^{10}\sum_{j=0}^{10} \text{cm}(i,j)},$$

$$\text{Common Precision} = \frac{\sum_{i=1}^{10} \text{cm}(i,i)}{\sum_{i=0}^{10}\sum_{j=1}^{10} \text{cm}(i,j)},$$

$$\text{Identity Precision} = \frac{\sum_{i=1}^{10} \text{cm}(i,i)}{\sum_{i=1}^{10}\sum_{j=1}^{10} \text{cm}(i,j)},$$

где $\text{cm}(i,j)$ – количество объектов, которые были отнесены к классу j, и на самом деле относятся к классу i. $\text{cm}(0,j)$ означает, что количество объектов, отнесённых к классу j, которых не существует (ложное срабатывание), $\text{cm}(i,0)$ означает количество объектов класса i, которые были пропущены.

Завершающим этапом является группировка данных в виде объединения цифр в числа. После первичной агломеративной кластеризации, которая была реализована с помощью геометрического хеширования, была проведена фильтрация образовавшихся кластеров, отвечающая за контекстный и семантический анализ:

– удаление отметок глубин, состоящих из одной цифры и не располагающихся на голубом фоне, означающем мелководье;

– удаление отметок, состоящих из трёх и более цифр;

– удаление отметок со значениями, отличающимися больше чем на 15 от соседних, найденных с помощью триангуляции Делоне.

В таблице 6 показаны итоговые результаты применения фильтраций. Критерии качества показаны для чисел в целом, а не для отдельных цифр. Время работы, включая все этапы фильтрации, составило 0.14 с/Мп.

ТАБЛИЦА 5
РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ ВТОРОГО КАСКАДА РАСПОЗНАВАНИЯ

Метод распознавания	**Recall, %**	**Common Precision, %**	**Identity Precision, %**
ANN, 10–10–10	84.4	82.2	92.8
ANN, 10–20–10	84.8	82.6	93.3
ANN, 10–100–10	**85.0**	**82.9**	**93.6**
DT, 10	59.2	57.7	65.1
DT, 20	66.2	64.5	72.8
DT, 100	83.6	81.4	91.9
DT, 1000	83.3	81.2	91.6
kNN, 1	83.2	81.1	91.5
kNN, 3	84.3	82.1	92.7
SVM	82.0	79.9	90.2
max	79.8	77.8	87.8

ТАБЛИЦА 6
РЕЗУЛЬТАТЫ ГРУППИРОВКИ И ФИЛЬТРАЦИИ ОБРАЗОВ

Этап	**Recall, %**	**Common Precision, %**	**Identity Precision, %**
Первичная кластеризация	77.4	70.6	78.4
Удаление отметок, состоящих из одной цифры на неголубом фоне	77.4	80.9	86.6
Удаление отметок, состоящих из трёх и более цифр	77.4	82.9	87.7
Удаление сильно отличающихся отметок по сравнению со соседними	77.4	85.0	89.1

V. Обсуждение результатов

Результаты сегментации показали, что усложнение интеллектуальных моделей приводит к увеличению времени работы, но не всегда приводит к значимому улучшению критерия качества. Это, в том числе, можно объяснить тем, что в данном исследовании важным считается только один класс. При многоклассовой сегментации показатели качества в среднем будут ниже, поэтому там имеет смысл усложнять обрабатывающие модели. В качестве базовой модели для сегментации была выбрана ИНС со структурой 27–27–1.

Критерии качества, полученные для этапа локализации соответствуют тому, что из всех отметок глубин только 3.9% будут потеряны, что является достаточно хорошим результатом, так как ССИ обычно обладают значительной степенью избыточности, значения в точках этих отметок могут быть восстановлены с достаточной точностью при помощи интерполяции. Значение функции приспособленности в 64.4% означает, что ложноположительных ошибок будет не больше 50% от количества пикселей, принадлежащих идеальному размеченному изображению.

По анализу таблицы 4 видно, что ИНС, обученные на центрах, взятых из окрестности Мура, и дерево решений сохраняют достаточно высокий уровень полноты, но точность гораздо выше у нейросетей. Отметим, что низкий уровень точности (~30%) будет улучшен после второго каскада и дальнейшей фильтрации на уровне группировки. Полнота при увеличении аугментации значительно росла только у ИНС, точность увеличивалась у всех моделей, что говорит о том, что данный способ увеличения тренировочных данных имеет положительный эффект. Так как на втором каскаде обрабатываются более высокоуровневые объекты с небольшой размерностью входных параметров, то время работы всех моделей оказалось примерно равным, разница в показателях точности не является существенной, лучшие показатели соответствуют ИНС со структурой 10–100–1.

Результатом этапа группировки и всего исследования является то, что из всех отметок с помощью полученного алгоритма можно верно определить 89%, среднее время работы на обработку одной карты составило 18 секунд (приведённое время 4.3 с/Мп). Данный результат сравним с качеством распознавания человека для отметок с печатным шрифтом в два раза большего размера: в [13] точность равна 82% для отметок высотой до 25 пикселей. Также стоит учитывать размеры отметок, используемых в данном исследовании, и способ их размещения: в настоящее время почти все научные работы касаются обработки цифр из базы MNIST, не масштабируя их и не учитывая пересечения их друг с другом.

VI. Выводы и заключение

В ходе работы были произведены теоретические и экспериментальные исследования проблемы анализа и обработки сложноструктурных изображений. В результате были получены следующие результаты.

1. Учитывая правила изображения морских навигационных карт, были смоделированы их изображения и получен датасет со всеми размеченными данными, с помощью которых можно оценить работу отдельных этапов.

2. Была сформирована методика, состоящая из нескольких этапов, которую можно использовать для анализа и обработки сложноструктурных растровых изображений, полученных автоматизировано, её работоспособность продемонстрирована для смоделированных изображений морских навигационных карт.

3. Разработанные алгоритмы исследованы экспериментально с использованием различных интеллектуальных моделей с получением следующих результатов на всех этапах:

3.1. сегментация – лучший результат (97.3%) показала нейронная сеть прямого распространения с одним скрытым слоем и структурой 27/27/1, что соответствует скользящему окну с шириной 3 и 27 нейронами на скрытом слое;

3.2. локализация – в результате запуска генетического алгоритма была получена последовательность морфологических и фильтрующих операций, благодаря которым была обеспечена полнота в нахождении центров отметок в 96.1%;

3.2. распознавание – для первого каскада лучший результат по точности показала нейросеть прямого распространения со структурой 143/400/10, что соответствует скользящему окну размера 13×11 и 400 нейронам на скрытом слое; для второго каскада лучший результат показала нейронная сеть прямого распространения со структурой 10/100/10.

4. группировка была произведена с помощью подбора параметров первичной кластеризации, выполненной с помощью геометрического хеширования, произведено удаление некорректных по размеру отметок и семантическая фильтрация с помощью отсеивания значений с использованием триангуляции Делоне.

В целом можно заключить, что поставленная цель исследования была достигнута и аналогичная методика может быть использована и для анализа и обработки сложноструктурных изображений других типов при необходимой доработке, учитывая особенности отображения этих ССИ.

Список литературы

1. Дмитриев Н. В. Методика исследования устойчивости обучения нейронных сетей (на примере задачи классификации образов на сложноструктурных изображениях) // Научно-технический вестник Поволжья. 2022. № 12. С. 108–111.

2. Szendrey R. A. A Knowledge-based Approach to Rastervector Conversion of Large Scale Topographic Maps. Acta Cybernetica. 2011. Vol. 20. No. 1. P. 145–165.

3. Broelemann K. Automatic understanding of sketch maps using context-aware classification / K. Broelemann, X. Jiang, A. Schwering // Expert Systems with Applications. 2016. V. 45. P. 195–207.

4. Egmont-Petersen M. Image processing with neural networks - a review / M. Egmont-Petersen, D. De Ridder, H. Handels // Pattern Recognition. 2002. V. 35. Pp. 2279–2301.

5. Доросинский Л. Г. Сравнительный анализ классических методов и методов машинного обучения при решении задач классификации радиолокационных изображений / Л. Г. Доросинский, С. С. Иванов // Ural Radio Engineering Journal. 2022. Т. 6. № 3. С. 310–323.

6. Nazari N. H. Integrating text recognition for overlapping text detection in maps / N. H. Nazari, T. Tan, Y. Y. Chiang // Proc. of Document Recognition and Retrieval, International Symposium on Electronic Imaging. 2016. P. 1–8.

7. Bhatt P. M. et al. Image-Based Surface Defect Detection Using Deep Learning: A Review // J. Comput. Inf. Sci. Eng. 2021. Vol. 21. 15 p.

8. Czimmermann T. et al. Visual-Based Defect Detection and Classification Approaches for Industrial Applications—A Survey // Sensors. 2020. Vol. 20. No. 5. P. 1459.

9. Дмитриев, Н. В. Метод восстановления значений на топографических картах для создания цифровой модели рельефа // Прикладная информатика. 2019. Т. 14. № 6(84). С. 71–79.

10. Fournier A. Et al. Computer rendering of stochastic models // Communications of the ACM. 1982. V. 25(6). P. 371–384.

11. Дмитриев, Н. В. Автоматизированная система распознавания символов на топографических картах // Научно-методический электронный журнал "Концепт". 2016. № Т11. С. 1881–1885.

12. Гонзалес, Р. Цифровая обработка изображений в среде Matlab / М.: Техносфера, 2006. – 616 с.

13. Netzer Y. et al Reading Digits in Natural Images with Unsupervised Feature Learning // NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2011.

УДК 004+91+165
DOI: 10.25206/978-5-8149-3873-2-2024-725-730

Методология визуального моделирования как компонент междисциплинарного исследования геопространства
Visual Modeling Methodology as a Component of Interdisciplinary Geospatial Research

Д. М. Кошлаков
ФГБОУ ВО «Брянский государственный технический университет», Брянск, Россия

D. M. Koshlakov
Bryansk State Technical University, Bryansk, Russia

Аннотация. В контексте проблем развития неогеографии в исследовании используются социотехноприродный подход к исследованию земной жизни и некоторые научные результаты, полученные в сфере визуальной эпистемологии. В качестве одного из ключевых элементов методологии мультидисциплинарного исследования геопространства рассматривается методология визуального моделирования. Делается вывод о том, что возможности визуального моделирования должны быть встроены в перспективные системы геоаналитики и поддержки принятия управленческих решений. Отмечается значение визуального моделирования в контексте решения исследовательских, аналитических, управленческих и учебных задач. Обсуждается идея симуляторов геопространственного управления. Указывается, что симуляторы такого рода могут применяться в учебной практике, а также в реальной управленческой практике при оптимизации сценариев геопространственного развития.

***Ключевые слова*:** визуализация, визуальное моделирование, геоаналитика, геовизуализация, геоинформатика, геопространственное управление, геопространство, мультидисциплинарное исследование, симулятор геопространственного управления, системы поддержки принятия управленческих решений, Цифровая Земля.

Annotation. In the context of the problems of neogeography development, the study uses a socio-techno-natural approach to the study of terrestrial life and some scientific results in the field of visual epistemology. The methodology of visual modeling is considered as one of the key elements of the methodology of multidisciplinary research of geospace. It is concluded that the capabilities of visual modeling should be built into promising systems of geoanalytics and support for management decision-making. The importance of visual modeling in the context of solving research, analytical, management and educational problems is noted. The idea of geospatial management simulators is discussed. It is indicated that simulators of this kind can be used in educational practice, as well as in real management practice when optimizing geospatial development scenarios.

***Keywords*:** Digital Earth, geoanaltics, geoinformatics, geospace, geospatial management, geospatial management simulator, geovisualization, management decision support systems, multidisciplinary research, visual modeling, visualization.

I. Введение

В нашем исследовании будет предпринята попытка совмещения двух актуальных философских подходов. Первый из этих исследовательских подходов – это визуальная эпистемология, направление эпистемологических изысканий, в рамках которого исследуется роль визуального компонента в познавательном процессе, в том числе при осуществлении реальных научных исследований. В качестве значимых для нас сочинений визуально-эпистемологической направленности можно отметить работы [1; 11; 15].

Второй подход, который будет взят за основу в нашем исследовании, – это социотехноприродный подход к исследованию земной жизни, разработанный в трудах Э.С. Демиденко, Е.А. Дергачевой, Н.В. Попковой и других авторов [4–7]. Обзор этого подхода дан в статье [12], а его разнообразные проекции на проблематику Цифровой Земли представлены в исследованиях [6–9, 13; 14].

II. Постановка задачи

Концепция Цифровой Земли стала принципиально новым этапом развития феномена геопространственной визуализации [8]. При этом принципиальность произошедших в этой связи эпистемических трансформаций столь высока, что имеются значимые основания к тому, чтобы говорить о «геопространственной революции» (в смысле теории Т.Куна, описывающей структуру научных революций) [10].

В любом случае, в самом первом приближении Цифровая Земля может рассматриваться как новая фаза развития феномена геовизуализации. При этом параллельно с развитием геовизуальной составляющей

концепции Цифровой Земли и реализуемых в ее рамках «разноскоростных» и «разнонаправленных» проектов (самым известным из которых является Google Earth) происходит развитие их геоаналитической составляющей.

Понятно, что Цифровая Земля строится не только как некий уникальный образовательный и эстетический ресурс, но и как некое новое средство познания мира, в том числе средство описания геопространства, его моделирования и анализа, а также прогнозирования его развития. В этих условиях, в принципе, возможны, по меньшей мере, два дополняющих друг друга направления развития обсуждаемой нами неогеографической проблематики. Первое направление развития будет, в частности, связано с усилением и углублением в рамках конкретных геоинформационных проектов их геоаналитической составляющей (и, соответственно, их геоаналитической направленности). В свою очередь, второе возможное направление развития будет обусловлено развитием таких внешних по отношению к Цифровой Земле геоаналитических ресурсов, которые в определенном смысле будут некой ее надстройкой, то есть, проще говоря, они будут черпать из Цифровой Земли информацию, по определенным алгоритмам ее перерабатывать и на выходе генерировать какие-то новые сведения и знания, в том числе актуальные в контексте принятия управленческих решений (социальных, экономических, градостроительных, политических и т.д. и т.п.). Примером такого рода управленческих задач и соответствующих им управленческих практик может служить геолого-экономический мониторинг в системе управления фондом недр, в том числе осуществляемый, например, в целях обеспечения воспроизводства минерально-сырьевой базы региона [2; 3]. Само собой, в этих условиях подобного рода деятельность должна замыкаться, в том числе, и на соответствующие системы геоаналитической деятельности и поддержки принятия управленческих решений.

Таким образом, есть основания говорить о множестве разнообразных геоаналитических систем и систем поддержки принятия управленческих решений, которые нуждаются и/или в перспективе будут нуждаться во взаимодействии с Цифровой Землей. При этом речь в данном случае может идти как о реально существующих системах, так и системах перспективных. По всей вероятности, особую значимость такого рода системы будут иметь в сферах экономики, маркетинга и менеджмента, государственного и муниципального управления, образования, а также в сфере экологического мониторинга и мониторинга процессов социотехноприродного развития [6; 7].

Организация взаимодействия и взаимоувязка подобного рода систем геоаналитики и поддержки принятия управленческих решений с концепцией Цифровой Земли поставит перед научным сообществом и специалистами в сферах управления и геоинформатики некоторое множество философских и методологических проблем и задач, от качества решения которых во многом будет зависеть эффективность управления (социального, экономического, политического, военного и т.д. и т.п.).

Как представляется, одной из ключевых эпистемологических задач при этом окажется задача развития методологии мультидисциплинарного анализа геопространства. Кроме того, возникнут и некоторые иные проблемы философского и методологического характера. Именно обсуждению некоторых из проблем такого рода будет посвящено наше исследование. При этом в качестве центрального звена методологии мультидисциплинарного анализа геопространства будет рассмотрена методология визуального моделирования.

Именно в таких общих формулировках может быть очерчена задача нашего исследования. Если же обратиться к более конкретным формулировкам, то ключевую задачу нашего исследования можно декомпозировать на некоторое множество частных задач:

- исследовать возможность применения методологии визуального моделирования в практике мультидисциплинарных исследований геопространства;
- выявить некоторые новые возможности, возникающие в связи с развитием концепции Цифровой Земли и расширением сферы ее применения в управлении и образовании, аналитической и научной деятельности;
- рассмотреть концептуальные основы развития и применения в образовательной и управленческой практике симуляторов геопространственного управления.

III. Теоретические аспекты проблемы

В контексте затронутой нами проблематики для прояснения нашей исходной методологической позиции мы сформулируем четыре семантически созвучных, но при этом принципиально значимых тезиса:

1. Геопространство является сложным и целостным феноменом, нуждающимся в глубоком и комплексном исследовании;

2. Сложность такого исследования приводит к тому, что у нас имеется целое множество различных вариантов географического знания. В этой связи принято говорить о своеобразном сосуществовании социальной, естественной (природной) и техногенной ветвей географического знания [8, с. 200].

3. Принципиальное значение для науки и аналитики имеет интеграция этих ветвей географического знания, причем одним из возможных вариантов такой интеграции может служить социотехноприродный подход к исследованию трансформации земной жизни [4; 5].

4. На частном, дисциплинарном уровне анализа можно выделить некоторое множество географических дисциплин, также нуждающихся в мультидисциплинарной интеграции. К числу таких дисциплин, например, могут быть отнесены такие области научного знания, как военная география, историческая география, лингвогеография, политическая география, социальная география, экономическая география и т.д. и т.п.

Проще говоря, целостность изучаемых нами объектов, в том числе нашей планеты и соответствующего ей геопространства, требует от нас применения таких подходов к познанию этих объектов, которые бы в ходе самого процесса познания (в том числе в ходе аналитической деятельности) не ликвидировали бы познаваемую нами целостность. Как представляется, к этому призывали нас и некоторые великие естествоиспытатели прошлого, и некоторые философы. К числу таковых можно, например, отнести И.В. Гёте, А. Гумбольдта, В.И. Вернадского, А.Ф. Лосева и некоторых других.

В любом случае, в рассматриваемом нами контексте поддержание целостности и даже комплексности и системности познавательного процесса является сложной и нетривиальной эпистемологической задачей. Одним из вариантов ее частичного решения является интеграция научного знания с той или иной философской доктриной, обладающей мощным интегративным потенциалом. Другим из возможных вариантов является осуществление мультидисциплинарного исследования того или иного типа. Естественно, все это сравнительно легко декларировать, но достаточно трудно осуществлять в реальной исследовательской и аналитической практике.

По нашему глубокому убеждению, помочь решению познавательных задач такого рода может методология визуального моделирования, в том числе рассматриваемая в качестве одной из составляющих методологии мультидисциплинарного исследования геопространства. Визуальное моделирование, с одной стороны, способно содействовать нам в осуществлении комплексных, системных и мультидисциплинарных исследований (в том числе исследований геопространства), а с другой стороны, может рассматриваться в качестве одного из компонентов методологии геоаналитической деятельности и соответственно должно быть встроено в перспективные системы геоаналитики, а также в системы поддержки принятия управленческих решений.

Визуальное моделирование систем и процессов (в том числе таких систем и процессов, которые плохо поддаются формализации) позволяет перевести подвергаемые нами изучению проблемы в некую совокупность более или менее интуитивно понятных и зрительно воспринимаемых (графических) образов, посредством чего стимулировать наши визуальное и образное мышление, интуицию и воображение.

Сам по себе концепт визуального моделирования можно ввести в оборот по аналогии с идеей математического моделирования. При этом в контексте обсуждения понятия визуального моделирования следует отметить, что данное понятие может быть тем или иным образом связано со следующими тремя типами визуализации: визуализацией знаний, визуализацией информации и визуализацией данных [1]. Учитывая, что из этих типов наименее распространенным и наименее проработанным является визуализация знаний, то наиболее важной представляется связь визуального моделирования именно с данным типом визуализации.

Визуальное моделирование (во всяком случае, в аспекте его связи с визуализацией знаний) предполагает построение на основе имеющихся у нас знаний, представлений и сведений некого множества визуальных моделей, описывающих тот или иной фрагмент реальности. В простейшем варианте визуальные модели представляют собой некоторые схемы, состоящие из какого-то количества блоков и тех или иных связей между блоками. При этом в реальной практике приходится строить не одну визуальную модель, а эволюционным образом разворачивать целые ансамбли и серии визуальных моделей, описывающих тот или иной фрагмент реальности под различными углами зрения и в своей совокупности образующих своего рода многомерные визуальные модели.

Зачем такого рода модели нам нужны? Как минимум, по тем причинам, что они:

- стимулируют мыслительный процесс, а также познавательную активность и аналитическую деятельность;
- позволяют увязывать воедино системы, события, процессы и феномены различной природы, чем помогают осуществлять комплексные и мультидисциплинарные исследования, что, например, может иметь существенное значение для социотехноприродного подхода к исследованию земной жизни;
- повышают уровень системности наших представлений о мире и о составляющих его объектах и явлениях.

Естественно, построение подобных моделей можно связывать с такими феноменами как извлечение знаний и их представление. В данном случае важно, пожалуй, то, что визуализация может быть входным ресурсом анализа, средством осуществления анализа и, наконец, его промежуточным или итоговым результатом. В каждом из этих случаев визуализация реализует в познавательном процессе свою специфическую функцию.

В любом случае, представляется, что визуальное моделирование должно нами рассматриваться как компонент методологии мультидисциплинарного исследования геопространства, а также некоторый значимый информационный ресурс, который в перспективе должен быть встроен в системы геоаналитики и системы поддержки принятия управленческих решений.

IV. Практические аспекты проблемы

Значимость визуального моделирования обуславливается тем, что оно позволяет или, во всяком случае, помогает осуществлять «сцепку» различных предметных сфер и существующих в пределах этих сфер феноменов, визуализировать эту «сцепку» и тем самым обеспечить возможность выхода на уровень междисциплинарного и в ряде случаев метадисциплинарного исследования.

В качестве элементарного примера визуальной модели приведем схему, изображенную на рис. 1. С помощью данной модели мы можем в самом первом приближении отразить те связи, которые возникают между общественным сознанием, развитием научного знания, развитием техносферы и экономическим развитием, а

также экологической проблематикой. При этом помимо прямых связей между некоторыми парами обозначенных нами на схеме феноменов предлагается также учитывать и обратные связи между ними.

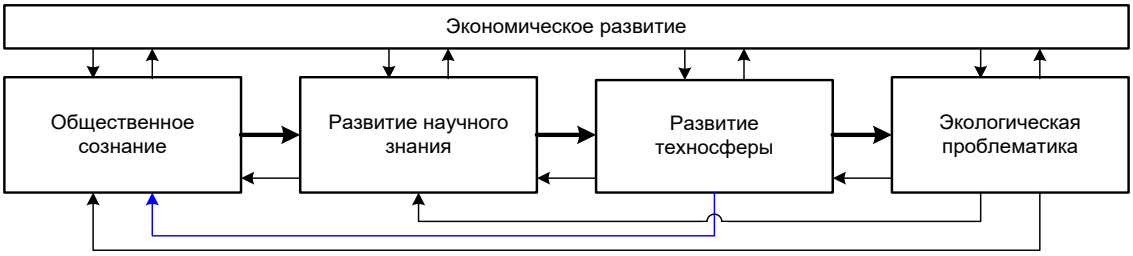


Рис. 1. Пример визуальной модели

Очевидно, что приведенная нами визуальная модель носит междисциплинарный характер и в зависимости от контекста решаемых нами задач (исследовательских, аналитических, управленческих или учебных) может быть преобразована в целое множество более точных, развернутых и конкретных визуальных моделей. Примерно таким образом может осуществляться визуальное моделирование систем и процессов в реальной научной, аналитической и управленческой практике.

В контексте нашего обсуждения следует отметить, что, по большому счету, речь идет еще и о том, что Цифровая Земля как некоторое множество динамических систем геопространственной визуализации, а также присущий им функционал, определенным образом должна быть встроена и в процессы принятия управленческих решений, и в актуальные системы поддержки принятия управленческих решений. Коль скоро все это так, то возникает определенное множество дополнительных вопросов, обсуждение которых в этой связи представляется необходимым.

Надо понимать, что Цифровая Земля в перспективе трансформирует и выведет на новый уровень и процессы геопространственного управления, и науки о таком управлении (равно как, вероятно, потребует создания новых научных дисциплин), и системы поддержки принятия управленческих решений, и системы геоаналитической деятельности. Вне всякого сомнения, все это потребует также и существенного пересмотра некоторых аспектов стратегии образования (образовательной деятельности).

Ключевой вопрос здесь состоит в том, что Цифровая Земля даст человеку огромную власть над миром, в силу чего придется готовить специалистов по геопространственному управлению, а также разрабатывать этику такого управления. Вопросы подобного рода являются вопросами двух ключевых задач образования – обучения и воспитания. Вполне очевидно, что эти вопросы ни в коей мере не должны выпасть из поля нашего профессионального внимания [7].

Касательно проблемы обучения специалистов в области геопространственного управления следует признать, что классического подхода к обучению (то есть классической дидактики) здесь будет принципиально недостаточно. Так, в классической дидактике в свое время был сформулирован и детально проработан принцип наглядности. В современной дидактике нам уже принципа наглядности становится мало, и педагоги говорят о том, что в настоящее время феномен наглядности дополняется явлением визуализации (в ее самых разнообразных аспектах и проявлениях).

Представляется, что в дидактике будущего (в том числе в связи с развитием проекта и концепции Цифровой Земли) принцип наглядности и феномен визуализации придется чем-то дополнять и укреплять, в том числе с учетом развития систем геовизуализации и геоаналитической деятельности.

Вполне вероятно, что на новый уровень своего развития выйдет использование в реальной учебной практике симуляторов, в том числе и бизнес-симуляторов. При этом надо отметить, что Цифровая Земля, в принципе, может служить базой для создания игровых и обучающих симуляторов нового поколения. В данном случае мы акцентируем внимание на том обстоятельстве, что задачи *геопространственного управления* потребуют создания на основе Цифровой Земли симуляторов нового поколения – условно говоря, симуляторов геопространственного управления. В круг задач симуляторов такого рода должно входить не только обучение специалистов в области геопространственного управления, но и имитационное моделирование различных сценариев развития Земли в аспекте проблемы геопространственного управления, а также применение методологии мультидисциплинарного анализа геопространства, в том числе с использованием средств визуального моделирования.

Говоря иными словами, задача симуляторов геопространственного моделирования состоит не только в оптимизации процесса обучения «менеджеров геопространства», но и в оптимизации самого процесса геопространственного менеджмента. Ясно, что обе эти задачи не просто переплетаются и перекликаются между собой, но, по сути, являются одной (двуединой) задачей. И обучение, и моделирование сценариев геопространственного развития, и управление геопространственным развитием (в том числе выбор оптимальных управленческих решений) будут требовать мощных систем имитационного моделирования, встроенных в симулятор, имеющий свои различные версии, в том числе учебную и профессиональную. Первая из названных версий необходима для эффективной организации процессов обучения, а вторая – для решения

профессиональных задач в сферах геоинформатики, геоаналитической деятельности и геопространственного управления.

В данном случае весьма важно также указать на проблему этики геопространственного управления и на необходимость ее воплощения и своеобразного воспроизводства в геопространственном управлении, а также в сфере образования, в том числе при освоении обучающимися навыков этичного геопространственного управления.

Указанная нами этика геопространственного управления пока еще только ждет своих исследователей, но ясно, что развитие и применение возможностей Цифровой Земли вопросы такого рода сделает чрезвычайно актуальными. В более же широком аспекте нельзя исключать и того, что потребуется создание таких новых областей философского знания, как онтология геопространственного управления, эпистемология геопространственного управления и аксиология геопространственного управления.

V. Обсуждение результатов

В свете развития концепции Цифровой Земли чрезвычайно значимой оказывается проблема развития методологии мультидисциплинарного исследования геопространства. Сюда же следует включить проблему взаимоувязки различных концепций и научных дисциплин в их геопространственном аспекте. Проще говоря, значимой становится увязка неогеографии с самыми различными отраслями научного знания, в том числе такими как, например, архитектура, биология и экология, ветеринария, военные науки, геология и минералогия, искусствоведение, политология, аграрные науки, социология, фармацевтические науки, экономика и т.д. и т.п. В каждом из случаев подобной междисциплинарной увязки необходимо возникает свой специфический круг методологических вопросов и научных проблем. Более того, по всей вероятности, возникает также возможность проектирования если не новых отраслей знания, то, во всяком случае, новых научных дисциплин.

В любом случае, развитие концепции Цифровой Земли будет предъявлять новые требования к образовательной, аналитической и управленческой деятельности, ставить в этой сфере принципиально новые задачи. Одной из таких задач неизбежно станет развитие методологии мультидисциплинарного исследования геопространства. Развитие данной методологии будет иметь как существенное аналитическое значение, так и управленческое значение. При этом одним из значимых инструментов такой методологии, как представляется, следует считать методологию визуального моделирования.

VI. Заключение

Визуальное моделирование как некое средство описания и исследования явлений окружающего нас мира имеет смысл рассматривать в качестве одного из значимых компонентов методологии мультидисциплинарного исследования, в том числе исследования геопространства.

Применение в этой сфере визуального моделирования как некого методологического инструмента оказывается весьма актуальным, в результате чего следует говорить о том, что тем или иным образом данный инструмент должен быть встроен в перспективные системы геоаналитики и поддержки принятия управленческих решений, а также может использоваться в реальной исследовательской, аналитической, управленческой и учебной практике, особенно в тех случаях, когда речь идет о познании сложных геопространственных систем и процессов и об управлении такими системами и процессами. Данный методологический инструмент способен оказать позитивное влияние на развитие методологии мультидисциплинарного исследования геопространства, а также методологии геопространственного управления.

В обсуждаемом нами контексте достаточно перспективным следует признать создание симуляторов геопространственного управления, которые были бы связаны с Цифровой Землей и могли бы осуществлять исследование различных сценариев геопространственного развития, в том числе на основе методов имитационного моделирования с ориентацией на принятие оптимальных управленческих решений. При этом, как можно предположить, использование подобного рода симуляторов может иметь существенное значение для решения задач как образовательной деятельности (например, при подготовке специалистов в области геопространственного менеджмента), так и деятельности исследовательской, аналитической и управленческой.

Более чем очевидным представляется то, что построение и развитие таких симуляторов геопространственного управления, а также самой практики геопространственного менеджмента должно включать в себя в качестве одного из эффективных эпистемических инструментов методологию визуального моделирования систем и процессов.

Благодарности

Автор выражает глубокую и искреннюю благодарность В.В. Бураго, В.С. Дадыкину, Е.А. Дергачевой, Е.Н. Ерёмченко и Ю.Т. Трифанкову за плодотворные дискуссии, оказавшие существенное влияние на развитие концепции данного исследования.

Список литературы

1. Ардашкин И.Б. К вопросу о визуализации знания и информации: роль смарт-технологий // ΠΡΑΞΗΜΑ. Проблемы визуальной семиотики. 2018. Т. 18. № 4. С. 12-48. DOI: 10.23951/2312-7899-2018-4-12-48.

2. Дадыкин В.С. Методика организации геолого-экономического мониторинга в управлении недропользованием на территории Центрального федерального округа. Брянск: Новый проект, 2017. 197 с.

3. Дадыкин В.С. Методология геолого-экономического мониторинга в управлении государственным фондом недр, горнодобывающими предприятиями и минерально-сырьевым потенциалом. Брянск: Новый проект, 2017. 207 с.

4. Демиденко Э.С., Дергачева Е.А. От социально-техногенного к социально-биосферному повороту в аграрной истории России // Гуманитарные исследования Центральной России. 2023. № 2 (27). С. 34-44.

5. Демиденко Э.С., Дергачева Е.А., Попкова Н.В. Философия социально-техногенного развития мира: статьи, понятия, термины. Брянск: БГТУ; М.: Всемирная информ-энциклопедия, 2011. 387 с.

6. Дергачева Е.А. Цифровая Земля: от идей к воссозданию виртуальной копии планеты // Геоконтекст. 2020. Т. 8. № 1. С. 5-16. DOI: 10.30987/2686-8326-2020-5-16.

7. Дергачева Е.А., Колесник Т.А. Научно-философские проблемы образования в системе Цифровой Земли // Вестник Южно-Российского государственного технического университета. Серия: Социально-экономические науки. 2022. Т. 15, № 2. С. 233–246. http://dx.doi.org/10.17213/2075-2067-2022-2-233-246.

8. Ерёмченко Е.Н. Неоплатонизм, изложенный ясно, как солнце, и Цифровая Земля // Вестник Южно-Российского государственного технического университета. Серия: Социально-экономические науки. 2024. Т. 17, № 2. С. 197–209. http://dx.doi.org/10.17213/2075-2067-2024-2-197-209.

9. Ерёмченко Е.Н. Основания социо-техно-природного кризиса и Цифровая Земля / Гуманитарий Юга России. 2024. Т. 13. № 3 (67). С. 66–76. DOI: 10.18522/2227-8656.2024.3.5.

10. Ерёмченко Е.Н. Цифровая Земля: геопространственная революция и ее мировоззренческие последствия // Логос. 2023. Т. 33. № 1. DOI: 10.22394/0869-5377-2023-1-221-240.

11. Латур Б. Визуализация и познание: изображая вещи вместе // Логос. 2017. Т. 27. № 2. С. 95-156.

12. Трифанков Ю.Т., Дергачев К.В. Обзор трудов ученых брянской научно-философской школы исследований социально-техногенного развития мира и жизни // Научное обозрение. 2016. № 6. С. 90–111.

13. Dergacheva E. Visualizing Socio-techno-natural Processes: Issues and Challenges // CEUR Workshop Proceedings of the 29th International Conference on Computer Graphics and Vision (Graphicon 2019). Vol. 2485. P. 168-172. URL: http://ceur-ws.org/Vol-2485/paper38.pdf.

14. Dergacheva E.A., Demidenko E.S. Visualizing Global Socio-Technogenic Human Transformation: Digital Challenges of Living Earth // CEUR Workshop Proceedings of the 30th International Conference on Computer Graphics and Vision (Graphicon 2020). Proceedings of the 30th International Conference on Computer Graphics and Machine Vision. Vol. 2744. P. 1–14. DOI: https://doi.org/10.51130/graphicon-2020-2-3-44.

15. Klinke H. Introduction: the image and the mind // Art Theory as Visual Epistemology / ed. by H.Klinke. Cambridge: Cambridge Scholars Publishing, 2014. P. 1-10.

УДК 101.1:316:004.91
DOI: 10.25206/978-5-8149-3873-2-2024-731-733

Методология визуального моделирования социотехноприродных процессов
Methodology of visual modeling of sociotechnological processes

В.С. Калинин
Брянский государственный технический университет, Брянск, Россия

V.S. Kalinin
Bryansk State Technical University, Bryansk, Russian

Аннотация. Исследование развития глобальных цивилизационных процессов в современной социотехноприродной среде делает необходимым выработку адекватных средств его представления и моделирования. Оптимальным подходом является геопространственная визуализация, позволяющая консолидировать разнородную информацию в едином пространственно-временном континууме. В работе кратко рассматриваются возможные способы визуализации, делается вывод об отсутствии универсального подхода. Указывается, что оптимальной средой геопространственной визуализации является «Цифровая Земля», позволяющая обеспечить целостность образа глобальной обстановки и возможность использования как непосредственного представления геоконтекста, так и опосредованных знаками данных.

Ключевые слова: социотехноприродные процессы, визуальное моделирование, безопасность.

Аннотация. The study of the development of global civilizational processes in the modern sociotechnological environment makes it necessary to develop adequate means of its representation and modeling. The optimal approach is geospatial visualization, which allows you to consolidate heterogeneous information in a single space-time continuum. The paper briefly discusses possible visualization methods, and concludes that there is no universal approach. It is indicated that the optimal environment for geospatial visualization is the "Digital Earth", which allows to ensure the integrity of the image of the global situation and the possibility of using both the direct representation of geocontext a well as sign-mediated data.

Ключевые слова: sociotechnonatural processes, visual modeling, security.

I.Введение

Современная цивилизация развивается в находящихся в непрерывном взаимодействии глобальных социальной, технической и природной средах, причём техника и технологии активно преображают их. Характер и долгосрочные последствия изменений, вызванных техногенными процессами, трудно предсказуемы, поэтому изучение и анализ эволюционных процессов является актуальной и одновременно сложной задачей, требующей выработки специфических инструментов. Одним из них является визуализация — мощный инструмент наглядного представления трансформационных процессов, происходящих в мире. Наглядное представление и моделирование изменений в социальных, технических и природных средах позволяет максимально эффективно и быстро взаимодействовать учёным из разных тематических областей, формируя междисциплинарное поле для исследований и обеспечивая взаимодействие учёных, управленцев и общественности.

С методологической точки зрения визуальное моделирование процессов вообще и в социотехноприродной среде в частности основывается на прямой, не опосредованной знаковыми или иными условностями восприятии обстановки в актуальном пространстве, в котором локализация сущностей (явлений, объектов, процессов и т. д.) становится определяющим эвристическим фактором. В особенности преимущества подобного подхода проявляются при изучении сложных систем — таких, как единая социотехноприродная среда нашей планеты.

Визуализация не сводится только лишь к отображению уже собранных коллекций данных, представленных в знаковой форме и к упрощению их восприятия — она призвана выявлять тенденции, закономерности и перспективы эволюции системы, т. е. создавать новую информацию и способствовать генерированию новых идей. Этот подход используется для анализа, планирования и управления сложными системами, где различные компоненты взаимосвязаны. В настоящее время визуализация широко используется в частных предметных и особенно прикладных областях, позволяя отображать социально-экономическую, технологическую, экологическую и другую информацию при моделировании и глобальных, и локальных процессов. Существенной перспективной задачей для визуализации является, во-первых, формирование основы для очень широко междисциплинарного синтеза, вне которого изучение социотехноприродной среды нашей планеты невозможно. Во-вторых, она призвана решить задачу бесшовного представления обстановки в глобальной среде на всех масштабных уровнях. Средства визуализации активизируют мыслительные процессы

учёных из разных областей, направляя их к новым исследовательским задачам. Таким образом, визуализация способствует развитию междисциплинарных тем и сотрудничеству в научных исследованиях.

Решение данных задач представляет собой существенную научную проблему. Важный шаг к её разрешению сделал в конце XX века тогдашний вице-президент США (1993-2001) Альберт Гор, предложивший проект «Цифровая Земля» (Digital Earth) [1]. Его целью стало создание (т. е. визуализация) модели геопространственной среды, базирующейся не на привычных картах, но на данных дистанционного зондирования — космических и аэроснимках. В отличие от традиционных карт, «Цифровая Земля» предоставляет интерактивный, внемасштабный и всеракурсный вид различных объектов, с возможностью произвольного выбора наблюдателем позиции в геоцентрическом пространстве.

Первые реализации проекта «Цифровая Земля» появились ещё в 1990-е гг., однако настоящим его стартом стало появление в 2005 г. онлайнового сервиса Google Earth [2]. Он сразу стал и до сих пор остаётся эталоном практической реализации концепции. Во многом этого качества удалось достичь благодаря очень широкому использованию космических и аэроснимков высокого качества. Важнейшую роль в развитии «Цифровой Земли» сыграла огромная аудитория Google Earth, быстро перешагнувшая рубеж 1 млрд. пользователей [3]. Благодаря очень широкому использованию сервиса быстро выявился его коммуникационный и коллаборативный потенциал — модель Земли в компьютере, точная, полная и визуально достоверная и в целом, и в деталях позволила организовывать взаимодействие учёных, управленцев и простых пользователей на недостижимом прежде уровне и с гораздо более глубоким понимание социотехноприродного контекста, в котором оно происходит. Цифровая Земля стала новым подходом к решению широкого круга задач, в первую очередь междисциплинарных [4].

В настоящее время «Цифровая Земля» широко используется при решении различных задач — от геополитических до прикладных и частных. Однако, по мнению ряда учёных, остаётся недостаточно прояснённой научно-философская методология «Цифровой Земли». Это осложняет понимание угрозы со стороны перехода биосферы в искусственную постбиосферную жизнь и проблем развития человека и социума под постоянным воздействием техногенных факторов. Включение отображения социально-техногенных трансформаций в проект «Цифровая Земля» позволит более грамотно планировать переход к устойчивому (социотехноприродному) развитию [5-7].

Поэтому, грамотная и понятная, методологически подготовленная и обоснованная визуализация и моделирование позволит своевременно и эффективно принять решения по стабилизации и обеспечению глобального устойчивого развития в современных социальных, технологических и природных процессах.

II. Теория

В визуальном моделировании можно выделить пять последовательных шагов: сбор данных, построение модели, её анализ, принятие решений и мониторинг с последующим обновлением модели. Среда моделирования включает в себя три компоненты: социальные, технические и природные процессы. Взаимодействие этих процессов в социотехноприродных системах требует интегративного подхода для достижения гармоничного состояния в отношении «общество-техносфера-биосфера». Поскольку моделирование неизбежно осуществляется в пространственной (в случае динамических процессов — в пространственно-временной) среде, средства моделирования следует типологизировать по критерию режимов представления в них геопространственных данных.

Одним из таких типов следует считать геоинформационные системы (ГИС). Они представляют собой построенные на базе картографического подхода системы по сбору, хранению, анализу и визуализации географических данных. Современные ГИС в полной мере гипермедийны, т. е. позволяют агрегировать в систему любую произвольную информацию об отображённых на географической карте объектах. Будучи в своей картографческой основе двумерными, они могут представлять пользователю также перспективный вид с учётом рельефа (т. н. 2,5D). Использование цифровых форматов представления данных позволяет интегрировать ГИС в цифровую среду интернет (так называемые картографические веб-сервисы). Существует огромное количество программных сред для ГИС (в частности, следует отметить в первую очередь открытое программное обеспечение Open GIS, проприентарные системы ArcGIS производства американской компании ESRI или аналогичную по функциональным возможностям систему Super Map производства одноимённой китайской компании). ГИС применяются во всех областях науки, техники и общественного развития практически без исключения: например, существуют ресурсные ГИС, экологические ГИС, транспортные ГИС, ГИС для общественной безопасности и обороны и т.д.

Визуальные модели, которые можно использовать для анализа опасностей в социотехноприродных процессах, могут включать следующие типы:

- карты рисков;
- графики и диаграммы;
- тепловые карты.

Рассмотрим каждый тип визуального моделирования и его применение.

Карты рисков представляют собой текстовое или графическое представление различных видов рисков, и угроз, в т.ч. ещё не идентифицированных с определёнными факторами (например, в случае предупреждения

эпидемий и пандемий). Этот вид визуализации удобно будет использовать, например, при отображении влияния техносферы на здоровье человека, ущерба, наносимого почвенному покрову технической деятельностью, и т.д.

Диаграммы сценариев представляют собой графические схемы, которые описывают различные возможные развития событий на основе определенных исходных условий и допущений. Одним из преимуществ использования данной диаграммы является оценка потенциальных рисков. С помощью них можно отображать, например, воздействие изменения климата на водные ресурсы и на биологическое разнообразие. Графики и диаграммы в силу формата представления данных в условных координатах менее наглядно отображают обстановку, нежели, например, 3D-модели, фактически представляющие собой те же трёхмерные графики, в которых данные погружены в богатый геопространственный контекст. Вместе с тем графики и диаграммы позволяют эффективно визуализировать большие объемы данных, определять тенденции и аномалии. С помощью них можно, например, отобразить динамику роста населения, изменение уровня моря. Тепловые карты – это мощный инструмент, который помогает анализировать и интерпретировать сложные взаимодействия в системах. Позволяет эффективно отображать интенсивность и распределение различных параметров на географических или абстрактных пространствах. Тепловые карты можно использовать при визуализации очаговых вспышек инфекционных заболеваний, повреждения почвенного покрова посредством применения токсичных удобрений и технических средств, распространения различных веществ от электростанций, предприятий и т.д.

Интегральным средством представления и моделирования процессов в глобальном пространственно-временном континууме является неогеография. Согласно российскому определению, она объединяет в себе средства и методы работы с геопространственной информацией, отличающиеся от предыдущих (карт и ГИС) использованием географических систем координат, растровой графики и использованием открытых форматов представления геоданных [8].

III. Выводы и заключение

Ни один из выше перечисленных методов визуализации и представления процессов в социотехноприродной среде не является универсальным, эвристически значимого результата можно ожидать только от комплексного применения этих и иных перспективных средств визуализации и подходов к ней. Наиболее перспективной средой такого моделирования можно считать Цифровую Землю (неогеографию) в силу её гетерогенного формата и возможности представления геопространственного контекста как в опосредованной знаками, так и в воспринимаемой непосредственно, с использованием не-знаковых средств моделирования обстановки [9].

Список литературы

1. Gore A. The Digital Earth: Understanding our planet in the 21st Century [Electronic resource]. 1998. 3 p. URL: https://www.asprs.org/wp-content/uploads/pers/99journal/may/1999_may_highlight.pdf (date accessed: 19.02.24).
2. Google Earth https://earth.google.com
3. Askay S. Google Earth turns 10 today // Google Earth. 29.06.2015. URL: https://blog.google/products/earth/google-earth-10th-birthday/
4. Annoni A. Digital Earth: yesterday, today, and tomorrow / A. Annoni, S. Nativi, A. Çöltekin, C. Desha, E. Eremchenko, C. M. Gevaert, … S. Tumampos // International Journal of Digital Earth. – 2023. – № 16(1). – P. 1022–1072. – DOI: 10.1080/17538947.2023.2187467.
5. Дергачева, Е. А. Новые закономерности развития в условиях социотехноприродного развития мира и смены направленности эволюции жизни / Е. А. Дергачева // Глобалистика-2020: Глобальные проблемы и будущее человечества : Сборник статей Международного научного конгресса, Москва, 18 мая – 24 октября 2020 года. – М: Межрегиональная общественная организация содействия изучению, пропаганде научного наследия Н.Д. Кондратьева, 2020. – С. 876-880.
6. Дергачева Е.А., Колесник Т.А. Научно-философские проблемы образования в системе Цифровой Земли // Вестник Южно-Российского государственного технического университета. Серия: Социально-экономические науки. 2022. Т. 15, № 2. С.233–246.
7. Dergacheva, Elena & Demidenko, Eduard. Visualizing Global Socio-Technogenic Human Transformation: Digital Challenges of Living Earth. Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020). Part 2. paper44-1.
8. Аноприенко, А. Я. Неогеография и постбинарный компьютинг / А. Я. Аноприенко, Е. Н. Еремченко // Наукові праці Донецького національного технічного університету. Серія: Проблеми моделювання та автоматизації проектування. – 2008. – № 7(150). – С. 249-257.
9. Komac B., Gašperič P. Cartographic time travel: Reflecting the past, defining the present, and challenging the future using old maps. Acta Geographica Slovenica. 2023; 63(2):9–21. DOI: https://doi.org/10.3986/AGS.13292

Секция 7

ГЕОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ. КОМПЬЮТЕРНАЯ ГРАФИКА

Geometric Modeling. Computer Graphics

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 621.01
DOI: 10.25206/978-5-8149-3873-2-2024-735-740

Использование перемещений оси схватоносителя по линейчатой поверхности ограничивающей телесный угол сервиса антропоморфных роботов для преодоления тупиковых ситуаций

The use of displacement of the contraction axis along a linear surface limiting the bodily angle of service of anthropomorphic robots to overcome deadlocks

Ф.Н. Притыкин [1]
[1] *Омский государственный технический университет, проспект Мира 11, город Омск, 644050, Россия*

F.N. Pritykin [1]
[1] *Omsk State Technical University, Mir Avenue 11, Omsk, 644,050, Russia*

Аннотация. Предложен алгоритм синтеза движений руки антропоморфных роботов по вектору скоростей при установке объектов манипулирования заданных, в виде прямоугольных призм в контейнер. Алгоритм позволяет осуществлять выход из возникающих тупиковых ситуаций при компьютерном моделировании движений. Суть метода состоит в использовании перемещений руки, при которых ось схватоносителя перемещается и образует с некоторым приближением линейчатую поверхность, задающую телесный угол сервиса. С этой целью предложено использовать базу данных конфигураций, задающих определенные положения центра выходного звена и осей схватоносителя совпадающих с образующими указанных линейчатых поверхностей. Представлены результаты расчетов промежуточных конфигураций при компьютерном моделировании движения антропоморфных роботов в организованной среде с использованием разработанного алгоритма.

Ключевые слова: Механизмы манипуляторов, мгновенные состояния механизмов, вектор обобщенных скоростей, линейчатые поверхности, компьютерное моделирование движений антропоморфных роботов.

Abstract. An algorithm for synthesizing hand movements of anthropomorphic robots by velocity vector during the installation of objects of manipulation given, in the form of rectangular prisms in a container, is proposed. The algorithm makes it possible to solve deadlock situations in computer simulation of movements. The essence of the method consists in the use of hand movements, at which the axis of the clash carrier moves and forms with some approximation a rule red surface, which specifies the body angle of service. For this purpose, it is proposed to use a database of configurations that specify certain positions of the output link center and of the accumulation carrier axes which coincide with the above-mentioned line surfaces. The results of calculations of intermediate configurations in computer modeling of movement of anthropomorphic robots in an organized environment using the developed algorithm are presented.

Keywords: Mechanisms of manipulators, instantaneous states of mechanisms, vector of generalized speeds, linear surfaces, computer simulation of movements of anthropomorphic robots.

Введение

В настоящее время актуальна разработка интеллектуальных систем управления механизмами рук антропоморфных роботов способных заменить человека при выполнении работ, связанных с оценкой рабочей сцены и принятия в соответствии с этим тех или иных решений [1]. При компьютерном моделировании движений роботов применяют метод, основанный на использовании матриц частных передаточных отношений [2,3]. С помощью данных матриц устанавливают взаимосвязь скоростей выходного звена (ВЗ) и обобщенных скоростей. При использовании указанного метода часто могут возникать тупиковые ситуации. Поэтому существует необходимость в модернизации указанного метода синтеза движений по вектору скоростей позволяющей преодолевать тупиковые ситуации при компьютерном моделировании движений антропоморфных роботов.

Постановка задачи

Исследуем алгоритм компьютерного моделирования движения антропоморфных роботов на примере, где необходимо размещать объекты манипулирования, имеющие форму прямоугольных параллелепипедов в

контейнер заданных размеров. На рисунке 1 представлены объекты манипулирования 3 различных размеров, которые перемещаются по конвейеру 4. Данные объекты необходимо размещать в контейнере 5 с использованием руки антропоморфного робота 2. На рисунке 1 приведено условное изображение руки антропоморфного робота с обозначением осей и поворотов во вращательных кинематических парах, задающих значения обобщенных координат q_2 - q_8. Рука антропоморфного робота также может смещаться поступательно вдоль направляющей 1 за счет изменения обобщенной координаты q_1.

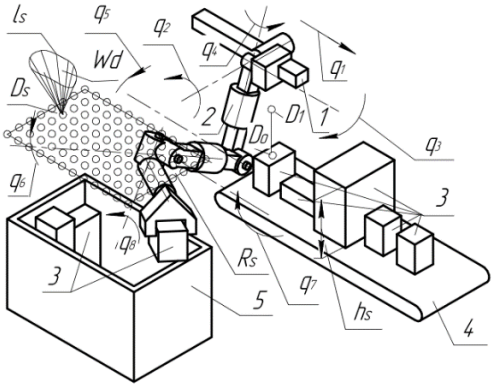
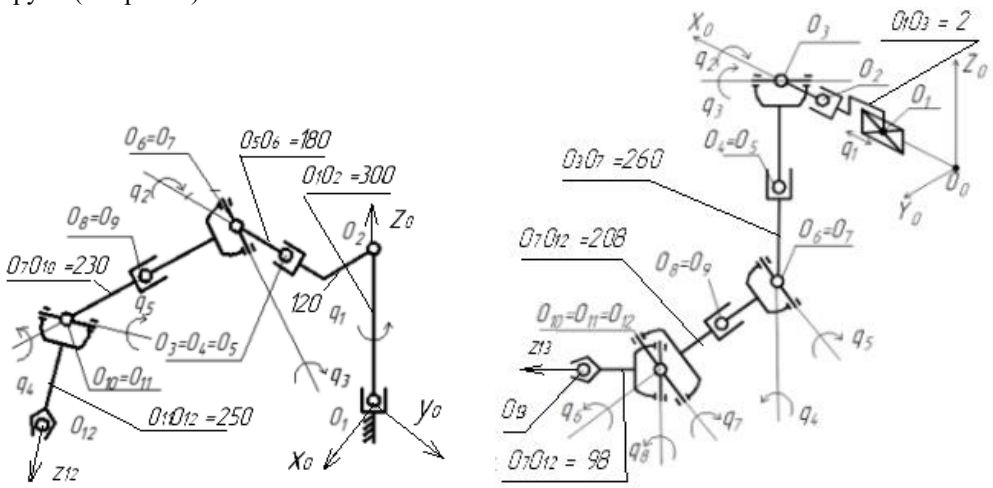


Рисунок 1 – Положение объектов манипулирования, контейнера, конвейера и руки антропоморфного робота

Кинематическая схема этого механизма изображена на рисунке 2б. При автоматизированном управлении движением механизма руки антропоморфного робота в сложно организованном пространстве необходимо обеспечивать достижимость захватом заданных объектов манипулирования на конвейере и установку их в контейнер. Заметим, что боковые стенки контейнера и установленные в контейнере объекты манипулирования (прямоугольные призмы) выступают в качестве запретных зон при компьютерном управлении движением механизма руки (см. рис. 1).



а б

Рисунок 2 – Кинематические схемы: а – механизм руки и туловища робота AR-600E (М3-12-10-8-3-12-2-12-3-12-2-12), nm = 12, n = 5;
б – механизм руки робота, nm = 13, n = 8 (М4-1-2-12-3-12-1-12-3-12-1-2-12)

Решение указанной задачи при возникновении тупиковых ситуаций при автоматизированном синтезе движений по вектору скоростей возможно с помощью осуществления движения механизма руки, при котором ось схватоносителя описывает линейчатую поверхность W_d, которая приближенно ограничивает телесный угол сервиса [4]. Тупиковая ситуация возникает, когда алгоритм синтеза движений не способен вычислить следующую промежуточную конфигурацию при смещении центра ВЗ в следующую точку заданной траектории. После возникновения тупиковой ситуации компьютерное моделирование движения осуществляется вновь от ближайшей точки D_s. При этом движении ось схватоносителя (оси $O_{12}z_{12}$ или $O_{13}z_{13}$ см. рис. 2) перемещается по образующим линейчатой поверхности W_d . Точка D_s находится в горизонтальной плоскости R_s, которая удалена от контейнера на расстояние h_s. Расстояние h_s задается в соответствии с наибольшей возможной высоте объектов манипулирования.

На рисунке 1 представлена плоскость R_s в которой находятся совокупность целевых точек D_s в которые первоначально перемещается центр ВЗ заданный точкой D_1. Совокупность точек D_s образуют равномерную сеть точек в плоскости R_s. На рисунке точка D_0 задает точку, в которую смещается центр ВЗ при захвате объекта манипулирования. После захвата объекта манипулирования центр ВЗ из точки D_0 перемещается в точку D_1.

Плоскость R_s и совокупность точек D_s используют для создания базы данных множества S конфигураций K_s (заданных значениями обобщенных координат q_i) руки антропоморфного робота полученных при перемещении центра ВЗ из точки D_1 в соответствующие точки D_s. При смещении центра ВЗ из точки D_1 в точку D_s (с использованием перемещения по критерию минимизации объема движения [5]) однозначно определяются конечные значения обобщенных координат q_i задающих множество S конфигураций K_s количество которых равно числу точек $D_s \in R_s$. База данных также содержит информацию о положениях множеств S_{DS} конфигураций K_{su} задающих положения механизма руки, при которых ось схватоносителя находится на линейчатой поверхности W_d для каждой отдельной конфигурации K_s множества S. Положение центра ВЗ заданного точкой D_1 обеспечивается строго заданной конфигурацией K_1. При создании базы данных множества S_{DS} (состоящих из двенадцати конфигураций K_{su} в соответствии с исследуемым тестовым примером) для каждой отдельной точки D_s используется метод, основанный на исследовании реализаций мгновенных состояний механизма руки антропоморфного робота [4]. Движение оси схватоносителя задающего угол сервиса (с вершиной в точке D_s) необходимо для изменения исходного положения звеньев руки по отношению к граням боковых стенок контейнера и граням установленных объектов манипулирования с целью обеспечения выхода из тупиковых ситуаций.

При определении конфигураций K_{su} руки при которых ось схватоносителя совпадает с образующими поверхности W_d и задает телесный угол сервиса, центр ВЗ остается в заданной точке D_s с некоторым допущением. Это допущение задается принадлежностью центра ВЗ заданной сфере при реализации мгновенных состояний [4]. При этом радиус сферы равен заданной величине точности позиционирования центра ВЗ δ. С целью сокращения времени вычислений промежуточных конфигураций (при возникновении тупиковых ситуаций) разработан алгоритм, при котором исключается итерационный поиск значения вектора обобщенных скоростей (вектора приращений обобщенных координат за одну итерацию).

Построение движений механизмов рук антропоморфных роботов

Рассмотрим кинематические схемы антропоморфных роботов, представленных на рисунке 2аб. На рисунке 2а задана кинематическая схема шестизвенного механизма руки и туловища антропоморфного робота AR-600E. В соответствии с методикой обозначения структурных моделей открытых кинематических цепей, принятой в работе [5] данный механизм, имеет обозначение М3-12-10-8-3-12-2-12-3-12-2-12. Соответственно на рисунке 2б представлен девятизвенный механизм руки антропоморфного робота (М4-1-2-12-3-12-1-12-3-12-1-2-12).

На указанном рисунке узловые точки задающие начала систем координат, связанных со звеньями механизма обозначены точками O_1, O_2, … , O_{nm}. Где параметр nm определяет число систем координат, используемых при задании модели кинематической цепи [5]. Расстояния между некоторыми узловыми точками на рисунке заданы длинами отрезков O_1O_2 = 300 (см. рис. 2а), и т.п. в миллиметрах. Известно, что совокупность мгновенных состояний, удовлетворяющих заданной точности позиционирования δ, определяет область $\Omega^Q \subset \Gamma^Q$ в многомерном пространстве обобщенных скоростей Q. Где Γ^Q обозначена p-плоскость заданная линейной системой уравнений задающей зависимость скоростей простейших движений ВЗ от обобщенных скоростей $\dot{q}_i$ (верхний индекс обозначений ^{Q} означает исследование и задание геометрических объектов Ω^Q и Γ^Q в многомерном пространстве Q, а знак $\subset$ определяет принадлежность). Указанная линейная система уравнений имеет следующий вид [3,5]:

$$\boldsymbol{JQ} = \boldsymbol{V}.$$

Или в развернутом виде:

$$\begin{gathered} J_{11}\dot{q}_1 + J_{12}\dot{q}_2 + \cdots + J_{1n}\dot{q}_n = v_x, \\ \cdots, \\ J_{r1}\dot{q}_1 + J_{r2}\dot{q}_2 + \cdots + J_{rn}\dot{q}_n = \omega_{z,} \end{gathered} \tag{1}$$

где $\boldsymbol{J}$ – матрица частных передаточных отношений размерности $r \times n$ задающая мгновенные передаточные отношения J_{11}, … , $J_{r\,n}$ [3]. Параметр r задает размерность вектора $\boldsymbol{V}$ простейших движений ВЗ и для общего случая эта размерность равна шести $\boldsymbol{V}(V_x, V_y, V_z, \omega_x, \omega_y, \omega_z)$ [3]. В зависимости от решаемой двигательной задачи параметр r может также принимать значения равные трем или пяти [5]. $\boldsymbol{Q}(\dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n)$ – вектор обобщённых скоростей, размерности n. Параметр n задает число обобщенных координат механизма q_1, q_2, … , q_n.

Область Ω^Q определяется множеством точек N^Q задающих допустимые мгновенные состояния механизма при которых $\delta_d < \delta$, где δ определяет заданную точность позиционирования (при проведении вычислительных экспериментов значение этого параметра принято δ = 5 мм). δ_d – погрешность линеаризации при использовании

системы уравнений (1) при расчете конфигураций. Исследованию размеров и формы областей Ω^Q для различных механизмов манипуляторов посвящены работы [6,7].

В настоящей работе предложен метод решения проблемы возникновения тупиковых ситуаций при поиске промежуточных конфигураций, когда возникают ситуации пересечения исполнительного механизма с запретными зонами. Суть метода состоит в использовании перемещений осей $O_{12}z_{12}$ и $O_{13}z_{13}$ при которых данные оси движутся по линейчатым поверхностям W_d приближенно ограничивающих телесный угол сервиса с вершинами в точках D_s. После возникновения тупиковых ситуаций начальное положение манипулятора выбирается такое, когда центр ВЗ совпадает с одной из точек D_s. Далее осуществляется синтез движений при котором оси $O_{12}z_{12}$ и $O_{13}z_{13}$ движутся по линейчатым поверхностям W_d до тех пор пока проекции узловых точек механизма O_{11} или O_{12} (задающие соответственно звенья $O_{11}O_{12}$ и $O_{12}O_{13}$ см. рис. 2аб) будут удалены на заданные расстояния l_g от проекций боковых стенок контейнера или боковых граней объектов манипулирования g_3 и т.п. (рисунок 3). Компоненты вектора $\boldsymbol{L_t}$ в подрисуночной надписи задают значения обобщенных координат q_i.

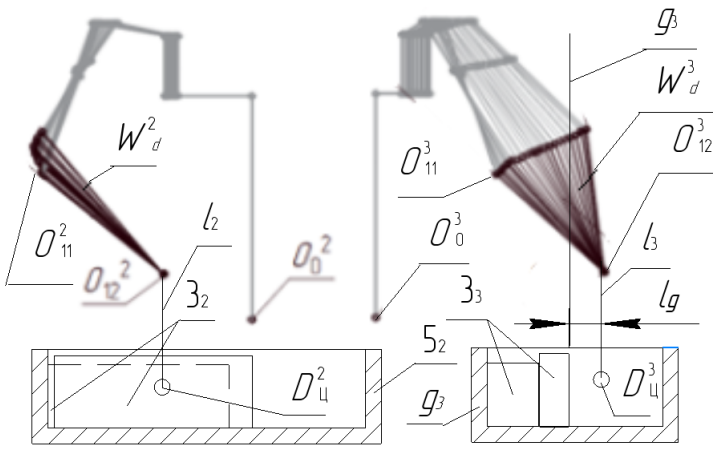


Рисунок 3 – Изменение положения и ориентации оси $O_{12}z_{12}$ робота М3-12-10-8-3-12-2-12-3-12-2-12 по отношению к боковым граням контейнера и граням объектов манипулирования с использованием перемещения данной оси по линейчатой поверхности W_d для конфигурации $\boldsymbol{L_t}$ (25°, 20°, -65°, 25°, -65°)

Расстояния l_g определяют во взаимосвязи с положением и размерами объектов манипулирования и положения боковых стенок контейнера. Заметим, что задачи определения телесных углов сервиса, ранее решались в работе [8].

Для обеспечения указанных движений осей $O_{12}z_{12}$ и $O_{13}z_{13}$ необходимо создание базы данных, в которой хранится информация для определенного множества конфигураций S заданных значениями векторов $\boldsymbol{L_s}$ при которых центр ВЗ совпадает с различными точками $D_s \in R_s$. В базе данных также содержится информация о множествах S_{DW} конфигураций $\boldsymbol{L_{sw}}$ для которых оси $O_{12}z_{12}$ и $O_{13}z_{13}$ являются образующими линейчатой поверхности W_d ограничивающей телесный угол сервиса. На рисунке 3 представлен синтез движения механизма руки антропоморфного робота AR-600E для случая, когда центр ВЗ совпадает с одной из точек D_s и ось $O_{12}z_{12}$ перемещается по линейчатой поверхности W_d. Синтез движений осуществляется до положения механизма, при котором узловая точка механизма O_{11} смещается в сторону следа профильной проецирующей плоскости g_3 до тех пор пака не займет противоположное положение. При этом след профильной проецирующей плоскости g_3 располагается на расстоянии l_g от боковой соседней грани ранее установленной призмы (объекта манипулирования). На рисунке 3 проекции точки O_{11} обозначены $O^2{}_{11}$ и $O^3{}_{11}$. Расстояние l_g определяется размерами перемещаемой в текущий момент времени призмы.

После осуществления указанного перемещения оси $O_{12}z_{12}$ по линейчатой поверхности W_d , и определение конечной конфигурации удовлетворяющей указанным требованиям вновь осуществляется синтез движений при котором центр ВЗ смещается по траектории движения $l\ (l_2,\ l_3)$ до целевой точки $D_ц$ ($D_ц^2$, $D_ц^3$). Алгоритм синтеза движений с преодолением тупиковых ситуаций и использования линейчатых поверхностей W_d представлен на рисунке 4.

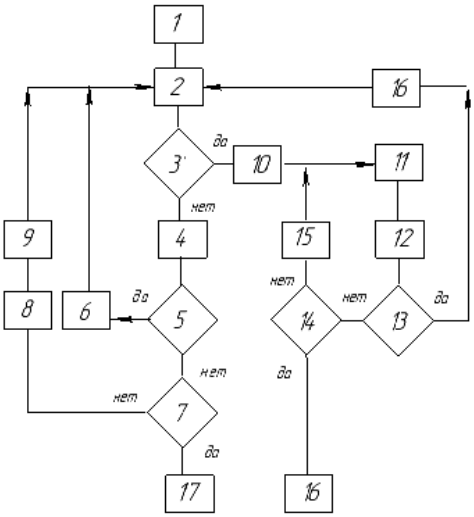


Рисунок 4 – Алгоритм вычисления промежуточных конфигураций с преодолением тупиковых ситуаций

На рисунке 4 приняты следующие обозначения: 1 – задание информации о начальном положении механизма руки, модели кинематической цепи ($q_1 \div q_n$, nm, n_{kod}, l_i , r, n, где n_{kod} – коды преобразований систем координат [5]), задание последовательности целевых точек движения центра ВЗ (D_1, D_s, $D_ц$ см. рис. 1, рис. 3, рис. 5), задание значений k_i^{max} (максимальных значений параметров определяющих максимальные размеры областей Ω^Q [5]), вычисление первоначального значения вектора $\boldsymbol{V}$; 2 – вычисление вектора $\boldsymbol{Q}$ обеспечивающего направление движения центра ВЗ к целевым точкам (D_1, D_s, $D_ц$); 3 – все значения вектора $\boldsymbol{Q}$ использованы $k_i > k_i^{max}$ (возникает тупиковая ситуация ?); 4 – реализация значений вектора $\boldsymbol{Q}$ и вычисление следующей конфигурации ($q_1 = q_1 + \Delta q_1, \ldots, q_n = q_n + \Delta q_n$, при этом принято допущение $\dot{q_1} \approx \Delta q_1$, $\dot{q_2} \approx \Delta q_2$ ..., $\dot{q_n} \approx \Delta q_n$), $k_s = 0$ (k_s – номер образующей линейчатой поверхности W_d в базе данных); 5 – расчетная конфигурация пересекает запретные зоны, или не удовлетворяет предельным значениям обобщенных координат; 6 – изменение значений параметров k_i ($k_i = k_i + 1$) задающих координаты точки N^Q в p-плоскости Γ^Q ; 7 – целевая точка $D_ц$ достигнута ; 8 – $k_i = 0$; 9 – вычисление матрицы $\boldsymbol{J}$ и вектора $\boldsymbol{V}$; 10 – возврат к синтезу движения от точки D_s (определение ближайшей конфигурации $\boldsymbol{L_s}$ и горизонтальной проекции точки D_s имеющей минимальное расстояние от горизонтальной проекции центра ВЗ (точек O_{12} или O_{13}) в которой произошла тупиковая ситуация, $k_s = 1$; 11 – Определение в соответствии с базой данных для положения руки $\boldsymbol{L_s}$ конфигурации $\boldsymbol{L_{sw}}$ задающей положение осей $O_{12}z_{12}$ и $O_{13}z_{13}$ совпадающих с положением k_s-ой образующей линейчатой поверхности W_d; 12 – определение значения вектора $\boldsymbol{Q_w}$ ($\Delta q_1 = (q_1 - q_{1\text{tsw}})/n_k$, $\Delta q_2 = (q_2 - q_{2\text{tsw}})/n_k$ и т. п.) при котором обеспечен синтез движения осей $O_{12}z_{12}$ и $O_{13}z_{13}$ (конфигурации $\boldsymbol{L_{tw}}$) до совмещения с k_s -ой образующей линейчатой поверхности W_d . Где $n_k = ((\sum_{i=0}^{n}(qi - qitsw))/n)/\Delta q^{max}$ (Δq^{max} – заданное максимальное приращение обобщенной координаты за одну итерацию при движении осей $O_{12}z_{12}$ и $O_{13}z_{13}$ по поверхности W_d. q_{itsw} – значения обобщенных координат механизма задающие конфигурацию и k_s-ую образующую поверхности W_d. Реализация значений вектора $\boldsymbol{Q_w}$ и вычисление следующей конфигурации $q_1 = q_1 + \Delta q_1, \ldots, q_n = q_n + \Delta q_n$ (значения векторов $\boldsymbol{Q_w}$ позволяют последовательно перемещать оси $O_{12}z_{12}$ и $O_{13}z_{13}$ в k_s - ое положение образующей линейчатой поверхности W_d ограничивающей телесный угол сервиса (при этом используется база данных значений векторов $\boldsymbol{L_{sw}}$); 13 – проекции точек O^2_{12} , O^2_{13} , O^3_{12} и O^3_{13} удовлетворяю заданным требованиям (положение относительно g_3); 14 – $k_s > k^{max}{}_s$ (построены все положения образующих линейчатой поверхности W_d текущей конфигурации $\boldsymbol{L_s}$) ; 15 – $k_s = k_s + 1$; 16 – $k_s = 0$; 16 – целевая точка не может быть достигнута; 17 – вывод значений обобщенных координат определяющих промежуточные конфигураций позволяющие достигнуть целевую точку.

Результаты экспериментов

На рисунке 5 на фронтальной и профильной плоскостях проекций представлен фрагмент синтеза движений руки антропоморфного робота кинематическая схема которого представлена на рисунке 2б. Компьютерное моделирование движения руки антропоморфного робота выполнено с использованием языка программирования AutoLISP а среде AutoCAD. На рисунке изображения проекций объектов манипулирования обозначены 3^2 и 3^3 . Соответственно изображения проекций точек O_1, $D_ц$ и контейнера 5 обозначены O^2_1 , O^3_1 , $D^2_ц$, $D^2_ц$, 5^2, и 5^3. При решении тестовой задачи были приняты следующие размеры геометрических объектов – размеры контейнера (соответственно длина, ширина и высота в миллиметрах) 500x325x200, размеры объектов манипулирования

210x110x85 - 4 шт. и 110x50x50 - 4 шт. Компьютерное моделирование движения руки позволило определить что при значениях l_{kx} = 0, l_{ky} = -140, l_{kz} = -200 и l_{mx} = -50, l_{my} = -380 и l_{mz} =330 рука антропоморфного робота может обеспечить установку заданных объектов манипулирования внутри контейнера.

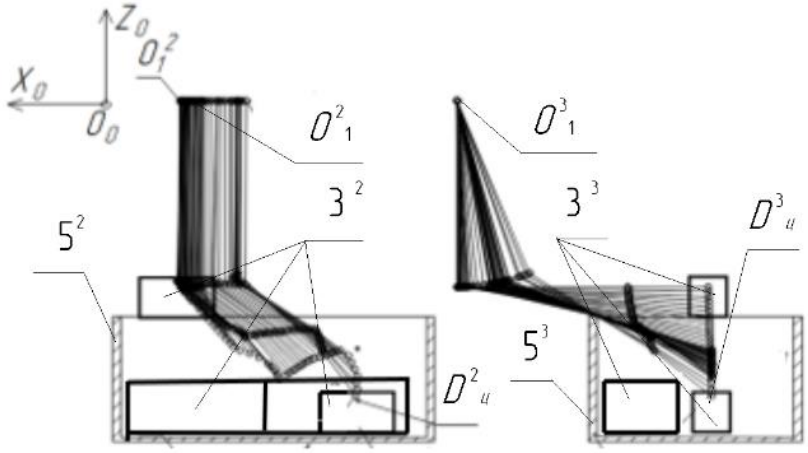


Рисунок 5 – Результат фрагмента компьютерного моделирования движения руки антропоморфного робота М4-1-2-12-3-12-1-12-3-12-1-2-12 на фронтальной и профильной плоскостях проекций при установке объекта манипулирования в контейнер

Выводы и заключение

Результаты компьютерного моделирования движения антропоморфных роботов с использованием разработанного алгоритма показали возможность преодолевать тупиковые ситуации. Создание и использование базы данных конфигураций задающих образующие линейчатых поверхностей образующих телесные углы сервиса позволяет сократить время вычислений тестовых заданий связанных с установкой объектов манипулирования в контейнер. Проведенные исследования могут быть использованы при разработке информационно-управляющих комплексов подвижных объектов, в частности при разработке интеллектуальных систем управления автономно функционирующих антропоморфных роботов в организованных средах.

Список источников

1. Афонин, В. Л., Макушкин В. А. Интеллектуальные робототехнические системы М.: Издательство «Интернет-университет информационных технологий», 2010. 208 с.

2. Whitney D. E. The Mathematics of Coordinated Control of Prosthetic Arms and Manipulators // Journal of Dynamic Systems, Measurement, and Control. 2010. № 94 (4). P. 303–309.

3. Манипуляционные системы роботов / Корендясев А. И., Саламандра Б. Л., Тывес Л. И. М: Издательство Машиностроение, 1989. 472 с.

4. Притыкин Ф. Н. Определение сферической кривой, задающей угол сервиса руки андроидного робота методом синтеза малых движений / Ф. Н. Притыкин, В. И. Небритов. // Омский научный вестник. Серия «Авиационно-ракетное и энергетическое машиностроение» 2018. т. 2, №3. С. 71–75.

5. Притыкин Ф. Н. Виртуальное моделирование движений роботов, имеющих различную структуру кинематических цепей Омск: Изд-во ОмГТУ, 2014. 172 с.

6. Ескенин, Р. Н. Геометрическое моделирование и оптимизация процессов управления адаптивным промышленным роботом: автореф. дис. … канд. .техн. наук: 05.01.01 / СибАДИ. Омск: 2010. 20 с.

7. Гордеев О. И. Исследование формы и положения областей в многомерном пространстве заданных допустимых значений вектора обобщенных скоростей плоского семизвенного манипулятора // Информационные технологии в науке и производстве: сб. трудов всерос. молод. науч.-техн. конф.. Омск: Изд-во ОмГТУ, 2015. С. 216-222.

8. Небритов В. И. Определение максимальных значений параметров задающих угол сервиса и множество конфигураций андроидного робота реализацией мгновенных состояний Информационные технологии в науке и производстве: сб. трудов III всерос. молод. науч.-техн. конф. Омск: Изд-во ОмГТУ, 2016 С. 215–220.

УДК 514.1
DOI: 10.25206/978-5-8149-3873-2-2024-741-746

Кубическая кривая Фергюсона на плоскости
Cubic Ferguson curve on a plane

В.А. Короткий, Е. А. Усманова, Л.И. Хмарова
Южно-Уральский государственный университет, пр. Ленина, 76, Челябинск, 454080, Россия

V.A. Korotkiy, E. A. Usmanova, L.I. Khmarova
South Ural State University, Lenina prospekt 76, Chelyabinsk, 454080, Russia

Аннотация. Предлагаются графоаналитические алгоритмы конструирования плоской кубической кривой, проходящей через наперед заданные начальную и конечную точки R_1, R_2 с указанными в этих точках единичными касательными векторами $\mathbf{T}_1$, $\mathbf{T}_2$. Касательные векторы и модули первых векторных производных w_1, w_2 в точках R_1, R_2 рассматриваются как элементы управления формой конструируемой кривой. Условия инцидентности и касания $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$ дополняются условием получения наперед заданной кривизны в начальной или конечной точке кривой. В статье решены две задачи: 1) построение кривой Фергюсона с фиксированной кривизной в начальной точке R_1; 2) построение кривой Фергюсона с фиксированной кривизной в конечной точке R_2. Такие задачи возникают при формировании плоской составной кривой с непрерывным изменением наклона и кривизны. Предлагаемые алгоритмы основаны на графическом представлении вектора кривизны как векторного произведения некоторых вспомогательных векторов и графическом определении ориентированных площадей параллелограммов, построенных на этих векторах. Представленные в статье примеры конструирования кривых Фергюсона транспарентны и могут быть использованы в учебном процессе.

Ключевые слова: Вектор и знак кривизны, условия инцидентности и касания, уравнения Эрмита, ориентированная площадь.

Abstract. The paper proposes graph analytic algorithms for constructing a flat parameterized cubic curve passing through preset initial and end points R_1, R_2 with unit tangent vectors $\mathbf{T}_1$, $\mathbf{T}_2$ specified at these points. The tangent vectors and modules of the first vector derivatives w_1, w_2 at points R_1, R_2 are considered as control elements of the constructed curve shape. Incidence and tangency $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$ conditions are supplemented by the condition of obtaining a preset curvature at the initial or end point of the curve. The paper solves two problems: 1) constructing the Ferguson curves with fixed curvature at the initial point R_1; 2) constructing the Ferguson curves with fixed curvature at the end point R_2. Such problems arise when forming a geometrically smooth compound curve with a continuous change in slope and curvature. The proposed algorithms are based on the graphical representation of the curvature vector as a vector product of some auxiliary vectors and the graphical determination of the signed areas of parallelograms constructed on these vectors. The presented examples of constructing Ferguson curves are transparent and can be used in the educational process.

Keywords: Curvature vector and sign, incidence and tangency conditions, Hermite equations, signed area.

Введение и постановка задачи

Кубическая параметризация в проектировании впервые введена Фергюсоном в 1963 году [1]. Кривая Фергюсона имеет вид

$$\mathbf{r}(t) = \mathbf{a}_0 + \mathbf{a}_1 t + \mathbf{a}_2 t^2 + \mathbf{a}_3 t^3, \quad t \in [0, 1]. \tag{1}$$

В докладе рассматриваются две задачи.

Задача 1. Через указанные на плоскости xy точки R_1, R_2 с заданными в этих точках единичными направляющими векторами $\mathbf{T}_1$, $\mathbf{T}_2$ требуется провести кривую (1) с наперед заданным вектором кривизны $\mathbf{K}_1$ в начальной точке R_1. Граничные условия символически записываются в виде $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2, \mathbf{K}_1\}$. Эта задача возникает при формировании составной плоской кривой, когда имеется сегмент $R_0 - R_1$ с известной кривизной $\mathbf{K}_1$ в своей конечной точке R_1, к которому надо присоединить следующий сегмент $R_1 - R_2$ с сохранением кривизны $\mathbf{K}_1$ в стыковой точке R_1.

Задача 2. Через указанные на плоскости xy точки R_1, R_2 с заданными в этих точках единичными направляющими векторами $\mathbf{T}_1$, $\mathbf{T}_2$ требуется провести кривую (1) с наперед заданным вектором кривизны $\mathbf{K}_2$ в

конечной точке R_2. Граничные условия имеют вид $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2, \mathbf{K}_2\}$. Эта задача возникает при формировании составной плоской кривой, когда имеется сегмент $R_2 – R_3$ с известной кривизной $\mathbf{K}_2$ в своей начальной точке R_2, перед которым надо вставить сегмент $R_1 – R_2$ с сохранением кривизны $\mathbf{K}_2$ в стыковой точке R_2.

Для решения задач 1 и 2 разработаны графоаналитические алгоритмы, не требующие решения систем алгебраических уравнений и пригодные к реализации в любой графически ориентированной среде. В докладе не рассматривается задача формирования кривой (1) с наперед заданной кривизной в обеих граничных точках. Эта задача решена алгебраическим способом в работе [2].

Задача 1

Ищем кривую (1), удовлетворяющую граничным условиям $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2, \mathbf{K}_1\}$. Из условий инцидентности и касания $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$ находим коэффициенты уравнения (1):

$$\begin{aligned} \mathbf{a}_0 &= \mathbf{R}_1 \\ \mathbf{a}_1 &= \omega_1 \mathbf{T}_1 \\ \mathbf{a}_2 &= 3(\mathbf{R}_2 - \mathbf{R}_1) - 2\omega_1 \mathbf{T}_1 - \omega_2 \mathbf{T}_2 \\ \mathbf{a}_3 &= \omega_1 \mathbf{T}_1 + \omega_2 \mathbf{T}_2 - 2(\mathbf{R}_2 - \mathbf{R}_1) \end{aligned} \quad . \qquad (2)$$

Условия (2) называют уравнениями Эрмита [3]. В этих уравнениях содержатся две независимые переменные – модули ω_1, ω_2 первых векторных производных функции (1) в ее граничных точках R_1, R_2. При заданных условиях инцидентности и касания коэффициенты уравнений (2) и форма кривой (1) определяется величинами ω_1, ω_2. Например, на рис. 1 показаны кривые (1) с различными значениями ω_1, ω_2, удовлетворяющие условиям $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$, но при этом имеющие различную кривизну в начальной точке R_1. Требуется найти сочетание значений ω_1, ω_2, обеспечивающее наперед заданную кривизну в этой точке. Схема решения: произвольно зафиксировав переменную ω_2, будем подбирать значение ω_1, при котором достигается требуемое значение кривизны. Задавая различные значения ω_2, всякий раз будем подбирать соответствующее значение ω_1. Получим однопараметрическое множество кривых Фергюсона, удовлетворяющих граничным условиям $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2, \mathbf{K}_1\}$.

Графоаналитический алгоритм решения задачи 1

Кривизна $\mathbf{K}_1$ в начальной точке R_1 определяется выражением

$$\mathbf{K}_1 = \frac{\dot{\mathbf{r}}(0) \times \ddot{\mathbf{r}}(0)}{|\dot{\mathbf{r}}(0)|^3} \ , \qquad (3)$$

где

$$\begin{aligned} \dot{\mathbf{r}}(0) &= \mathbf{a}_1 = \omega_1 \mathbf{T}_1, \\ \ddot{\mathbf{r}}(0) &= 2\mathbf{a}_2 = 6\mathbf{R}_1\mathbf{R}_2 - 4\omega_1 \mathbf{T}_1 - 2\omega_2 \mathbf{T}_2. \end{aligned} \qquad (4)$$

Здесь $\mathbf{R}_1\mathbf{R}_2$ – вектор, соединяющий начальную и конечную точки R_1, R_2 кривой (1). Подставляя (4) в (3) и учитывая, что $|\dot{\mathbf{r}}(0)| = \omega_1$, получаем:

$$\mathbf{K}_1 = \frac{6\mathbf{T}_1 \times \mathbf{R}_1\mathbf{R}_2 - 2\omega_2 \mathbf{T}_1 \times \mathbf{T}_2}{\omega_1^2} \ . \qquad (5)$$

Введем обозначение

$$\mathbf{m}(\omega_2) = 3\mathbf{R}_1\mathbf{R}_2 - \omega_2 \mathbf{T}_2 \ . \qquad (6)$$

При фиксированных векторах $\mathbf{R}_1\mathbf{R}_2$ и $\mathbf{T}_2$ вектор $\mathbf{m}$ зависит только от модуля ω_2 первой векторной производной в конечной точке R_2.

Подставляя (6) в (5), получаем:

$$\mathbf{K}_1 = \frac{2\mathbf{T}_1 \times \mathbf{m}(\omega_2)}{\omega_1^2} \ . \qquad (7)$$

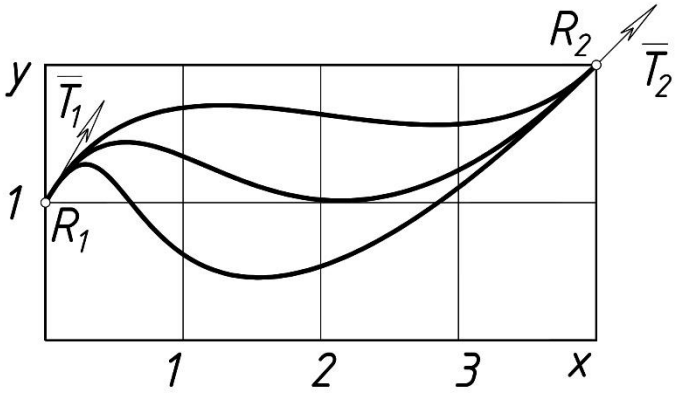


Рисунок 1 – Кривые Фергюсона, удовлетворяющие уравнениям Эрмита

Шаг 1. Найдем "критическое" значение $\omega_2 = \omega_2^*$, при котором кривизна $\mathbf{K}_1$ становится равна нулю. Векторное произведение $\mathbf{T}_1 \times \mathbf{m}$, входящее в (7), равно нулю, если векторы $\mathbf{T}_1$ и $\mathbf{m}$ коллинеарны. В соответствии с (6) начертим треугольник R_1AB, сторона R_1B которого равна утроенному вектору $\mathbf{R}_1\mathbf{R}_2$, а стороны R_1A и AB параллельны векторам $\mathbf{T}_1$, $\mathbf{T}_2$ (рис. 2,а). При таком расположении векторов выполняется равенство (6), достигается коллинеарность векторов $\mathbf{T}_1$, $\mathbf{m}$ и обеспечивается нулевое значение кривизны $\mathbf{K}_1$. Измерив длину вектора $\mathbf{AB} = \omega_2^*\mathbf{T}_2$, получим критическое значение ω_2^*.

Шаг 2. При фиксированном значении ω_2 можно, согласно (7), подобрать такую величину ω_1, при которой кривизна $\mathbf{K}_1$ достигает требуемого значения. Величина ω_2 может быть задана произвольно, но при этом необходимо учитывать правило знаков векторного произведения.

Правило знаков. Если в точке R_1 задана положительная кривизна, то, согласно (7), векторы $\mathbf{T}_1$, $\mathbf{m}$, $\mathbf{K}_1$ должны образовать правую тройку векторов. При этом вектор $\mathbf{T}_1$ поворачивается к $\mathbf{m}$ против часовой стрелки, а вектор $\mathbf{K}_1$ "смотрит вверх" перпендикулярно плоскости чертежа. На рис. 2,б показана именно такая ситуация: значение

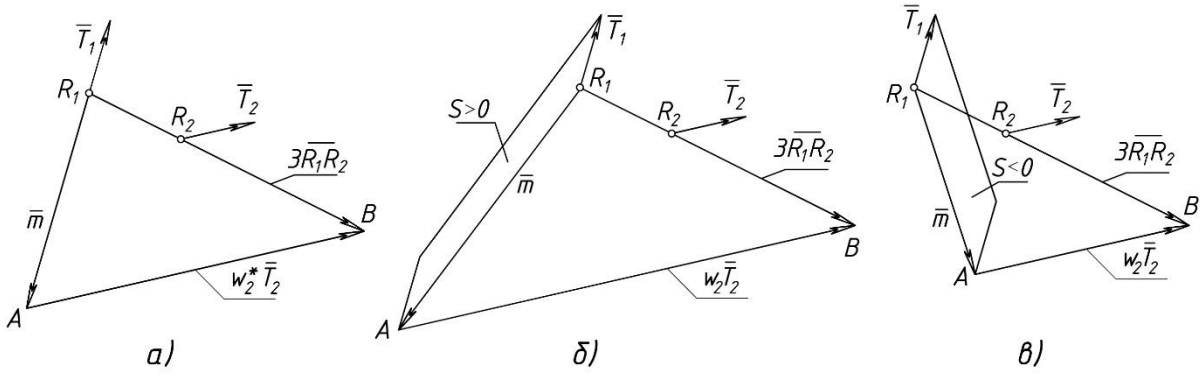


Рисунок 2 – К алгоритму расчета кривой Фергюсона с заданной кривизной в начальной точке

ω_2 выбрано хотя и произвольно, но больше критического значения ω_2^*. Получаем правую тройку векторов $\mathbf{T}_1$, $\mathbf{m}$, $\mathbf{K}_1$ и положительную кривизну K_1. Если в точке R_1 задана отрицательная кривизна, то векторы $\mathbf{T}_1$, $\mathbf{m}$, $\mathbf{K}_1$ должны образовать левую тройку векторов. Для этого надо выбрать значение ω_2 меньше, чем ω_2^* (рис. 2,в). Наглядное представление о знаке кривизны дает рис. 3.

Зафиксировав произвольное значение ω_2, на векторах $\mathbf{T}_1$ и $\mathbf{m}(\omega_2)$ вычерчиваем параллелограмм и определяем его ориентированную площадь S, которой присваиваем знак плюс или минус в соответствии с правилом знаков: если вектор $\mathbf{T}_1$ вращается к вектору $\mathbf{m}$ против часовой стрелки (векторы $\mathbf{T}_1$, $\mathbf{m}$, $\mathbf{K}_1$ образуют правую тройку), то $S>0$ (см. рис. 2,б), и наоборот (см. рис. 2,в).

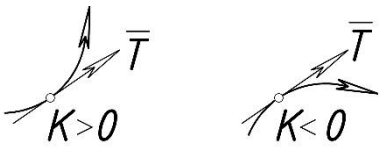


Рисунок 3 – Знак кривизны

Шаг 3. Величина S – это модуль входящего в (7) векторного произведения $\mathbf{T}_1 \times \mathbf{m}$, взятый со знаком плюс или минус. Подставляя величину S в (7), находим значение ω_1, доставляющее (при фиксированном значении ω_2) заданное значение кривизны в точке R_1:

$$\omega_1^2 = \frac{2S}{K_1} \cdot \tag{8}$$

Здесь K_1 – модуль вектора кривизны $\mathbf{K}_1$, взятый, согласно условию задачи, со знаком плюс или минус. Задача 1 решена.

Примечание. Если при выбранном значении ω_2 (например, при ω_2 бóльшем, чем ω_2^*) скалярные величины S и K_1 имеют разные знаки, то задача 1 не имеет решения. В этом случае следует задать значение ω_2 меньше, чем ω_2^*.

Пример решения задачи 1

В точках $R_1(0; 2)$ и $R_2(4; 0)$ заданы направляющие векторы $\mathbf{T}_1$, $\mathbf{T}_2$, наклоненные к оси x под углами 60° и 0° соответственно (рис. 4). Требуется построить кривые Фергюсона (1) с кривизной K_1=+0,25 и K_1= – 0,25 в начальной точке R_1, удовлетворяющие условиям инцидентности и касания $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$.

Шаг 1. Начертив, согласно рис. 2,а, треугольник R_1AB, находим критическое значение $\omega_2 = \omega_2^* = 15{,}4641$, при котором кривизна K_1 равна нулю.

Шаг 2. Чтобы обеспечить положительное значение кривизны K_1 =+0,25, задаем значение ω_2 больше его критического значения, например, ω_2=30. На векторах $\mathbf{T}_1$ и $\mathbf{m}$ строим параллелограмм и определяем его ориентированную площадь S=+12,5885 (см. рис. 2,б).

Шаг 3. По формуле (8) вычисляем ω_1=10,0353. При сочетании ω_2=30, ω_1=10,0353 достигается требуемая кривизна K_1=+0,25 в точке R_1. Подставляя найденные значения ω_1, ω_2 в (2), вычисляем декартовы координаты входящих в (1) векторов: $\mathbf{a}_0(0; 2)$, $\mathbf{a}_1(5{,}0177; 8{,}6908)$, $\mathbf{a}_2(–28{,}0353; –23{,}3817)$, $\mathbf{a}_3(27{,}0177; 12{,}6908)$. Получаем уравнение кривой Фергюсона (1) в координатной форме:

$$\begin{aligned} x(t) &= 5{,}0177t - 28{,}0353t^2 + 27{,}0177t^3, \\ y(t) &= 2 + 8{,}6908t - 23{,}3817t^2 + 12{,}6908t^3, \; t \in [0,\ 1]. \end{aligned} \tag{9}$$

Кривая (9) показана на рис. 4, поз. 1.

Проверка. Вычисляем скалярное значение кривизны параметрически заданной кривой (9) по формуле [4]:

$$K = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{(\dot{x}^2 + \dot{y}^2)^{3/2}}. \tag{10}$$

Дифференцируя (9) по параметру t, вычисляем первые и вторые производные при t=0 (в начальной точке R_1). Подставляя рассчитанные значения производных в (10), получаем K_1=+0,25, что соответствует условию задачи.

Задавая различные значения ω_2 (больше критического) и повторяя шаги 1…3, получаем однопараметрическое семейство кривых Фергюсона (1) с положительной кривизной K_1=+0,25 в начальной точке.

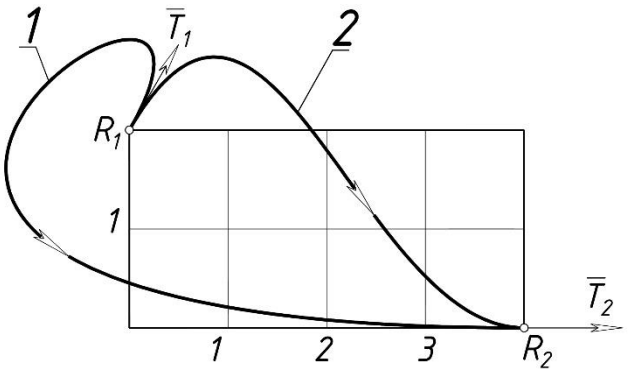


Рисунок 4 – Кривые Фергюсона с фиксированной кривизной в начальной точке

Чтобы найти кривую (1) с отрицательной кривизной K_1= – 0,25 в начальной точке, задаем ω_2 меньше его критического значения, например, ω_2=5. На векторах $\mathbf{T}_1$ и $\mathbf{m}$=R_1A строим параллелограмм и определяем его ориентированную площадь S= – 9,0622 (см. рис. 2,в). По формуле (8) находим ω_1=8,5145. При найденном сочетании ω_1, ω_2 достигается требуемая кривизна K_1= – 0,25 в точке R_1 (рис. 4, поз. 2). Множеству значений ω_2 (меньше критического) отвечает однопараметрическое множество кривых Фергюсона (1) с отрицательной кривизной K_1= – 0,25 в точке R_1.

Задача 2

Ищем кривую (1) с наперед заданной кривизной $\mathbf{K}_2$ в конечной точке R_2. Кривая должна удовлетворять граничным условиям $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2, \mathbf{K}_2\}$. Коэффициенты уравнения (1), удовлетворяющие условиям инцидентности и касания $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$, определяются согласно уравнениям Эрмита (2). Эти уравнения содержат две независимые переменные – модули ω_1, ω_2 первых векторных производных функции (1) в ее граничных точках R_1, R_2. Требуется подобрать сочетание значений ω_1, ω_2, обеспечивающее наперед заданную кривизну $\mathbf{K}_2$ в конечной точке R_2. Схема решения: произвольно зафиксировав переменную ω_1, будем подбирать значение ω_2, при котором достигается требуемое значение кривизны. Задавая различные значения ω_1, всякий раз будем находить соответствующее значение ω_2. Получим однопараметрическое множество кривых Фергюсона, удовлетворяющих граничным условиям $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2, \mathbf{K}_2\}$.

Графоаналитический алгоритм решения задачи 2

Кривизна $\mathbf{K}_2$ в конечной точке R_2 (при t=1) определяется выражением

$$\mathbf{K}_2 = \frac{\dot{\mathbf{r}}(1) \times \ddot{\mathbf{r}}(1)}{|\dot{\mathbf{r}}(1)|^3}, \tag{11}$$

где

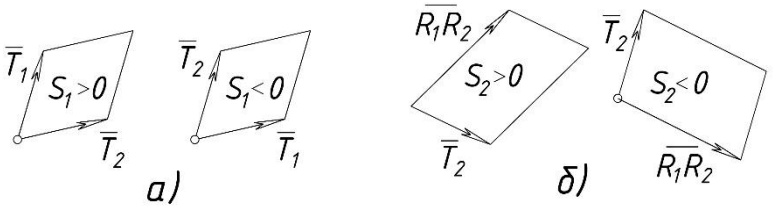


Рисунок 5 – К алгоритму расчета кривой Фергюсона с фиксированной кривизной в конечной точке

$$\begin{aligned}\dot{\mathbf{r}}(1) &= \mathbf{a}_1 + 2\mathbf{a}_2 + 3\mathbf{a}_3 = \omega_2\mathbf{T}_2,\\ \ddot{\mathbf{r}}(1) &= 2\mathbf{a}_2 + 6\mathbf{a}_3 = 2\omega_1\mathbf{T}_1 + 4\omega_2\mathbf{T}_2 - 6\mathbf{R}_1\mathbf{R}_2.\end{aligned} \tag{12}$$

Подставляя (12) в (11), получаем:

$$\mathbf{K}_2 = \frac{2\omega_1\mathbf{T}_2\times\mathbf{T}_1 - 6\mathbf{T}_2\times\mathbf{R}_1\mathbf{R}_2}{\omega_2^2}\ . \tag{13}$$

Шаг 1. Находим критическое значение $\omega_1 = \omega_1^*$, при котором кривизна $\mathbf{K}_2$ становится равна нулю. Приравнивая нулю числитель выражения (13), получаем:

$$\omega_1^*\mathbf{T}_2\times\mathbf{T}_1 = 3\mathbf{T}_2\times\mathbf{R}_1\mathbf{R}_2. \tag{14}$$

Очевидно, условие (14) выполняется только в том случае, если векторные произведения в левой и правой частях (14) имеют одинаковый знак. Вычерчиваем векторы $\mathbf{T}_2$, $\mathbf{T}_1$ и определяем ориентированную площадь S_1 параллелограмма, построенного на этих векторах. Знак площади S_1 проиллюстрирован на рис. 5,а.

Вычерчивая векторы $\mathbf{T}_2$ и $\mathbf{R}_1\mathbf{R}_2$, определяем ориентированную площадь S_2 параллелограмма, построенного на этих векторах. Знак площади S_2 проиллюстрирован на рис. 5,б. Согласно (14), вычисляем критическое значение $\omega_1 = \omega_1^*$:

$$\omega_1^* = \frac{3S_2}{S_1}. \tag{15}$$

Если величины S_1 и S_2 имеют разные знаки, то при заданных граничных условиях $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$ нулевая кривизна в конечной точке кривой (1) не может быть достигнута ни при каком сочетании параметров ω_1, ω_2.

Шаг 2. Находим значение ω_2, доставляющее (при фиксированном значении ω_1) требуемое значение кривизны в точке R_2. Для этого заменяем в (13) векторные произведения ориентированными площадями S_1, S_2. Получаем скалярное выражение

$$K_2 = \frac{2\omega_1 S_1 - 6S_2}{\omega_2^2}\ . \tag{16}$$

Зафиксировав произвольное значение ω_1 (большее или меньшее его критического значения) и задав требуемую кривизну K_2, из (16) находим соответствующее значение ω_2:

$$\omega_2^2 = \frac{2\omega_1 S_1 - 6S_2}{K_2} \tag{17}$$

Найденное сочетание ω_1, ω_2 обеспечивает заданную кривизну K_2 (положительную или отрицательную) в конечной точке конструируемой кривой (1).

Примечание. Если при выбранном значении ω_1 (например, при ω_1 большем, чем ω_1^*) числитель и знаменатель выражения (17) имеют разные знаки, то при таком значении ω_1 задача не имеет решения. В этом случае следует задать значение ω_1 меньше, чем ω_1^*.

Пример решения задачи 2

В точках $R_1(0; 2)$ и $R_2(4; 0)$ заданы направляющие векторы $\mathbf{T}_1$, $\mathbf{T}_2$. Вектор $\mathbf{T}_1$ наклонен к оси x под углом 45°. Вектор $\mathbf{T}_2$ направлен вертикально вверх (рис. 6). Требуется построить кривые Фергюсона (1) с кривизной K_2=+0,25 и K_2= – 0,25 в конечной точке R_2, удовлетворяющие условиям инцидентности и касания $\{R_1, R_2, \mathbf{T}_1, \mathbf{T}_2\}$.

Шаг 1. Находим критическое значение $\omega_1 = \omega_1^*$, при котором кривизна K_2 становится равна нулю. Для этого вычерчиваем векторы $\mathbf{T}_2$, $\mathbf{T}_1$ и определяем ориентированную площадь S_1= – 0,7071 параллелограмма,

построенного на этих векторах (см. рис. 5,а). Вычерчивая векторы $\mathbf{T}_2$ и $\mathbf{R}_1\mathbf{R}_2$, определяем ориентированную площадь S_2= – 4,0 параллелограмма, построенного на этих векторах (см. рис. 5,б). Находим, согласно (15), критическое значение ω_1^*, при котором кривизна K_2 становится равна нулю:

$$\omega_1^* = \frac{3S_2}{S_1} = 16{,}9705 \cdot$$

Шаг 2. Чтобы в конечной точке конструируемой кривой обеспечить положительную кривизну K_2=+0,25, задаем значение ω_1 меньше его критического значения, например, ω_1=4. Из (17) находим ω_2=8,5658. При сочетании ω_1=4, ω_2=8,5658 достигается требуемая кривизна K_2=+0,25 в конечной точке кривой (1).

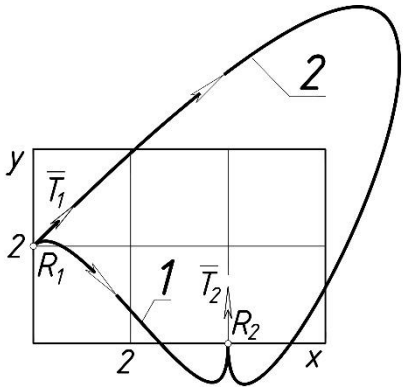


Рисунок 6 – Кривые Фергюсона с фиксированной кривизной в конечной точке

Подставляя найденные значения ω_1, ω_2 в (2), вычисляем декартовы координаты входящих в (1) векторов: $\mathbf{a}_0$(0; 2), $\mathbf{a}_1$(2,8284; 2,8284), $\mathbf{a}_2$(6,3431; –20,2226), $\mathbf{a}_3$(-5,1716; 15,3942). Получаем уравнение кривой Фергюсона (1) в координатной форме:

$$\begin{aligned} x(t) &= 2{,}8284t + 6{,}3431t^2 - 5{,}1716t^3, \\ y(t) &= 2 + 2{,}8284t - 20{,}2226t^2 + 15{,}3942t^3, \; t \in [0,\ 1]. \end{aligned} \tag{18}$$

Кривая (18) показана на рис. 6, поз. 1.

Задавая различные значения ω_1 (меньше критического значения $\omega_1^* = 16{,}9705$) и повторяя шаги алгоритма, получаем однопараметрическое семейство кривых Фергюсона (1) с положительной кривизной K_2=+0,25 в конечной точке R_2.

Чтобы найти кривую (1) с отрицательной кривизной K_2= – 0,25 в точке R_2, задаем ω_1 больше его критического значения, например, ω_1=60. Подставляя в (17) значения S_1= – 0,7071, S_2= – 4, ω_1=60 и K_2= – 0,25, получаем ω_2=15,6015. При найденном сочетании значений ω_1, ω_2 достигается требуемая кривизна K_2= – 0,25 в точке R_2 (рис. 6, поз. 2). Множеству значений ω_1 (больше критического) отвечает однопараметрическое множество кривых Фергюсона (1) с отрицательной кривизной K_2= – 0,25 в точке R_2.

Выводы

Предложены графоаналитические алгоритмы конструирования плоской параметризованной кубической кривой, проходящей через две наперед заданные точки с указанными в этих точках направляющими векторами и с наперед заданной кривизной в одной из заданных точек. Предлагаемые алгоритмы основаны на графическом представлении вектора кривизны как векторного произведения некоторых вспомогательных векторов и графическом определении ориентированных площадей параллелограммов, построенных на этих векторах.

В статье решены две задачи: 1) построение семейства кривых Фергюсона с фиксированной кривизной в начальной точке; 2) построение семейства кривых Фергюсона с фиксированной кривизной в конечной точке. Представленные в статье примеры конструирования кривых Фергюсона транспарентны и могут быть использованы в учебном процессе.

Список источников

[1] Фокс А., Пратт М. Вычислительная геометрия. Применение в проектировании и на производстве. Пер. с англ. М.: Мир, 1982. 304 с.

[2] Короткий В. А. Конструктивные алгоритмы формирования составных кубических кривых Безье в пространстве и на плоскости // Омский научный вестник, № 2 (182), 2022, С. 10–16.

[3] Шикин Е. В., Плис А. И. Кривые и поверхности на экране компьютера. Руководство по сплайнам для пользователей. М.: Диалог-МИФИ, 1996. 240 с.

[4] Бронштейн И.Н., Семендяев К.А. Справочник по математике. Санкт-Петербург: Лань, 2022. 608 с.

УДК 378.147:514.18
DOI: 10.25206/978-5-8149-3873-2-2024-747-754

Единый конструктивный алгоритм построения окружностей Мальфатти
Unified Constructive Algorithm for Constructing Malfatti Circles

Е.П.Бояшова, Д.В.Волошинов, Т.В.Мусаева
СПбГУТ им. проф. М.А. Бонч-Бруевича, 193232, Санкт-Петербург, Россия

E.P.Boyashova, D.V.Voloshinov, T.V.Musaeva
SPbSUT im. prof. M.A. Bonch-Bruevich, 193232, St. Petersburg, Russia

Аннотация. Статья посвящена рассмотрению нового конструктивного подхода к решению задачи о построении трех взаимно касающихся друг с другом окружностям, вписанным в треугольник, и известной под названием «задача об окружностях Мальфатти». Существующие в настоящее время методы построения таких окружностей основаны на выполнении элементарных геометрических построений, впервые предложенных Якобом Штейнером. На данную задачу обращали внимание многие исследователи, однако широкую известность получили только ее аналитические интерпретации. Данная работа направлена на поиск единых геометрических закономерностей, присущих этой задаче. Полученные результаты позволяют констатировать, что задача об окружностях Мальфатти основывается на многократном применении алгоритмов построения линий, равноудаленных к парам окружностей. Полученный алгоритм расширяет арсенал функций системы конструктивного геометрического моделирования Симплекс.

***Ключевые слова*:** окружности Мальфатти, кривые второго порядка, алгоритмы построения коник, конструктивная геометрия, Симплекс

Аннотация. The article is devoted to the consideration of a new constructive advice for solving problems of constructing three mutual neighbors with other neighbors inscribed in a triangle, and a representation called the “Malfatti environment problem”. Currently existing methods for constructing such circles are based on elementary geometric constructions first proposed by Jacob Steiner. Many researchers have paid attention to this problem, but only its analytical interpretations have become widely known. This work is aimed at searching for uniform geometric patterns inherent in this problem. The results obtained allow us to state that the problem of Malfatti circles is based on the repeated application of algorithms for constructing lines equidistant to pairs of circles. The resulting algorithm expands the arsenal of functions of the Simplex constructive geometric modeling system.

***Ключевые слова*:** Malfatti circles, second order curves, algorithms for constructing conics, constructive geometry, Simplex

I. Введение

Задача о построении окружностей Мальфатти имеет относительно простую формулировку, вследствие чего многие исследователи полагают, что она может быть отнесена к разряду базовых геометрических построений. Ее сущность заключается во вписывании в треугольник трех взаимно касающихся окружностей таким образом, чтобы каждая из них касалась пары сторон треугольника. Иными словами, такие окружности должны создавать «плотную упаковку» окружностей в треугольнике.

Задача была впервые Франческо Мальфатти в 1803 году [1], но им самим решение получено не было. Геометрически задача была решена Якобом Штейнером в 1826 году с помощью простых геометрических инструментов [2], но доказательство представлено не было. Доказательство корректности решения Якоба Штейнера было выполнено Петерсоном [3]. Решение можно найти в [4, 5] и ряде других работ, посвященной тематике данной задачи. Большой вклад в решение задачи внесен в работе Беленького В.З. и Заславского А.А. [6], предложивших оригинальный подход к ее интерпретации.

Обращает, однако, на себя внимание тот факт, что работы, посвященные решению задачи Мальфатти, так или иначе сводятся к выполнению тех или иных вычислительных действий, инициированных ее геометрической природой. В то же время исследований, посвященных исключительно конструктивному подходу, обосновывающего общую геометрическую природу этой задачи, не обнаруживается. Целью данной статьи является поиск единого конструктивного обоснования решения данной задачи.

II. Алгоритм построения коник, равноудаленных от двух кривых, по фокусам и точке

Пусть изначально заданы две точки F_1 и F_2 – фокусы пары сопряженных кривых второго порядка и произвольная точка плоскости P, через которую эти кривые должны проходить (рис. 1). Для построения этой пары выполним следующие действия:

Построим отрезок прямой линии $z_1 = F_1 \circ F_2$. Найдем точку O – центр отрезка z_1 и проведем прямую линию $z_2 : z_2 \perp z_1, z_2 \sim O$ – суть оси строящихся кривых. Проведем окружность $d = F_1 \circ F_2 \circ P$. Построим точки $K, L = d \times z_2$ и соединим их с точкой P прямыми линиями $k = K \circ P$ и $l = L \circ P$. Найдем точки $A = S_{z_2}(P)$, $B = S_{z_1}(A)$, $C = S_{z_1}(P)$ где символами S_{z_1} и S_{z_2} обозначено преобразование осевой симметрии относительно прямых z_1 и z_2 соответственно. Полученные объекты позволяют провести две коники: $v = C \circ l \circ P \circ B \circ A; B \sim A$ и $u = C \circ k \circ P \circ B \circ A; B \sim A$, реперами которых и являются две фокальные точки F_1 и F_2 и произвольная точка плоскости P. Обратим внимание на то, что касательные к коникам в точках их пересечения перпендикулярны, вследствие чего полученные коники будем называть ортогональными. Произвольность положения точки P позволяет говорить о задании двух пучков ортогональных коник, точки F_1 и F_2 для которых являются их центрами.

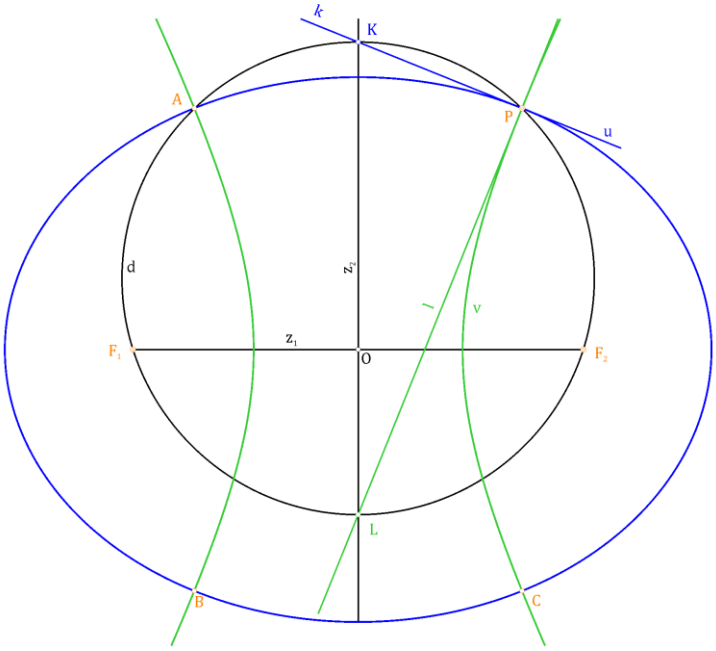


Рис. 1. Конструктивная схема построения двух комплементарных ортогональных коник

В работе [7] была рассмотрена конструктивная геометрическая схема для построения линии, равноудаленной по отношению к двум произвольным окружностям. Особенностью данной схемы является то, что в этих алгоритмах построение равноудаленных кривых (коник) проведено с использованием заранее известных их фокальных точек.

Пусть имеются две окружности p и q (рис. 2). Воспользовавшись алгоритмом, приведенным выше, построим две коники u и v, равноудаленные от этих окружностей. Задача в общем случае имеет четыре решения.

Пусть $A \sim u$. Построим $ap = A \circ P$ и $aq = A \circ Q$, где точки P и Q – центры окружностей p и q соответственно. Найдем точки $AP_1, AP_2 = ap \times p$, $AQ_1, AQ_2 = aq \times q$.

Проведем окружность $a : A = C(a), a \sim AP_1$. Очевидно, что $a \sim AQ_1$, вследствие чего точка $A \sim u$ равноудалена от окружностей p и q.

Пусть $B \sim u$. Построим $bp = B \circ P$ и $bq = B \circ Q$. Найдем точки $BP_1, BP_2 = bp \times p$, $BQ_1, BQ_2 = bq \times q$. Проведем окружность $b : B = C(b), b \sim BP_2$. Очевидно, что $b \sim BQ_2$, вследствие чего точка $B \sim u$ равноудалена от окружностей p и q.

Пусть $C \sim \nu$. Построим $cp = C \circ P$ и $cq = C \circ Q$. Найдем точки $CP_1, CP_2 = cp \times p$, $CQ_1, CQ_2 = cq \times q$. Проведем окружность $c : C = \mathrm{C}(c), c \sim CP_1$. Очевидно, что $c \sim CQ_2$, вследствие чего точка $C \sim \nu$ равноудалена от окружностей p и q.

Пусть $D \sim \nu$. Построим $dp = D \circ P$ и $dq = D \circ Q$. Найдем точки $DP_1, DP_2 = dp \times p$, $DQ_1, DQ_2 = dq \times q$. Проведем окружность $d : D = \mathrm{C}(d), d \sim DP_2$. Очевидно, что $d \sim DQ_1$, вследствие чего точка $D \sim \nu$ равноудалена от окружностей p и q.

Приведенные построения демонстрируют наличие четырех различных пар рядов окружностей с едиными центрами P и Q и парой общих окружностей p и q.

$$pa : P = \mathrm{C}(pa), pa \sim A,\ qa : Q = \mathrm{C}(qa), qa \sim A;$$

$$pb : P = \mathrm{C}(pb), pb \sim B,\ qb : Q = \mathrm{C}(qb), qb \sim B;$$

$$pc : P = \mathrm{C}(pc), pc \sim C,\ qc : Q = \mathrm{C}(qc), qc \sim C;$$

$$pd : P = \mathrm{C}(pd), pd \sim D,\ qd : Q = \mathrm{C}(qd), qd \sim D.$$

Таким образом, имеем четыре решения поставленной задачи.

Принцип выполнения построений равноудаленных кривых в случае позиционного отношения исходных окружностей, при котором одна окружность полностью лежит в круге, ограниченном другой окружностью, тот же, за исключением того, что равноудаленные кривые будут представлены эллипсами, поэтому этот случай здесь не рассматривается.

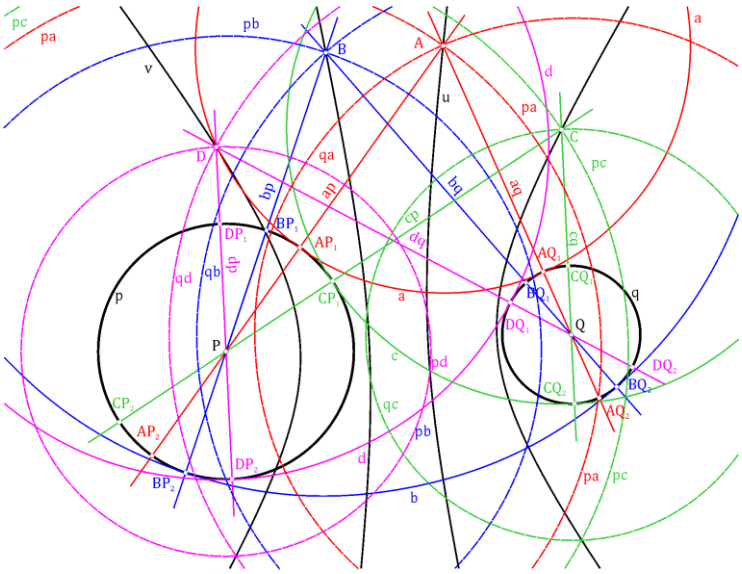


Рис. 2. Пояснения к построению коник, равноудаленных от двух окружностей на примере гиперболических кривых

В качестве примеров в работе [7] продемонстрированы два случая, в которых исходные окружности располагаются вне друг друга или частично пересекаются явно (т.е. в вещественных точках). Примечательно то, что в этих двух случаях линиями, эквидистантными к двум окружностям являются гиперболы. В том случае, если одна из окружностей полностью находится в другой, то равноудаленными линиями для них являются эллипсы, причем общая схема построения эквидистант не меняется. Непринципиальным отличием с точки зрения общетеоретического подходя является лишь то, что часть геометрических образов становятся мнимыми, что может представлять проблему лишь для инструментального исполнения алгоритма традиционными геометрическими средствами. Принимая во внимание то обстоятельство, что все эти ограничения легко преодолеваются компьютерными средствами поддержки геометрического моделирования и скрытыми от их пользователей, будем считать способ построения эквидистантных кривых к двум произвольным окружностям единым.

III. Единая конструктивная схема построения окружностей Мальфатти

Для обоснования принципов, предлагаемых в конструктивной схеме построения окружностей Мальфатти, рассмотрим классическое решение этой задачи, сформулированное Якобом Штейнером.

Пусть заданы три вершины A, B, C треугольника $\triangle ABC$. Проведем через них три прямые линии – стороны треугольника $ab = A \circ B$, $bc = B \circ C$ и $ac = A \circ C$ (рис. 3). Центры искомых окружностей должны касаться каждой их пар сторон треугольника, следовательно, они лежат на линиях, равноудаленных от этих сторон, т.е. на биссектрисах углов, образованных этими сторонами. Построим биссектрисы углов при вершинах A – $a_1 : \angle a_1, ab = \angle a_1, ac$, $a_2 : \angle a_2, ab = \angle a_2, ac$; B – $b_1 : \angle b_1, ab = \angle b_1, bc$, $b_2 : \angle b_2, ab = \angle b_2, bc$ и C – $c_1 : \angle c_1, ac = \angle c_1, bc$, $c_2 : \angle c_2, ac = \angle c_2, bc$. Найдем точки пересечения биссектрис $S = a_1 \times b_1 \times c_1$, $AB = a_2 \times b_2$, $BC = b_2 \times c_2$ и $AC = a_2 \times c_2$.

В соответствии с методом, предложенным Штейнером, выполним построение трех вспомогательных окружностей: k (вписана в треугольник, образованный линиями ab, a_1, b_1); l (вписана в треугольник, образованный линиями bc, b_1, c_1) и m (вписана в треугольник, образованный линиями ac, a_1, c_1). Каждая из биссектрис касается пары начерченных окружностей. Достроим также комплементарные касательные для каждой такой пары: km, ml и kl и найдем искомые окружности $u = \triangle_{ins}(ab, ac, ml)$, $v = \triangle_{ins}(ab, bc, km)$, $w = \triangle_{ins}(bc, ac, kl)$. Обратим внимание на то, что прямые линии km, ml и kl – суть радикальные оси соответственных пар искомых окружностей (k,m), (m,l) и (k,l).

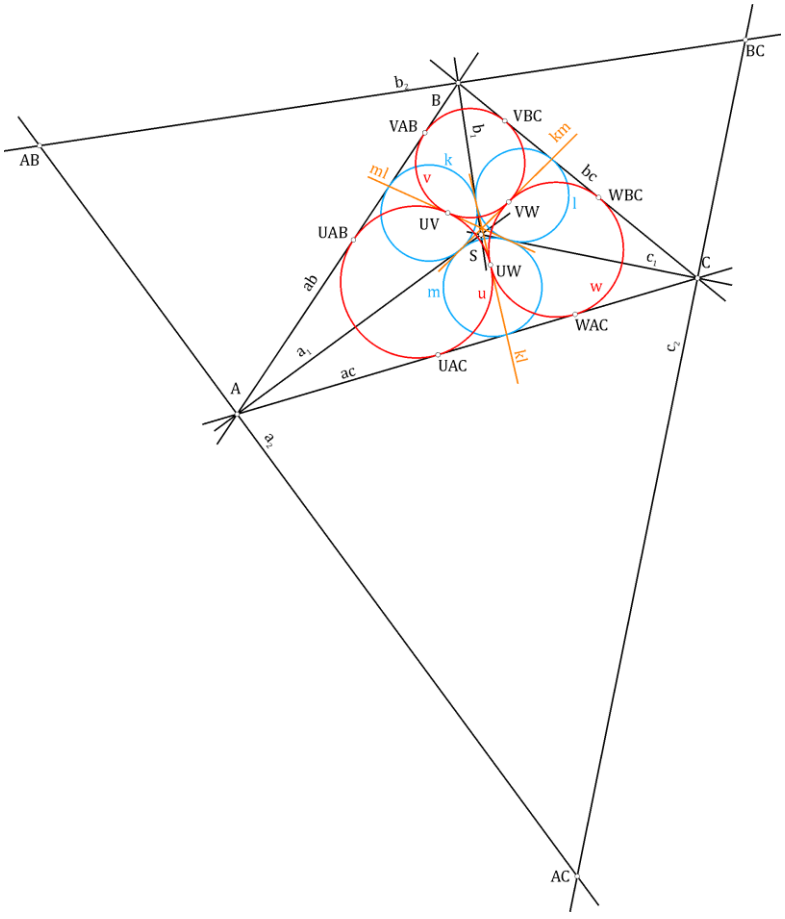


Рис. 3. Классическая схема построения окружностей Мальфатти

Несмотря на корректность полученного решения, оно не дает ясного представления о геометрической природе полученных окружностей Мальфатти. Для более подробного анализа осуществим ряд дополнительных построений, которые позволят изучить эту природу в достаточной полноте.

Проведем окружности $p = UAB \circ VAB$, $q = VBC \circ WBC$, $r = UAC \circ WAC$ через точки касания UAB, VAB, VBC, WBC, WAC и UAC окружностей Мальфатти со сторонами треугольника $\triangle ABC$ (рис. 4). Нетрудно увидеть, что $p : p \perp u; p \perp v; p \perp ab$, $q : q \perp v; q \perp w; q \perp bc$ и $r : r \perp u; r \perp w; r \perp ac$. Построим также окружности $a \sim UAB$ с центром в точке A, $b \sim VAB$ с центром в точке B и $c \sim WBC$ с центром в точке C. Из построения следует, что центр $Cp = \mathrm{C}(p)$ окружности p равноудален от окружностей a и b, центр $Cq = \mathrm{C}(q)$ окружности q равноудален от окружностей b и c, а центр $Cr = \mathrm{C}(r)$ окружности r равноудален от окружностей c и a. В то же время точка пересечения биссектрис S треугольника $\triangle ABC$ также равноудалена от окружностей a, b и c. В соответствии с вышеизложенным линия – коника e, равноудаленная от окружностей a и b, может быть построена, и она, в свою очередь, будет проходить через точку S. Соответственно, можно определить и точку Cp центра окружности p как одну из двух точек пересечения коники с прямой ab. Аналогичные рассуждения справедливы и в отношении коник f и g, а также точек Cq и Cr.

Найдем теперь положение точки R – радикального центра искомых окружностей Мальфатти u, v и w. Вследствие того, что эти окружности касаются друг друга, расстояние от точки R до всех трех точек касания одинаковое, следовательно, через точки касания можно провести окружность z, ортогональную к окружностям u, v и w. С другой стороны, окружности p, q и r касаются окружности z в тех же соответственных точках, вследствие чего точка R является равноудаленной по отношению к этим окружностям, а окружность z может рассматриваться как вписанная в окружности p, q и r.

Рассмотрим, к примеру, пару окружностей p и r. Построим их равноудаленную h и заметим, что она проходит через точку A. Действительно, точка центра окружности p инцидентна со стороной треугольника $ab \sim Cp$; точка центра окружности r инцидентна со стороной треугольника $ac \sim Cr$, а расстояния от точки A до точек UAB и UAC равны. Те же рассуждения справедливы и в отношении коник i и j. Таким образом, $R, R' = h \times i \times j$, где R – центр окружности z, вписанной в окружности p, q, r, а R' – центр окружности z', описанной возле тех же окружностей.

Теперь не составляет труда построить радикальные оси $km = R \circ Cp$, $ml = R \circ Cq$ и $kl = R \circ Cr$, а уже с их помощью найти окружности $u = \triangle_{ins}(ab, ac, km)$, $v = \triangle_{ins}(ab, bc, ml)$, $w = \triangle_{ins}(bc, ac, kl)$. Заметим, что задача о вписывании окружности в треугольник также опирается на построение равноудаленных линий – биссектрис пар сторон треугольников, исходящих из общих для них вершин.

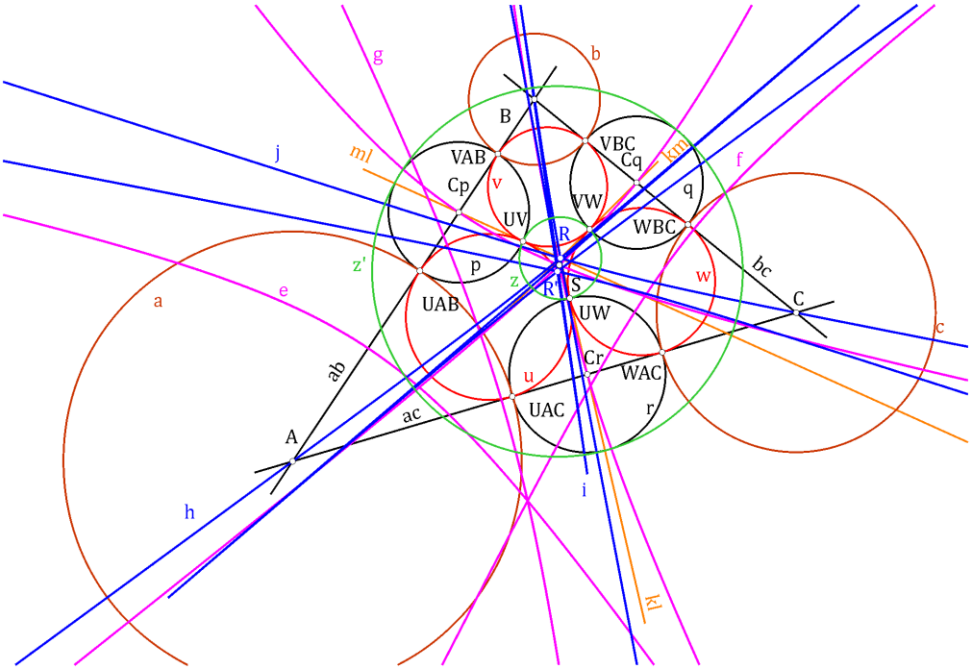


Рис. 4. Схема для анализа конструктивных свойств задачи об окружностях Мальфатти

Выполнив анализ приведенной схемы, приходим к выводу, что построение окружностей Мальфатти можно выполнить по следующему общему алгоритму (рис. 5).

1. Имея заданными три исходные вершины A, B, C, строим треугольник $\triangle ABC$ со сторонами $ab = A \circ B$, $bc = B \circ C$ и $ac = A \circ C$.
2. Строим биссектрисы углов при вершинах A – $a_1 : \angle a_1, ab = \angle a_1, ac$, $a_2 : \angle a_2, ab = \angle a_2, ac$; B – $b_1 : \angle b_1, ab = \angle b_1, bc$, $b_2 : \angle b_2, ab = \angle b_2, bc$ и C – $c_1 : \angle c_1, ac = \angle c_1, bc$, $c_2 : \angle c_2, ac = \angle c_2, bc$. Найдем точки пересечения биссектрис $S = a_1 \times b_1 \times c_1$, $AB = a_2 \times b_2$, $BC = b_2 \times c_2$ и $AC = a_2 \times c_2$. Точки AB, BC и AC по аналогии с точкой S позволяют получить кратные решения задачи об окружностях Мальфатти. Поскольку принципиально алгоритм построения на основе этих точек не отличается от алгоритма с использованием точки S, будем рассматривать только его.
3. Изначально окружности a и b неизвестны, поэтому построение равноудаленной к ним линии ($e : a, b$) будем выполнять посредством точек A и B, представляющих фокусы этой кривой, и точки $S \sim e$. Заметим, что одновременно с линией e будет получена комплементарная с ней коника e', которая является равноудаленной для ряда пар окружностей a' и b' с теми же центрами A и B. Аналогичные рассуждения следует провести и в отношении линий g, g', h и h'. Рассмотрение результатов взаимодействия системы линий e, e', g, g', h и h' позволяет получить все кратные решения задачи в отношении выбранного центра S.
4. Из чертежа видно, что на основе построения линий e, e', g, g', h и h' можно выделить четыре точки трехкратного пересечения коник $T = e \times f \times g$, $U = e' \times g' \times f$, $V = e' \times f' \times g$, $W = f' \times g' \times e$. Выбор любой из них позволяет определить пары точек пересечения коники со стороной треугольника $\triangle ABC$, проходящей через фокусы коники и, следовательно, получить одно из кратных решений задачи.
5. Пусть для примера выбрана точка $T = e \times f \times g$. Следовательно, для дальнейшего построения потребуются точки $Cp, Cp' = ab \times e$, $Cq, Cq' = bc \times f$ и $Cr, Cr' = ac \times g$.
6. Для дальнейших построений требуется выбрать две тройки точек для построения аналогов центров ранее рассмотренных окружностей p, q и r. Для осуществления такого выбора следует через две тройки провести две коники, вписанные в треугольник $\triangle ABC$. На чертеже в качестве примера показаны две коники $y = ac \circ bc \circ Cp$; $ac \sim Cq$; $bc \sim Cr$ и $y' = ac \circ bc \circ Cp'$; $ac \sim Cq'$; $bc \sim Cr'$. Существует

только две коники из восьми возможных, которые будут удовлетворять требованию касания к сторонам треугольника $\triangle ABC$, следовательно, выбор одной точки T повлечет за собой получение двух решений. Соответственно, на основе одного центра S может быть получено 8 решений, а учитывая также центры AB, BC и AC все 32 решения задачи.

7. Поскольку изначально окружности вида p и r неизвестны, но доступны точки Cp и Cr, а также точка A, то построить конику h можно по двум фокусам – центрам Cp и Cr недоступных окружностей p и r и точке A – $h:Cp,Cr,A$. Выполним также построение коник $h':Cp,Cr,A$; $i:Cp,Cq,B$; $i':Cp,Cq,B$; $i':Cp,Cq,B$; $i':Cp,Cq,B$ по аналогичной схеме.
8. Найдем радикальный центр искомых окружностей $R=h\times i\times j$, а также радикальные оси $ml=Cp\circ R$, $km=Cq\circ R$, $kl=Cr\circ R$.

Построения искомых окружностей во вписывании окружностей в прямые $u=\triangle_{ins}(ab,ac,ml)$, $v=\triangle_{ins}(ml,ab,bc)$ и $w=\triangle_{ins}(bc,ac,kl)$.

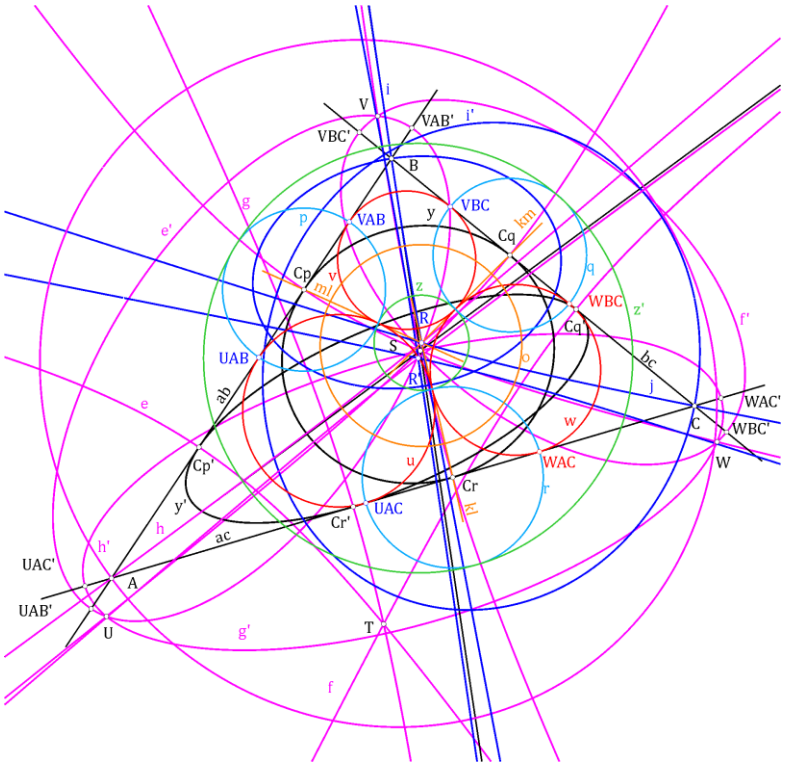


Рис. 5. Единая конструктивная схема построения окружностей Мальфатти

Построим окружности $z:z\perp u;z\perp v;z\perp w$; $o:o\perp q;o\perp p;o\perp r$; $z'=\mathrm{I}_o(z)$. Как видно из построения, окружность z с центром в R является одним из решений задачи Аполлония относительно p, q и r. Следовательно, комплементарное решение z' может быть получено как результат инверсии z относительно окружности 0, ортогональной к p, q и r, при этом центром окружности z' является точка R'. Из данного построения видно, что в общей конструктивной схеме решения задачи Мальфатти естественным образом присутствует ее формулировка и в инверсионной постановке. То есть решение задачи о построении трех взаимно касающихся друг друга окружностей и сопряженных с тремя другими окружностями, имеющими общую точку, сводится к решению классической задачи Мальфатти.

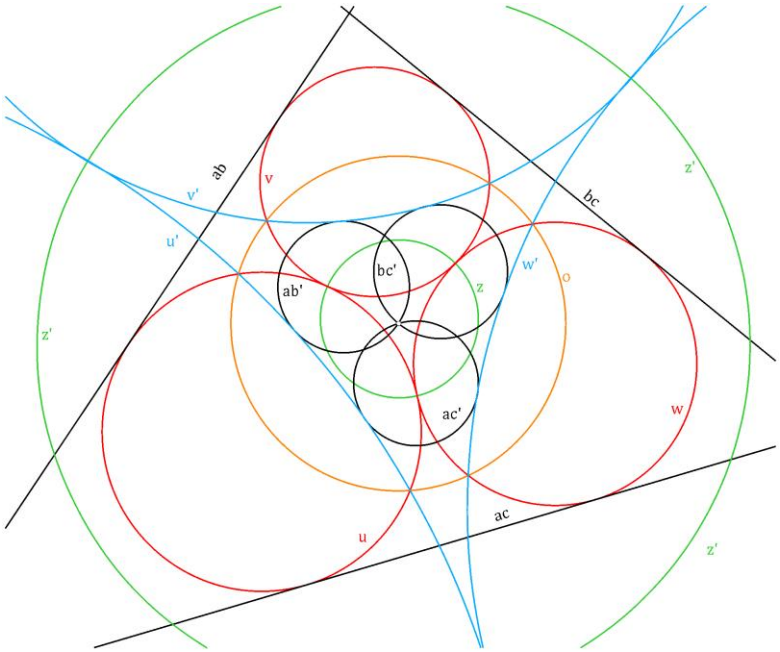


Рис. 6. Задача об окружностях Мальфатти в инверсной постановке

IV. Заключение

В результате проведенного исследования получены следующие результаты:

- дано единое конструктивное обоснование природы задачи об окружностях Мальфатти, заключающееся в нахождении ряда геометрических структур методами построения линий, равноудаленных от пары окружностей;
- представлены примеры рабочего использования алгоритма построения конических кривых, заданных парой фокальной точек и проходящих через заданную точку;
- сформулировано расширенное толкование задачи об окружностях Мальфатти, распространяющееся не только на линейные треугольники, но и на пары сопряженных окружностей, имеющих общую точку; показано место этой задачи в общей классической схеме.

Список литературы

1. Malfatti. Memoirie di matematica. Tomo X. Parte I. Modena. 1803.
2. Steiner J. Einige geometrische Betrachtungen // Crelle J. Parte I. 1826.
3. Petersen J. Methodes et theories pour la resolution des problemes des constructions geometriques. Paris: Gautier-Villars. 1880.
4. Адлер А. Теория геометрических построений / пер. с нем. Г. М. Фихтенгольца. — Л. : Учпедгиз, 1940. С. 15
5. Шклярский Д.О., Ченцов Н.Н., Яглом И.М. Избранные задачи и теоремы математики. Ч. 2. Геометрия (планиметрия). М.: Гостехиздат. 1952. С. 204.
6. В. З. Беленький, А. А. Заславский. Решение обобщенной задачи Мальфатти с помощью комплексной (гиперболической) тригонометрии // Математическое просвещение. — 1998. — Вып. 2. — С. 141-154.
7. Voloshinov D.V. Constructive Geometry of Conics Principal Conjugations. Proceedings of the International Conference on Computer Graphics and Vision “Graphicon”. 2023. № 33. С. 814-825.

УДК 515.74
DOI: 10.25206/978-5-8149-3873-2-2024-755-760

Determining new higher order curves using biquadratic transformation methods

Auyez Baidabekov[1], Elmira Kemelbekova[2],

[1,2] L.N. Gumilev Eurasian National University, Astana, Kazakhstan
a.baydabekov@mail.ru, e.kemelbekova@mail.ru

Abstract. The article deals with fourth-order curves and their construction methods. For the first time, the concept of four-order curves is mentioned in the works of ancient Greek scientists. One such scientist was Perseus who lived in the second century. He obtained fourth-order curves by cutting a torus with planes parallel to the axis. And Nicomedes, a Greek geometer who lived in the third century, used fourth-order curves (Nicomedes shells) to solve segment problems. At the same time, he used fourth-order curves to solve trigonometric problems, so he called this curve the "Nicomedian envelope". Since the eighteenth century, scientists have tried to group fourth-order curves with different symbols. An earlier attempt to group fourth-order curves was made by the English mathematician Edward Waring. In 1694, fourth-order curves were also described in an article by the Swiss Daniel Bernoulli. He found a new fourth-order curve by equating the parameters a and c in this equation and studying the Cassini ellipse, which has two inflection points at the node. This curve was later called "Bernoulli's lemniscate".

D.A. Gudkov and his students studied the shape of fourth-order curves, but the individual figure of fourth-order curves with separate material points and different geometric shapes. D.A. Gudkov obtained 24 types of the complete set of inseparable special forms of the fourth-order curve, whose singular points do not differ in this form, and obtained about 600 curves for these curves. At the beginning of the last century, Zeitana obtained 42 types of fourth- order curves without bending. In the article, the biquadratic transformation method was invented by the multivariate matching method, so we considered how to draw and obtain fourth-order curves using these biquadratic transformation methods. For this purpose, the method of some biquadratic transformations is used. Using some of these biquadratic transformations to obtain a fourth-order curve, the next case shows how the curve is constructed. In this article, we use the graphical model of the biquadratic transformation using the L_8 method to obtain the fourth- order curve. To obtain the fourth-order curve, we first considered a straight line and then a circle as the initial image of the curve under different conditions.

Keywords: second-order curve, third-order curve, fourth-order curve, quadratic transformation, biquadratic transformation.

I. Introduction

The article discusses fourth-order curves and methods for their construction. The concept of curves was first mentioned in the works of ancient Greek scientists. Along with the doctrine of second-order curves, ancient Greek scientists were aware of some third- and fourth-order curves. Apollonius determined the foci, chords, diameters and asymptotes of second-order curves. In the second century, Perseus obtained curves of the fourth order by cutting the torus with planes parallel to the axis. Additionally, the third-century Greek geometer Nicomedes (280–210) used fourth-order curves (Nycomedes shells) to solve problems on three angular segments. Thanks to the formation of R. Descartes in the 17th century. analytical geometry (Cartesian coordinate system), many mathematicians, when solving practical problems, obtained new curves of the fourth order and studied the properties and shapes of these curves.

The ancient Greek scientist Meneme (380-320), who lived in the 4th century BC, studied second-order curves. Euclid (325-265) and Archimedes (287-212) did much work on second-order curves. Ancient Greek scientists studied second-order curves through sections of cones and planes and called them "conic sections." The great Greek geometer Apollonius (250-200) wrote a valuable eight-volume work on second-order curves, which he systematized and developed. Apollonius spoke the language of fragments and circles.

At the end of the 17th century, the English physicist and mathematician J. Newton (1642-1727) created a third set of singular points and curves. He stated that there are three types of singular points and five types of curves of the third order. In addition, for each of the five types of curves, Newton defined the number of curves and the location of the object points. The Greek geometer Nicomedes (280-210) used fourth-order curves to solve three- angle problems in the 3rd century BC, so he called this curve the "Nicomedes envelope". In the 2nd century B.C. Perseus obtained fourth-order curves by cutting the grid with planes parallel to the axis, which are now called "Perseus curves" [1].

Some individual fourth-order curves became known even to Greek scientists. In the 17th century, the French mathematician René Descartes (1596-1650) laid the foundations of analytic geometry, obtained new fourth-order curves in solving practical problems, and studied the properties and shape of these curves. In 1637, R. Descartes defined the equation of fourth-order curves, which were later called "Descartes' ovals," whose equation is as follows

$$((x2 + y2 - 2rx)2 - 22(x2 + y2) -k =0).$$

And in 1662, Sluze obtained a new fourth-order curvilinear equation

$$((x2 + y2)2 y2 = a2 x2),$$

he called this curve the Kappa curve.

The next curve of the fourth order is called the Cassini oval, invented by Italian astronomer Giovanni Domenico Cassini (1625-1712):

$$((x2 + y2)2 - 2c2 (x2 - y2) = a4 - c4).$$

This equation produces curves of different shapes for different ratios of parameters *a* and *c*. One such curve was described in 1694 in an article by Daniel Bernoulli (1700-1782) of Switzerland. He found a new fourth-order curve by equating the parameters a and c in this equation and by studying the Cassini oval with two inflection points at *a* node. The new curve was later renamed Bernoulli's lemniscata.

A cardioid is a fourth-order curve in the projection plane, given by the following equation:

$$((x2 + y2 + 2rx)2 = 4r2 (x2 + y2)).$$

The Swiss mathematician Jakob Steiner (1696-1763) studied a new fourth-order curve, which is given by the following equation:

$$((x2 + y2)2 + 8rx (3y2 - x2)+18r2 (x2 + y2) - 27r4 =0),$$

later called this curve the "Steiner curve. The French mathematician Blaise Pascal (1623-1662) studied Descartes' ovals and determined the individual state of Descartes' ovals:

$$((x2 + y2 - 2rx)2 -l2 (x2 + y2) =0).$$

This fourth-order curve is called "Pascal's cochlea". If *l* is less than *2r*, the "circle of Pascal" (with nodes) passes through the starting point. If *l* and *2r* are equal, then "Pascal's snail" passes through the vertex [9].

Beginning in the eighteenth century, geometric scientists attempted to group fourth-order curves using various symbols. An early attempt to group fourth-order curves was made by the English mathematician (Edward Waring) Waring (1736-1798) (1792). He divided curves into 12 groups, combining 84551 individual curves. The Swiss mathematician Leonhard Euler (1700-1783), 1748 In his work An Introduction to the Analysis of Infinite Sets, he obtained 146 curves. In 1839, the German scientist Julius Plücker (1801-1868) obtained 152 types of fourth-order curves according to his theory of algebraic curves and showed the form of the canonical equation for each of these types. In the late 19th and early 20th centuries, the English mathematician Arthur Kelly (1821-1895) and the Danish geometer Jerome Georg Zeithen (1839-1920) studied fourth-order curves. They distinguished 36 types of fourth- order curves, grouped into 13 groups.

In the 1960s, D.A. Gudkov (1918-1992) [8] and his students compiled a complete classification of types of fourth-order curves on the computational plane. In 1966, in "A complete classification of indivisible 4th-order curves," D. A. Gudkov gave a collection of all types of non-damped 4th-order curves without singular points, and in 1988 he created a series of curve types with imaginary singular points on the constructive plane. In 1990, D. A. Gudkov proposed types of indecidable fourth-order curves. Thus, D.A. Gudkov proved that there are 99 types of indivisible fourth-order curves without singular points, 18 types with imaginary singular points, and 96 types of reduced fourth-order curves.

I. Zeiten [6] and D.A. Gudkov and their students studied the form of fourth-order curves, but they studied separate forms of fourth-order curves with separate material points and different geometric forms. At the beginning of the last century, I. Zeitana obtained 42 types of fourth-order curves without inflection. In 1981, a study of the shape of the fourth-order curve was carried out by O. J. Viro Klein, who published a scientific paper generalizing the formula. Professor D.A. Gudkov obtained 24 kinds of complete set of indivisible singular forms of the fourth- order curve, whose singular points do not differ in this form, and obtained about 600 curves for these curves.

II Formulation of the problem

D.A. Gudkov, having developed several types of curves, formed a geometrical definition of the set of higher- order curves by constructive correspondence of elements. On the basis of this definition, the problem of constructing a higher-order curve was solved. It was found that the solution of the problem depends on a geometric curve whose elements consist of two rectangular clusters by virtue of a multidimensional correspondence [1, 2, 7].

Twelve kinds of biquadratic transformations have been developed using this multivariate correspondence. To obtain a fourth-order curve, we use some of these biquadratic transformations and see how the curve is formed in the following case. To do this, we first get a graphical model of the equation and a fourth-order curve. In scientific work we obtain equation and graphical model of biquadratic transformation by L_8 method [3, 4].

III. Theory

Translated with Translator (free version)To obtain a fourth-order curve, we first take a straight line in various situations as the initial image of the curve.

Let us define the equation of the biquadratic transformation method of the resulting fourth-order curve L_8 in the following order:

$$\sqrt{\frac{x_1'^2+x_2'^2}{2}}-k\sqrt{\frac{x_1'^2-x_2'^2}{2}}-m=0.$$

Then we obtain a graphical model of the quadratic transformation method L8 (Fig. 1). In this graphical model we consider a straight line as the initial image of the Cartesian coordinate system in various situations.

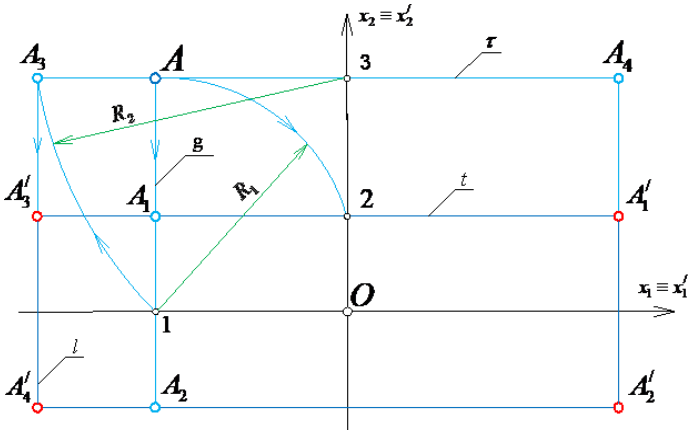


Fig. 1. A graphical model of L_8 biquadratic transformation

III. Experimental results

If in the general case the initial image is a straight line which makes an angle $\alpha=15$ with the axis OX_1, we obtain a fourth-order curve according to Fig. 2.

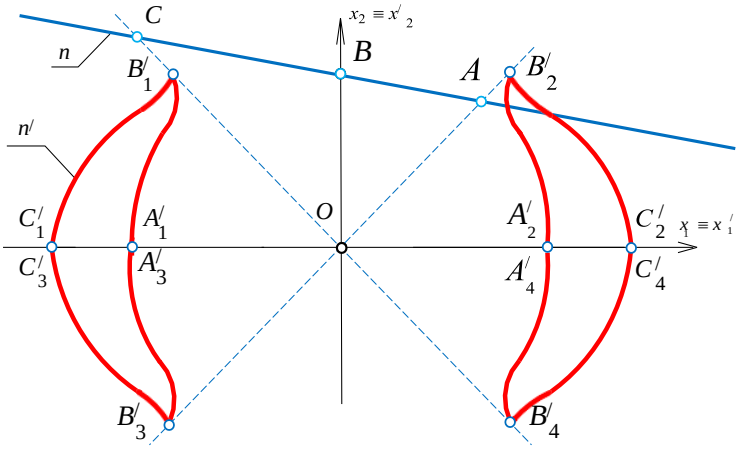


Fig. 2. Fourth-order curve

If the initial image of the line n is located in the general case and the angle of inclination to the axis OX_1 is equal to $\alpha=20^0$ gadus, then we get a new fourth-order curve n' according to Fig. 3.

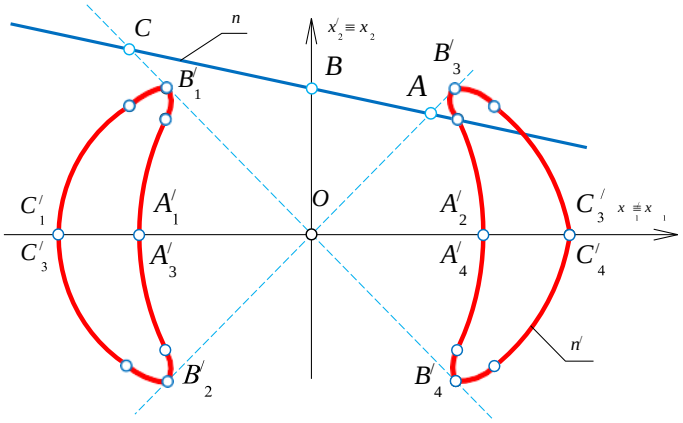


Fig. 3. Fourth-order curve

If the straight line *n* located in the general case has an angle α = 30⁰ to the axis OX_1, then we get a new curve n' of the fourth order according to Fig. 4.

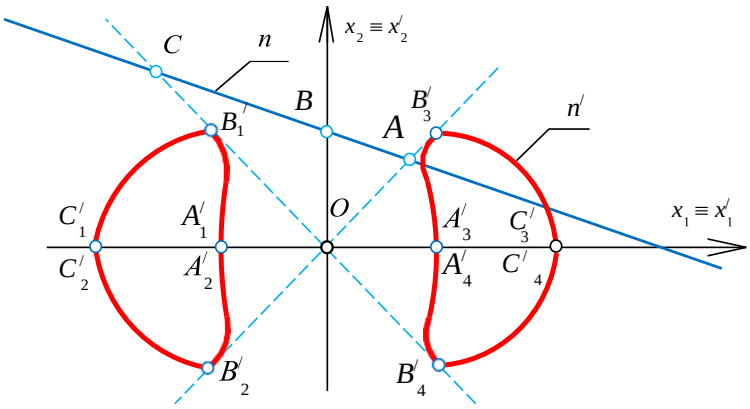


Fig. 4. Fourth order curve

If in the general case the *n* line is located at an angle of α=30⁰ degrees to the OX_1 axis, then according to Fig. 5 we get a new curve n' of the fourth order.

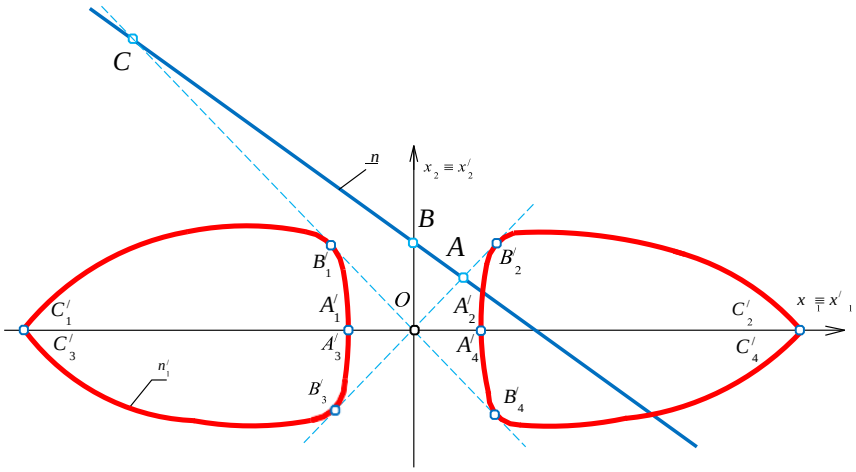


Fig. 5. Fourth order curve

If the prototype is a circle *n* near the axis $0X_2$ and the distance from the center of the prototype to the axis OX_1 *is* $4r$, we get a fourth-order curve n', which is divided into two curves according to Fig. 6.

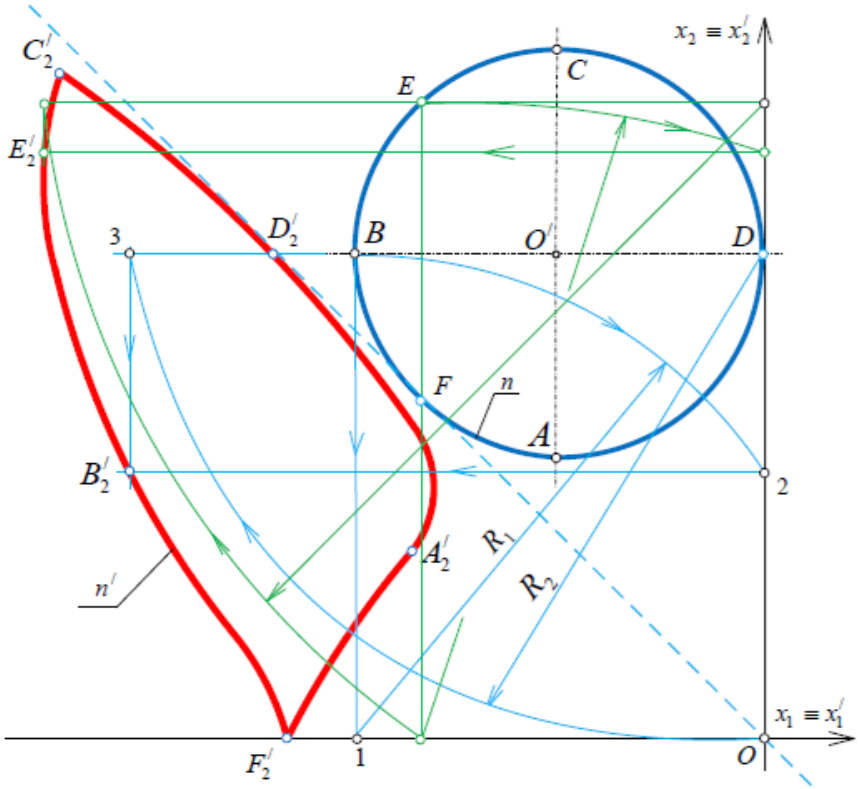


Fig. 6. Fourth order curve

V. Discussion of results

Now we will consider how to obtain the fourth-order equation of this curve that we are studying in this way. For this, we define the prototype of equation *n* as follows:

$$(x_1 + a)^2 + (x_2 + b)^2 = r^2. \tag{1}$$

Let's write down the equations of the biquadratic transformation $L_8{}'$:

$$\begin{cases} x_1 = \sqrt{\dfrac{x_1'^2 - x_2'^2}{2}} \\ x_2 = \sqrt{\dfrac{x_1'^2 + x_2'^2}{2}} \end{cases} \tag{2}$$

If we substitute the values of x_1 and x_2 in equation (1), we obtain the equation of the desired fourth-order curve

$$\left(\sqrt{\frac{x_1'^2 - x_2'^2}{2}} + a\right)^2 + \left(\sqrt{\frac{x_1'^2 + x_2'^2}{2}} + b\right)^2 = r^2, \tag{3}$$

where: *a*, *b*, *r* - are parameters of the circle.

Thus, it was found that the use of graphical models of biquadratic transformations allows drawing new fourth-order curves in applied geometry.

VI. Conclusions and Conclusion

The main result of this work is the development of methods for obtaining fourth-order curves using biquadratic transformations. The proposed biquadratic transformation methods allow obtaining fourth-order curves and determining their equations, that is, a new direction has been opened in the theory of plane curves. Thus, the developed methods of curve modeling by means of biquadratic transformations differ from other methods of obtaining curves in that they are defined by the curve prototype and biquadratic transformation. It is determined that the properties of the obtained curves depend on the parameters of the prototype and the biquadratic transformation device. Using graphical models of biquadratic transformations, new types of fourth-order curves are obtained, which can be used to construct surfaces of various shapes in applied geometry.

Reference

1. Andreev K.A. On the geometric formation of plane curves. -Kharkov, 1875.
2. Andreev, K.A. On geometric correspondences in application to the question of constructing curved lines / K.A. Andreev. -Moscow, 1879.
3. Hayashi T. Notes on algebraical plane curves/ T.Hayashi// Toh. Math. Journ., No. 2, 1912, pp. 100-107.
4. Smogorzhevsky, A.S. Handbook on the theory of plane curves of the third order / A.S. Smogorzhevsky, E.S. Dining room. -M.: FIZMALTIT, 1961, -271 p.
5. Nurmakhanov B.N. Development of algorithms for modeling nonlinear point correspondences of a plane generated by the establishment of binary models of surfaces, and their practical application: abstract of Ph.D. tech. Sciences: 01/05/01. -Kiev: 1978. -18 p.
6. Fomenko A. T., Visual geometry and topology. Mathematical images in the real world. -M.: Publishing house Moscow. Univ., 1992. -432 p.
7. Rosendfeld, B.A. Apollonius of Perga. -M.: MTsNMO, 2004, -176 p.
8. Baidabekov A.K. Biquadratic transformations. Monograph. -Minsk: BNTU, 2012. -188 p.
9. Adamova R.S. Curves of fourth order in the complex projective plane. -Voronezh: Bulletin of VSU. Series: Physics. Mathematics. 2013. No. 2. P. 114-117.

УДК 514.122.3
DOI: 10.25206/978-5-8149-3873-2-2024-761-765

Сопряжение набора Безье кривых с анализом точности в САПР «Сарус»
Conjugation of a set of Bezier curves with accuracy analysis in CAD "Sarus"

С.Н. Ганчук[1,2], О.В. Кривошеев[1,2], С.В. Маврин[1,2], С.А. Рыжов[1,2]
[1] Саровский физико-технический институт - филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский ядерный университет "МИФИ" (СарФТИ НИЯУ МИФИ), ул. Духова, д.6, г. Саров, 607186, РФ
[2] Российский федеральный ядерный центр - Всероссийский научно-исследовательский институт экспериментальной физики (РФЯЦ-ВНИИЭФ) - федеральное государственное унитарное предприятие Государственной корпорации по атомной энергии "Росатом", пр. Мира, 37, г. Саров, 607188, РФ

S.N. Ganchuk[1,2], O.V. Krivosheev[1,2], S.V. Mavrin[1,2], S.A. Ryzhov[1,2]
[1] Sarov Institute of Physics and Technology - branch of the Federal State Autonomous Educational Institution of Higher Education "National Research Nuclear University "MEPhI" (SarPhTI NRNU MEPhI), Dukhova str., 6, Sarov, 607186, Russian Federation
[2] Russian Federal Nuclear Center - All-Russian Scientific Research Institute of Experimental Physics (RFNC-VNIIEPh) - Federal State Unitary Enterprise of the State Atomic Energy Corporation "Rosatom", 37 Mira Ave., Sarov, 607188, Russian Federation

Аннотация. В современных CAD- и CAE-системах одной из геометрических операций является полностью гладкая аппроксимация набора смежных кривых Безье произвольной степени. Под полностью гладкой аппроксимацией понимается линия порядок непрерывности которой совпадает со степенью Безье кривой. Для выполнения этого требования в работе формулируется оптимизационная задача с ограничениями в виде равенств, решение которой сводится в системе линейных алгебраических уравнений. Дополнительными ограничениями могут выступать заданные значения производных в граничных точках.

Для контроля отклонения исходной кривой от аппроксимирующей используются три метрики. Метрика Хаусдорфа (максимальное линейное расстояние), метрика изменения кривизны и квадратичная разность.

Алгоритм непрерывного и гладкого соединения кривых был реализован на языке программирования C++ и интегрирован в CAD-систему «Сарус», разрабатываемую предприятием Госкорпорации «Росатом» РФЯЦ-ВНИИЭФ.

Алгоритм непрерывного и гладкого соединения кривых был реализован на языке программирования C++ и интегрирован в CAD-систему «Сарус» [1], разрабатываемую предприятием Госкорпорации «Росатом» РФЯЦ-ВНИИЭФ.

Ключевые слова: кривая Безье; система автоматизированного проектирования; сопряжение; имплементация.

Abstract. In modern CAD and CAE systems, one of the geometric operations is a completely smooth approximation of a set of adjacent Bezier curves of arbitrary degree. By completely smooth approximation we mean a line whose order of continuity coincides with the Bezier degree of the curve. To fulfill this requirement, the work formulates an optimization problem with constraints in the form of equalities, the solution of which is reduced in a system of linear algebraic equations. Additional restrictions may include specified values of derivatives at boundary points.

Three metrics are used to monitor the deviation of the original curve from the approximating one. Hausdorff metric (maximum linear distance), curvature change metric and quadratic difference.

The algorithm for continuous and smooth connection of curves was implemented in the C++ programming language and integrated into the Sarus CAD system, developed by the Rosatom State Corporation RFNC-VNIIEF.

Keywords*:* Bezier curve; computer-aided design system; conjugation; implementation.

I. Постановка задачи

Рассмотрим $(n+1)$ Безье кривую степени p[24]:

[24] если заданы кривые различных степеней, то степени кривых меньших степеней необходимо увеличить до максимальной степени, процедурой, которая не искажает исходную кривую Безье [4].

$$\begin{matrix} P^0(u_0) = \sum_{l=0}^{p} P_l^0 N_l^p(u_0), & u_0 \in [0,1] \\ P^1(u_1) = \sum_{l=0}^{p} P_l^1 N_l^p(u_1), & u_1 \in [0,1] \\ \vdots & \\ P^n(u_n) = \sum_{l=0}^{p} P_l^n N_l^p(u_n), & u_n \in [0,1] \end{matrix} \quad (1)$$

где $P^j(u_j)$ – j-я Безье кривая, u_j – параметрическая переменная j-й кривой Безье, P_l^j - l-я контрольная точка j-й кривой Безье, $N_l^p(u_j)$ - l-я базисная функция степени p.

j-й узловой вектор определим следующим образом:

$$U_j = (\underbrace{0, \dots, 0}_{p}, 0,1, \underbrace{1, \dots, 1}_{p})$$

Набор кривых (1) представим как одну составную кривую:

$$R(t) = \begin{cases} P^0\left(\frac{t}{\lambda_0}\right) = \sum_{l=0}^{p} P_l^0 N_l^p\left(\frac{t}{\lambda_0}\right), & t \in [0, \lambda_0], \\ P^1\left(\frac{t-\lambda_0}{\lambda_1-\lambda_0}\right) = \sum_{l=0}^{p} P_l^1 N_l^p\left(\frac{t-\lambda_0}{\lambda_1-\lambda_0}\right), & t \in [\lambda_0, \lambda_1], \\ \vdots & \\ P^n\left(\frac{t-\lambda_{n-1}}{1-\lambda_{\lambda_{n-1}}}\right) = \sum_{l=0}^{p} P_l^n N_l^p\left(\frac{t-\lambda_{n-1}}{1-\lambda_{\lambda_{n-1}}}\right), & t \in [\lambda_{n-1}, 1], \end{cases}$$
$$t \in [0,1],$$
$$0 < \lambda_0 < \cdots < \lambda_{n-1} < 1,$$

где λ_i, $i = \overline{0,(n-1)}$ – возрастающая последовательность чисел.

Постановка задачи заключается в последовательном выполнении 2-х этапов. Этап 1 - реализовать полностью гладкое сопряжение кривых (1). Этап 2 – рассчитать параметры кривой $Q(t)$ (Безье кривая степени p), которая минимизирует метрику $d(R,Q)$ между кривыми $R(t)$ и $Q(t)$.

Термин полностью гладкое сопряжение означает аппроксимацию кривых $P^j(u_j)$ кривыми $\hat{P}^j(u_j)$, $j = \overline{0,n}$ таким образом, что вплоть до p-го порядка:

$$\hat{P}^{j^{(k)}}(u_j)|_{u_j=1} = \hat{P}^{j+1^{(k)}}(u_{j+1})|_{u_{j+1}=0}, \quad j = \overline{0,(n-1)} \quad (2)$$

где $\hat{P}^{j^{(k)}}(u_j)$ – k-я производная j-й Безье кривой, которая аппроксимирует кривую $P^j(u_j)$.

Условия (2) могут быть расширены наложением дополнительных условий неизменности производных в начальной и предельной точках кривой $R(t)$:

$$\begin{matrix} \hat{P}^{0^{(k1)}}(u_0)|_{u_0=0} = P^{0^{(k1)}}(u_0)|_{u_0=0}, \\ \hat{P}^{n^{(k2)}}(u_n)|_{u_n=1} = P^{n^{(k2)}}(u_n)|_{u_n=1}, \end{matrix} \quad (3)$$

где $k1$, $k2$ –означает порядок производных.

Дополнительные ограничения (3) уменьшают порядок непрерывности «k» в условиях (2) на величину (max ($k1$, $k2$) + 1) [1], то есть:

$$k_{new} = k - [\max(k1, k2) + 1],$$

где k_{new} - порядок непрерывности гладкого сопряжения (2) при дополнительных условиях (3).

$\hat{P}^j(u_j)$ в (2) запишем через приращение координат:

$$\hat{P}^j(u_j) = \sum_{l=0}^{p} \hat{P}_l^j N_l^p(u_j) = \sum_{l=0}^{p} (P_l^j + \varepsilon_l^j) N_l^p(u_j), \quad (4)$$

где ε_l^j – приращение координат контрольных точек P_l^j.

Метрику $d(R,Q)$ по аналогии с [2] представим в виде дискретной нормы:

$$d(R,Q) = \sum_{j=0}^{n} \sum_{l=0}^{p} \left\| \varepsilon_l^j \right\|^2 \to min. \quad (5)$$

Оптимизационная задача (5) решается при ограничениях (2), (3).

II. Метод решения

Используя представления (4) для записи условий (2), получаем систему n равенств:

$$\sum_{l=0}^{p}\left(P_l^j+\varepsilon_l^j\right)N_l^{p^{(k)}}(1)=\sum_{l=0}^{p}\left(P_l^{j+1}+\varepsilon_l^{j+1}\right)N_l^{p^{(k)}}(0),\quad j=\overline{0,(n-1)} \tag{6}$$

где $N_l^{p^{(k)}}(1)$, $N_l^{p^{(k)}}(0)$ – k-я производная l-й базисной функции степени p в точках 1 и 0, соответственно.

Для решения оптимизационной задачи (5) при ограничениях (4) используется метод неопределенных множителей Лагранжа [3], с функцией Лагранжа вида:

$$L=\sum_{j=0}^{n}\sum_{l=0}^{p}\left\|\varepsilon_l^j\right\|^2+\sum_{j=0}^{n-1}\sum_{k=0}^{p}\lambda_k^j\begin{bmatrix}\sum_{l=0}^{p}\left(P_l^j+\varepsilon_l^j\right)N_l^{p^{(k)}}(1)-\\ \sum_{l=0}^{p}\left(P_l^{j+1}+\varepsilon_l^{j+1}\right)N_l^{p^{(k)}}(0)\end{bmatrix} \tag{7}$$

где λ_k^j – множители Лагранжа, $j=\overline{0,(n-1)}$.

Приравнивая их нулю частные производные функции Лагранжа (7) по ε_l^j, $j=\overline{0,n}$ и λ_k^j, $j=\overline{0,(n-1)}$, приходим к системе линейных уравнений:

$$\frac{\partial L}{\partial\varepsilon_k^0}=2\varepsilon_k^0+\sum_{i=0}^{p}\lambda_i^0\,N_k^{p^{(i)}}(1)=0,\quad k=\overline{0,p} \tag{8}$$

$$\vdots$$

$$\frac{\partial L}{\partial\varepsilon_k^j}=2\varepsilon_k^j+\sum_{i=0}^{p}\left[-\lambda_i^{j-1}N_k^{p^{(i)}}(0)+\lambda_i^jN_k^{p^{(i)}}(1)\right]=0,\quad j=\overline{1,(n-1)},k=\overline{0,p} \tag{9}$$

$$\vdots$$

$$\frac{\partial L}{\partial\varepsilon_k^n}=2\varepsilon_k^n+\sum_{i=0}^{p}\lambda_i^{n-1}\,N_k^{p^{(i)}}(0)=0,\quad k=\overline{0,p} \tag{10}$$

$$\frac{\partial L}{\partial\lambda_k^j}=\sum_{l=0}^{p}\left(P_l^j+\varepsilon_l^j\right)N_l^{p^{(k)}}(1)-\sum_{l=0}^{p}\left(P_l^{j+1}+\varepsilon_l^{j+1}\right)N_l^{p^{(k)}}(0)=0,\quad j=\overline{1,(n-1)}. \tag{11}$$

После перегруппировки в (11) имеем:

$$\sum_{l=0}^{p}\varepsilon_l^jN_l^{p^{(k)}}(1)-\sum_{l=0}^{p}\varepsilon_l^{j+1}N_l^{p^{(k)}}(0)=\sum_{l=0}^{p}\left[-P_l^jN_l^{p^{(k)}}(1)+P_l^{j+1}N_l^{p^{(k)}}(0)\right] \tag{12}$$

Таким образом получили систему $[(p+1)(2n+1)]$ линейных уравнений (8)-(10) и (12) относительно ε_l^j, $j=\overline{0,n}$ и λ_k^j, $j=\overline{0,(n-1)}$.

III. Метрики сравнения кривых

Для определения точности и сравнения результатов алгоритма в данной работе используются следующие метрики: метрика Хаусдорфа, кривизна сплайна и квадратичная разность между кривыми. Рассмотрим каждую метрику подробнее. Сразу определим, что кривая Безье будет вычисляться как частный случай NURBS.

Метрика Хаусдорфа – перпендикуляр (наименьшее расстояние) между двумя сплайн-кривыми. Для определения этой метрики рассмотрим задачу нахождения расстояния от некоторой точки до сплайн-кривой. Пусть нам задана точка P и кривая $C(u)$, тогда расстояние от точки P до кривой $C(u)$, будет соответствовать длине минимального отрезка:

$$\|P-C(u)\|=\min_{Q\in C(u)}\|P-Q\|,$$

где P – заданная точка;

$C(u)$ – сплайн-кривая;

Q – точка сплайн-кривой $C(u)$.

Это расстояние находится с помощью итеративного численного метода Ньютона [4]:

$$u_{i+1}=u_i-\frac{f(u_i)}{f'(u_i)}=u_i-\frac{C'(u_i)\cdot(C(u_i)-P)}{C''(u_i)\cdot(C(u_i)-P)+|C'(u_i)|^2},$$

где $P=\begin{pmatrix}x\\y\end{pmatrix}$ – координаты точки, от которой ищется расстояние до кривой $C(u)$;

u_i – параметрическая переменная сплайн-кривой.

Вторая метрика – кривизна, – определяет степень отклонения от прямой линии и рассчитывается как интеграл второй производной кривой:

$$\sum_{i=1}^{n}\left(C''(u_i)\right)^2 * \frac{1}{numPoints},$$

где $C''(u_i)$ – вторая производная кривой;

$numPoints$ – количество точек построения кривой на параметрическом интервале [0,1].

Квадратичная разность между кривыми – это площадь отклонения между двумя кривыми, которая вычисляется как интеграл квадрата разности:

$$\int_0^1\left(C_1(u_i) - C_2(u_i)\right)^2 du,$$

где $C_1(u_i)$ – исходная кривая;

$C_2(u_i)$ – аппроксимирующая кривая.

IV. Имплементация в CAD «Сарус» и тестирование

Данный алгоритм был имплементирован как дополнительный модуль в импортонезависимую CAD «Сарус» с использованием языка программирования C++ и современной парадигмы ООП.

Интерфейс разработанного модуля изображен на рисунке 1. В нём мы можем селектировать объект на 3D сцене и все необходимые параметры в интерфейсе будут заполнены автоматически. Также мы можем считать контрольные точки кривой и ее узловой вектор из файла, в случае того, если у нас не изображена кривая на 3D сцене.

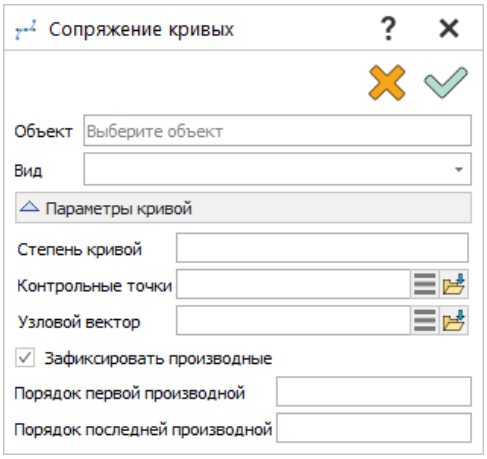


Рис. 1. Интерфейс модуля «Сопряжение кривых» в CAD «Сарус»

Рассмотрим пример работы реализованного модуля. Построим составную кривую из 4-х кривых Безье 3-й степени (см. рисунок 2).

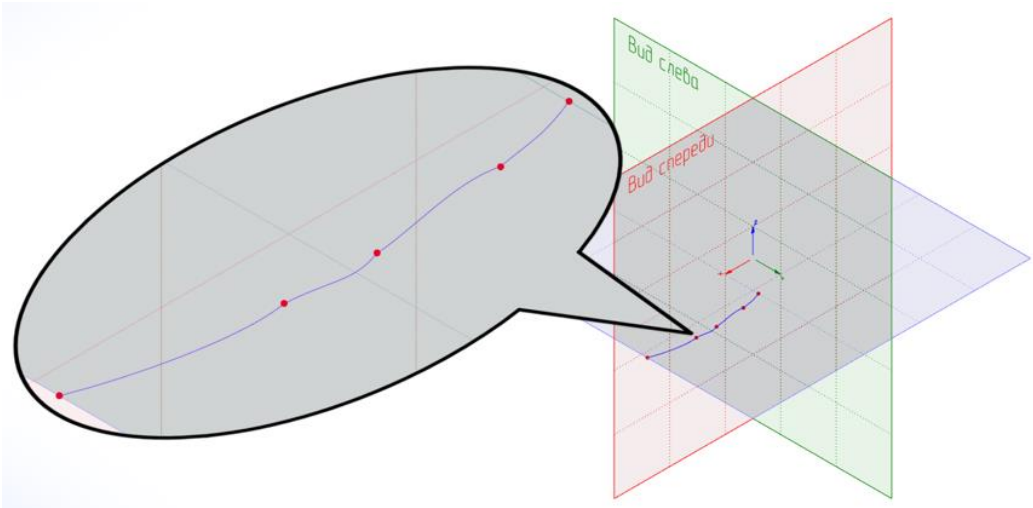


Рис. 2. Составная кривая из 4-х кривых Безье 3-й степени в 3D

Из рисунка 2 мы можем заметить, что в точках соединения кривых Безье сами кривые являются не гладкими, то есть имеется разрыв в точках производных некоторых порядков и тем самым получается острый угол на границе кривых. Особенно это заметно при сильном приближении (см. рисунок 3).

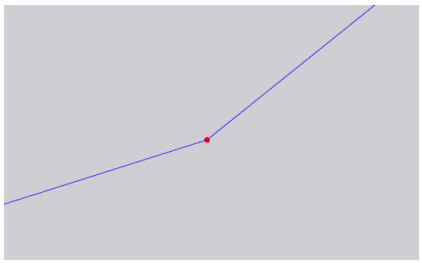


Рис. 3. Приближенная точка соединения двух кривых из рисунка 2

После применения алгоритма сопряжения мы перестроим текущую кривую и построим новую, гладкую непрерывную кривую (см. рисунок 4). Мы получили новые контрольные точки составной кривой, чтобы кривые Безье сдвинулись и точки сопряжения были сглажены.

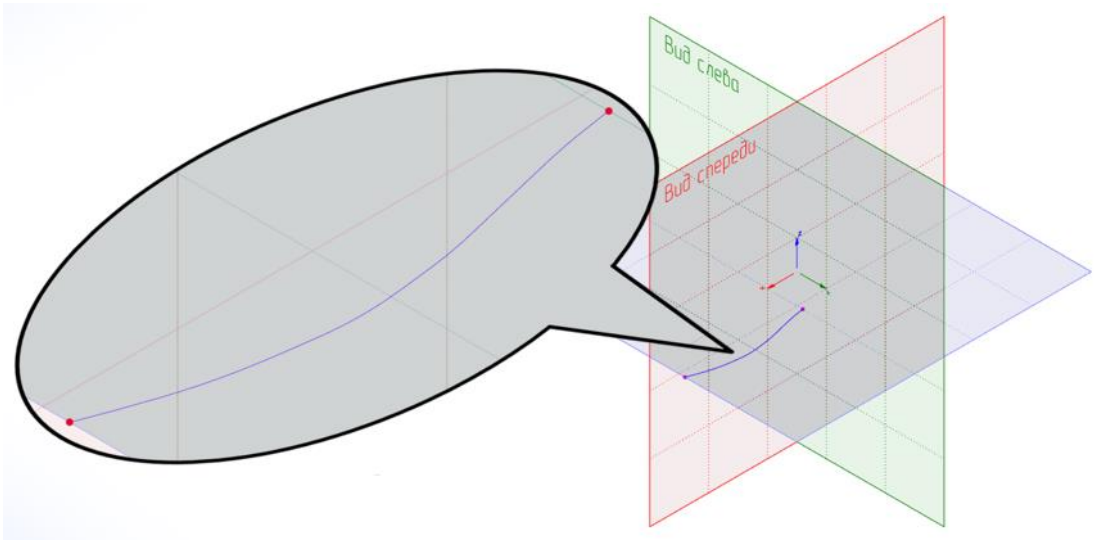


Рис. 4. Составная кривая из 4-х кривых Безье 3-й степени после применения алгоритма сопряжения

С помощью реализованных метрик сравнения кривых был проведен сравнительный анализ и получили следующие результаты: кривизна изменилась с 593 до 1135, квадратичная разность между исходной и непрерывно гладкой кривой равна 410.65, а метрика Хаусдорфа равняется 1.27355.

IV. Выводы и заключение

В результате проделанной работы был реализован и запрограммирован алгоритм непрерывного гладкого сопряжения кривых Безье в 3D. Также были разработаны три метрики сравнения кривых, что позволило контролировать точность изменения новой кривой от исходной. Разработанный алгоритм сопряжения был имплементирован в импортонезависимую CAD систему «САРУС» как дополнительный функциональный модуль.

Список литературы

1. Hu S.-M., Tong R.-F., Ju T., Sun J.-G. [Approximate merging of a pair of Bezier curves]. *Computer-Aided Desing*. (33) 2001, pp. 125–136.

2. Аппроксимация сопряжения кривых Безье с сохранением порядка гладкости и дополнительными ограничениями / С.Н. Ганчук, О.В. Кривошеев, С.В. Маврин, С.А. Рыжов // Вестник ЮУрГУ. Серия «Строительство и архитектура». 2024. Т. 24, № 1. С. 59–69. DOI: 10.14529/build240108

3. Алексеев В.М., Тихомиров В.М., Фомин С.В. Оптимальное управление. 2-е изд. М.: ФИЗМАТЛИТ, 2005. 384 с.

4. Piegl L., Tiller W. The NURBS book. 2nd ed. New York: Springer; 1997. C. 646.

УДК 004.925.8
DOI: 10.25206/978-5-8149-3873-2-2024-766-775

Исследование операции «Скругление» в программном обеспечении «Система полного жизненного цикла «САРУС»
Investigation of the "Rounding" operation in the software "SARUS Full Life Cycle System"

Е.А. Самарина[1], Н.А. Денисова[1]
[1] Саровский физико-технический институт - филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский ядерный университет "МИФИ" (СарФТИ НИЯУ МИФИ), ул. Духова, д.6, г. Саров, 607186, РФ

E.A. Samarina[1], N.A.Denisova[1]
[1] Sarov Institute of Physics and Technology - branch of the Federal State Autonomous Educational Institution of Higher Education "National Research Nuclear University "MEPhI" (SarPhTI NRNU MEPhI), Dukhova str., 6, Sarov, 607186, Russian Federation

Аннотация. Одна из самых труднорешаемых задач в системах автоматизированного проектирования – построение скруглений при моделировании объектов сложных форм. За построение скруглений, как и за всю геометрию в САПР, отвечает геометрическое ядро.

В рамках исследования программного модуля «Система конструкторского проектирования» (ПМ CAD) комплекса программ «СПЖЦ «Цифровое предприятие» («САРУС») изучение функции «Скругление» является достаточно актуальной задачей, так как при проектировании деталей сложных форм возникают проблемы с данной функцией. Таким образом, результаты данной работы позволят в дальнейшем более эффективно использовать функцию «Скругление» при моделировании деталей сложных форм.

В данной работе исследован метод построения скруглений средствами ПМ CAD, проведен сравнительный анализ с другими известными САПР (например, КОМПАС 3D), подробно описан оптимальный порядок построения скруглений, с учетом выявленных особенностей в работе команд.

Ключевые слова. САПР, Программный модуль CAD, 3D-модель, функции построения, операции скруглений.

Abstract. One of the most difficult tasks in computer – aided design systems is the construction of fillets when modeling objects of complex shapes. The geometric core is responsible for the construction of fillets, as well as for all geometry in CAD.

As part of the study of the software module "Design Engineering System" (PM CAD) of the software package "FLCS "Digital Enterprise" ("SARUS"), the study of the "Rounding" function is quite an urgent task, since problems arise with this function when designing parts of complex shapes. Thus, the results of this work will make it possible to use the "Rounding" function more effectively in the future when modeling parts of complex shapes.

In this paper, the method of constructing fillets by means of PM CAD is investigated, a comparative analysis with other well-known CAD systems (for example, COMPASS 3D) is carried out, the optimal order of constructing fillets is described in detail, taking into account the identified features in the work of the command.

Keywords. CAD, CAD software module, 3D model, construction functions, rounding operations.

I. Введение

Одна из самых труднорешаемых задач в системах автоматизированного проектирования – построение скруглений при моделировании объектов усложненных форм. За построение скруглений, как и за всю геометрию в САПР, отвечает геометрическое ядро, позволяющее создавать 3D-модели в CAD. [1]

В рамках исследования программного модуля «Система конструкторского проектирования» (далее ПМ CAD) комплекса программ «СПЖЦ «Цифровое предприятие» (далее - «САРУС») изучение функции «Скругление» является достаточно актуальной задачей, так как при проектировании деталей сложных форм возникают проблемы с данной функцией. Ожидается, что результаты данной работы позволят в дальнейшем более эффективно использовать функцию «Скругление» при моделировании деталей с повышением конструктивной сложности элементов их поверхностей.

II. Постановка задачи

Целью данной работы является изучение функции «Скругление» в СПЖЦ «САРУС» для более эффективного использования при моделировании деталей II группы конструктивной сложности с усложнением элементов форм.

Для достижения данной цели поставлены следующие задачи:

- Проанализировать алгоритмы функции «Скругление» в СПЖЦ «САРУС» и проверить их работоспособность на 3D-моделях деталей простой формы (I, II групп конструктивной сложности).
- Продемонстрировать работу данных алгоритмов на деталях с усложнением элементов формы в СПЖЦ «САРУС».
- Провести сравнительный анализ с другими известными САПР, в частности КОМПАС 3D.

III. Теоретическое обоснование математической основы исследования

На рис. 1 показаны детали, построенные с использованием операции «Скругление» в «САРУС» и КОМПАС 3D соответственно.

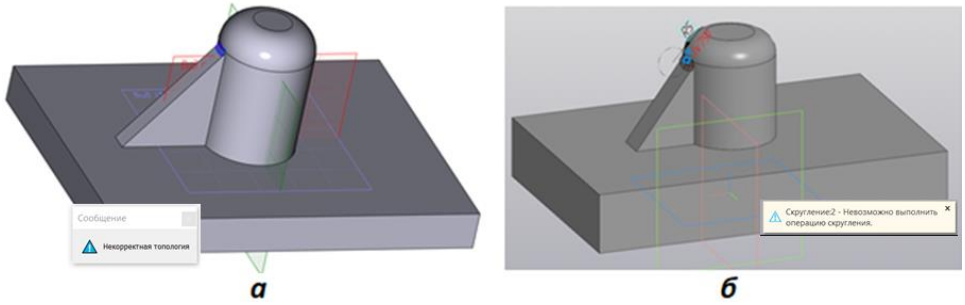


Рис. 1. 3D-модели: *а* – построена в СПЖЦ «САРУС»; *б* – построена в КОМПАС-3D

Как видно из данного простого примера – программные сбои по операции «Скругление» встречаются в обеих CAD-системах.

Задача данной работы – проанализировать работоспособность алгоритмов построения скруглений в «САРУС» [2] и сравнить реализацию функций с ПО «КОМПАС-3D [3].

IV. Результаты экспериментов

IV.1. Анализ алгоритмов функции «Скругление в СПЖЦ «САРУС»

В основной версии программного модуля СПЖЦ «САРУС» ПЗ ТП 07623615.00423-04 81 01-6 Часть 6 [1] предложен ряд алгоритмов по созданию скруглений для ряда функций.

Функция 1 (рис. 2). Построение скруглений рёбер твёрдого тела (базовая реализация). Функция должна обеспечивать:

1. возможность сформировать вдоль выбранного ребра поверхность плавного перехода от одной грани к другой, с условием, чтобы в каждом сечении этой поверхности, построенном перпендикулярно ребру, получалась дуга окружности заданного радиуса;
2. назначение радиуса скругления разным ребрам модели, при выполнении операции скругления (все ребра принадлежат одному телу);
3. изменение значений радиусов скругления одним из двух способов:
   - перемещением маркера радиуса;
   - введением значения в динамическом поле ввода (строка редактирования с поддержкой ЕИ и переменных);
4. последующее редактирование радиусов скругления.

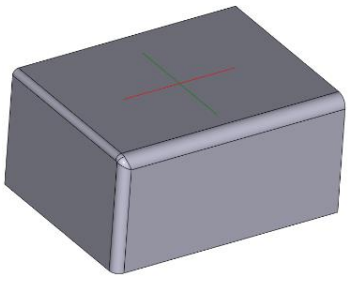


Рис. 2. Реализация функции «Построение скруглений рёбер твёрдого тела»:
а – окно редактирования функции; *б* – результат реализации функции

Функция 2 (рис. 3). Возможность создавать несколько скруглений разного радиуса в одной команде (базовая реализация). Функция должна обеспечивать:

1) задание разных радиусов скругления в одной команде;
2) редактирование радиусов скругления (рисунок 4).

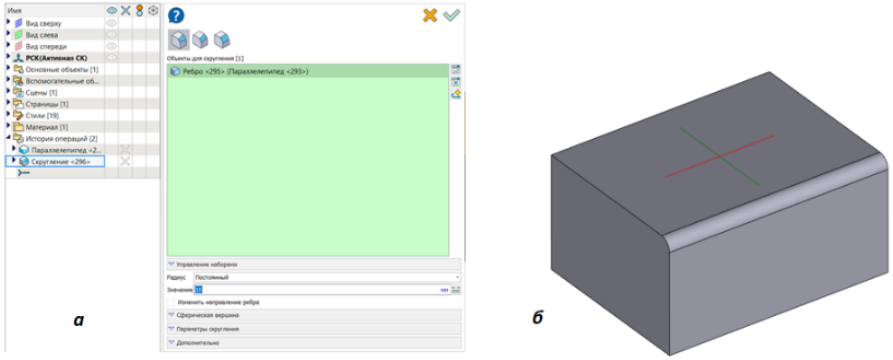


Рис. 3. Результат работы функции «Скругление» с несколькими ребрами с различными значениями радиуса и с использованием сферической вершины

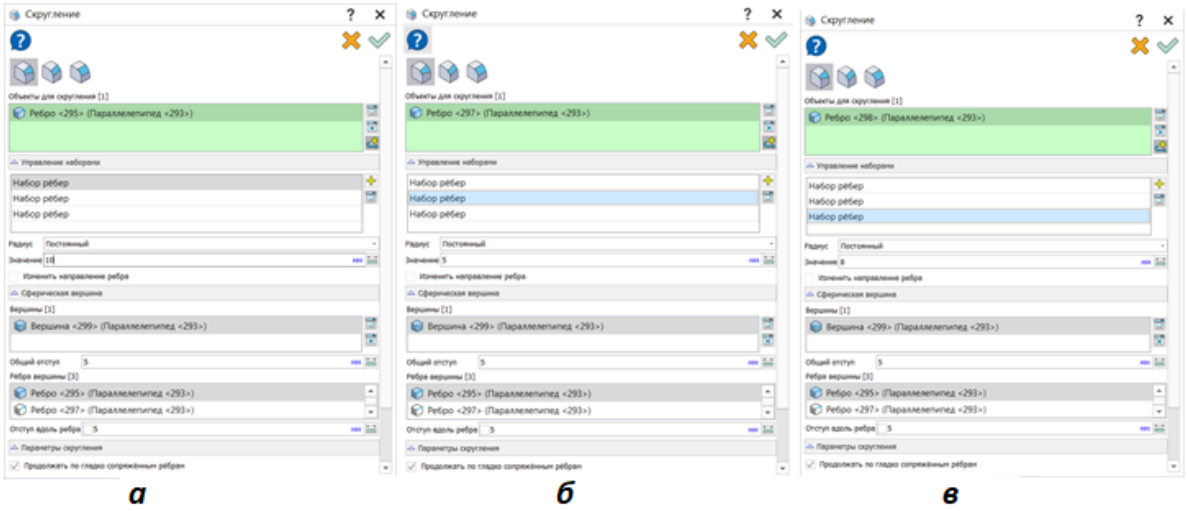


Рис. 4. Задание в одной функции разных радиусов скругления:
а - радиус скругления – 10 мм; *б* - радиус скругления – 5 мм; *в* - Радиус скругления – 8 мм

Функция 3 (рис. 5). Построение неполных скруглений с опорой на ребро или на продолжение грани. Функция должна обеспечивать:

1) выбор ребра;
2) изменение значения радиуса скругления;
3) изменение значений перемещений от граней;
4) редактирование заданных значений радиуса скругления;
5) редактирование заданных значений перемещений от граней.

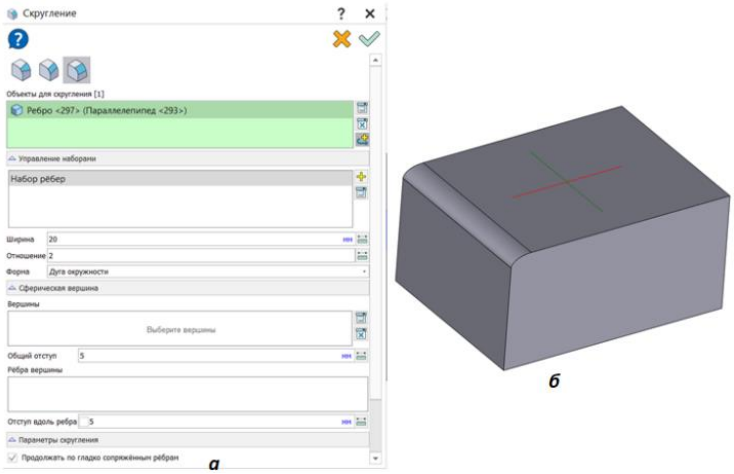


Рис. 5. Реализация функции «Построение неполных скруглений с опорой на ребро или на продолжение грани»: *а* – окно редактирования функции; *б* – результат реализации функции

Функция 4. (рис. 6). Возможность создавать несколько скруглений разного радиуса в одной команде. Функция должна обеспечивать:

1) выбор нескольких ребер модели с использованием клавиши «Ctrl» на клавиатуре;
2) назначение разных радиусов скругления разным ребрам изделия при выполнении операции скругления;
3) изменение значения радиусов скругления одним из двух способов: перемещением маркера радиуса или введением значения в динамическом поле ввода;
4) возможность дальнейшего редактирования радиусов скругления.

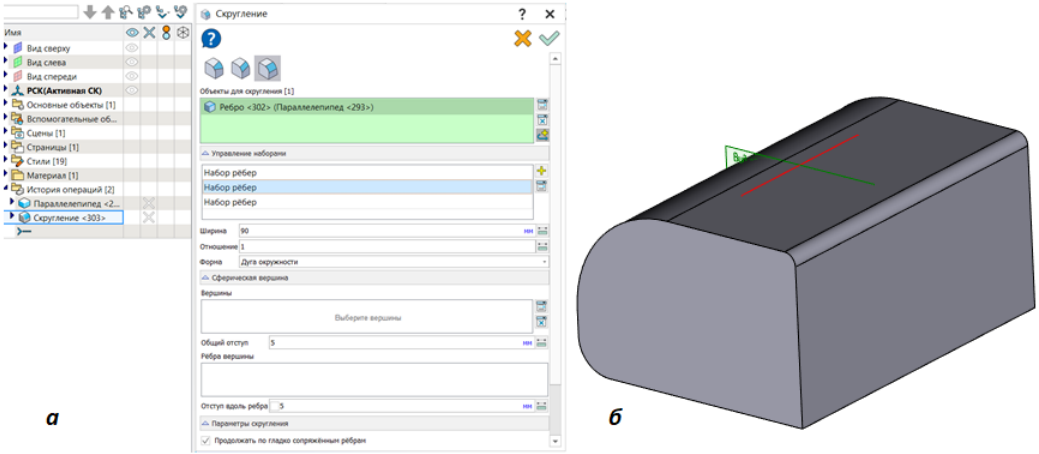


Рис. 6. Реализация функции «Возможность создавать несколько скруглений разного радиуса в одной команде»: *а* – окно редактирования функции; *б* – результат реализации функции

Функция 5 (рис. 7). Возможность назначать участки с переменным радиусом скругления. Функция должна обеспечивать:

1) назначение разных радиусов скругления разным участкам ребер модели при выполнении операции скругления;
2) изменение значения радиусов скругления одним из двух способов: перемещением маркера радиуса или введением значения в динамическом поле ввода;
3) дальнейшее редактирование радиусов скругления.

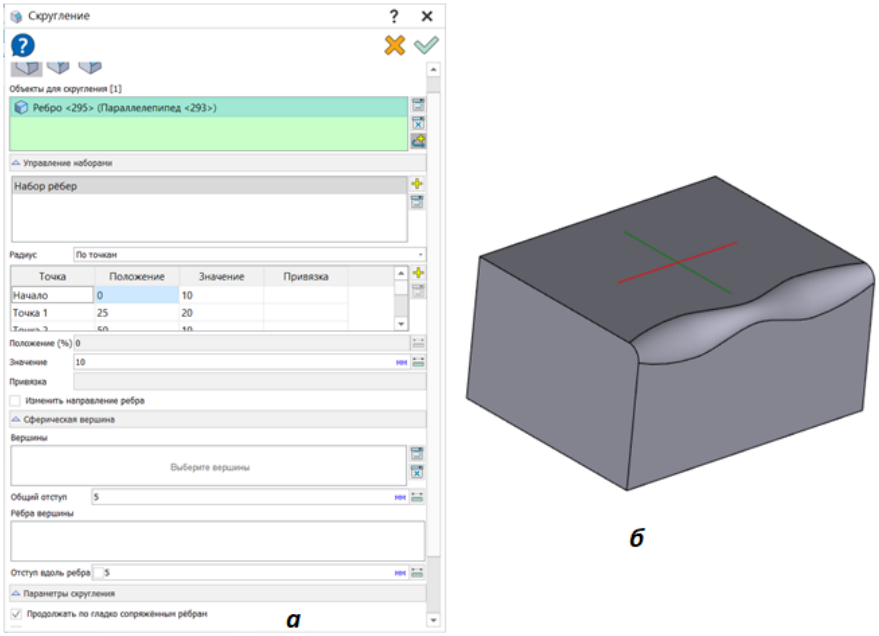


Рис. 7. Реализация функции «Возможность назначать участки с переменным радиусом скругления»: *а* – окно редактирования функции; *б* – результат реализации функции

Функция 6 (рис. 8). Возможность задавать резкий и плавный (S-образный) переход радиуса для участков переменного радиуса Функция должна обеспечивать:

1) назначение линейного перехода радиуса на участках переменного радиуса при выполнении операции скругления;
2) назначение плавного перехода радиуса на участках переменного радиуса при выполнении операции скругления.

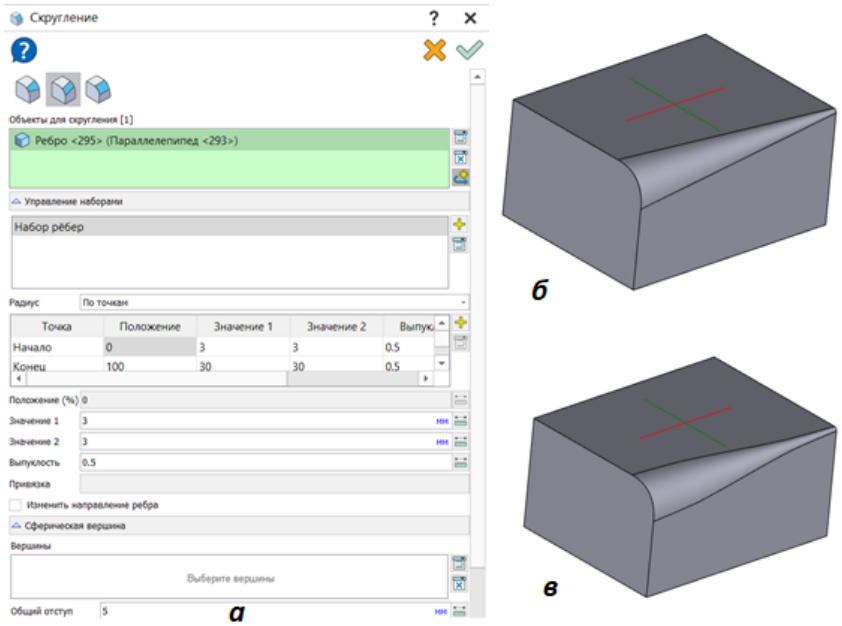


Рис. 8. Реализация функции «Возможность задавать резкий и плавный (S-образный) переход радиуса для участков переменного радиуса»: *а* – окно редактирования функции; *б* – результат реализации функции с линейным переходом; *в* - результат реализации функции с плавным переходом

Функция 7 (рис. 9). Возможность управлять формой скругления для вершин путём задания отступа от вершины до начала регулярного скругления по каждому направлению. Функция должна обеспечивать:

1) задание отступов для ребер, сходящихся в одной вершине;
2) задание разных значений отступов одним из двух способов: перемещением маркера значения отступа или введением значения в динамическом поле ввода;
3) редактирование значений отступов;
4) назначение разных радиусов скругления ребрам изделия;

5) изменение значений радиусов скругления одним из двух способов: перемещением маркера радиуса или введением значения в динамическом поле ввода;
6) редактирование радиусов скругления ребер.

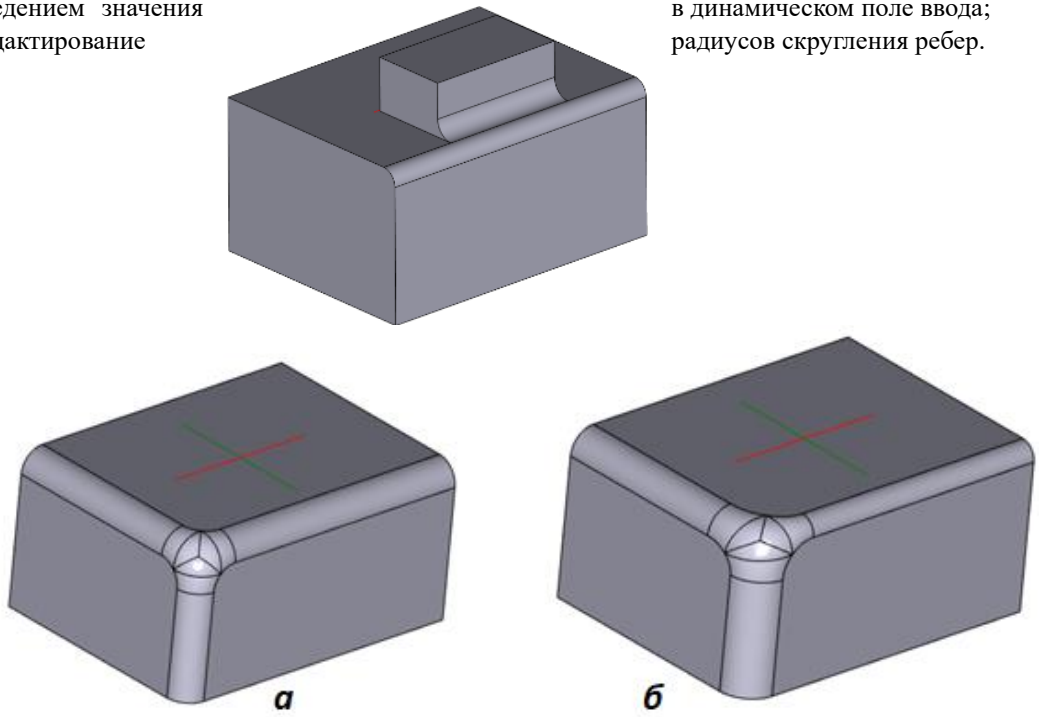


Рис. 9. Реализация функции «Возможность управлять формой скругления для вершин путём задания отступа от вершины до начала регулярного скругления по каждому направлению»: *а* – для одинаковых отступов и радиусов; *б* – для разных значений отступов и радиусов

Функция 8 (рис. 10). Управление вариантами перекрытия скругления - прокатывание по ребру, прокатывание по сопрягаемой поверхности. Функция должна обеспечивать:

1) прокатывание по ребру, когда требуется сохранить ребро неизменным и скорректировать поверхность сглаживания так, чтобы это требование соблюдалось;
2) прокатывание по сопрягаемой поверхности.

Рис. 10. Результат работы функции «Управление вариантами перекрытия скругления - прокатывание по ребру, прокатывание по сопрягаемой поверхности»

Функция 9 (рис. 11, 12) - Управление опциями перекрытия скруглений - взаимная обрезка или сопряжение скруглений с наращиванием материала. Функция должна обеспечивать:

1) обработку ситуаций при перекрытии скругления одной выпуклости (формирование не гладкого перехода между гранями скругления, а грани скругления должны обрезаться друг по другу);
2) обработку ситуаций при перекрытии скругления разной выпуклости.

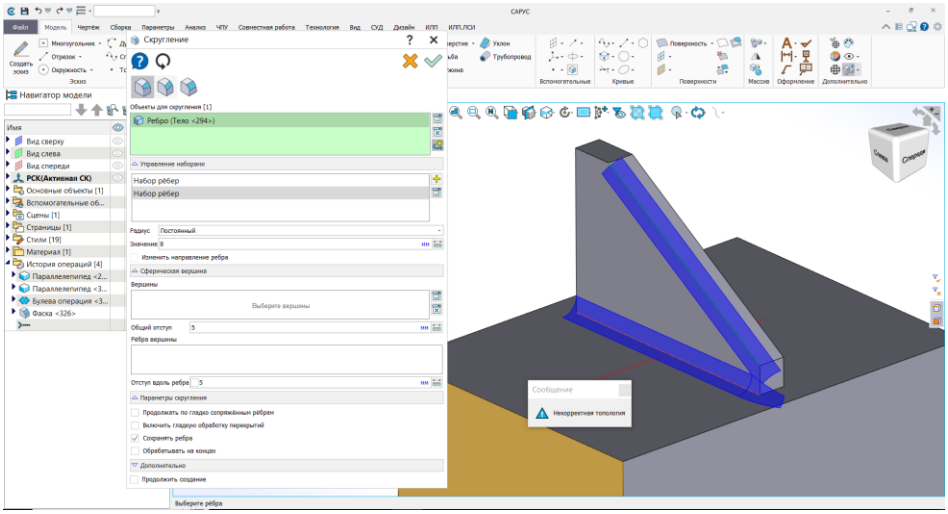


Рис. 11. Результат работы функции «Управление опциями перекрытия скруглений - взаимная обрезка или сопряжение скруглений с наращиванием материала», п.2)

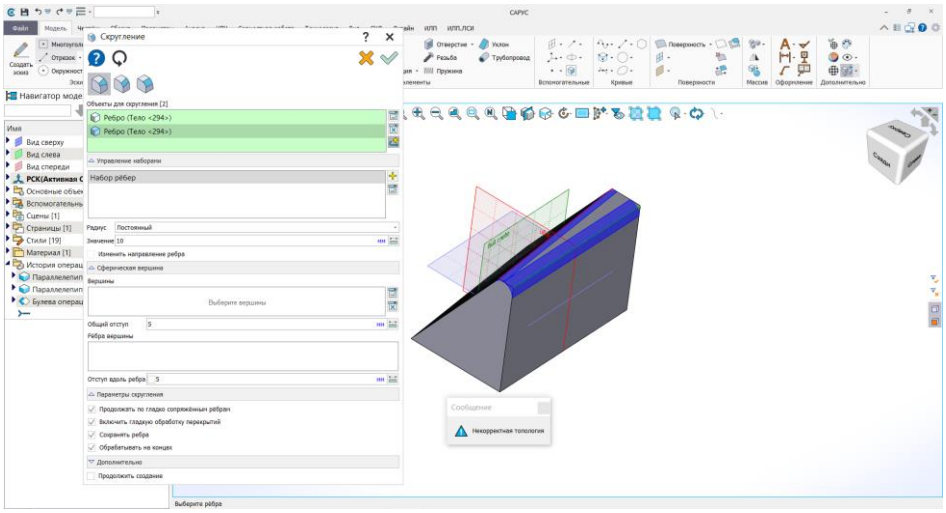


Рис. 12. Результат работы функции «Управление опциями перекрытия скруглений - взаимная обрезка или сопряжение скруглений с наращиванием материала», п.2)

Данный метод реализовать в СПЖЦ «САРУС» не удалось, программа выдает ошибку «Нарушение топологии». Данная ошибка устраняется лишь уменьшением значения радиуса скругления до момента, когда 2 ребра скругления не пересекаются друг с другом.

Вывод. Таким образом на простейшей фигуре «Параллелепипед» удалось построить все алгоритмы, за исключением «Управление опциями перекрытия скруглений - взаимная обрезка или сопряжение скруглений с наращиванием материала». Можно заключить, что построение скруглений на поверхностях деталей простой формы (I, II групп конструктивной сложности) в основном возможно.

IV. 2 Анализ работы алгоритмов функции «Скругление» на деталях с усложнением элементов формы в СПЖЦ «САРУС»

Для демонстрации работы алгоритмов сначала была взята композиция примитивов со сложными переходами, а именно:

- Скругление на границе сфера-параллелепипед
- Внутренние и внешние скругления с пересечением в вершине
- Скругление на криволинейной поверхности
- Пересечение скруглений
- Скругление с пересечением отверстия

Результаты работы функции «Скругление» с заданными выше параметрами показаны на тестовой детали, приведенной на рис. 13.

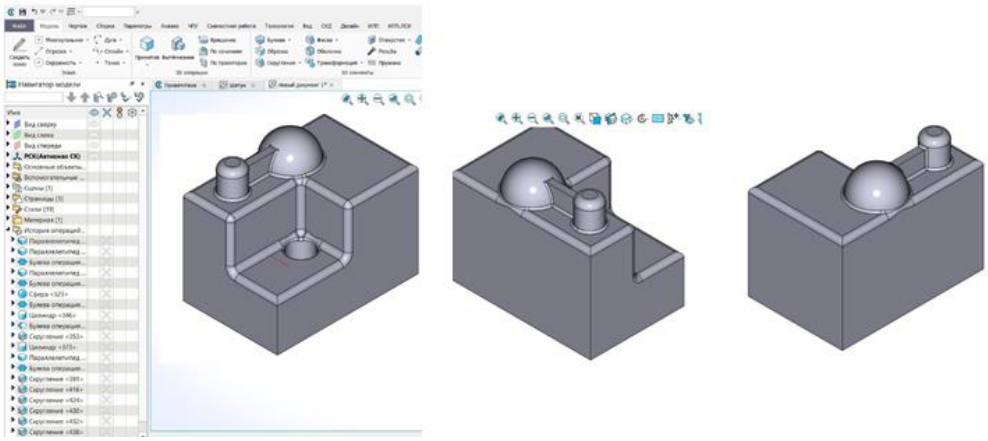


Рис. 13. Тестовая деталь в трех видах для демонстрации функции «Скругление»

На рис. 13 видно, что в данном построении не возникло проблем с пересечениями скруглений, СПЖЦ «САРУС» корректно построил все скругления ребер.

Чтобы продемонстрировать работу скруглений с переменным радиусом, была построена деталь «Шатун» (рис. 14).

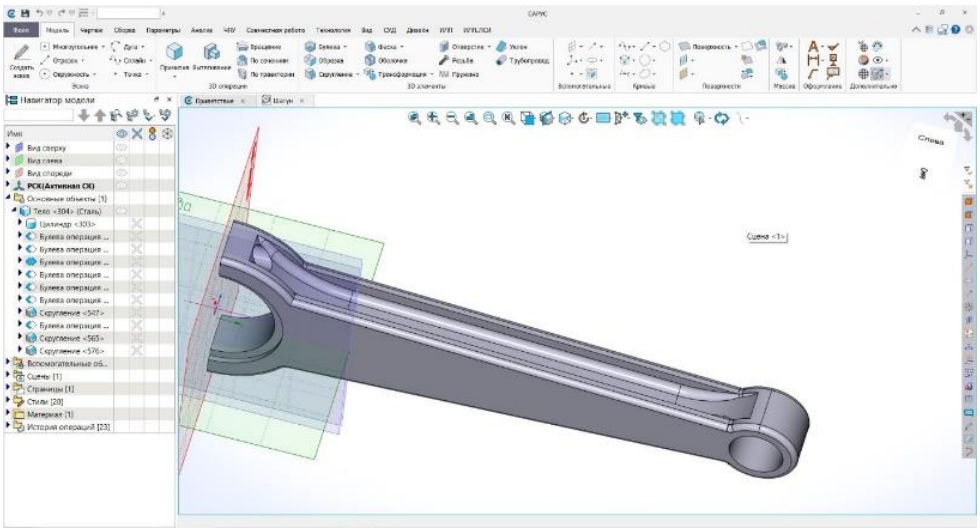


Рис. 14. 3D-модель детали «Шатун»

На концах ребер скругления был использован алгоритм переменного радиуса по точкам, представленным в п. 2, функция 5. На данной детали алгоритм работал верно.

Выводы: таким образом, в деталях с усложнением элементов форм все алгоритмы функции «Скругление», представленные в п. 2, работают корректно.

V. Обсуждение результатов. Сравнительный анализ с КОМПАС-3D

Сравнительный анализ работы алгоритмов функции «Скругление» показал следующие отличия в работе алгоритмов в КОМПАС-3D относительно СПЖЦ «САРУС».

При построении скруглений переменного радиуса (рис. 15) на 3D-модели детали «Шатун» КОМПАС-3D корректно построил пересечения скруглений в отличии от «САРУС» (рис.14).

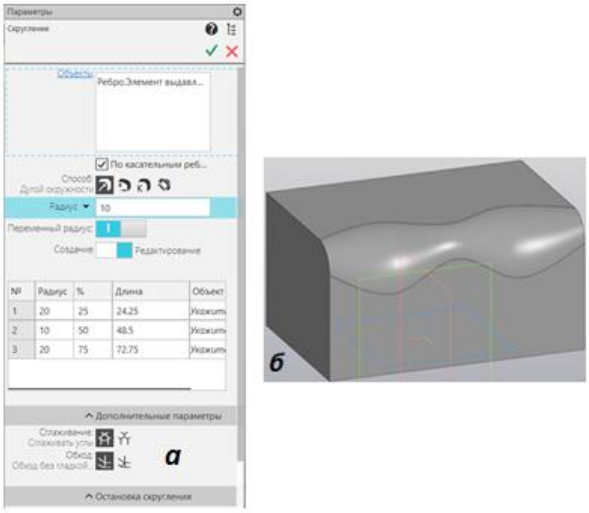


Рис. 15. Результат работы функции: *а* – окно редактирования функции; *б* – результат работы функции

Надо отметить, что в КОМПАС-3D есть дополнительный функционал, который отсутствовал в СПЖЦ «САРУС», а именно обход с гладкой стыковкой при пересечении 2-х ребер, остановка скругления и полное скругление (рис. 16).

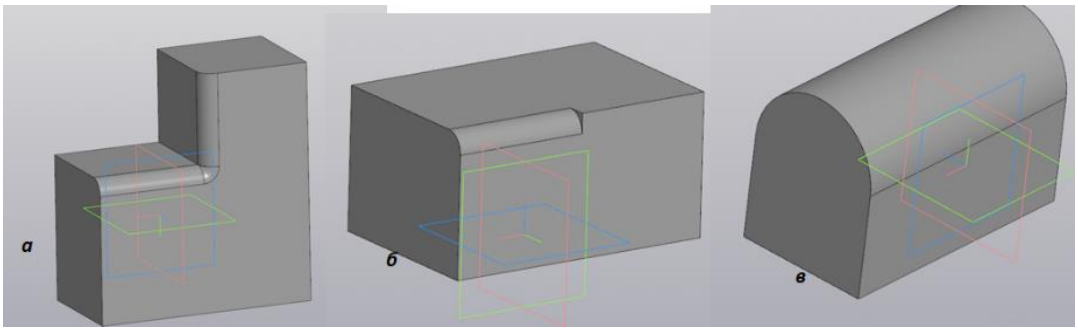


Рис. 16. Дополнительный функционал КОМПАС-3D: *а* - Скругление 2-х ребер с обходом с гладкой стыковкой; *б* - Остановка скругления в заданной точке; *в* - Полное скругление

В остальных алгоритмах функции «Скругление», представленных в п. 2, работают идентично СПЖЦ «САРУС».

Для сравнения в КОМПАС 3D была построена деталь «Шатун» (рис. 17).

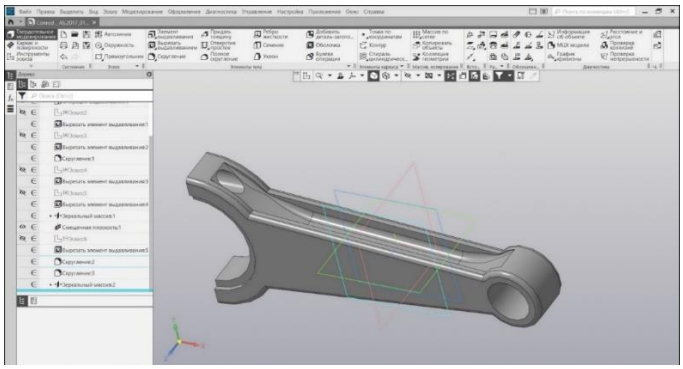


Рис. 17. Деталь «Шатун» в КОМПАС-3D

При построении детали «Шатун» в КОМПАС-3D возникли некоторые сложности при построении. Во-первых, здесь оказался менее удобным функционал выбора точки на ребре для построения скругления с переменным радиусом, так как нужная длина в мм или % неизвестна, а вручную точку найти проблематично. Также по чертежу скругление должно было зайти на отверстие и по его периметру обрезаться. В СПЖЦ «САРУС» такое скругление построилось, в то время как в КОМПАС-3D САПР выдавал ошибку, пока значение величины скругления не оказалось меньшим и не перестало пересекать отверстие.

Это привело к заключению, что деталь «Шатун» в СПЖЦ «САРУС» строится корректнее, чем в КОМПАС-3D.

VI.Выводы и заключение

Данная исследовательская работа была выполнена в связи с проблемой построения скруглений в более ранних версиях «САРУС». Необходимо было проверить работоспособность ряда функций «Скругление» в СПЖЦ «САРУС», получившего сертификат САПР базового уровня.

Все алгоритмы, кроме управления опциями перекрытия скруглений – взаимной обрезки или сопряжения скруглений с наращиванием материала, оказались работоспособными.

Также была продемонстрирована работа данных алгоритмов на деталях с усложненными элементами формы в СПЖЦ «САРУС». Здесь все алгоритмы работали исправно, даже в местах пересечений скруглений, чего, кстати, не удалось сделать на детали простейшей формы (I группы конструктивной сложности).

Был проведен сравнительный анализ с известными САПР КОМПАС-3D, зарекомендовавшим себя результативным в работе на базовом предприятии. Здесь у СПЖЦ «САРУС» были свои достоинства и недостатки. К первому относятся более корректное построение скругления с переменным радиусом, а также возможность построения скругления с перекрытием отверстия (скругление повторяет форму отверстия), что в КОМПАС-3D не получилось сделать. Из недостатков СПЖЦ «САРУС» по отношению к КОМПАС-3D можно отнести отсутствие некоторого функционала, который есть в КОМПАС-3D, а именно: полное скругление, остановка скругления в точке и обход с гладкой стыковкой на пересечении 2-х ребер. Также КОМПАС-3D корректно выполнил алгоритм управления опциями перекрытия скруглений – взаимной обрезки или сопряжения скруглений с наращиванием материала. Однако и в той, и в другой САПР есть факты несрабатывания функций.

Работать с функциями «Скругление» в СПЖЦ «САРУС» возможно на базовом уровне программного обеспечения CAD.

Источник финансирования. Благодарности

Работа выполнена в соответствии:

- с Рабочим учебным планом по направлению 15.04.05- Конструкторско-технологическое обеспечение машиностроительных производств;
- с Договором от 25 марта 2024 г. № 96-2024/244 на выполнение научно- исследовательской работы по теме: «Проведение научных исследований в интересах ФГУП «РФЯЦ-ВНИИЭФ» с использованием результатов для актуализации образовательных программ подготовки специалистов с высшим профессиональным образованием для ФГУП «РФЯЦ-ВНИИЭФ», Задача 10 «Исследование программно-аппаратных решений в области цифровых технологий, возможностей их отраслевого применения, разработка и версионное тестирование компонентов полностью защищенной системы полного жизненного цикла среднего класса СПЖЦ V3 «Сарус».

Научный руководитель Денисова Наталья Александровна, СарФТИ-НИЯУ МИФИ, кафедра Технологии специального машиностроения

Авторы благодарят:

- Кафедру цифровых технологий СарФТИ-НИЯУ МИФИ под руководством Кривошеева О.В., к.т.н., заведующий кафедрой, зам. директора ФГУП "РФЯЦ-ВНИИЭФ" по технологиям полного жизненного цикла - директор института цифровых технологий (ИТЦ) - главный конструктор СПЖЦ;

Список литературы

1. Комплекс программ в защищённом исполнении «Система полного жизненного цикла изделий «Цифровое предприятие». Основная версия программного модуля. «Система конструкторского проектирования» (версия 1)/ПЗ ТМ 07623615.00423-04 81 01-6 – 2021 г., 720 с.
2. Комплекс программ в защищённом исполнении «Система полного жизненного цикла изделий «Цифровое предприятие». Основная версия программного модуля. «Система конструкторского проектирования» (версия 2)/Руководство оператора 07623615.00423-06 34 01. – 2021 г., 338 с.
3. Голованов Н.Н. Геометрическое моделирование. Учебное пособие. – М.: КУРС: ИНФРА-М, 2016. 400 с.

УДК 004.925.8
DOI: 10.25206/978-5-8149-3873-2-2024-776-784

Оптимизация создания сборочных моделей в СПЖЦ «САРУС»
Optimization of the creation of assembly models in the SARUS Full Life Cycle System

А.М. Рыжов[1], Н.А.Денисова[1],
[1]Саровский физико-технический институт - филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский ядерный университет "МИФИ" (СарФТИ НИЯУ МИФИ), ул. Духова, д.6, г. Саров, 607186, РФ

A.M. Ryzhov[1], N.A. Denisova[1],
[1]Sarov Institute of Physics and Technology - branch of the Federal State Autonomous Educational Institution of Higher Education "National Research Nuclear University "MEPhI" (SarPhTI NRNU MEPhI), Dukhova str., 6, Sarov, 607186, Russian Federation

Аннотация. Одна из основных задач современного проектирования во многих технических областях — моделирование сборок. Сборочная модель — трехмерная модель, объединяющая модели деталей, подсборок и стандартных изделий – компонентов сборки, а также информацию о взаимном расположении и зависимостях между параметрами компонентов.

На современном производстве зачастую возникает необходимость работы со сборками, содержащими сотни, тысячи или даже десятки тысяч деталей. Современные системы геометрического моделирования позволяют работать с такими объемными сборками. Однако процесс создания сборочной модели достаточно долгосрочен и трудоемок.

Автор принимает участие в версионном функциональном тестировании вновь разрабатываемого комплекса программ «Система полного жизненного цикла изделий «Цифровое предприятие» («САРУС»). Цель исследования – оптимизация процесса 3D-сборки в программном модуле CAD. При анализе опыта работы в CAD было выявлено, что наиболее эффективным способом соединения элементов в сборке является сопряжение по плоскостям, осям и точкам. Также выявлено, что возможность использования данного способа в «САРУС» на данный момент отсутствует. Поэтому выдвинуто предложение по улучшению функциональной части данного ПО для оптимизации создания сборочных моделей.

Ключевые слова. САПР, Программный модуль CAD, сборочная 3D-модель, функции построения сборок.

Abstract. One of the main tasks of modern design in many technical fields is assembly modeling. An assembly model is a three dimensional model, that combines models of parts, subassemblies and standard assembly component products, as well as information about the relative location and dependencies between component parameters.

In modern manufacturing, it is often necessary to work with assemblies containing hundreds, thousands, or even tens of thousands of parts. Modern geometric modeling systems allow you to work with such volumetric assemblies. However, the process of creating an assembly model is quite long-term and time-consuming.

The author participates in the versioned functional testing of the newly developed software package "Digital Enterprise Product Full Life Cycle System" ("SARUS"). The purpose of the study is to optimize the 3D assembly process in the CAD software module. When analyzing the experience of working in CAD, it was revealed, that the most effective way to connect elements in an assembly is to interface along planes, axes and points. It has also been revealed, that the possibility of using this method in SARUS is currently absent. Therefore, a proposal has been put forward to improve the functional part of this software to optimize the creation of assembly models.

Keywords. CAD, CAD software module, 3D assembly model, assembly-building functions.

I. Введение

В условиях жесткой конкуренции на машино- и приборостроительных предприятиях активно разрабатываются и внедряются системы автоматизированного проектирования (CAD). Их целью является моделирование технических объектов и создание проектно-конструкторской документации в соответствии с требованиями стандартов и нормативных актов. Одним из важных элементов этой документации являются сборочные модели, необходимые для полного представления о конечном продукте.

Сборочная модель представляет собой трехмерное представление, которое объединяет в себе детали, подсборки и стандартные изделия, называемые компонентами сборки. Она также содержит информацию о взаимном положении компонентов и зависимостях между их параметрами.

Одна из основных задач современного проектирования во многих технических областях — создание сборочных моделей. Сборки создаются для проверки созданной конструкторской документации и наглядного представления внешнего вида будущего изделия. Кроме этого, сборочная модель изделия позволяет производить различные анализы конструкции — оценка размеров, расчет массово-инерционных характеристик, разборка изделия, анимация механизмов и др. На современном производстве достаточно часто возникает необходимость работы со сборками, содержащими большое количество деталей. На данный момент системы геометрического моделирования позволяют работать с объемными сборками. Однако, процесс создания сборочной модели достаточно долгосрочен и трудоемок.

В CAD-системах используются различные способы соединения компонентов в сборке. Каждый способ имеет свои плюсы и минусы. Поэтому, важной задачей является определение наиболее эффективного способа или комбинации способов соединения элементов при создании сборочной модели для оптимизации этого процесса. Исходя из этого, в системе проектирования должен быть реализован различный функционал, среди которого можно выбрать наиболее эффективный способ. Кроме этого, часть времени затрачивается на поиск и вызов необходимой для пользователя команды. Поэтому, интерфейс системы проектирования должен быть понятен, логичен и прост.

По мнению ведущих мировых аналитиков, ключевым фактором успеха в современном промышленном производстве является уменьшение времени выхода продукции на рынок. Существует несколько способов сократить этот срок, в том числе путем оптимизации. Оптимизация – это процесс улучшения чего-либо с целью достижения наивысших результатов. Конкретно, в данной работе будет рассмотрена оптимизация создания сборочных моделей, т.е. сокращение времени, затрачиваемого на проектирование сборок, путем выдвижения предложений по введению новых команд или доработки интерфейса в программном обеспечении (ПО) «САРУС» [1].

II. Постановка задачи

Цель текущей работы: выдвинуть предложения по доработке функционала и интерфейса ПО САРУС для оптимизации процесса создания сборочных моделей.

Основные задачи работы:

1) Выбрать сборочную модель, на примере процесса сборки которой будут выдвинуты предложения для его оптимизации;
2) Выбрать систему проектирования, сопоставляемую ПО САРУС;
3) Собрать сборочную модель в одной системе и произвести замер времени, затрачиваемого на процесс сборки;
4) Собрать сборочную модель в другой системе и произвести замер времени, затрачиваемого на процесс сборки;
5) Сравнить полученные результаты замеров времени, затрачиваемого на процесс создания сборочной модели в сравниваемых системах;
6) Выдвинуть предложения по оптимизации процесса создания сборочной модели в ПО САРУС.

III. Теоретическое описание создания 3D-сборок

Рассмотрим основные этапы создания сборок.

1. Добавление компонента из файла. Данное действие осуществляется с помощью соответствующей команды.

2. Создание компонента в процессе сборки. При составлении сборки возможно не только включать уже готовые компоненты, но и формировать новые непосредственно в рамках текущего файла сборки.

3. Задание положения элементов в сборке. После вставки компонента в сборку необходимо задать его положение и ориентацию. В CAD-системах существует несколько способов задания положения элементов в сборке: поворот компонента вокруг оси, сдвиг компонента в любом направлении, а также с использованием различных сопряжений. При работе со сборкой иногда требуется фиксировать определенный компонент, чтобы исключить его случайное перемещение в системе координат. Это позволит обеспечить предсказуемость перемещения других компонентов при наложении сопряжений. При добавлении первого компонента в новую сборку, он автоматически зафиксируется.

4. Создание массивов компонентов. Если в сборке присутствуют одинаковые компоненты (детали или подсборки), то они могут быть определенным образом упорядочены с помощью использования различного рода массивов. Наиболее распространенными являются следующие массивы:

- «Массив по образцу» позволяет создать массив компонентов, расположив их так же, как в уже существующем массиве (образце);
- «Массив по концентрической сетке» позволяет создать массив компонентов сборки, расположив их в узлах концентрической сетки;
- «Массив вдоль кривой» позволяет создать массив компонентов сборки, расположив их вдоль указанной кривой;

- «Массив по сетке» позволяет создать массив компонентов сборки, расположив их в узлах параллелограммной сетки;
- «Зеркальный массив» позволяет создать копию выбранных элементов, симметричную любой выбранной плоскости или грани.

5. Внесение сопряжений в сборку. Сопряжение — это параметрическая связь между выбранными компонентами сборки, которая формируется за счет задания взаимного положения между этими компонентами. В сопряжениях могут участвовать тела, грани, ребра, вершины.

IV. Результаты эксперимента. Построение 3D-модели сборки в СПЖЦ “САРУС”

IV.1. Сборочная модель

Для выполнения данной работы была выбрана сборочная модель «Нагнетатель», (рис. 1). Различные сечения сборочной модели «Нагнетатель» представлены на рис. 2.

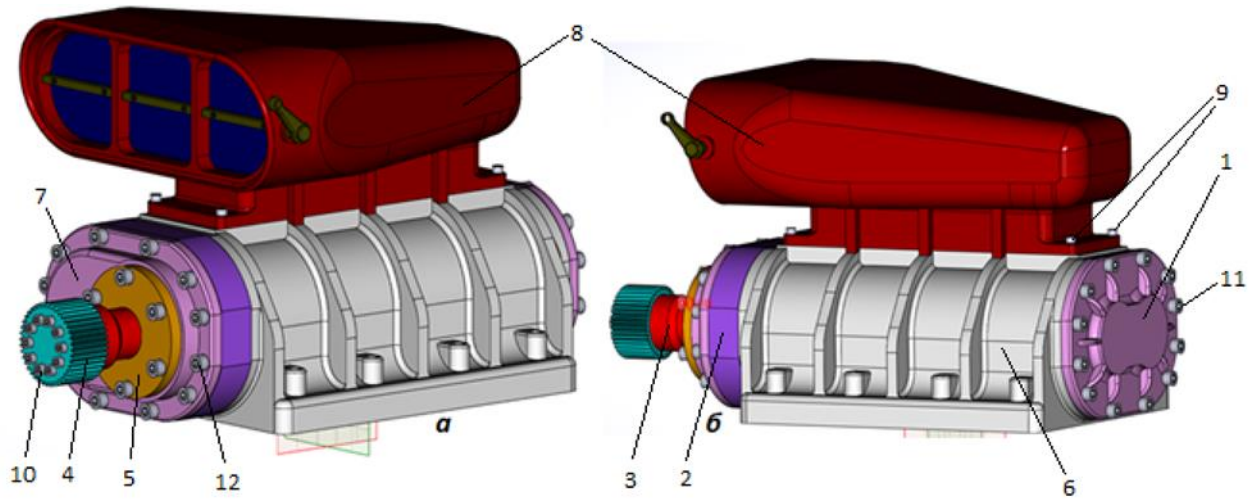


Рис. 1. Сборочная модель «Нагнетатель»

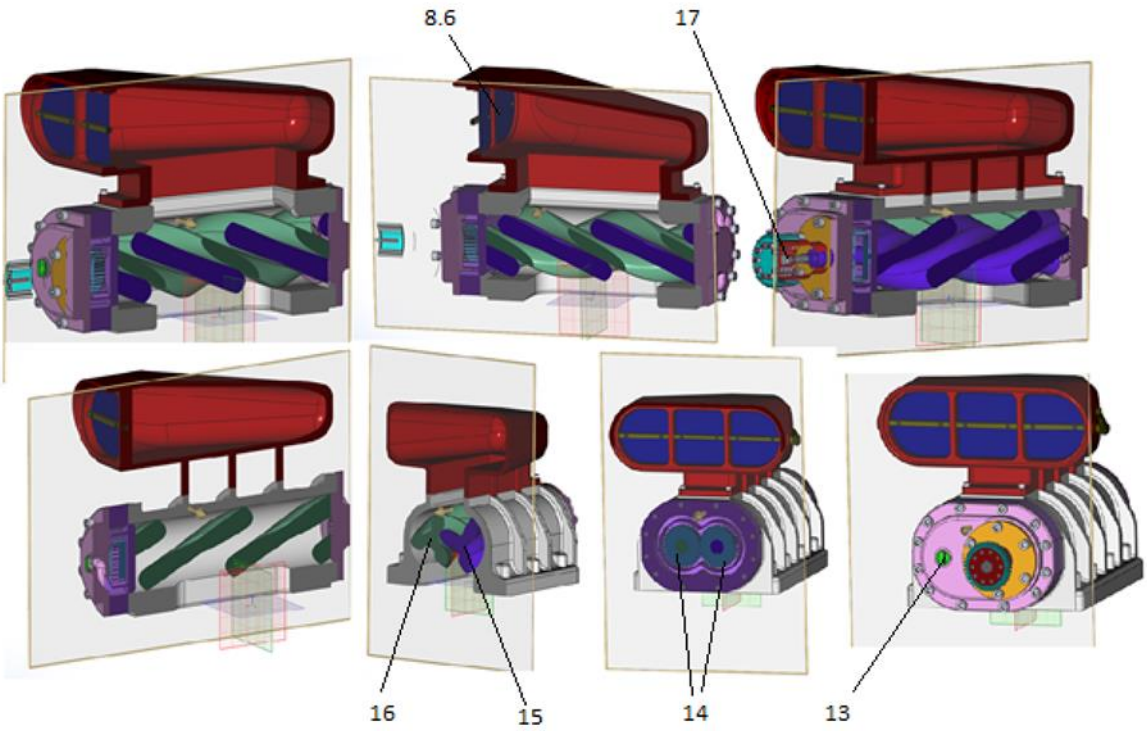


Рисунок 2 – Различные сечения сборочной модели «Нагнетатель»

Сборочная модель «Нагнетатель» состоит из следующих компонентов:
1 - «Крышка задняя» x1 (рис. 1);
2 - «Крышка средняя» x1 (рис. 1);

3 - «Деталь сцепления» x1 (рис. 1);
4 - «Шестерня» x1 (рис. 1);
5 - «Кольцо» x1 (рис. 1);
6 - «Корпус» x1 (рис. 1);
7 - «Крышка передняя» x1 (рис. 1);
8 - «Воздуховод» (рис. 1);
9 - «Винт M5-6gx20 ГОСТ 11738-84» x4 (рис. 1);
10 - «Винт M5-6gx30 ГОСТ 11738-84» x12 (рис. 1);
11 - «Винт M8-6gx20 ГОСТ 11738-84» x18 (рис. 1);
12 - «Винт M8-6gx55 ГОСТ 11738-84» x12 (рис. 1);
13 - «Заглушка» x1 (рис. 2);
14 - «Колесо зубчатое» x2 (рис. 2);
15 - «Ротор» x1 (рис. 2);
16 - «Ротор» x1 (рис. 2);
17 - «Болт M10-6gx35 ГОСТ 7796-70» x1 (рис. 3).

Сборочная единица «Воздуховод» представлена на рис. 3.

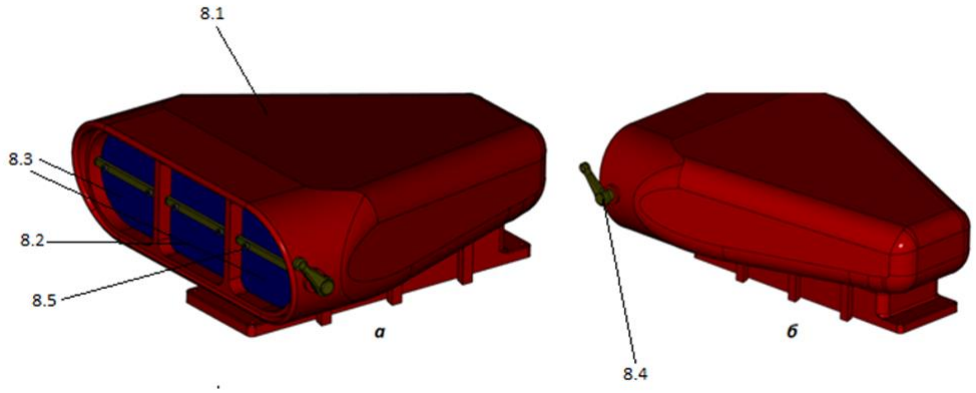


Рисунок 3 – Сборочная единица «Воздуховод»

Сборочная единица «Воздуховод» (8) состоит из следующих компонентов:
8.1 - «Воздухозаборник» x1 (рис. 3);
8.2 - «Закрылка центральная» x1 (рис. 3);
8.3 - «Закрылка» x2 (рис. 3);
8.4 - «Стержень» x1 (рис. 3);
8.5 - «Винт M3-6gx12 ГОСТ 11738-84» x6 (рис. 3);
8.6 - «Гайка M3 ГОСТ 15526-70» x6 (рис. 2).

IV.2. Сравниваемые системы проектирования

САРУС – технологически независимая система управления полным жизненным циклом изделий для промышленных предприятий. Система представляет собой программное обеспечение (ПО) PLM-класса (Product Lifecycle Management, управление жизненным циклом изделий) входит в число основных классов индустриального ПО, разрабатываемого в соответствии с Дорожной картой развития высокотехнологичных областей «Новые производственные технологии», утвержденной Правительством РФ в 2021 году [2].

Система САРУС разработана предприятием Госкорпорации «Росатом» РФЯЦ-ВНИИЭФ. Брендовое название цифрового продукта созвучно месту его создания – городу Сарову Нижегородской области [2].

Система САРУС создавалась с 2018 года в интересах атомной отрасли и оборонно-промышленного комплекса (ОПК) как моноплатформенное решение, полностью независимое от зарубежных технологий и элементов. Подтверждена её кроссплатформенность – продукт может работать как на операционной системе Windows, так и на российской операционной системе Astra Linux [2].

В числе преимуществ также стоит отметить применение суверенного российского геометрического ядра RGK (Russian Geometric Kernel) для работы с 3D-объектами. Полная импортонезависимость и защищенное исполнение являются ключевыми преимуществами САРУС сегодня [2].

Система САРУС может применяться на предприятиях, использующих различные типы производства. Осуществленная апробация САРУС на предприятиях Росатома позволяет уверенно предлагать её заказчикам из других отраслей машиностроения в качестве цифрового продукта, имеющего опыт практического применения в условиях высоких требований атомной индустрии [2].

Как было сказано выше, САРУС находится на стадии разработки и местом его создания является Саров. Кроме этого, Саровский Физико-Технический Институт выступает в качестве тестировщика данного ПО. Поэтому, оптимизация создания сборочных моделей путем выдвижения предложений по введению новых команд или доработки интерфейса будет производиться в программном обеспечении (ПО) САРУС в версии 1.0.23.80.

Для объективного выдвижения предложений по введению новых команд или доработки интерфейса, рационально будет сравнить САРУС с ПО, которое уже представлено на рынке. На основе этого сравнения и будут выдвинуты предложения для оптимизации создания сборок. Студенты в СарФТИ НИЯУ МИФИ, среди множества CAD-систем, наиболее подробно изучают российскую систему трехмерного проектирования КОМПАС-3D, которая была выпущена на рынок в 1989 г. Поэтому САРУС будет сравниваться с КОМПАС-3D в версии V20.

КОМПАС-3D разработан российской компанией «Аскон» и предназначен для проектирования изделий основного и вспомогательного производств в таких отраслях промышленности, как машиностроение (транспортное, сельскохозяйственное, энергетическое, нефтегазовое, химическое и т.д.), приборостроение, авиастроение, судостроение, станкостроение, вагоностроение, металлургия, промышленное и гражданское строительство, товары народного потребления и т. д. В основе КОМПАС-3D лежит российское геометрическое ядро C3D (создано C3D Labs, дочерней компанией АСКОН) и собственные программные технологии. Но для установки системы применяется автономный установщик Microsoft.NET Framework 4.8, являющийся разработкой американской компании Microsoft. Поэтому КОМПАС-3D считается отечественной разработкой, но с применением зарубежных составляющих [3].

IV.3. Создание сборочной модели в КОМПАС-3D

Процесс создания сборки в КОМПАС-3D включает в себя следующие общие шаги [4]:

1) Создание сборки начинается с изучения конструкторской документации (КД), по которой будет создаваться сборка.

2) Создание файла сборки в ПО: «Файл» → «Создать» → «Сборка».

После создания файла становится доступной инструментальная панель «Сборка», предназначенная для редактирования сборки.

3) Добавление первой детали с помощью команды «Вставить компонент из файла» в поле «Компоненты». Рекомендуется, чтобы первой деталью в сборке была та, на которую базируются больше всего остальных деталей. Например, в сборке «Нагнетатель» это деталь «Корпус».

4) Задание пространственного положения детали. После чего подтверждается выполнение команды «Вставить компонент из файла».

5) Добавление следующей детали в файл аналогично пунктам 3,4.

6) Задание взаимного положения добавляемой детали с базовой деталью сборки с помощью различного рода сопряжений, которые доступны в Компас-3D [5]. К ним относятся: совпадение, соосность, параллельность, перпендикулярность, расположение элементов на заданном расстоянии, расположение элементов под заданным углом, касание, симметрия и зависимое положение.

7) Повторение пунктов 5 и 6 для всех оставшихся деталей.

В сборочной модели присутствует большое количество стандартных изделий. Задание определенного положения каждого из них займет большое количество времени. Поскольку, большинство из стандартных изделий в сборке «Нагнетатель» расположены по периметру окружностей, то для ускорения процесса задания их положения можно воспользоваться массивом, а именно «Массив по концентрической сетке». Для этого достаточно вручную задать положение одного из элементов, а положение остальных с помощью массива. Эта функция заметно упрощает и ускоряет процесс создания сборочной модели. Кроме этого, для задания положения болтов крепления сборочной единицы «Воздуховод» к детали «Корпус» логично воспользоваться командой «Зеркальный массив», что также ускорит процесс создания сборочной модели.

Время, за которое удалось создать сборочную модель «Нагнетатель» в результате использования вышеописанных шагов и применения массивов при данной конфигурации ПО Компас-3D представлено в Табл. 1. Замер времени производился 3 раза.

ТАБЛИЦА 1
ЗАМЕР ВРЕМЕНИ СОЗДАНИЯ 3D-МОДЕЛИ В КОМПАС-3D

Замер времени / Элемент	1	2	3	Среднее время
«Воздуховод»	8,53 мин.	8,28 мин.	8,46 мин.	8,42 мин.
«Нагнетатель»	26,24 мин.	25,57 мин.	25,33 мин.	25,71 мин.
Суммарное время	34,77 мин.	33,85 мин.	33,79 мин.	34,13 мин.

В итоге среднее время создания сборочной модели «Нагнетатель» в ПО Компас-3D равно 34,13 минутам.

IV.4. Создание сборочной модели в «САРУС»

В целом, процесс создания сборки в ПО САРУС схож с процессом создания сборки в ПО Компас-3D, но, тем не менее, имеет ряд отличий и включает в себя следующие общие шаги:

1) Создание сборки начинается с изучения конструкторской документации (КД), по которой будет создаваться сборка.

2) Создание файла сборки в ПО: «Файл» → «Новый документ» → «Сборочная модель».

После создания файла открывается инструментальная панель «Сборка», предназначенная для редактирования сборки.

3) Добавление первой детали, с помощью команды «Вставить компонент». После активации команды открывается диалоговое окно. В поле «Файл», с помощью пиктограммы «Выбрать файл» выбирается необходимый файл из памяти компьютера, содержащий деталь. Рекомендуется, чтобы первой деталью в сборке была та, на которую базируются больше всего остальных деталей. Например, в сборке «Нагнетатель» это деталь «Корпус».

4) Задание пространственного положения детали. После чего подтверждается выполнение команды «Вставить компонент».

5) Открыть в инструментальной панели вкладку «Модель». В графе «Вспомогательные» выбрать команду «Система координат по объекту». Теперь необходимо создать локальную систему координат (ЛСК) в том месте, где будет располагаться следующая деталь для задания ее пространственного положения (рис. 4).

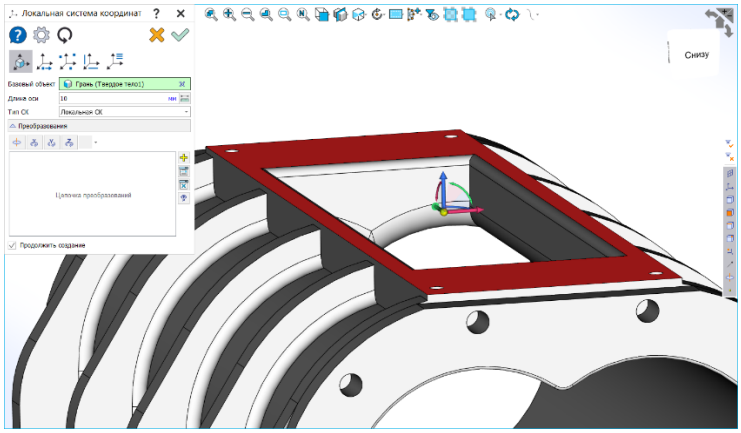


Рис. 4. Создание ЛСК

6) Добавление следующей детали в файл, аналогично пункту 3.

7) Задание пространственного положения добавленной детали производится следующим образом. Во вкладке «Привязка» рабочего окна команды «Вставить компонент» необходимо в качестве целевой ЛСК, в соответствующем поле, выбрать ЛСК, созданную в п.5 (рис. 5).

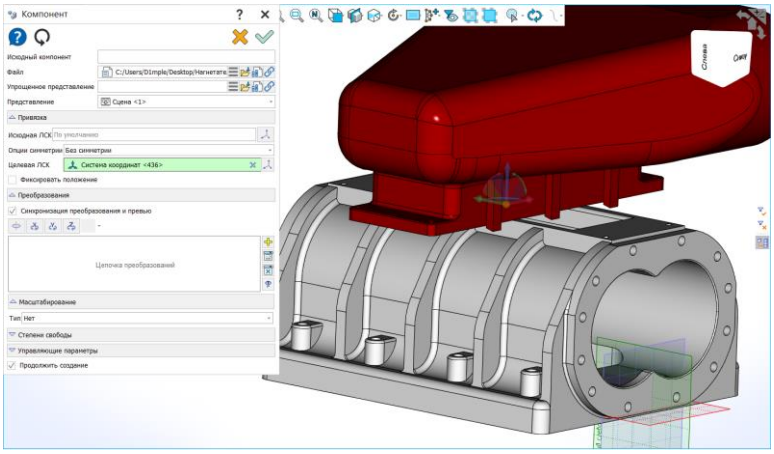


Рис. 5. Задание целевой ЛСК

После этого необходимо провести корректировку пространственного положения, если это необходимо. Данное действие выполняется во вкладке «Преобразования» рабочего окна команды «Вставить компонент». В

приведенном случае необходимо повернуть вставляемую деталь на 90° против часовой стрелки (рис. 6). После чего подтвердить выполняемую команду «Вставить компонент».

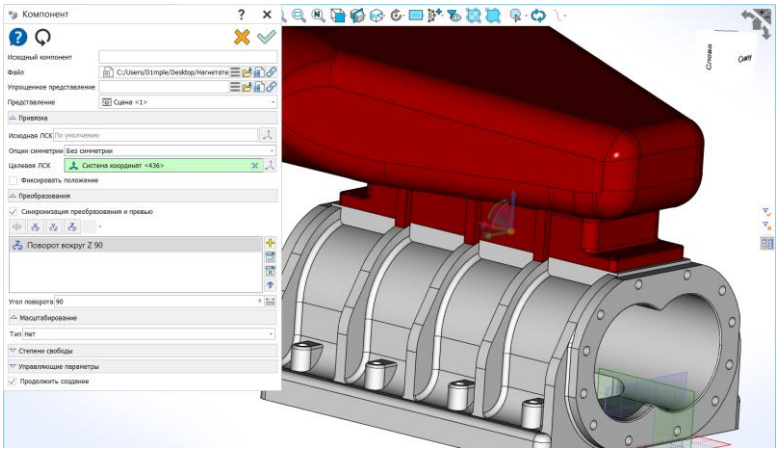


Рис. 6. Задание положения детали

8) Повторение пунктов 6 и 7 для всех оставшихся деталей.

В сборочной модели присутствует большое количество стандартных изделий. Задание определенного положения каждого из них займет большое количество времени. Поскольку, большинство из стандартных изделий в сборке «Нагнетатель» расположены по периметру окружностей, то для ускорения процесса задания их положения можно воспользоваться массивом, а именно «Круговой». Для этого достаточно задать положение одного из элементов, аналогично п. 6 и 7, а положение остальных с помощью массива. Эта функция заметно упрощает и ускоряет процесс создания сборочной модели. Кроме этого, для задания положения болтов крепления сборочной единицы «Воздуховод» к детали «Корпус» логично воспользоваться массивом «Симметрия», что также ускорит процесс создания сборочной модели.

Время, за которое удалось создать сборочную модель «Нагнетатель» в результате использования вышеописанных шагов и применения массивов при данной конфигурации ПО САРУС представлено в Табл. 2. Замер времени производился 3 раза.

ТАБЛИЦА 2
ЗАМЕР ВРЕМЕНИ СОЗДАНИЯ 3D-МОДЕЛИ В САРУС

Замер времени / Элемент	1	2	3	Среднее время
«Воздуховод»	10,13 мин.	9,32 мин.	9,49 мин.	9,64 мин.
«Нагнетатель»	30,86 мин.	30,43 мин.	29,91 мин.	30,4 мин.
Суммарное время	40,99 мин.	39,75 мин.	39,4 мин.	40,04 мин.

В итоге среднее время создания сборочной модели «Нагнетатель» в ПО «САРУС» равно 40,04 минутам.

V. Обсуждение результатов. Сравнение процесса создания сборочных моделей в КОМПАС-3D и САРУС

Для наглядного представления полученных результатов, сведем их в Табл. 3 в виде разницы времени, затрачиваемого на создание сборочной модели «Нагнетатель» в ПО САРУС и КОМПАС-3D:

ТАБЛИЦА 3
РАЗНИЦА ВРЕМЕНИ, ЗАТРАЧИВАЕМОГО НА СОЗДАНИЕ СБОРОЧНОЙ МОДЕЛИ «НАГНЕТАТЕЛЬ» В ПО «САРУС» И КОМПАС-3D

Замер времени / Элемент	1	2	3	Среднее время
«Воздуховод»	1,60 мин.	1,04 мин.	1,03 мин.	1,22 мин.
«Нагнетатель»	4,62 мин.	4,86 мин.	4,58 мин.	4,69 мин.
Суммарное время	6,62 мин.	5,90 мин.	5,61 мин.	5,91 мин.

Отсюда следует, что процесс создания сборки в САРУС затрачивает больше времени на 5,91 минуты, чем в КОМПАС-3D, то есть на 17,3%.

Полученный результат вызван, в первую очередь, отсутствием команд для обеспечения различного рода сопряжений между составляющими сборку деталями. Поэтому перед добавлением каждой детали в сборку приходится создавать дополнительную ЛСК и только после этого добавлять деталь, с указанием конечной ЛСК. Это несет в себе весомые затраты времени, поскольку приходится либо на этапе создания сборочной модели корректировать положение деталей из-за несовпадения необходимого и фактического положений деталей в сборке после совмещения ЛСК самой детали с конечной ее ЛСК в сборке, либо заранее продумывать расположение ЛСК самих деталей и их конечные ЛСК в сборке.

Кроме этого, часть времени затрачивается на поиск или вызов необходимых команд. Примером служат массивы. Большая часть команд, необходимых для создания сборки находятся на инструментальной панели во вкладке «Сборка», но массивы расположены во вкладке «Модель». Поэтому для вызова массивов приходится переходить из вкладки «Сборка» во вкладку «Модель», вызывать необходимый вид массива, и после завершения работы с массивами возвращаться обратно во вкладку «Сборка» для продолжения дальнейшей работы.

В ПО САРУС команда «Вставить компонент» объединяет вставку компонента из файла, библиотеки и исходного компонента. Поэтому в диалоговом окне команды приходится выбирать необходимый источник файла и, соответственно, вызывать его. На это, также, затрачивается часть времени за счет излишних передвижений курсора и процесса выбора.

При необходимости добавить в модель несколько одинаковых деталей в САРУС, за один сеанс команды «Вставить компонент» возможно добавить только одну деталь. Для добавления остальных таких же деталей необходимо снова использовать команду «Вставить компонент», но в этот раз в рабочем окне в поле «Исходный компонент» необходимо выбрать нужную деталь и только после этого возможно ее скопировать. На это действие также затрачивается некоторое количество времени.

В КОМПАС-3D удается сэкономить часть времени за счет возможности быстрого вызова некоторых команд, например, «Вставить компонент», с помощью нажатия правой клавиши мыши и открытия дополнительной миниатюрной панели инструментов, тем самым минимизировать передвижение курсора.

VI.Выводы и заключение

Исходя из представленных выше особенностей процесса создания сборочных моделей на примере сборки «Нагнетатель», можно выдвинуть следующие предложения для оптимизации процесса моделирования сборочных моделей:

1) Добавление команд, обеспечивающих различного рода сопряжения между составляющими сборку деталями;

2) Добавление команд для вызова различного рода массивов во вкладку «Сборка»;

3) Команду «Вставить компонент», объединяющую в себе вставку компонента из файла, библиотеки и исходного компонента, стоит разделить на отдельные команды, такие как «Вставить компонент из файла», «Вставить компонент из библиотеки», «Вставить исходный компонент» или «Копировать объект».

4) Реализовать возможность вставки необходимого числа одинаковых деталей в сборочную модель за один сеанс работы команды «Вставить компонент», аналогично КОМПАС-3D;

5) Реализовать возможность быстрого вызова наиболее часто используемых команд, например, «Вставить компонент» или, в дальнейшем, различных сопряжений, путем нажатия правой клавиши мыши и открытия дополнительной миниатюрной панели инструментов, аналогично КОМПАС-3D.

Таким образом, в ходе исследования было выявлено, что на создание сборочных моделей в ПО САРУС затрачивается больше времени, чем в системе КОМПАС-3D. Замер времени создания сборочной модели производился на примере сборки «Нагнетатель». В результате было получено, что на создание сборки «Нагнетатель», в среднем, в ПО КОМПАС-3D затрачивается 34,13 минут. На создание той же сборки в ПО САРУС затрачивается 40,04 минут, что на 5,91 минут больше, чем в ПО КОМПАС-3D, что составляет 17,3%. Исходя из этого, были выдвинуты предложения по оптимизации процесса создания сборочных моделей в ПО САРУС, связанные с доработкой функционала и интерфейса системы. Поставленные задачи были выполнены.

Источник финансирования. Благодарности

Работа выполнена в соответствии:

- с Рабочим учебным планом по направлению 15.04.05- Конструкторско-технологическое обеспечение машиностроительных производств;

- с Договором от 25 марта 2024 г. № 96-2024/244 на выполнение научно- исследовательской работы по теме: «Проведение научных исследований в интересах ФГУП «РФЯЦ-ВНИИЭФ» с использованием результатов для актуализации образовательных программ подготовки специалистов с высшим профессиональным образованием для ФГУП «РФЯЦ-ВНИИЭФ», Задача 10 «Исследование программно-аппаратных решений в области цифровых технологий, возможностей их отраслевого применения, разработка и версионное тестирование компонентов полностью защищенной системы полного жизненного цикла среднего класса СПЖЦ V3 «Сарус».

Научный руководитель Денисова Наталья Александровна, СарФТИ-НИЯУ МИФИ, кафедра Технологии специального машиностроения

Авторы благодарят:
- Кафедру цифровых технологий СарФТИ-НИЯУ МИФИ под руководством Кривошеева О.В., к.т.н., заведующий кафедрой, зам. директора ФГУП «РФЯЦ-ВНИИЭФ» по технологиям полного жизненного цикла - директор института цифровых технологий (ИТЦ) - главный конструктор СПЖЦ;

Список литературы

1. Комплекс программ в защищенном исполнении «Система полного жизненного цикла изделий «Цифровое предприятие»; Программный модуль «Система конструкторского проектирования». ПЗ ТП 07623615.00423-04 81 01-1 часть 1. 2021г. 889 с.
2. PLM-система Росатома «САРУС». Сайт «atommedia.online». [Электронный ресурс]. Режим доступа:URL: https://atommedia.online/reference/plm-sistema-rosatoma-sarus/.
3. КОМПАС-3D. Система трехмерного моделирования. Сайт «ascon.ru». [Электронный ресурс]. Режим доступа:URL: https://ascon.ru/products/kompas-3d/
4. СОЗДАНИЕ 3D-МОДЕЛЕЙ СБОРОК В РАЗНЫХ CAD-СИСТЕМАХ. Сайт «izron.ru». [Электронный ресурс]. Режим доступа: URL: https://izron.ru/articles/aktualnye-voprosy-tekhnicheskikh-nauk-v-sovremennykh-usloviyakh-sbornik-nauchnykh-trudov-po-itogam-m/sektsiya-1-inzhenernaya-grafika-sapr-cad-cae-spetsialnost-05-01-01/sozdanie-3d-modeley-sborok-v-raznykh-cad-sistemakh/
5. Голованов Н.Н. Геометрическое моделирование. Учебное пособие. – М.: КУРС: ИНФРА-М, 2016. 400 с.

УДК 087.2
DOI: 10.25206/978-5-8149-3873-2-2024-785-794

Подготовка кадров для дальнейшего развития геометрического ядра САПР «САРУС»
Training of personnel for the further development of the geometric core of the SARUS CAD system

Н.А.Денисова[1], Г.А. Федоренко[1], С.Н. Ганчук[1,2], О.В. Кривошеев[1,2], С.В. Маврин[1,2], С.А. Рыжов[1,2]
[1] Саровский физико-технический институт - филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский ядерный университет "МИФИ" (СарФТИ НИЯУ МИФИ), ул. Духова, д.6, г. Саров, 607186, РФ
[2] Российский федеральный ядерный центр - Всероссийский научно-исследовательский институт экспериментальной физики (РФЯЦ-ВНИИЭФ) - федеральное государственное унитарное предприятие Государственной корпорации по атомной энергии "Росатом", *пр. Мира, 37, г. Саров, 607188, РФ*

N.A.Denisova[1], G.A. Fedorenko[1], S.N. Ganchuk[1,2], O.V. Krivosheev[1,2], S.V. Mavrin[1,2], S.A. Ryzhov[1,2]
[1] Sarov Institute of Physics and Technology - branch of the Federal State Autonomous Educational Institution of Higher Education "National Research Nuclear University "MEPhI" (SarPhTI NRNU MEPhI), Dukhova str., 6, Sarov, 607186, Russian Federation
[2] Russian Federal Nuclear Center - All-Russian Scientific Research Institute of Experimental Physics (RFNC-VNIIEPh) - Federal State Unitary Enterprise of the State Atomic Energy Corporation "Rosatom", 37 Mira Ave., Sarov, 607188, Russian Federation

Аннотация. В настоящее время разработка и производство изделий машиностроения (как и любых других) невозможна без применения САПР, при этом постоянно возрастающая сложность изделий требует, в свою очередь, непрерывного развития САПР и, в частности, геометрического ядра CAD-систем. Это требование в полной мере относится и к геометрическому ядру программного модуля CAD, входящему в комплекс программ «Система полного жизненного цикла изделий «Цифровое предприятие», разработка которого выполнена во «РФЯЦ-ВНИИЭФ». Работа посвящена описанию процесса подготовки кадров для развития существующих и разработки новых математических методов геометрического ядра, приводится описание учебных дисциплин, примеры решаемых студентами задач, научные публикации, выполненные с активным участием студентов.

Ключевые слова: подготовка кадров**,** CAD-системы, геометрическое моделирование, непрерывность, кривые Безье, B-сплайны, NURBS.

Abstract. Currently, the development and production of mechanical engineering products (as well as any other) is impossible without the use of CAD, while the ever-increasing complexity of products requires, in turn, continuous development of CAD and, in particular, the geometric core of CAD systems. This requirement fully applies to the geometric core of the CAD software module, which is part of the software package "Complete Product Lifecycle System "Digital Enterprise", the development of which was carried out at "RFNC-VNIIEPh". The work is devoted to the description of the personnel training process for the development of existing and new mathematical methods of the geometric core, a description of academic disciplines, examples of tasks solved by students, scientific publications carried out with the active participation of students.

Keywords: personnel training, CAD systems, geometric modeling, continuity, Bezier curves, B-splines, NURBS

I. Введение

В современном мире разработка и производство изделий машиностроения, как и любых других изделий, невозможны без использования систем автоматизированного проектирования (САПР). Это обусловлено тем, что сложность изделий постоянно возрастает, и для обеспечения их качества и конкурентоспособности необходимо применять современные технологии проектирования и производства. В свою очередь, постоянно возрастающая сложность изделий требует непрерывного развития САПР и, в частности, геометрического ядра CAD-систем (CAD - *Computer-Aided Design*). Геометрическое ядро является основой CAD-системы, так как в нем выполняются математические преобразования для представления и моделирования трёхмерных и двухмерных объектов.

Необходимость постоянного развития геометрического ядра в полной мере относится и к геометрическому ядру программного модуля CAD, входящему в комплекс программ «Система полного жизненного цикла изделий «Цифровое предприятие» (коммерческое наименование – «Сарус») [1]. Разработка этого комплекса выполнена во ФГУП «РФЯЦ-ВНИИЭФ». Система полного жизненного цикла изделий «Цифровое предприятие»

представляет собой интегрированную платформу, которая позволяет автоматизировать все этапы жизненного цикла изделия: от проектирования и производства до эксплуатации и утилизации. Она включает в себя модули для управления проектами, инженерного анализа, управления данными об изделии и другие. Геометрическое ядро программного модуля CAD является ключевым компонентом системы «Цифровое предприятие», поскольку оно обеспечивает создание точных и реалистичных трёхмерных моделей изделий.

Геометрическое ядро является симбиозом геометрического моделирования и объектно-ориентированного программирования (ООП). Геометрическое моделирование включает в себя различные разделы математики – дифференциальную геометрию, аналитическую геометрию, численные методы, методы математического программирования, линейную алгебру и др. [2]. ООП – это методология и соответствующие языки программирования, предназначенные для разработки больших программных комплексов. ООП предполагает абстрагирование от конкретных деталей реализации и основывается на таких понятиях, как инкапсуляция, наследование и полиморфизм [3]. Из этого чрезвычайно краткого описания геометрического моделирования и ООП уже видно, что эти направления математики и программирования довольно сложны. Таким образом, первой проблемой, присущей геометрическому ядру CAD-систем, является его сложность.

Вторая проблема развития геометрического ядра – отсутствие выпускников ВУЗов по этой специальности (по крайней мере в РФ).

Третьей проблемой является отсутствие полноценных учебников и монографий, посвященных всему кругу вопросов, связанных с разработкой и развитием геометрических ядер (по крайне мере на русском языке). Конечно, здесь следует указать многократно переиздававшуюся монографию [2], но использовать эту монографию как основной учебник для разработчиков геометрического ядра CAD «Сарус», к сожалению, нецелесообразно по следующим причинам:

- во-первых, монография описывает математические методы и геометрические объекты, положенные в основу геометрического ядра C3D (CAD «Компас»), при этом в геометрическом ядре CAD «Сарус» основным геометрических объектом является NURBS (*Non-Uniform Rational B-Splines – Не Равномерные Рациональные B-Сплайны*) и его частные случаи B-сплайны и кривые Безье [4]. В последнем издании этой монографии [2] параграфы, посвященные NURBS отсутствуют, хотя в первых изданиях они были.

- во-вторых, монография скорее представляет собой краткое введение в геометрическое моделирование и не раскрывает каких-либо деталей, что является совершенно разумным. Так, геометрическое ядро C3D – это коммерческий продукт.

Наряду с [2] можно указать переведенные с английского монографии [5,6], но эти монографии отражают уровень развития геометрического моделирования 70 - х годов прошлого века. Классическая монография по NURBS [7] существует только на английском языке, что делает невозможным работу с ней каждым студентом (как показал опыт - Google Переводчик не помогает на начальном этапе освоения), и самый основной недостаток этой монографии как учебника – сложность изложения материала.

II. Постановка задачи

Исходя из острой потребности РФЯЦ-ВНИИЭФ в специалистах в области геометрического моделирования с учетом вышеизложенных проблем, кафедрой ЦТ (Цифровые Технологии) и ТСМ (Технология специального машиностроения) СарФТИ-НИЯУ МИФИ и ИЦТ (Институт Цифровых Технологий) РФЯЦ-ВНИИЭФ было принято совместное решение по подготовке специалистов в области сопровождения и развития геометрического ядра CAD «Сарус». Для этого наряду с общеобразовательными и специальными курсами, представленными в разделе IV, ведется индивидуальная научная работа с некоторыми перспективными студентами, представленная в разделе III.

III. Результаты экспериментов. Решенные задачи и научные публикации, выполненные при активном участии студентов

Для более полноценного формирования высоких профессиональных качеств студентов, обучающихся по направлению «Информационные системы и технологии» (выпускающая кафедра ЦТ), студенты привлекаются к решению конкретных практически значимых задач, которые входят в рабочие планы ИЦТ РФЯЦ-ВНИИЭФ. Особенность подготовки этих студентов состоит в том, что на производственной практике после 3-го курса бакалавриата (длительность практики ~ 1 месяц) к студенту прикрепляется научный руководитель, с которым студент будет работать до окончания магистратуры. На этой практике все студенты выполняют практически одну и туже работу – программируют на языке C++ построение кривых Безье, B-сплайнов, NURBS, рассчитывают производные и строят графики этих объектов. Начиная с первого семестра 4-го курса (в рамках курсового проекта) начинается индивидуальна научная работа. Тема работы выбирается таким образом, что начальная часть работы будет представлена в дипломной работе бакалавра, а дальнейшее развитее – в магистерской диссертации. На этом этапе бакалавриата студенты знакомятся с архитектурой программного комплекса САПР «Сарус» и имплементируют решение своей задачи в рабочий код комплекса САПР «Сарус». При продолжении учебы в магистратуре студенты трудоустраиваются в ИЦТ на 0.5 ставки техника, при этом продолжают научную работу с научным руководителем и дополнительно решают производственные задачи, как программисты (язык C++).

Для выработки навыков самостоятельной научной работы студентам представляются математические описания алгоритмов (практически всегда на английском языке), студенты самостоятельно изучают алгоритм, при необходимости консультируются с научным руководителем, и полностью самостоятельно реализуют его в виде программного кода. Так же выработки самостоятельности студентам может быть представлен руководителем частный случай алгоритма, например сопряжение 3-х кривых 2-й степени, а студент реализует общий случай – произвольное количество кривых любой степени.

Актуальность студенческих работ, приведенных ниже, обусловлена практической необходимостью создания и преобразования кривых и поверхностей высокого порядка гладкости (вплоть до произвольно заданного порядка гладкости), так как в настоящее время САПР «Сарус» может создавать линии и поверхности и выполнять их сопряжение до 2-го порядка гладкости.

Для обеспечения высокого научного уровня ИЦТ и созданию в будущем научной школы к студентам предъявляется обязательное требование по выпуску научных публикаций. В настоящем разделе приводятся аннотации выполненных работ и некоторые иллюстративные рисунки.

III.1. Аппроксимация сопряжения кривых Безье с сохранением порядка гладкости и дополнительными ограничениями [8]

Кривые Безье являются обязательной составляющей геометрического ядра современных систем автоматизированного проектирования (CAD - *Computer-Aided Design*). В статье предлагается математический подход, позволяющий выполнить аппроксимацию сопряжения (соединения) кривых Безье произвольной степени, таким образом, чтобы в точке сопряжения выполнялись условия гладкости (непрерывности) до порядка, равного степени заданных Безье кривых. Данный подход позволяет представить сопряженные кривые одной кривой Безье, со степенью, равной степеням заданных кривых. На сопряженные кривые и аппроксимирующую кривую могут быть наложены дополнительные ограничения в виде полного совпадения с одной из заданных кривых или прохождения аппроксимирующей кривой через заданную точку и равенства производных заданным значениям в этой точке.

Для решения указанных задач вводятся две различные метрики разности между заданными кривыми и аппроксимирующей кривой, формулируются оптимизационные задачи с ограничениями в виде равенств, для решения которых применяется метод множителей Лагранжа, который сводится к решению соответствующей системы линейных алгебраических уравнений. Для представления кривых Безье предлагается использовать базисные функции B-сплайнов, что позволяет пользоваться программными функциями, входящими в геометрическое ядро современных CAD систем. Это существенно упрощает получение производных всех степеней для кривых и, без существенных изменений в последующем, позволит распространить результаты на задачи сопряжения B-сплайнов. Приводятся примеры аппроксимаций с использованием различных метрик и с учетом ограничений (рис. 1, 2, 3).

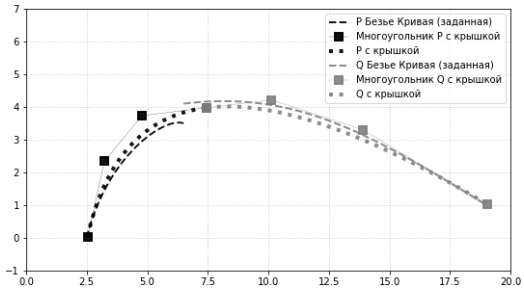


Рис 1. Заданные кривые ($P(u)$ и $Q(v)$) и эти же кривые после сопряжения ($\hat{P}(u)$ и $\hat{Q}(v)$) (дискретная норма)

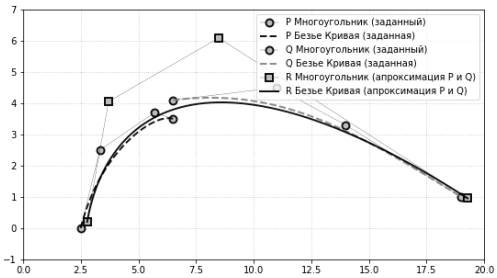


Рис. 2. Замена двух гладко сопряженных кривых одной кривой $R(t)$ (интегральная норма)

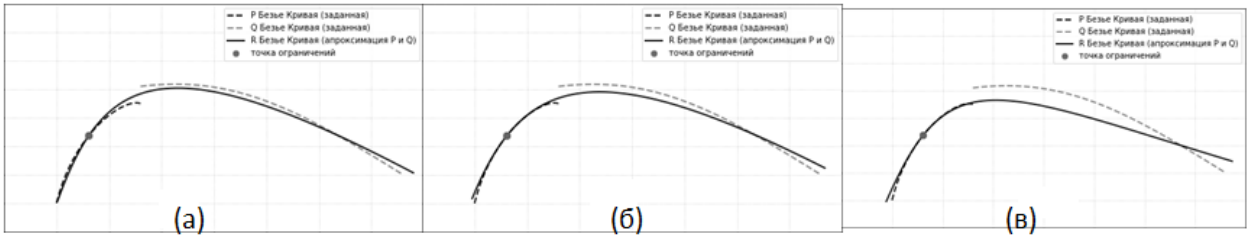


Рис. 3. Сопряжение с ограничениями (интегральная норма), (а) – ограничение – задана точка,
(б) – ограничение – задана точка и первая производная,
(в) – ограничение – задана точка, первая и вторая производные

III.2. Аппроксимация полностью гладкого сопряжения B-сплайнов с учетом дополнительных ограничений

В геометрическом ядре современных систем автоматизированного проектирования (CAD - *Computer-Aided Design*) одним из основных геометрических объектов являются *B*-сплайны. В статье предлагается к рассмотрению аналитический метод для аппроксимации полностью гладкого сопряжения *B*-сплайнов произвольной степени, определенных на любом числе контрольных точек. Под полностью гладким сопряжением понимается выполнение условий непрерывности в точке сопряжения вплоть до максимально возможного порядка (вплоть до $(p-1)$, p – степень сплайна). Для решения этой задачи формулируется экстремальная задача по минимизации отклонения сопряженных сплайнов от исходных. Предлагается аналитический подход, позволяющий перейти от двух полностью сопряженных *B*-сплайнов к одному аппроксимирующему той же степени, но определенному на меньшем числе контрольных точек. Переход не вносит каких-либо погрешностей. На сопряженные *B*-сплайны и, соответственно, аппроксимирующий могут быть наложены дополнительные ограничения в виде полного совпадения с одним из исходных сплайнов или прохождения через заданную точку, для которой могут быть заданы производные, вплоть до $(p-1)$ порядка. Для решения указанных задач в работе вводится метрика разности между сопряженными и исходными сплайнами, значение которой минимизируется. Задача минимизации решается методом неопределенных множителей Лагранжа, которая сводится к системе линейных алгебраических уравнений. Замена двух полностью сопряженных сплайнов одним аппроксимирующим выполняется в два этапа. На первом этапе выполняется переход от закрытого узлового вектора к открытому, для чего записывается равенство производных этих сплайнов в граничных точках области определения. На втором этапе проводится «объединение» двух сплайнов в один с перерасчетом нового закрытого узлового вектора. Приводятся примеры сопряжения с дополнительными ограничениями (рис. 4, 5).

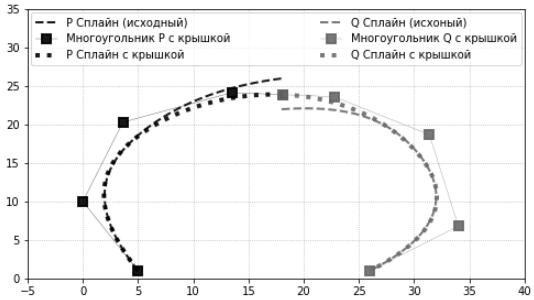


Рис. 4. Исходные *B*-сплайны $P(u)$, $Q(v)$ и сопряженные $\hat{P}(u)$, $\hat{Q}(v)$

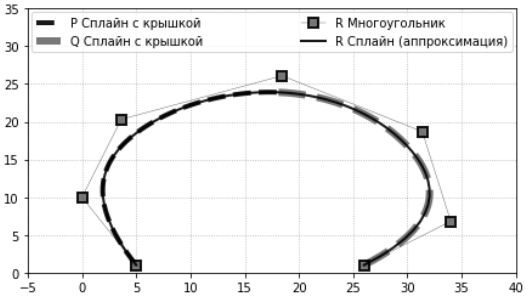


Рис. 5. «Объединение» сплайнов $\hat{P}\_UnClR(u)$ и $\hat{Q}\_UnClL(v)$ в один аппроксимирующий сплайн $R(t)$

III.3. Гладкое сопряжение двух плоских кривых Безье [10]

В работе рассматривается математический метод, посредством которого решается задача сопряжения (соединения) кривой Безье с двумя заданными кривыми в граничных точках. В точках сопряжения должен сохраняться порядок гладкости (непрерывности), соответствующий степени заданных сопрягаемых кривых. Для решения поставленной задачи составляется система линейных уравнений, в которой уравнения являются

условиями равенства производных сопрягаемых кривых до порядка равного степени заданных кривых Безье. Решением этой системы являются координаты вершин определяющего многоугольника кривой сопряжения. Численные вычисления реализованы на языке программирования Python (рис. 6).

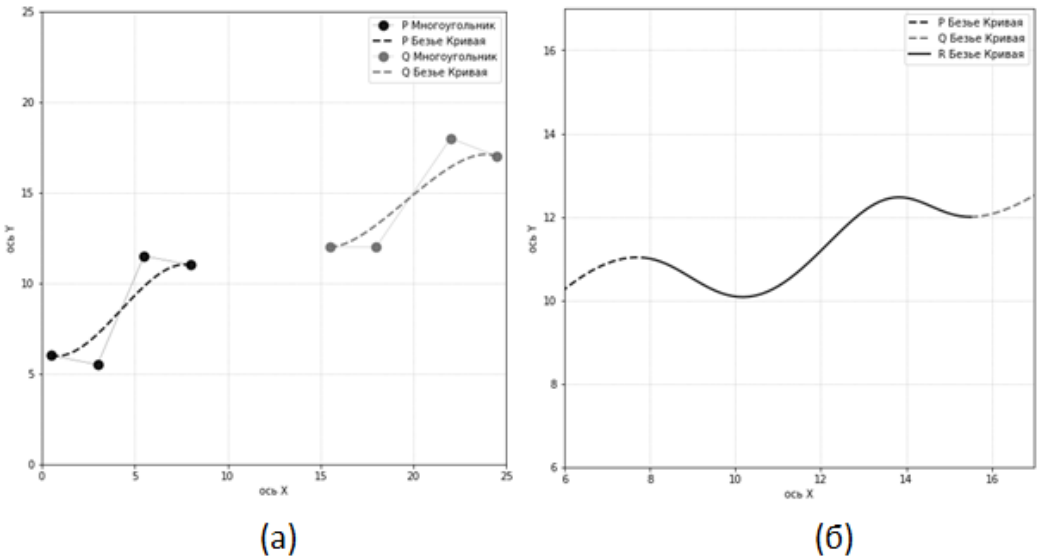


Рис. 6. Графики гладкого сопряжения двух плоских кривых Безье:
(а) – заданные кривые, для которых необходимо выполнить сопряжение с сохранением непрерывности,
(б) – результаты сопряжения с непрерывностью соответствующей непрерывности заданных кривых

III.4. Скругление углов с помощью кривых Безье

В работе рассматривается способ скругления углов, образованных кривыми Безье. Во внутренних точках области определения заданные кривые сопрягаются (соединяются) с кривой скругления. При этом на кривую скругления накладывается условие, что в точках сопряжения должен сохраняться порядок гладкости (непрерывности), соответствующий степени кривых, образующих угол (рис. 7).

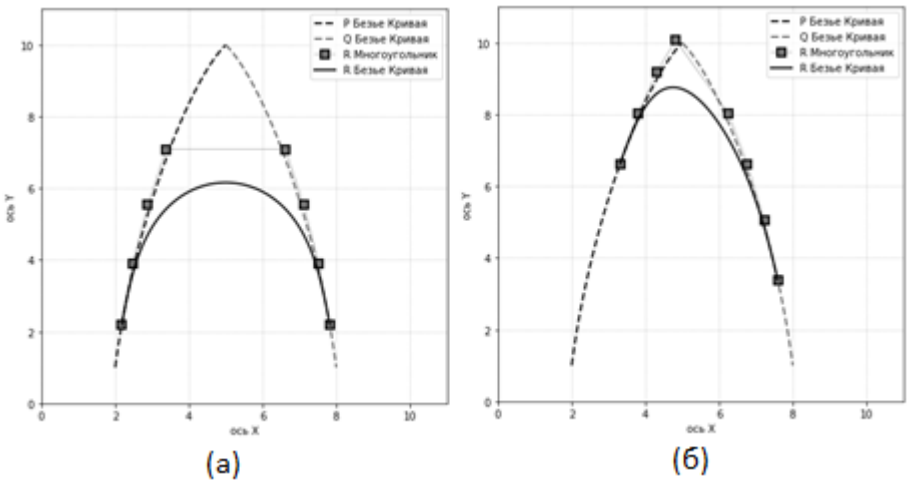


Рис. 7. Симметричное (а) и ассиметричное (б) скругление угла, образованного кривыми Безье 3-ей степени (в точках сопряжения непреравность не нурушается)

III.5. Разработка и сравнительный анализ алгоритмов аппроксимации кривых [10, 11]

В работе представлено описание разработанных алгоритмов понижения степени B-сплайнов. Для получения аппроксимирующего сплайна пониженной степени предлагаются два алгоритма – нативный (простейший) и алгоритм на основе интегральной метрики. Для оценки допустимой степени понижения вводятся три метрики – метрика Хаусдорфа, кривизна сплайна и квадратичная разность между кривыми. Выполненные численные эксперименты показали превосходство (по точности аппроксимации по всем трем метрикам) алгоритма на основе интегральной метрики по сравнению с нативным (рис. 8).

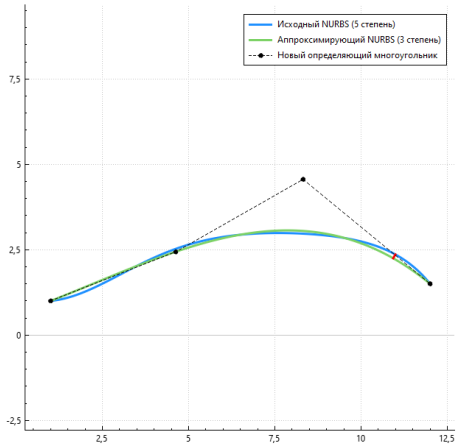


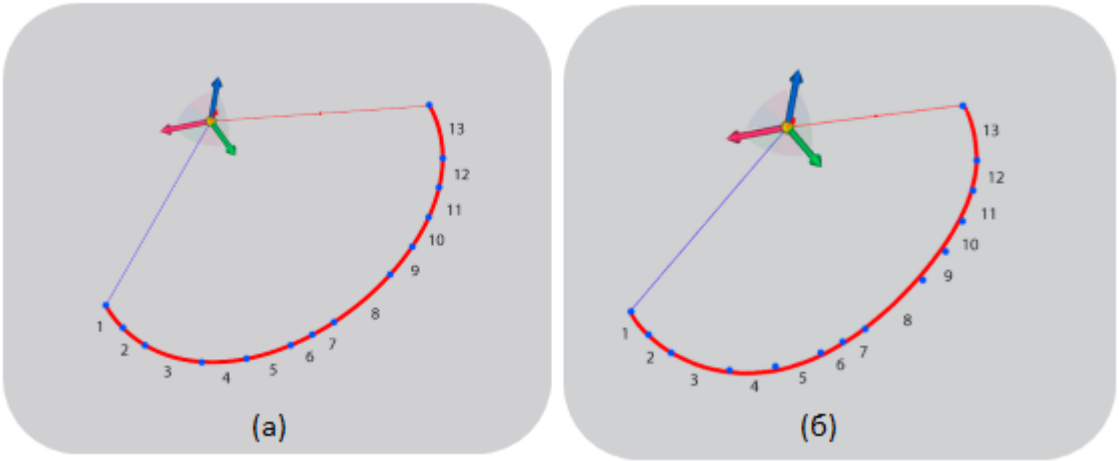
Рис. 8. Понижение степени B-сплайна 5-й степени до 3-й степени

III.6. Разработка алгоритма непрерывного сопряжения кривых Безье [12]

В работе описывается алгоритм полностью гладкого сопряжения (без нарушения непрерывности в точках сопряжения) нескольких кривых Безье произвольной степени. Для оценки точности и сравнения кривых Безье используются три метрики: метрика Хаусдорфа (максимальное линейное расстояние), метрика кривизны (интеграл второй производной) и квадратичная разность (интеграл квадрата разности).

Основа данной работы состоит в решении оптимизационной задачи при наличии ограничений. Для решения этой задачи используется метод множителей Лагранжа, который сводится к системе линейных алгебраических уравнений, и расчёта метрик сравнения кривых.

Данный алгоритм непрерывного гладкого сопряжения кривых реализован на языке C++ и имплементирован в CAD «Сарус» (рис. 9).



Метрика Хаусдорфа = **0.0420225**
Кривизна изменилась на **10.6874**
Квадратичная разность = **0.000741222**

Рис. 9. Проекция на плоскость 13 Безье кривых 5-й степени. (а) – до сопряжения, (б) – после полностью гладкого сопряжения

III.7. Алгоритм и программная реализация корреляционных сплайнов высокой степени [13]

В работе реализованы алгоритмы построения корреляционных сплайнов (сплайнов, проходящих через множество точек и отображающих их генеральную зависимость (тренд, тенденцию, регрессию)). В основу алгоритма положен метод наименьших квадратов, который приводит к системе линейных алгебраических уравнений. В работе реализованы: три метода расчета узлового вектора, учет весов контрольных точек и учет заданных значений производных. Алгоритмы реализованы на языке C++ и имплементирован в CAD «Сарус» (рис. 10, 11).

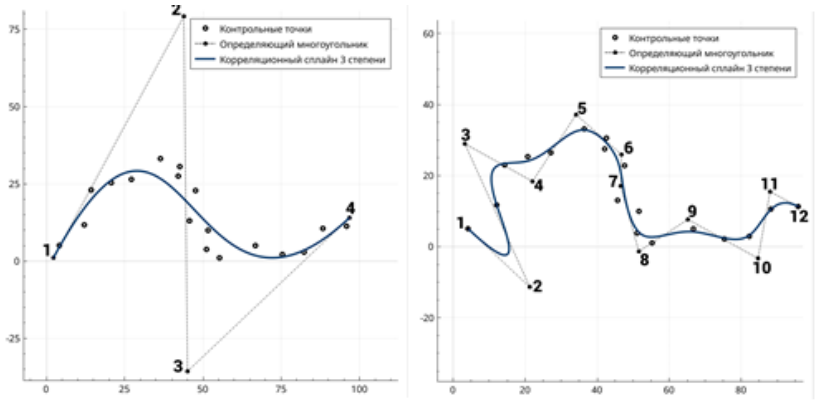


Рис. 10. Изменение корреляционного сплайна в зависимости от количества контрольных точек:
(а) – 4 контрольные точки, (б) –12 контрольных точек

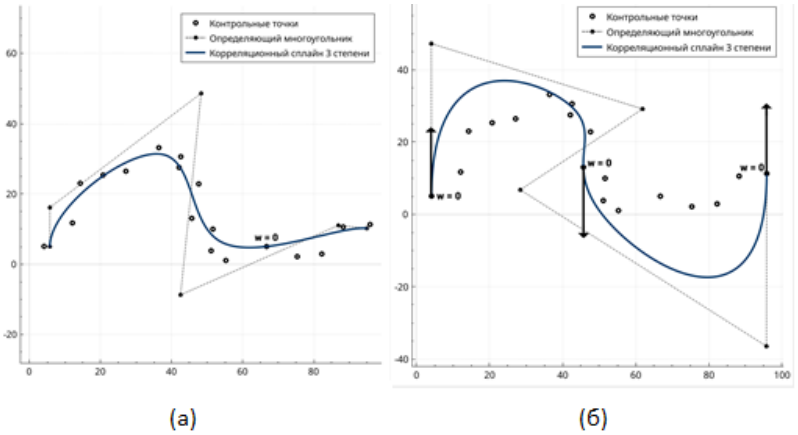


Рис. 11. Изменение корреляционного сплайна: (а) – при весе точки w=0 сплайн интерполирует эту точку,
(б) – влияние заданных производных на корреляционный сплайн

III.8. Алгоритмы построения поверхностей высокой степени гладкости [14]

В работе реализован алгоритм построения NURBS произвольного порядка непрерывности (произвольная степень NURBS) и реализован алгоритм расчета частных производных, вплоть до максимально порядка. Алгоритмы реализованы на языке C++ и имплементирован в CAD «Сарус» (рисунок 12).

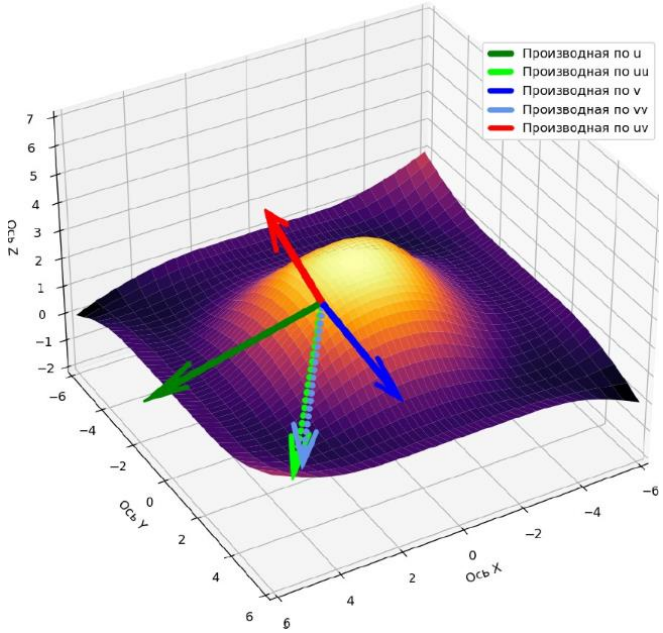


Рис.12 – Построенная поверхность и ее производные

IV. Обсуждение результатов

Исходя из выводов по рациональности проведения подобных работ при обучении студентов и в целях подготовки специалистов для сопровождения и развития геометрического ядра CAD «Сарус», ведущие специалисты ИЦТ РФЯЦ-ВНИИЭФ и научно-преподавательский состав кафедры ЦТ и ТСМ СарФТИ-НИЯУ МИФИ выполнили анализ имеющихся и разработали дополнительные учебные курсы. Было определено, что подготовка специалистов будет основываться на основных (читаемых преподавателями СарФТИ-НИЯУ МИФИ) курсах:

- по направлению программирования, математического моделирования и других информационных технологий:
  - математический анализ;
  - аналитическая геометрия и линейная алгебра;
  - информационные технологии;
  - математическая логика и теория алгоритмов;
  - вычислительная математика;
  - программирование на языке С;
- по направлению конструкторско-технологического обеспечения машиностроительных производств:
  - 3D-моделирование в машиностроении;
  - разработка конструкторской документации.

Дополнительно к базовым предметам были разработаны следующие курсы, которые ведут сотрудники ИЦТ РФЯЦ-ВНИИЭ:

- по направлению программирования, математического моделирования и других информационных технологий:

1. Курс «Объектно-Ориентированное Программирование на С++», который включает в себя:
   - функции, функции высшего порядка, чистые функции;
   - разделение команд и запросов, лямда-функции;
   - структуры, методы, конструкторы и деструкторы;
   - классы и объекты, перегрузка и переопределение методов;
   - полиморфизм;
   - основные шаблоны проектирования.
2. Курс «Численные методы анализа и линейной алгебры», который включает в себя:
   - источники и классификация погрешностей;
   - приближение функций;
   - численное дифференцирование и интегрирование;
   - вычисление корней уравнений;
   - решение задач линейной алгебры;
   - алгебраическая проблема собственных значений;
   - задача Коши для ОДУ. Методы решения.
3. Курс «Геометрическое моделирование» (находится в стадии разработки), который включает в себя:
   - аффинные преобразования;
   - проецирование;
   - Безье кривые, B-сплайны, NURBS;

- построение поверхностей;
- гладкие сопряжения.

- по направлению конструкторско-технологического обеспечения машиностроительных производств

4. Формирование и функционирование проектных модулей, который включает в себя:

- описание процесса жизненного цикла изделий в технологии сквозного проектирования;
- компоненты инфраструктуры – центры обработки данных, рабочие места, сервера, линии связи, коммуникационное оборудование;
- преобразование бизнес-процессов в процессы информационных систем;
- программное обеспечение, форматы данных, интеграционные взаимосвязи;
- электронная структура изделия (ЭСИ), электронная модель изделия (ЭМИ), электронный конструкторский документ (ДЭ);
- функционирование проектных модулей PDM, CAD, CAE, CAM;
- принципы организации единого информационного пространства (ЕИП) предприятия

V. Выводы и заключение

В работе были рассмотрены некоторые вопросы, касающиеся разработки и сопровождения геометрического ядра современных CAD систем, показано, что одной из ключевых проблем является подготовка кадров для этой области. В этой связи приведено описание процесса подготовки таких специалистов для геометрического ядра CAD «Сарус». Подготовка ведется совместно преподавательским составом ВУЗ-а (кафедрой ЦТ и ТСМ СарФТИ-НИЯУ МИФИ) и специалистами базового предприятия (ИЦТ ФГУП РФЯЦ-ВНИИЭФ). Студенты, как бакалавры, так и магистры активно привлекаются к выполнению реальных задач по развитию геометрического ядра CAD «Сарус».

Теоретическая значимость предлагаемого подхода состоит в совершенствовании организации учебного процесса подготовки специалистов в области программной архитектуры и математических методов, используемых в геометрических ядрах современных САПР. Практической значимостью является выпуск специалистов обладающих детальными (полноценными) знаниями некоторых задач, необходимых для дальнейшего развития геометрического ядра CAD «Сарус» и имеющих представление о архитектуре программного комплекса.

Источник финансирования

Договор от 25 марта 2024 г. № 96-2024/244 на выполнение научно- исследовательской работы по теме: «Проведение научных исследований в интересах ФГУП «РФЯЦ-ВНИИЭФ» с использованием результатов для актуализации образовательных программ подготовки специалистов с высшим профессиональным образованием для ФГУП «РФЯЦ-ВНИИЭФ». Задача 10 «Исследование программно-аппаратных решений в области цифровых технологий, возможностей их отраслевого применения, разработка и версионное тестирование компонентов полностью защищенной системы полного жизненного цикла среднего класса СПЖЦ V3 «Сарус».

Научный руководитель: Маврин Сергей Валентинович, Институт цифровых технологий РФЯЦ-ВНИИЭФ, СарФТИ-НИЯУ МИФИ

Список литературы

1. Отечественная система полного жизненного цикла «Сарус» обеспечит импортонезависимость и безопасность. //САПР и графика, №12, 2023, с. 68-71.
2. Голованов Н. Н. Геометрическое моделирование. – М.: ДМК Пресс, 2020.-406 с.: ил.
3. Прата С. Язык программирования C++. Лекции и упражнения, 6-е изд.: Пер. с англ.-М.: ООО «И.Д. Вильямс», 2012. 1248 с.: ил.
4. Ушаков Д. NURBS и САПР: 30 лет вместе //Интернет-ресурс: http://isicad. ru/ru/articles. php.
5. Роджерс Д.Ф., Адамс Дж. А. Математические основы машинной графики. М.: Мир, 2001. 602 с.
6. Фокс А., Пратт М. Вычислительная геометрия. Применение в проектировании и на производстве: Пер. с англ.-М.: Мир, 1982.-304 с., ил.

7. Piegl, L. The NURBS Book / L. Piegl, W. Tiller. Springer, 1997. 646 pp.

8. Аппроксимация сопряжения кривых Безье с сохранением порядка гладкости и дополнительными ограничениями / С.Н. Ганчук, О.В. Кривошеев, С.В. Маврин, С.А. Рыжов // Вестник ЮУрГУ. Серия «Строительство и архитектура». 2024. Т. 24, № 1. С. 59–69. DOI: 10.14529/build240108

9. Гладкое сопряжение двух плоских кривых Безье / С.Н. Ганчук, С.В. Маврин, В.В. Сёмина, А. С. Старкова // Вестник Липецкого государственного технического университета. 2024. №1(54), С. 5-8.

10. Разработка и сравнительный анализ алгоритмов аппроксимации кривых / С.А. Рыжов, С.В. Маврин, О.В. Кривошеев // Сборник трудов конференции «Молодежь в науке 2023» (принята к публикации)

11. Алгоритмы и программная реализация понижения степени B-сплайнов / С.А. Рыжов. Научный руководитель: С.В. Маврин // Всероссийская научно-практическая конференция им. Жореса Алфёрова: сборник тезисов статей. М.: Издательство «Директ-Медиа», 2023. С. 27.

12. Разработка алгоритма непрерывного сопряжения кривых Безье / С.А. Рыжов, С.В. Маврин // Сборник материалов XVIII Всероссийской научно-инновационной школы. Саров: «Интерконтакт», 2024. С. 233-234.

13. Алгоритм и программная реализация интерполяционных сплайнов высокой степени / Д.А. Гришин, Е.Н. Ерофеев, С.Н. Ганчук, С.В. Маврин // Сборник материалов XVIII Всероссийской научно-инновационной школы. Саров: «Интерконтакт», 2024. С. 191.

14. Алгоритмы построения поверхностей высокой степени гладкости / Е.Н. Ерофеев, Д.А. Гришин, С.Н. Ганчук, С.В. Маврин // Сборник материалов XVIII Всероссийской научно-инновационной школы. Саров: «Интерконтакт», 2024. С. 210.

УДК 004.925.8
DOI: 10.25206/978-5-8149-3873-2-2024-795-804

Роль системы координат в построении 3D-модели в СПЖЦ «САРУС»
The role of the coordinate system in the construction of a 3D model in the SARUS Full Life Cycle System

Ю.Б. Прыткова[1], Н.А.Денисова[1],
[1]Саровский физико-технический институт - филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский ядерный университет "МИФИ" (СарФТИ НИЯУ МИФИ), ул. Духова, д.6, г. Саров, 607186, РФ

U.B. Prytcova [1], N.A. Denisova[1],
[1]Sarov Institute of Physics and Technology - branch of the Federal State Autonomous Educational Institution of Higher Education "National Research Nuclear University "MEPhI" (SarPhTI NRNU MEPhI), Dukhova str., 6, Sarov, 607186, Russian Federation

Аннотация. В ядерном центре успешно реализуется пилотный проект по созданию и внедрению типовой информационной системы отраслевых предприятий ядерно-оборонного комплекса. Одной из таких разработок является комплекс программ «Система полного жизненного цикла изделий «Цифровое предприятие» САРУС, в версионном функциональном тестировании программного модуля «Система конструкторского проектирования» (ПМ CAD) которого приняли участие студенты магистратуры кафедры Технологии специального машиностроения.

Данная работа посвящена анализу результатов тестирования, где приводится аналитический обзор особенности применения базовых операций/построений 3D-модели в «САРУС» с использованием алгоритмов функций, заявленных разработчиками. В работе рассматривается «Построение системы координат» (СК). В результате аналитического сравнения предлагаемого алгоритма функций и выполняемого алгоритма построения выявлено, что все СК, созданные в процессе работы с моделью, доступны в «Навигаторе» модели. К каждой системе координат можно обратиться в любой момент для выполнения различных задач моделирования. Сделан вывод, что точная привязка к конкретной системе координат на начальном этапе создания 3D-модели делает дальнейшую работу с моделью с использованием принципа сквозного проектирования на платформе «САРУС» более удобной.

Ключевые слова. САПР, Программный модуль CAD, 3D-модель, функции построения, Локальная система координат.

Abstract. The nuclear center is successfully implementing a pilot project to create and implement a standard information system for industrial enterprises of the nuclear defense complex. One of such developments is the software package "Full life cycle system of products "Digital Enterprise" SARUS, in the version functional testing of the software module "System of Design Engineering" (PM CAD), which was attended by students of the Master's degree of the Department of Special Engineering Technology.

This work is devoted to the analysis of test results, which provides an analytical overview of the features of the application of basic operations/constructions of a 3D model in SARUS, using algorithms of functions declared, by the developers. The paper considers the "Construction of a coordinate system" (CS). As a result of an analytical comparison of the proposed algorithm of functions and the construction algorithm being performed, it was revealed, that all the CS, created in the process of working with the model, are available in the "Navigator" of the model. Each coordinate system can be accessed at any time to perform various modeling tasks. It is concluded, that accurate reference to a specific coordinate system at the initial stage of creating a 3D model, makes further work with the model using the principle of end-to-end design on the SARUS platform more convenient.

Keywords. CAD, CAD software module, 3D model, construction functions, Local coordinate system.

I. Введение

В последние годы отраслевые предприятия начинают процессы активного внедрения инновационных информационных технологий комплексной автоматизации управленческой и производственной деятельности на основе современных методологий. В ядерном центре успешно реализуется пилотный проект по созданию и внедрению типовой информационной системы отраслевых предприятий ядерно-оборонного комплекса. Одной из таких разработок является комплекс программ «Система полного жизненного цикла изделий «Цифровое предприятие» (далее – «САРУС»), в версионном функциональном тестировании программного модуля «Система

конструкторского проектирования» (далее – ПМ CAD) которого приняли участие студенты магистратуры кафедры ТСМ, в том числе автор данной работы.

II. Постановка задачи

Целью работы является выполнение анализа результатов тестирования, где приводится аналитический обзор особенности применения базовых операций/построений 3D-модели в «САРУС» с использованием алгоритмов функций, заявленных разработчиками. В данной работе рассматривается в частности «Построение системы координат» (далее – СК) [1].

Объект исследования: комплекс программ «Система полного жизненного цикла изделий «Цифровое предприятие» («САРУС»).

Предмет исследования: функция «Построение системы координат».

Поставленные ***задачи***:

- Построить 3D-модель простой детали «Опора» в СПЖЦ «САРУС».
- При построении в качестве тестирования использовать алгоритмы функции «Построение системы координат».
- Выполнить аналитический обзор особенностей применения операций.

III. Теоретическое обоснование математической основы исследования

В САПРах система координат (далее - СК) играет важнейшую роль при построении 3D-моделей. На начальном этапе моделирования конструктор задаёт СК, которой следует на протяжении всего процесса. Эта геометрическая особенность определяет ***противоречие*** – своеобразную «разность потенциалов»: привязанность к СК является несомненным плюсом при создании операций механической обработки модели и управляющей программы обработки в модуле «Система разработки управляющих программ с ЧПУ» (ПМ CAM), но повышает трудоёмкость работы конструктора, так как все размеры модели также выдерживаются с помощью координат. Устранение этого противоречия не является очевидным, поэтому мы имеем ***проблему***, для которой надо найти решение.

III.1. Локальная система координат

В 3D-моделировании локальная система координат (ЛСК) является неотъемлемой частью математического аппарата. С её помощью определяется положение точки в пространстве. ЛСК задаётся точкой начала системы координат и направлениями её осей X, Y, Z. Объект ЛСК может быть создан:

- по умолчанию – система расположена в точке (0,0,0), а оси координат направлены вдоль глобальных осей X, Y, Z;
- копированиеv исходной ЛСК;
- по точке начала системы координат;
- по точке начала системы координат и направлению оси X;
- по точке начала системы координат, направлению оси X и направлению нормали к плоскости XY;
- по точке начала системы координат, направлениям осей X, Y, Z.

На рис. 1, *а* показано, как задаётся СК по умолчанию при построении призмы, а на рис.1, *б* – автоматическое создание ЛСК при способе задания «Система координат по объекту». Как видно системы координат по умолчанию совпадают.

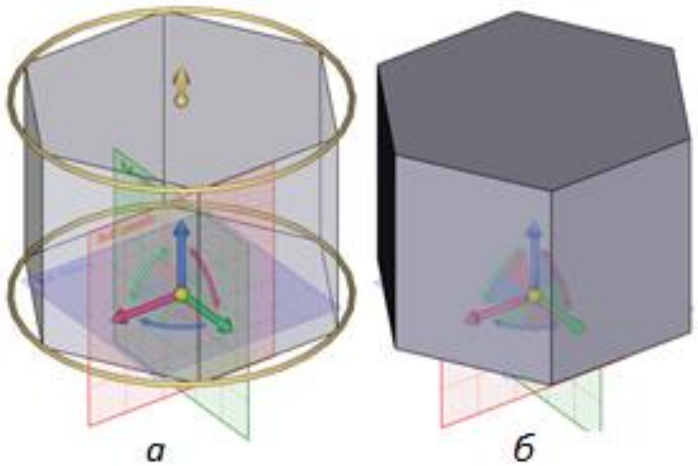


Рис. 1. СК по умолчанию и автоматическое создание ЛСК

Разработчиками также представлены требования к функции «Построение системы координат». По документам разработчика функция должна обеспечивать как общие требования, так и особые требования, присущие каждому из четырех способов построения СК.

IV. Результаты экспериментов. Построение модели в СПЖЦ «САРУС»

Одна из особенностей построения 3D-моделей заключается в способе задания поверхностей, которое осуществляется через указание координат в пространстве. Конструктору задаёт СК, которой следует на протяжении всего процесса.

Разработчики предлагают свои алгоритмы построения СК. В СПЖЦ «САРУС» есть 4 способа построения системы координат. Рассмотрим два алгоритма построения системы координат: построение СК по трем точкам и СК координат по двум отрезкам/осям.

В Пояснительной записке технологического процесса, часть 4 представлены требования к функциям, описанным в разделе 1.2, которые должны выполняться при работе пользователей в «САРУС» [2]. Реализацию этих функций проверим в процессе построения модели ***простой*** детали (II группа конструктивной сложности) по заданному алгоритму.

IV.1. Построение системы координат по двум отрезкам/осям

Построим 3D-модель детали «Опора» и зададим ЛСК, следуя алгоритму. Для того, чтобы Построить ЛСК по двум отрезкам, сначала построим основание детали, которое является параллелепипедом.

Разработчиком предлагается следующий алгоритм функции [2]:

- активизируем инструментальную панель «Вспомогательная геометрия»;
- на инструментальной панели «Вспомогательная геометрия» нажимаем на пиктограмму «ЛКС»;
- в диалоговом окне выбираем в качестве способа построения – «построение по двум отрезкам/осям»;
- выбираем первый отрезок;
- выбираем второй отрезок. После этого система должна построить фантом ЛКС с точкой начала в месте пересечения осей/отрезков, при этом ось Y будет лежать в плоскости выбранных осей/отрезков, а ось Z будет перпендикулярна этой плоскости;
- выбираем ориентацию координатных осей ЛКС.

Любую систему координат можно сделать активной в «Навигаторе» модели и продолжить построение, в новой системе координат.

На первом этапе построения создаём примитив для основания детали с заданными параметрами (рис. 2). Построение ведётся в системе координат, созданной по умолчанию (начало в точке с координатами (0,0,0), а оси координат направлены вдоль глобальных осей X, Y, Z).

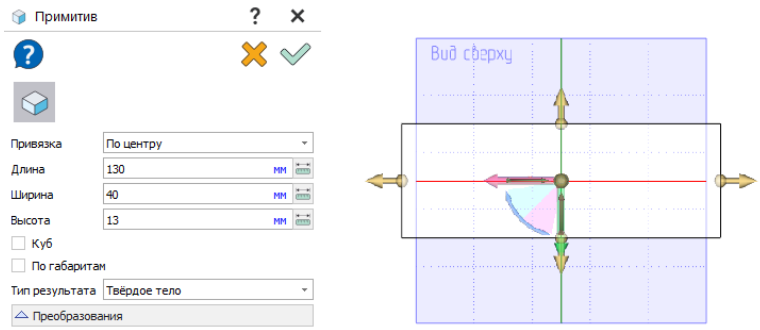


Рис. 2. Первый этап построения

На втором этапе создаём ЛСК по двум отрезкам. Для этого выделяем необходимые отрезки (рёбра параллелепипеда) как показано на рис. 3.

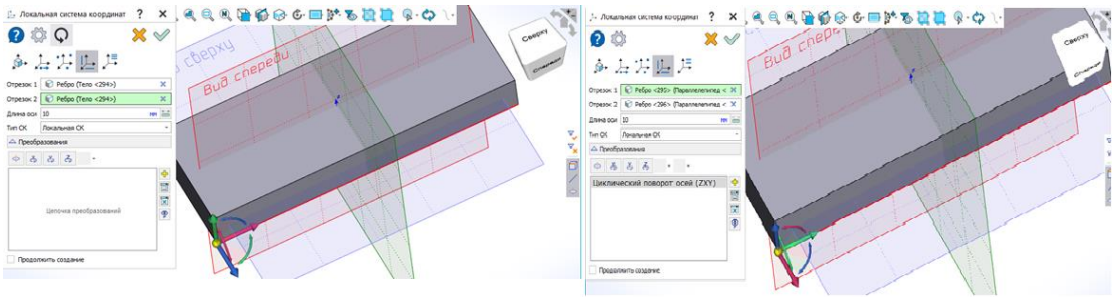


Рис. 3. Второй этап построения

При необходимости можно поменять ориентацию так, чтобы OZ смотрела вверх, используя «Преобразования», «Циклический поворот осей».

Делаем новую ЛСК активной и продолжаем построение детали «Опора». На этапе три делаем скругление радиусом 20 мм на гранях примитива с обеих сторон (рис. 4).

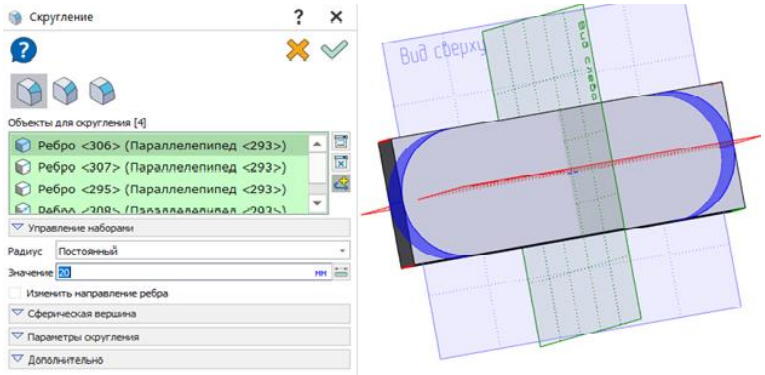


Рис. 4. Третий этап построения

На четвёртом этапе нужно построить на основании два цилиндра, пользуясь «Примитивом». Чтобы поставить цилиндр в нужное место, необходимо посчитать его положение относительно ЛСК. Перемещение по OY будет равно 20 мм. По OX перемещение 20 мм со знаком минус. По OZ перемещение равно 13 мм, так как учитывается высота основания (рис. 5).

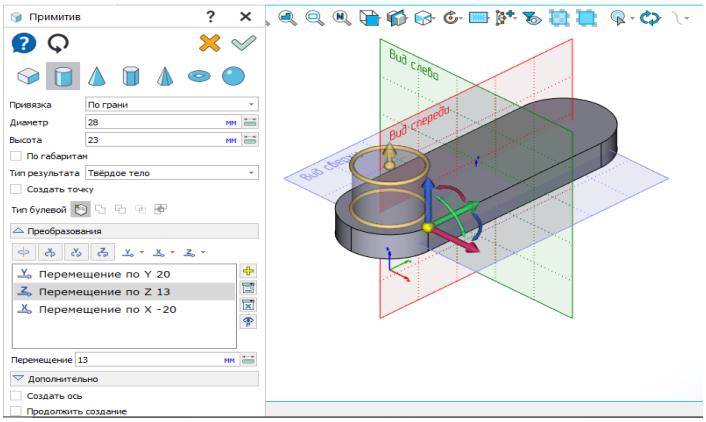


Рис. 5.Четвёртый этап построения

Далее аналогично строится второй цилиндр, с перемещением по OУ на 110 мм. Затем необходимо применить «Булеву операцию», чтобы цилиндры и основание стали одним целым.

На пятом этапе создаём отверстия. Задать расположение для отверстия теперь легче, так как мы уже рассчитали некоторые координаты, когда строили цилиндры на предыдущем этапе (рис. 6). Два отверстия ступенчатые, поэтому при построении используем команду «отверстие с цековкой» (рис. 7).

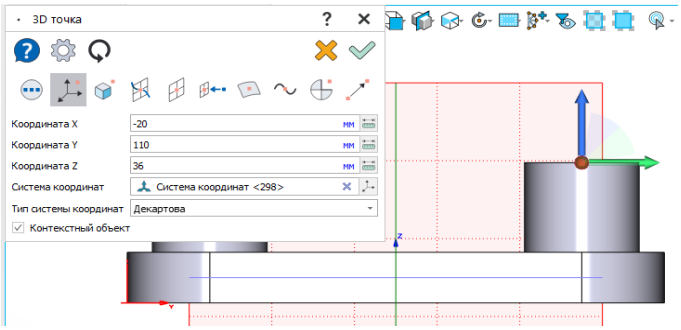


Рис. 6. Координаты для отверстия

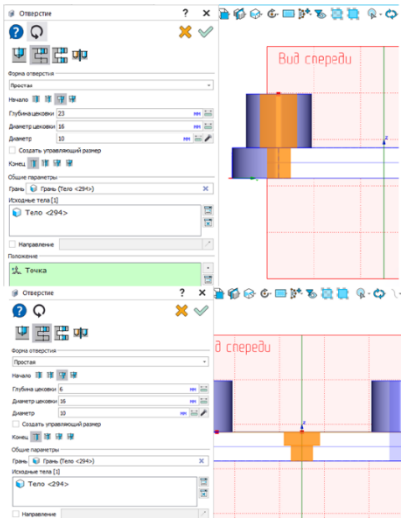


Рисунок 7 – Пятый этап построения

На шестом этапе создаем резьбу М12 в ранее построенном гладком отверстии (рис. 8).

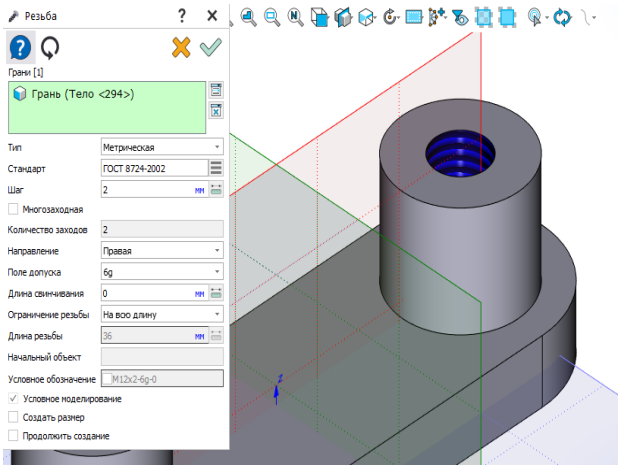


Рис. 8. Шестой этап построения

На седьмом этапе предстоит построить рёбра жёсткости. Это можно сделать разными способами. Например, с помощью примитива строим параллелепипед со сторонами 15х15 мм. Высчитываем правильные координаты размещения. А затем с параллелепипеда снимем фаску и воспользуемся операцией «Симметрия». (рис. 9).

Далее обязательно выполняется «Булева операция» для объединения рёбер с остальной деталью (рис. 10).

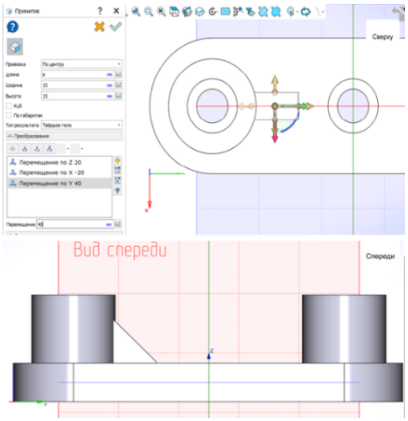


Рис. 9. Седьмой этап построения

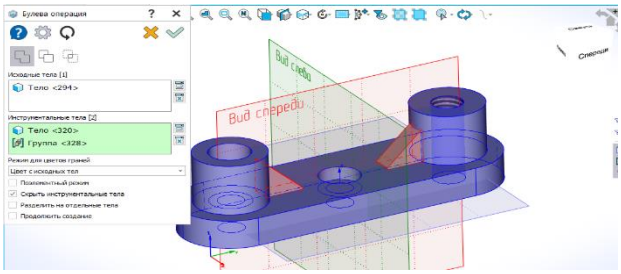


Рисунок 10 – Булева операция

На этапе восемь делаем необходимые скругления радиусом 1 мм по требованию чертежа. Если скругления не получаются, то стоит проверить, выполнена ли ранее «Булева операция объединения тел» (рис. 11).

Все использованные операции, тела и системы координат отображены в «Навигаторе» получившейся 3D-модели, что показано на рис. 12.

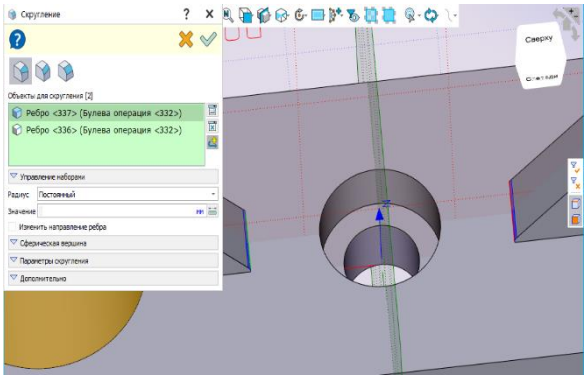


Рисунок 11 – Восьмой этап построения

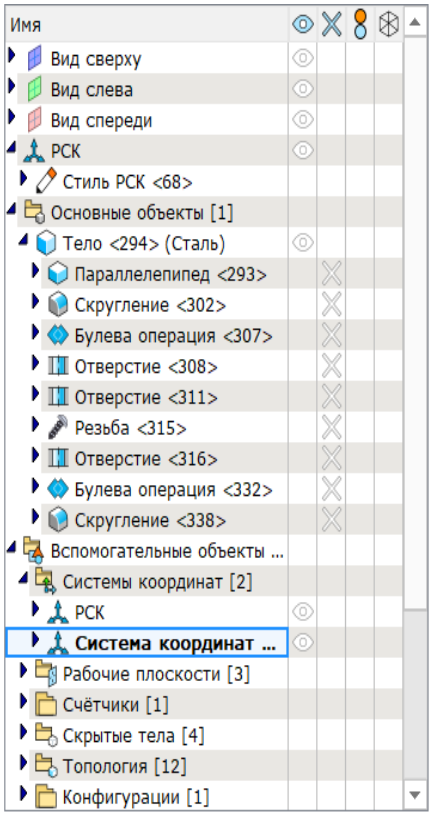


Рис. 12. «Навигатор» 3D-модели

IV.2. Построение системы координат по трем точкам

Алгоритмы рассматриваемых функций во многом похожи, но имеют свои особенности построения.

Алгоритм функции:

- активизируем инструментальную панель «Вспомогательная геометрия»;
- нажимаем на инструментальной панели «Вспомогательная геометрия» на пиктограмму «СК»;
- выбираем в диалоговом окне в качестве способа построения - «СК по трем точкам»;
- выбираем в графическом окне 3D-модели первую точку - точку начала СК;

- выбираем в графическом окне 3D-модели вторую точку. Вектор, направленный от первой выбранной точки до второй выбранной точки, будет определять направление оси X (вектор OX);
- выбираем в графическом окне 3D-модели третью точку. После выбора третьей точки будет построена СК, при этом вектор OY (ось Y) будет перпендикулярен оси OX и лежать в плоскости, проходящей через ось OX и третью выбранную точку, а вектор OZ (ось Z) будет построен перпендикулярно плоскости XY;
- при необходимости, выбрать ориентацию в диалоговом окне преобразований СК.

Этапы построения детали «Опора» с помощью функции построения СК по трем точкам будут по большей части схожи. Поэтому рассмотрим особенности.

Основные отличия будут на первом этапе – собственно построение ЛСК. В первую очередь необходимо создать основание детали. В данном случае можно сразу сделать скругления параллелепипеда и приступить к выбору в графическом окне 3D-модели трёх точек необходимых для построения ЛСК по трём точкам (рис.13).

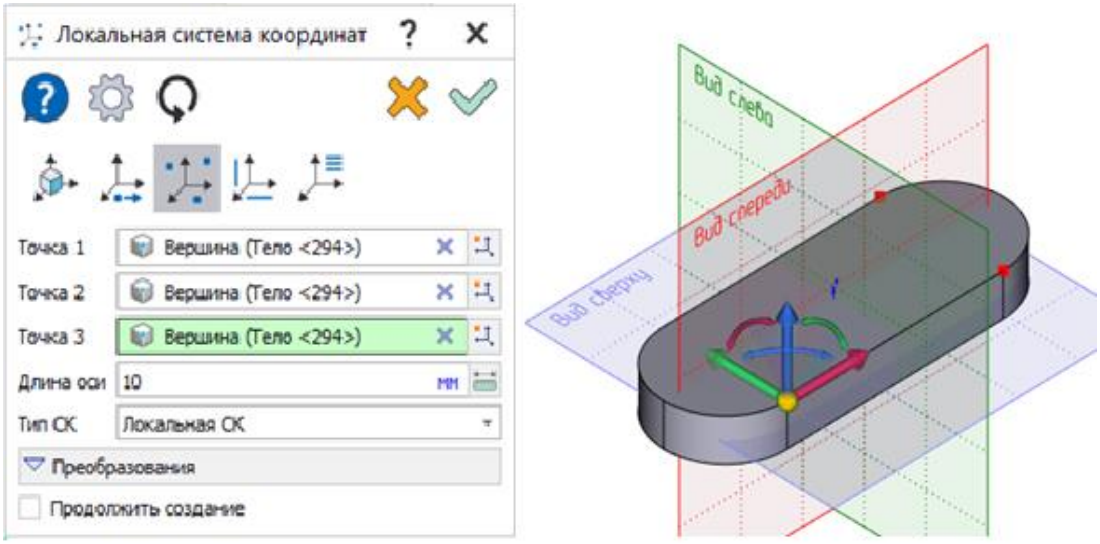


Рис. 13. Задание ЛСК по трём точкам

Далее в «Навигаторе» модели выбираем новую ЛСК и делаем её активной. Продолжаем построение детали «Опора» аналогично уже рассмотренному выше алгоритму.

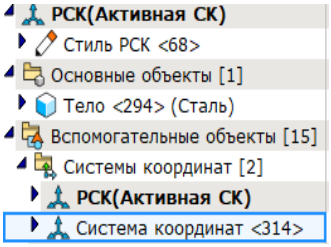
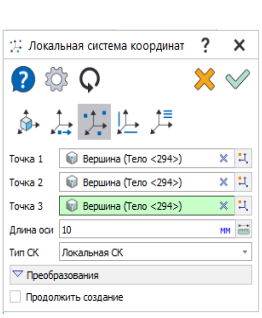
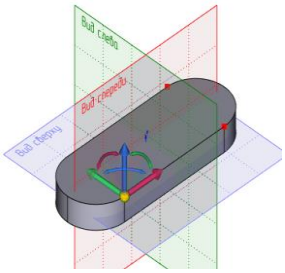
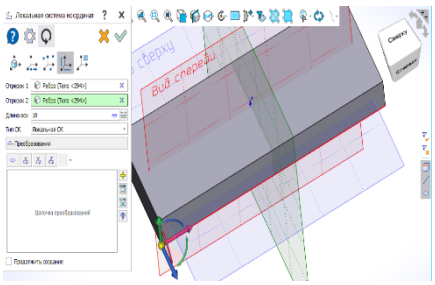
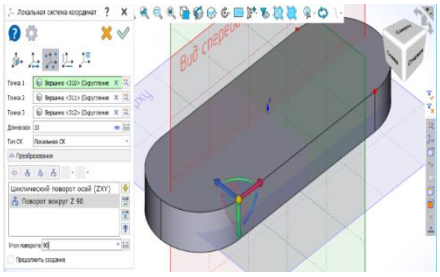
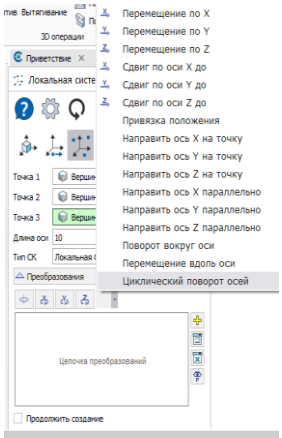
V. Обсуждение результатов. Требования к функции

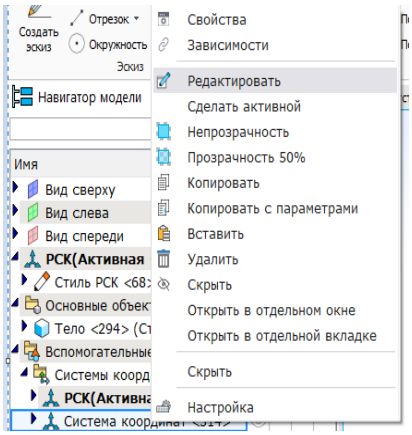
В Пояснительной записке разработчика перечислены все требования, которым должна соответствовать функция «Построение системы координат» [2]. Существуют общие требования к обеим функциям, но также и отличия, обусловленные особенностями построения. Функции должны обеспечивать:

- появление пиктограммы ЛСК в дереве модели;
- возможность назначения ЛСК имени, отличного от имени, назначаемого системой по умолчанию;
- в графическом окне 3D-модели: выбор точек; выбор пересекающихся отрезков/осей. Отрезки могут быть выбраны и в том случае, если продолжения этих отрезков так же пересекаются;
- построение точки начала ЛСК в месте пересечения отрезков/осей или их (отрезков) продолжений, при этом ось X СК будет коллинеарна первому выбранному отрезку/оси;
- изменение ориентации векторов координатных осей;
- отображение в свойствах ЛСК координат в АСК в качестве справочной информации; редактирование параметров ЛСК после ее создания через контекстное меню в дереве модели и удаление СК при необходимости.

Проанализируем выполнение требований к функциям в Табл. 1.

ТАБЛИЦА 1
АНАЛИЗ РЕАЛИЗАЦИИ ТРЕБОВАНИЙ К ФУНКЦИЯМ

Требование	**Реализация**
Появление пиктограммы ЛСК в дереве модели.	
Возможность назначения ЛСК имени, отличного от имени, назначаемого системой по умолчанию.	Не удалось реализовать данное требование.
а) выбор точек в графическом окне 3D-модели.	 
б) выбор пересекающихся отрезков/осей. Отрезки могут быть выбраны и в том случае, если продолжения этих отрезков так же пересекаются; в) построение точки начала ЛСК в месте пересечения отрезков/осей или их (отрезков) продолжений, при этом ось X СК будет коллинеарна первому выбранному отрезку/оси.	Требование к функции реализуется. В данном случае были выбраны два ребра параллелепипеда, пересекающихся в вершине. Также возможно выбрать точку начала ЛСК на продолжении отрезков. 
Изменение ориентации векторов координатных осей.	Выбирается ось и ЛСК поворачивается на нужный угол. Можно осуществить в окне операции.  

Требование	Реализация
Отображение в свойствах ЛСК координат в АСК в качестве справочной информации.	Не удалось реализовать данное требование
Редактирование параметров ЛСК после ее создания через контекстное меню в дереве модели и удаление СК при необходимости.	

По результатам анализа можно заключить, что требования, предъявляемые к функциям, в основном, выполняются. Функции работают, построение ЛСК простое и понятное, не требует большого количества времени.

Таким образом очевидно, что реализация функции «Построение системы координат» для создания простых деталей (I, II групп конструктивной сложности) выполняется.

VI.Выводы и заключение

В результате работы были выполнены поставленные задачи:

- Построена 3D-модель детали в СПЖЦ «САРУС».
- Использованы два алгоритма функции «Построение системы координат».
- Выполнен аналитический обзор особенностей применения операций.

В результате тестирования установлено, что при создании простой детали (I, II степени сложности) функции работают исправно, быстро, удобный функционал дает возможность не затрачивать на построение ЛСК много времени. Если сравнивать построение в ПО «САРУС» и другой распространенной САПР, например, КОМПАС-3D [3], очевидно, что затрачивается больше времени, чтобы строить детали в соответствии с координатами, чтобы выдержать все требуемые размеры безошибочно. Ожидается, что в дальнейшем, при корректировке и доведении СПЖЦ «САРУС» до совершенства производительность, комфортность и эффективность работы конструктора будут достигнуты.

В результате выполненного алгоритма построения выявлено, что все СК, созданные в процессе работы с моделью, доступны в «Навигаторе» модели. К каждой системе координат можно обратиться в любой момент для выполнения различных задач моделирования [4]. Сделан вывод, что точная привязка к конкретной системе координат на начальном этапе создания 3D-модели делает дальнейшую работу с моделью с использованием принципа сквозного проектирования на платформе «САРУС» более удобной.

Источник финансирования. Благодарности

Работа выполнена в соответствии:

- с Рабочим учебным планом по направлению 15.04.05- Конструкторско-технологическое обеспечение машиностроительных производств;

- с Договором от 25 марта 2024 г. № 96-2024/244 на выполнение научно- исследовательской работы по теме: «Проведение научных исследований в интересах ФГУП «РФЯЦ-ВНИИЭФ» с использованием результатов для актуализации образовательных программ подготовки специалистов с высшим профессиональным образованием для ФГУП «РФЯЦ-ВНИИЭФ», Задача 10 «Исследование программно-аппаратных решений в области цифровых технологий, возможностей их отраслевого применения, разработка и версионное тестирование компонентов полностью защищенной системы полного жизненного цикла среднего класса СПЖЦ V3 «Сарус».

Научный руководитель Денисова Наталья Александровна, СарФТИ-НИЯУ МИФИ, кафедра Технологии специального машиностроения

Авторы благодарят:

- Кафедру цифровых технологий СарФТИ-НИЯУ МИФИ под руководством Кривошеева О.В., к.т.н., заведующий кафедрой, зам. директора ФГУП «РФЯЦ-ВНИИЭФ» по технологиям полного жизненного цикла - директор института цифровых технологий (ИТЦ) - главный конструктор СПЖЦ;

Список литературы

1. Комплекс программ в защищённом исполнении «Система полного жизненного цикла изделий «Цифровое предприятие»: Программный модуль «Система конструкторского проектирования». Описание программы 07623615.00423-05 13 01. 2021 г. 80 с.
2. Комплекс программ в защищенном исполнении «Система полного жизненного цикла изделий «Цифровое предприятие». Основная версия программного модуля. «Система конструкторского проектирования» (версия 1)/ПЗ ТП 07623615.00423-04 81-4. Часть 4. 2021 г. 282 с.
3. Голованов Н.Н. Геометрическое моделирование. Учебное пособие. – М.: КУРС: ИНФРА-М, 2016. 400 с.
4. Комплекс программ в защищенном исполнении «Система полного жизненного цикла изделий «Цифровое предприятие». Основная версия программного модуля. «Система конструкторского проектирования» (версия 2)/Руководство оператора 07623615.00423-06 34 01. – 2021 г. – 338 с.

УДК 004.925.8
DOI: 10.25206/978-5-8149-3873-2-2024-805-812

Особенности применения булевых операций в программном модуле «Система конструкторского проектирования» ПО «САРУС»

Features of the application of Boolean operations in the software module "Design engineering system" by "SARUS"

Н.А. Жарков[1], Н.А. Денисова[1]
[1] Саровский физико-технический институт - филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский ядерный университет "МИФИ" (СарФТИ НИЯУ МИФИ), ул. Духова, д.6, г. Саров, 607186, РФ

N.A. Zharkov[1], N.A.Denisova[1]
[1]Sarov Institute of Physics and Technology - branch of the Federal State Autonomous Educational Institution of Higher Education "National Research Nuclear University "MEPhI" (SarPhTI NRNU MEPhI), Dukhova str., 6, Sarov, 607186, Russian Federation

Аннотация. Настоящая работа посвящена исследованию, направленному на развитие и оптимизацию алгоритмов построения 3D-моделей с применением булевых операций средствами ПМ CAD вновь разрабатываемой программы «Система полного жизненного цикла изделий «Цифровое предприятие» («САРУС»). Булевы операции (БО) – это математические операции, которые выполняются над геометрическими объектами в CAD-системах с целью создания и модификации форм и конструкций, посредством объединения, вычитания, пересечения. БО являются одним из основных инструментов, циклично применяемых в процессе проектирования, что в свою очередь накладывает определенные требования к результатам выполнения БО. Результатом выполнения сравнительного анализа алгоритмов построения с использованием БО (ПМ CAD, «Компас-3D») является выявление отличий, включая определенные ограничения. Поэтому функциональность БО выделена нами в отдельную проблематику исследования. Эта, казалось бы, «простая» теоретико-множительная операция в геометрическом моделировании, показала неоднозначность её применения в сравниваемых CAD-системах.

Ключевые слова. САПР, Программный модуль CAD, 3D-модель, Булевы операции.

Abstract. This work is devoted to a study aimed at the development and optimization of algorithms for building 3D models using Boolean operations using PM CAD tools of the newly developed program "Digital Enterprise Product Full Life Cycle System" ("SARUS"). Boolean operations (BO) are mathematical operations that are performing on geometric objects in CAD systems in order to create and modify shapes and structures by combining, subtracting, and intersecting. BO is one of the main tools used cyclically in the design process, which in turn imposes certain requirements on the results of BO execution. The result of performing a comparative analysis of construction algorithms using BO (PM CAD, Compass-3D) is the identification of differences, including certain limitations. Therefore, we have singled out the BO functionality as a separate research issue. This seemingly "simple" multiplication-theoretic operation in geometric modeling has shown the ambiguity of its application in the compared CAD systems.

Keywords. CAD, CAD software module, 3D model, Boolean operations.

I. Введение

На базе центра цифровых технологий, в сотрудничестве с кафедрами «Технология специального машиностроения» и «Цифровые технологии» организована экспериментальная площадка, работа которой направлена на подготовку пользователей, изучение функциональных возможностей и алгоритмов команд, проведение функционального тестирования разрабатываемой ФГУП «РФЯЦ-ВНИИЭФ» системы полного жизненного цикла изделий «Цифровое предприятие»» (далее по тексту - «САРУС»). В рамках площадки студенты кафедры ТСМ выполняют проблемно-ориентированные исследования, направленные на развитие системы. Программный модуль «Система конструкторского проектирования» (далее по тексту – ПМ CAD), является компонентом «САРУС» и представляет собой CAD-систему.

Настоящая работа посвящена исследованию, направленному на развитие и оптимизацию алгоритмов построения 3D-моделей с применением булевых операций средствами ПМ CAD. Булевы операции (БО) – это математические операции, которые выполняются над геометрическими объектами в CAD-системах с целью создания и модификации форм и конструкций, посредством объединения, вычитания, пересечения [1]. БО

являются одним из основных инструментов, циклично применяемых в процессе проектирования, что в свою очередь накладывает определенные требования к результатам выполнения БО.

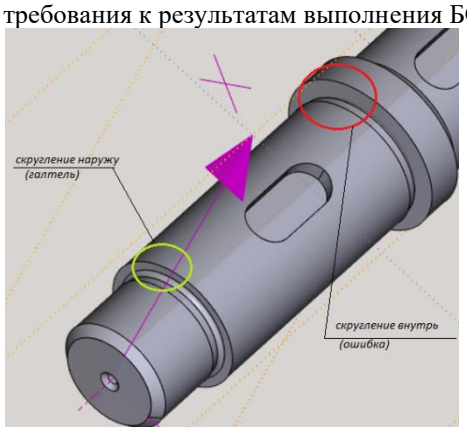


Рис. 1. Результаты операции «Скругление»

Булевы операции в CAD имеют высокую актуальность, поскольку они предоставляют широкий спектр возможностей для моделирования, редактирования и анализа 3D-моделей. Они позволяют создавать сложные объекты из простых примитивов, объединять, разделять и модифицировать их различными способами. Булевы операции также используются для оптимизации дизайна, проверки коллизий и автоматизации процесса проектирования. Можно утверждать, что БО играют ключевую роль в современных системах автоматизированного проектирования.

При тестировании ранних версий ПО CAD СПЖЦ «САРУС» были установлены факты невозможности выполнения некоторых элементов 3D-моделей при объединении этих элементов в одно тело – выполнение Булевой операции. Пример результата такой операции приведен на рис. 1.

Исходя из аналитического обзора отчетов студентов-тестировщиков, мы обратили внимание не некоторые интересные факты применения базовых операций/построений 3D-модели в «САРУС» с использованием алгоритмов функций, заявленных разработчиками. Студентов заинтересовали характерные при применении данного ПМ CAD особенности применения булевых операций, отличающиеся при построении 3D-моделей в сравниваемом ПО «КОМПАС-3D»: нет автоматического объединения элементов модели, построенных с помощью операции «Примитив», в одну деталь, поэтому их надо постоянно объединять с помощью булевой операции в одно тело. Возможно, такой подход является плюсом при изготовлении сложных деталей или в случаях, когда автоматическое объединение в один объект может помешать проектированию [2].

II. Постановка задачи

Исходя из создавшихся предпосылок была определена цель работы - раскрыть значение и возможности применения булевых операций при построении 3D-моделей в программном обеспечении «САРУС», обозначить основные инструменты и алгоритмы, используемые для выполнения этих операций, а также сравнить реализацию булевых операций в ПО «САРУС» и «КОМПАС-3D».

III. Теоретические основы исследования

В мире геометрического моделирования существует некоторое классическое, «стандартное» множество операций, позволяющих конструктору, не привлекая сложной математики, создавать объект почти произвольной формы. Одно из таких множеств - Булевы операции: это математические операции, которые выполняются над геометрическими объектами в CAD-системах (САПР) с целью создания и модификации форм и конструкций. В типичной САПР булевы операции имеют несколько основных функций:

- Объединение (Union): Эта операция объединяет два или более объекта в один. Например, объединение двух прямоугольников может создать новый объемный объект – параллелепипед.
- Вычитание (Subtraction): Эта операция удаляет один объект из другого. Например, можно вычесть отверстие из поверхности, чтобы создать полость.
- Пересечение (Intersection): Пересечение находит общую область двух или более объектов. Например, при пересечении двух цилиндров получится кольцо или усеченный конус.
- Инверсия (Inversion): Инверсия меняет внутреннюю и внешнюю области объекта. Например, инверсия сферы создаст полость внутри сферы.
- Зеркальное отражение (Mirror): Эта операция создает зеркальное отражение объекта относительно заданной плоскости или оси. Например, можно зеркально отразить модель автомобиля, чтобы создать зеркально симметричную версию.

- Угловое скругление (Fillet): Угловое скругление добавляет скругленные углы к объектам, чтобы смягчить их форму. Например, можно добавить скругление к острым углам куба, чтобы получить более гладкую форму.

Эти операции позволяют проектировщикам выполнять сложные геометрические манипуляции и создавать сложные формы, используя основные простые объекты и булевы операции над ними.

В «САРУС» Булева операция определяется как генератор, который принимает два набора тел (твёрдых или листовых) и выполняет над ними регуляризованную булеву операцию [1: 563-564]. Система должна обеспечивать три типа операций [3]: объединение (рис. 2, *а*), вычитание (рис. 2, *б*), пересечение (рис. 2, *в*).

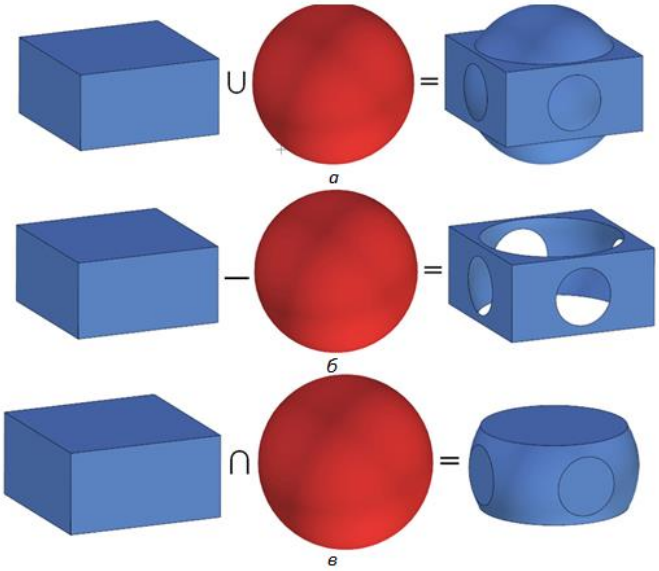


Рис. 2. Типы БО в ПМ CAD: *а* – объединение; *б* – вычитание; *в* - пересечение

Булева операция предназначена для создания нового тела на основе двух уже существующих тел (операндов). В результате создаётся новое тело, являющееся комбинацией исходных тел. Исходные тела могут быть как твёрдыми, так и листовыми. Первый операнд — это тело, которое в результате операции будет преобразовано, второй операнд — это тело, с помощью которого выполняется данное преобразование. Вторых операндов может быть множество.

Система должна поддерживать возможность проведения булевой операции между различными типами операндов (листовых и твёрдых тел). Булевы операции должны позволять:

- интерпретацию листовых операндов как листовых тел, так и как ориентированных полупространств;
- получение результата в требуемой форме, там, где это возможно (переключатель - листовое/твёрдое тело);
- управление точностью операции;
- управление результатом в части выбора результирующих частей операции по различным геометрическим маркерам (точкам, рёбрам, граням);
- управление обработкой касательных случаев при помощи явного указания контактирующих пар объектов и типов контактов;
- производить операцию между явно выбранными подмножествами граней операндов (локальная булева);
- производить автоматическое построение скруглений рёбер, получающихся в результате операции заданным радиусом;
- специально обрабатывать случаи операций с наборами операндов "по образцу" - при неизменности конфигурации пересечения разных экземпляров операндов строить пересечения копированием для оптимизации скорости вычислений (типичный случай вычитание из одного объекта набора (массива) отверстий);
- оставлять операнды в модели в виде неизменных тел;
- идентификацию результирующей топологии на основе идентификаторов исходной геометрии;
- опционально копировать свойства материала из операндов на результат. Это касается как визуальных, так и физических свойств.

Можно определить основной функционал БО при создании 3D-модели - они применяются к трехмерным объектам для управления их формой: с помощью БО можно вырезать в объекте отверстие, получить деталь сложной формы, скруглить углы и ребра и сделать многое другое. Кроме того, БО применяют, чтобы упростить геометрию объекта посредством удаления невидимых поверхностей.

Как отмечают специалисты, «основной задачей булева алгоритма является конструирование границ финальной модели. Иногда эти алгоритмы называют алгоритмами вычисления и слияния границ, что явно

подчеркивает их чисто топологическую натуру. Дело в том, что «сверхзадачей» булева алгоритма является способность работать и выдавать адекватный результат на произвольных пользовательских данных» [4].

Для булевых алгоритмов выделяют три ключевых критерия качества [4]:

- **Точность.** Это первое требование к любой вычислительной схеме. Оно заключается в том, что результат операции должен быть близок к «истинному» результату, причем степень близости (точность) контролируется тем или иным способом. В некоторых случаях, особенно, если операнды соприкасаются, точное вычисление произвести непросто. Другим источником неточности вычислений являются привычные ошибки округления, связанные с ограниченным представлением вещественных чисел в памяти компьютера.
- **Надежность** (устойчивость по входным данным) — важнейший критерий качества булева алгоритма. Требование состоит в том, чтобы при незначительном изменении входных данных результат также изменялся незначительно.
- **Эффективность**, понимаемая в смысле расхода памяти и времени процессора. Булевы операции — традиционно тяжеловесные алгоритмы, которые не могут похвастаться быстрым временем отклика, а потому их интенсивное использование в интерактивном режиме на сегодняшний день практически невозможно. Если геометрия модели нетривиальна (т.е. состоит не из одних плоских зон), то частые вызовы булева алгоритма могут катастрофически сказаться на производительности системы.

Таким образом можно заключить, что значительное множество моделирующих функций (но далеко не все) технически представляют собой не что иное, как специальным образом «заряженные» булевы операции, причем некоторые выполняются автоматически (мы даже не знаем, что реализация какой-либо команды произошла при применении булевой операции), а что-то выполняется вручную, непосредственно по решению конструктора.

Как отмечают исследователи CAD, «Булева операция дает инженеру полную свободу в формообразовании, не привлекая каких-либо кинематических соображений (как, например, операция «заметания» вдоль траектории). Платой за гибкость является повышенная сложность подготовки операндов, так как булев алгоритм не является предметно-ориентированным, т.е. формируя конструктивный элемент посредством БО, инженер должен сам позаботиться о соблюдении всевозможных ограничений моделирования» [4].

IV. Результаты экспериментов. Алгоритмы практической реализации функций БО

Реализацию ряда функций БО рассмотрим на примере построения 3D-модели детали «Опора» (рис. 3).

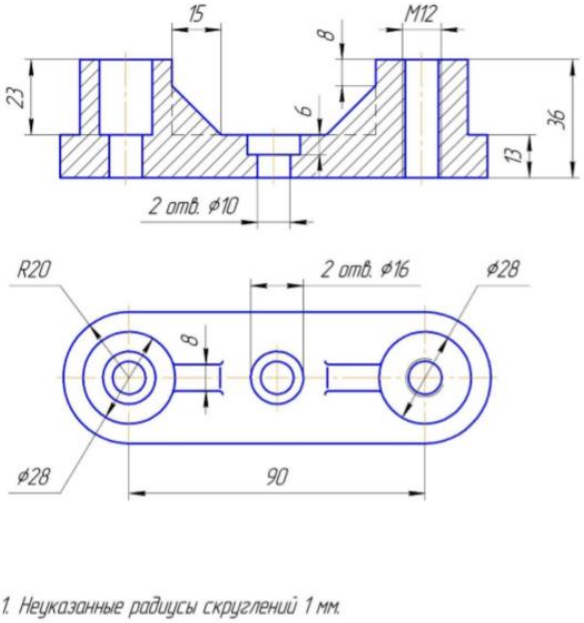
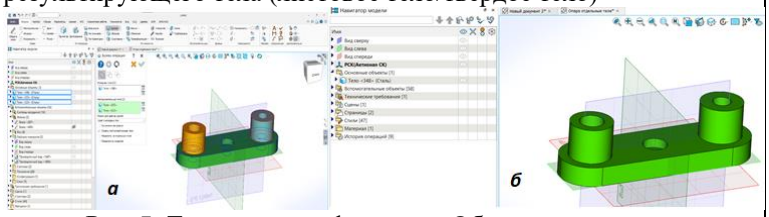
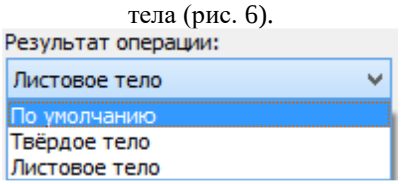
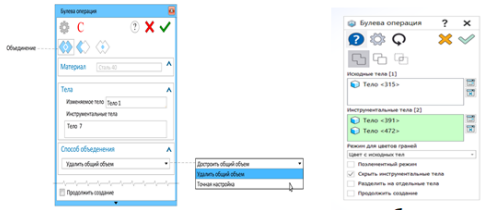
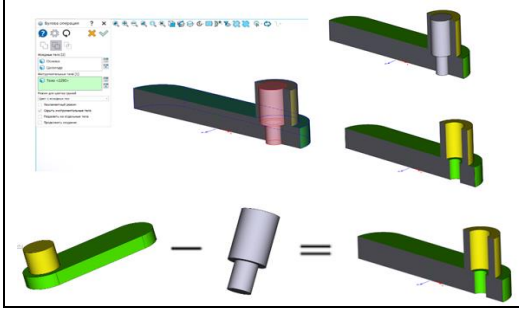


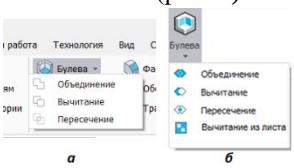
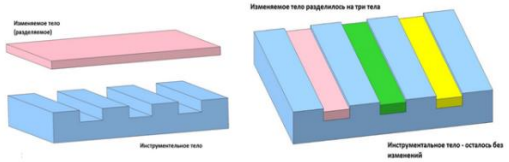
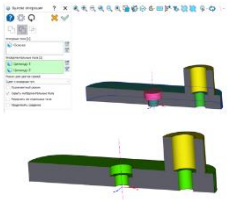
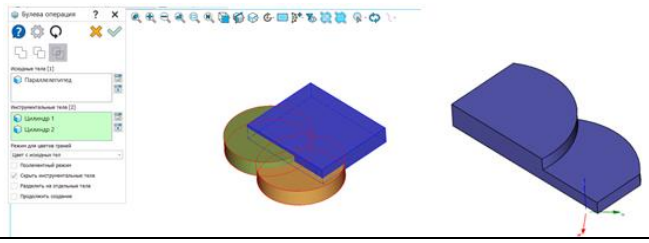
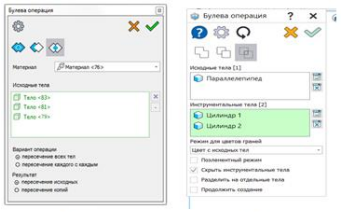
Рис. 4. Чертеж детали «Опора», выполненный в ПО Компас-3D

IV.1 Тестирование работоспособности функций БО

Необходимо определить работоспособность функций при их выполнении по алгоритму разработчика [3]. Итоги реализации функций в соответствии с их описанием сведены в Табл. 1.

ТАБЛИЦА 1
АНАЛИЗ РЕАЛИЗАЦИИ ТРЕБОВАНИЙ К ФУНКЦИЯМ РЕАЛИЗАЦИИ БУЛЕВЫХ ОПЕРАЦИЙ

Описание функции	Итог реализации функции
1. Возможность объединения тел, объединение нескольких тел (три и более) за одну операцию. (рис. 5) Функция должна обеспечивать: объединение двух тел для создания нового тела; объединение между набором тел; выбор результирующего тела (листовое тело/твердое тело)  Рис. 5. Тестирование функции «Объединение тел»: ***а*** – выполнение функции; ***б*** – реализация работы функции	Функция выполняется, но нет опции «Результат операции», поэтому нет возможности выбора результирующего тела (рис. 6).  Рис. 6. Окно для выбора типа результирующего тела из ПЗ разработчика
2. Объединение с удалением объёма Функция должна обеспечивать: - объединение двух и более тел (элементов модели) для создания нового тела с возможностью удаления объема(ов), образованного при пересечении тел; - выбор объема(ов) для удаления по принадлежности к телам.	По факту функция не реализована, нет опции «Способ объединения» (рис. 7).  Рис.7. Диалоговое окно БО: ***а*** – В ПЗ разработчика; ***б*** – В ПО «Сарус»
3. Объединение с достроением объёма Функция должна обеспечивать: - объединение двух и более тел (элементов модели) для создания нового тела с возможностью достроения объема(ов), образованного при пересечении тел; - выбор объема(ов) для достроения по принадлежности к телам; - отображение в дереве построения с возможностью редактирования.	По факту функция не реализована, нет опции «Способ объединения» (рис. 7)
4. Возможность одновременно удалять объёмы и достраивать объёмы Функция должна обеспечивать: - удаление объемов или заполнение замкнутых объемов, полученных в результате пересечения тел при их объединении.	По факту функция не реализована, нет опции «Способ объединения» (рис. 7)
5. Возможность продолжения граней при удалении объёма Функция должна обеспечивать: - выбор граней в операции объединение и достраивание их до поверхности и на расстояние; - отображение фантома измененной грани; -удаление объема с учетом новой геометрии тела.	По факту функция не реализована, нет опции «Способ объединения» (рис. 7)
6. Возможность одновременного вычитания из нескольких тел  Функция должна обеспечивать вычитание одного тела из группы тел за одну операцию Рис. 8. Результат реализации работы функции	Функция выполняется, но нет опции «Результат операции», поэтому нет возможности выбора результирующего тела (рис. 8).

Описание функции	Итог реализации функции
7. Вычитание твердых тел из листового Функция должна обеспечивать: - выбор тел для операции и создание листового тела как результат вычитания нескольких твердых тел (два и более) из листового тела; - корректное построение линий среза листового тела. Результирующее тело по умолчанию должно иметь материал исходного листового тела, с возможностью определения другого материала; - выделение тел в произвольном порядке, система сама должна определить исходное листовое тело по его типу.	Функция не реализована, нет опции «Вычитание из листа» (рис. 9).  *а* *б* Рис. 9. Диалоговое окно БО со списком свойств: ***а*** – В ПО «Сарус»; ***б*** – В ПЗ разработчика
8. Возможность разделения тела после вычитания Функция должна обеспечивать возможность разделять тело на несколько после вычитания согласно геометрии (рис. 10)  Рис. 10. Результат работы функции	Функция выполняется
9. Вычитание из тела одного или нескольких тел  Функция должна обеспечивать: - вычитание одного тела из другого; - вычитание между наборами тел, часть которых может быть выбрана как первые операнды, а другие тела как вторые (рис. 11) Рис. 11. Результат работы функции	Функция выполняется
10. Возможность построения тел пересечения. Пересечение нескольких тел (три и более) за одну операцию (рис.12) Функция должна обеспечивать: - выбор тел для операции; - создание нового тела как результат пересечения исходных тел; - создание нового тела как результат пересечения копий исходных тел. Реализация возможна в двух вариантах: пересечение всех тел; пересечение каждого с каждым (в данном варианте строится пересечение тел попарно, затем результаты объединяются); - корректное построение линий пересечения тел; - назначение новому телу по умолчанию материала первого выбранного тела, с возможностью определения другого материала. Рис. 12. Результат работы функции 	Функция реализована не полностью. Нет выбора вариантов реализации функции (рис. 13).  *а* *б* Рис. 13. Диалоговое окно БО: ***а*** – В ПЗ разработчика; ***б*** – В ПО «Сарус»
11. Возможность удаления/сохранения исходных тел и тел инструментов для всех операций Функция должна обеспечивать сохранение или удаление выбранных исходных тел и тел инструментов при выполнении булевых операций	Функция не реализована, нет пункта «Сохранение исходных тел после операций».

2.2 Применение БО при создании 3D-модели детали «Опора»

Используя Булевы операции, удалось построить 3D-модель детали «Опора» (рис. 14). Эту деталь можно отнести к II группе конструктивной сложности – простые детали. Необходимо отметить, что при построении 3D-модели детали «Опора» со сравнительно простой геометрией трудностей не возникло. Однако вероятно, некоторые функциональные возможности, применяемые в других САПР, а здесь в настоящее время отсутствуют, облегчили бы процесс построения и, с большой вероятностью, были бы необходимы при построении деталей со сложной геометрией.

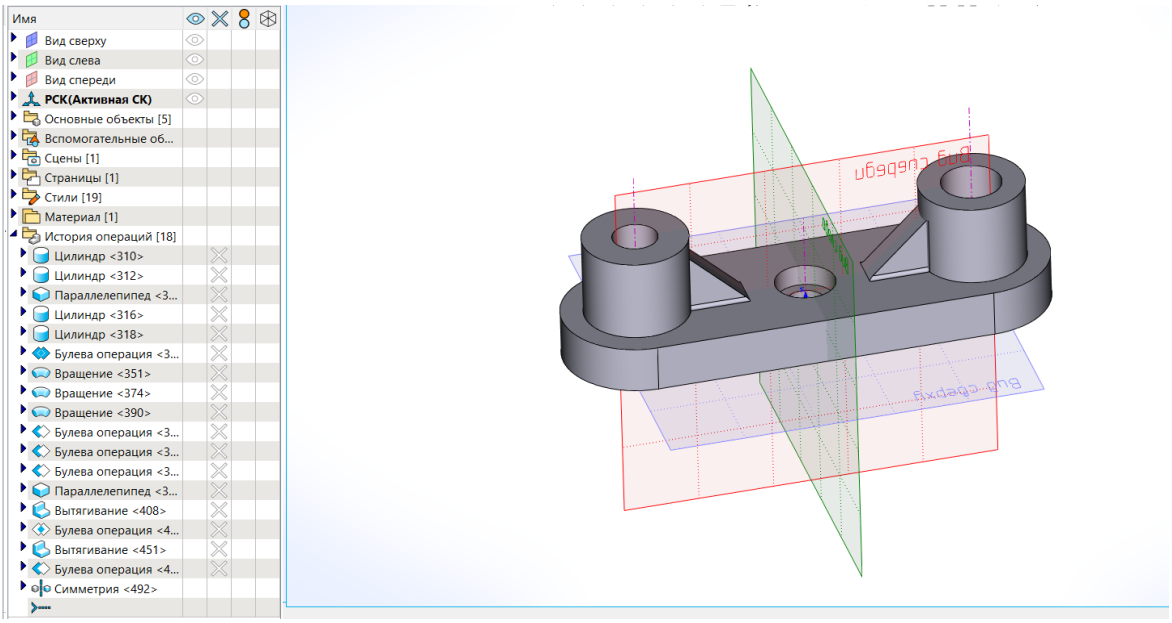


Рис. 14. 3D-модель детали «Опора», созданная в ПО «САРУС» с применением БО

V. Обсуждение результатов

После работы с программой «САРУС» и изучения особенностей применения булевых операций, можно сделать вывод о том, что программа потенциально пригодна для работы, и особенность применения функций БО по сравнению с КОМПАС-3D [5] вызывает интерес. Это отличает «САРУС» от других программ, где БО выполняются автоматически. Кроме того, программа предоставляет возможность детального контроля над процессом создания детали, что может быть полезно для инженеров, работающих со сложными проектами.

В сравнении с «КОМПАС-3D», где БО выполняются автоматически, в «САРУС» при построении нестандартной, неформальной детали, программа заставляет нас подумать и самим принять решение, когда их использовать. При работе с простыми деталями между программами разница незначительна, но при работе с деталями сложной геометрии, возможно, будет проще работать с набором тел. ПМ CAD «САРУС» обеспечивает в дальнейшем внесение корректировок в случае изменений. Однако надо понимать, что при этом могут возникнуть сложности с выполнением других операций.

Большой функционал БО в системах CAD безусловно предоставляет конструкторам больше возможностей для творчества. Булевы операции, такие как объединение, пересечение и вычитание, позволяют создавать сложные трехмерные объекты из более простых форм, что может быть полезно при проектировании различных деталей и конструкций. Например, инженер-конструктор может использовать булевы операции для создания сложных деталей из нескольких простых компонентов. Это может включать в себя создание корпуса автомобиля из листового металла или проектирование пространственной рамы из различных металлических профилей.

VI.Выводы и заключение

Большой набор БО позволяет конструкторам экспериментировать с различными формами и структурами, что может привести к более эффективному и оригинальному решению. Это особенно важно в условиях жесткой конкуренции на рынке, когда необходимо предлагать инновационные продукты и услуги.

В целом функционал БО делает CAD-систему в СПЖЦ «САРУС» незаменимым инструментом для конструктора, позволяя ему проявить свое творчество и умение находить нестандартные решения. Это способствует развитию новых технологий и улучшению качества продукции.

Сделанный вывод коррелируется с мнением специалистов, что в CAD функциональность БО дает пользователю полную свободу в формообразовании, не привлекая каких-либо кинематических соображений, она

больше рассчитана на творчество инженера конструктора. Их функционал очень гибок, инженер использует его для достижения нужных результатов. Он может получить то, что хочет, потому что автоматическое выполнение операций может не дать нужного результата [4].

Гибкость же требует более сложной подготовки операндов, поскольку булев алгоритм не является предметно-ориентированным. Инженеру, создающему элемент с помощью БО, приходится самому заботиться о соблюдении всех возможных ограничений моделирования. Также данная программа требует от конструктора иного подхода к работе. Нужно не просто создать какую-либо конструкцию, но и тщательно продумать все детали, поскольку работа конструктора оказывает влияние на последующие этапы программы, зафиксированные на общей платформе. В данной программе инженер-конструктор должен знать не только систему автоматизированного проектирования, но и задумываться о дальнейшем маршруте детали, поскольку может оказаться, что созданная модель будет непригодна для дальнейшей работы.

Источник финансирования. Благодарности

Работа выполнена в соответствии:

- с Рабочим учебным планом по направлению 15.04.05- Конструкторско-технологическое обеспечение машиностроительных производств;
- с Договором от 25 марта 2024 г. № 96-2024/244 на выполнение научно- исследовательской работы по теме: «Проведение научных исследований в интересах ФГУП «РФЯЦ-ВНИИЭФ» с использованием результатов для актуализации образовательных программ подготовки специалистов с высшим профессиональным образованием для ФГУП «РФЯЦ-ВНИИЭФ», Задача 10 «Исследование программно-аппаратных решений в области цифровых технологий, возможностей их отраслевого применения, разработка и версионное тестирование компонентов полностью защищенной системы полного жизненного цикла среднего класса СПЖЦ V3 «Сарус».

Научный руководитель Денисова Наталья Александровна, СарФТИ-НИЯУ МИФИ, кафедра Технологии специального машиностроения

Авторы благодарят:

- Кафедру цифровых технологий СарФТИ-НИЯУ МИФИ под руководством Кривошеева О.В., к.т.н., заведующий кафедрой, зам. директора ФГУП «РФЯЦ-ВНИИЭФ» по технологиям полного жизненного цикла - директор института цифровых технологий (ИТЦ) - главный конструктор СПЖЦ;

Список литературы

6. Комплекс программ в защищенном исполнении «Система полного жизненного цикла изделий «Цифровое предприятие»; Программный модуль «Система конструкторского проектирования». ПЗ ТП 07623615.00423-04 81 01-1 часть 1. 2021г. 889 с.
7. Отчёт о НИР «Исследование цифровых технологий и возможностей их отраслевого применения, разработка и тестирование компонентов СПЖЦ V3» (Договор на выполнение НИР от 2 марта 2023 г. № 96-2023/203, Задача 10). Инв. номер НО/38-278.10-2023-О от 15.11.23. С.8-27
8. Комплекс программ в защищенном исполнении «Система полного жизненного цикла изделий «Цифровое предприятие»; Программный модуль «Система конструкторского проектирования». ПЗ ТП 07623615.00423-04 81 01-6 часть 6. 2021г. 720 с.
9. Многообразная геометрия: https://quaoar.su/blog/page/bulevy-operacii-v-b-rep-vvedenie
10. Голованов Н.Н. Геометрическое моделирование. Учебное пособие. – М.: КУРС: ИНФРА-М, 2016. 400 с.

УДК: 004.925.8
DOI: 10.25206/978-5-8149-3873-2-2024-813-819

Применение методов геометрического моделирования и средств компьютерной графики при формировании геометрического облика вертолета
Application of geometric modeling methods and computer graphics tools in the formation of the geometric appearance of the helicopter

А.Н. Дергачев [1]
[1] *Московский авиационный институт (Национальный исследовательский университет), Волоколамское шоссе, д. 4, г. Москва, 125993, Россия*

A.N. Dergachev [1]
[1] *Moscow Aviation Institute (National Research University), 4, Volokolamskoe shosse, Moscow, 125993, Russia*

Аннотация. В статье представлен методический подход визуализации внешних факторов, оказывающих влияние на геометрический облик проектируемого вертолета. Визуализация различных ограничений, определяемых условиями эксплуатации вертолета, позволяет на ранних стадиях проектирования предусмотреть необходимые конструктивные решения и уменьшить трудоемкость выполняемых работ по обслуживанию в обособленных случаях. Данный подход также применим при компоновке кабины экипажа, пассажирского салона и отсеков с оборудованием. Полученная система геометрических ограничений формирует дополнительные требования при проектировании фюзеляжа вертолета, агрегатов и систем, в т.ч. геометрические характеристики которых требуют уточнения по результатам проведенных расчетов и моделирования. Унификация предъявляемых требований к проектируемым вертолетам позволяет сформировать параметрическую модель внешних факторов, которая позволит существенно сократить сроки проектирования.

Ключевые слова: Вертолет, проектирование, геометрические ограничения, компоновка, компьютерная графика.

Abstract. The article presents a methodological approach to visualization of external factors influencing the geometric appearance of the designed helicopter. Visualization of various limitations determined by the operating conditions of the helicopter makes it possible to provide the necessary design solutions at the early stages of design and reduce the complexity of the maintenance work performed in isolated cases. This approach is also applicable to the layout of the crew cabin, passenger compartment and equipment compartments. The resulting system of geometric constraints forms additional requirements for the design of the helicopter fuselage, assemblies and systems, including geometric characteristics that require clarification based on the results of calculations and modeling. The unification of the requirements for the designed helicopters makes it possible to form a parametric model of external factors, which will significantly reduce the design time.

Keywords: Helicopter, design, geometric constraints, layout, computer graphics.

Описание системы внешних ограничений

Совокупность факторов, оказывающих влияние на геометрический облик проектируемого вертолета можно условно разделить на несколько категорий, определяемых:

–заданными летно-техническими характеристиками;
– целевым назначением;
– требованиями транспортабельности;
– удобством технического обслуживания;
– антропометрическими показателями членов экипажа и пассажиров;
– условиями эксплуатации;
– нормами летной годности;
– прочими руководящими требованиями.

Заданные летно-технические характеристики проектируемых вертолетов в первую очередь оказывают влияние на диаметр несущего винта, размеры двигателей и редуктора, объемы топливных баков и геометрические характеристики хвостового оперения.

К целевому назначению можно отнести количество перевозимых пассажиров, массогабаритные характеристики транспортируемого груза в транспортной кабине, необходимость применения морских спасательных костюмов, съемное целевое оборудование, размещаемое как снаружи, так и в специальном отсеке вертолета, и т.д.

Обеспечение транспортабельности при формировании геометрического облика вертолета в первую очередь предусматривает уменьшение габаритных характеристик для обеспечения размещения и погрузки в транспортных кабинах самолетов и вертолетов, на автомобильных и железнодорожных платформах и т.д. В процессе выполнения проработок на некоторых типах вертолетов для обеспечения погрузки вертолета в транспортную кабину самолета отмечена необходимость демонтажа или складывания хвостовой балки или хвостового оперения. На ранних стадиях проектирования данные особенности позволят обеспечить быстросъемные соединения как элементов конструкции, так и систем (электроснабжение аэронавигационных огней, антенн, привод рулевого винта и т.д.), а также исключить необходимость проведения процедур, требующих обособленных технологических условий (например, удаление старого и нанесение нового герметика).

Ограничения, формируемые требованиями к удобству технического обслуживания, в большей степени определяются обеспечением монтажно-демонтажных работ (с учетом заданной периодичности и требуемых зазоров), доступом к лючкам, разъемам и т.д. Требования к свободному пространству при работе с инструментом (в т.ч. заданному разработчиком агрегата при формировании электронной модели) позволит определить очередность монтажно-демонтажных процедур или установить запрет на размещение оборудования в заданной области.

Применение актуализированных среднестатистических антропометрических показателей членов экипажа и пассажиров позволят оптимизировать размещение оборудования и органов управления вертолетом, уточнить требуемые размеры кресел и расстояния между ними, расположение подножек и рукояток. Данный вопрос становится наиболее актуальным при большом количестве изготавливаемых вертолетов, в т.ч. различных модификаций.

К факторам, оказывающим влияние на геометрический облик вертолета, в части условий эксплуатации можно отнести следующее:

– посадка на площадки ограниченных размеров (в т.ч. на морских судах);
– посадка на неподготовленные площадки с уклоном;
– посадка с использованием лыж;
– унифицированные средства технического обслуживания;
– размещение в ангарах и укрытиях;
– сопряжение с системами буксировки, подвески съемного оборудования и т.д.;
–применение аварийных баллонет, спасательных лебедок и т.д.

Требованиями действующих нормативных документов регламентируются требования к конструкции вертолета, характеристикам систем и агрегатов, эргономики и т.д.

Совокупность указанных факторов формируют ограничения геометрического облика вертолета, обеспечивающего предъявляемым требованиям. Простота формирования системы геометрических ограничений позволяет на самых ранних этапах проектирования не только учесть особенности конструкции и условий эксплуатации, но и сформулировать технические требования к агрегатам и системам, в т.ч. допустимые габаритные характеристики, расположение такелажных узлов, крепежных элементов, разъемов и т.д.

Визуализация системы внутренних и внешних ограничений

Для каждого из элементов конструкции вертолета (рисунок 1) формируются индивидуальные внешние и внутренние ограничения, формирующие диапазон допустимых габаритных характеристик. При этом, при заданных предельных граничных значениях общих габаритных характеристик, могут возникать случаи, при которых требуется изменение геометрических характеристик смежных элементов конструкции и агрегатов.

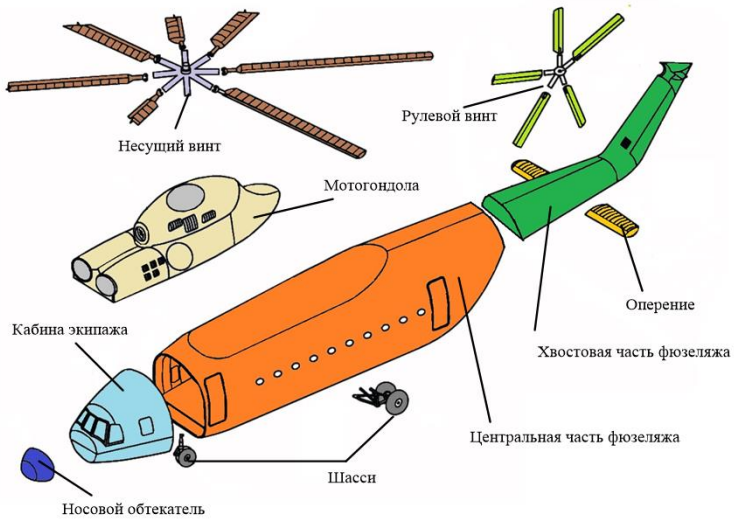


Рисунок 1 – Декомпозиция конструкции вертолета

В первом приближении, на начальных этапах формирования системы геометрических ограничений, элементы конструкции представляются в виде простейших геометрических форм (рис. 2).

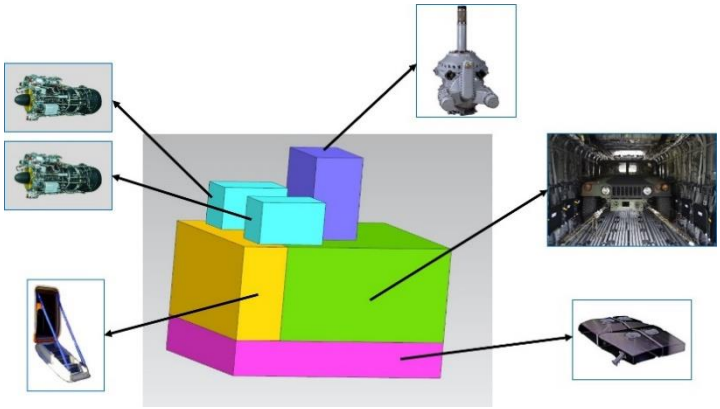


Рисунок 2 – Представление элементов конструкции и агрегатов в виде простейших геометрических форм

Далее, каждый из элементов приобретает более сложную форму, как исходя из условий аэродинамических свойств, так и по мере дополнения системы геометрических ограничений дополнительными элементами, а также выполнения различных расчетов.

Геометрический облик вертолета можно представить в следующем виде:

$$L = f(m_0) + f(U_{нс}) + f(U_{пн}) + f(U_{об}) + f(U_{нс}), \quad (1)$$

$$^{a}B = f(m_0) + f(U_{нс}) + f(U_{пн}) + f(U_{об}) + f(U_{ш}) + f(U_{лх}), \quad (2)$$

$$^{b}H = f(m_0) + f(U_{нс}) + f(U_{пн}) + f(U_{об}) + f(U_{ш}) + f(U_{топл}) \quad (3)$$

где L – длина вертолета, B – ширина вертолета, H – высота вертолета, m_0 – максимальная масса вертолета, $U_{нс}$ – характеристики несущей системы, $U_{пн}$ – характеристики полезной нагрузки, $U_{об}$ – характеристики размещаемого оборудования, $U_{ш}$ – параметры шасси, $U_{лх}$ – заданные летно-технические характеристики, $U_{топл}$ – характеристики топливных баков.

Характеристики несущей системы $U_{нс}$ оказывают существенное влияние на геометрический облик проектируемого вертолета. В первую очередь влияние оказывает применяемая схема – одновинтовая схема с рулевым винтом или винтом в кольце, соосная, продольная или поперечная схема.

Влияние устанавливаемого на геометрический облик вертолета в первую очередь рассматривается в части внешнего оборудования (грузовые лебедки, датчики, антенны и т.д.), а также оборудование, размещаемое в носовом обтекателе (локатор) и в кабине экипажа. Габаритные характеристики кабины экипажа определяются количеством членов экипажа, размещаемым оборудованием и обеспечением требуемого обзора из кабины экипажа [1] (рисунок 2).

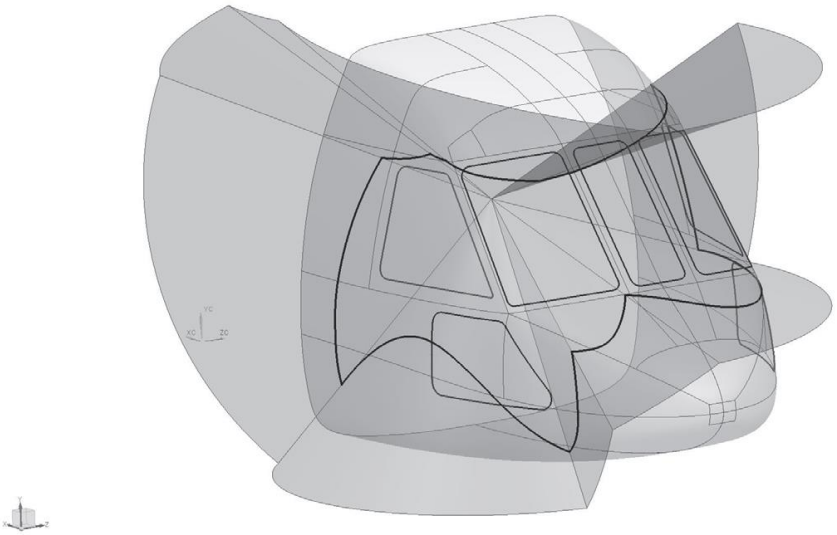


Рисунок 2 – Моделирование диаграммы обзора из кабины экипажа

Геометрические характеристики центральной части фюзеляжа можно представить в виде

$$^{c}U_{\text{цчф}} = f(n_{\text{пасс}}) + f\left(U_{\text{гр}}\right) + f\left(U_{\text{загр}}\right) + f(U_{\text{об}}) + f(U_{\text{топл}}), \quad (4)$$

где $\mathrm{U}_{\text{цчф}}$ – геометрические характеристики центральной части фюзеляжа, $\mathrm{U}_{\text{гр}}$ – массогабаритные характеристики груза, $\mathrm{U}_{\text{загр}}$ – характеристики проемов для загрузки/выгрузки грузов и посадки/высадки пассажиров (в т.ч. аварийные выходы).

При определении характеристик груза $U_{\text{гр}}$ учитываются габаритные характеристики наибольшего типа груза (например контейнер, автомобиль и т.д.), а также приведенные габаритные характеристики груза исходя из максимально допустимой массы, высоты грузовой кабины и допустимой нагрузки на пол грузовой кабины.

Характеристики проемов и дверей $\mathrm{U}_{\text{загр}}$ определяют не только возможность загрузки грузов исходя из габаритных характеристик, размещения пассажирских кресел, но и накладывают ограничения по размещению оборудования, например в районе хвостовой балки при загрузке крупногабаритных грузов через рампу.

Геометрические характеристики топливных баков напрямую зависят от их размещения. При размещении топливных баков под полом транспортной (грузовой) кабины доступная площадь для размещения топливных баков определяется с учетом площади пола транспортной (грузовой) кабины. При необходимости размещения дополнительного оборудования (например, элементов внешней грузовой подвески) доступная площадь для размещения топливных баков уменьшается, что приводит к увеличению общей высоты центральной части фюзеляжа при постоянных характеристиках пассажирской (грузовой) кабины. Размещение топливных баков снаружи фюзеляжа существенно увеличивают ширину вертолета, что может потребовать упрощения их монтажа/демонтажа при жестком ограничении ширины вертолета при транспортировке автомобильным или железнодорожным транспортом. Размещение топливных баков над грузовой кабиной может потребовать увеличения общей высоты центральной части фюзеляжа, а также изменения геометрии мотогондолы, что в свою очередь окажет влияние на общие аэродинамические качества проектируемого вертолета.

При проектировании вертолета также учитывается возможность его транспортирования в частично разобранном виде [2] (рисунок 3).

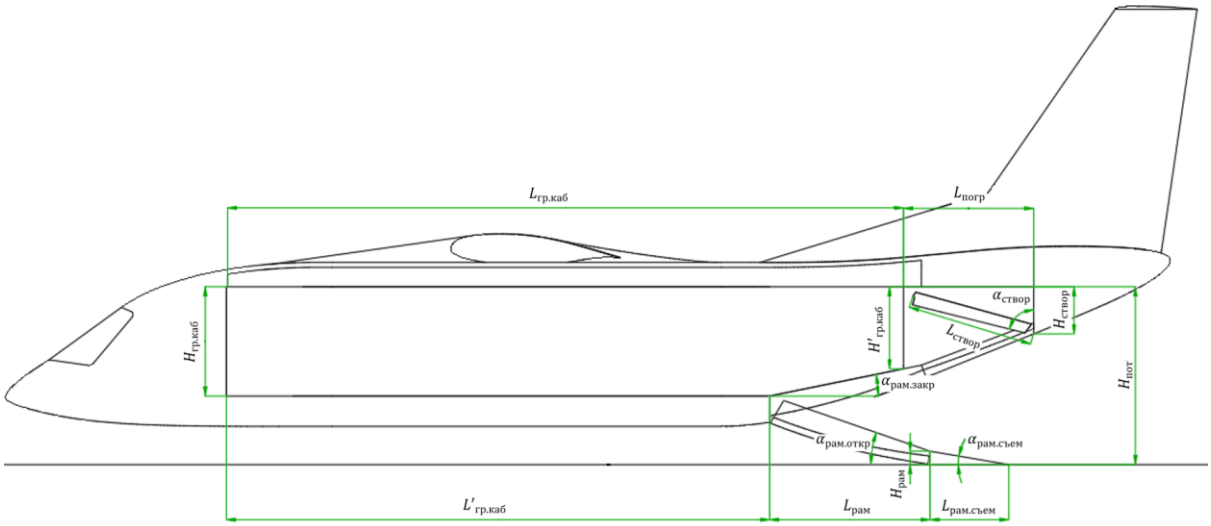


Рисунок 3 – Параметры грузовой кабины транспортного самолета

Моделирование процесса погрузки вертолета в частично разобранном виде позволяет не только определить объем подготовки вертолета к транспортированию (демонтаж лопастей несущего винта, редуктора и т.д.), но и сформировать требования к конструкции хвостовой балки (упрощенная процедура демонтажа или складывания при необходимости) и размещению оборудования (антенны и т.д.). По результатам геометрического моделирования транспортной кабины самолета (рис. 4) в первом приближении сформированы геометрические характеристики подготовленного к транспортированию проектируемого вертолета (поз. 11) с учетом угла наклона «земли стояночной) (поз. 1) относительно пола грузовой кабины самолета, геометрических размеров грузовой кабины транспортного самолета (поз. 2), размеров проема (поз. 3), размещения грузоподъемных устройств (поз. 4) и диапазона их размещения (поз. 5, 6), угла наклона рампы (поз. 7), требуемых зазоров при размещении транспортируемого вертолета перед рампой (поз. 8) и проходных зазоров (поз. 9) и необходимого клиренса вертолета (поз. 10).

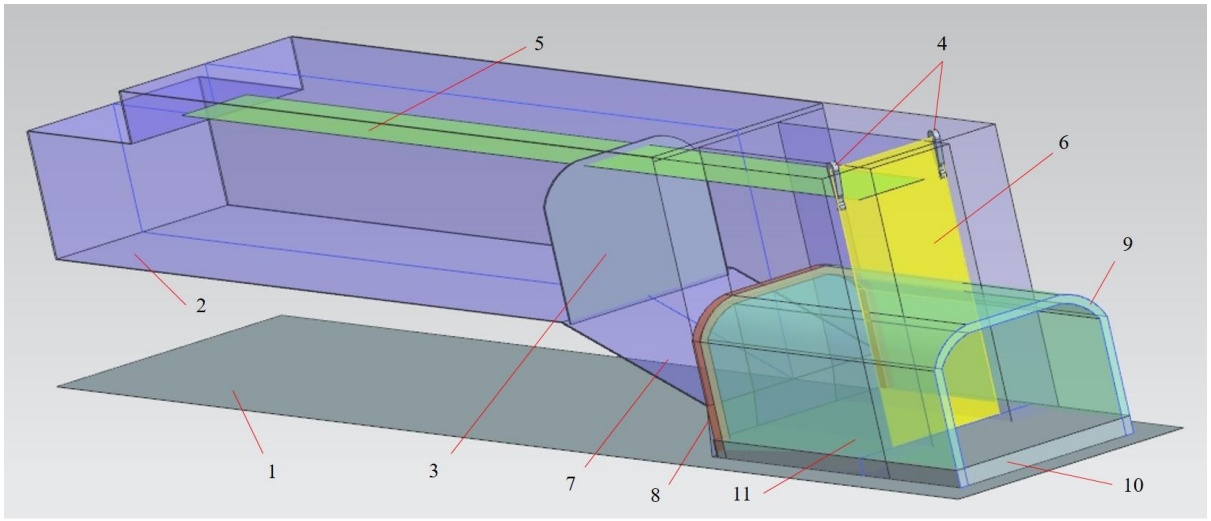


Рисунок 4 – Параметры грузовой кабины транспортного самолета

В случае базирования проектируемого вертолета на кораблях необходимо учитывать и размеры взлетно-посадочной площадки (ВППл) [3], ангара, конструктивные особенности палубы и систем буксировки вертолета в ангар. Размещение вертолета в ангаре предполагает реализацию решений, обеспечивающих уменьшение габаритов вертолета (складывание лопастей несущего винта, хвостовой балки и т.д.). При этом, в качестве внешних ограничений выступают габариты ангара и его проема, характеристики системы буксировки и место подстыковки к вертолету, наличие низких препятствий, а также необходимость выполнения отдельных работ по обслуживанию вертолета и т.д.

В процессе предварительного моделирования геометрического облика вертолета при корабельном базировании (рисунок 5) сформированы требования к габаритам вертолета (поз. 7, 10) с учетом общие габариты вертолета исходя из размеров ВППл (поз. 1, 3), возможное смещение от центра ВППл при посадке (поз. 2), комингс (поз. 4), требуемые зазоры (поз. 5, 6, 8, 9, 15, 17), изменение клиренса вертолета (с колесным шасси) (поз. 12), необходимости уменьшения габаритов элементов конструкции вертолета (поз. 11) и характеристик системы буксировки (поз. 14, 16).

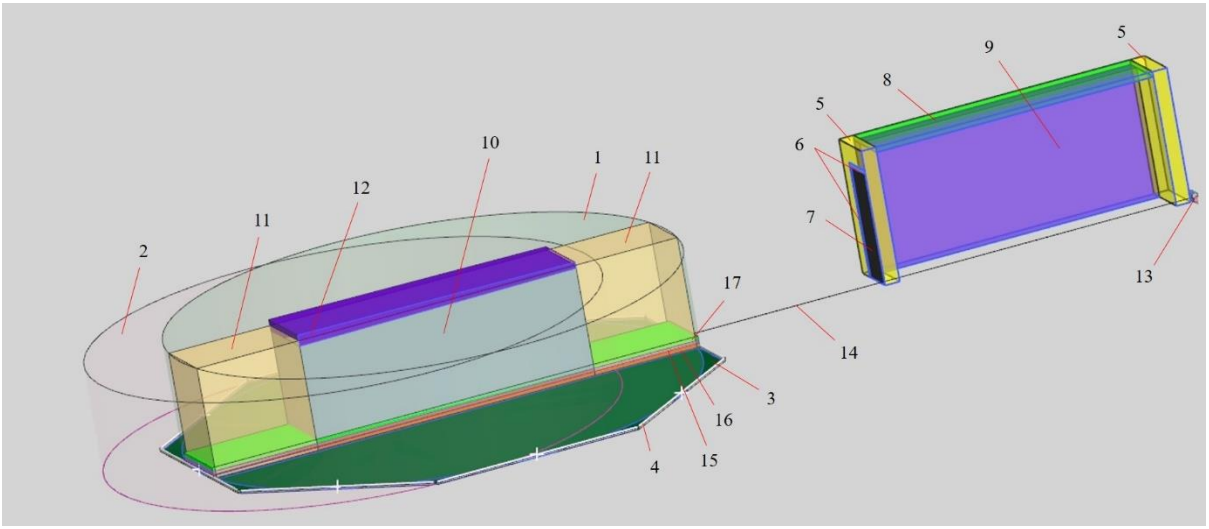


Рисунок 5 – Моделирование ограничений при корабельном базировании

Формирование геометрического облика во втором приближении осуществляется путем объединения результатов частного моделирования (рис. 6) (на примере корабельного базирования (поз. 1, 3), транспортировки самолетом (поз. 2) и построения проекционных связей (поз. 4)).

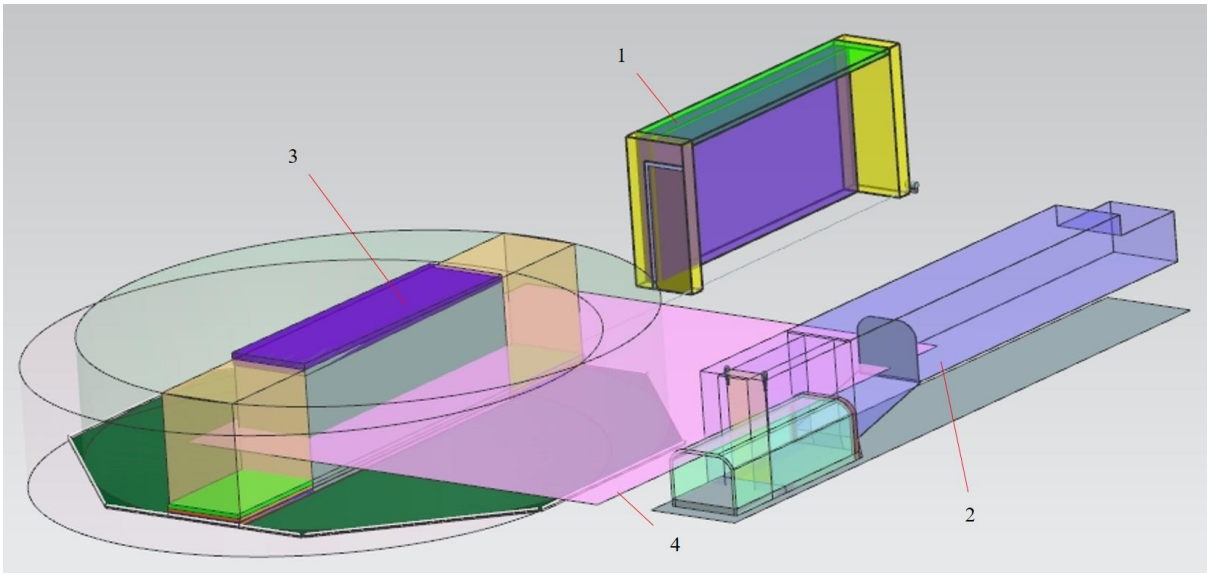


Рисунок 5 – Объединение результатов частного моделирования

В третьем приближении геометрический облик вертолета уточняется с учетом аэродинамических и прочих расчетов, а также проведения оценки возможности и удобства обслуживания (рис. 6).

Заключение

Процесс формирования геометрического облика проектируемого вертолета является совокупностью результатов частного моделирования с учетом предъявляемых к вертолету требований, а также внешних и внутренних ограничивающих факторов. Технические возможности современных систем автоматизированного

проектирования (САПР), различные методики автоматизированной компоновки оборудования, а также применение статистических данных [4-7] в процессе проектирования позволяют существенно сократить сроки формирования облика вертолета в первом приближении. Сформированная система геометрических ограничений позволяет не только уточнить предъявляемые требования к массогабаритным характеристикам комплектующих изделий, но и существенно уменьшить количество итераций компоновок оборудования и систем в процессе взаимной увязки электронного макета вертолета, что положительным образом сказывается на общих сроках проектирования.

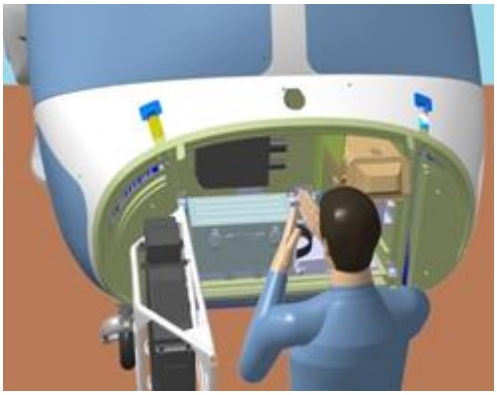


Рисунок 6 – Моделирование процедур технического обслуживания вертолета

Список источников

Гельвиг М.Ю., Метод построения диаграмм обзора внекабинного пространства пилотом летательного аппарата // Вестник Московского авиационного института. 2016. Т. 23 № 3. С. 129–135.

Дергачев А.Н., Геометрический расчет возможности погрузки техники в грузовые кабины самолетов и вертолетов // Приволжский научный журнал. 2024. № 1 (69). С. 269-277.

Приложение 14 к Конвенции о международной гражданской авиации, т. 2 «Вертодромы», 5-е изд. – Монреаль: ИКАО, 2020

Кривцов В.С., Карпов Я.С., Лосев Л.И. Проектирование вертолетов. Харьков.: Нац. Аэрокосм. Ун-т «Харьк. Авиаци. Ин-т», 2003. 344 с.

Проскурин В.Д. Расчет параметров вертолета на этапе предварительного проектирования: учебное пособие. Оренбург: ОГУ, 2014. 199 с.

Дудник В.В. Конструкция вертолетов. – Ростов н/Д: издательский дом ИУИ АП, 2005. – 158 с.

Тищенко М.Н., Некрасов А.В., Радин А.С. Вертолеты. Выбор параметров при проектировании. М., Машиностроение, 1977. 368 с.

УДК 004: 921
DOI: 10.25206/978-5-8149-3873-2-2024-819-826

Цифровые инструменты цветового конструирования в компьютерной графике
Digital tools for color design in computer graphics

Ю. В. Дианова[1], Л. С. Тарасова[2]
[1,2] ФГАОУ ВО «Пермский национальный исследовательский политехнический университет», Пермь, РФ

Yu. V. Dianova[1], L. S. Tarasova[2]
[1, 2]Perm National Research Polytechnic Universit, Perm, Russia

Аннотация. Технологии компьютерного моделирования диктуют необходимость привлечения смежных цифровых ресурсов в процессе работы. На этапе разработки дизайна продукта важно подготовить цветовое оформление и выполнить его в электронном формате. Правила и законы цветовых сочетаний, применяемые в рисунке на бумаге, трансформируются в среде графических редакторов. Онлайн-инструмент для подбора цветов и генерации цветовых схем ColorScheme позволяет погрузиться в цифровое пространство цветовых сочетаний, опробовать различные комбинации при разработке дизайн-проекта, понять логику работы с классическим цветовым кругом – опорной схемой дизайнеров и художников. В процессе обучения студентов по программе профессиональной переподготовке «Информационные технологии в дизайне» данный ресурс используется для цветового анализа и разработки колористического решения интерьера общественного транспорта.

***Ключевые слова*:** колористика, цветоведение, цвет, растровая графика, цветовые модели, компьютерная графика

Abstract. Computer modeling technologies dictate the need to involve related digital resources in the process of work. At the stage of product design development, it is important to prepare a color scheme and present it in electronic format. The rules and laws of color combinations applied in the drawing are transformed in the environment of graphic editors. The online color matching and color scheme generation tool ColorScheme allows you to immerse yourself in the digital space of color combinations, try out various combinations when developing a design project, and understand the logic of working with the classic color circle - the reference scheme of designers and artists. In the process of training students under the professional retraining program "Information Technologies in design", this resource is used for color analysis and the development of a coloristic solution for the interior of public transport.

Keywords: coloristics, color science, color, raster graphics, color models, computer graphics

I. Введение

В ходе разработки и создания любого продукта, будь то электронная модель или готовое изделие, этапу презентации готового образца уделяется особое внимание. Подготовка к демонстрации в электронном формате предполагает выбор выгодных ракурсов, разработку бэкграундов, рендеринг модели, оформление фирменного стиля, подготовку полиграфических материалов и пр. [1] Возможности графического дизайна и компьютерной графики предоставляют необходимые инструменты для достижения запланированных результатов.

Подобные практики востребованы и в обучении студентов [2]. Для высшего образования на примере дисциплин, связанных с компьютерным моделированием, проектированием, презентация проектной разработки имеет значение при оценке работы, перспектив ее внедрения, коммерциализации. Практические навыки и приемы в дизайне помогают грамотно организовать данный процесс [3]. Применение знаний композиции, цветоведения, подбор шрифта определяют оформление готового изделия [4].

II. Постановка задачи

В творческие виды деятельности, традиционно выполняющиеся вручную, в которых проектное решение является продуктом творческой мысли, сегодня проникает искусственный интеллект и предоставляет возможность цифровых методов обработки всех видов информации. Исключением не стала и художественная сфера [5]. Генерация всевозможных сюжетных вариантов, композиций, стилей на сегодняшний день является общедоступным инструментом. Замена произведения авторской мысли цифровыми вариациями «на тему» не дает возможности осознать личностный вклад творца, ощутить его настроение и разгадать идеи, заложенные в

его труде. Проблема соотношения вклада цифровых ресурсов и творческого гения должна быть обозначена, а ее решения проработаны.

В качестве предложения выскажем позицию о вспомогательной роли цифровых программных продуктов в деятельности художника, дизайнера, колориста. Прикладной задачей в таком случае ставится осознанный поиск вариантов и рациональный вклад электронных ресурсов в проектный результат с возможностью снижения временных и производственных затрат в сфере компьютерной графики, в частности ее раздела – колористики.

III. Теория

1. Колористика – наука о цвете

В жизни мы постоянно находимся в окружении цвета. Восприятие цвета обычно и естественно для человека. С другой стороны осознанное видение цвета, вычленение оттенка среди остальных, способность к его описанию, цветовому анализу, конструирование гармоничных цветовых сочетаний являются специальными навыками, требующими определенной подготовки.

Правила работы с цветом определяет наука колористика. В основе любого цвета лежит цветовой тон. Обозначим некоторые характеристики цвета, являющиеся необходимыми для анализа и конструирования цветовых решений.

Все цвета делятся на хроматические и ахроматические. Хроматические (от др.-греч. χρῶμα – цвет; краска) цвета и их оттенки – это все те, которые мы различаем и называем «цветными»: красный, оранжевый, бордовый, цвет пыльной розы и т.п. Ахроматические (от др.-греч. ἀ- – не; отрицание) цвета черно-белого оттенка, так называемые «бесцветные»: черный, белый и все серые оттенки. Для точности описания оттенка вводят характеристики яркости и насыщенности. Под яркостью (или светлотой) цвета понимают степень отличия выбранного оттенка по отношению к белому. Например, алый и розовый цвет образованы от одного тона – красного, но имеют разную яркость. Алый – более яркий, розовый – менее яркий. Чем больше белого участвует в образовании цвета, тем ниже его яркость. Параметр насыщенности цвета характеризует его чистоту, а именно «количество» чистого хроматического оттенка, участвующего в образовании цвета. Эпитетом «грязный» характеризуется ненасыщенный цвет (грязно-желтый). Диапазон цветов разной насыщенности можно представить на примере оттенков джинсовой ткани.

С перечисленными характеристиками каждый сталкивается в повседневной жизни, они понятны и используемы в быту. Однако профессиональное владение цветом, знание законов создания цветовых оттенков, цветовое обновление, коррекция пространства требует подготовки и практических навыков. Классическим примером отработки цветовых вариаций служит цветовой круг, представленный на рис. 1. Акцентируем внимание в данном разделе на важности «ручной» техники конструирования цветов «на бумаге» на основе красок всего трех оттенков – синего, красного и желтого [6]. На основе цветового круга создаются гармоничные сочетания по правилам: дополнительные (компланарные) цвета, триады, четырехугольники.

Рис. 1. Цветовой круг, выполненный красками

Базовые навыки работы с цветом есть у каждого. Однако при детальном изучении теории цвета, во-первых, привлекает интересными фактами историческая тематика, излагающая опыты И. Ньютона (XVII в.) по разложению цветов в спектр, художественное произведение И.В. фон Гёте «К теории цвета», описывающее восприятие цвета и природу цвета. Во-вторых, современные фундаментальные исследования, применяемые физические методы, математический аппарат, цифрованные данные и их обработка, описывают законы и вводят точные характеристики для понятия «цвет».

2. Методы цифрового конструирования

С применением компьютерных технологий подходы и методы к формированию цвета изменились. Несложно предположить черно-белую «картинку» в виде матричной структуры из 0 и 1 (где 1 – цвет (пикселя),

0 – его отсутствие). На основе кодирования цветов сформировались новые подходы к их генерации и обозначению.

Классические понятия из теории цвета остались прежними: цветовой тон, яркость, насыщенность. При этом появились точные числовые значения этих параметров. Важную роль стали играть цветовые модели, позволяющие создавать цвета через наборы числовых комбинаций (кортежей). Наиболее актуальные на сегодня – CMYK, RGB, HSB, использование каждой из них определяется поставленными задачами, выбором графического редактора, постобработкой (например, для печати). Проиллюстрируем цветовое пространство и способ цветообразования на примере модели RGB на рис. 2.

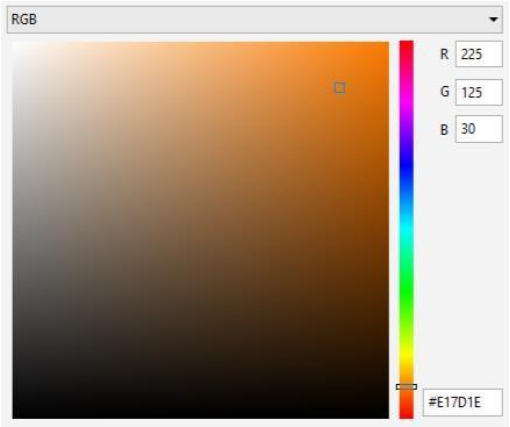


Рис. 2. Цветовое пространство RGB

Принцип формирования цвета в указанной модели строится на основе комбинаций трех цветов, отличных от цветовой палитры красок, – красного, зеленого, синего, каждый из которых принимает значение от 0 до 255. Цветовой круг представлен здесь линейно, градиентным переходом чистых оттенков. Изменение яркости и насыщенности определяется «на глаз» через положение указателя: из верхнего правого угла (чистый цвет) влево (уменьшение яркости, добавление белого), вниз (уменьшение насыщенности, добавление черного). Отметим, что, например, в модели HSB за эти параметры отвечают точные числовые процентные значения (0% – 100%). Важно понимать при таком подходе ограниченность сгенерированных оттенков и разницу в их отображении при переходе от одной модели к другой. Точечное заполнение графического объекта цветом реализует принцип растровой графики [7].

IV. Результаты экспериментов

Навыки работы с цифровыми инструментами цветового анализа и, как следующий этап, цветового конструирования формируются в процессе выполнения практических упражнений и заданий. Приведем примеры внедрения подобной методики, применяемой в рамках реализации профессиональной программы переподготовки «Информационные технологии в дизайне» (квалификация «Промышленный дизайнер») для студентов Пермского Политеха.

На основе авторской фотографии интерьера (салона) общественного транспорта в графическом редакторе требуется выделить (инструментом типа «пипетка») пиксели 5–7 цветов и разбить их по группам:

- Основные цвета (заполняющие наибольшее пространство),
- Нюансы (промежуточные цвета),
- Акценты («пятна»).

Данные занести в таблицу, содержащую: а) образец цвета, б) выделенную группу, в) название цвета, г) принадлежность к цветовой гамме (хроматическая/ахроматическая, е) яркость, ж) насыщенность, и) подбор дополнительных цветов (компланарные цвета, классическая и аналоговая триада). Рекомендуется представить личностную характеристику прорабатываемых цветов, основанную на эмоционально-образном восприятии.

Продемонстрируем результат выполнения предложенного задания (рис. 3).

Рис. 3. Типовой интерьер салона автобуса

ТАБЛИЦА 1
АНАЛИЗ ЦВЕТОВОЙ ПАЛИТРЫ ИНТЕРЬЕРА САЛОНА АВТОБУСА

	Цвет Название	Гамма Яркость / Насыщенность	Дополнительный цвет	Триада: Классическая, Аналоговая
Основные	02497B Насыщенный синий	Хроматическая Средняя / Высокая		
	B6A89B Светлый синевато-серый	Хроматическая Средняя / Низкая		
Нюансы	020308 Глубокий фиолетово-черный	Ахроматическая Низкий / Высокая	-	- -
	Af0414 Бордо (Красно-бордовый)	Хроматическая Средняя / Высокая		
	D3DCE5 Барвинок Крайола	Хроматическая Высокая / Низкая		
Акценты	0325A3 Синяя пыль	Хроматическая Высокая / Высокая		
	EED465 Кожа буйвола (Палевый)	Хроматическая Высокая / Средняя		

Цветовой анализ фотографии выполнен на основе сервиса ColorScheme.ru, представляющего онлайн-инструмент для подбора цветов и генерации цветовых схем. Программа реализует широкий спектр задач в области колористики: позволяет выбирать для работы востребованные сегодня цветовые модели: RGB, PANTONE, RAL; составлять различные виды цветовых сочетаний: моно, контраст, триада, тетрада, аналогия, акцент аналогия; регулировать яркость и насыщенность цветовой схемы в ручном режиме или автоматическом режиме; экспортировать значения цвета в различные форматы данных; пользоваться калькулятором цветов; определять названия цветов. Использование данного ресурса в образовательной программе обеспечивает погружение слушателя в цветовое пространство с возможностью почувствовать многогранность цветопередачи в цифровом эквиваленте. При выполнении основной учебной задачи информационно-технологической направленности, состоящей в анализе цвета цифровыми методами, формируются не менее важные понятийные установки и жизненные практики. Так, работа с цветовыми образцами показывает бесконечность цветового ряда в реальном мире и ограниченность цветопередачи в цифровом пространстве. Подход к формированию цвета и его характеристик в «цифре» позволяет применить математический аппарат к экспрессивной области знаний, «просчитать» изменения оттенков, сформированных на основании субъективного взгляда. И, наконец, осознанный подход к выбору цифрового решения, «безопасных» цветов в различных ситуациях (web-дизайн, цветная печать), обеспечивает качество, стоимость и трудоемкость разрабатываемого продукта.

Остановимся подробнее на настройках и интерфейсе ColorScheme в рамках выполнения задания по колористике (рис. 4, 5, 6).

1. Задание цветовой «пробы» в режиме «Моно», определение положения на цветовом круге.

Рис. 4. Определение цвета в программе ColorScheme

2. Определение насыщенности, яркости и названия цветового образца.

Рис. 5. Определение насыщенности, яркости и названия цветового образца

Рис. 6. Определение классической триады на основе выделенного цвета

V. Обсуждение результатов

Законы конструирование цветовых решений подробно описаны в книге И. Иттена «Искусство цвета». Цветовое восприятие уникально для каждого человека. «Понятие цветовой гармонии должно быть изъято из области субъективных чувств и перенесено в область объективных закономерностей» [8], особенно, когда дело касается профессиональной деятельности дизайнера. Презентация и обсуждение авторских решений на основе реальных примеров на учебном практическом занятии знакомит участников смыслами и идеями, заложенными в проекте. Так, на рис. 7 предложена проработка цветового решения салона автобуса в следующих вариантах: ахроматическая гамма, исходный вариант, авторский вариант на основе выбранного оттенка в Табл. 1 (здесь, красный оттенок триады цвета «Синяя пыль»).

Рис. 7. Варианты цветовых решений интерьера общественного транспорта

VI. Выводы и заключение

Креативные идеи творческого человека не поддаются программному описанию и алгоритмическому сопровождению. Однако без компьютерных технологий сегодня сложно реализовать современный продукт. В компьютерной графике важным умением является грамотная разработка гармоничного цветового решения. С применением цифровых методов обработки изображений процесс выполнения организуется проще и быстрее. Доступные инструменты коррекции и редактирования дают возможность разработать альтернативный вариант. Итоговое решение остается за специалистом. В компьютерной графике роль электронных ресурсов заключается в эффективности и упрощении процесса поиска, например, цветовой гаммы, снижении временных и затрат при печати или производстве продукта.

Важным фактором в применении цифровых методов обработки цвета является использование компьютерных технологий в процессе обучения. При дистанционных подходах в общении со студентами зачастую невозможно конструктивно организовать работу в ручном исполнении (работа с красками, карандашами и пр.). В таком случае цифровые технологии дают альтернативный инструментарий, позволяющий «механически» отработать определенные навыки в колористике.

Список литературы

1. Дианова Ю.В., Столбова И.Д., Носов К.Г. Трехмерная визуализация цифровых учебных моделей в инженерной подготовке // Вестник компьютерных и информационных технологий. 2024. Т. 21, № 3(237). С. 20–29.
2. Карпова Е.А. Особенности профильной подготовки основам колористики на кафедре «Архитектура» НИУ МГСУ // Педагогический журнал. 2022. Т. 12. № 2А. С. 425–432.
3. Салтанова, Ю. С. Методы проблемного обучения на занятиях "Цветоведение и колористика" / Ю. С. Салтанова // Современные тенденции развития науки и технологий. 2017. № 3-10. С. 105-110.
4. Долгих Н. Н. Колористика в дизайн-образовании // Вестник Томского государственного педагогического университета. 2016. № 8(173). С. 57–63.
5. Пилюгайцева Ю. И. Внедрение компьютерных технологий в процесс изучения колористики студентами - дизайнерами // Новая наука как результат инновационного развития общества : сборник статей Международной научно-практической конференции: в 17 частях, Сургут, 22.04.2017. Том Часть 16. Сургут: Общество с ограниченной ответственностью "Агентство международных исследований", 2017. С. 95– 98.
6. Яньшина М. М. Практический опыт профессиональной подготовки художественно-технических редакторов по дисциплине «Цветоведение и колористика» // Дизайн-образование: преемственность и динамика: материалы Всеросс. науч.-практ. конф., Оренбург, 13 октября 2023 года. Оренбург: Оренбургский государственный университет, 2023. С. 98–101.
7. Колодников А. И. Ранние формы компьютерного дизайна: пиксельная графика и растровая система // Terra Artis. Искусство и дизайн. 2022. № 3. С. 36–41.
8. Иттен И. Искусство цвета / пер. с немецкого: 3-е изд.. М.: Изд. Д. Аронов, 2004. 96 с.

УДК 514:142
DOI: 10.25206/978-5-8149-3873-2-2024-827-836

Анализ топологии полиэдров в задаче аппроксимации замкнутых поверхностей полиэдрами с группами конгруэнтных граней
Analysis of the topology of polyhedra in the problem of approximation of closed surfaces by polyhedra with groups of congruent faces

В. В. Рустамян МИРЭА – Российский технологический университет, пр-т Вернадского, д. 78, Москва, Россия

V. V. Rustamyan [1]
[1] *MIREA – Russian Technological University, 78 Vernadsky Avenue, Moscow, 119454, Russia*

Аннотация. Подходы к решению задачи аппроксимации поверхностей свободной формы полиэдрами с группами конгруэнтных граней могут быть разнообразными. Один из возможных подходов заключается в восприятии задачи как оптимизационной и решении её эволюционными методами, такими как генетический алгоритм. Для этого требуется понимание свойств алгоритма, правильная формализация условий задачи и исследование поведения аппроксимирующего полиэдра как динамической системы. В статье представлены результаты эксперимента, целью которого является выявление влияния топологии полигональной сети на способность группирования граней по признаку конгруэнтности. Экспериментальные данные демонстрируют, как сети с отличающейся топологией могут по-разному влияют на способность образования групп конгруэнтных граней, что важно для оптимизации процесса аппроксимации и улучшения качества получаемых полиэдров. Полученные результаты позволяют глубже понять взаимодействие между топологией полигональной сети и группированием граней, а также предлагают новые пути для дальнейших исследований в области эволюционных методов оптимизации для аппроксимации сложных поверхностей.

***Ключевые слова*:** топология многогранников, генетический алгоритм, икосаэдр, функция приспособленности

Аннотация. Approaches to solving the problem of free-form surface approximation using polyhedra with groups of congruent faces can vary widely. One possible approach is to treat the problem as an optimization task and solve it using evolutionary methods, such as a genetic algorithm. This requires an understanding of the algorithm's properties, proper formalization of the problem conditions, and an investigation of the behavior of the approximating polyhedron as a dynamic system. The article presents the results of an experiment aimed at identifying the influence of the polygonal mesh topology on the ability to group faces based on congruency. Experimental data demonstrate how networks with differing topologies can affect the capacity for forming groups of congruent faces in various ways, which is crucial for optimizing the approximation process and improving the quality of the resulting polyhedra. The results provide deeper insight into the interaction between polygonal mesh topology and face grouping, as well as suggest new avenues for further research in the field of evolutionary optimization methods for approximating complex surfaces.

***Ключевые слова*:** topology of polyhedra, genetic algorithm, icosahedron, fitness function

I. Введение

Аппроксимация поверхностей свободной формы является современной задачей, на которую направлены взгляды многих исследователей. В основном подобные аппроксимации применимы в архитектуре, потому задача является комплексной и рассматривается со многих точек зрения. Есть исследования, в которых рассматриваются аппроксимации в зависимости от физических свойств материалов, используемых для конструирования [1, 2]. В некоторых исследованиях уделяют внимание эстетике, практичности и массово-прочностным характеристикам конструкций, содержащих поверхности свободной формы [3, 4]. Ведутся исследования в данной области для создания архитектурных конструкций, с учетом множественности влияющих факторов на будущий конструктив [5 - 8]. Это могут быть геометрические факторы (вписание объекта в ограниченную среду), природные (учет направления ветра, движения солнца), оптические свойства (при использовании зеркальных поверхностей), внутренняя планировка зданий и множество других параметров. Подобные решения опираются на параллельно развивающиеся машинное обучение, искусственный интеллект и аналитику больших данных. В следствие чего появляются различное программные решения с возможностью генеративного дизайна. Еще одним направлением исследования является создание математического (аналитического, геометрического, численного) аппарата и программных решений для аппроксимации поверхностей свободной формы гранными поверхностями. Существуют публикации [9 - 11], в которых ставится задача исследования аппроксимации поверхностей свободной формы полиэдрами с группами конгруэнтных граней и в некоторой степени, с допущениями решается.

Также исследуется вопросы расчета и возведения опор, ферм и сопутствующего конструктива [12]. Анализ литературных источников показывает, что задача аппроксимации поверхностей свободной формы полиэдрами с группами конгруэнтных граней является актуальной и относится к сложно формализуемым задачам.

В рамках решения задачи аппроксимации поверхностей свободной формы полиэдрами с группами конгруэнтных граней в данной статье рассматривается вопрос влияния топологии полиэдра на скорость и результативность группирования по признаку конгруэнтности граней полиэдра. Для этого производится ряд ограничений. Вместо произвольной поверхности свободной формы выбирается сферическая поверхность. За образец полиэдра, аппроксимирующего сферическую поверхность, выбирается из списка известных замкнутых выпуклых полиэдров с группами конгруэнтных граней. Ограничения связанные с работой программы [13]: замкнутость поверхности; полиэдр состоит исключительно из треугольников. Также ограничением является количество граней многогранника. Связанно это со скоростью проведения экспериментов и наглядностью визуализации образцов.

Известен ряд замкнутых выпуклых полиэдров с группами конгруэнтных граней. Правильные полиэдры - тела Платона. Полуправильные полиэдры - тела Архимеда и двойственные им Каталановы тела. Призмы и антипризмы. Тела Джонсона [14] – замкнутые и выпуклые многогранники, содержащие правильные грани (треугольники, квадраты, пятиугольники и т.д.). Тела Джонсона не принадлежат к предыдущим группам многогранников. Alain Lobel (Алан Лобель) разработал математический аппарат и алгоритм, аппроксимирующий аналитические поверхности правильными треугольниками для создания архитектурных форм – Lobel mashes [15]. Существуют различные модификации сеток Лобеля [16]. Из данного многообразия форм требуется произвести выбор, удовлетворяющий заданным ограничениям. Используя выбранный образец и, созданный при помощи алгоритмов программы [13], не изоморфный аналог (одинаковый по количеству граней, но имеющий отличия по количеству ребер в многогранных углах при вершинах полиэдра) требуется проанализировать поведение и результаты процесса оптимизации.

II. Постановка задачи

Выдвигается гипотеза о том, что от топологии полиэдра, являющегося первым индивидом генетического алгоритма, зависит скорость и результативность поиска полиэдра с группами конгруэнтных граней.

Ставятся задачи исследования:

1. Осуществить выбор образца для исследования.
2. Усовершенствовать функцию приспособленности (fitness function) генетического алгоритма для рекомбинации положений вершин полиэдров, изоморфных образцу, во время работы алгоритма.
3. Выполнить сравнительный анализ аппроксимации поверхности полиэдрами с теоретически верной топологией и топологией генерируемой программой [13].

III. Теория

10. Описание алгоритма генерации первого индивида

Здесь описываются теоретические основы и алгоритмы, заложенные в работу части программы, которая отвечает за создание первого индивида (полиэдра) для дальнейшей оптимизации положения вершин с помощью генетического алгоритма.

Программа принимает stl-файл с поверхностью свободной формы, то есть набор координат вершин триангулированной поверхности достаточной точности. В блоке кода по созданию первого индивида (полиэдра с заданным количеством граней) реализован алгоритм «схлопывания по ребру». Блок-схема алгоритма продемонстрирована на рис. 1. На вход подаются реляционные таблицы «список вершин (vertices)», «список ребер (edges)» и «список граней (faces)», созданные по данным из stl-файла. Если V – множество вершин, E – множество ребер и F – множество граней триангулированной модели поверхности свободной формы из stl-файла, то реляционные таблицы организованы следующим образом:

$$vertices = \{(v_i, (x_i, y_i, z_i), E_{v_i}, F_{v_i}) | v_i \in F\},$$
$$v_i - \text{индекс вершины},$$
$$E_{v_i} - \text{множесво ребер, инцидентных вершине } v_i,$$
$$F_{v_i} - \text{множество граней, инцидентных вершине } v_i;$$

$$edges = \left\{\left(e_j, V_{e_j}, F_{e_j}\right) \middle| e_j \in E\right\},$$
$$e_j - \text{индекс ребра},$$

V_{e_j} – множество вершин, инцидентных ребру e_j,
F_{e_j} – множество граней, инцидентных ребру e_j;

$$faces = \{(f_k, V_{f_k}, E_{f_k}) | f_k \in F\},$$
f_k – индекс грани,
V_{f_k} – множество вершин, инцидентных грани f_k,
E_{f_k} – множество ребер, инцидентных грани f_k.

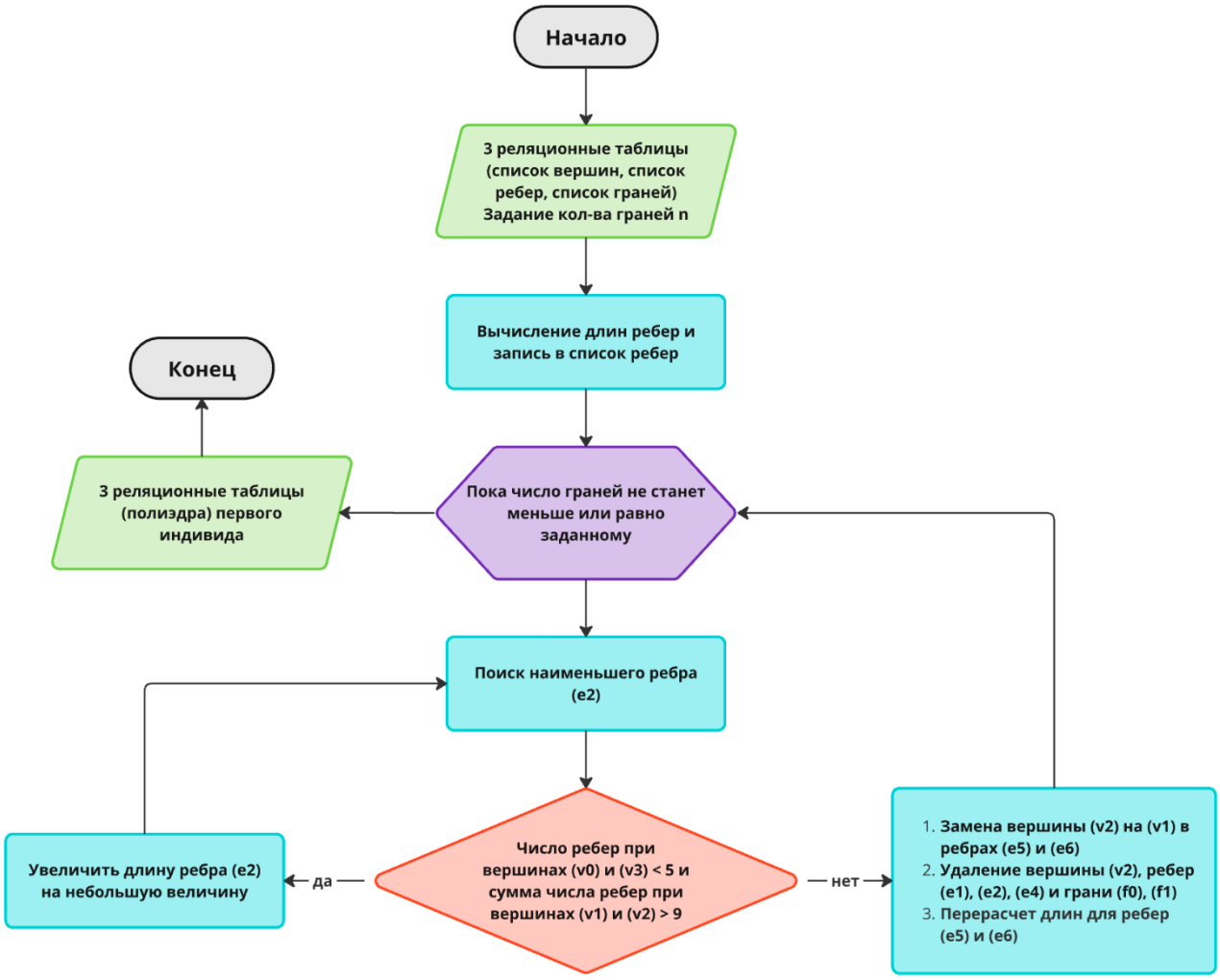


Рис. 1. Блок-схема части программы. Алгоритм «схлопывания по ребру»

Задается число граней (полиэдра) первого индивида - n. Далее происходит процесс расчета и записи в «список ребер» длин всех ребер. Обозначим инцидентные ребру вершины V_{e_j} как (v_a, v_b), где:

$$v_a = (x_a, y_a, z_a),$$
$$v_b = (x_b, y_b, z_b).$$

Длина ребра e_j рассчитывается как евклидово расстояние между точками v_a и v_b:

$$l(e_j) = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2}.$$

Реляционная таблица «список ребер» выглядит:

$$edges = \left\{ \left(e_j, V_{e_j}, F_{e_j}, l(e_j) \right) \middle| e_j \in E \right\}.$$

В цикле, пока число граней не будет меньше или равно заданному

$$|F_p| \leq n$$

в несколько этапов происходит процесс «схлопывания по ребру».

Первый шаг – поиск наименьшего ребра

$$e_{min} = arg \min_{e_j \in E} \{l(e_j)\},$$

например, e_2 (рис.2).

Во втором шаге используется условный оператор: если число ребер при вершинах v_0 и v_3 строго меньше 5 и сумма числа ребер при вершинах v_1 и v_2 строго больше 9, то … Данное условие служит ограничением нижней и верхней границы количества ребер инцидентных вершинам полиэдра. Логично предположить, что требуется короткий диапазон ребер инцидентных вершинам, например, 4–5 или 5–6. Но, используя такую реализацию алгоритма и условный оператор, нельзя добиться иного диапазона, нежели 4–7 ребер при вершинах. И связано это с возникновением ситуаций отсутствия возможности инициации процесса «схлопывания по ребру», исходя из геометрии полиэдра. При выполнении условия, то есть не найдено подходящее ребро, увеличивается длина ребра

$$l(e_{min}) \leftarrow l(e_{min}) + \beta,$$

$$\beta - \text{небольшая относительно длины ребра величина,}$$

и процесс возвращается к первому шагу. Иначе, заменяется вершина v_2 на v_1 в ребрах e_5 и e_6,

$$\forall e_j = (v_2, v_t) \in E, e_j \leftarrow (v_1, v_t),$$

удаляются вершины v_2, ребра e_1, e_2 и e_4 и грани f_0 и f_1,

$$V \leftarrow V \setminus \{v_2\},$$

$$E \leftarrow E \setminus \left\{E_{v_2} \cap \{E_{f_0}, E_{f_1}\}\right\},$$

$$F \leftarrow F \setminus \{F_{v_2}\},$$

и пересчитываются длины для ребер e_5 и e_6 [17].

По завершению условия общего цикла, когда созданы реляционные таблицы первого индивида $vertices_p$, $edges_p$ и $faces_p$, процесс данного блока завершится.

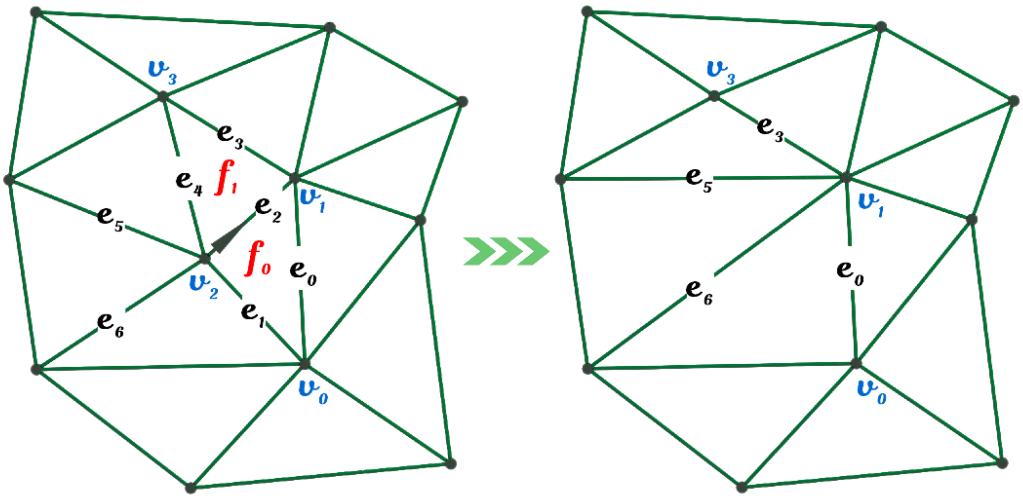


Рис. 2. Демонстрация метода «схлопывания по ребру»

11. Выбор образца и генерация первого индивида

Выбор образца для исследования основывается на ограничениях программы и удовлетворению рамок задачи. Ограничения программы:

1) поверхности аппроксимации замкнута;
2) отсутствие самопересечений поверхности аппроксимации;
3) тип непрерывности поверхности аппроксимации выше G0;
4) аппроксимирующий полиэдр образован треугольными гранями;
5) вершины в многогранных углах инцидентны количеству ребер в диапазоне 4–7 включительно.

Рамки задачи:

1) аппроксимирующий полиэдр должен содержать небольшое количество граней (обеспечение скорости проведения эксперимента, наглядность образца);
2) аппроксимирующий полиэдр должен содержать достаточное количество граней (алгоритм «схлопывания по ребру» должен быть способен создать сетку топологически отличную от теоретически верной сетки образца);

3) топологические свойства аппроксимирующего полиэдра должны быть известны.

Из 5-и тел Платона куб и додекаэдр не подходят по количеству вершин в грани. Тетраэдр не подходит по второму условию рамок задачи и 5-му ограничению программы. Подходят октаэдр и икосаэдр.

Из 4-х тел Кплера-Пуансо может подойти малый звездчатый додекаэдр, но он не является выпуклым и не ясно какую поверхность аппроксимирует, что может осложнить поиск конгруэнтных групп. Остальные тела самопересекающиеся или не удовлетворяют ограничениям программы.

Ни одно тело Архимеда не подходит по 4-му ограничению программы.

Из двойственных архимедовым телам 13-и тел Каталана по 4-му ограничению программы подходят 7. Это триакистетраэдр, тетракисгексаэдр, пентакисдодекаэдр, триакисоктаэдр, триакисикосаэдр, гекзакисоктаэдр и гекзакисикосаэдр. Но по 5-му ограничению останутся только тетракисгексаэдр и пентакисдодекаэдр.

Из 92-х тел Джонсона всем условиям удовлетворяют 4 многогранника. Это пятиугольная бипирамида, скрученно удлинённая четырёхугольная бипирамида, трижды наращённая треугольная призма и плосконосый двуклиноид. Могут аппроксимировать поверхность эллипсоида вращения.

Сетки Алана Лобея состоящие из правильных треугольников содержат 9 групп, состоящие из сотен сеток. Основная масса сеток незамкнута. Замкнутые сетки, как правило, сетки, полученные зеркальной симметрией не замкнутых сеток. Также, из атласов Лобеля не всегда понятно какую поверхность аппроксимирует сетка и мало информации о свойствах и методе создания таких сеток. Потому сетки Лобеля не рассматриваются в качестве кандидатов в образцы.

ТАБЛИЦА 1
СВОЙСТВА ПОДХОДЯЩИХ ПОЛИЭДРОВ

№	Наименование	Кол-во граней	Конфигурация вершин
Тела Платона			
1	Октаэдр	8	$6(3^4)$
2	Икосаэдр	20	$12(3^5)$
Тала Каталана			
3	Тетракисгексаэдр	24	$6(3^4)8(3^6)$
4	Пентакисдодекаэдр	60	$12(3^5)20(3^6)$
Тела Джонсона			
5	Пятиугольная бипирамида	10	$5(3^4)2(3^5)$
6	Скрученно удлинённая четырёхугольная бипирамида	16	$2(3^4)8(3^5)$
7	Трижды наращённая треугольная призма	14	$3(3^4)6(3^5)$
8	Плосконосый двуклиноид	12	$4(3^4)4(3^5)$

Существуют еще масса многогранников с группами конгруэнтных граней, не входящие в вышеописанный список и удовлетворяющие заданным ограничениям, например, шести- и семиугольные бипирамиды. Но найденных 8-и полиэдров вполне достаточно для анализа и поиска подходящего образца. По количеству граней в средний диапазон попадают икосаэдр и тетракисгексаэдр, см. Табл. 1. К сферической поверхности ближе пентакисдодекаэдр с 60 гранями, но обладает меньшей наглядностью. Остальные полиэдры не очень подходящие, в силу меньшего количества граней, что уменьшает вероятность возникновения отличия в топологии решетки относительно теоретических образцов при работе метода «схлопывания по ребру».

Из двух наиболее подходящих, выбор пал на икосаэдр. На рис. 3 показан икосаэдр и привычное представление графа изоморфного икосаэдру.

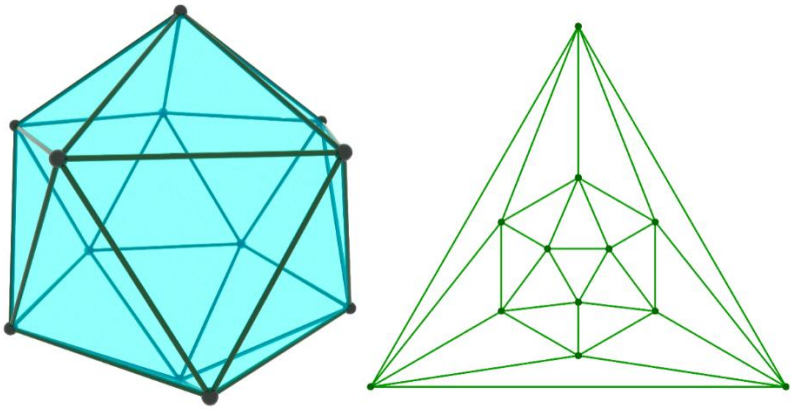


Рис. 3. Правильный икосаэдр и изоморфный ему граф

На рис. 4. Показана ретопология икосаэдра, образующая икосаэдр Йессена. Икосаэдр Йессена изоморфен правильному икосаэдру [18]. Если ретопология не завершена или ретопологии подвержены другие пары граней, возникает множество вариантов неизоморфных, но эквивалентных по составу полиэдров. Рассмотрение данного множества может быть задачей для отдельного исследования.

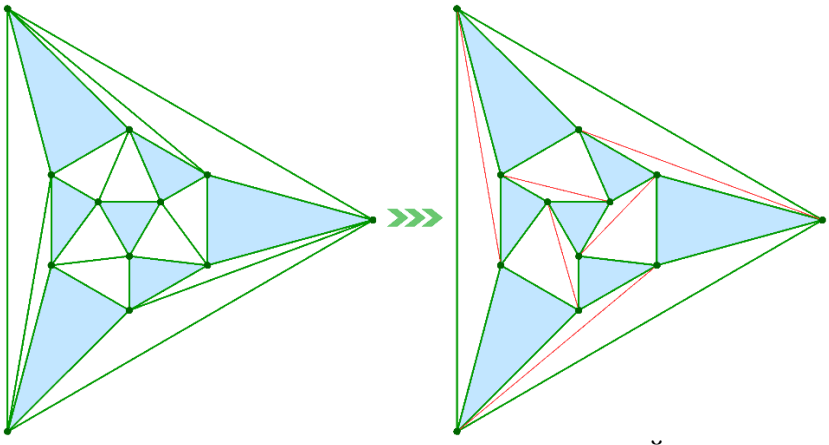


Рис. 4. Граф правильного икосаэдра и икосаэдра Йессена

Программой был сгенерирован первый индивид, назовем его «сгенерированный образец». На рис. 5 показан изоморфный ему граф, в удобном для сравнения представлении, и сам объект. На представлении в виде графа хорошо видно, по отношению к топологии икосаэдра, ретопологию двух пар граней.

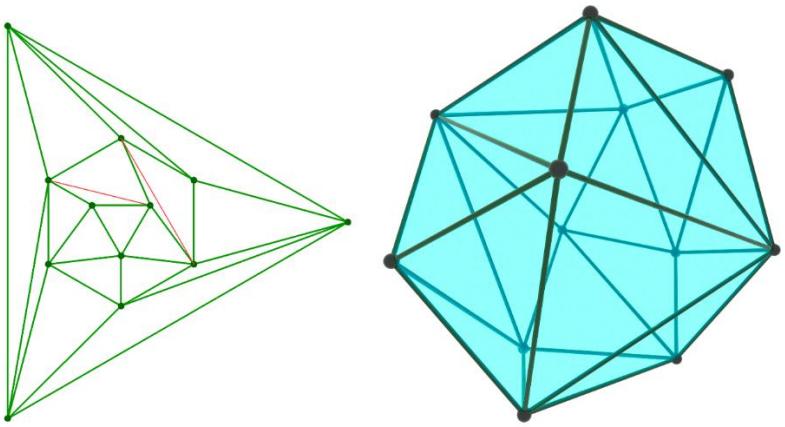


Рис. 5. Граф изоморфный сгенерированному образцу и сам образец

Для чистоты эксперимента правильный икосаэдр был подвержен деформации так, чтобы его соответствующие вершины совпали со сгенерированным образцом, как показано на рис. 6. Такую форму икосаэдра назовем «оригинальным образцом».

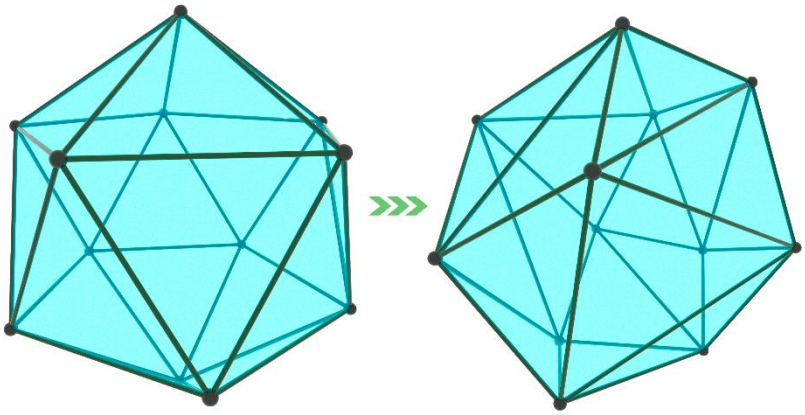


Рис. 6. Получение оригинального образца

12. Усовершенствование функции приспособленности и экспериментальная процедура

В блоке программы с генетический алгоритм есть раздел, где представлена функция приспособленности. В этой функции формализованы условия выживания особей при работе генетического алгоритма. В зарегистрированной программе для ЭВМ [13] содержится два условия. Первое - условие принадлежности вершин полиэдра поверхности свободной формы. Второе – условие сгруппированности граней по признаку конгруэнтности. Оба условия выражены численно, что позволяет характеризовать индивида (особь) в популяции и выявлять лучшего. В связи с тем, что образцы основаны на правильном икосаэдре, было принято решение дополнить функцию приспособленности еще одним условием, связанным со стремлением граней индивида к равносторонности.

Основными проблемами формализации любой задачи оптимизации для использования генетического алгоритма являются выявление гена индивида и создание соответствующей гену и условиям задачи функции приспособленности. При этом функция может быть дискретной, но генетический алгоритм работает лучше, если функция приспособленности непрерывна. Все функции приспособленности в программе дискретны. Для уменьшения шага дискретизации была разработана функция, основанная на изменении углов треугольника. Выведен показатель ***D*** (1) – сумма модулей разности каждого угла треугольника и угла 60°. Рассмотрены предельные случаи треугольников (рис. 7). Показатель ***D*** для равностороннего треугольника равен 0°. Если угол β_2 стремится к 0°, то углы β_1 и β_3 стремятся к 90°, значит в предельном случае показатель ***D*** равен 120°. Если угол γ_2 стремится к 180°, тогда углы γ_1 и γ_3 стремятся к 0°, значит показатель ***D*** равен 240°. Показатель ***D*** будет находиться в диапазоне 0°…240°. Далее, для рассчитанных показателей ***D*** всех граней полиэдра, получается среднее арифметическое значение. Среднее значение делится на 60°, чтобы получить численное представление, меняется знак на противоположный и смещается на 1. В результате этих действий получается показатель ***m*** (2), который находится в диапазоне 1…-3. Показатель ***m*** является признаком стремления к равносторонности всех граней индивида. При ***m=1*** все грани полиэдра будут равносторонними.

$$D = \sum_{i=1}^{3} |\alpha_i - 60°| \tag{1}$$

$$m = 1 - \frac{\sum_{i=1}^{n} D_i}{n \cdot 60°}, n - \text{число граней полиэдра} \tag{2}$$

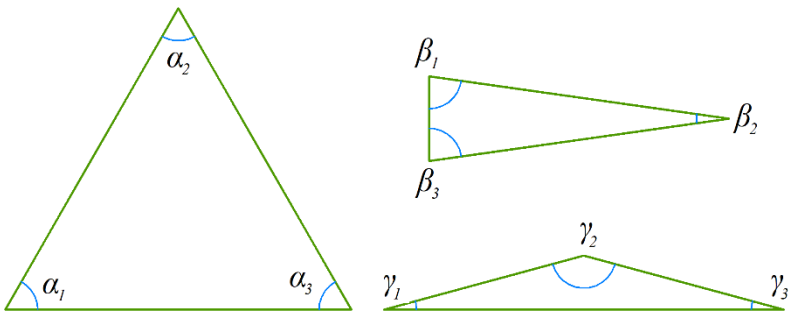


Рис. 7. Предельные случаи углов треугольника

В эксперименте участвуют оригинальный образец и сгенерированный образец. Так как генетический алгоритм является стохастическим, то для каждого образца проводится по 10 испытаний с установлением среднего значения сгруппированных граней. Эксперимент производится двумя способами: с помощью оригинальной программы; с использованием усовершенствованной функции приспособленности. Все гиперпараметры системы остаются неизменными на протяжении всего эксперимента.

Основные гиперпараметры:

1) Количество индивидов в популяции – 500
2) Количество поколений – 5000
3) Вероятность скрещивания – 50%
4) Вероятность мутации индивидов в поколении – 90%
5) Вероятность мутации гена индивида – 10%

Эксперимент проводился на персональном компьютере с характеристиками: ОС – Windows 10; процессор - Intel(R) Core(TM) i5-7400 CPU @ 3.00GHz, 3001 МГц, ядер: 4, логических процессоров: 4; оперативная память – 16 Gb.

IV. Результаты экспериментов

Результатом экспериментов является сводная таблица, где показаны результаты сгруппированности в каждом испытании, средние значения и значения в процентах от общего количества граней, табл. 2. Также на рис. 8 приведен график зависимости функции приспособленности от количества поколений одного из процесса расчетов для оригинального образца с модифицированной функцией приспособленности.

ТАБЛИЦА 2
РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Сгенерированный образец без модификации функции приспособленности											Среднее значение	% от общего
Граней в группах	2	2	4	4	2	4	2	4	2	4	3	15
Оригинальный образец без модификации функции приспособленности											Среднее значение	% от общего
Граней в группах	2	4	2	6	4	2	2	2	6	4	3,4	17
Сгенерированный образец с модификацией функции приспособленности											Среднее значение	% от общего
Граней в группах	8	6	4	4	4	4	4	6	6	8	5,4	27
Оригинальный образец с модификацией функции приспособленности											Среднее значение	% от общего
Граней в группах	9	10	8	9	10	6	11	6	11	11	9,1	45,5

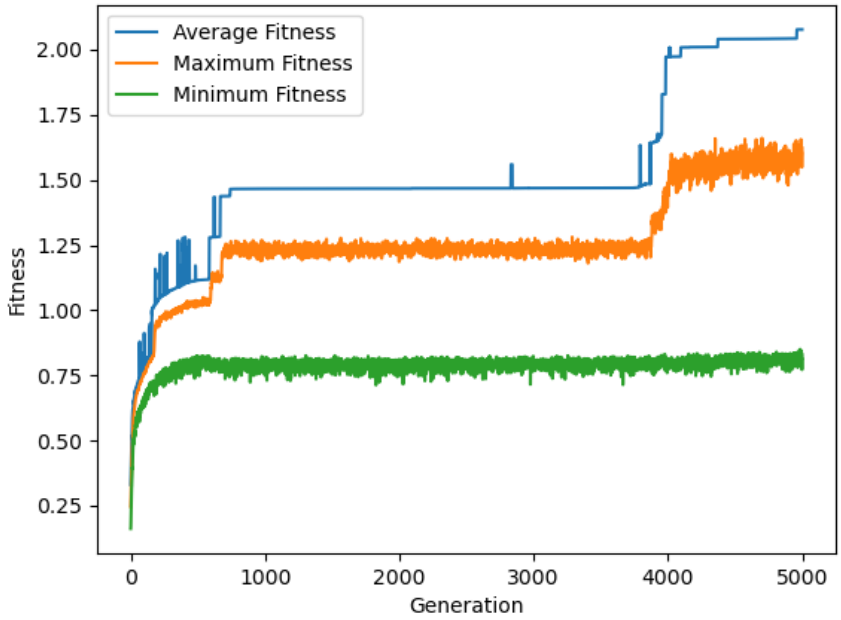


Рис. 8. График зависимости значения функции приспособленности (Fitness) от количества поколений (Generation) одного расчета оригинального образца с модифицированной функцией приспособленности

V. Обсуждение результатов

При проведении эксперимента без модификации функции приспособленности наблюдается отличие в сторону оригинального образца в 2%, см. табл. 2. Показатель не велик, но дает понятие о тенденции. Возможно, что этот показатель в пределах погрешности? Проведение большего количества испытаний или увлечение топологических изменений в образце, могли бы дать ответ на этот вопрос точнее. При сравнении результатов в случае с модификацией функции приспособленности наблюдается преимущество оригинального образца почти в 1.7 раза. Если посчитать количество граней, отличных от икосаэдра (рис. 5), то их всего 4. Но если учесть грани, смежные по ребрам с этим участком, то в сумме их уже 10. Возможно полученный результат каким-то образом коррелирует с этим фактом. Но в роли контраргументов выступает работа совокупности трех функций приспособленности и их приоритеты. Функции, виляющие на принадлежность вершин полиэдра поверхности и стремление к равносторонности, заложены в программе, примерно, одинакового приоритета. То есть 0.5, от возможной оценки в 1, выделяется на одну функцию и те же 0.5 на другую. Приоритет функции сгруппированности гораздо выше. Она помножает возможную оценку индивида в 1 на величину (1 + 0.2 * количество сгруппированных граней), что наблюдается в виде резких скачков на рис. 8. Единственное, эта величина с малым шагом дискретизации, так как граней всего 20, из-за чего сгруппированность является редким событием в процессе работы генетического алгоритма. Еще контраргументом является то, что не ставилось задачи получить равносторонние грани и алгоритм осуществляет поиск конгруэнтных. Стоит заметить, что модификация функции приспособленности повысила результативность практически в 2 раза в случае со сгенерированным образцом и в 2.7 раза в случае теоретически верной сеткой.

VI. Выводы и заключение

Усовершенствована функция приспособленности программы [13], что привело к повышению результативности программы. Проведен эксперимент, в ходе которого можно сделать вывод, что есть однозначное влияние топологии полигональной сети сгенерированного полиэдра, аппроксимирующего поверхность свободной формы, на конгруэнтность граней в нем. Как именно влияет топология сети требует более детального рассмотрения. Представляет интерес дальнейшее проведение испытания над полигонами с большим количеством граней, например, с пентакисдодекаэдром (60 граней). Используя теорию многогранных углов, можно изучить влияние изменения положения вершин грани на соседние грани, например, при фиксации части граней. Также следует рассмотреть другие технологии генерации первого индивида или, возможно, задать способность изменения топологии в гене индивида во время работы генетического алгоритма.

Список литературы

1. Jimenez M., Müller C., Pottmann H. Discretizations of Surfaces with Constant Ratio of Principal Curvatures. // Discrete & Computational Geometry. 2020. 63. 10.1007/s00454-019-00098-7.

2. Мишанин И. Н. Геометрическое моделирование многогранных конструкций с плоской разверткой поверхности из модульных элементов: специальность 05.01.01 "Инженерная геометрия и компьютерная графика": диссертация на соискание ученой степени доктора технических наук / Мишанин Иван Никифорович. Пенза, 2002. 336 с.

3. Киричков И. В. Формообразующие принципы отеля "Morpheus" в Макао // Современная архитектура мира. 2022. № 2(19). С. 67-80. DOI 10.25995/NIITIAG.2022.19.2.003.

4. Коновалова Н. А. Архитектура Оперного театра в Дубае: цели и компромиссы // Современная архитектура мира. 2022. № 1(18). С. 217-232. DOI 10.25995/NIITIAG.2022.18.1.011.

5. Акшов Э. А. Использование вычислительного проектирования и искусственного интеллекта при моделировании архитектурных объектов // Архитектура и современные информационные технологии. 2023. № 2(63). С. 298-315.

6. Кравченко Г. М., Труфанова Е. В., Полетаев М. В., Пуданова Л. И. Эволюция формообразования здания параметрической архитектуры с учетом аэродинамики // Инженерный вестник Дона. 2021. № 9(81). С. 268-277.

7. Липкин С. М., Михалин Е. С., Губий В. Д., Чипко А. Г. Оптимизация расположения датчиков в сенсорной сети на основе генетического алгоритма // Успехи современной науки и образования. 2017. Т. 5, № 2. С. 78-81.

8. Салех М. С. Методика поиска архитектурной формы путем применения принципов генетического алгоритма с помощью цифровых технологий на примере общественного центра в городе Истре // Наука, образование и экспериментальное проектирование. Труды МАРХИ: Материалы международной научно-практической конференции, Москва, 08–12 апреля 2019 года. Москва: Московский архитектурный институт (государственная академия), 2019. С. 433-436.

9. Liu Y., Lee T.-U., Rezaee Javan A., Pietroni N., Xie Y. Reducing the Number of Different Faces in Free-Form Surface Approximations Through Clustering and Optimization. // Computer-Aided Design. 2023. 166. 103633. 10.1016/j.cad.2023.103633.

10. Singh M., Schaefer S. Triangle Surfaces with Discrete Equivalence Classes. // ACM Trans. Graph. 2010. 29. 10.1145/1833351.1778783.

11. Pellis D., Kilian M., Wang H., Jiang C., Müller C., Pottmann H. Architectural freeform surfaces designed for cost-effective paneling through mold re-use. // Conference: Advances in Architectural Geometry. 2021.

12. Lu H., Xie Y. Reducing the number of different members in truss layout optimization. // Structural and Multidisciplinary Optimization. 2023. 66. 10.1007/s00158-023-03514-y.

13. Свидетельство о государственной регистрации программы для ЭВМ № 2024612472 Российская Федерация. Программа создания триангулированных полиэдров, грани которых группируются по свойству конгруэнтности треугольников, аппроксимирующих замкнутые не самопересекающиеся поверхности свободной формы с помощью генетического алгоритма: № 2023685692: заявл. 24.11.2023: опубл. 01.02.2024 / В. В. Рустамян.

14. Johnson N. W. Convex Polyhedra with Regular Faces. // Canadian Journal of Mathematics 18. 1966. С. 169–200.

15. Lobel A. Lobel mashes. // https://www.equilatere.net/frame.php?page=en/fondamentals/families.php. (дата обращения 02.07.2024)

16. Jiang C., Tang Ch., Tomicic M., Wallner J., Pottmann H. Interactive Modeling of Architectural Freeform Structures. // Combining Geometry with Fabrication and Statics. 2014. 10.1007/978-3-319-11418-7_7.

17. Александров П.С. Комбинаторная топология [Текст] / Ленинград: Гостехиздат, 1947 (Ленинград: тип. им. Евг. Соколовой), 660 с.

18. Jessen B. Orthogonal Icosahedra. // Nordisk Matematisk Tidskrift 1967. Vol. 15. no. 2/3 P. 90–96.

УДК 004.921
DOI: 10.25206/978-5-8149-3873-2-2024-837-842

Динамическое редактирование геометрических образов с использованием логических инструментов в программной среде Симплекс
Dynamic editing of geometric images using logical tools in the Simplex software environment

С. Ю. Щур Санкт-Петербургский политехнический университет Петра Великого, ул. Политехническая, 29, Санкт-Петербург, 195251, РФ

S. Yu. Shchur
[1]Peter the Great St.Petersburg Polytechnic University (SPbPU), Polytechnicheskaya, 29 , St.Petersburg, 195251, Russia

Аннотация. В работе рассматриваются вопросы динамического редактирования геометрических структур с использованием логического, декларативного подхода. Результаты исследования могут лечь в основу разработки новых интерфейсов взаимодействия с геометрическими структурами для решения задач проектирования поверхностей. Использование динамического редактирования в решении геометрических задач позволяет добиваться более наглядного результата. В работе рассматривается возможность генеративного проектирования объектов окружения через геометрическое моделирование.

***Ключевые слова*:** геометрическое моделирование, генеративное проектирование, язык логического программирования, пролог, предикат

Аннотация. The paper discusses the issues of dynamic editing of geometric structures using a logical, declarative approach. The results of the study can form the basis for the development of new interfaces for interaction with geometric structures to solve surface design problems. The use of dynamic editing in solving geometric problems allows you to achieve a more visual result. The paper examines the possibility of generative design of environmental objects through geometric modeling..

***Ключевые слова*:** geometric modeling, generative design, logical programming language, prolog, predicate.

I. Введение

Подходы генеративного проектирования сегодня особенно популярны. Технологии генерации вариантов решений как в конструкторской, так и в дизайнерской области, демонстрируют возможности, способные вести работу над новой моделью значительно быстрее, а также принимать решения на основе запросов быстрее и точнее. Из множества вариантов проектировщик выбирает наиболее подходящий, что говорит о возможности создавать проекты в ограниченных и неограниченных проектных пространствах. [1] Связь декларативного программирования и геометрического моделирования отображается в серии работ [2, 3], в частности описывается использование языка PROLOG для доказательства теорем евклидовой геометрии. В качестве основной программы была выбрана система Симплекс [4].

В связи с современными требованиями производства возникает необходимость разработки новых подходов к формированию структуры продукта. Это обусловлено рядом факторов: индивидуальные запросы клиентов, требующие быстрой адаптации и модификации структуры продукта; усиление конкуренции, требующее повышения эффективности производства и планирования продукции на ранних этапах; ограничения исходной статичной структуры продукта, предполагающей ограниченное количество вариантов; потребность в определении характеристик продукта, таких как стоимость и вес; а также необходимость соблюдения плановых сроков запуска новой продукции для максимизации прибыли.

Индивидуальные спецификации заказчиков, требующих быстрой адаптации продукта и модификации его структуры, могут быть связаны с изменением предпочтений потребителей, появлением новых технологий или рыночных трендов. Примеры таких спецификаций включают: разработку мобильных приложений с учётом особенностей использования на разных устройствах; адаптацию веб-сайтов для работы на различных платформах и устройствах;изменение дизайна продукта в соответствии с новыми трендами и предпочтениями потребителей; модификацию функционала продукта для удовлетворения специфических потребностей клиентов; внедрение новых технологий и инноваций для улучшения пользовательского опыта и конкурентоспособности продукта.

II. Постановка задачи

Целью работы является формирование алгоритмического комплекса решения задач, связанных с генеративным проектированием поверхностей. Актуальность данного исследования заключается в поиске нового подхода к проектированию конструкций, структур, а также объектов промышленного дизайна через использование методов геометрического моделирования.

В литературе можно найти определенную последовательность действий в рамках проектирования структур и поверхностей, что в основном является следствием применения отработанных технологий, всеобще известных и наиболее используемых. Это подходы в автоматизации проектирования, и информационные системы поддержки жизненного цикла изделия, и принципы работы станков с ЧПУ и особенности обработки современных материалов с учетом их стоимости, характеристик, эксплуатационных свойств и т.д.

С одной точки зрения появление, развитие и популяризация аддитивных технологий позволило проектировщикам уйти от ограничений, связанных с классическими подходами в проектировании и производстве изделий. С другой стороны, появились вопросы интеграции новых технологий в производственную среду. Новые подходы требуют новых решений также и в программной части, в том числе в области разработки САПР и информационных систем иного порядка, значительно отличающихся от стандартных представлений о проектировании. В данном случае речь уже идет о более тонких настройках с точки зрения геометрического моделирования, о возможности динамической корректировки структур с учетом технологий изготовления.

Также целью данной работы является определение возможности программного комплекса автоматизированного проектирования Симплекс в решении задач генеративного проектирования. Это позволит определить подходы в решении области задач, связанных с генеративным проектированием динамических структур. В основе исследования лежит программный компонент системы Симплекс - Логика, которая основана на принципах декларативного программирования языка Пролог [5].

На сегодняшний день с помощью Пролога реализуется разработка динамических баз данных, создание прототипов прикладных программ, управление производственными процессами. Также Prolog применяется в экспертных системах, символьных пакетах вычислений и исследованиях в области искусственного интеллекта.

На базе Prolog разработаны экспертные системы в области планирования, перевода текста, интеллектуальных игры и т.д. Примеры экспертных систем на основе Prolog включают: PIE (Prototype Integrated Environment), Tubo-Prolog, SWI-Prolog, Visual-Prolog. Приницпиы работы на языке пролог легли в основу декларативного программирования и других языков, использующих логический подход к программированию, таких как Mercury, Oz и ECLiPSe.

III. Теория

Работа инженера-разработчика с системой автоматизированного проектирования проходит через четыре ключевых этапа, на первом из которых определяются цели проектирования, соответствующие алгоритмы и ограничения. После чего создаются параметрические модели, которые передаются разработчику для проверки и оценки на соответствие поставленной задаче. Если требуется изменить модель, инженер-проектировщик корректирует ограничения и алгоритмы. (рис 1). [6]

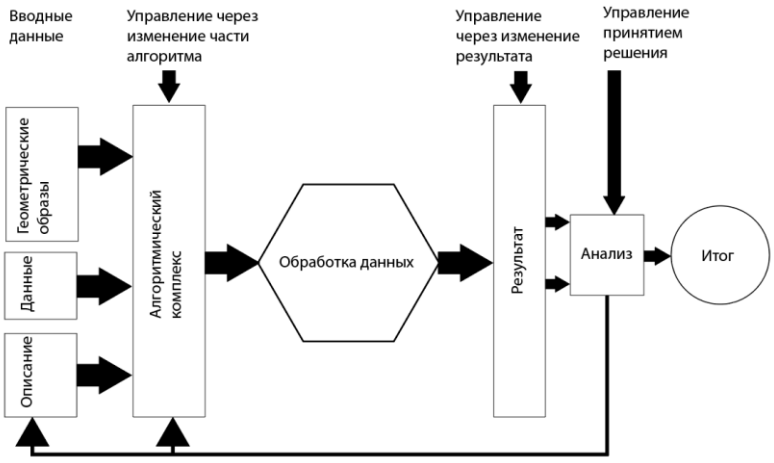


Рис. 1. Связь между компонентами генеративного проектирования и его запуска

Процесс оптимизации топологии можно считать итеративным процессом. Его цель — улучшить характеристики конструкции. Во время оптимизации определяется наилучшее распределение материалов и оптимальные изменения структуры с учётом нагрузок. Внешний вид конструкции меняется под влиянием граничных условий и ограничений, установленных разработчиками.

В последние годы было разработано множество методов оптимизации проектирования, направленных на объединение подходов к оптимизации топологии, формы и размера для избежания локальных оптимумов. Основные категории включают предопределённую практику проектирования пространства, максимально возможную практику проектирования пространства с учётом граничных условий и комплексную оптимизацию формы и топологии.

Оптимизация топологии — это итеративный процесс, направленный на улучшение характеристик конструкции. Он позволяет определить наиболее подходящее распределение материалов и оптимальные изменения структуры с учётом нагрузок. Внешний вид конструкции изменяется под воздействием граничных условий и ограничений, заданных разработчиками.

За последнее время разработано множество методов оптимизации проектирования, направленных на объединение подходов к оптимизации топологии, формы и размера во избежание локальных оптимумов. Выделяют три основных категории: предопределённая практика проектирования пространства, максимально возможная практика проектирования пространства с учётом граничных условий и комплексная оптимизация формы и топологии.(рис. 2). [7]

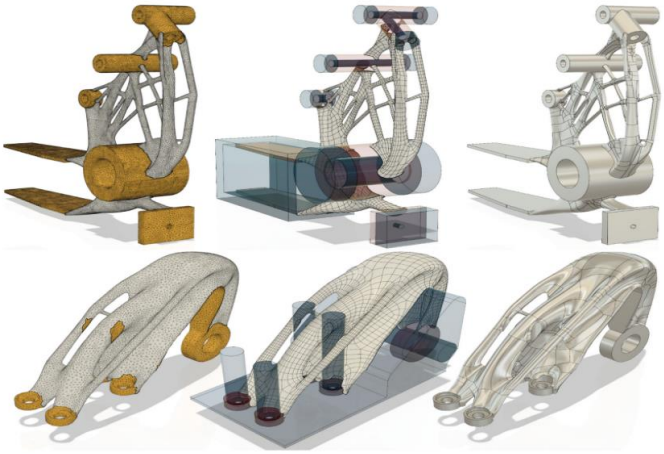


Рис. 2. Пример оптимизации топологии

С помощью геометрического моделирования возможно управлять процессом генерации форм объектов, изменять параметры в геометрическом представлении, что позволит с высокой точностью влиять не только на выбор решения в результате определенной итерации, но и на процесс генерации формы в целом.

IV. Результаты экспериментов

В качестве примера для исследования взята геометрическая структура инволюции. Используя программный модуль Логика была написана небольшая последовательность действий для редактирования геометрической структуры.

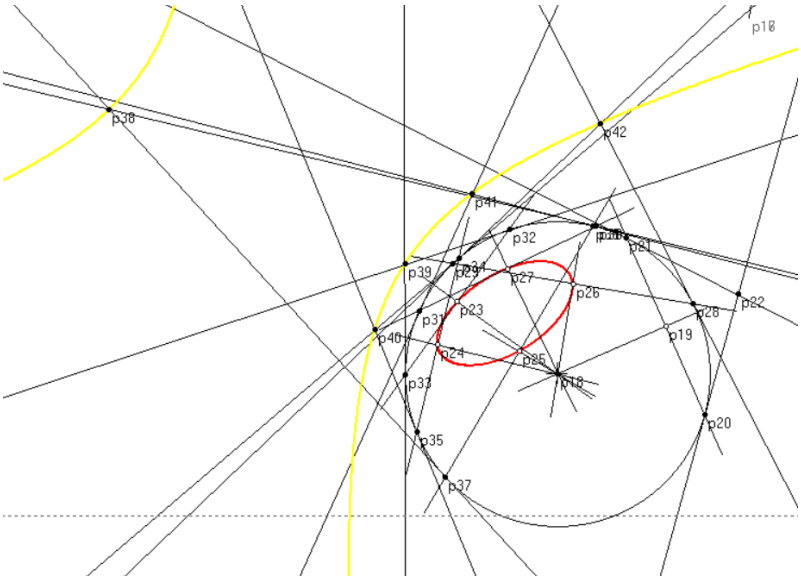


Рис. 3. Геометрическая модель в Симплексе

На рисунке 3 представлена модель, собранная в Симлексе. Для изменения геометрического образа программа остается прежней. Меняются только точки. Со сменой точек меняется и геометрический образ. Таким образом мы программно задаем изменение геометрии, не влияя на нее непосредственно с помощью курсора

Программа составлена следующим образом:

```
member(Y,[25,30,40,50,80]),
writeln(Y),
addline([[Y], [20]],   [p27],   '0', 'Точка задана координатами', 'Главный').
make if
member(X,[25,30]),
writeln(X),
addline([[X], [20]],   [p24],   '0', 'Точка задана координатами', 'Главный').
make if
member(Y,[25,30]),
writeln(Y),
addline([[Y], [20]],   [p24],   '0', 'Точка задана координатами', 'Главный').
```

Где

member(Y,[25,30,40,50,80]) является предикатом проверки того, что Y входит в состав списка, представленного в квадратных скобках.

writeln(Y) является предикатом вывода на экран значения аргумента с переходом на следующую строку.

addline([[Y], [20]], [p27], '0', 'Точка задана координатами', 'Главный') - добавляет отношения в алгоритм Симплекса, где

[[Y], [20]] - это входной параметр отношения,

[p27] - выходной параметр отношения

'0' - тип согласования

'Точка задана координатами' - функция

'Главный' - алгоритм

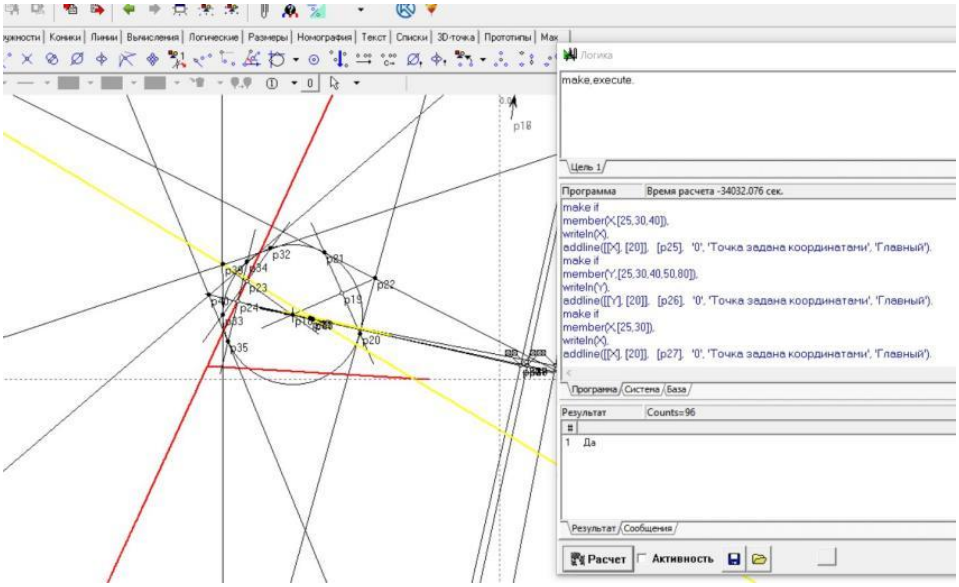


Рис. 4. Трансформация модели с использованием точек p25, p26, p27

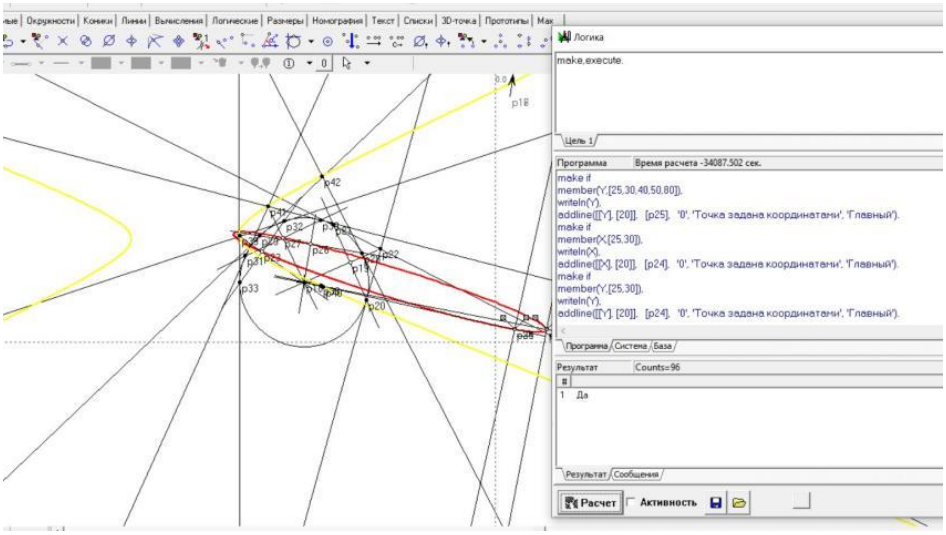


Рис. 5. Трансформация модели с использованием точек p24 и p25

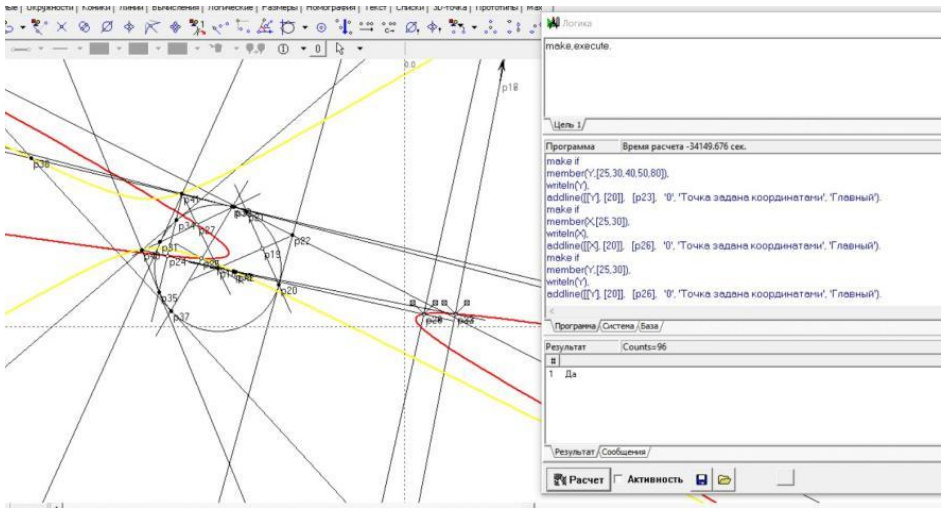


Рис. 6. Трансформация модели с использованием точек p23 и p26

V. Обсуждение результатов

Использование модуля логика в процессах динамического преобразования геометрических структур позволяет получать результаты, которые можно исследовать и использовать в генеративном проектировании новых поверхностей и форм. Учитывая декларативный характер способа получения новых геометрических образов, можно рассуждать о формировании экспертной системы нового порядка, способную анализировать загружаемые чертежи и определять геометрические структуры, которые требуются для решения определенных задач.

VI. Выводы и заключение

Использование модуля логика в процессах динамического преобразования геометрических структур оправдана и позволяет рассуждать о интересных результатах в дальнейшем. С помощью данной системы можно анализировать сложные геометрические образы, находить соответствия между имеющимися алгоритмами и разрабатываемыми. Таким образом в работе рассматривается возможность редактирования геометрических образов с помощью логических функций в программе Симплекс.

В дальнейшем планируется продолжить исследования в представленной области для получения практических результатов.

Источник финансирования. Благодарности

Научный руководитель Волошинов Денис Вячеславович, СПбГУТ им. Бонч-Бруевича, СПбПУ Петра Великого.

Список литературы

1. Shahroz Khan, Muhammad Junaid Awan, A generative design technique for exploring shape variations, Advanced Engineering Informatics,Volume 38, 2018, Pages 712-724, ISSN 1474-0346, https://doi.org/10.1016/j.aei.2018.10.005.

2. Coelho H., Pereira L. M. Automated reasoning in geometry theorem proving with Prolog //Journal of Automated Reasoning. – 1986. – Т. 2. – С. 329-390.

3. Replogle J. A. et al. Extracting isovolumes from three-dimensional torso geometry using PROLOG //IEEE transactions on information technology in biomedicine. – 1998. – Т. 2. – №. 1. – С. 10-19.

4. Волошинов, Д. В. Разработка алгоритмического комплекса для решения задач конструктивной геометрии / Д. В. Волошинов, А. В. Соловьева // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2020) : Сборник научных статей IX Международной научно-технической и научно-методической конференции. В 4-х т., Санкт-Петербург, 26–27 февраля 2020 года. Том 2. – Санкт-Петербург: Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, 2020. – С. 182-186. – EDN VRYFMK.

5. Bratko I. Prolog programming for artificial intelligence. – Pearson education, 2001.

6. Ma, W.; Wang, X.; Wang, J.; Xiang, X.; Sun, J. Generative Design in Building Information Modelling (BIM): Approaches and Requirements. Sensors 2021, 21, 5439. https:// doi.org/10.3390/s21165439.

7. Marinov M. et al. Generative design conversion to editable and watertight boundary representation //Computer-Aided Design. – 2019. – Т. 115. – С. 194-205.

УДК 087.2
DOI: 10.25206/978-5-8149-3873-2-2024-843-853

Имплементация в учебный процесс вуза тестирования нового САПР «САРУС»
Implementation of testing of the new SARUS CAD system into the educational process of the university

Н.А.Денисова[1], Г.А. Федоренко[1]
[1] Саровский физико-технический институт - филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский ядерный университет "МИФИ" (СарФТИ НИЯУ МИФИ), ул. Духова, д.6, г. Саров, 607186, РФ

N.A.Denisova[1], G.A. Fedorenko[1],
[1] Sarov Institute of Physics and Technology - branch of the Federal State Autonomous Educational Institution of Higher Education "National Research Nuclear University "MEPhI" (SarPhTI NRNU MEPhI), Dukhova str., 6, Sarov, 607186, Russian Federation

Аннотация. В Институте цифровых технологий РФЯЦ-ВНИИЭФ разрабатывается Комплекс программ в защищенном исполнении «Система полного жизненного цикла изделий «Цифровое предприятие» - система автоматизированного проектирования (САПР), разработанной на базе интегрированной инженерной программной платформы и программно-математического ядра трехмерного моделирования «RGK», Россия. На данный момент версия 3 «СПЖЦ «Цифровое предприятие», получившая коммерческое название «САРУС», не запущена как инструмент работы на предприятии, она находится на тестировании и в интенсивном развитии по его результатам. Студенты кафедры технологии специального машиностроения СарФТИ-НИЯУ МИФИ принимают участие в тестировании Программного модуля CAD. Тестирование проводится в сравнении нескольких версий «САРУС» по мере их выхода. Превалирующая цель исследования – возможности применения операций, предлагаемых интерфейсом ПМ CAD, по построению криволинейных поверхностей в 2D и 3D режимах.

Ключевые слова: САПР, Программный модуль CAD, 3D моделирование, Конструирование в машиностроении, Тестирование, Учебный процесс

Abstract. The Institute of Digital Technologies of the RFNC-VNIIEPh is developing a set of programs in a secure version "Digital Enterprise Product Full Life Cycle System" - a computer-aided design (CAD) system, developed based on the integrated engineering software platform and the software and mathematical core of three-dimensional modeling "RGK", Russia. At that moment, version 3 of the “FLCS "Digital Enterprise", which has received the commercial name "SARUS", has not been launched as a tool for working at the enterprise, it is being tested and under intensive development based on its results. Students of the SarPhTI Department of Special Engineering Technology take part in testing the CAD Software Module. Testing is underway in comparison of several versions of SARUS as they are releasing. The prevailing purpose of the study is the possibility of using the operations offered by the PM CAD interface to construct curved surfaces in 2D and 3D modes.

Keywords: Automated system CAD, CAD software module, 3D modeling, Design in mechanical engineering, Testing, Educational process.

I. Введение

В докладе на 33-й Международной конференции по компьютерной графике и машинному зрению GraphiCon в 2023 году было заявлено, что на базе центра цифровых технологий СарФТИ-НИЯУ МИФИ в сотрудничестве с кафедрами «Технология специального машиностроения» и «Цифровые технологии» организована экспериментальная площадка, работа которой направлена на подготовку пользователей, изучение функциональных возможностей и алгоритмов команд, проведение функционального тестирования разрабатываемой ФГУП «РФЯЦ-ВНИИЭФ» системы полного жизненного цикла изделий «Цифровое предприятие»» (далее по тексту - «САРУС»). В рамках площадки студенты кафедры ТСМ выполняют проблемно-ориентированные исследования, направленные на развитие системы в области программного модуля «Система конструкторского проектирования» (далее по тексту – ПМ CAD), который является компонентом «САРУС» и представляет собой CAD-систему [1].

В результате начального тестирования ПМ CAD отмечено положительное отличие «САРУС» – наличие технологической платформы (ТП). Все необходимые программные модули расположены на одной ТП, что дает

возможность разрабатывать новые и корректировать уже работающие программные модули на единой и открытой архитектуре. ТП обеспечивает возможность расширения состава задач за счёт разработки дополнительных программных модулей (приложений), решающих новые задачи на основе базовых платформенных инструментов. Средства ТП должны предоставлять инструменты, обеспечивающие встраивание в конечную систему или комплекс систем дополнительных программных модулей, и средства, обеспечивающие их совместное (интегрированное) функционирование.

По окончании первого этапа тестирования сделано заключение, что «САРУС» нам понравился как «головоломка» - интересно искать решение проблем нестандартными способами. Такая деятельность полезна и в учебном плане – она ведет к рефлексивному восприятию создания 3D-моделей изделий машиностроения при применении цифровых технологий. По мнению большинства студентов, при работе в ПМ возникает интерес (даже азарт) поиска решений. Чтобы осуществить многие операции, нужно знать алгоритм действий. Интуитивный поиск мало что дает. В этом случае можно обратиться к справке, предусмотренной в модуле, или искать решение самостоятельно [1].

Исходя из аналитического обзора в отчетах студентов-тестировщиков, мы обратили внимание не некоторые интересные факты применения базовых операций/построений 3D-модели в «САРУС» с использованием алгоритмов функций, заявленных разработчиками. Студентов заинтересовали характерные при применении данного ПМ CAD особенности, отличающиеся при построении в сравниваемом ПО «КОМПАС-3D»:

- *Роль системы координат* (далее - СК). Одна из особенностей построения 3D-моделей заключается в способе задания поверхностей, которое осуществляется через указание координат в пространстве. На начальном этапе моделирования конструктор задаёт СК, которой следует на протяжении всего процесса. Эта геометрическая особенность, ставит проблему: привязанность к СК является несомненным плюсом при создании операций механической обработки модели и управляющей программы обработки в модуле «Система разработки управляющих программ с ЧПУ» (ПМ CAM), но повышает трудоёмкость работы конструктора, так как все размеры модели также выдерживаются с помощью координат.
- *Применение булевой операции (далее – БО)*. Нет автоматического объединения элементов модели, построенных с помощью операции «Примитив», в одну деталь, поэтому их надо постоянно объединять с помощью булевой операции в одно тело. Возможно, такой подход является плюсом при изготовлении сложных деталей или в случаях, когда автоматическое объединение в один объект может помешать проектированию.
- *Построение скруглений* средствами ПМ CAD. При проектировании деталей сложных форм возникают проблемы с функцией «Скругление». Выявлено, что программные сбои по операции «Скругление» встречаются в обеих сравниваемых CAD-системах («САРУС» и «КОМПАС-3D»). Проанализировать работоспособность алгоритмов построения скруглений является достаточно актуальной задачей.

Выявленные факты при их осознании и аналитическом обосновании при исследовании имеют существенное значение не только для организации процесса работы непосредственно ПМ CAD, но и в повышении образовательного уровня студентов.

II. Постановка задачи

Для ориентации тестирования на учебную работу определили ***цель***: имплементация тестирования в учебный процесс при исследовании возможности применения операций, предлагаемых интерфейсом ПМ CAD, входящему в систему «САРУС», по построению криволинейных поверхностей в 2D и 3D- режимах в деталях *простой формы* с теоретическим анализом предлагаемых функций построения.

Задачи:

- Тестирование ПМ CAD по срабатыванию выполнения функций, операций и команд построения, рекомендуемых разработчиком при построении 3D моделей, исходя из конструктивной сложности деталей.
- Исследование оптимальных алгоритмов построения криволинейных поверхностей в ПМ CAD «САРУС» в 3D-моделях простых деталей.
- Анализ роли СК при построении 3D-моделей в сравнении с «КОМПАС-3D».
- Анализ применения булевых операций при построении 3D-моделей в сравнении с «КОМПАС-3D».
- Анализ работоспособности алгоритмов построения скруглений при построении 3D-моделей в сравнении с «КОМПАС-3D».
- Подготовка методических материалов по обучению работы с ПМ CAD «САРУС» студентов по направлению подготовки 15.03.04 – Конструкторско-технологическое обеспечение машиностроительных производств, а также специалистов, имеющих опыт работы с рядом САПР в направлении конструкторской подготовки.

III. Теория

III.1. Описание классификации деталей по конструктивной сложности

На данном этапе в соответствии с задачами объектом нашего исследования определено построения 3D-моделей деталей I и II групп конструктивной сложности:

- Детали геометрических форм, не требующие расчетов и имеющие вспомогательное значение (кольца, прокладки, втулки, шайбы, болты и др.).
- Детали простых геометрических форм, но требующие несложных расчетов (рычаги, фланцы и др.).

Здесь нас интересуют методы и приёмы формообразования поверхностей деталей. Из рассмотренных классификационных характеристик можно сделать вывод, что вся совокупность деталей разбивается на классы, подклассы, группы по определенным конструкторским и технологическим признакам. Этими признаками являются конфигурация детали, ее размеры, точность обработки, качество поверхности, материал детали и его свойства. В группе деталей рекомендуется выбирать типовую деталь, на которую разрабатывается типовой технологический процесс. Процесс типизации не стандартизирован ввиду большой номенклатуры изделий, различных по форме, размерам и свойствам. Задачи типизации решаются на отраслевом уровне или каждым предприятием самостоятельно. Например, классификация деталей, предложенная профессором А.П. Соколовским, предусматривает 14 классов. Эта классификация касается деталей общего назначения и включает следующие классы: валы, втулки, диски, эксцентричные детали, крестовины, рычаги, плиты, стойки, угольники, бабки, зубчатые колеса, фасонные кулачки, ходовые винты и червяки, мелкие крепежные детали. Каждый класс может быть также классифицирован в зависимости от конструктивной сложности и точностных характеристик. Классификация позволит нам ограничить число деталей, для которых требуется создать 3D-модели.

В CAD-системах метод создания моделей можно сравнить с обработкой заготовок деталей резанием, где постепенно удаляется припуск на обработку и, при достижении заданных чертежом размеров, получаются элементарные поверхности, которые при их сочетании формируют деталь. Есть ещё один метод: применение операции «Примитив», где модель создается постепенно добавлением простейших геометрических фигур – элементов детали, при котором целостная модель получается сложением, вычитанием или пересечением данных элементов друг с другом с помощью булевых операций.

Тогда классификация создания 3D-моделей может выглядеть следующим образом:

1. Группа команд «3D-операции», объединяющая традиционные команды твердотельного моделирования
   1.1 Команда «Вытягивание» предназначена для создания твердого или поверхностного тела, которое представляет собой след от движения контура в заданном направлении на заданную длину.
   1.2 Команда «Вращение» предназначена для создания тела как результата вращения контура относительно заданной оси на заданный угол.
   1.3 Команда «По сечениям» предназначена для создания тел со сложной геометрией.
   1.4 Команда «По траектории» предназначена для создания тела путем перемещения контура вдоль пространственной траектории.
2. Команда «Примитив» предназначена для создания базовых твердых тел.
3. Команда «Булева» предназначена для создания тела на основе двух или более ранее созданных тел.
4. Другие команды («Обрезка», «Скругление», «Фаска», «Оболочка», «Отверстие», пр.), способные окончательно сформировать форму детали согласно чертежу.

Исходя из предложенных классификаций выбрали для исследования функций по созданию 3D-моделей *простые* корпусные детали: крестовины, рычаги, плиты, стойки, примеры которых показаны на рис. 1. Основные команды при их выполнении – 1.1, 2, 3, 4.

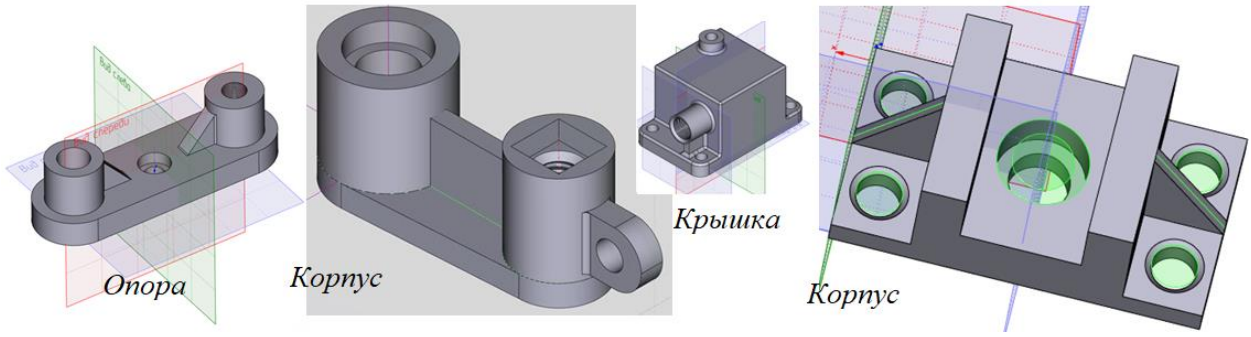


Рис. 1. Примеры 3D-моделей простых корпусных деталей, созданных в ПМ CAD ПО «САРУС»

III.2. Основы математического механизма ПМ CAD, требуемые для исследования

Поставленная нами цель исследования требует сделать акцент на математический уровень. В Пояснительной записке технологического процесса (далее – ПЗ ТП) [2; 176] дано подробное его описание. Мы сделали выборку компонентов математического уровня, интересную нам при проведении второго этапа и формулированию аналитических заключений.

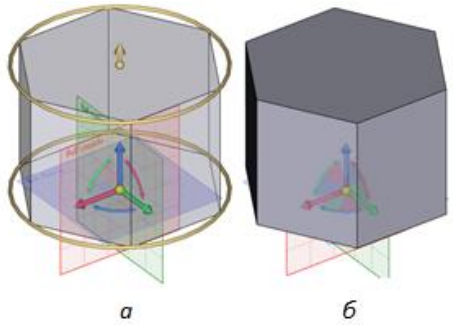
В состав CAD системы на уровне технологической платформы реализован базовый математический компонент – 3D ядро, предназначенное для использования в прикладных системах, включая CAD систему, для создания, хранения, восстановления и расчётов трёхмерных моделей физических объектов. Оно обеспечивает работу с 3D геометрией и топологией объектов моделирования. 3D ядро является программной библиотекой, не имеющей какого-либо пользовательского интерфейса. Имеет набор программных интерфейсов (классов, функций, структур данных) для создания приложений [каркасного (проволочного), поверхностного и] твердотельного моделирования.

Из трёх основных типов моделей – проволочные, поверхностные и твердотельные [2], в нашей работе интерес представляют твердотельные модели. Это наиболее часто используемый тип моделей объектов проектирования. Твердотельными моделями наиболее адекватно представляются объекты реального мира. Они служат основой для описания объектов проектирования изделий не только машиностроительного, но и прочих промышленных объектов.

IV. Результаты экспериментов

IV.1. Локальная система координат [3]

В САПРах система координат (далее - СК) играет важнейшую роль при построении 3D-моделей. На начальном этапе моделирования конструктор задаёт СК, которой следует на протяжении всего процесса. Эта геометрическая особенность определяет ***противоречие*** – своеобразную «разность потенциалов»: привязанность к СК является несомненным плюсом при создании операций механической обработки модели и управляющей программы обработки в модуле «Система разработки управляющих программ с ЧПУ» (ПМ CAM), но повышает трудоёмкость работы конструктора, так как все размеры модели также выдерживаются с помощью координат. Устранение этого противоречия не является очевидным, поэтому мы имеем ***проблему***, для которой надо найти решение.



В 3D-моделировании локальная система координат (ЛСК) является неотъемлемой частью математического аппарата. С её помощью определяется положение точки в пространстве. На рис. 1, *а* показано, как задаётся СК по умолчанию при построении призмы, а рис.1, *б* – автоматическое создание ЛСК при способе задания “Система координат по объекту”. Как видно, системы координат по умолчанию совпадают.

Рис. 1. СК по умолчанию и автоматическое создание ЛСК

Разработчиками также представлены требования к функции «Построение системы координат». По документам разработчика функция должна обеспечивать как общие требования, так и особые требования, присущие каждому из четырех способов построения СК.

Одна из особенностей построения 3D-моделей заключается в способе задания поверхностей, которое осуществляется через указание координат в пространстве. Конструктор задаёт СК, которой следует на протяжении всего процесса. В ПЗ ТП, часть 4 представлены требования к функциям, описанным в разделе 1.2, которые должны выполняться при работе пользователей в «САРУС» [4]. Реализация этих функций была проверена в процессе построения модели ***простой*** детали (II группа конструктивной сложности).

В результате аналитического сравнения предлагаемого алгоритма функций и выполняемого алгоритма построения выявлено, что все СК, созданные в процессе работы с моделью, доступны в «Навигаторе» модели. К каждой системе координат можно обратиться в любой момент для выполнения различных задач моделирования [5]. Сделан вывод, что точная привязка к конкретной системе координат на начальном этапе создания 3D-модели делает дальнейшую работу с моделью с использованием принципа сквозного проектирования на платформе «САРУС» более удобной. Ожидается, что в дальнейшем, при корректировке и доведении СПЖЦ «САРУС» до совершенства производительность, комфортность и эффективность работы конструктора будут достигнуты.

IV.2. Булевы операции [6]

Исходя из аналитического обзора в отчетах студентов-тестировщиков, мы обратили внимание не некоторые интересные факты применения базовых операций/построений 3D-модели в «САРУС» с использованием алгоритмов функций, заявленных разработчиками. Студентов заинтересовали характерные при применении данного ПМ CAD особенности применения булевых операций, отличающиеся при построении 3D-моделей в сравниваемом ПО «КОМПАС-3D»: нет автоматического объединения элементов модели, построенных с помощью операции «Примитив», в одну деталь, поэтому их надо постоянно объединять с помощью булевой операции в одно тело. Возможно, такой подход является плюсом при изготовлении сложных деталей или в случаях, когда автоматическое объединение в один объект может помешать проектированию [5].

Создавшиеся предпосылки определили цель практико-ориентированного исследования - раскрыть значение и возможности применения булевых операций при построении 3D-моделей в программном обеспечении «САРУС», обозначить основные инструменты и алгоритмы, используемые для выполнения этих операций, а также сравнить реализацию булевых операций в ПО «САРУСс» и «КОМПАС-3D».

Как отмечают специалисты, «основной задачей булева алгоритма является конструирование границ финальной модели. Иногда эти алгоритмы называют алгоритмами вычисления и слияния границ, что явно подчеркивает их чисто топологическую натуру. Дело в том, что «сверхзадачей» булева алгоритма является способность работать и выдавать адекватный результат на произвольных пользовательских данных» [6].

Используя булевы операции, построена 3D-модель деталь «Опора» (рис. 2). Эту деталь можно отнести к II группе конструктивной сложности – простые детали. Необходимо отметить, что при построении 3D-модели детали «Опора» со сравнительно простой геометрией трудностей не возникло. Однако вероятно, некоторые функциональные возможности, которые в настоящее время отсутствуют, облегчили бы процесс построения и, с большой вероятностью, были бы необходимы при построении деталей со сложной геометрией.

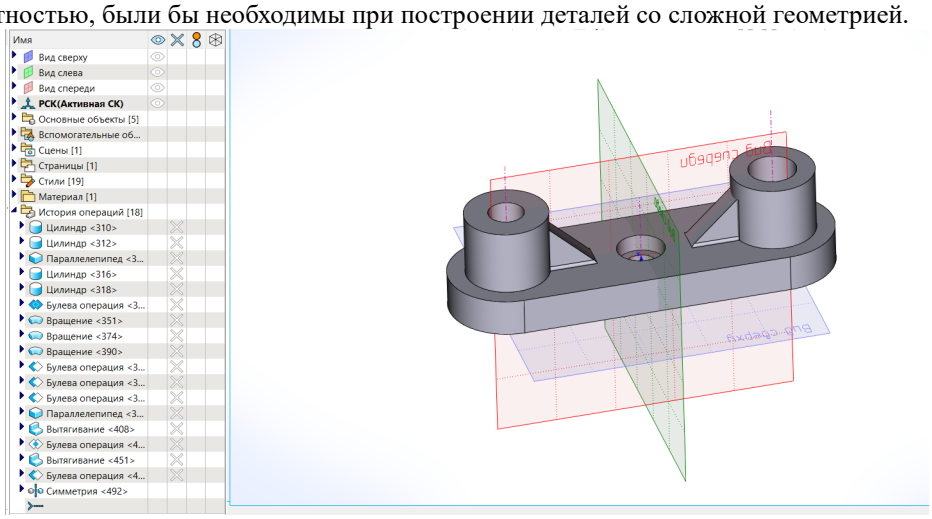


Рис. 2. 3D-модель детали «Опора», созданная в ПО «САРУС» с применением булевых операций

После работы с программой «САРУС» и изучения особенностей применения булевых операций, можно сделать вывод о том, что программа потенциально пригодна для работы, и особенность применения функций БО по сравнению с КОМПАС-3D вызывает интерес. Это отличает «САРУС» от других программ, где БО выполняются автоматически. Кроме того, программа предоставляет возможность детального контроля над процессом создания детали, что может быть полезно для инженеров, работающих со сложными проектами.

Сделанный вывод коррелируется с мнением специалистов, что в CAD функциональность БО дает пользователю полную свободу в формообразовании, не привлекая каких-либо кинематических соображений, она больше рассчитана на творчество инженера конструктора. Их функционал очень гибок, инженер использует его для достижения нужных результатов. Он может получить то, что хочет, потому что автоматическое выполнение операций может не дать нужного результата [5]. Большой набор булевых операций позволяет конструкторам экспериментировать с различными формами и структурами, что может привести к более эффективному и оригинальному решению. Это особенно важно в условиях жесткой конкуренции на рынке, когда необходимо предлагать инновационные продукты и услуги.

В целом, такой функционал БО делает CAD-систему в СПЖЦ «САРУС» незаменимым инструментом для конструктора, позволяя ему проявить свое творчество и умение находить нестандартные решения. Это способствует развитию новых технологий и улучшению качества продукции.

IV.3. Операции скруглений [7]

Математики отмечают, что одна из самых труднорешаемых задач в системах автоматизированного проектирования – построение скруглений при моделировании объектов сложных форм. За построение скруглений, как и за всю геометрию в САПР, отвечает геометрическое ядро. [8]

В процессе тестирования исследован метод построения скруглений средствами ПМ CAD, проведен сравнительный анализ с другими известными САПР (КОМПАС 3D), подробно описан оптимальный порядок построения скруглений, с учетом выявленных особенностей в работе команд.

На рисунке 3 показаны детали, построенные с использованием операции «Скругление» в «САРУС» (*а*) и КОМПАС 3D (*б*) соответственно. Как видно из данного простого примера – программные сбои по операции «Скругление» встречаются в обеих CAD-системах. Задача данной работы – проанализировать работоспособность алгоритмов построения скруглений.

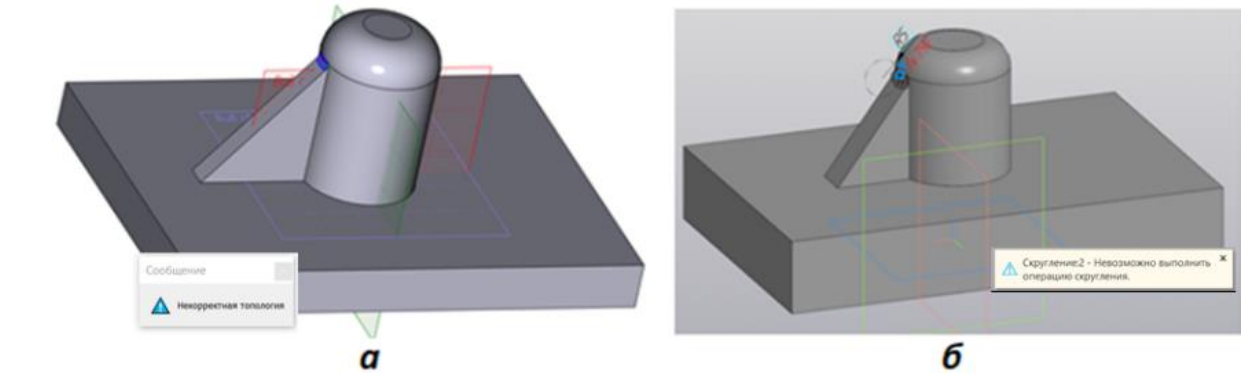


Рис. 3. 3D-модели: ***а*** – построена в СПЖЦ «САРУС»; ***б*** – построена в КОМПАС-3D

На простейшей фигуре «Параллелепипед» удалось реализовать 9 алгоритмов построения скруглений различной топологии, за исключением «Управление опциями перекрытия скруглений - взаимная обрезка или сопряжение скруглений с наращиванием материала». Можно заключить, что построение скруглений на поверхностях деталей простой формы (I, II групп конструктивной сложности) в основном возможно.

Для демонстрации работы алгоритмов была взята композиция примитивов со сложными переходами. Результаты работы функции «Скругление» с заданными выше параметрами показаны на тестовой детали, приведенной на рис. 4.

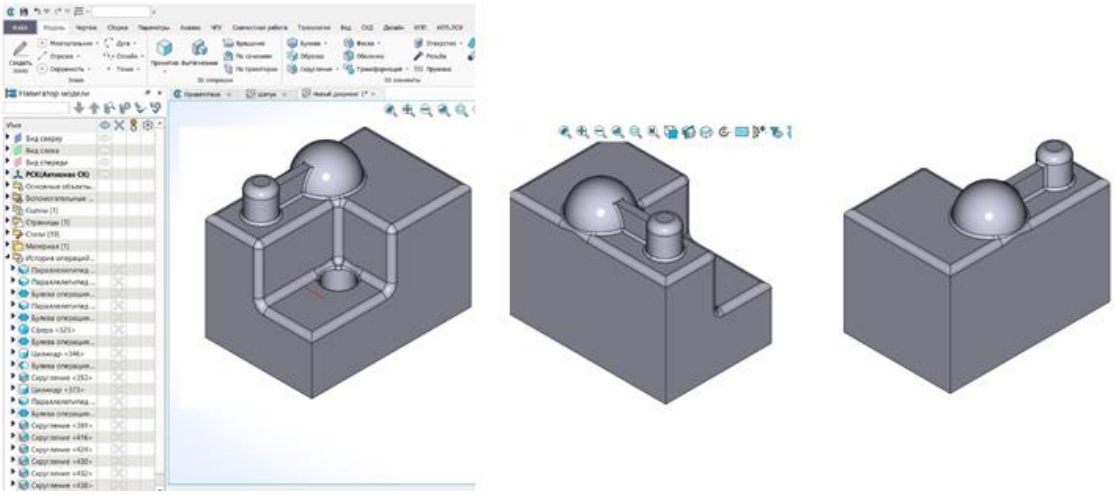


Рис. 4. Тестовая деталь в 3 видах для демонстрации функции «Скругление»

На рисунках видно, что в данном построении не возникло проблем с пересечениями скруглений, СПЖЦ «САРУС» корректно построил все скругления ребер.

Чтобы продемонстрировать работу скруглений с переменным радиусом, была построена деталь «Шатун» (рис. 5).

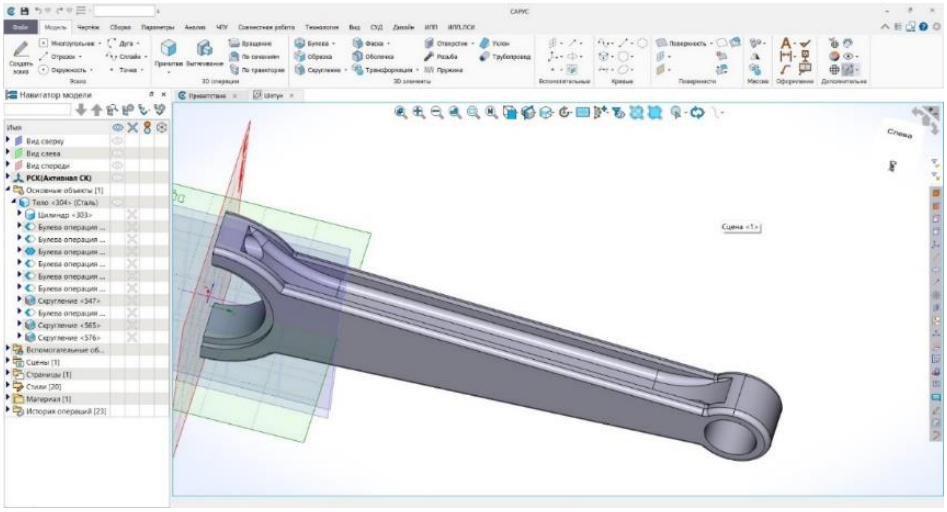


Рис. 5. 3D-модель детали «Шатун»

Показано, что в деталях с усложнением элементов форм все алгоритмы функции «Скругление», работают корректно, даже в местах пересечений скруглений, чего, кстати, не удалось сделать на детали простейшей формы (I группы конструктивной сложности).

Был проведен сравнительный анализ с известными САПР КОМПАС 3D, зарекомендовавшим себя результативным в работе на базовом предприятии. Здесь у СПЖЦ «САРУС» были свои достоинства и недостатки. К первому относятся более корректное построение скругления с переменным радиусом, а также возможность построения скругления с перекрытием отверстия (скругление повторяет форму отверстия), что в КОМПАС-3D не получилось сделать. Из недостатков СПЖЦ «САРУС» по отношению к КОМПАС-3D можно отнести отсутствие некоторого функционала, который есть в КОМПАС-3D, а именно: полное скругление, остановка скругления в точке и обход с гладкой стыковкой на пересечении 2 ребер. Также КОМПАС-3D корректно выполнил алгоритм управления опциями перекрытия скруглений - взаимной обрезки или сопряжения скруглений с наращиванием материала. Однако и в той, и в другой САПР есть факты несрабатывания функций.

Работать с функциями «Скругление» в СПЖЦ «САРУС» возможно на базовом уровне программного обеспечения CAD.

V. Обсуждение результатов

На данном этапе в соответствии с задачами исследования выполнен ряд построений 3D-моделей деталей I и II групп конструктивной сложности. По заданным чертежам было построено 14 моделей с применением различных алгоритмов построения в ПО «САРУС» и ПО «КОМПАС-3D». По результатам построения было проведено аналитическое сравнение алгоритмов.

На этапе 2 у нас уже появились алгоритмы построения 3D-моделей, выполненных и описанных студентами, а также рекомендации после исследований, описанных в разделе IV. Поэтому определили следующие критерии: функциональное назначение конкретных операций и исследование двух алгоритмов построения в «САРУС» и сравнение их с алгоритмом построения в «КОМПАС-3D»; удобство и понятность интерфейса, интуитивный уровень его освоения; производительность - быстрота построения одинаковых деталей.

Аналитическое сравнение в данном случае также имеет качественный характер. Даже производительность зависит не только от количества операций, необходимых для построения 3D-модели. Здесь большую роль играет опыт конструктора, умение применить оптимальный алгоритм построения, ведь одну и ту же модель можно создать, используя разные операции.

Приведем анализ практико-ориентированного исследования построения 3D-модели детали «Корпус» (рис. 6). Критериями сравнения назначим трудоёмкость и удобство (комфортность) работы. Алгоритмы построения сведём в Табл. 1.

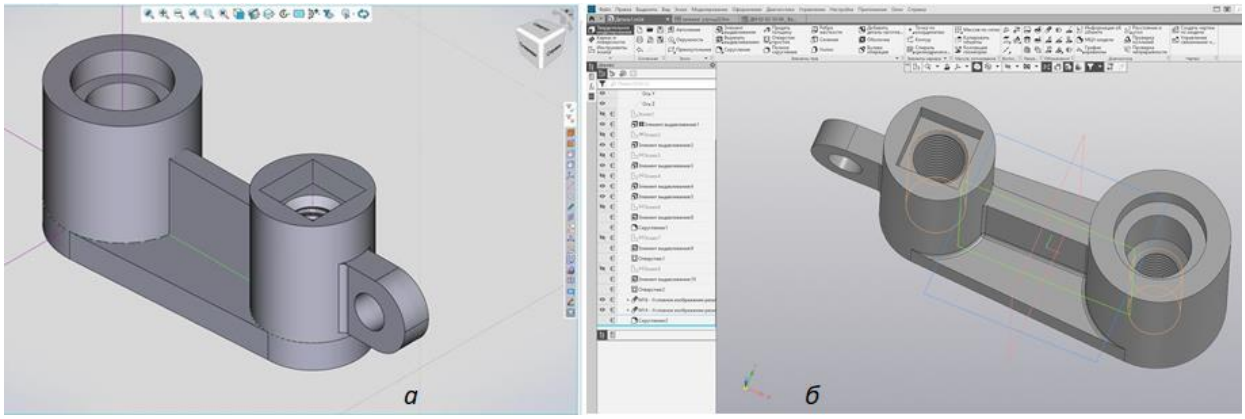
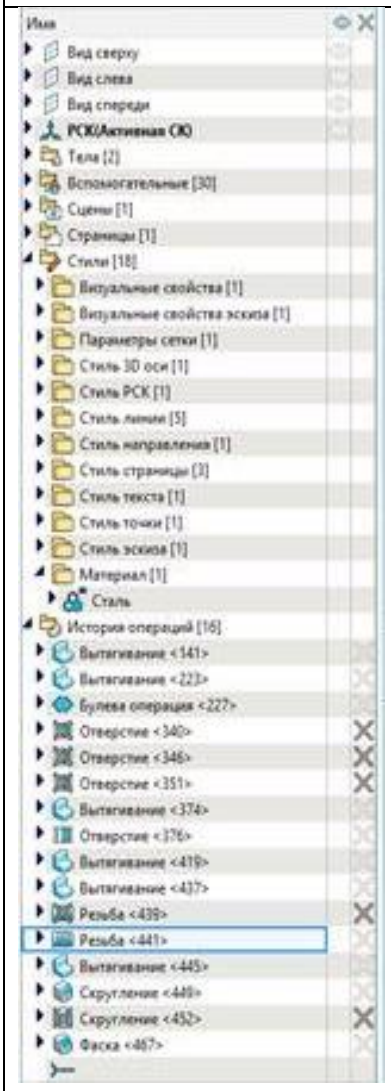
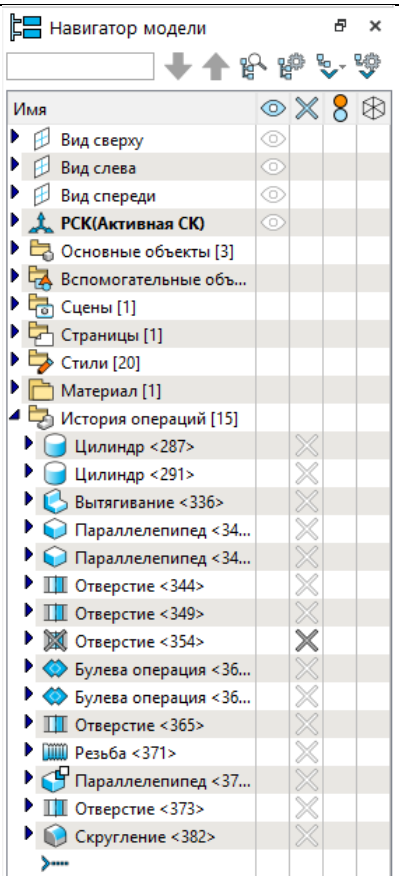
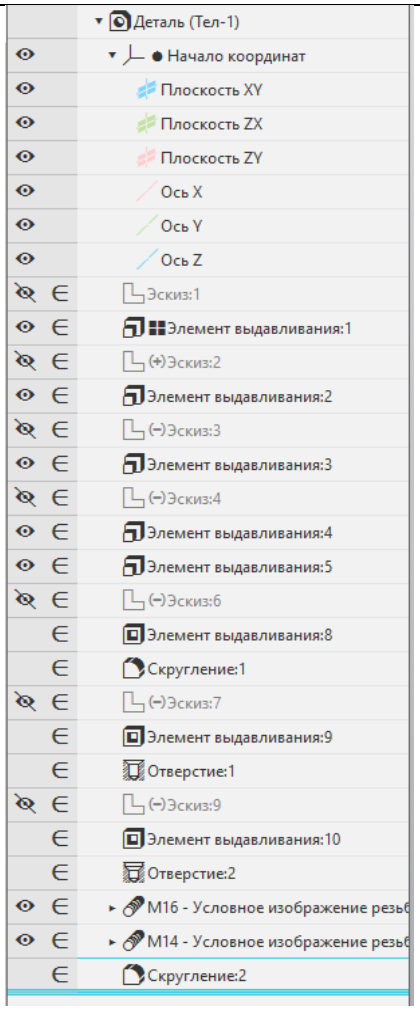


Рис. 6. 3D-модель детали «Корпус»: а – выполнена в ПМ CAD «САРУС»; б – выполнена в «КОМПАС-3D»

ТАБЛИЦА 1
СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ ПОСТРОЕНИЯ 3D-МОДЕЛИ ДЕТАЛИ «КОРПУС»

СПЖЦ способ I	**СПЖЦ способ II**	**«Компас»**
		

Количество операций в способе 1 –16; в способе 2 – 15; в «КОМПАС» 17. Как видим, трудоёмкость примерно одинаковая. Однако, по мнению студентов-тестировщиков, выполнение детали «Корпус» проще сделать в СПЖЦ путем создания примитивов.

Аналогично отработано построение 3D-моделей 16-ти корпусных деталей, соответствующих II группе конструктивной сложности с использованием двух алгоритмов построения: по эскизу и по примитивам, после чего проведено аналитическое сравнение трудоёмкости построений. Кроме того, в анализ включено построение тех же моделей в САПР «КОМПАС», как практически наиболее освоенного студентами ПО CAD.

Оценка студентов-тестировщиков разошлась по поводу алгоритмов построения в «САРУС»:

1. Алгоритм построения по примитивам был проведён в версии программного комплекса СПЖЦ «САРУС» 1.0.17.9; Алгоритм построения по эскизам был проведён в версии программного комплекса СПЖЦ «САРУС» 2023.1.0.1-RC16.

В первом способе использовались примитивы с дальнейшим выполнением команды «Булева операция», а именно «Сложение» или «Вычитание». Данный способ построения показался достаточно сложным, т.к. необходимо знать все координаты точек детали. При построении по эскизам использовались лишь построения эскизов с дальнейшим выполнением команды «Вытягивание», далее - операциями скругление, фаска и резьба. По мнению тестировщика, данный способ показался удобнее и быстрее.

Субъективно оценивая два способа, можно сделать вывод о том, что интуитивно понятнее использовать алгоритм построения с помощью эскизов, так как в ПО «КОМПАС», используется именно этот способ. Хочется отметить, что построение модели с помощью примитивов показалось достаточно интересным, так как это принципиально новый способ построения 3D-моделей.

2. Способ построения по примитивам позволяет не только построить модель со значительно меньшим количеством операций (например, 25 операций по сравнению с 45 операциями), но и более качественно, поскольку при построении отсутствуют смещения эскизов. В КОМПАС-3D та же модель строится за 11 операций. При построении модели работа в САРУС усложнена отсутствием некоторых команд (например, отсутствием возможности построить резьбовое отверстие без использования эскизов).

Проведя сравнение количества операций и затраченного на выполнения моделирования времени выяснено, что наиболее производительным на данный момент является КОМПАС-3D. Однако, сравнивать две эти программы не совсем корректно, ведь КОМПАС-3D активно используется уже не один десяток лет, тогда как «САРУС» начал разрабатываться всего несколько лет назад, и на данный момент находится на стадии доработки.

3. Построение 3D-модели по оптимальности одинаково как в СПЖЦ «САРУС» посредством создания примитивов и при создании эскизов, так и в САПР «КОМПАС». Но сравнивать «КОМПАС» и «САРУС», наверное, не является объективным, т.к. это два совершенно разных подхода к построению моделей. По мнению тестировщика, преимуществом СПЖЦ «САРУС» является совершенно другая логика построения по сравнению с САПР «КОМПАС-3D», акцент на различных привязках к линиям и поверхностям, но при этом не хватает наличия такой функции, как вспомогательные прямые, кривые и точки.

4. В СПЖЦ «САРУС» требуются многие доработки функций. Наиболее явные из них – создание скруглений, фасок и выполнение контура детали. При выполнении поправок в проекционном виде чертежа велика вероятность «вылета» программы, впоследствии восстановление наработки является трудоёмкой задачей. На взгляд тестировщика, наиболее оптимальными путями развития программы являются: проработка связей «Модель – Чертеж», доработка редактирования и внесения данных в основную надпись чертежа и свойства 3D-модели, доработка мелких замечаний (отображение резьбы на чертеже, привязка размеров, возможность ввода символов греческого алфавита, символа «°», наличие показа угла конусности.

5. В результате построения модели и создания чертежа по данной модели можно сравнить две программы. В «Компас-3D» построение делается проще и быстрее. Это может быть связано в первую очередь с тем, что программа более привычна и имеется опыт работы в ней.

В программе «САРУС» есть много интересных вещей. Создание примитивов, булева операция и другие. Однако, есть и свои минусы. Многие вещи работают некорректно. Крайне неудобно заполнение основной надписи и выбор материала детали при создании чертежа. При выполнении детали часто «вылетает» программа и появляется окно критической ошибки, при этом деталь не сохраняется.

Более производительным было создание детали в «КОМПАС-3D», т.к. время создания детали в системе «САРУС» - 4 пары (примерно 6 часов), а в системе «КОМПАС-3D» - 10-15 минут. Для работы на производстве это имеет огромное значение. Однако «САРУС» удобен для работы технологам, т.к. в этой программе предусмотрены CAD и CAM системы. При дальнейшей разработке и модернизации, ПО СПЖЦ «САРУС» способно заменить широко известные CAD-программы.

VI. Выводы и заключение

Результатом исследования на данном этапе явилась проверка алгоритма срабатывания функций при построении деталей II группы конструктивной сложности. Отчеты о проведенных исследованиях приведены в работах [3], [6], [8].

Сделаны следующие выводы.

1. При операций механической обработки модели и управляющей программы обработки в модуле «Система разработки управляющих программ с ЧПУ» (ПМ CAM) и других возможностей «САРУС» требует привязанности к единой системе СК, точности выдержанных размеров и следование стандартам при 3D-моделировании. Конструктору необходимо осознавать, какая технология применяется для создания детали, как именно обрабатывается деталь.

В результате тестирования установлено, что при создании простой детали (I, II степени сложности) функции работают исправно, быстро, удобный функционал дает возможность не затрачивать на построение ЛСК много времени. Если сравнивать построение в ПО «САРУС» и другой распространенной САПР, например, КОМПАС-3D, очевидно, что затрачивается больше времени, чтобы строить детали в соответствии с координатами, чтобы выдержать все требуемые размеры безошибочно.

В результате аналитического сравнения предлагаемого алгоритма функций и выполняемого алгоритма построения выявлено, что все СК, созданные в процессе работы с моделью, доступны в «Навигаторе» модели. К каждой системе координат можно обратиться в любой момент для выполнения различных задач моделирования. Сделан вывод, что точная привязка к конкретной системе координат на начальном этапе создания 3D-модели делает дальнейшую работу с моделью с использованием принципа сквозного проектирования на платформе «САРУС» более удобной [3].

2. Можно определить основной функционал БО при создании 3D-модели - они применяются к трехмерным объектам для управления их формой: с помощью БО можно вырезать в объекте отверстие, получить деталь сложной формы, скруглить углы и ребра и сделать многое другое. Кроме того, БО применяют, чтобы упростить геометрию объекта посредством удаления невидимых поверхностей. Значительное множество моделирующих функций (но далеко не все) технически представляют собой не что иное, как специальным образом «заряженные»

булевы операции, причем некоторые выполняются автоматически (мы даже не знаем, что реализация какой-либо команды произошла при применении булевой операции), а что-то выполняется вручную, непосредственно по решению конструктора.

Как отмечают исследователи CAD, и это подтверждается результатами проведенного исследования, «Булева операция дает инженеру полную свободу в формообразовании, не привлекая каких-либо кинематических соображений (как, например, операция «заметания» вдоль траектории). Платой за гибкость является повышенная сложность подготовки операндов, так как булев алгоритм не является предметно-ориентированным, т.е. формируя конструктивный элемент посредством БО, инженер должен сам позаботиться о соблюдении всевозможных ограничений моделирования» [6].

3. При проверке работоспособности ряда функций «Скругление» в СПЖЦ «САРУС» установлено, что все алгоритмы, кроме управления опциями перекрытия скруглений - взаимной обрезки или сопряжения скруглений с наращиванием материала оказались работоспособными. Также была продемонстрирована работа данных алгоритмов на деталях с усложненными элементами формы в СПЖЦ «САРУС». Здесь все алгоритмы работали исправно, даже в местах пересечений скруглений, чего, кстати, не удалось сделать на детали простейшей формы (I группы конструктивной сложности).

Был проведен сравнительный анализ с известными САПР КОМПАС-3D. Здесь у СПЖЦ «САРУС» были свои достоинства и недостатки. К первому относятся более корректное построение скругления с переменным радиусом, а также возможность построения скругления с перекрытием отверстия (скругление повторяет форму отверстия), что в КОМПАС-3D не получилось сделать. Из недостатков СПЖЦ «САРУС» по отношению к КОМПАС-3D можно отнести отсутствие некоторого функционала, который есть в КОМПАС-3D, а именно: полное скругление, остановка скругления в точке и обход с гладкой стыковкой на пересечении 2 ребер. Также КОМПАС-3D корректно выполнил алгоритм управления опциями перекрытия скруглений - взаимной обрезки или сопряжения скруглений с наращиванием материала. Однако и в той, и в другой САПР есть факты несрабатывания функций.

Работать с функциями «Скругление» в СПЖЦ «САРУС» возможно на базовом уровне программного обеспечения CAD [8].

С целью долговременного применения СПЖЦ «САРУС» в учебном процессе в вузе, а также обучения специалистов базового предприятия при переходе на новое ПО студентами – тестировщиками предложен ряд методических рекомендаций по построению 3D-моделей простых деталей.

Источник финансирования. Благодарности

Договор от 25 марта 2024 г. № 96-2024/244 на выполнение научно- исследовательской работы по теме: «Проведение научных исследований в интересах ФГУП «РФЯЦ-ВНИИЭФ» с использованием результатов для актуализации образовательных программ подготовки специалистов с высшим профессиональным образованием для ФГУП «РФЯЦ-ВНИИЭФ», Задача 10 «Исследование программно-аппаратных решений в области цифровых технологий, возможностей их отраслевого применения, разработка и версионное тестирование компонентов полностью защищенной системы полного жизненного цикла среднего класса СПЖЦ V3 «САРУС».

Научный руководитель Денисова Наталья Александровна, СарФТИ-НИЯУ МИФИ, кафедра Технологии специального машиностроения

Авторы благодарят:

- Кафедру цифровых технологий СарФТИ-НИЯУ МИФИ под руководством Кривошеева О.В., к.т.н., заведующий кафедрой, зам. директора ФГУП "РФЯЦ-ВНИИЭФ" по технологиям полного жизненного цикла - директор института цифровых технологий (ИТЦ) - главный конструктор СПЖЦ;
- Научно-исследовательский Отдел поддержки и внедрения Института цифровых технологий РФЯЦ-ВНИИЭФ, руководитель Лазарев Г.С.

Список литературы

1. Денисова Н.А., Федоренко Г.А. Применение в учебном процессе вуза тестирования программного модуля CAD вновь создаваемого САПР «САРУС» //GraphiCon 2023: труды 33-й Междунар. конф. по компьютерной графике и машинному зрению (Москва, 19–21 сент. 2023 г.). М.: Институт прикладной математики им. М.В. Келдыша РАН, 2023. С. 848-859.
2. Комплекс программ в защищённом исполнении «Система полного жизненного цикла изделий «Цифровое предприятие»: Программный модуль «Система конструкторского проектирования». Пояснительная записка технического проекта- 07623615.00423-12 81 01.Часть 1. 2021 г. 924 с.
3. Прыткова Ю.Б., Денисова Н.А. Роль системы координат в построении 3D-модели в СПЖЦ «САРУС» // Отчёт о НИР «Исследование цифровых технологий и возможностей их отраслевого применения, разработка и тестирование компонентов СПЖЦ V3» от 25 марта 2024 г. Инв. номер № НО/6-286.10-2024-

О. С. 13-44.
4. Комплекс программ в защищённом исполнении «Система полного жизненного цикла изделий «Цифровое предприятие»: Программный модуль «Система конструкторского проектирования». ПЗ ТП 07623615.00423-04 81 01-4. Часть 4. 2021 г. 282 с.
5. ОТЧЕТ о научно-исследовательской работе ИССЛЕДОВАНИЕ ЦИФРОВЫХ ТЕХНОЛОГИЙ И ВОЗМОЖНОСТЕЙ ИХ ОТРАСЛЕВОГО ПРИМЕНЕНИЯ, РАЗРАБОТКА И ТЕСТИРОВАНИЕ КОМПОНЕНТОВ СПЖЦ V3.- (Договор на выполнение НИР от 2 марта 2023 г. № 96-2023/203, Задача 10). Инв. № НО/37-278.10-2023-О. С. 8 – 27
6. Жарков Н.А., Денисова Н.А. Особенности применения булевых операций в программном модуле «Система конструкторского проектирования» // Отчёт о НИР «Исследование цифровых технологий и возможностей их отраслевого применения, разработка и тестирование компонентов СПЖЦ V3» от 25 марта 2024 г. Инв. номер № НО/6-286.10-2024-О. С. 13-44.
7. Многообразная геометрия: https://quaoar.su/blog/page/bulevy-operacii-v-b-rep-vvedenie
8. Самарина Е.А., Денисова Н.А. Исследование операции «Скругление» в программном обеспечении «Система полного жизненного цикла «САРУС»//Отчёт о НИР «Исследование цифровых технологий и возможностей их отраслевого применения, разработка и тестирование компонентов СПЖЦ V3» от 25 марта 2024 г. Инв. номер № НО/6-286.10-2024-О. С. 13-44.

Секция 8

ГРАФИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ИММЕРСИВНЫЕ ТЕХНОЛОГИИ НА СТАДИЯХ ЖИЗНЕННОГО ЦИКЛА ПРОДУКТА

Graphic Information Systems and Immersive Technologies in the Product Life Cycle Stages

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 004.89; 004.35
DOI: 10.25206/978-5-8149-3873-2-2024-855-862

Интеллектуализация графических и биомеханических интерфейсов управления коллаборативными роботами

С. А. К. Диане 1, А. С. Алымова 2, И. А. Москалев 2, Д. И. Крохмальный 2, А. И. Буравцова 2
1Институт проблем управления им. В. А. Трапезникова РАН, Москва, Россия
2МИРЭА – Российский технологический университет, Москва, Россия

S. A. K. Diane 1, A. S. Alymova 2, I. A. Moskalev 2, D. I. Krokhmalny 2, A. I. Buravtsova 2
[1]V. A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia
[2]MIREA — Russian Technological University, Moscow, Russia

Аннотация. Доклад посвящен вопросу совместного применения графического и биомеханического интерфейсов коллаборативного робота. По результатам проведенного обзора и анализа предметной области обосновывается потребность в комплексной интеллектуализации мультимодальных интерфейсов по направлениям, связанным с распознаванием командной информации, адаптивной подстройкой хранимых управляющих процедур и интеграцией интерфейсных сигналов в общую схему технологического процесса. Предлагается структура информационно-управляющей системы с интеллектуальным двухмодальным интерфейсом, позволяющая реализовать взаимодействие с коллаборативным роботом. Описаны результаты экспериментальных исследований, раскрывающие особенности реализации визуально-жестикулярного интерфейса для управления автономным мобильным роботом при решении задачи обслуживания складского помещения. Обсуждаются пути совершенствования разработанного алгоритмического обеспечения.

***Ключевые слова*:** интеллектуальная обработка информации, мультимодальный интерфейс, коллаборативный робот, мобильный робот, автоматизация технологических процессов

Аннотация. The report is devoted to the issue of joint application of graphical and biomechanical interfaces of a collaborative robot. Based on the results of the review and analysis of the subject field, the need for a comprehensive intellectualization of the multimodal interface in areas related to the recognition of command information, adaptive adjustment of stored control procedures and integration of interface signals into the overall scheme of the technological process is justified. The structure of an information and control system with an intelligent two-modal interface is proposed, which allows to implement interaction with a collaborative robot. The results of experimental studies are described, revealing the features of the implementation of a visual-gestural interface for controlling an autonomous mobile robot when solving the problem of warehouse maintenance. The possibilities to enhance the developed algorithmic software are discussed.

***Ключевые слова*:** intelligent information processing, multimodal interface, collaborative robot, mobile robot, automation of technological processes.

I. Введение

В современном мире всё большее количество задач передаётся на выполнение роботам. Они позволяют повысить точность выполнения производственных операций, улучшить качество выпускаемой продукции, снизить необходимые при этом временные и финансовые затраты. Но, наряду с выполнением тяжелых физических или монотонных процессов, они становятся востребованными и в более широком спектре прикладных областей, относящихся к сферам услуг, легкой промышленности, сельского хозяйства и т.д.

Вместе с тем применение обычных роботов всё ещё вызывает трудности ввиду того, что оператору необходимо обучаться тонкостям работы с конкретной робототехнической системой, а зачастую – обладать навыками ее программирования. Поэтому более эффективным является внедрение роботов, управление которыми похоже на общение между людьми, включает в себя какие-то привычные и повседневные действия, интуитивно понятные оператору и легкие в освоении. При этом потенциал интерфейсного взаимодействия с роботами огромен – число степеней подвижности человека исчисляется десятками, приумножаясь еще и за счет комбинированных перемещений [1].

Именно по этой причине переход от теоретических и опытно-экспериментальных исследований автономных роботов к их массовому применению так или иначе сопряжен с полным или частичным отказом от компьютерной мыши и клавиатуры как единственных средств целеуказания в пользу джойстиков, графических сенсорных

панелей и других разновидностей портативных человеко-машинных интерфейсов, построенных на мультимодальных и биомеханических принципах [2, 3].

Подобное замещение технологий вовсе не является данью моде или погоней за футурологическими концепциями научно-технического прогресса, но обусловлено прагматической потребностью в повышении оперативности, функциональности и надежности средств взаимодействия человека-оператора с роботами полуавтоматического и автоматического типов.

Несмотря на то, что классические символьно-кнопочные средства ввода информации по-прежнему необходимы на этапе программирования и отладки робототехнических систем, при штатном использовании все большее распространение получают биомеханические и речевые интерфейсы. Зачастую они применяются совместно со средствами дополненной реальности, когда задача для робота формируется оператором в интерактивном режиме с привязкой к видеоизображению рабочей зоны.

Еще большую актуальность вопрос разработки удобных человеко-машинных интерфейсов приобретает с постепенным распространением коллаборативных роботов, которые способны действовать в тесном контакте с человеком. Такие роботы подстраиваются не только под условия внешней среды, но и под движения оператора, достигая согласованности в выполнении технологических операций.

В контексте коллаборативных роботов особый интерес представляют биомеханические интерфейсы жестикулярного и окулографического типа как наиболее простые в реализации и эксплуатации. Они допускают, с одной стороны, формирование высокоуровневых управляющих команд, а с другой – прецизионное телеоператорное управление роботом. Данные типы интерфейсов успешно применяются в сложных условиях внешней среды при решении задач, требующих от человека и робота повышенной оперативности реагирования. При этом отдельные каналы управления зачастую объединяются в мультимодальный интерфейс, обеспечивающий достаточную надежность системы.

Сопоставление потенциально применяемых способов ввода информации с решаемыми прикладными задачами приведено в табл. 1. Для удобства приняты следующие сокращения наименований интерфейсов: К – кнопочный, С – сенсорный, В – визуально-жестикулярный, П – потенциометрический, И – инерциальный, О – окулографический, Р – речевой.

ТАБЛИЦА 1
ОСОБЕННОСТИ ОРГАНИЗАЦИИ ИНТЕРФЕЙСОВ КОЛЛАБОРАТИВНЫХ РОБОТОВ

№	Прикладная задача	Особенность ввода информации	Альтернативы						
			К	С	В	П	И	О	Р
1	Ремесленное и мелкосерийное производство [4]	Нежелательность переключения внимания мастера с обрабатываемого изделия на средства ввода/вывода; загрязнение ладоней в процессе работы	+	-	+	+	+	+	+
2	Обслуживание складских помещений [5]	Необходимость оперативной координации совместных перемещений нескольких роботов при транспортировке крупногабаритных объектов	+	+	+	+	+	-	-
3	Техобслуживание транспортных средств [6]	Занятость ладоней техника инструментами; загрязнение ладоней горюче-смазочными материалами; неудобство ввода/вывода информации	+	-	+	-	-	-	+
4	Хирургия, стоматология, травматология [7]	Занятость ладоней врача медицинскими инструментами и материалами; нежелательность контактного взаимодействия с нестерильными поверхностями средств ввода/вывода	-	-	+	-	-	+	+
5	Строительство и ремонт зданий [8]	Загрязнение ладоней рабочего строительными материалами; невозможность оборудования стационарной поверхности для размещения средств ввода/вывода ввиду динамичности перемещений рабочего; наличие акустического шума	-	-	+	-	+	-	-
6	Пожаротушение в помещениях [9]	Плохие условия освещенности; низкая видимость средств ввода/вывода по причине задымленности помещения; неудобство ввода информации в строительной экипировке	+	-	-	+	+	-	+
7	Поиск и спасение людей [10, 11]	Невозможность оборудования стационарной поверхности для размещения средств ввода/вывода ввиду динамичности перемещений спасателя; непредсказуемость погодных условий	+	+	+	+	+	-	+

8	Подводные монтажные работы [12]	Низкая видимость средств ввода/вывода; неудобство ввода информации в снаряжении для подводного плавания	-	-	+	+	+	-	-

II. Постановка задачи

Как видно из представленной таблицы, перечень факторов неопределенности, присущих средам функционирования коллаборативных роботов, весьма обширен. Это обуславливает необходимость привлечения технологий искусственного интеллекта для разработки алгоритмов человеко-машинного взаимодействия, робастных к изменениям в параметрах целевых объектов и вариативности движений пользователя.

Можно выделить по меньшей мере пять направлений подобной интеллектуализации:

1. классификация команд оператора на основе комплексного анализа сенсорной информации;
2. накопление процедур управления коллаборативным роботом и их адаптация под текущую задачу;
3. прогноз вариантов развития рабочей обстановки и формирование рекомендаций по управлению роботом;
4. обобщение и наглядная визуализация многомерной сенсорной информации, поступающей от интерфейса;
5. интеграция сигналов, поступающих от интерфейса в общий технологический процесс.

В настоящем исследовании сосредоточимся подробнее на сформулированных выше задачах 1, 2 и 5.

Отталкиваясь от хорошо изученных принципов организации интеллектуальных систем управления, основанных на применении знаний [13], сформируем концепцию применения и структуру интерфейса, подходящего для управления коллаборативным роботом в задаче обслуживания складского помещения.

III. Теория

Обобщенная схема организации интеллектуального мультимодального интерфейса представлена на рис. 1. Его неотъемлемыми компонентами являются базы знаний, регламентирующие как порядок предобработки сенсорной информации, поступающей от графической или биомеханической подсистемы интерфейса, так и способ сопоставления обработанной сенсорной информации с известным набором ситуаций функционирования робота для формирования целесообразных управляющих воздействий на робота.

Дополнительно надежность и функциональность интерфейса может достигается за счет комплексирования нескольких модальностей. При этом требуется закладывать базы знаний и по анализу отдельных сенсорных каналов, и по анализу обобщенного информационного образа на выходе мультимодального интерфейса (рис. 1).

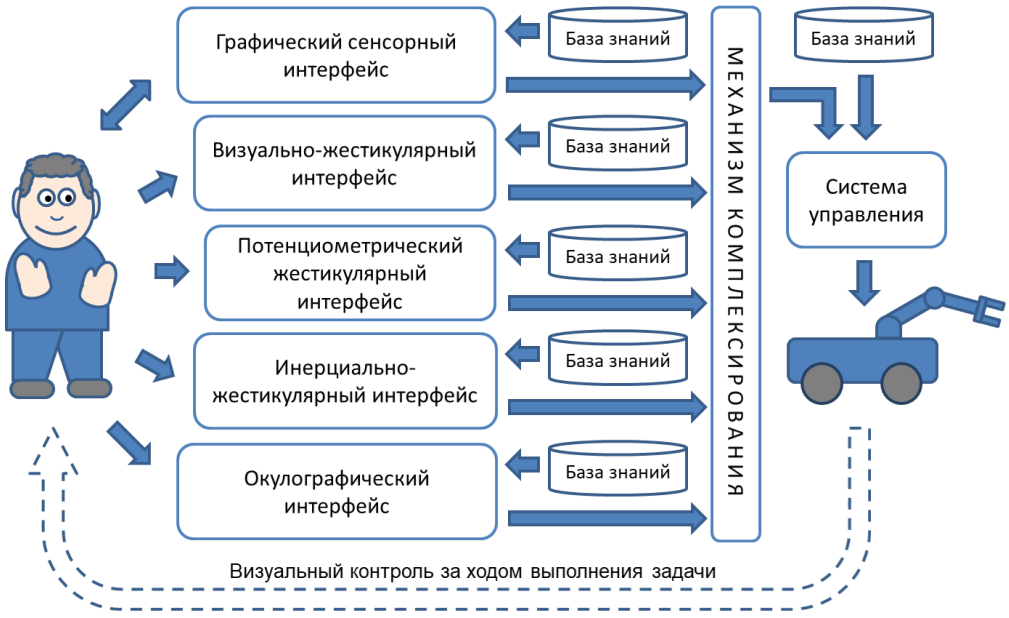


Рис. 1. Концепция управления роботом с применением мультимодального интерфейса

На практике количество информационных каналов интерфейса может уменьшено. В частности, возможны реализации системы за счет совмещения: потенциометрической перчатки и кнопочного интерфейса; окулографического интерфейса и сенсорной панели; визуально-жестикулярного и командно-текстового

интерфейсов. В качестве теоретического обобщения этих вариантов на рис. 2 представлена предлагаемая архитектура двухмодальной интеллектуальной системы для взаимодействия с коллаборативным роботом.

Отметим, что элементы искусственного интеллекта целесообразно внедрять не только непосредственно в блоки биомеханического и графического интерфейсов, но и в блоки формирования движений и коррекции движений робота. В режиме эксплуатации системы, формирователь движений анализирует вектор сенсорных данных X, получаемый от биомеханического интерфейса, сопоставляет его с заранее подготовленными циклограммами движения робота и определяет порядковый номер i подходящей циклограммы движения (фактически решается задача классификации сенсорного образа). Затем данный блок взаимодействует с корректором движений, получая от него модифицированную циклограмму, адаптированную для наблюдаемого состояния рабочей зоны робота.

Базовые циклограммы движения либо закладываются на этапе разаработки, либо же формируются в режиме обучения показом. В последнем случае пользователь при помощи телеоператорного управления формирует управляющие сигналы для приводной системы робота. Причем, исходя из соображений удобства, управляющие уставки могут формироваться либо на основе биомеханического средства ввода-вывода – $g(X)$, либо при помощи графического интерфейса – $s(Y)$, где X и Y – сенсорные образы соответственных каналов двухмодального интерфейса.

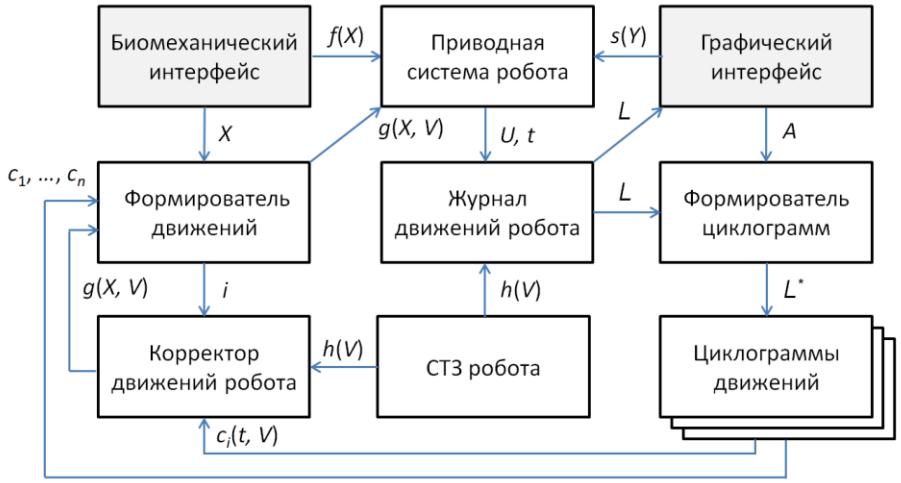


Рис. 2. Структура информационно-управляющей системы с интеллектуальным двухмодальным интерфейсом

Фактические сигналы U, поданные на приводную систему, совмещенные с временными отметками, сохраняются в журнал L движений робота. Затем они фильтруются в интерактивном режиме с применением функциональных возможностей A графического интерфейса и преобразуются в более лаконичный формат циклограмм.

Однако такого рода циклограммы не подходят для ситуаций, когда робот действует в среде с неопределенностями. Поэтому наряду с интеллектуализацией самого интерфейса, необходимо предусмотреть достаточную степень автономности коллаборативного робота. Она может быть обеспечена за счет оснащения робота системой технического зрения (СТЗ).

В задачи СТЗ входит детекция целевых объектов и передача вектора с их координатами и визуальными параметрами в подсистему коррекции движений робота. Кроме того, на этапе настройки векторы данного типа должны записываться и в журнал движений робота – дополнительно к сигналам управления. В таком случае коррекция выбранной циклограммы c_i с учетом результатов детекции объектов $h(V)$ на видеокадре V камеры робота позволит принять скорректированное управляющее решение $g(X, V)$.

IV. Результаты экспериментов

Для проверки эффективности предлагаемого подхода был исследован сценарий взаимодействия человека-оператора с мобильным робототехническим комплексом в задаче сбора объектов в рабочей зоне на основе сочетания графического и визуально-жестикулярного интерфейсов.

На основе функциональных возможностей моделирующего комплекса CoppeliaSim была разработана тестовая виртуальная сцена, представляющая собой фрагмент помещения. Целевые объекты – кубики разных цветов – располагались случайным образом на полу рабочей зоны. Робот, оснащенный системой технического зрения по жестовым командам оператора выполнял поиск, захват и перемещение кубиков в коробку. При этом

использовались четыре жеста, инициирующие автономное выполнение отдельных технологических операций или остановку робота (табл. 2).

Распознавание жестов на видеоизображении осуществлялось с применением связки из двух сверточных нейронных сетей MediaPipe Hands [14], Первая нейронная сеть работает на полном изображении и возвращает ограничивающую рамку ладони оператора. Вторая модель получает на вход уже обрезанную область изображения и с помощью регрессии находит на ладони пространственные координаты 21 ключевой точки.

ТАБЛИЦА 2
СЛОВАРЬ ЖЕСТОВ ДЛЯ УПРАВЛЕНИЯ КОЛЛАБОРАТИВНЫМ РОБОТОМ

Номер жеста	1	2	3	4
Графическое обозначение				
Значение	Команда на остановку робота	Команда на упаковку красного кубика	Команда на упаковку синего кубика	Команда на упаковку желтого кубика

Знания о допустимой очередности выполнения технологических операций по сбору объектов и их упаковке в коробку закладывались в систему управления роботом с применением графического интерфейса и объектно-ориентированного языка программирования Python. В качестве формы представления знаний был выбран аппарат конечных автоматов. Структура разработанного контролирующего автомата и логика переключений между его состояниями приведены на рис. 3.

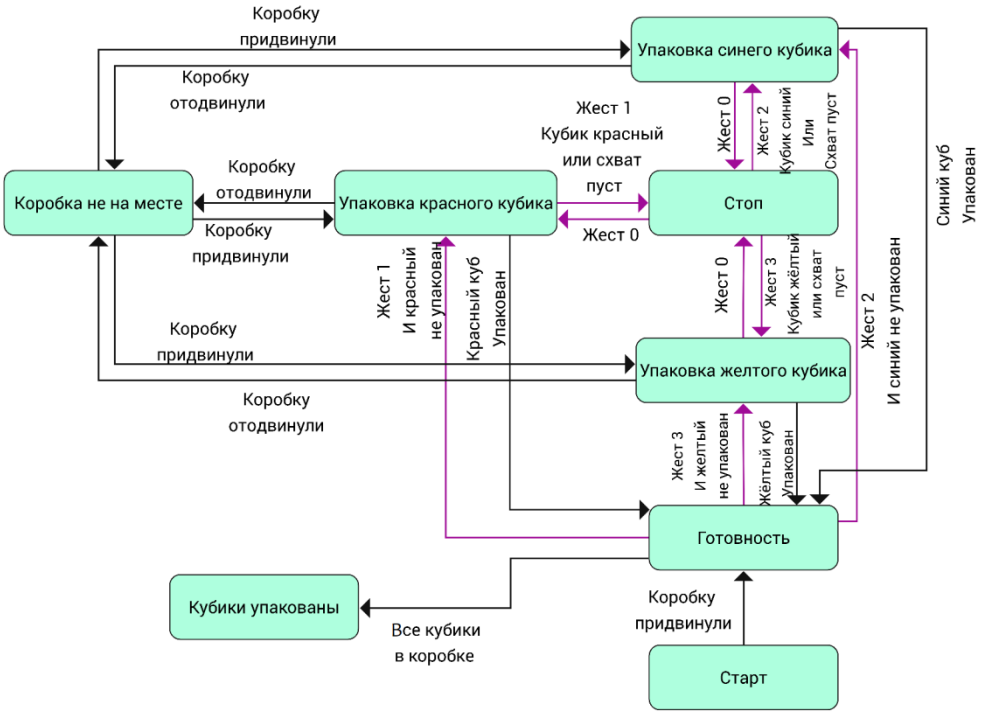
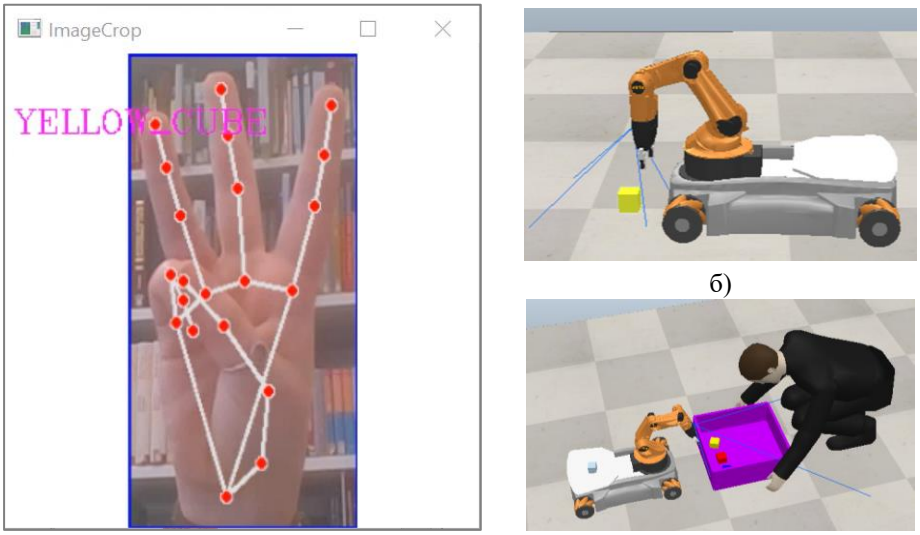


Рис. 3. Конечный автомат, описывающий процесс сбора целевых объектов

В целях обеспечения повторяемости запуска технологического процесса наряду с состояниями упаковки целевых объектов и остановки, введены дополнительные состояния, отвечающие за исходную готовность робота, наблюдение коробки в ожидаемой позиции, укомплектованность коробки кубиками. Фиолетовым цветом помечены переходы, обусловленные использованием визуально-жестикулярного интерфейса и требующие активного участия оператора.

Фрагмент моделирования перемещений коллаборативного робота в процессе выполнения поставленных задач представлен на рис. 4. Оператор показывает один из допустимых жестов в поле зрения видеокамеры, присоединенной к моделирующему компьютеру (рис. 4, а). Коллаборативный робот осуществляет поиск и захват объекта (рис. 4, б). Затем робот перемещает объект в коробку, заблаговременно размещенную человеком-участником технологического процесса в целевой позиции (рис. 4, в). Данная последовательность действий повторяется, пока все требуемые объекты не окажутся в коробке, в чем можно убедиться по графику переключения состояний робота, изображенному на рис. 5.

Отметим, что автономность коллаборативного робота достигается слаженной работой его системы технического зрения и алгоритмов планирования движений колесной платформы и манипулятора. Оценка положения центра масс кубика на видеокадре, формируется на основе цветофильтрации и пересчитывается в локальную систему координат робота с учетом параметров крепления видеокамеры. Данная информация используется в качестве визуальной обратной связи для выравнивания платформы робота и позиционирования конечного звена манипулятора при захвате целевого объекта.



а) б) в)

Рис. 4. Моделирование процесса сбора объектов с применением визуально-жестикулярного интерфейса

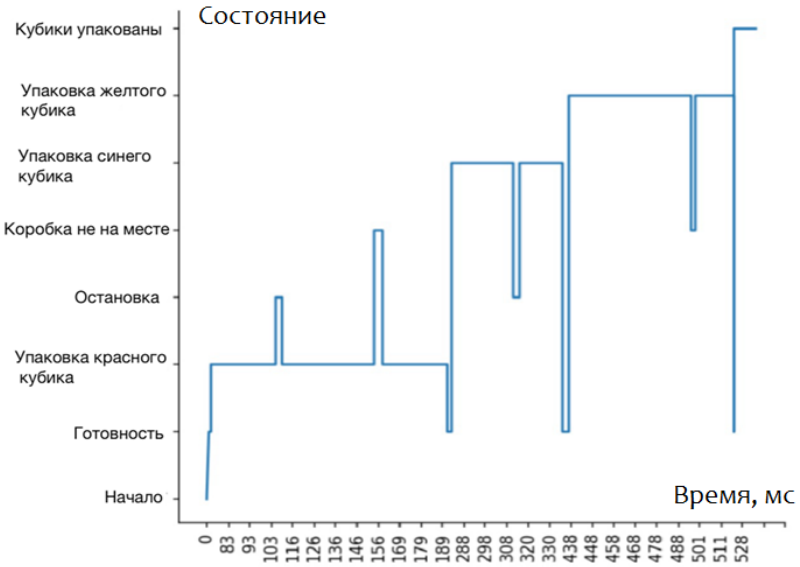


Рис. 5. Переключение состояний конечного автомата в рамках технологического процесса

V. Обсуждение результатов

Анализируя результаты экспериментов, можно заключить, что программная реализация модели технологического процесса по коллаборативному сбору и упаковке кубиков в коробку потребовала проработки значительной части подсистем в рамках структуры, представленной на рис. 2.

При этом подтвердился положительный эффект от интеллектуализации интерфейсов коллаборативного робота по обозначенным в постановке задач направлениям, которые связаны с необходимостью классификации сенсорно-командной информации, поступающей от оператора; адаптацией процедур управления роботом под выполняемую технологическую операцию; интеграцией интерфейсных и логических сигналов в общий технологический процесс.

Следует, однако, отметить, что ввиду относительной простоты решаемой задачи сбора кубиков, набор циклограмм, поступающих на вход формирователя движений моделируемого робота, был редуцирован и закладывался на стадии разработки. При захвате целевого объекта манипулятор принимает лишь два опорных положения – предзахватное и захватное. Выравнивание манипулятора в промежуточных позициях достигается за счет визуальной обратной связи.

VI. Выводы и заключение

Приведенный в начале работы обзор показал, что многообразие прикладных задач и вариативность современных графических и биомеханических интерфейсов определяют потребность в развитии совместимого с ними программно-алгоритмического обеспечения. С одной стороны, оно должно быть эффективным в обработке сигналов поступающих от оператора, а с другой – обладать достаточной степенью универсальности для интеграции в состав многосенсорных систем. Предложенный в настоящем исследовании подход во многом решает обозначенную проблему – за счет принципов взаимодополняемости интерфейсов различного типа и их комплексной интеллектуализации.

Практическое применение подобных интеллектуальных многосенсорных интерфейсов возможно в приложениях коллаборативной робототехники – мелкосерийном производстве, обслуживании складских помещений, медицине, строительстве и целом ряде других сфер, автоматизация которых упирается в способность человека формировать целеуказания робототехническим системам в режиме реального времени.

Дальнейшее развитие разработанных моделей и алгоритмического обеспечения связано с углубленным рассмотрением вопросов обучения робота показом при автоматическом выполнении многоэтапных технологических операций и вопросов подстройки его движений в соответствии с априорно неизвестными условиями внешней среды. Кроме того, существенный интерес представляет внедрение технологий виртуальной и смешанной реальности в графический интерфейс оператора для наглядного отображения информационных элементов технологического процесса.

Список литературы

1. Лукьяненко В.П. Физическая культура: основы знаний: учебное пособие. 2-е изд., стереотипное. М.: Советский спорт, 2005. 224 с.

2. Лохин В. М., Манько С. В., Александрова Р. И., Романов М. П., Диане С.А.К. Принципы построения и программно-алгоритмическое обеспечение человеко-машинного интерфейса для автономных роботов и мультиагентных робототехнических систем // Мехатроника, автоматизация, управление. М.: Новые технологии. 2016. №9. Том 17. С. 606 – 614.

3. Туровский Я.А., Алексеев В.Ю., Мурадова Л.Г., Миронкин А.П. Особенности генерации команд для окулографических интерфейсов в условиях вестибулярных воздействий // Сенсорные системы. - 2023. - №37. - С. 49-59.

4. Vanderborght B. et al. Towards an acceptable socially collaborative robot for the manufacturing industry. In ICRA workshop IC3-Industry of the future: Collaborative, Connected, Cognitive: Novel approaches stemming from Factory of the Future & Industry. 2017. Vol. 4, pp. 1-2.

5. Kennel-Maushart F., Poranne R., Coros S., Interacting with Multi-Robot Systems via Mixed Reality, 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 11633-11639

6. Ullah S. et al. An Automated Robot-Car Control System with Hand-Gestures and Mobile Application Using Arduino. 2019. Preprints 2019, 2019010029

7. Yang G., Lv H., Chen F., Pang Z., Wang J., Yang H., Zhang J. A Novel Gesture Recognition System for Intelligent Interaction with a Nursing-Care Assistant Robot. Applied Sciences. 2018; 8(12):2349.

8. Wang X., Zhu Zh. Vision-based hand signal recognition in construction: A feasibility study. Automation in Construction. 2021. 125. 103625.

9. Alon O., Rabinovich S., Fyodorov C., Cauchard J.R. First Step toward Gestural Recognition in Harsh Environments. Sensors. 2021; 21(12):3997

10. Meshcheryakov R.V. et al., An application of swarm of quadcopters for searching operations, IFAC-PapersOnLine, Volume 52, Issue 25, 2019, Pages 14-18.

11. Li S. et al. A Mobile Robot Hand-Arm Teleoperation System by Vision and IMU, 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020, pp. 10900-10906.

12. Chiarella D. et al. A Novel Gesture-Based Language for Underwater Human–Robot Interaction. Journal of Marine Science and Engineering. 2018. 6. 91

13. Искусственный интеллект и интеллектуальные системы управления / Макаров И. М., Лохин В. М., Манько С. В., Романов М. П.; Отд-ние информ. технологий и вычислит. систем РАН; отв. ред. Макаров И. М. – М.: Наука, 2006. – 333 с.

14. Vakunov A., Chang C. L., Zhang F., Sung G., Grundmann M., Bazarevsky V. Mediapipe hands: On-device real-time hand tracking. In Workshop on Computer Vision for AR/VR. Vol. 2. No. 4. p. 5.

УДК 004.94, 621.7
DOI: 10.25206/978-5-8149-3873-2-2024-863-869

Применение технологий геометрического моделирования и VR/AR для разработки инструкций по сборке и ремонту производственного оборудования
Application of geometric modeling and VR/AR technologies to manuals' development aimed the assembling and repairing production equipment

М. А. Петров[1], А. С. Вивтоненко[1], С. А. Богданов[2]
[1]ФГАОУ ВО «Московский политехнический университет», кафедра «Обработка материалов давлением и аддитивные технологии», г. Москва, Российская Федерация
[2]ООО «Топ-Системы», г. Москва, Российская Федерация

M. A. Petrov[1], A. S. Vivtonenko[1], S. A. Bogdanov[2]
[1]FSAEI HE "Moscow Polytechnic University", Department "Material Forming & Additive Technologies", Moscow, Russian Federation
[2]Top Systems Ltd., Moscow, Russian Federation

Аннотация. В статье рассматривается применение технологий виртуальной (VR) и дополненной (AR) реальности при решении машиностроительных задач. Разработаны инструкции по сборке и ремонту исполнительного механизма механического горячештамповочного пресса и настольного фотополимерного LCD-принтера. Рассмотрены особенности реализации инструкций в среде VR и AR.

***Ключевые слова*:** виртуальная реальность, дополненная реальность, инструкция по эксплуатации, ремонт оборудования, механический пресс, настольный LCD-принтер, обработка материалов давлением, аддитивные технологии.

Аннотация. The presented article deals with the application of virtual reality (VR) and augmented reality (AR) technologies in good for mechanical engineering tasks. Instructions for assembly and maintenance of the executive mechanism of mechanical hot stamping press and desktop photopolymer LCD-printer are prepared. The peculiarities of instructions realization in VR and AR environment are considered.

***Ключевые слова*:** virtual reality, augmented reality, user's manual, maintenance, mechanical press, desktop LCD-printer, material forming, additive technologies.

I. Введение

С усложнением конструкций машин и механизмов, а также появлением мехатронных устройств, требуется более продолжительный цикл обучение работы на них и их обслуживания, сборки и ремонта. Появление носимых для решения различных инженерных задач устройств претерпевают ряд трансформаций, активная фаза которой приходится на 1980-ые годы XX века [1]. Изобретатели проектировали носимые на голове устройства преимущественно для развития сферы развлечения. Каждое устройство предлагало пользователю выполнить одну задачу, имело разные функциональные назначения, делался акцент на восприятие событие определённым органом чувств (зрительное, звуковое, обонятельное). В 1979 году был разработан прототип устройства с глубоким, объёмным визуальным восприятием транслируемого изображения и фильма, который в дальнейшем был усовершенствован и расширен на другие органы восприятия (осязание и слух). Системы непрерывно совершенствовались на протяжении более сорока лет. На сегодняшний день помимо виртуальной реальности (Virtual Reality или VR) существует дополненная (Augmented Reality или AR) и смешанная реальность (Mixed Reality или MR). Современные системы способны изолировать пользователя от окружающей визуальной среды и переносить его в виртуальное пространство (VR), добавлять в окружающую среду объекты (AR) и взаимодействовать с виртуальными объектами в окружающем пространстве (MR). Сочетание всех инструментов трёх различных технологий (VR&AR&MR) привело к появлению новой объединённой технологии расширенной реальности (Extended Reality или XR).

В последнее время появилось весьма много практических примеров применения технологий VR/AR, основными целями которых является демонстрация возможностей этих технологий в практической работе в роли помощника или учителя/преподавателя (рис. 1а) [2, 3, 4]. Инженеры применяют технологии VR/AR для совместной разработки на этапах проектирования различных систем сложного устройства, например, системы

кондиционирования, энергетического и электронного управления целого завода, космического аппарата или самолёта. В медицинских учреждениях весьма распространены технологии AR на этапах планирования операций (рис. 1б) [5, 6].

а)

б)

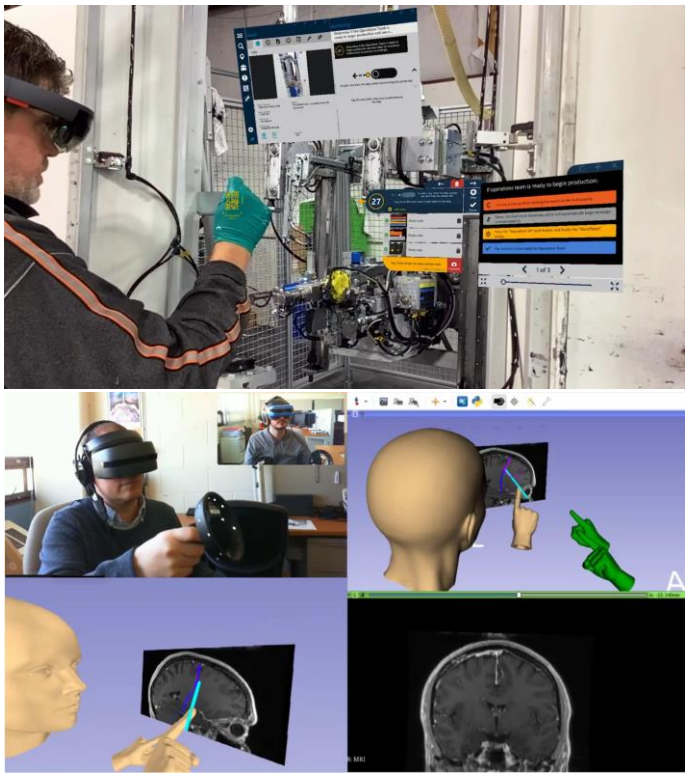


Рис. 1. Пример применения системы AR для обслуживания/замены динамического смесителя машины герметизации стеклопакетов (а) и системы VR при проведении совместного планирования операции (б)

II. Постановка задачи

Современное производственное оборудование весьма сложно в установке, настройке, работе и обслуживании. С целью ускорения процесса обучения на нём и преодоления психологического барьера, перехода от теории к практике, предлагается применять технологии VR/AR. Это позволит на месте проведения работ или до работ, в дистанционном формате, познакомиться с оборудованием и заранее провести виртуальные действия, правильная последовательность выполнения которых будет приводить к аналогичному в реальности результату.

Создание инструкции (по эксплуатации, обслуживанию или ремонту) подразумевает последовательное выполнение некоторых действий. В качестве примера были рассмотрены два направления заготовительного производства: инструкция по ремонту механического пресса (инструкция 1) и инструкция по сборке (инструкция 2) и ремонту (инструкция 3) настольного LCD-принтера.

Рассмотрим последовательность выполнения задач по инструкции 1 с применением технологии VR.

1) Загружаем модель механического пресса (не детализированная модель, полученная, например, по технологии 3D-сканирования);
2) Выбираем место разборки при помощи наведения манипулятора шлема виртуальной реальности, при этом происходит загрузка детализированных моделей деталей для выбранной области;
3) Выбор размеров заменяемой детали с учётом, например, износа сопрягаемых деталей, при помощи инструментов параметризации, постановка и расчёт оптимизационной задачи и итоговая актуализация сборки;
4) Установка новой детали в сборку;
5) Проверка на наличие статических пересечений;
6) Визуализация работы механизма с новой деталью с анализом пересечений в процессе динамической работы;
7) В дальнейшем возможно проводить компьютерное моделирование по, например, методу конечных элементов, с целью определения полей статического нагружения и/или построения графиков динамической нагрузки заменённой детали;
8) Составление отчёта о выполненных виртуальных ремонтных работах.

Задачи к выполнению инструкции 2 в режиме VR сводятся к трём пунктам.

1) Загрузка деталей в виртуальное пространство;
2) Сборка LCD-принтера;
3) Проверка на наличие статических пересечений.

Рассмотрим последовательность выполнения задач по инструкции 3 с применением технологии AR. Разделы инструкции 1 и 2 будут идентичны инструкции 1, поэтому начнём сразу с пункта 3.

3) Выбор заменяемой детали по каталогу запасных частей, представленной в виде библиотеки;
4) Установка новой детали в сборку;
5) Проверка на наличие статических пересечений;
6) Составление отчёта о выполненных виртуальных ремонтных работах.

На первый взгляд может показаться, что инструкции повторяются. Однако, как видно, там, где требуется только собрать объект изучения не требуется проведения проверочных работ и составления протокола/отчёта. Кроме этого, принципы работы в режимах VR и AR отличаются друг от друга.

Идентификация загружаемых деталей может выполняться либо по названиям, найденным в базе данных деталей машины, либо загружаться из внешнего хранилища, например, из облака по наведению манипулятора шлема виртуальной реальности на QR-код или гиперссылку. На рис. 2 показана сборка исполнительного механизма (ИМ) горячештамповочного механического пресса с номинальной силой 40 МН [7], для которой созданы два QR-кода: первый даёт краткое описание сборки, а по второму доступна ссылка на облачное хранение этой же 3D-модели. Сборка выполнена в программе параметрического проектирования Т-Флекс CAD, результат сборки экспортирован в файл полигональной (stl-) модели и импортирован в программу Blender, в которой имеется возможность запуска VR-среды.

Для визуализации результатов геометрического проектирования в среде VR/AR были использованы две системы, Oculus Quest 3 (VR/AR) и HTC Vive Pro (VR), представленные на рис. 3. Подготовка геометрических данных проводилась в программах Т-Флекс и Blender. В этих же программах присутствует возможность загрузки 3D-моделей в среду виртуальной реальности (VR). Применение дополненной реальности (AR) необходимо при реализации сценария, при котором пользователю необходимо контролировать совершаемые действия по сборке и ремонту по отношению к соседним объектам, например, чтобы не допустить чрезмерного пересечения тел. Технологии AR также удобны для условий не обучения сборки и ремонту, а для проведения реальной сборки и ремонта по месту расположения техники (рис. 1а и рис. 4). Для распознания отдельных деталей конструкции оборудования и активации определённых разделов инструкции в режиме AR возможно применение специальных меток (QR-коды, штрихкоды, гиперссылки, активные метки и т.д.), выбор которых осуществляется не при помощи контроллеров, а движением рук и определёнными жестами пальцев, так как системой AR ведётся непрерывное слежение за движением рук, что делает работу более комфортной. В режиме VR все манипуляции совершаются контроллерами через программное меню оболочки VR. Так, например, в оболочке Т-Флекс VR можно взять объект, создать сечение, определить размер, провести разборку и т.д. (рис. 5а). Аналогичным образом ведётся работа и в программе Blender, но инструменты не вызываются по контекстному меню, как в Т-Флекс VR, а требуется активное перемещения указателей контроллеров по меню и окнам основного интерфейса программы (рис. 5б).

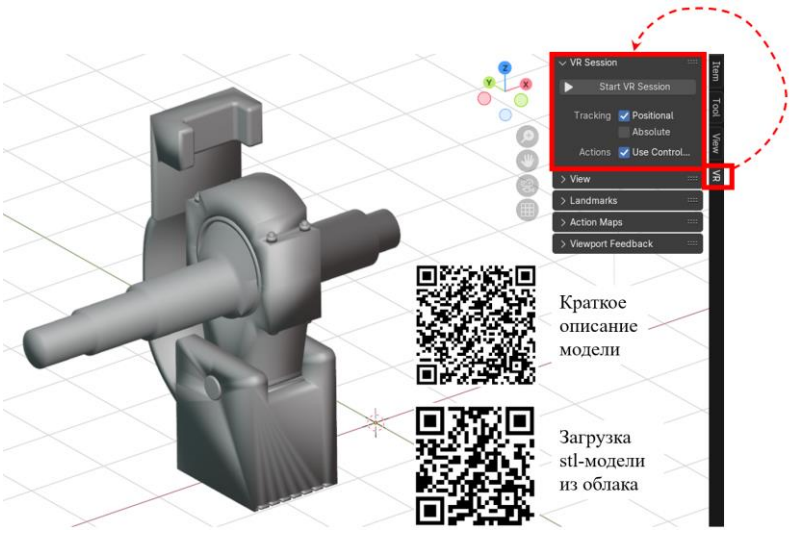


Рис. 2. Сборка исполнительного механизма

а) б)

Рис. 3. Системы виртуальной реальности с датчиками движения и контроллерами: Oculus Quest 3 (а) и HTC Vive Pro (б)

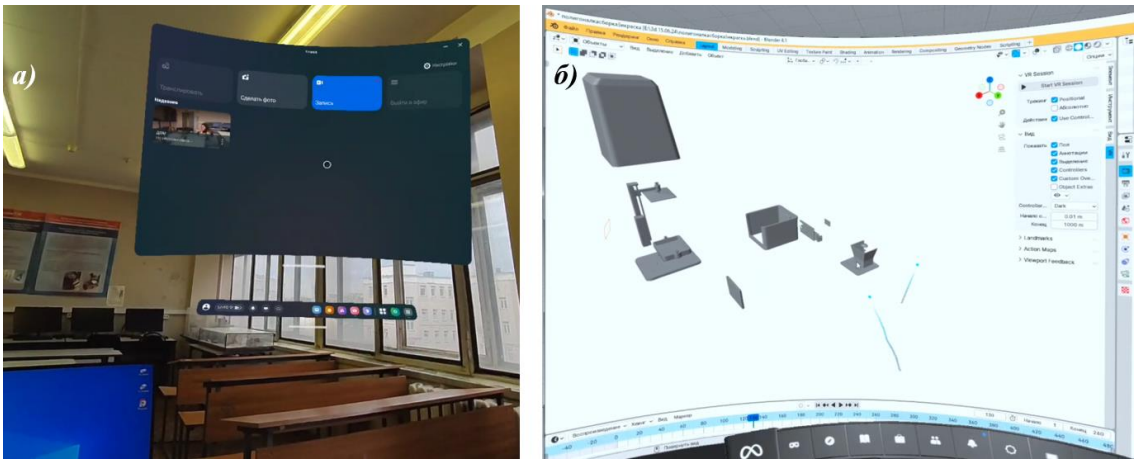
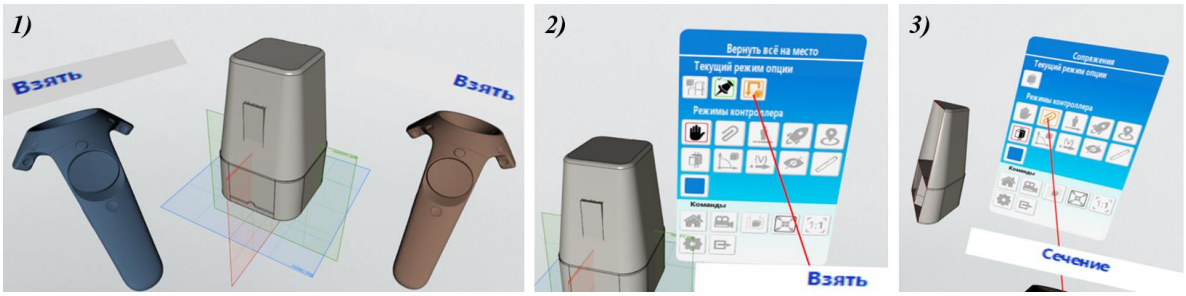
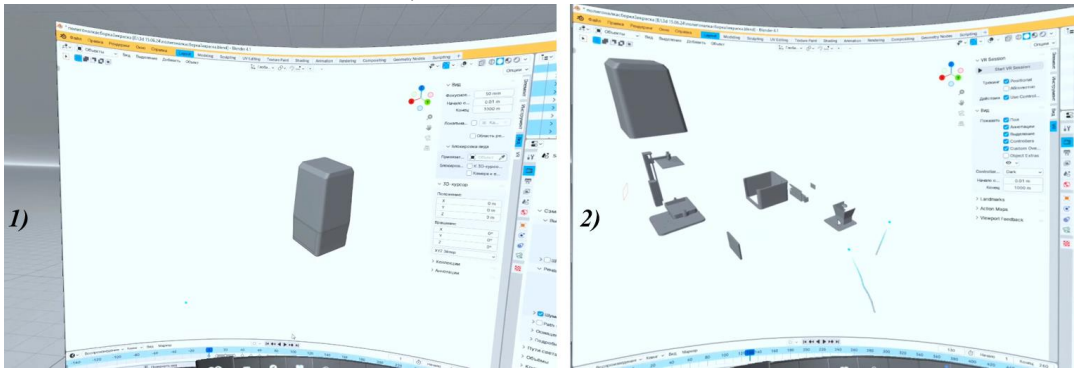


Рис. 4. Пример переключения функции дополненной реальности (а) и виртуальной реальности (б) в системе Oculus Quest 3

Среда Т-Флекс VR более глубоко погрузиться в виртуальное пространство, так как нет границ окон. Кроме этого, при работе в Blender весьма сильно ощущается зависание системы, даже при работе с небольшими сборками.



а) система HTC Vive Pro



б) система Oculus Quest 3

Рис. 5. Работа в среде VR в программах Т-Флекс (а) и Blender (б)

III. Способы создания 3D-моделей

13. *Параметрическое проектирование*

С развитием технологий компьютерного геометрического проектирования стоит остановиться на двух наиболее распространённых подходах моделирования – это параметрическое твердотельное моделирование и полигональное (оболочковое) моделирование. Возможна высокоточная конвертация геометрических твердотельных моделей в полигональные. Однако, обратное действие осложняется тем, что точность модели определяется алгоритмом триангуляции модели, зависящего от размера ячейки полигональной сетки и количества полигонов.

В отличие от непараметрического твердотельного моделирования, параметрическое моделирование позволяет изменять все параметры модели через окно переменных, без необходимости редактирования значений параметров на чертежах и/или эскизах, перезадания значений в отдельных меню каждой операций. Все параметры выведены в одно окно, через которое также возможно их управление. При желании, можно поставить и оптимизационную задачу с подключением численного моделирования, например, по методу конечных элементов (КЭ). Таким образом, проводя поиск оптимальных значений параметров для заданной целевой функции будет осуществляться и расчёт статической/динамической КЭ-задачи, по результатам решения которой значение параметра будет изменено автоматически. На рис. 6а перечислены параметры, введённые для контроля размеров деталей исполнительного механизма пресса: ширина шатуна (S1), ширина ползуна (S2), длина вала (L1), диаметр эксцентрика (2×R1) и величина эксцентриситета (exc) главного вала. На рис. 6б показано решение оптимизационной задачи по методу быстрого поиска для целевой функции «масса главного вала», при варьируемом значении переменной «exc».

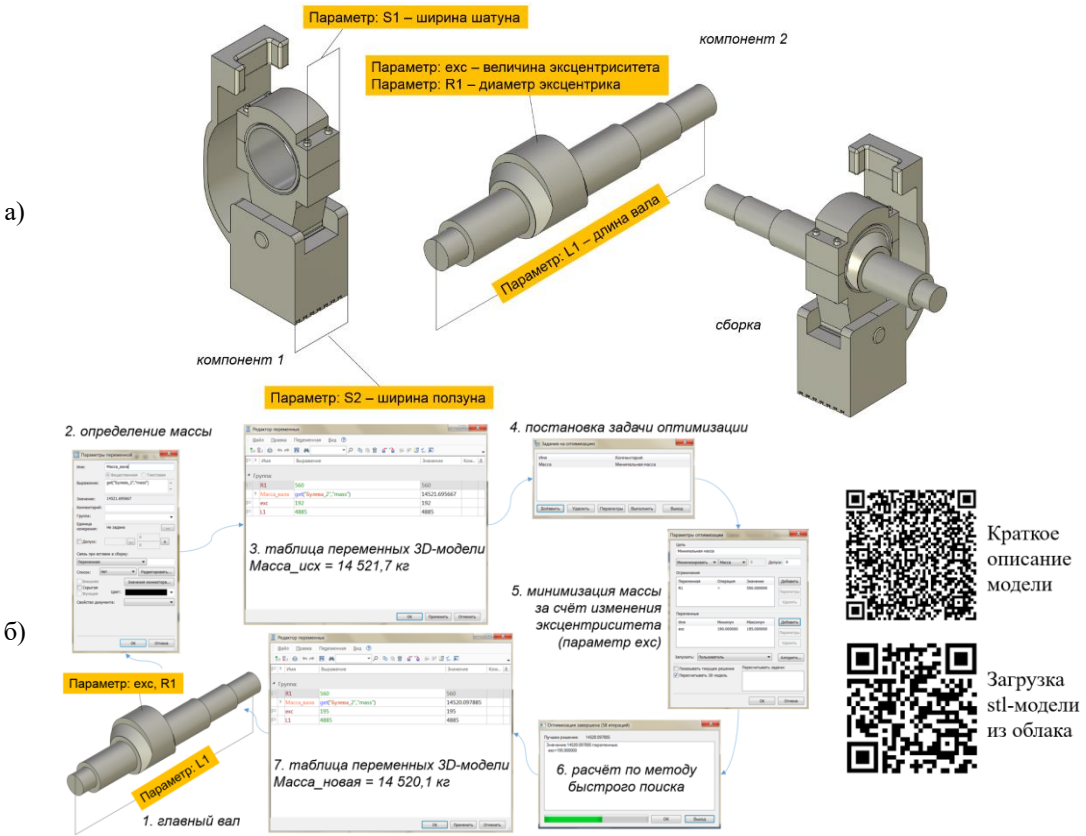


Рис. 6. Параметризация деталей сборки ИМ

На рис. 7 представлена частично параметризованная твердотельная модель LCD-принтера. Перед выбором главы раздела инструкции, например, по сборке, целесообразно применять низкодетализированные 3D-модели. При раскрытии целевого параграфа такая модель будет локально заменяться на высокодетализированную. Такой

подход позволит снизить требования к вычислительной мощности рабочего компьютера. Параметризация позволяет создавать библиотечные экземпляры деталей, размеры которых можно настроить перед вставкой объекта в сборку.

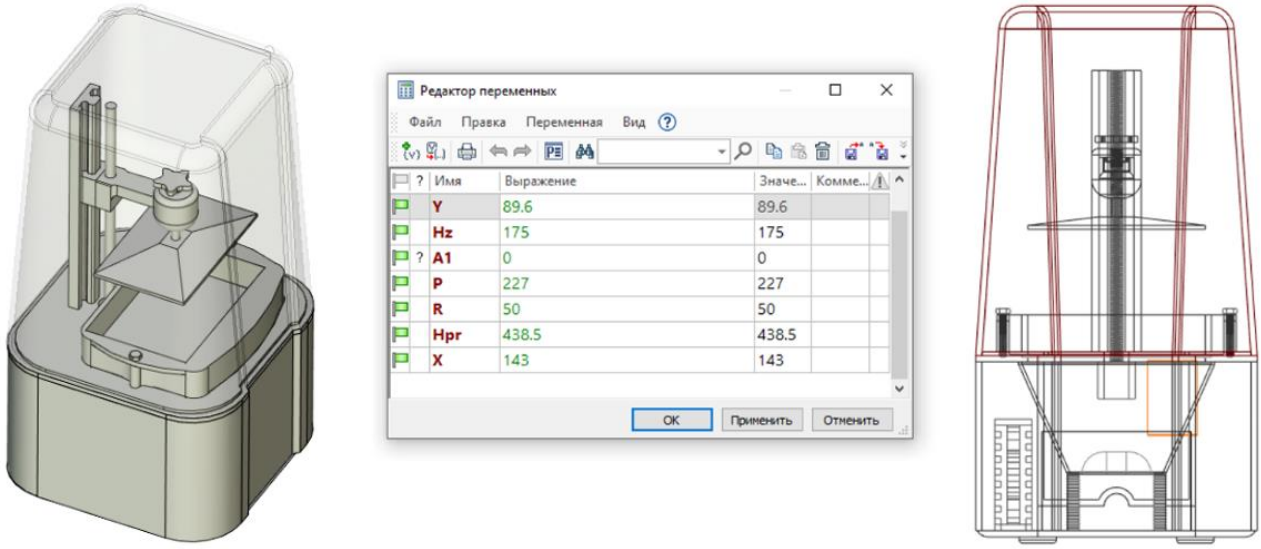
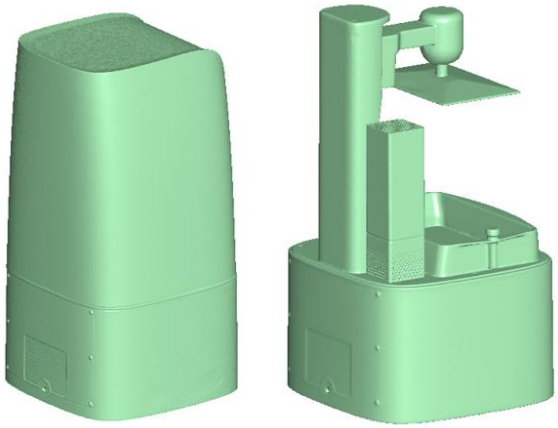


Рис. 7. Частичная параметризация твердотельной модели 3D-принтера Elegoo Mars 3

14. 3D-сканирование

Проектирование геометрии возможно и при помощи инструментов полигонального моделирования, например, в программах Blender, Rhino, 3DSMax и других, которые относятся к так называемым инструментам прямого моделирования. Существует и обратный подход, т.е. от объекта к модели. Он может быть реализован при помощи бесконтактного 3D-сканирования. На рис. 8 показаны упрощённые 3D-модели фотополимерного LCD-принтера Elegoo Mars 3, полученные при помощи лазерного сканера SimScan 30 (ScanTech).



а) с крышкой б) без крышки

Рис. 8. Полигональные модели LCD-принтера, с количеством полигонов: 463292 (а) и 964486 (б)

IV. Результаты экспериментов

По результатам проделанного исследования была разработана методика подготовки интерактивных инструкций для сборки и/или ремонта ИМ механического пресса и фотополимерного 3D-принтера с применением технологий AR/VR. Благодаря проведенным анализам была получена информация о необходимом наполнении виртуальной инструкции.

V. Обсуждение результатов

Основным препятствием на пути широкого внедрения технологий VR/AR является носимое на голове оператора оборудование, вес которого не всегда позволяет вести долговременную работу. На первый взгляд

простота подхода требует всё же уверенной работы с системой VR/AR, так как устройства обладаю высокой функциональностью, требующей периодического программного обновления системы. Поэтому применение этих систем не снижает требования по квалификации оператора, а ещё больше расширяет их.

VI. Выводы и заключение

Применение сред дополненной и виртуальной реальности позволит ускорить процесс обучения персонала на при работе и обслуживании высокотехнологичного оборудования. Работа в VR позволит снизить риск возникновения ошибок и их последствий на ранней стадии, что достигается многократным повторением методики работы или ремонта. Применение AR позволяет работать в непосредственной близи к оборудованию, однако этот этап работы должен идти за VR, так как он максимально приближен к реальным условиям работы.

Благодарность

Авторы выражают благодарность представителям компании ООО «НПО 3D-Интеграция» и ООО «СПИН» за помощь при проведении работ по 3D-сканированию.

Список литературы

1. https://en.wikipedia.org/wiki/Virtual_reality *(дата обращения: 25.08.2024)*
2. Nee A.Y.C. (Ed.). Augmented Reality - Some Emerging Application Areas, InTech, 2011. — 280 p.
3. Ma D., Gausemeier J., Fan X., Grafe M. (eds.) Virtual Reality & Augmented Reality in Industry, Springer, 2011. — 215 p.
4. https://www.youtube.com/watch?v=G1xNd_n1QyE *(дата обращения: 25.08.2024)*
5. https://www.youtube.com/watch?v=rG9ST6xv6vg *(дата обращения: 25.08.2024)*
6. Chheang, V., Saalfeld, P., Joeres, F., Boedecker, C., Huber, T., Huettl, F., … Hansen, C. (2021). A collaborative virtual reality environment for liver surgery planning. Computers & Graphics, 99, 234–246. doi:10.1016/j.cag.2021.07.009
7. А.А. Игнатов, Т.А. Игнатова, Кривошипные горячештамповочные прессы, М.: Машиностроение, 1974, с. 352

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-870-878

Исследование восприятия графической информации в интерфейсе веб-сайта
Study of the perception of graphic information in the website interface

В.Э.Янчус[1], А.Е. Хейфиц[1], Е.В.Боревич[1]
[1]Санкт-Петербургский политехнический университет Петра Великого 1, ул. Политехническая, 29, Санкт-Петербург, 195251, Россия
V.E. Yanchus[1], A.E. Kheifits[1], E.V. Borevich[1]
[1]Peter the Great St.Petersburg Polytechnic University (SPbPU), Polytechnicheskaya, 29, St.Petersburg, 195251, Russia

Аннотация. В статье описана разработанная и опробованная методика проведения экспериментального исследования восприятия графических свойств интерфейса веб-сайта человеком. Изучение восприятия информации может позволить сделать интерфейс более эффективным. В работе использована стационарная система, фиксирующая глазодвигательную активность с применением технологии айтреккинга. С помощью экспериментальной установки, с использованием айтреккингового оборудования, возможна фиксация параметров шаблона рассматривания проводится. В статье описан метод подготовки стимульного материала. Методика направлена на исследование различных факторов, таких как фактор цветового решения, композиции и расположения меню. В эксперименте участвовали две группы испытуемых: обладающие художественной подготовкой и студенты технических специальностей, у которых отсутствовала специальная художественная подготовка. Для обработки результатов эксперимента используются методы математической статистики. Полученные в результате проведения эксперимента данные выявили статистически значимое влияние фактора композиции. Для группы испытуемых, обладающих художественной подготовкой, выявлено влияние также фактора цветового решения.
Ключевые слова: Веб-сайт, Зрительное восприятие, Графический интерфейс, Айтрекинг, Стимульный материал, Цветовое решение, Композиция, Статистический анализ.

Abstract. The article describes a developed and tested methodology for conducting an experimental study of human perception of the graphical properties of a website interface. Studying the perception of information can make the interface more effective. The work used a stationary system that records oculomotor activity using eye tracking technology. With the help of an experimental setup, using eye tracking equipment, it is possible to fix the parameters of the viewing pattern. The article describes a method for preparing stimulus material. The methodology is aimed at studying various factors, such as color scheme, composition and menu layout. Two groups of subjects participated in the experiment: those with artistic training and technical students who did not have special artistic training. Methods of mathematical statistics are used to process the experimental results. The data obtained of the experiment revealed a statistically significant influence of the composition factor. For the group of subjects with artistic training, the influence of the color scheme factor was also revealed.
***Keywords*:** Website, Visual perception, Graphical interface, Eye tracking, Stimulus material, Color scheme, Composition, Statistical analysis.

Введение

С момента своего появления графические пользовательские интерфейсы (GUI) произвели революцию во взаимодействии человека с компьютером. От первых пиксельных экранов до современных сенсорных дисплеев GUI прошли значительный путь эволюции, став неотъемлемой частью нашей повседневной жизни. Понимание основных принципов функционирования GUI является критически важным для разработки удобных, эффективных и эстетически привлекательных программных продуктов.

Разработка первых GUI, таких как Xerox Alto (1973) и Apple Macintosh (1984), ознаменовала собой переход от командной строки к более интуитивно понятным способам взаимодействия с компьютерами. Концепции, предложенные в работах Алана Кея, Батлера Лэмпсона и других пионеров в области GUI, легли в основу современных операционных систем, приложений, а также и веб-сайтов [1].

Графические пользовательские интерфейсы состоят из различных элементов, таких как окна, кнопки, меню и значки, которые организованы в соответствии с определёнными принципами дизайна. Важную роль играют метафоры реального мира (например, рабочий стол), а также принципы видимости, обратной связи, согласованности и доступности [2]. Существуют различные модели взаимодействия пользователя с GUI, такие как «объект-действие» (например, щелчок мышью по файлу для его открытия) и модель «прямого манипулирования» (например, перетаскивание объекта на экране). Понимание этих моделей позволяет создавать более интуитивные и эффективные интерфейсы [3]. Существует множество методов и инструментов для

разработки и оценки GUI, включая прототипирование, юзабилити-тестирование и анализ эвристик. Целью этих методов является создание удобных, эффективных и удовлетворяющих пользователей интерфейсов [4]. Современные тенденции в области GUI включают в себя использование сенсорных экранов, голосового управления, жестов, виртуальной и дополненной реальности. Эти технологии открывают новые возможности для взаимодействия человека с компьютером, но также ставят перед разработчиками новые вызовы [5].

Систематизация информации в графическом виде (инфографика), упрощает ее восприятие и позволяет оператору-человеку эффективно считывать информацию о состоянии объекта наблюдения [6]. Зрение человека – самый информативный канал получения информации [7], изучение процессов восприятия зрительной информации может расширить эффективность взаимодействия пользователя с графическим интерфейсом.

Ускорение восприятия данных во многом зависит от способности пользователя получать обратную информацию от интерфейса. Это расширение границ работы между пользователем и графическим интерфейсом, построение этого улучшения зависит от сбора данных после экспериментов, в которых человек выступает подопытным и решает определенные задачи. Считывание и анализ этих данных позволит улучшить управление, восприятие и передачу информации [8].

В одном из экспериментов был проведен анализ чтения визуальной информации без явной смысловой нагрузки. Исследования проводились с применением технологии ай-трекинга. Представленная в статье методика показала свою эффективность и позволила получить определенные результаты [9]. В следующем эксперименте была установлена статистически значимая зависимость времени рассматривания стимула от фактора стилизации изображения [10]. В следующей статье определена функциональность инфографики и указывается оптимальный формат для использования данного средства как инструмента [11]. Продуманность технической эстетики инфографических материалов позволяет человеку свободно считывать информацию в сложных условиях [12,13].

[d] Графический пользовательский интерфейс использует монитор, который ограничивает обзор, тем самым сужая область восприятия информации до своих рамок. Информационно-управляющее поле современных веб-сайтов может оперировать довольно большим количеством данных. В связи с этим, является актуальной задача исследования и разработки эффективного интерфейса, позволяющего оптимизировать информационную загруженность пользователя и, как следствие, улучшить его впечатление от взаимодействия с сайтом [14].

Результаты исследования позволят создавать более удобные, эффективные и интуитивно понятные графические пользовательские интерфейсы, которые будут отвечать потребностям пользователей и способствовать более эффективному взаимодействию человека с компьютером.

II. Постановка задачи

Для исследования влияния элементов графического интерфейса на восприятие испытуемым зрительной информации в информационном поле веб-сайта необходимо разработать методику проведения эксперимента, а также стимульный материал.

Сформулированы следующие задачи исследования:

1. Проанализировать существующие распространённые типы интерфейсов в веб-дизайне. Влияние на эффективность работы различных элементов графического пользовательского интерфейса. Сравнить эффективность различных моделей в зависимости от типа задачи и контекста использования.
2. Разработать стимульный материал для постановки эксперимента с целью выявления зависимостей влияния различных элементов интерфейса на зрительное восприятие зрителя и на эффективность работы пользователя в зависимости от этих элементов.
3. Разработать рекомендации по проектированию графических пользовательских интерфейсов с учётом когнитивных способностей пользователей

III. Теория

Для более глубокого понимания механизмов взаимодействия человека с машиной целесообразно использовать информационную модель восприятия зрительной информации человеком на перцептивном уровне (рис.1) [15].

Постановка задачи для испытуемого должна позволить исследовать в данном эксперименте влияние трех факторов: композиции, цвета и расположения меню. Данные факторы обуславливают проектирование и создание стимульного материала:

– подбор цветового решения выполнен в соответствии с теорией Геринга и цветовыми контрастами И.Итена [7];

– композиционное построение интерфейса основано на законах гештальт психологии [16].

IV. Результаты эксперимента

Разработанная лабораторная экспериментальная установка, с использованием стационарного айтрекингового аппаратного комплекса подробна описана в предыдущих опубликованных экспериментальных исследованиях [17]. Методика проведения эксперимента включает разработку стимулов в соответствии с определенными

факторами исследования, определение задачи испытуемым в ходе эксперимента, проведение эксперимента на установке айтрекинга и статистический анализ полученных экспериментальных данных.

Для создания стимулов был разработан условный интерфейс приветственного экрана главной страницы сайта. По наполнению контентом все стимулы одинаковые: одно изображение, меню из 5-ти пунктов, приветствие и две кнопки-действия. Порядок пунктов меню варьировался от стимула к стимулу, чтобы разнообразить задачу, стоящую перед испытуемыми. Всего подготовлено 36 стимулов.

Одним из исследуемых фактором была композиция. Соответственно, все стимулы можно однозначно разделить на три группы: со смещением композиционного центра в правую сторону относительно центра стимула (рис. 2а), с центральной композицией (рис. 2б) и со смещением композиционного центра в левую сторону относительно центра стимула (рис. 2в).

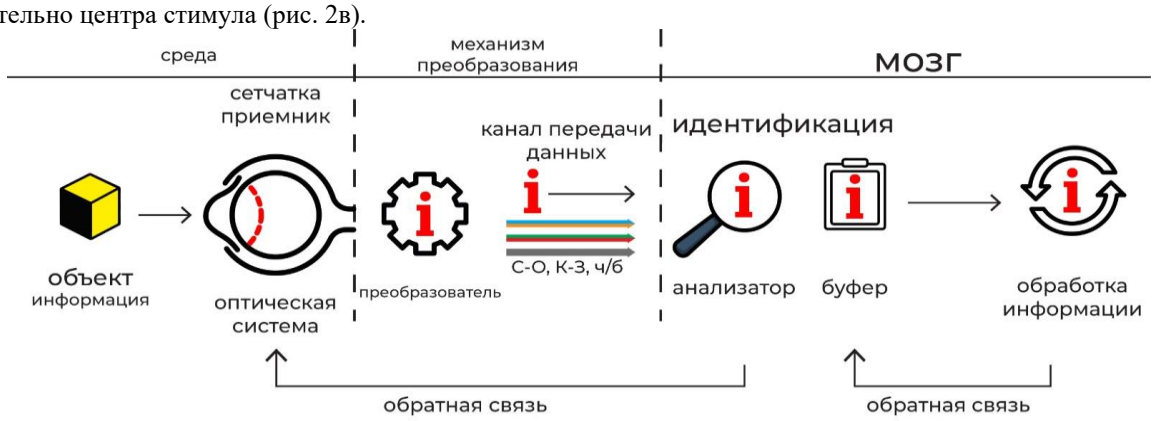


Рис. 1. Информационная модель восприятия зрительной информации человеком на перцептивном уровне

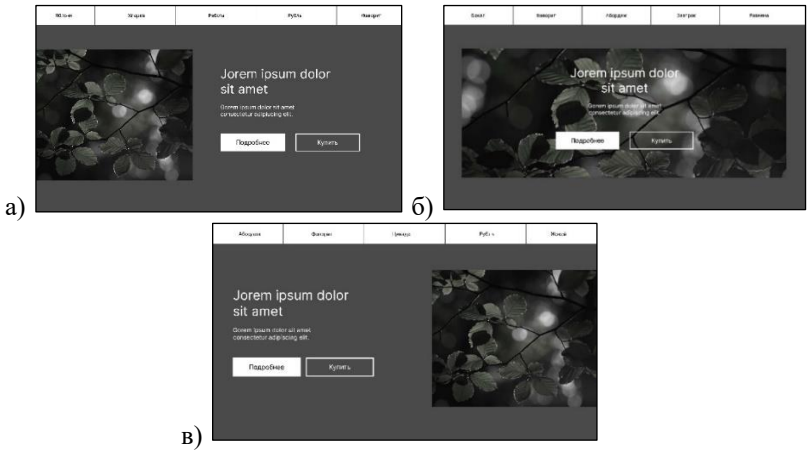


Рис. 2. Пример стимульного материала с различными вариантами композиционного построения: а) «правая», б) центральная, в) «левая»

Следующий фактор – разное расположение меню интерфейса веб-сайта: горизонтальное (рис. 3а) и вертикальное (рис. 3б). Стоит отметить, что в стимулах с «правой» композицией вертикальное меню располагалось справа, а с центральной и с «левой» композицией – слева.

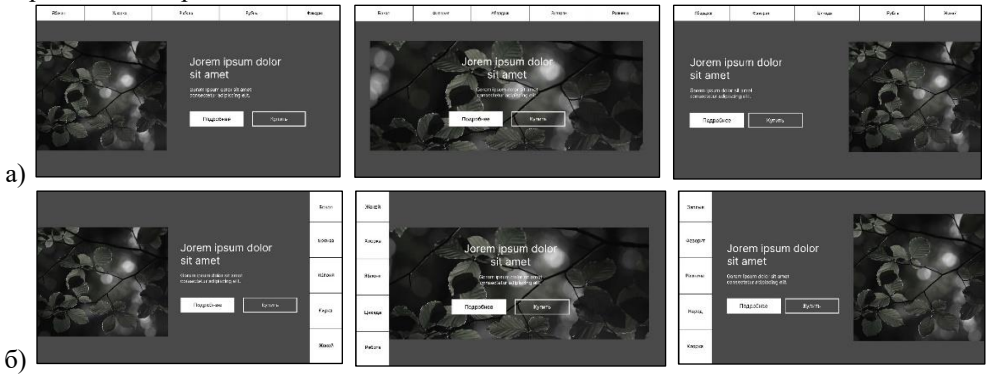


Рис. 3. Пример стимульного материала с различными вариантами расположения меню: а) горизонтальное, б) вертикальное

В соответствии с задачей эксперимента, цветовое решение стимульного материала выбрано таким образом, что все стимулы можно однозначно разделить на 6 групп: фон стимулов светлый, элементы управления тёмные (рис. 4а); фон – тёмный, элементы – светлые (рис. 4б); фон – оранжевый, элементы – синие (рис. 4в); фон – синий, элементы – оранжевые (рис. 4г); фон – красный, элементы – зелёные (рис. 4д); фон – зелёный, элементы – красные (рис. 4е).

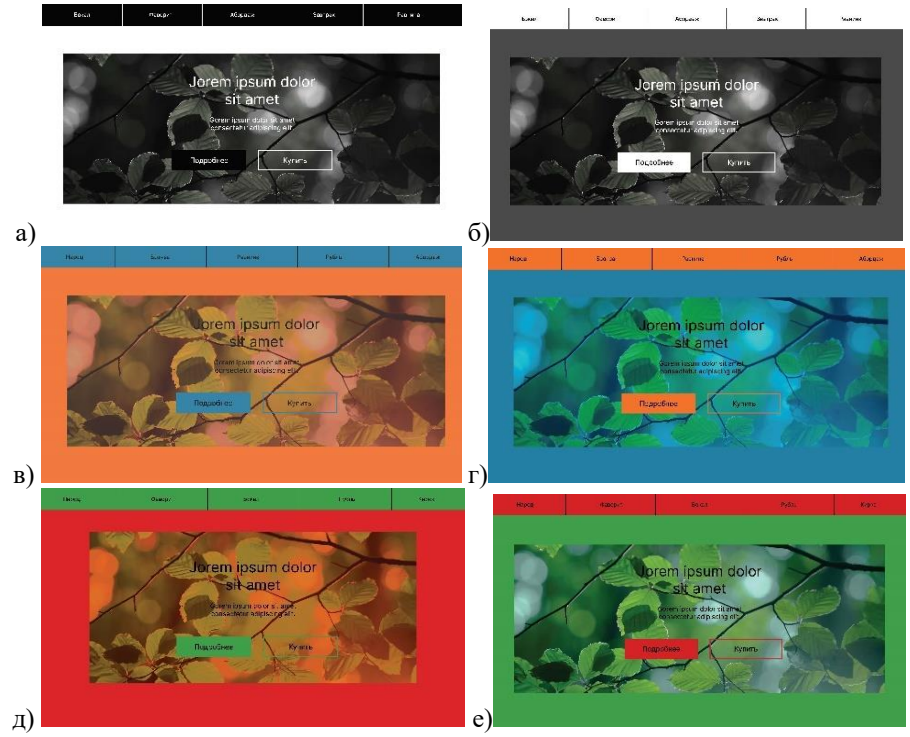


Рис. 4. Пример стимульного материала с различными вариантами цветового решения: а) фон стимулов светлый, элементы управления тёмные; б) фон – тёмный, элементы – светлые; в) фон – оранжевый, элементы – синие; г) фон – синий, элементы – оранжевые; д) фон – красный, элементы – зелёные; е) фон – зелёный, элементы – красные

Перед испытуемым в эксперименте ставилась следующая задача: посмотреть на одну из кнопок действия, затем, найти определённый заданный пункт меню и отметить ее наведением курсора мыши и нажатием левой кнопки мыши. Координаты нажатия кнопки компьютерной мыши фиксируется в базе эксперимента, после чего происходит переход к следующему стимулу. Все испытуемые были проинструктированы о поставленной задаче перед прохождением эксперимента. Время для решения поставленной задачи было неограниченно, и испытуемый выбирал его самостоятельно.

V. Обсуждение результатов

В эксперименте участвовало 30 человек от 18 до 45 лет из числа студентов и преподавателей Санкт-Петербургского политехнического университета Петра Великого. Испытуемые были подобраны таким образом, чтобы их можно было разделить на фокус-группы, соответствующие исследуемым признакам: половой, образование (гуманитарное – техническое), наличие художественной подготовки.

Анализировались параметрические данные шаблона рассматривания 30ти испытуемых. Было собрано 10610 фиксаций и 10181 саккад. Статистическая обработка результатов эксперимента производилась посредством дисперсионного анализа [18]. Анализировались следующие параметры шаблона рассматривания: время наблюдения стимула (time), средняя длительность фиксаций при наблюдении одного стимула, среднее количество фиксаций при наблюдении одного стимула, средняя длительность саккад и среднее количество саккад при наблюдении одного стимула, максимальная амплитуда саккад при наблюдении одного стимула. Задачей анализа было выявить влияние факторов композиционного построения, цветового решения и расположения меню, а также факторов художественной подготовки испытуемых (наличие / отсутствие), типа образования (гуманитарное / техническое) и гендерного признака (мужчины / женщины). Уровень значимости p-value для принятия гипотезы был выбран 0,05 [19]. Значения p-value, полученные в результате выполнения многофакторного дисперсионного анализа представлены в таблице 1. Красным выделены значения, позволяющие принять гипотезу о влиянии фактора.

ТАБЛИЦА 1
ВЫЧИСЛЕННЫЕ ЗНАЧЕНИЯ P-VALUE ДЛЯ ПАРАМЕТРА TIME

Фактор	**p-value**
Цветовое решение	0,7479
Композиция	0,04182
Положение Меню	0,5037
Половой признак	0,7891
Художественная подготовка	0,0001

Статистический анализ выявил статистически значимую зависимость времени рассматривания стимула испытуемым от фактора композиции и фактора художественной подготовки.

Анализируя зависимость времени рассматривания стимула испытуемыми, от фактора композиции (рис. 5), можно увидеть повышение эффективности работы с интерфейсом веб-сайта с центральной композицией, где все элементы подчинены симметрии вдоль центральной вертикальной оси.

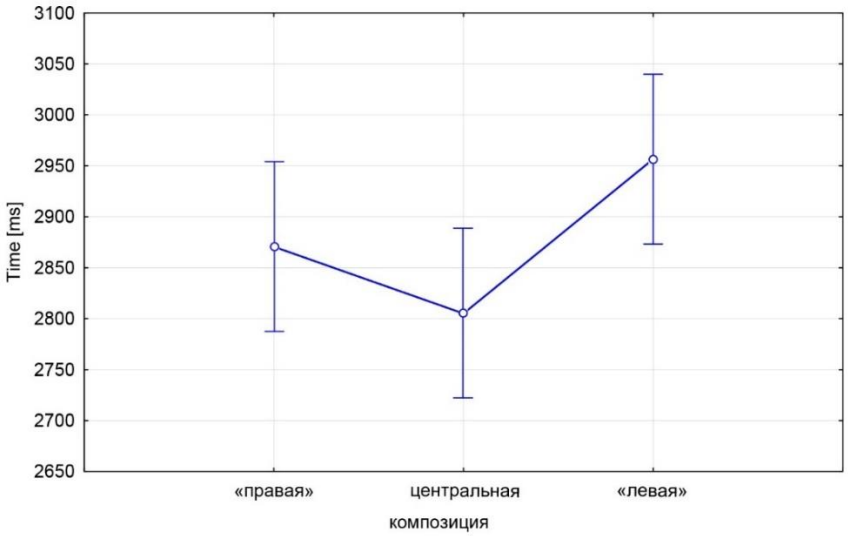


Рис. 5. График распределения значений времени рассматривания стимула испытуемыми в зависимости от фактора композиции

Анализируя график распределения значений среднего времени рассматривания стимулов в зависимости от фактора цветового решения (рис.6), можно наблюдать незначительное снижение эффективности в цветовых решениях с использованием сочетания красного и зелёного цветов. Однако статистически значимой зависимости от фактора цветового решения не наблюдается.

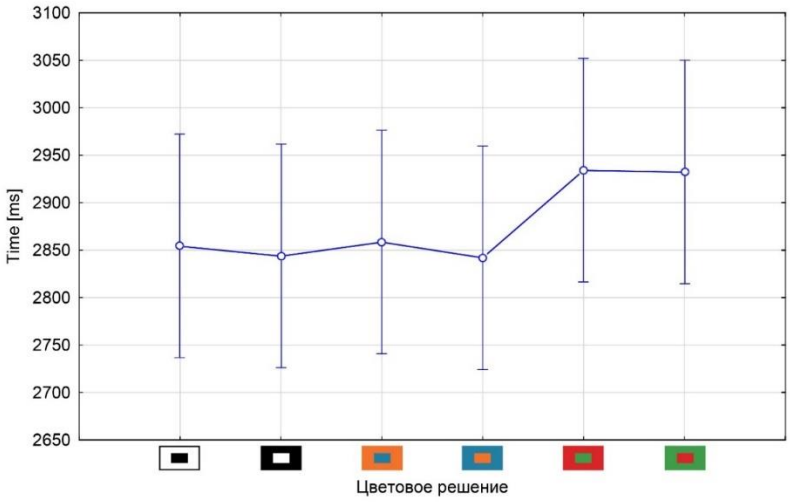


Рис. 6. График распределения значений времени рассматривания стимула испытуемыми в зависимости от фактора цветового решения

Анализируя зависимость времени рассматривания стимула испытуемыми, от фактора художественной подготовки испытуемых (рис. 7), можно увидеть влияние наличия художественного образования на эффективность работы с интерфейсом веб-сайта. Испытуемые, обладающие художественной подготовкой, решали задачу дольше.

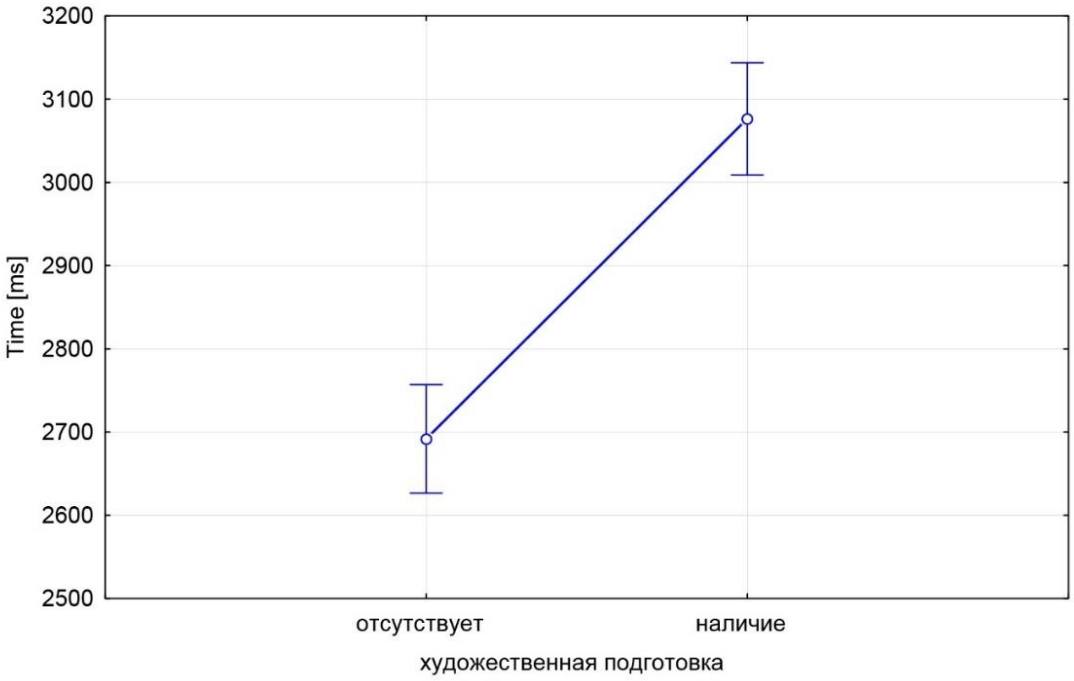


Рис. 7. График распределения значений времени рассматривания стимула испытуемыми в зависимости от фактора художественной подготовки

Следующим этапом обработки экспериментальных данных стало рассмотрение выборочных данных по испытуемым. Анализировалось влияние исследуемых фактором в выборке по испытуемым, обладающим художественной подготовкой. В таблице 2 приведены вычисленные значения значений p-value для различных факторов. Красным выделены значения, позволяющие принять гипотезу о влиянии фактора.

ТАБЛИЦА 2
ВЫЧИСЛЕННЫЕ ЗНАЧЕНИЯ P-VALUE ДЛЯ ПАРАМЕТРА TIME. ИСПЫТУЕМЫЕ, ОБЛАДАЮЩИЕ ХУДОЖЕСТВЕННОЙ ПОДГОТОВКОЙ

Фактор	**p-value**
Композиция	0,0776
Цветовое решение	0,04804
Положение Меню	0,4267
Половой признак	0,0002

Статистический анализ выявил статистически значимую зависимость времени рассматривания стимула испытуемым от всех факторов, кроме расположения меню.

Анализируя зависимость времени рассматривания стимула испытуемыми, обладающими художественной подготовкой, от фактора цветового решения (рис. 8), можно увидеть повышение эффективности работы с интерфейсом веб-сайта в цветовом решении, где основной преобладающий цвет чёрный, а элементы управления белого цвета.

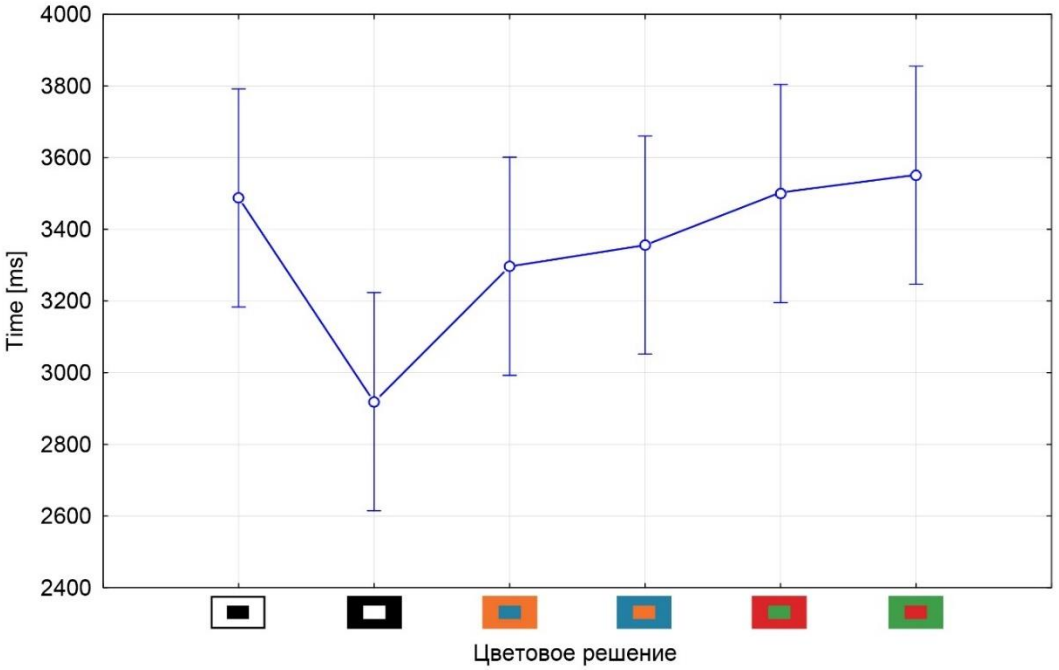


Рис. 8. График распределения значений времени рассматривания стимула испытуемыми с художественной подготовкой в зависимости от цветового решения

У испытуемых мужчин среднее время рассматривания, как правило, больше, кроме случая с цветовым решением белые элементы на тёмном фоне (рис. 9). У испытуемых женщин среднее время рассматривания меньше, чем у мужчин, однако статистически значимой зависимости от цветового решения не наблюдается.

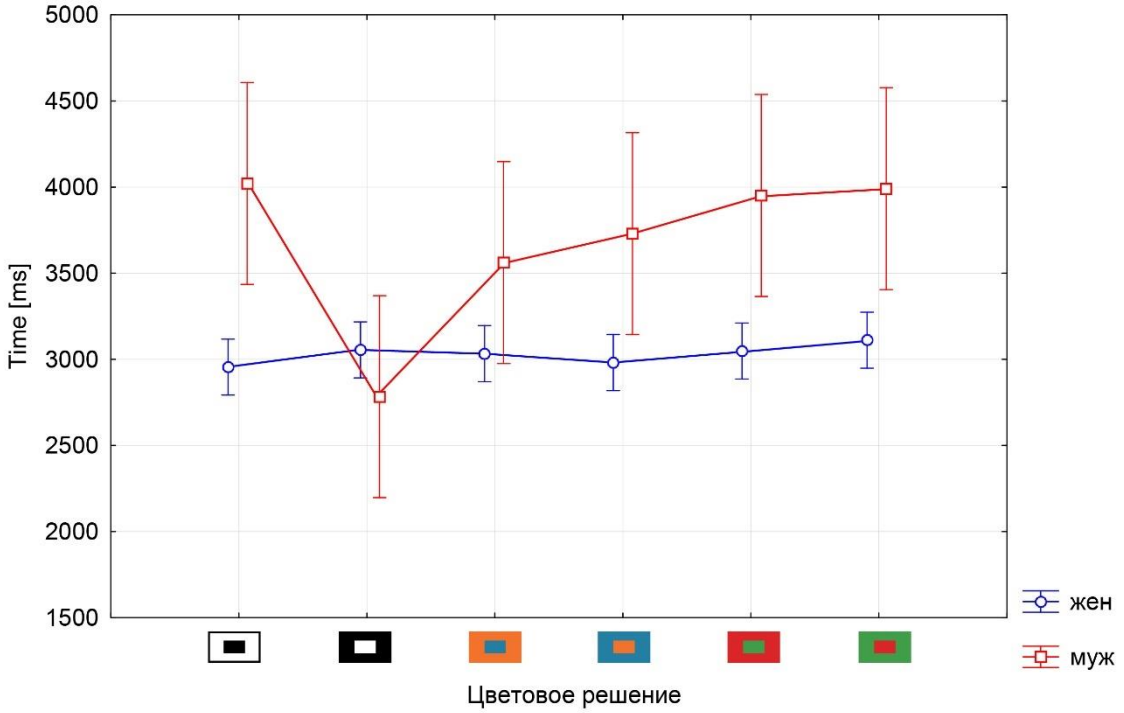


Рис. 9. График распределения значений времени рассматривания стимула испытуемыми с художественной подготовкой в зависимости от цветового решения и полового признака

Более детальный анализ позволяет выявить некоторые особенности при решении испытуемыми поставленной в эксперименте задачи. В таблице 3 представлены вычисленные значения p-value для фактора Scanpath Length [px], который отражает путь взгляда испытуемого, измеренный в пискелях.

ТАБЛИЦА 3

ВЫЧИСЛЕННЫЕ ЗНАЧЕНИЯ P-VALUE ДЛЯ ПАРАМЕТРА SCANPATH LENGTH [PX]

Фактор	**p-value**
Цветовое решение	0.1739
Композиция	0.0024
Положение Меню	0.00001
Половой признак	0.9556

Статистический анализ выявил статистически значимую зависимость пути взгляда в процессе рассматривания стимула испытуемым от фактора композиции и фактора положения меню.

Анализируя зависимость пути взгляда рассматривания стимула испытуемыми, от фактора положения меню (рис. 10), можно увидеть значительное сокращение пути взгляда при взаимодействии с интерфейсом веб-сайта с вертикальном расположением меню.

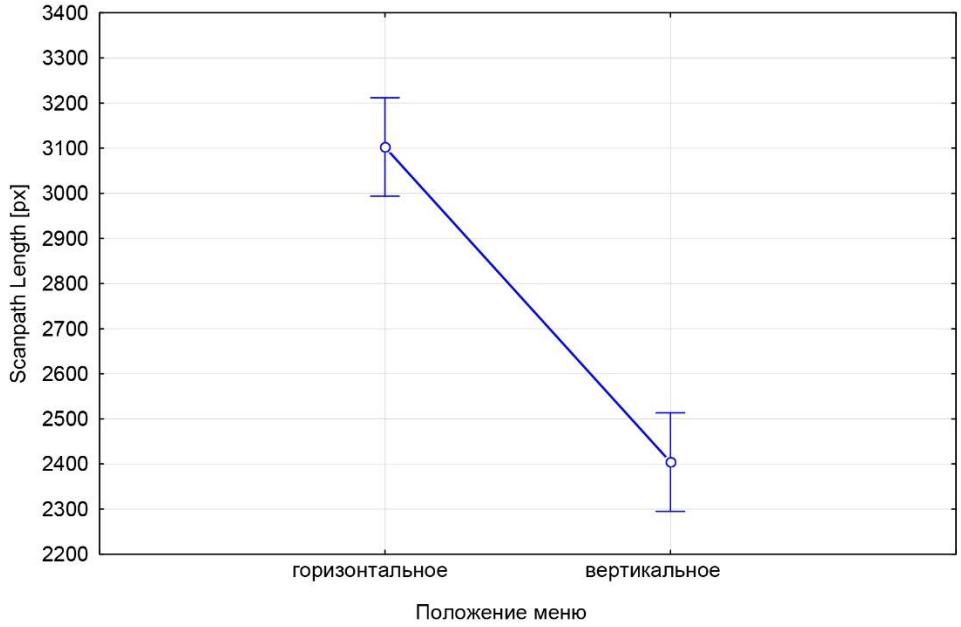


Рис. 10. График распределения значений пути взгляда испытуемых в пикселях в зависимости от фактора расположения меню в интерфейсе

Время не меняется, но путь взгляда испытуемых в случае с вертикальным меню гораздо меньше, что позволят косвенно говорить о снижении когнитивной нагрузки. Итак, мы увидели влияние всех исследуемых факторов элементов интерфейса на шаблон рассматривания.

VI. Выводы и заключение

Существует статистически значимое влияние композиции на идентификацию элементов в стимульном материале. При разработке интерфейса вебсайта для ускорения взаимодействия пользователя с сайтом стоит использовать центральную композицию макета.

Если говорить про цветовое решение сайта, в результате эксперимента было выявлено статистически значимое влияние фактора цвета для испытуемых, обладающих художественной подготовкой. Стоит отметить, что для более глубокого анализа и более полных выводом касательно разработки интерфейса вебсайта необходимо проведение дополнительных экспериментов.

Вертикальное меню позволяет снизить длину пути взора по экрану, измеренную в пикселях. Однако, время выполнения задачи испытуемыми в случае горизонтального меню приблизительно такое же. Следовательно, скорость движения взора в горизонтальном меню выше, т.е. выше физическая нагрузка на глаза. Можно сделать предположение, что при длительной работе горизонтальное меню будет приводить к большей утомляемости пользователя веб-сайта.

Список литературы

1. Smith, T. W., Snyder, C. R., & Handelsman, M. M. (1982). On the self-serving function of an academic wooden leg: Test anxiety as a self-handicapping strategy. *Journal of Personality and Social Psychology, 42*(2), 314–321. https://doi.org/10.1037/0022-3514.42.2.314
2. Norman, D. A. (1988). *The psychology of everyday things.* Basic Books.
3. Ben Shneiderman, "A Computer Graphics System for. Polynomials," The Mathematics Teacher, Vol. 67, No. 2,. Feb. 1974, pp. 111-113
4. Nielsen, J. (1993) Usability Engineering
5. Rogers, E.M. (2010) Diffusion of Innovations. 4th Edition, Simon and Schuster, New York.
6. Сергеев С.Ф. Интеллектные симбионты организованных техногенных средств управления подвижными объектами // Мехатроника, автоматизация, управление. 2013. № 9. С. 30-36.
7. Юрьев Ф.И. Цветовая образность информации. Том второй. Гармония сфер / Киев. — 2007. — 327 с.
8. Сергеев С. Ф.: Санкт-Петербургское отделение Научного совета по методологии искусственного интеллекта и когнитивных исследований РАН в 2020 году. Философские науки. 2021;64(1):149-154. https://doi.org/10.30727/0235-1188-2021-64-1-149-154]
9. Орлов П. А., Лаптев В. В., Иванов В. М.: К вопросу о применении систем ай-трекинга// Информатика, телекоммуникации и управление.URL: https://cyberleninka.ru/article/n/k-voprosu-o-primenenii-sistem-ay-trekinga
10. Borevich E., Mescheryakov S., Yanchus V.:Computer Eye-Tracking Model to Investigate Influence of the Viewer's Perception of the Graphic Information. pp.720-728. 10.20948/graphicon-2021-3027-720-728.
11. Bruce N., Tsotsos J. Weiss Y., Schölkopf B., Platt J.:(2006). Saliency based on information maximization. Advances in neural information processing systems 18 pp.155–162. Cambridge, MA: MIT Press
12. Bundesen C.: A theory of visual attention. Psychological Review, 1990. pp.523–547.
13. Carmi R., Itti L.:(2006). The role of memory in guiding attention during natural vision. Journal of Vision, pp. 898–914, http://journalofvisionorg/6/9/4/, doi:101167/694
14. Грешников И.И., Лаврова Г.А., Сальников Т.Д., Златомрежев В.И., Сергеева Г.В. Методика субъективных оценок информационно-управляющего поля кабины пилотов / Нейрокомпьютеры: разработка, применение. 2020. Т. 22. № 3. С. 18-25.
15. В.Э. Янчус Информационная модель восприятия визуальной информации человеком ГрафиКон 2023: 33-я Международная конференция по компьютерной графике и машинному зрению, 19-21 сентября 2023 г., Институт проблем управления им. В.А. Трапезникова Российской академии наук, г. Москва, Россия DOI: 10.20948/graphicon-2023-969-975
16. Koffka, K. Principles of Gestalt psychology /K. Koffka. -New-York., Routledge, 1935. - 720 p.
17. Янчус В.Э., Боревич Е.В., Авдеева А.А. Применение технологии ай-трекинга в вопросах исследования восприятия графической информации // Программные системы и вычислительные методы. 2021. №1. URL: https://cyberleninka.ru/article/n/primenenie-tehnologii-ay-trekinga-v-voprosah-issledovaniya-vospriyatiya-graficheskoy-informatsii (дата обращения: 02.07.2024).
18. Goodman, S. N.[англ.]. Aligning statistical and scientific reasoning: Misunderstanding and misuse of statistical significance impede science : [англ.] // Science. — 2016. — Vol. 352, no. 6290. — P. 1180—1181. — doi:10.1126/science.aaf5406.
19. Nuzzo, R. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume : [англ.] // Nature. — 2014. — Vol. 506, no. 7487. — P. 150—152. — doi:10.1038/506150a.

УДК 7.011.2
DOI: 10.25206/978-5-8149-3873-2-2024-879-887

Исследование влияния композиционного решения на восприятие зрительной информации в кадре

RESEARCH ON THE INFLUENCE OF COMPOSITIONAL SOLUTION ON THE PERCEPTION OF VISUAL INFORMATION IN THE FRAME

Федулова А.В., Янчус В.Э.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия.

Fedulova A.V., Yanchus V.E.
Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia

Аннотация. В статье проводится исследование влияния композиционного решения на восприятие зрительной информации в кадре. Авторы рассматривают факторы, влияющие на скорость взаимодействия человека с иллюстративным материалом. Для постановки эксперимента был разработан стимульный материал, представленный в трех комплементарных цветовых решениях, а также с разными композиционными доминантами. Всего было подготовлено 24 иллюстрации. В работе подробно описана методика проведения эксперимента, представлена интерпретация полученных результатов. Проведенное айтрекинг-исследование позволило определить композиционные условия, влияющие на среднее время выполнение задач респондентами. Данные результаты могут быть использованы для разработки анимации, рекламы и т.п.

Ключевые слова: комплиментарные цветовые схемы, стимульный материал, постобработка, айтрекинг, восприятие цвета, композиционное решение, превалирование, фон, графический образ**.**

Annotation. The article studies the influence of compositional decisions on the perception of visual information in the frame. The authors consider factors influencing the speed of human interaction with illustrative material. To set up the experiment, stimulus material was developed, presented in three complementary colour schemes, as well as with different compositional dominants. A total of 24 illustrations were prepared. The work describes in detail the experimental methodology and presents the interpretation of the results obtained. The conducted eye-tracking study made it possible to determine the compositional conditions that influence the average time it takes respondents to complete tasks. The conducted eye-tracking research made it possible to determine the compositional conditions that influence the average time it takes respondents to complete tasks. These results can be used to develop animation, advertising, etc.

Key words: complementary colour schemes, stimulus material, post-processing, eye tracking, colour perception, compositional solution, prevalence, background, graphic image.

I. Введение

Композиция является системой, предполагающей определенную формализацию. Р. Арнхейм определяет композицию и другие изобразительные средства как «визуальные силы, заложенные в направлении взгляда». Автор утверждает, что благодаря гармоничному согласованию с фоном элемент композиции становится по-настоящему видимым [1].

Композиционное решение не может быть обусловлено только эстетическими характеристиками, поскольку оно влияет на процесс рассматривания [2]. Поэтому условия организации элементов внутри визуальной среды должны быть научно обоснованы. За счет композиционных акцентов, доминант, сочетаний, контрастов зритель может быстрее воспринимать информацию или удерживать фокус внимания на самых важных для него объектах. Более того, композиционный строй может обеспечить необходимый для эффективной обработки информации уровень эмоционального напряжения у человека.

В эпоху ускоренного темпа жизни возникла проблема неустойчивой коммуникации с визуальной информацией. Создание осознанного, логичного и эргономичного для человеческого восприятия дизайна становится актуальной задачей. Распространение в обществе такого феномена как “баннерная слепота” [3] требует поиска способов привлечения внимания человека и формирования зрительского интереса за короткий период времени. Затруднения в визуальной коммуникации могут быть преодолены через регуляцию с низшего,

сенсорно-перцептивного уровня [4]. Среднее время реакции человека на внешние стимулы возрастает соответственно количеству альтернатив [5]. Данный закон Хика находит свое практическое подтверждение и при взаимодействии зрителя с художественным пространством. Чем больше в иллюстрации деталей, знаков, тем больше времени и ресурсов потребуется от зрителя [6]. Соответственно, для разного типа визуального взаимодействия графического изображения с человеком необходимо учитывать закономерно влияющие факторы. При контакте с визуальными объектами в городской среде или на web-ресурсе у людей формируются невзаимосвязанные по специфике запросы. Например, в уличном пространстве человеку потребуются более выразительные образы для привлечения внимания, однако для архитектуры сайта это может стать негативным фактором, отвлекающим от решения прямых задач.

“Пропускная” способность человека к восприятию визуальной информации определяется возможностями ее внутренней организации [7]. Организация логичного взаимодействия составляющих композиции может помочь зрителям осознать интегрированную структуру целого. Ведь многие гештальт-эксперименты доказывают, что “...проявление любого элемента зависит от его места и функции в модели целого. Всякий мыслящий человек не может не восхищаться активным стремлением к единству и порядку...”, а “мышление в борьбе за истинное понимание реальной действительности развивается (в своей логической эволюции) от наиболее простых, зрительно воспринимаемых моделей к наиболее сложным” [8].

Первый этап взаимодействия человека с визуальной информацией возникает на перцептивном уровне через органы зрения. Если при поступлении данных на первый уровень восприятия есть определенные неточности, то переход на второй этап (обработка), а также на третий (формирование эмоционального отклика, подключение личного опыта человека и т.д.) будет более сложным. В связи с этим необходимо изучать “зоны зрительского интереса” [9]. Выявить показатели визуальной привлекательности представляется возможным с помощью использования специального аппаратного комплекса фиксации глазодвигательной активности – айтрекинг-установки. Прохождение первых двух уровней можно отследить через айтрекинг-систему, однако завершающий этап включает в себя нестабильные психо-эмоциональные связи (настроение, самочувствие, уровень образования, память и т.д.). Подобное двухстороннее взаимодействие между объектом и субъектом усложняет получение достоверных результатов эксперимента.

II. Постановка задачи

Цель исследования – выявить факторы влияния композиционного решения и цветового сочетания на восприятие зрителем графического материала в кадре. Для решения поставленной цели была разработана методика эксперимента, включающая в себя следующие этапы:

1. Подготовка стимульного материала.
2. Создание эксперимента на программно-аппаратном комплексе SMI RED 250.
3. Определение задачи испытуемым во время прохождения эксперимента.
4. Статистическая обработка результатов экспериментов.

Подготовка стимульного материала. Были определены факторы исследования: 1) композиционный баланс между графикой и фоном, 2) сочетания комплементарных цветов. Стимулы можно разделить по шести признакам: “доминирование графики в сине-оранжевом цвете” (рис.1); “доминирование фона в сине-оранжевом цвете” (рис.2); “доминирование фона в красно-зеленом цвете” (рис.3); “доминирование графики в красно-зеленом цвете” (рис.4), и “доминирование фона в желто-фиолетовом цвете” (рис.5), “доминирование графики в желто-фиолетовом цвете” (рис.6).

Для проведения исследования было разработано 24 стимула. Выбор печатной графики обусловлен апробированной функцией коммуникации с аудиторией [10]. Креолизованный текст составляет единое визуальное целое благодаря наличию и вербальных (текст), и иконических (изобразительные символы) знаков [11]. Поэтому обязательным условием являлось наличие графических элементов (шрифт+образ) и фонового наполнения.

Оригинальные цвета иллюстративного материала были замещены на комплементарные сочетания: красный и зеленый, желтый и фиолетовый, оранжевый и синий. Выбор данных цветовых решений обусловлен теорией цветовых контрастов Иттена [12,13]. Раннее ряд ученых уже выяснил, что использование комплементарного цветового решения является способом ускорения восприятия зрительной информации [14].

Рис.1. Доминирование графики в сине-оранжевом цвете

Рис.2. Доминирование фона в сине-оранжевом цвете

Рис.3. Доминирование фона в красно-зеленом цвете

Рис.4. Доминирование графики в красно-зеленом цвет.

Рис.5. Доминирование фона в желто-фиолетовом цвете

Рис.6. Доминирование графики в желто-фиолетовом цвете

Дополнительно была проведена коррекция соотношения графики и фона. Выделено два композиционных решения: превалирование фона над графикой и графики над фоном соответственно.

Исследование ориентировано на скорость выполнения задач. Важной задачей стало избежание влияния косвенных факторов на чистоту эксперимента. Поэтому для сложности быстрого декодирования текста шрифтовое наполнение представлено либо с пониженной читабельностью, либо на иностранном языке. Иллюстративная составляющая плакатов также абстрактна, минимальна или искажена.

III. Теория.

Методика проведения эксперимента. Для проведения эксперимента был использован программно-аппаратный комплекс айтрекинга SMI RED 250. [15]. Размер экрана составлял 640x400 мм (1920*1080 pix), удаленность глаз соответствовала 600 мм.

20 респондентов были разделены на фокус группы по следующим признакам: гендер, художественная подготовка. Возраст участников составил от 18 до 25 лет. Перед началом исследования с использованием айтрекинговой системы с участниками была проведена беседа о ходе эксперимента. Перед респондентами были поставлены следующие задачи: 1. Посмотреть на изображение и запомнить его. 2. Определить те изображения, которые были просмотрены раннее (интервал: 30 минут).

Эксперимент проводился в 2 этапа: 1. Восприятие и запоминание; 2. Выявление усвоенного визуального материала по истечении 30 минут (после прохождения ключевой отметки согласно кривой забывания Эббингауза [16]; 3. Данные задачи позволяют изучить скорость взаимодействия со зрительной информацией в зависимости от цветовых и композиционных решений, пола, художественной подготовки, повторной демонстрации.

На первом этапе участнику тестирования на мониторе демонстрировались 12 стимулов: по 2 изображения согласно признакам шести групп (“красно-зеленое цветовое решение с доминированием графических элементов”; “красно-зеленое цветовое решение с доминированием фона”; “оранжево-синее цветовое решение с доминированием графических элементов”; “оранжево-синее цветовое решение с доминированием фонового наполнения”; “желто-фиолетовое цветовое решение с доминированием графических элементов”, “желто-фиолетовое цветовое решение с доминированием фонового наполнения”). Задачей респондента являлось запомнить увиденные картинки. Время эксперимента не было ограничено.

После прохождения первой части эксперимента был выдержан временной интервал в 30 минут, после чего респонденты приступали к прохождению следующего этапа эксперимента. На втором этапе к стимульному материалу было добавлено еще 12 картинок: по 2 стимула согласно признакам шести групп. Новые изображения были выведены на экран в хаотичном порядке с теми, которые раннее демонстрировались респондентам. Участникам исследования было необходимо определить из ряда изображений те, которые уже были просмотрены. От респондентов требовались закрытые ответы в форме “да” или “нет”.

Полученные в ходе исследования экспериментальные данные были обработаны методом математической статистики.

IV. Результаты экспериментов

Был произведен анализ следующих параметров шаблона рассматривания: время наблюдения стимула (time), средняя длительность фиксаций (период времени, в течение которого глаза фиксируются на зрительной цели, восприятие стабильно, и глаза воспринимают фиксированную информацию [17]), среднее количество фиксаций при наблюдении одного стимула, средняя длительность саккад (одновременные согласованные движения глаза от одной фиксации к другой [17]), среднее количество саккад при наблюдении одного стимула, максимальная длина взора. Задачей анализа было выявить влияние факторов демонстрации, композиционного и цветового решений, а также полового признака и наличия художественной подготовки.

Более подробный анализ демонстрирует закономерности при решении респондентами поставленных задач. Анализ выявил статистически значимую зависимость времени рассматривания стимула испытуемым от факторов демонстрации, композиционной доминанты (превалирование фона над графикой/графики над фоном), цвета, художественного образования и гендерного признака. Аналогичные показатели также статистически влияют на длину взора, за исключением фактора цвета.

Уровень значимости критерия p-value был принят равным 0.05.

Вычисленные значения p-value с помощью программной процедуры ANOVA для времени рассматривания стимула в зависимости от факторов демонстрации, композиции, цвета и пола указаны в следующих таблицах:

Табл. 1 демонстрирует зависимость времени выполнения задач от фактора демонстрации, цвета и пола.

ТАБЛИЦА 1

Данные зависимости времени выполнения задач от фактора демонстрации, цвета, пола.

Фактор	**Univariate Tests of Significance for Time [ms]**
	p-value
Demo	0.017
Colour	0.049
Gender	0.003

Табл. 2 демонстрирует зависимость скорости выполнения задач от фактора демонстрации, цветового решения и художественной подготовки.

ТАБЛИЦА 2

Данные зависимости времени выполнения задач от фактора демонстрации, цвета, художественного образования.

Фактор	**Univariate Tests of Significance for Time [ms]**
	p-value
Demo	0.023
Colour	0.038
Art	0.001

В Табл. 3 показаны результаты зависимости скорости выполнения задач от фактора демонстрации, цвета и композиции.

ТАБЛИЦА 3

Данные зависимости времени выполнения задач от фактора демонстрации, цвета, композиции.

Фактор	**Univariate Tests of Significance for Time [ms]**
	p-value
Demo	0.015
Colour	0.044
Composition	0.007

Табл. 4 демонстрирует зависимость длины взора от фактора демонстрации и композиции. Исходя из результатов данной статистики цвет не оказывает влияние на формирование общей закономерности.

ТАБЛИЦА 4

Данные зависимости длины взора от фактора демонстрации, цвета, композиции.

Фактор	**Univariate Tests of Significance for Scanpath Length [px]**
	p-value
Demo	0.001
Colour	0.536
Composition	0.005

V. Обсуждение результатов

Было выяснено, что цветовое решение не имеет статистического влияния на длину взора, однако композиционное решение, пол, повторная демонстрация, наличие художественного образования формируют закономерные связи (на время решения задачи эксперимента).

Было установлено, что превалирование графики над фоном способствует более быстрому решению задачи идентификации стимула. Результаты показали, что взаимодействие респондента с уже раннее просмотренным стимулом приводит к более длительному решению задачи. Данный факт подтверждает наличие дополнительной когнитивной нагрузки и удержание фокуса внимания для осуществления процесса вспоминания. Зелено-красный, фиолетово-желтый – при доминировании графики, а при доминировании фона – предпочтительно сине-оранжевое сочетание (рис.5).

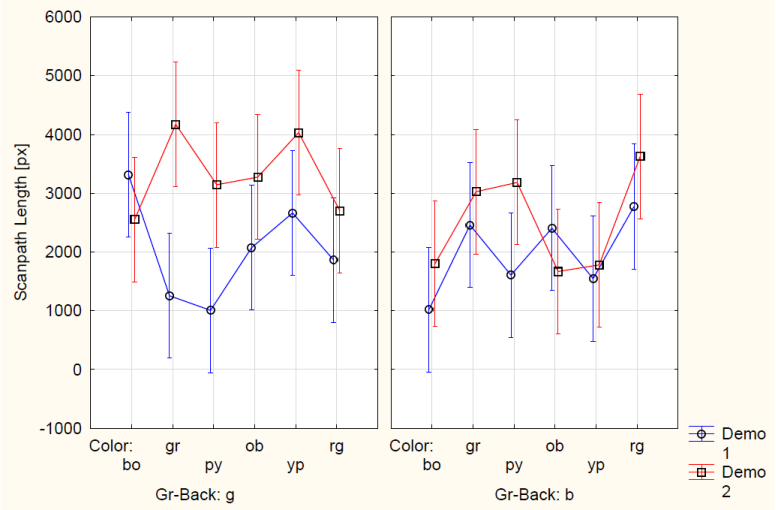


Рис. 5. Влияние композиционного баланса и повторной демонстрации изображений на длину взора.

В результате эксперимента было статистически установлено, что лица мужского пола справляются с задачами быстрее (рис.6).

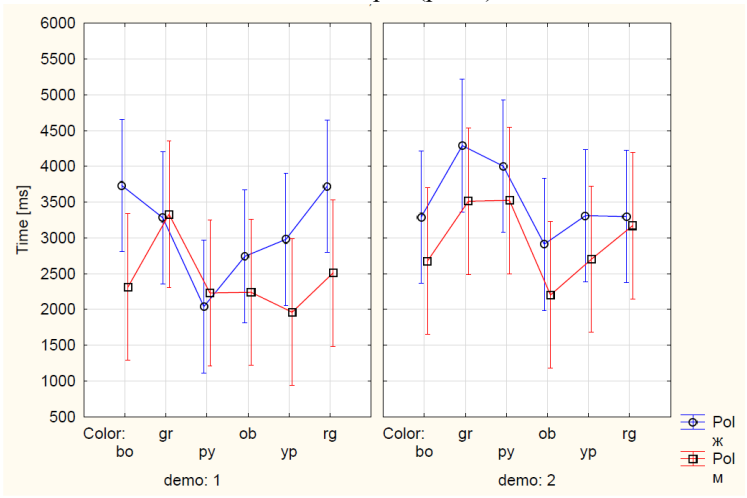


Рис.6. Влияние полового признака на скорость выполнения задач.

Также наличие художественного образования у респондентов стало фактором более длительного решения задач (рис.7).

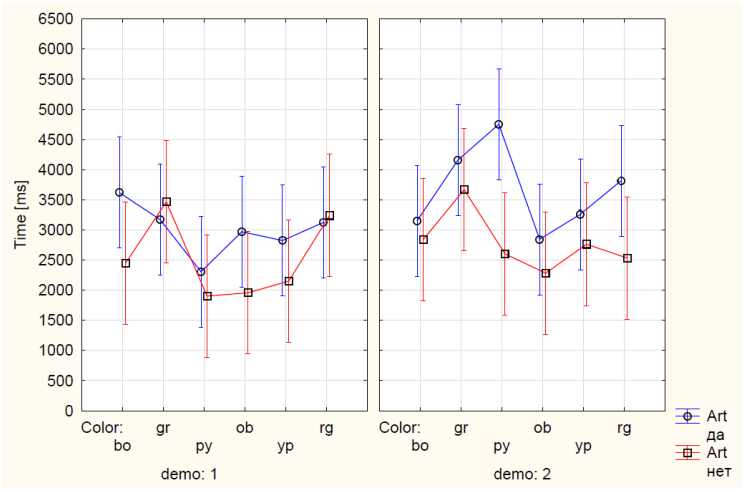


Рис.7. График зависимости скорости выполнения задач от наличия художественного образования.

Анализируя график на рис. 8, можно заключить, что средняя длительность фиксаций при повторной демонстрации изображений больше, за исключением некоторых сине-оранжевых комбинаций.

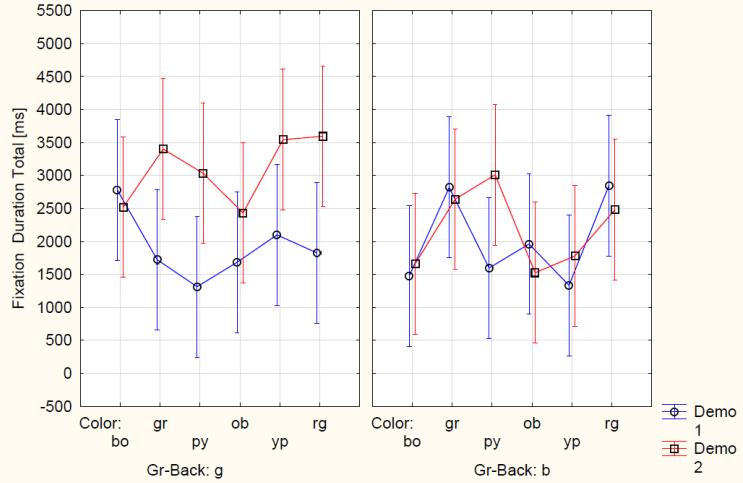


Рис 8. График зависимости фиксаций от фактора демонстрации.

VI. Выводы и заключение

При разработке интерфейса приложения или сайта для работы и подробного изучения информации подходят варианты с уклоном на визуальную эргономику, однако при взаимодействии с пользователем на первом перцептивном уровне необходимы более усложненные композиционные решения. Постановка более сложных задач может помочь привлечь зрительское внимание, но при работе с данными для преобразования в информацию человек будет эффективнее справляться с более простыми, абстрагированными от сложных стилистических и цветовых решений средами.

Исходя из результатов эксперимента при разработке видеоматериалов целесообразно использовать следующие решения: сине-оранжевое сочетание при преобладании фона, желтый фон и фиолетовая графика при превалировании графического образа в композиции.

Выводы.

1. Разработанная методика проведения экспериментального исследования показала свою работоспособность.
2. Были выявлены определенные статистические закономерности влияния определенных факторов на шаблон рассматривания.
3. Полученные результаты можно использовать при разработке визуально привлекательных и запоминающихся видеоматериалов.

Обсуждения:

Был проведен анализ тепловых карт (от англ. «heat map») — графическое представление данных, где индивидуальные значения в таблице отображаются при помощи цвета [18]. Стоит отметить, комплиментарный контраст может быть важным визуальным акцентом [19]. Значительная концентрация фиксаций была расположена не только на круглом объекте, но и на фоне, рядом с границей формы (краевой контраст) (рис.8).

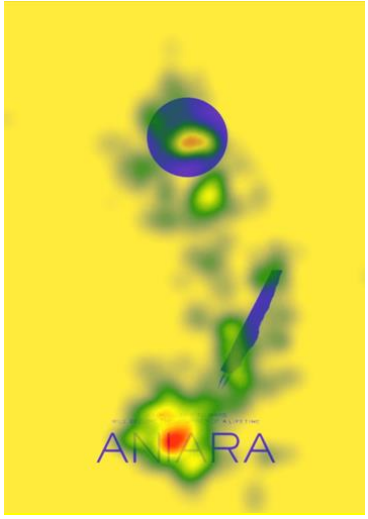


Рис.8. Тепловая карта 1.

Анализ тепловых карт также показал, что все испытуемые фиксировали больше внимания на тексте, а не на графическом образе. Первая и последняя буквы слов практически игнорировались в отличие от центральной или первой буквы последующего слова/части фразы (рис.9).

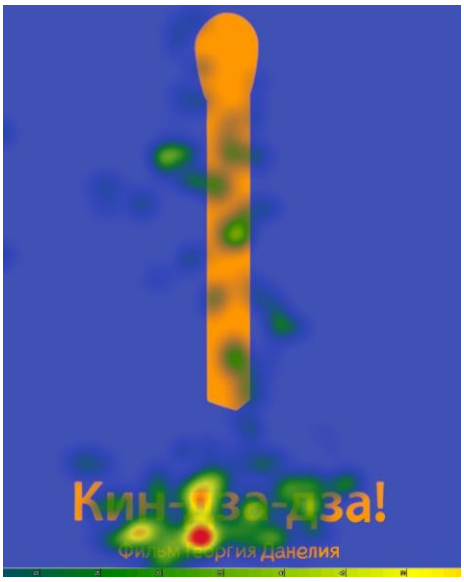


Рис.9. Тепловая карта 2.

Также графические элементы с пониженной интенсивностью цвета привлекают зрительское внимание, которое не ограничено окуломоторным диапазоном (рис10), [20].

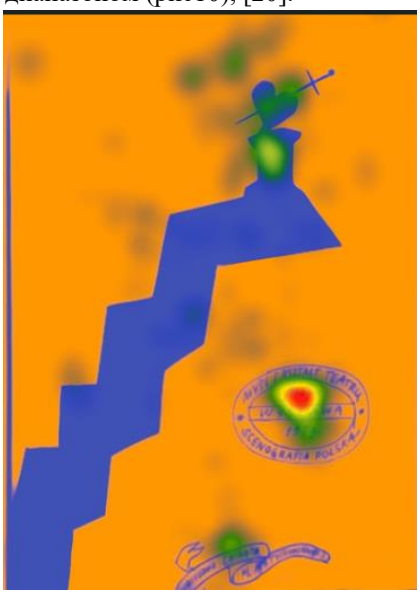


Рис.10. Тепловая карта 3.

Список литературы

1. Арнхейм Р. Новые очерки по психологии искусства. Пер. с англ. Москва: Прометей, 1994. С. 12-13.
2. Боревич Е.В. — Ай-трекинговое исследование влияния композиции на восприятие кинокадра // Программные системы и вычислительные методы. – 2023. – № 1. DOI: 10.7256/2454-0714.2023.1.39634 EDN: IWYBNX URL: https://nbpublish.com/library_read_article.php?id=3963/ (дата обращения: 21.06.2024).
3. Benway, J. P. Banner blindness: web searchers often miss 'obvious' links / J.P. Benway, D.M. Lane. – Houston, Texas.: Rice University, 1998. – P. 11. – URL: https://www.ruf.rice.edu/~lane/papers/banner_blindness.pdf (дата обращения: 27.06.2024).
4. Костин А.Н. Гибкое изменение степени автоматизации как средство распределения функций между человеком и автоматикой // Институт психологии российской академии наук. Организационная психология и психология труда. 2018. Т.3. №3. С. 218-232.

5. Hick W. E. On the rate of gain of information // Quarterly Journal of Experimental Psychology. 1952. 4 (1): P. 11-26.

6. Лотман Ю.М. Искусствознание и точные методы в современных зарубежных исследованиях // Семиотика и искусствометрия: современные зарубежные исследования: сборник переводов. 1972. С. 5-23.

7. Бакаев М.А. Определение сложности задач для зрительно-пространственной памяти и пропускной способности человека-оператора // Управление большими системами: Сборник трудов. 2017. №40. С. 25-57.

8. Арнхейм Р. Искусство и визуальное восприятие. Москва: Прогресс, 1974. 288 с.

9. Лаптев М.В., Янчус В.Э., Лаптев В.В. Выявление зон интереса пользователя при визуализации данных с использованием метода ай-трекинга // Программные системы и вычислительные методы. – №2. – 2023. – С. 49-62. – URL: https://nbpublish.com/library_read_article.php?id=43502 (дата обращения: 01.07.2024).

10. Ухова Л.В. Эффективность рекламного текста : монография. – Ярославль: Изд-во ЯГПУ, 2012. – 393 с.

11. Креолизованный текст: Смысловое восприятие. Коллективная монография / Отв. ред. И.В. Вашунина. Ред. колл.: Е.Ф. Тарасов, А.А. Нистратов, М.О. Матвеев. – М.: Институт языкознания РАН, 2020. – 206 с.

12. Матюшин М.В. Справочник по цвету /Закономерность изменяемости цветовых сочетаний// М.В. Матюшин. – М.: Аронов Д., 2007.

13. Иттен И. Искусство цвета / пер. с нем. Л. Монаховой. – 2е изд. – М. : Д. Аронов, 2001. – 96 с.

14. Янчус В.Э., Боревич Е.В. Исследование значения цветового решения в процессе гармонизации кинокадра / Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2016. – № 4(252). – С. 53-68.

15. Новый стиль. Система удаленного трекинга глаз RED 250 / RED 500 [Электронный ресурс] URL: https://newstyle-y.ru/high-school/group_2164/group_2166/item_11065/? (дата обращения 07.07.2024).

16. Боревич Е.В., Мещеряков С.В., Янчус В.Э. Эффективные методы и модели цифровой обработки киноматериала // Графикон-2017: тр. 27-й междунар. конф. по компьютерной графике и машинному зрению. – Пермь: Пермский государственный научно-исследовательский университет, 2017. – С. 51-54.

17. Rayner K. Eye movements and attention in reading, scene perception, and visual search. Q. J. Exp. Psychol. 2009. – V. 62 (8). – P. 1457-1506.

18. Бондарь О.В., Чмутин А.М., Чмутин М.А. Алгоритм управления контрастом насыщенностей цифровых изображений и его программная реализация. // Инженерный вестник Дона. 2020. № 5. 16 с. [Электронный ресурс]. - Режим доступа: http://ivdon.ru/ru/magazine/archive/n5y2020/6458 (дата обращения: 09.07.2024).

19. Wilkinson L., Friendly M. The History of the Cluster Heat Map // The American Statistician, 2009. – Vol. 63. – № 2 – P. 179-184.

20. Hanning N., Szinte M., Deubel H.: Visual attention is not limited to the oculomotor range. In: Proceedings of the National Academy of Sciences of the United States of America, USA. 2019. – P. 9665-9670.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-886-894

Исследование визуального восприятия дискретных цветовых шкал в сложных рельефах

Study of visual perception of discrete colour scales in complex reliefs

М. В. Лаптев, В. Э. Янчус
1,2 СПБПУ, Политехническая ул., 29, Санкт-Петербург, 195251, Россия

M. V. Laptev, V.E. Janchus
[1]SPBPU, Politekhnicheskaya st., 29, St. Petersburg, 195251, Russian Federation

Аннотация. *Целью данной статьи является проверка выдвигаемой гипотезы о восприятии цветовых дискретных шкал в сложных рельефах при их делении на 5, 7 и 9 сегментов. В основе исследования лежит эксперимент, проведенный с помощью системы ай-трекинга, по результатам которого были получены наглядные результаты. Выявлено преимущество визуального отображения сложных рельефов, поделенных на 5 и 7 сегментов, когда карты с разделением на 9 сегментов показали свою неэффективность. Для предложенных в эксперименте стимулов различие в тоне цвета не является статистически значимым признаком.*

Ключевые слова**:** *ай-трекинг, фиксации, саккады, сложные рельефы, оформление карт, цветовое решение, визуализация данных, дискретные шкалы.*

Аннотация. *The purpose of this paper is to test the hypothesis put forward about the perception of color discrete scales in complex reliefs when they are divided into 5, 7 and 9 segments. The study is based on an experiment conducted with the help of an eye tracking system, the results of which were visualized. The advantage of visual display of complex landforms divided into 5 and 7 segments was revealed, when maps with division into 9 segments were shown to be ineffective. For the stimuli offered in the experiment, the difference in color tone was not a statistically significant feature.*

Ключевые слова: *eye-tracking, fixations, saccade, complex reliefs, map design, colour scheme, data visualization, discrete scales.*

I. Введение

Графическое представление поверхности сложного рельефа активно встречается в различных областях деятельности человека. Так в повседневности ярким примером визуализации сложного рельефа можно назвать картографию, и в особенности тематическую картографию, где сложные данные распределены по карте, а принцип визуального отображения этих данных различен и выбирается автором заранее из наиболее подходящих методик. Еще одним примером различных сложных рельефов, где автор самостоятельно выбирает метод отображения данных, можно назвать инфографику. В ней также существует вариативность донесения до зрителя идеи через визуальную составляющую.

Значительной областью жизни человека, где не обойтись без графического отображения сложных рельефов, является промышленность. В большинстве диагностической аппаратуры, объем данных выводится к оператору именно в виде точного рельефа, распределенного по уровням, ясно дающий понять о структуре исследуемого объекта. Данный способ устоявшийся, например в таких приборах как дефектоскоп, где дефект в исследуемом объекте необходимо показать на карте поверхности, или в некоторых медицинских аппаратах, таких как томограф или аппарат УЗИ. Глубокая интеграция во множество общедоступных сфер жизни и быта человека говорит об актуальности вопроса исследования корректного отображения сложных рельефов поверхности.

Необходимо отметить, что корректность и достоверность проектных методик отображения научных и прикладных данных, распределенных по поверхности сложного рельефа, является предметом споров и дискуссий среди представителей научного сообщества. Так, последние два десятилетия идет активное обсуждение правильности использования спектральных карт для отображения полученных данных [1], которые получили наибольшее распространение и популярность в научных кругах. Хотя в последние годы количество научных трудов с представлением данных спектральным методов сокращается [2], есть много сторонников именно такого способа вывода информации [3]. Также научным сообществом ставится вопрос о количестве сегментов последовательных шкал и о предпочтительных характеристиках цвета [4]. Решению указанных научных проблем в части визуального восприятия дискретных цветовых шкал в сложных рельефах посвящена данная работа.

II. Постановка задачи

На основании исторического опыта отображения рельефов в картографии в противовес к спектральному отображению данных, существует метод отображения сложных рельефов с использованием одного цвета [5]. В связи с широким информационным полем вокруг неоднозначности применения спектрального метода при визуализации сложных рельефов, автор счел необходимым проверить на практике именно методику визуализации большого количества данных на карте с помощью использования всего одного спектрального тона. На базе именно этого метода и будет построен дальнейший эксперимент. Гипотеза настоящего исследования предполагает, что человек корректно может воспринимать деление индикаторной шкалы на 4-6 оттенков. Также предполагается, что большее количество делений индикаторной шкалы позволяет точнее определить значение показателя в некой точке наблюдения, но влечет за собой бóльшие временные затраты и повышенный риск ошибки. Это говорит о снижении эффективности восприятия графики, которая определена Ж. Бертеном в параметрах скорости и правильности решения поставленной задачи [6].

Первостепенной задачей в эксперименте является аргументированный выбор анализируемых цветов, на основании которых будут формироваться сложные рельефы. Вторым этапом необходимо сформировать индикаторные шкалы с использованием каждого выбранного цвета с разным количеством оттенков. На основании решения вышеупомянутых задач подготавливаются идентичные по рельефу двумерные карты анализируемых цветов с сегментными шкалами. Следующим шагом проводится эксперимент с использованием системы ай-трекинга при участии 30 испытуемых с техническим и художественным образованием.

Посредством решения вышеперечисленных задач, финальный этап включает выявление временных затрат, без фиксации возможных ошибок при определении значений оттенков одного цвета в неких точках двумерных карт при делении индикаторных шкал на 5, 7 и 9 делений и анализ полученных результатов.

III. Теория

В настоящий момент визуализация сложных рельефов в научных трудах представлена различным образом. Следует отметить, что интерес вызывает не только объект исследования — восприятие визуализации полученных данных в сложных рельефах, но и сам метод исследования — особенности использования системы окулографии (ай-трекинга).

При рассматривании графического представления сложных рельефов можно обратиться к опыту прошлых веков и взглянуть на тенденции в наиболее близкой области — картографии. За более чем двухсотлетний опыт эта наука прошла от многоцветных карт с нелогичным распределением цветов и карт с контрастным переходом между оттенками, до логично построенных семантических карт или карт с моноцветовым распределением, где разница высот задается изменением насыщенности или яркости в рамках одного цвета [7].

В качестве альтернативы хаотической колористики карт в середине XIX века австрийский военачальник Франц фон Гауслаб выдвигает принципиально новый метод по отображению рельефа. Его карты последовательно окрашивались от светлого цвета, до темного по принципу: чем выше, тем темнее. Результатом трудов Гауслаба, стали карты с последовательностью от белого, желтого до темно-синего через коричневый цвет [8]. Однако семантический эффект от этого метода был неудовлетворительным. Территория высокогорий, ледников и заснеженных гор визуализировалось пространством темных оттенков. Следующий этап ознаменовался изменением стратегии. Австрийский картограф Карл Пейкер предложил смысловое наполнение в отображении рельефов по принципу «адаптивно-цветной объемной перспективы» [9]: от яркого насыщенного к тусклому бледному. В качестве шкалы он впервые начал использовать спектральную линейку цветов. И хотя картография за прошедшие столетия полностью перешла на визуальное представление рельефов с использованием привычных по смыслу цветов, в научном мире идея спектральных (тепловых) карт закрепилась.

Спектральное отображение данных известное, но спорное решение при визуализации данных. За последние десятилетия в научных кругах всё сильнее падает популярность этого метода, но во многих других сферах этот классический способ отображения данных всё еще активно используются. В своей работе [2] авторы проводят анализ и сведение большого количества исследований, посвящённых спектральным картам, и приходят к выводу, что хотя процент использования «радужных карт» упал с 85.7% в 1994 году до 1.9% в 2020 году, в некоторых областях науки он всё еще высок. Например, в науках о Земле процент использования спектрального отображения данных выше в 14 раз чем средний по остальным сферам. При использовании спектральных цветов оттенки хорошо различимы, но такие карты плохи для понимания зависимостей, поскольку дают ложные границы и оператору тяжелее находить закономерности в полученных результатах. Так что такой способ в научном сообществе считается неподходящим при поставленных задачах сравнения величин [10].

В настоящее время при графическом представлении данных, распределенных по поверхности, наиболее активно используется метод «чем темнее, тем больше». Его описал Янсон в 1878 г. в своей монографии [11]. Он выдвинул идею об ограниченном количестве ступеней в цветовой шкале, предположив, что она должна быть

ограничена 5 оттенками одного цвета. Развитие этих идей можно увидеть у швейцарского картографа Э. Имгофа. Он добивался шкалы в 15 шагов из трех цветов, использовав только пять ступеней каждого цвета, которые последовательно располагал друг за другом [12]. Примерно тоже самое можно увидеть в отечественных гипсометрических и батиметрических картах.

Для решения поставленной задачи был использован метод окулографии. В отличие от простой фиксации скорости и правильности решения, это необходимо для выяснения когнитивной нагрузки. Количество фиксаций и длительность саккад может стать источником для выяснения дополнительных сведений об удобочитаемости карты. Выявление зон интереса пользователя (AOI), определенные под нагрузкой решения задачи на основе совокупности паттернов рассматривания, позволит определить композиционную стратегию составления карт. Такой подход был применен в исследовании различного вида диаграмм [13], и его можно распространить на настоящее исследование визуального восприятия дискретных цветовых шкал в сложных рельефах.

IV. Результаты экспериментов

Для проверки выдвигаемой гипотезы с помощью системы ай-трекинга были выбраны 6 цветов: красный и зеленый, синий и оранжевый, пурпурный и желтый. Все цвета чистые по модели RGB. Красный (255,0,0), Зеленый (0,255,0), Синий (0,0,255), Пурпурный (255,0,255), Желтый (255,255,0) и Голубой (0,255,255). Выбранные цвета показаны на рис.1.

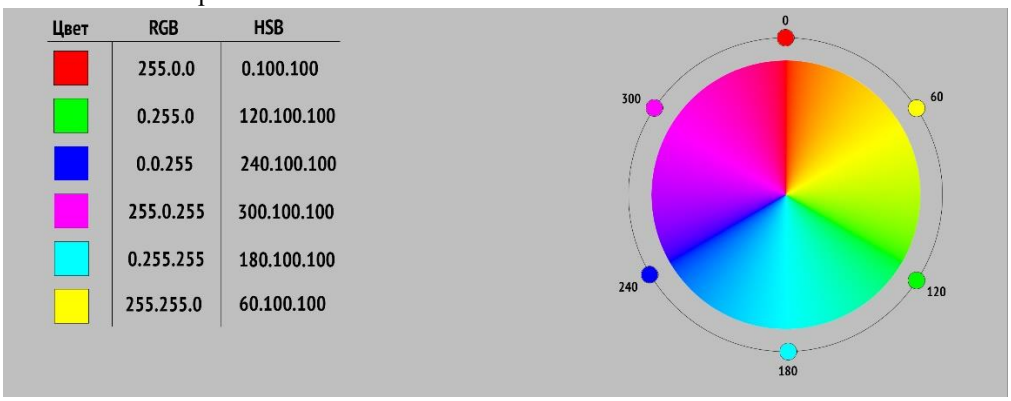


Рис. 1. Подобранные цвета

С полученными цветами были построены индикаторные шкалы в цветовой модели HSB с разделением на 5, 7 и 9 сегментов. Цвет в шкалах отличается по яркости («B» в HSB) в диапазоне от 10% до 100%. Шаг изменения равный, для индикаторных шкал с разделением на 5 сегментов — 23%, с разделением на 7 сегментов — 15% и с разделением на 9 сегментов — 11%. Индикаторные шкалы, построенные для эксперимента, представлены на рис. 2.

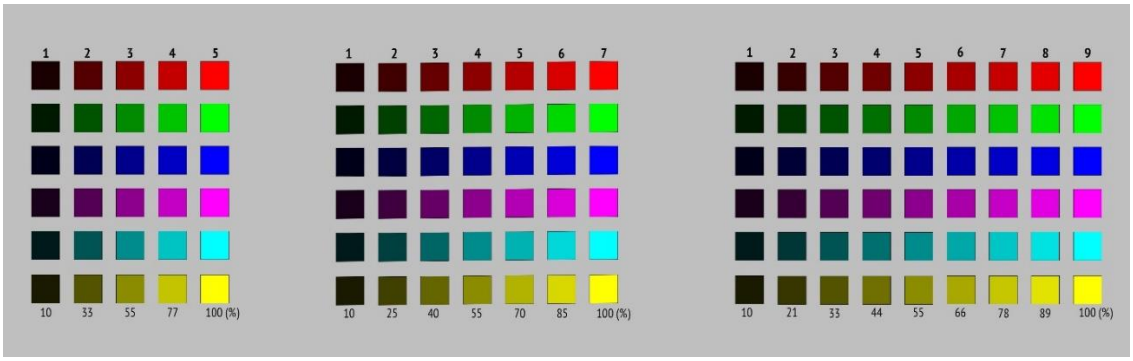


Рис. 2. Индикаторные шкалы

Для эксперимента была построена двумерная карта с изменением высот и несколькими вершинами. Карта поделена на 14 областей, где 1 — это самая низкая точка, а 14 соответствует самой высокой точке. Области разделены между собой непрерывными и замкнутыми изолиниями (рис.3). Для каждого из вариантов сегментирования шкалы карта была перестроена и выкрашена в ранее выбранный цвет. Пример изменения карты на рис.4 для зеленого цвета.

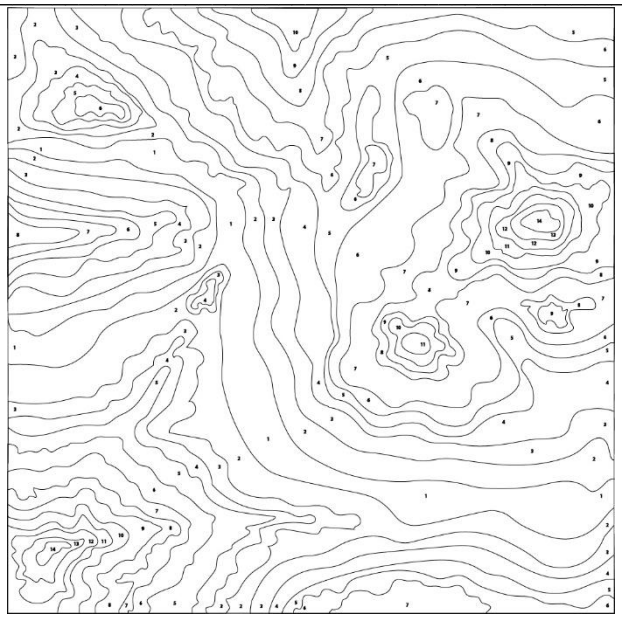


Рис.3 Двумерная карта с нанесенными изолиниями

Перед испытуемым ставилась задача, найти на карте, выкрашенной в определенный цвет, белую точку, выставленную в произвольном месте. Необходимо определить значение в этой области карты и с помощью клика компьютерной мыши выбрать нужный сегмент, соответствующий области размещения точки, на вертикальной индикаторной шкале, расположенной на сером фоне в небольшом отдалении от карты. Для исключения привыкания к карте испытуемых, с каждым новым стимулом карта поворачивалась против или по оси, а также отражалась по горизонтали или вертикали. Всего за время проведения эксперимента, испытуемый просматривал 54 стимульных материала, по три карты для каждого из цветов. При этом на каждой из трех карт для некого цвета, точка находилась в разных цветовых областях. В одном случае — точка в темной области карты, в другом — в середине, в третьем случае необходимая точка располагалась в высокой части карты, т.е на ярких участках.

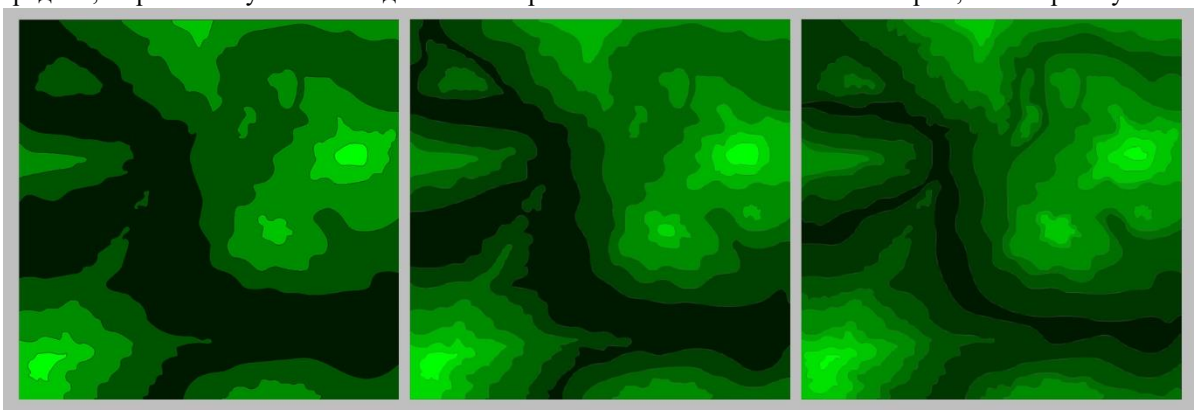


Рис.4. Изменение карты в зависимости от уровня сегментации

Эксперимент проводился на базе Лаборатории человеко-компьютерного взаимодействия Санкт-Петербургского Политехнического университета Петра Великого. Во время решения поставленной перед испытуемыми задачи глазодвигательная активность регистрировалась с помощью стационарного ай-трекера SMI RED250 и программного обеспечения SMI iViewX. Регистрировалось время решения поставленной задачи, количество фиксаций, количество саккад, путь глазодвигательной активности и итоговый выбор на карте. Предварительная подготовка испытуемых и калибровка оборудования соответствуют принятым техническими условиями, предъявляемым к экспериментам с использованием ай-трекингового оборудования [14].

В ходе эксперимента проверялась основная гипотеза исследования: что человек эффектней выполняет данные задачи при визуализации сложного рельефа на карте, поделенной на 5 сегментов; а также то, что при увеличении количества сегментов, вырастает время на решение поставленной задачи и повышается риск ошибки.

Всего, за время проведения эксперимента, в испытаниях приняли участия 30 студентов из Санкт-Петербургского Политехнического университета Петра Великого. Среди которых, 12 человек — женского пола

и 18 человек — мужского, все обладают нормальным цветовым зрением. Образование — гуманитарное и технического направление, у некоторых испытуемых есть дополнительная художественная подготовка.

Результаты проведенного эксперимента показали:

1. Уровень сегментации карты влияет на скорость решения поставленной перед испытуемым задачей. В некоторых областях карты и для некоторых выбранных цветов, например для зеленого цвета в темной области, скорость решения задачи отличается почти вдвое между картой разделенной на 5 сегментов и картой с делением на 9. При этом восприятие карт с 5 и 7 сегментами статистически не отличается, а для яркой области время решения задачи почти по всем стимулам имеет близкое значение (кроме зеленого). Как показал эксперимент, карты синего цвета более трудоемки для восприятия. Графики зависимости времени решения задачи для разных областей (темная, средняя и яркая) и выбранных цветов от уровня сегментации карты представлены на рисунке 5.

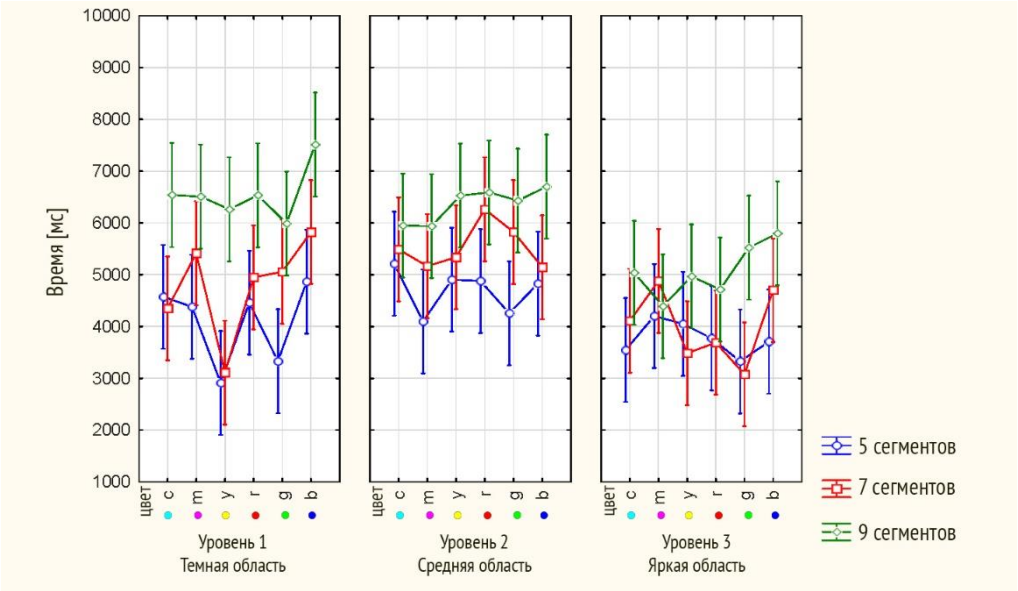


Рис. 5. Время решения задачи в зависимости от цвета и уровня сегментации

2. Количество фиксаций и количество саккад при решении задач также зависят от уровня сегментации карты. В особенности эта зависимость активно проявляется на темных и средних участках карты. Тогда как на ярких областях эта зависимость выражена слабее. Для темной области статистически значима разница восприятия 5 и 9 сегментных карт (кроме красного цвета). Визуальное восприятие 5 и 7 сегментных карт, выраженное в количестве фиксаций и саккад, статистически не различимо. Графики зависимости количества фиксаций и количества саккад на рисунках 6 и 7 соответственно.

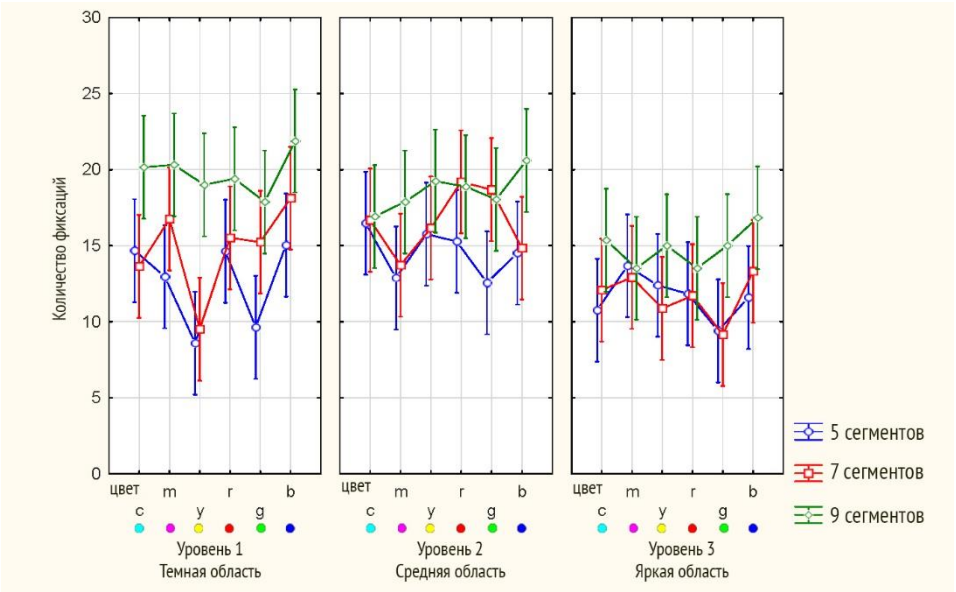


Рис. 6. Количество фиксаций при решении задачи в зависимости от цвета и уровня сегментации

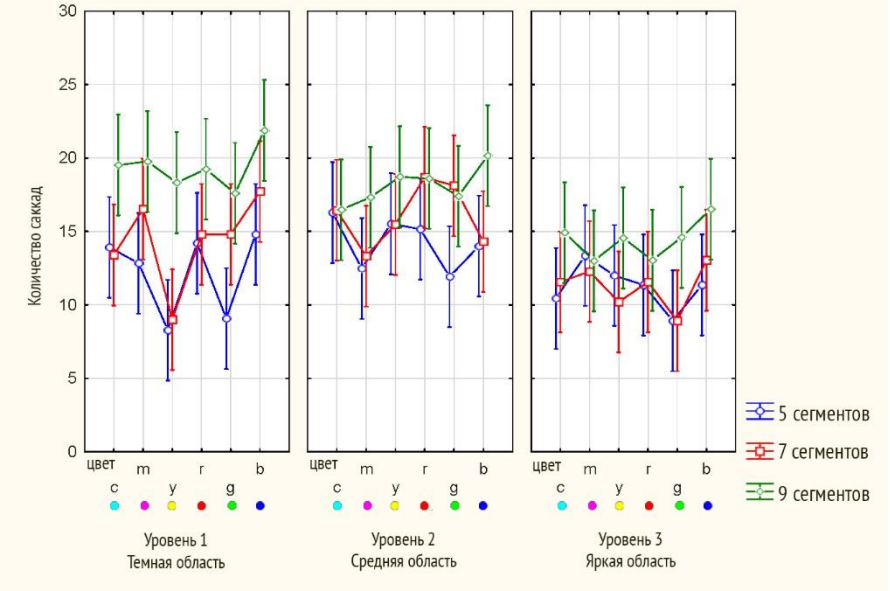


Рис. 7. Количество саккад при решении задачи в зависимости от цвета и уровня сегментации

V. Обсуждение результатов

Результаты эксперимента показали преимущество 5 и 7 сегментных карт перед делением шкалы на 9 ступеней. И хотя это в значительной степени было выявлено для отдельных цветов (зеленый и синий) и темной области карты поставленная гипотеза относительно количества ступеней в шкале в целом подтверждается. Было отмечено, что среднее время решения задачи для 5 сегментной карты составляет 4.18 секунды, тогда как для 7 сегментной в среднем 4.77 секунды, а 9 сегментной — 6 секунд. Использование сегментирования карты более чем на 7 ступеней является нецелесообразным для шкалы в один тон цвета. Остается вопрос с различием восприятия 5 и 7 сегментной карты.

Для предложенных в эксперименте стимулов различие в тоне цвета не является статистически значимым признаком, влияющим на визуальное восприятие (скорость решения задачи, количество фиксаций и саккад). Можно поставить вопрос об утомляемости пользователя при длительной работе с картами использующие возбуждающие (красный, желтый, пурпурный) или успокаивающие (зеленый, синий, голубой) оттенки цветов.

При выполнении задачи в яркой области карты время, количество фиксаций и количество саккад в среднем на всех стимульных материалах была ниже, чем при выполнении задач на средней и темной области карты. Несмотря на то, что эта разница статистически незначима, этот фактор следует учесть при составлении шкал для карт визуализирующие сложные рельефы.

VI. Выводы и заключение

Представляется важным продолжение экспериментов в данной области с выявлением зависимости восприятия карт не только с чистыми RGB цветами, но и с иными характеристиками цвета. Необходимо уделить внимание подготовке стимульного материала: формировать карты с двумя и более экстремумами различной сложностью, контурных изолиний. В последующих экспериментах следует проверить когнитивную нагрузку испытуемых с помощью специальных тестов, а также учесть их готовность к решению задачи после обучающей процедуры. Дальнейшие эксперименты предполагают выявление эффективности восприятия карт сложного рельефа, что означает не только время на решение задачи, но и ее правильность.

Список литературы

1. Borland D., Taylor II R. M. Rainbow color map (still) considered harmful // IEEE Comput. Graphics Appl. 2007. Vol. 27. No. 2, P. 14–17.

2. Gołębiowska I., Çöltekin A. What’s wrong with the rainbow? An interdisciplinary review of empirical evidence for and against the rainbow color scheme in visualizations // ISPRS Journal of Photogrammetry and Remote Sensing. 2022. № 194. P. 195–208.

3. Ware C., Stone M., Szafir D. A. Rainbow Colormaps Are Not All Bad // IEEE Computer Graphics and Applications. 2023. Vol. 43. Is. 3. P. 88–93.

4. Ермолова Т. К., Ваганова Ю. С., Лаптев В. В. Исследование градаций цвета в визуализации распределения числовых данных // Материалы Международной научно-практической конференции «Графический дизайн: история и тенденции современного развития» / Санкт-Петербургский государственный университет промышленных технологий и дизайна. СПб., 2016. С. 122–129.

5. Imhof E. Die Reliefkarte. Beitraege zur kartographischen Gelaendedarstellung / E. Imhof // Mitteilungen der Ostschweizerischen Geograph. Commerciellen Gesellschaft in St. Gallen. 1924 (1925). S. 59–150, spez. S. 69–91.

6. Bertin J. Semiology of Graphics. Diagrams. Networks. Maps / Jacques Bertin. Redlans: Esri Press, 2011. 438 p.

7. Лаптев В. В. Колористические схемы инфографики / В. В. Лаптев // Вестник Санкт-Петербургского государственного университета технологии и дизайна. Серия 2: Искусствоведение. Филологические науки. 2013. № 4. С. 32–39.

8. Kretschmer I. Naturnahe Farben kontra Farbhypsometrie // Cartographica Helvetica. 2000. No. 21–22. S. 39–47.

9. Peucker K. Der Farbenraum / Karl Peucker // Natur und Kultur. 1920. № 17. S. 481–546.

10. Wilkinson L. The Grammar of Graphics / Leland Wilkinson. Berlin: Springer, 2005. 690 p.

11. Янсон Ю. Э. Теория статистики : лекции проф. Ю. Э. Янсона 1886/87 / Ю. Э. Янсон. СПб.: тип. Шредера, 1891. 561 с.

12. Imhof E. Cartographic relief presentation / E. Imhof. Berlin: New York: Walter de Gruyter, 2007. 388 p.

13. Лаптев М. В. Выявление зон интереса пользователя при визуализации данных с использованием метода ай-трекинга / М. В. Лаптев, В. Э. Янчус, В. В. Лаптев // Программные системы и вычислительные методы. 2023. № 2. С. 49–62.

14. Орлов П. А., Лаптев В. В., Иванов В. М. К вопросу о применении систем ай-трекинга // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2014. № 5 (205). С. 82–92.

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-895-902

Автоматизация в задачах внутрицехового оперативно-календарного планирования производства
Automation in tasks of intra-shop operational scheduling of production

Д.Ю. Васин
Национальный исследовательский Нижегородский государственный университет им. Н.И.Лобачевского, Нижний Новгород, Россия

D.Yu.Vasin
National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia

Аннотация. Цель настоящей работы заключалась в разработке эффективной, импортозамещающей информационной технологии решения комплекса задач внутрицехового оперативно-календарного планирования на предприятии ООО «Язаки Волга» (г. Городец, Нижегородской обл.). Суть производственного процесса на предприятии заключается во взаимодействии всех производственных факторов, направленных на преобразование сырья в готовую продукцию установленного качества и в установленные сроки. Рассмотрены цели, задачи, особенности и проблемы внутрицехового планирования, влияющие на процессы его автоматизации и разработки эффективного импортозамещающего ПО в указанной предметной области. Предложена реализованная в среде «1С–Предприятие» версия системы оперативно–календарного внутрицехового планирования, внедренная на предприятии ООО «Язаки Волга». Приводятся оценки экономической эффективности от внедрения указанной подсистемы.

Ключевые слова: планирование производства, внутрицеховое оперативно–календарное планирование, автоматизация процессов планирования производства, информационные технологии планирования производства, импортозамещение

Аннотация. The purpose of this work was to develop an effective import-substituting information technology for solving a set of problems of intra-shop operational calendar planning at the enterprise OOO Yazaki Volga (Gorodets, Nizhny Novgorod Region). The essence of the production process at the enterprise lies in the interaction of all production factors aimed at converting raw materials into finished products of the established quality and within the established time frame. The goals, objectives, features and problems of intra-shop planning that affect the processes of its automation and the development of effective import-substituting software in the specified subject area are considered. A version of the operational calendar intra-shop planning system implemented in the 1C-Enterprise environment and implemented at the enterprise OOO Yazaki Volga is proposed. Estimates of the economic efficiency of implementing this subsystem are provided.

***Keywords*:** production planning, intra-shop operational scheduling, automation of production planning processes, information technologies for production planning, import substitution.

I. Введение

Современный производственный процесс, удовлетворяющий достаточно высоким требованиям по эффективности, а во многих отраслях производства и условиям импортозамещения, составляет базовую цель оперативного управления производственной деятельностью предприятия. При этом система внутрицехового планирования производства играет одну из решающих организационно-управляющих ролей при выполнении различных работ. Очевидно, что такая система призвана обеспечивать постоянную координацию действий организации по обеспечению многоплановой работы всех ее производственных мощностей, с учетом требования максимизации эффективности использования всех производственных ресурсов, направленной на создание и выпуск заданной номенклатуры изделий в определенном объеме и установленном временном диапазоне.

В результате соразмерной загрузки производственных мощностей предприятия возрастает эффективность использования оборудования и его фондоотдача. Это приводит к снижению различных внутрицеховых производственных затрат при неизбежном росте эффективности использования всех производственных активов.

В современных условиях повышение эффективности производства невозможно без перехода к концепции DDM (data-driven management – управление на основе данных). Это потребует формирования единого цифрового пространства (ЕЦП), что актуализирует в рамках реализации программы импортозамещения создание высокоэффективного безопасного отечественного софта.

Data-driven management – новейшая доктрина управления, в основе которой консолидация данных во все производственные процессы компании, основу которой составляет всеобъемлющая обработка колоссальных объемов разнотипных данных, полученных из различных источников, являющаяся организационным трендом, что позволяет организации перенаправить свои бизнес-процессы на постоянно изменяющиеся запросы рынка, обеспечить переход к работе по требованиям потребителей в режиме потоковой передачи данных (стриминга).

Отметим, что сейчас повышение эффективности производственных процессов компании, описание заказчиков ее продукции и их предпочтений, оперативное планирование, организация сервисных процессов (обслуживание/ремонт оборудования) и т.д. реализуется на основе всестороннего и глубокого анализа данных. Изменение роли данных, которые становятся инструментом решения управленческих задач, составляет базовое отличие DDM от устоявшихся подходов к управлению производством. Очевидно, что система внутрицехового планирования производства является одной из ведущих в реализации концепции DDM.

Однако, начиная с 2022 г. внедрение цифровых технологий в отечественную экономику существенно осложнилось в связи с введением западных ограничительных санкций. Это привело к целому ряду негативных последствий: сокращению числа приобретаемого и используемого импортного ПО, ужесточению лицензионной политики, прекращению сервисной поддержки и обновления уже имеющегося ПО. В итоге, в механизме управления наблюдается отказ от долгосрочного планирования по переводу производства в цифровую форму и запуск цифровых инициатив «по запросу» [2, 3]. Это также актуализирует ускорение решения проблемы импортозамещения, ставшего неотъемлемой частью перехода российских компаний на цифровые технологии.

В 2023 году импортозамещение ПО являлось актуальной проблемой цифровизации бизнеса. По экспертным оценкам порядка 24% компаний активно осуществляют замещение иностранного программного обеспечения (ПО) российским. При этом, около четверти из опрошенных руководителей компаний отметили сложности с адаптацией, вплоть до полной невозможности адаптировать отечественный аналог под специфику конкретного производства, около 16% опрошенных указали на сложности внедрения, связанные с большой несогласованностью (разрозненностью) данных у заказчика, еще порядка 15% отметили достаточно высокую стоимость российской IT-системы для решения задачи [2].

Национальная программа «Цифровая экономика» закрепляет необходимость проведения цифровой трансформации полностью опираясь на российские решения. Целевые показатели в паспорте проекта определяют увеличение в 2024 году доли закупаемого (арендуемого) органами власти отечественного софта до 90%, а для государственных корпораций и компаний с госучастием – до 70% [3].

Очевидно, что главная идея цифровизации – сделать данные полезными, поскольку становится очевидным, что успешные управленческие решения в любой производственной сфере опираются на те или иные данные. Реализация этой идеи диктует необходимость перехода к непрерывной автоматизации процессов с одновременным отказом от проектного подхода.

Главная цель цифровизации производства – на основе непрерывности, импортонезависимости, масштабируемости и широкого внедрения роботов в различные производственные процессы обеспечить создание ЕЦП в компании, причем используемые на предприятиях цифровые платформы способны обеспечить его создание, автоматизируя все шаги, необходимые для выполнения операций.

В [3] приводятся основополагающие принципы построения современной цифровой архитектуры. Отмечена необходимость четкого задания правил принятия архитектурных решений, использование возможностей быстрой адаптации архитектуры к меняющимся требованиям, с соблюдением гарантий на высокий уровень безопасности, обеспечение посредством перекомбинации компонентов легкой адаптации приложений к решению новых задач. В основе ее построения положена компонуемая бизнес-архитектура (PBC – Packaged Business Capabilites), состоящая из микросервисов автономных компонентов, используемых для создания приложения. Доступ к сервисам обеспечивается через любые каналы, основными из которых являются web- и мобильные приложения, различные call-центры, личные кабинеты сотрудников и пр.

К ключевым характеристикам PBC и компонуемой бизнес-архитектуры отнесем модульность, автономность, оркеструемость, обнаруживаемость [4].

В то же время очевидно, что производственный процесс большинства отечественных организаций требует серьезных обновлений, как на уровне его технологическом и управленческом уровнях. Это приводит к внедрению в производственный процесс лучших российских и зарубежных решений по улучшению и оптимизации производства, при этом совершенствование составляющих производственный процесс является первичной составляющей указанной модернизации. Вдобавок отметим, что не существует реального производства, не содержащего тех или иных производственных процессов, требующих определенной доработки, той или иной оптимизации для повышения общей эффективности.

По мнению экспертов, различные проблемы в производственном управлении, являющиеся первопричиной потери компанией потенциальных заказчиков, рынков сбыта готовой продукции, как следствие, снижение прибыли, тесно коррелирует с неэффективным планированием вследствие отсутствия или недостаточной автоматизации производственного планирования. Можно говорить, что автоматизация планирования производства кардинально изменяет всю производственную деятельность [5].

II. Цели и задачи работы

Цель настоящей работы заключалась в разработке эффективной, импортозамещающей информационной технологии решения комплекса задач внутрицехового оперативно-календарного планирования на предприятии ООО «Язаки Волга» (г. Городец, Нижегородской обл.)

Компания ООО «Язаки Волга» была основана в 2012 году. Основной вид деятельности – производство проводки и компонентов электрических систем для автомобилей. Месторасположение в г. Городец, Нижегородской области. Компания входит в состав мирового концерна Yazaki Corporation. На производственных площадях Язаки Волга расположены цеха по производству проводки и компонентов, обеспечивающие производство высококачественных продуктов электрических систем для автомобилей. Компания ежегодно производит значительное количество компонентов для автомобильной промышленности. В рамках деятельности Язаки Волга разрабатывает и производит широкий спектр электропроводки и разъемов для автомобилей различных классов, включая специализированные компоненты для улучшения функциональности и надежности электрических систем.

Для достижения заявленной цели необходимо решить следующие задачи:

– исследовать структуру предприятия ООО «Язаки Волга» также предметную область рассматриваемого комплекса задач;
– определить место выбранного комплекса задач в информационной системе предприятия;
– разработать необходимые математические модели и на их основе оригинальное информационное, алгоритмическое и программное обеспечение;
– провести тестирование и комплексную отладку предлагаемой технологии;
– рассчитать экономическую эффективность предлагаемого внедрения.

Указанная система разрабатывалась для последующего внедрения в отдел производственной логистики и предназначена для ежедневного оперативного анализа изготовления продукции, в зависимости от заказов покупателей, даты отгрузки готовых товаров и наличия или поступления материалов, необходимых для изготовления. Ее внедрение позволит обеспечить своевременное, ритмичное и экономичное движение производственных ресурсов внутри основного производства.

III. Производственный процесс и производственное планирование

Определим производственный процесс любого предприятия как совокупность операций, последовательное выполнение которых превращают сырьевые, материальные ресурсы и различные комплектующие в выходную продукцию. Любая промышленная производственная деятельность базируется на тех или иных производственных процессах.

В основу производственного процесса положен технологический процесс, представляющий собой фрагмент производственного процесса, включающий в себя упорядоченный набор действий, выполнение которых приведет к определению/изменению состояния предмета труда.

Технологическая операция – выделенная, законечная часть технологического процесса, реализуемая на одном рабочем месте одним или несколькими рабочими.

Кроме трудовых и автоматических процессов в производственный процесс включают еще естественные процессы, которые в большинстве случаев не трудозатратны. Современные предприятия, ориентированные на выпуск сложной, высокотехнологичной продукции, отличаются большой вариабельностью реализуемых производственных процессов. Поскольку не существует двух абсолютно идентичных производств, то уникальны и соответствующие производственные процессы.

По назначению производственные процессы разделяются на основные, вспомогательные и обслуживающие (рис. 1). Для выполнения основных процессов необходимо привлечение трудовых ресурсов, предметов и средств труда [6].

Суть основных производственных процессов заключается в прямом изменении формы или состояния исходного материала будущего изделия, а совокупность этих процессов образует основное производство.

Основное назначение вспомогательных производственные процессы – обеспечить непрерывное протекание основных процессов. В результате реализации вспомогательных процессов появляется продукция, используемая в основном производстве предприятия.

В задачу обслуживающих производственных процессов входит сервисное обеспечение основных и вспомогательных процессов необходимыми для их нормального функционирования услугами.

Не существует основных, вспомогательных и обслуживающих производственных процессов, не обладающих той или иной спецификой и разными тенденциями развития. Структура производственного процесса складывается из их состава.

Типичный производственный процесс состоит из значительного числа взаимозависимых процессов, операций и иных элементов, отличающихся для разных производств и имеющих определенные ограничения на применяемые подходы к производственному планированию. В зависимости от выпускаемой продукции производственные процессы будут существенно отличаться.

Сейчас эффективное производственное планирование является наиважнейшим условием обеспечения успешности, конкурентоспособности и прибыльности компании в долгосрочной перспективе [6].

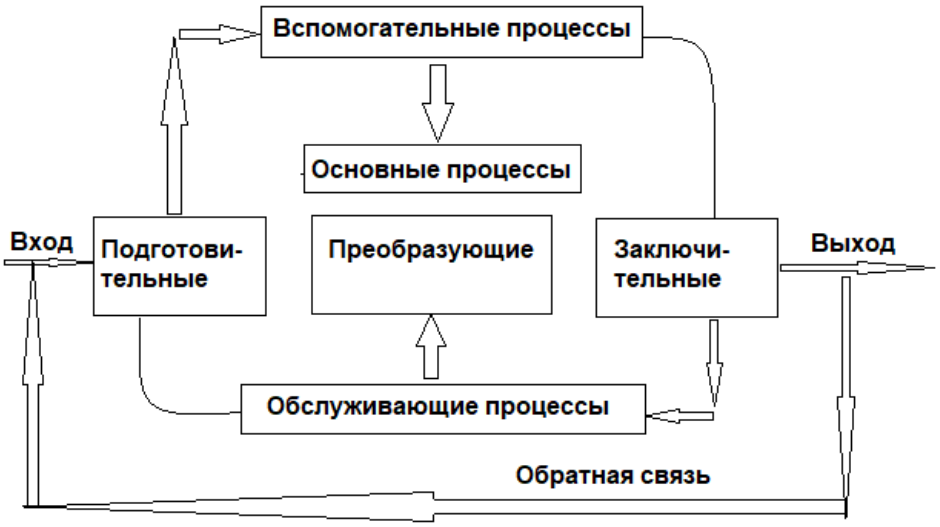


Рис. 1 Производственные процессы и их взаимосвязь

Выделим следующие основные виды планирования по: срокам достижения результата, обязательности, степени охвата, изменению базовой информации, объектам планирования. В плане оперативно-календарного планирования важным критерием является срок достижения результата. По этому критерию выделяют: оперативное (текущее) планирование, краткосрочное планирование, среднесрочное планирование и долгосрочное планирование:

- долгосрочное планирование – обычно охватывает период 5 лет и более, и устанавливает единую стратегию выхода компании на рынок. В основе этого вида планирования лежит расширение производства и снижение затрат. Для достижения этой цели необходимо менять номенклатуру выпускаемой продукции и корректировать функциональную сферу деятельности. В результате формулируются долгосрочные цели, на их основе принимается долгосрочная производственная политика с формированием планов на долгосрочные проекты;
- среднесрочное планирование – охватываемый временной интервал от 1 до 5 лет, принимаются во внимание возможности цехов и иных подразделений, основанные на их индивидуальной оценке. Как правило к этой группе относят план предприятия по маркетингу, план производства, план по труду и финансовый план;
- краткосрочное планирование – охватываемый временной интервал до 1 года, а текущее (оперативное) планирование как правило составляется ежедневно. Обычно включает план по труду и заработной плате, материально-техническому обеспечению, рентабельности, объему производства, себестоимости, прибыли и т.д. Представляет собой детальную разработку планов предприятия и его подразделений путем соединения планирования и диспетчеризации. При этом детальные планы и расписания составляются на временные интервалы обычно не превышающие месяца, а функции диспетчеризации обеспечивают контроль за производством, исправностью оборудования и при необходимости устранение текущих неполадок, организацию бесперебойных поставок на рабочие места необходимых сырьевых и материальных ресурсов, организацию контроля качества и экспорта готовой продукции, утилизацию отходов.

Внутри оперативного планирования выделяют внутрицеховое и межцеховое планирования. Внутрицеховое планирование ответственно за распределение работ на конкретных рабочих местах для обеспечения качественного производства и ритмичной работы компании в целом и ее отдельных подразделений вплоть до отдельных сотрудников, а межцеховое обеспечивает межцеховую координацию деятельности с учетом вклада в производственный процесс основных, вспомогательных и сервисных цехов.

В процессе оперативно-календарного планирования формируется согласованная, синхронная деятельность взаимодействующих цехов для устойчивой работы всего предприятия в целом. Оно подразделяется на два вида – внутреннее и межцеховое оперативное планирование. Основная задача внутреннего планирования – организация реализации поставленных перед цехом производственных целей, а также создание графиков работы на короткие промежутки времени, не превышающие декаду. Внутреннее планирование обеспечивает решение задач оперативного регулирования, контроля и учета выполнения планов.

Межцеховое оперативное планирование согласует функционирование всех подразделений для решения производственных задач на уровне завода. В процессе планирования используют сквозные расписания и календарные нормативы производственных процессов, что позволяет проверять исполнение цехами заводского расписания.

Основные задачи оперативно-календарного планирования весьма обширны и разнообразны и могут быть отнесены к организационным, диспетчерским, контролирующим. В результате создаются условия для внедрения инновационных форм организации работы, перехода к цифровым технологиям управления, автоматизации на основе использования современной вычислительной техники.

На вход системы оперативно-календарного планирования подаются разнообразные данные, среди которых оперативные планы выпуска продукции и ремонта оборудования, различные технологические данные, режимы работы предприятия и его отдельных участков и пр.

В механизме управления важнейшую роль играет система оперативного управления основным производством (ОУОП) – является важнейшей функциональной подсистемой любой АСУ, представляет собой набор информационных процедур, обеспечивающих синтез управляющих воздействий на производственный процесс, с целью достижения им плановых показателей в установленные сроки. Ее основными функциями являются планирование, оперативный анализ и учет. ОУОП – сложная организационно-плановая система, состоящая из организационной, функциональной и элементной подсистем, описывающих производство с разных точек зрения: организационная – с позиций построения системы управления, функциональная – характеризует функциональный набор, выполняемый системой управления, а элементарная – описывает основные составные элементы.

Внутрицеховое планирование является краткосрочным (разрабатывается на месяц, декаду, неделю, сутки), формирующимся на основе календарного плана, разрабатываемого на межцеховом уровне и завершающим звеном производственного планирования, обеспечивает регламентирование работ между исполнителями, доводит плановые задачи до каждого производственного участка и конкретного исполнителя, делая упор на точности и детализации. Различные уклонения от данного планы влекут за собой нарушение сроков выпуска продукции и невыполнение плана. Цель такого планирования заключается в соблюдении сроков выполнения заказов цехом в целом и его участками и обеспечении выполнения планов верхнего уровня. Объектами планирования выступают различные узлы и детали, их перемещения внутри цеха. Невыполнение указанных планов влечет за собой нарушение обязательств по поставкам продукции, ухудшение финансового состояния предприятия, снижение его рентабельности, конкурентного статуса.

Указанный вид планирования предполагает высокую точность, детализацию и динамичность поскольку производственная среда характеризуется изменчивостью и наличием множества ограничений [5]. В условиях реального производства календарный план представляет собой документ, подверженный оперативным изменениям и дополнениям. Поэтому разумна его сегментация на части и дальнейшее внесение изменений осуществляется уже в пределах выделенных частей. Таким образом указанные планы формируются/уточняются, персонализируются, при необходимости корректируются, детализируются для всех участников производства по каждой технологической операции, узлу, детали. Обеспечивается разработка разветвленной системы сроков движения продукции по цеху.

Для учета всех технологических и иных особенностей производства кроме календарного плана цеха исходными данными для внутрицехового планирования и выполнения различных расчетов являются: маршрутные карты, календарно-плановые нормативы, данные о производительности рабочих мест, текущей загрузке оборудования, а также график ремонта станков и оборудования.

Требование использования для планирования актуальной информации обо всех аспектах производственного процесса (реальные производственные возможности, регламентные и внеплановые остановы/ремонт оборудования, наличие необходимого сырья и материалов, квалификации рабочих и т.д.) влечет необходимость анализа и последующей оперативной работы с большими массивами исходных данных, что делает ручное формирование планов очень трудоемким, носящим приблизительный характер и требующим немедленной автоматизации.

Непосредственное влияние системы внутрицехового планирования в первую очередь сказывается на различных первичных показателях работы подразделений, изменяя себестоимость продукции и иные финансовые показатели.

Целью системы внутрицехового планирования является обеспечение выполнения планов вышестоящего уровня и сроков выполнения заказов цехом в целом и его участками, обеспечивая при этом устойчивое производство определенной продукции в установленные сроки и с наибольшей эффективностью.

На систему внутрицехового календарного планирования возложено решение достаточно обширного класса разнообразных задач планирования, формирования графиков работы и учета их выполнения, оперативного снабжения участков и рабочих мест необходимыми ресурсами, различные согласования и координацию производственного процесса для исключения отклонений от графика и пр.

Внутрицеховое планирование должно обеспечивать решение двух противоречащих друг другу задач: гибко реагировать на различные изменения условий внешней среды и внутренние события, но при этом адаптация под изменяющиеся условия не приводит к сбоям в существующих процессах и не сказывается негативно на ключевых показателях эффективности. Поэтому она по праву относится к классу сложных.

Любой производственный процесс должен учитывать возможности появления различных непредсказуемых форс–мажорных событий, напрямую влияющих на производственный процесс, среди прочего выделим простой оборудования из–за поломок, непреднамеренный выпуск некачественной и/или бракованной продукции,

отсутствие отдельных работников по болезни и пр. Кроме того, процесс календарного планирования должен учитывать технологические особенности конкретного производства, возможное разнообразие выходного ассортимента выпускаемой продукции, иные факторы, усложняющие процесс календарного планирования и приводящие к необходимости оперативного внесения изменений в первоначальные планы в условиях строгих временных ограничений. В противном случае исходные внутрицеховые планы, графики и сменно-суточные задания могут оказаться сорванными.

Появление внутрицехового оперативно-календарного планирования связано с тем, что на любой стадии производства экономика обязана определять характер времени и также затрат, то есть должна извлекать точный и целый производственный план с учетом работы, ее длительности, а также план требуемых ресурсов.

IV. Практическая реализация подсистемы внутрицехового планирования

В ООО «Язаки Волга» формирование выпуск продукции, определение материальных затрат на производство, контроль фактического выпуска готовой продукции можно представить в виде схемы информационной взаимосвязи задач. Взаимосвязь идет последовательно, на каждом шаге происходит определенное действие и формируется соответствующая документация (рис.2).

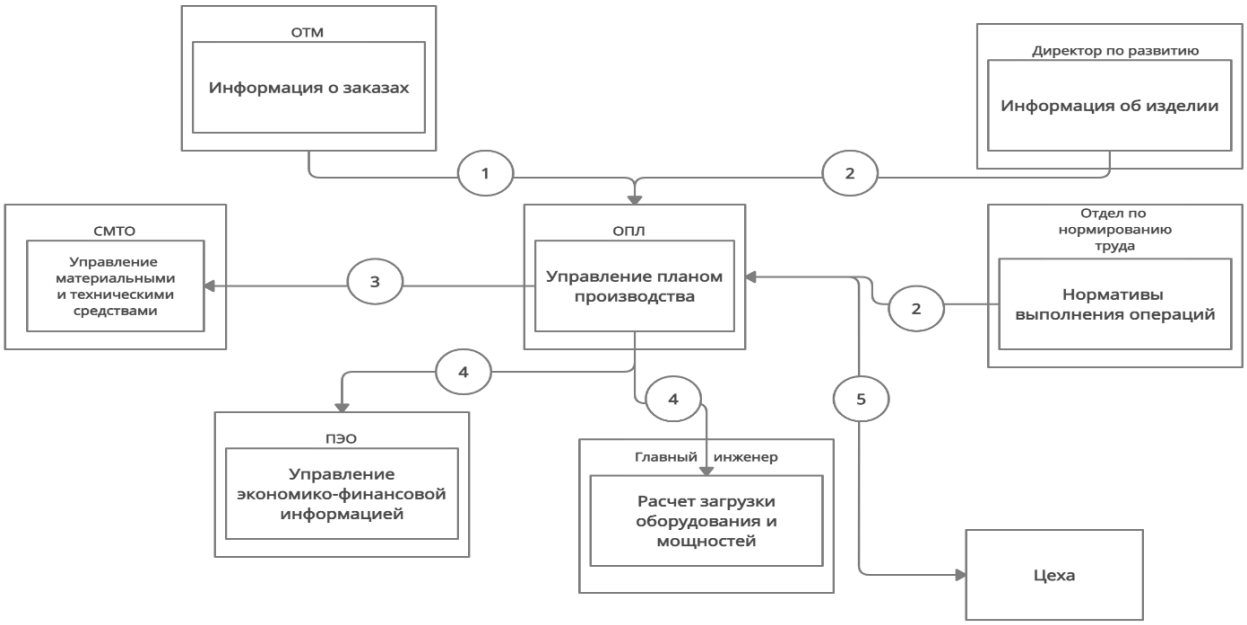


Рис. 2 Взаимодействие задач

Список сокращений на рис.2: ОТМ – отдел торговли и маркетинга; СМТО – служба материально-технического обеспечения; ОПЛ – отдел производственной логистики; ПЭО – планово-экономический отдел; Директор по развитию; Отдел по нормированию труда; Главный инженер; Цеха – производственные подразделения предприятия; 1 – информация о номенклатуре и количестве заказов; 2 – информация о составе изделия, ТМО; 3 – информация о потребности в ТМЦ; 4 – план выпуска готовой продукции; 5 – сформированный маршрутный лист выпуска продукции.

Анализ распространенных сред разработки, с учетом сложившейся обстановки с использованием в РФ импортного ПО, а также очевидная необходимость интеграции создаваемого ПО в среду уже используемого в ООО «Язаки–Волга», определило выбор базовой платформы разработки. При реализации задачи «Внутрицеховое оперативно-календарное планирование» будет использоваться среда «1С: Предприятие 8.3» [7].

Входная информация представлена следующими документами: заказ на производство; поступление материалов; отчет производства за смену, а также представлена в справочниках: контрагенты; должности; разряды; сотрудники; бригады; продукция; единицы измерения; материалы; участки; маршруты; рабочие центры.

Примеры описания в системе «1С: Предприятие 8.3» реквизитов документа «Поступление материалов» приведен в табл.1, реквизитов справочника «Продукция» – в табл.2, реквизитов документа «План фактный анализа выпуска» – в табл.3

В системе определены следующие классификаторы: идентификационный номер заказа; идентификационный номер материала; идентификационный номер товара; идентификационный номер должности; идентификационный номер бригады; идентификационный номер контрагента; табельный номер сотрудника.

Пример оформления в системе «1С: Предприятие 8.3» классификатора «Идентификационный номер заказа»:
– метод классификации – элементарный;
– метод кодирования – порядковый;
– длина кода – 2 знака.

На основании переменной и нормативно-справочной информации в конце отчетного периода формируется выходная информация.

Выходную информацию представляют в экранных формах и на бумажных носителях: план фактный анализ выпуска; оборотно-сальдовая ведомость материалов; остаточная стоимость материалов; маршрутный лист.

Периодичность и сроки выполнения: план фактного анализа выпуска – ежедневно, ведомости по материалам – по запросу, остаточной стоимости оборудования – по запросу. Пользователи: СМТО, ПЭО, цеха.

ТАБЛИЦА 1
ОПИСАНИЕ РЕКВИЗИТОВ ДОКУМЕНТА «ПОСТУПЛЕНИЕ МАТЕРИАЛОВ»

№ п/п	Наименование	Идентификатор	Формат
1	Номер	Номер	N(9)
2	Дата	Дата	D(10)
3	Контрагент	Контрагент	C(25)
4	Материал	МатериалыМатериал	C(25)
5	Единица измерения	МатериалыЕдиницаИзмерения	C(25)
6	Количество	МатериалыКоличество	N(3)
7	Цена	МатериалыЦена	N(10)
8	Сумма	МатериалыСумма	N(10)

ТАБЛИЦА 2
ОПИСАНИЕ РЕКВИЗИТОВ СПРАВОЧНИКА «ПРОДУКЦИЯ»

Наименование	Идентификатор	Формат
Код	Код	N(4)
Наименование	Наименование	C(25)
Вид продукции	ВидДетали	N(1)
Материал	МатериалыМатериал	C(25)
Единица измерения	МатериалыЕдиницыИзмерения	C(10)
Количество	МатериалыКоличество	N(10)
Маршрут	МаршрутМаршрут	C(10)

ТАБЛИЦА 3
ОПИСАНИЕ РЕКВИЗИТОВ ДОКУМЕНТА «ПЛАН ФАКТНЫЙ АНАЛИЗ ВЫПУСКА»

№ п/п	Наименование	Идентификатор	Формат
1	Идентификатор заказа	Номер	N(2)
2	Дата заказа	Дата	D(10)
3	Продукция	Продукция	C(25)
4	ДатаОтгрузки	ДатаОтгрузки	D(10)
5	Не выполненный план на начало	НеВыполненныйПланНаНачало	N(3)
6	Перевыполненный план на начало	ПеревыполненныйПланНаНачало	N(3)
7	Запланировано за период	ЗапланированоЗаПериод	N(3)
8	Выполнено за период	ВыполненноЗаПериод	N(3)
9	Не выполненный план на конец	НеВыполненныйПланНаКонец	N(3)
10	Перевыполненный план на конец	ПеревыполненныйПланНаКонец	N(3)

V. Экономическая эффективность проекта

Внедрение различных АИС в производственных процесс определяет увеличение производства продукции за счет повышения эффективности процессов управления. Внедрение механизмов оперативного управления

производством невозможно без весомых финансовых затрат. При этом затраты по предоставлению различной справочной информации о пользователях и вычислительных услугах должны быть возмещены как можно скорее.

Для формирования показателей экономической эффективности используем стандартную методику, включающую расчеты суммы годовой экономии, коэффициента экономической эффективности капитальных вложений и срока окупаемости капитальных вложений.

Проведенные расчеты показали, что годовой эффект от внедрения разработанной подсистемы составит порядка 250 тыс. рублей.

VI. Выводы

В результате реализации проекта были решены следующие задачи:

- изучена деятельность предприятия ООО «Язаки Волга» (г. Городец, Нижегородской обл.), его организационная структура, состав технических средств, реализующих задачу автоматизации отдельных производственных составляющих;
- изучена предметная область, принципы и основные задачи внутрицехового оперативно календарного планирования;
- в результате анализа ПО, предлагаемого на российском рынке ПО, выявлены базовые преимущества и недостатки существующих программных решений в указанной предметной области, определена актуальность разработки собственной уникальной информационной системы;
- осуществлен выбор платформа разработки АИС внутрицехового оперативно-календарного планирования
- определён функционал АИС внутрицехового оперативно-календарного планирования и сформирован список требований к ней;
- разработано информационное, алгоритмическое и программное обеспечение АИС внутрицехового оперативно-календарного планирования, описаны входные и выходные документы, сформированы технологические и инструкционные карты, предоставлена сопровождающая документация. Система содержит информацию о производстве жгутов в сборочном цехе ООО «Язаки Волга», о поступлении и затратах материалов, необходимых для производства, позволяет осуществлять анализ фактического и планового выпуска продукции, предоставляет возможность производить расчёт стоимости имеющихся в наличии и требующихся материалов, с целью последующего планирования закупок, и формировать необходимые настраиваемые отчёты со всей статистической информацией;
- осуществлена комплексная отладка и внедрение разработанной АИС в производственный процесс в сборочном цехе ООО «Язаки Волга»;

Список литературы

1. https://base.garant.ru/72296050/(Дата обращения: 13.07.24)
2. https://q.diasoft.ru/mediacenter/publications/tsifrovaya-transformatsiya-v-usloviyakh-importozameshcheniya-kak-razrabatyvat-po-bystro-kachestvenno/ (Дата обращения: 13.07.24)
3. https://research.naumen.ru/archive/tsifrovaya-transformatsiya-v-rossii-2023/ (Дата обращения: 13.07.24)
4. Экономика и управление на предприятии : учебник / А. П. Агарков, Р. С. Голов, В. Ю. Теплышев, Е. А. Ерохина ; под ред. А. П. Агаркова, Р. С. Голова. - 4-е изд., стер. - Москва : Дашков и К, 2022. - 398 с. - ISBN 978-5-394-04340-6. - Текст : электронный. - URL: https://znanium.com/catalog/product/2085953 (дата обращения: 13.07.2024). – Режим доступа: по подписке.
5. https://adeptik.com/blog/avtomatizaciya-proizvodstvennogo-planirovaniya (Дата обращения 13.07.24)
6. https://book2.diasoft.ru/part4/#rec514467031 (Дата обращения: 13.07.24)
7. https://gendalf.ru/news/review/obzor-vozmozhnostey-platformy-1spredpr (Дата обращения: 13.07.24)

УДК 004.94
DOI: 10.25206/978-5-8149-3873-2-2024-903-910

Исследование методов прикладного программирования Renga API для конечных пользователей информационных моделей
Research of applied programming methods Renga API for end users of information models

И.И. Серов 1, Т.Н. Томчинская 1, А.Д. Филинских 1
1 Нижегородский государственный технический университет им. Р.Е. АлексееваНижний Новгород, Россия

I.I. Serov 1, T.N. Tomchinskaya 1, A.D. Filinskikh 1
[1]Nizhny Novgorod State Technical University, n.a. R.E. AlekseevNizhny Novgorod, Russia

Аннотация. В статье описываются возможности взаимодействия с Renga API, используя различные языки программирования, такие как C++ и C#, а также динамически типизируемые языки(Python, VBScript и т.д). Описаны особенности работы в программной среде Power Shell с целью разработки плагинов и скриптов для работы с информационной моделью, разработанной в программном продукте Renga. Описаны методы, служащие для запуска самого программного продукта Renga, создания, изменения и сохранения под новым именем проекта Renga с информационной моделью. В статье также описываются варианты разработки скрипта, служащего для подсчета объектов информационной модели и получения их свойств.

***Ключевые слова*:** Renga, Renga API, интерфейс прикладного программирования, плагин, Информационное моделирование зданий, свойства объекта.

Аннотация. The article describes the possibilities of interaction with Renga API using various programming languages, such as C++ and C#, as well as dynamically typed languages (Python, VBScript, etc.). The article describes the features of working in the Power Shell software environment for the purpose of developing plug-ins and scripts for working with the information model developed in the Renga software product. The methods used to launch the Renga software product itself, create, change, and save a Renga project with an information model under a new name are described. The article also describes options for developing a script used to count objects of the information model and obtain their properties..

***Ключевые слова*:** Renga, Renga API, Application Programming Interface, Plugin, Building Information Modeling, Object Properties.

I. Введение и постановка задачи

Проектирование, строительство и эксплуатация строительного объекта – это процесс, который затрагивает множество специалистов, каждому из которых необходима специализированная документация. Формирование и агрегация проектной, строительной и эксплуатационной документации возможно по отдельности для каждой отрасли, но этот подход плодит разрозненность и путаницу при ее хранение и передаче. Проблему большого количества разрабатываемой по отдельность документации решает подход информационное моделирование строительных объектов (англ. Building Information Modeling). В результате такого подхода к проектированию и дальнейшему строительству и эксплуатации строительного объекта получается не только документация и трехмерные модели, но и цела общая структурированная база данных, нацеленная на повышение эффективности взаимодействия между специалистами, а также на уменьшение количества коллизий [1-7].

Но отсюда вытекает тот факт, что современные программные продукты, реализующие подход информационного моделирования, имеют высокую функциональную нагруженность и оттого требуют от пользователя специальных навыков. Такие навыки зачастую необходимо получать в специализированных образовательных учреждениях.

Упрощение процесса взаимодействия со сложным и объёмным в функциональном плане программным продуктом можно добиться разработкой внешних или внутренних надстроек и плагинов (расширений). Разработка таких расширений доступна при использование интерфейса прикладного программирования, того программного продукта, а которое и планируется установить разрабатываемое дополнение. На сегодняшний день разработчики многих систем, которые заточены на проектирование и разработку информационных моделей, добавляют доступ к интерфейсу прикладного программирования своего продукта.

Интерфейс прикладного программирования (англ. Application Programming Interface) – это набор протоколов, методов, команд, которые используются для взаимодействия между программными службами.

API-сервисы (Application Programming Interface сервисы) служат для доступа к данным одного программного продукта или программной службы из другого, при этом нет необходимости в понимание того, как именно работает первый программный продукт (рисунок 1) [8].

Существует несколько вариантов взаимодействия с API:

- Используя данные, передаваемые интерфейсу, с помощью которых он сможет выполнять методы.
- Используя процесс, который может выполнять приложение с помощью этого интерфейса.
- Используя данные, получаемые приложением после взаимодействия и работы с API.

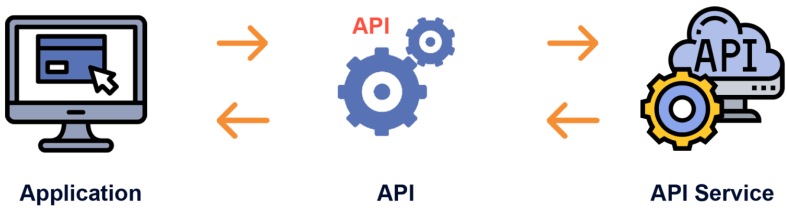


Рис. 1. Схема взаимодействия с API-сервисами

Интерфейс прикладного программирования может быть использован как фундамент для создания нового программного продукта, который будет упрощать, а в некоторых случаях дублировать функционал родительского сервиса, но этот функционал будет адаптирован под конкретные задачи. Так, например, для работы с информационной моделью, разработанной в программном продукте Renga от компании «Аскон», необходимы специфические знания, которых может не быть у конечного пользователя на этапе эксплуатации. Разработчики Renga добавили в свое программное обеспечение возможность взаимодействия с API. Взаимодействуя с Renga API, появляется возможность разработки адаптированного под конкретные задачи или под конкретную группу пользователей инструмента, который в понятном и в каком-то смыслу упрощённом виде позволяет взаимодействовать с информационной моделью и получать из неё необходимые данные[1-4, 9,10].

II. Теория

Разработчики программного продукта Renga непрерывно занимаются улучшением и усовершенствованием функционала своей BIM-системы (Building Information Modeling. Обновление функционала не может не затронуть и такой элемент Renga, как Application Programming Interface. На сегодняшний день, доработка этого функционала еще не закончена, но уже имеет большие возможности в автоматизации процесса проектирования и эксплуатации строительного объекта.

Функциональные возможности Renga API можно разделить на две группы в зависимости от используемого языка программирования и среды разработки. Использование C++ и C# дает возможность загружать плагины в процесс работы Renga, при этом при использовании динамически типизируемых языков, таких как Python, VBScript и т.д, загрузка плагина производится, как загрузка отдельного приложения, которое извне может управлять функционалом API Renga.

Плагин, разработанный на языках программирования C++ и C# имеет возможность расширения пользовательского интерфейса путем добавления новых иконок и кнопок в контекстное меню и панель инструментов, также имеет возможность экспорта геометрии объектов и экспорта файлов в такие форматы данных, как IFC, DWG/DXF.

Разработка же внешнего плагина не дает возможности изменения пользовательского интерфейса, но это не является критическим показателем эффективности данного подхода, так как могут быть сценарии работы, при которых нет необходимости работы в самом программном продукте Renga, а вся работа производится во внешней системе, которая имеет доступ к API [10].

III. Результаты экспериментов

1. Запуск проектов и дальнейшие сохранение

15. Выбор среды выполнения скриптов

Интерфейса прикладного программирования (англ. Application Programming Interface – API) Renga — это программный интерфейс к системе Renga, созданный с использованием технологии COM (Component Object Model). Этот стандарт возможно использовать во многих программных средах и применять разные языки

программирования, такие как C++, C#, динамически типизированные языки, такие как Python, VBScript и т.д. К тому же возможно использование программной среды Power Shell.

Оболочка Power Shell — это, во-первых, межплатформенное средство, используемое для автоматизации различных задач, которое является оболочкой над стандартной комодной строкой, включающей скриптовый язык и платформу управления конфигурацией. Во-вторых, компания Microsoft, которая является разработчиком этой оболочки, предоставляет всеобъемлющий набор документации, которая поможет разобраться во всех тонкостях работы [11,12].

16. Автоматический запуск Renga

Команды для запуска Renga представлены на рисунке 2. Далее подробно разберем назначение каждого из блоков плагина.

Для начала необходимо создать COM-объект Renga. Для этого создается новая переменная, в PowerShell для этого зарезервирован специальный символ – $. После объявления переменной необходимо присвоить ей значение, отвечающее за создание COM-объекта, для этого необходимо использовать команду *new-object -com*. Так как нам нужен доступ к API Renga, необходимо присвоить особое имя класса – *Renga.Application.1*. Важной особенности объявления переменной типа COM-объекта Renga является то, что необходимо добавление параметра *-strict*, который дает Power Shell сведения о том, что у вызываемого объекта есть своя оболочка времени выполнения. Отсутствие данного параметра может привести к появлению ошибок.

Далее необходимо использовать конструкцию *try- finally*, которая в похожем виде встречается практически во языках программирования. Данная конструкция поможет гарантировать закрытие Renga в том случае, если произойдет какая-то непредвиденная ошибка, так как в блоке *finally* вызывается метод *Quit()*. При отсутствии блока, обрабатывающего ошибки при работе, может возникнуть такая ситуация, когда экземпляр Renga останется открытым до момента перезагрузки устройства, что при многочисленных итерациях выполнения кода может наплодить десятки запущенных экземпляров Renga, что в свою очередь ведет к перегрузке системы.

Внутри блока *try-* необходимо написать код, который сделает Renga «видимой», для этого применяется особое свойство Renga API – $renga_app.Visible = $true. Далее выведем текстовое сообщение в процесс выполнения кода, для этого в оболочке PowerShell необходимо использовать следующий оператор: *«Run Renga»*.Для того, чтоб это сообщение было замято человеческому глазу, необходимо приостановить выполнение программы истпользуя оператор *pause;* [10,11].

```
1   $renga_app = new-object -com Renga.Application.1 -strict
2
3   try
4   {
5           $renga_app.Visible = $true;
6           "Run Renga";
7           pause;
8   }
9   finally
10  {$renga_app.Quit()}
```

Рис. 2. Листинг скрипта запуска Renga

Для выполнения написанного кода в среде PowerShell, фал должен иметь расширение ps1, это системное расгирение, которое уже связано с оболочкой выполнения данных, пример выполнения кода представлен на рисунке 3.

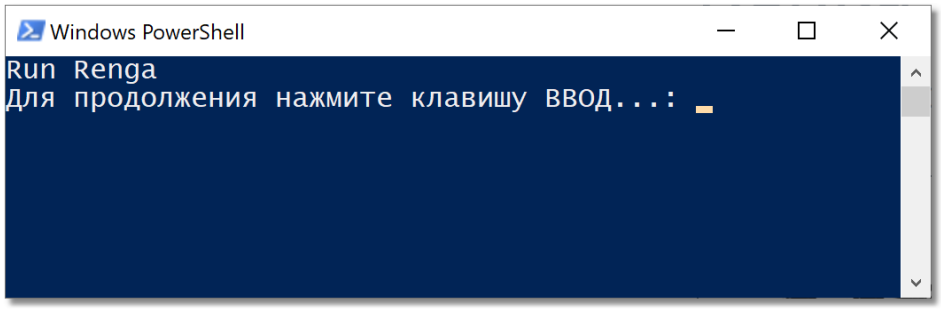


Рис. 3. Окно выполнения скрипта

Определение названия COM-объекта Renga, а также методы и свойства этого объекта отображены в документации самой регна, которую можно получить на официальном сайте, либо в архиве с комплектом разработки программного обеспечения Renga (англ. Renga Software Development Kit – Renga SDK).

Плагин запуск Renga может быть совмещен с созданием нового проекта а основе стандартного шаблона, для этого используется метод *.CreateProject()*, пример кода, в который добавлено создание нового проекта, представлен на рисунке 4 [10,11].

```
1  $renga_app = new-object -com Renga.Application.1 -strict
2
3  try
4  {
5          $renga_app.Visible = $true;
6          $renga_app.CreateProject();
7          "Run Renga and Create project";
8          pause;
9  }
10 finally
11 {$renga_app.Quit()}
```

Рис. 4. Листинг скрипта запуска Renga с созданием проекта

После запуска скрипта запустилась Renga и был создан новый проект на основе стандартного шаблона.

17. Сохранение проекта с помощью скрипта под другим именем

Открытие программного продукта и создание нового проекта является важной частью процесса автоматизации, так как является одним из основных этапов, но зачастую работа производится с уже готовыми проектами, на рисунке 5 представлен код, который открывает проект и сохраняет его под новым именем.

Для открытия и сохранения проекта, необходимо вести новые переменные – *$FN* и *$FNN*, которые хранят в себе путь до файла проекта, который необходимо открыть, и путь с именем файле, куда его нужно сохранить.

Для исключения непредвиденных ошибок и исключений, все операции необходимо производить внутри конструкции *try – finally*. Открытие существующего проекта производится при помощи метода *OpenProject($FN)*, который вызывается объектом *$renga_app* и в который в качестве аргумента передается переменная, содержащая путь до файла проекта. Также можно обработать случаи, в которых будут ошибки при открытии проекта, для этого используется конструкция *if-else*.

Строчкой кода *$project = $renga_app.Project;* создаётся переменная, которая имеет значение как проекта Renga, с помощью которого дальнейшая работа будет производится непосредственно с проектом.

Используя метод *SaveAs($FNN,0,$true)*, объекта *$project* появляется возможность сохранения текущего проекта под новым именем и по новому пути, для этого в качестве аргументов передаем новый путь, параметр отвечающий за сохранения проекта как новый или как шаблон текущего, а также параметр, который указывает методу на то, что нужно перезаписать файл, найденный по переданному пути.

Закрытие проекта производится при помощи метода *CloseProject($true)*, который вызывается объектом Renga, параметром которого является логическое значение, которое отвечает за игнорирование выполненных изменений.

Стоит отметить тот факт, что в данном случае рабочее окно Renga остается невидимым для пользователя, так как никакие изменения вручную выполнять нет необходимости [10,11].

Результат работы программы представлен на рисунке 6.

```
$renga_app = new-object -com Renga.Application.1 -strict
$FN = "D:\test.rnp";
$FNN = "D:\test_new.rnp";
try
{
    $result = $renga_app.OpenProject($FN)
    if ($result -eq 0) #Проверяем открылся ли проект, если 0 - открылся, иначе ошибка
    {
        $project = $renga_app.Project;
        $result = $project.SaveAs($FNN,0,$true);
        if ($result -eq 0) #Проверяем сохранился ли проект, если 0 - сохранился, иначе ошибка
        {
            "Project saved"
        }
        else
        {
            "Error saving project as"
        }
        $result =$renga_app.CloseProject($true);
        if ($result -eq 0) #Проверяем закрылся ли проект без ошибок, если 0 - закрылся, иначе ошибка
        {
            "Project closed"
        }
        else
        {
            "Error closing project"
        }
        pause;
    }
    else
    {
        "Error opening project"
    }
}
finally
{$renga_app.Quit()}
```

Рис. 5. Сохранение проекта с новым именем

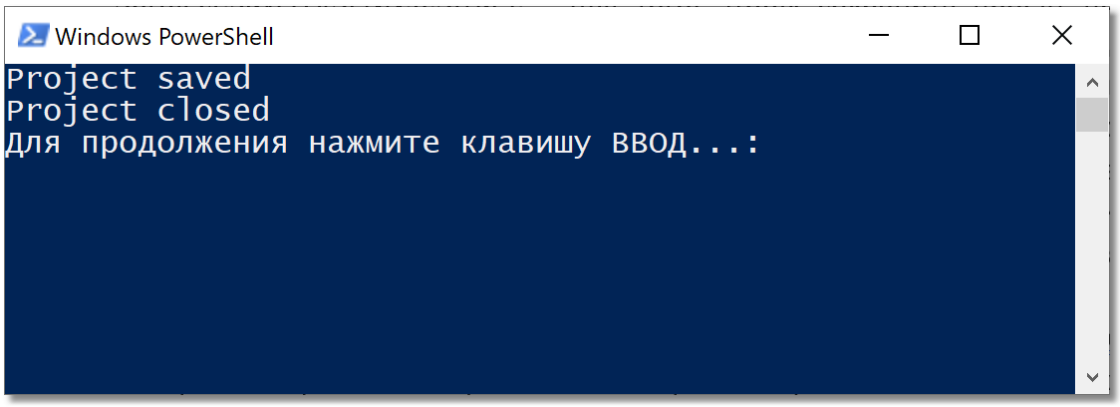


Рис. 6. Окно выполнения скрипта сохранения

2. Подсчет объектов

Подсчет объектов производится при помощи обращения как к методам проекта, так и модели, для загрузки модели, которая содержится в проекте используется параметр *model*, который вызывается из объекта *project*: *$Model3D = $project.model*. Для получения списка объектов в модели используется метод *GetObjects()*.

Для подсчета количества объектов, создаются переменные для каждого из объектов и обнуляются:

$CntWalls = 0;
$CntColumns = 0;
$CntRooms = 0;
$CntDoors = 0;
$CntWindows = 0;
$CntBeams = 0;
$CntFloors = 0; $
CntRamps = 0;
$CntStairs = 0;
CntWallFoundation =0;
$CntIsolatedFoundation =0;

Определение типа объектов производится при использование уникальных идентификаторов объектов Renga, которые перечислены в документации API Renga, которая хранится внутри SDK Renga [10,11].

Перебор объектов производится при помощи цикла for:

for ($i = 0; $i -lt $ModelObjs.Count(); $i++)

В теле цикла перебираем элемнты и получаем следующий при помощи строчки кода: *$object = $ModelObjs.GetByIndex($i);*

Для определения типа объекта производится сравнение свойств объекта *ObjectTypeS* с идентификатором искомого объекта:

if ($object.ObjectTypeS -eq "{4329112A-6B65-48D9-9DA8-ABF1F8F36327}")

После подсчета всех искомых объектов и суммирования их количества выводим данные на экран пользователя, используя следующие конструкции:

Write-Host "Стен: " $CntWalls;
Write-Host "Колонн: " $CntColumns;
Write-Host "Балок: " $CntBeams;
Write-Host "Перекрытий: " $CntFloors;
Write-Host "Помещений: " $CntRooms;
Write-Host "Дверей: " $CntDoors;
Write-Host "Окон: " $CntWindows;
Write-Host "Лестниц: " $CntStairs;
Write-Host "Пандусов: " $CntRamps;
Write-Host "Ленточных фундаментов: " $CntWallFoundation;
Write-Host "Столбчатых фундаментов: " $CntIsolatedFoundation;

Результат работы плагина представлен на рисунке 7.

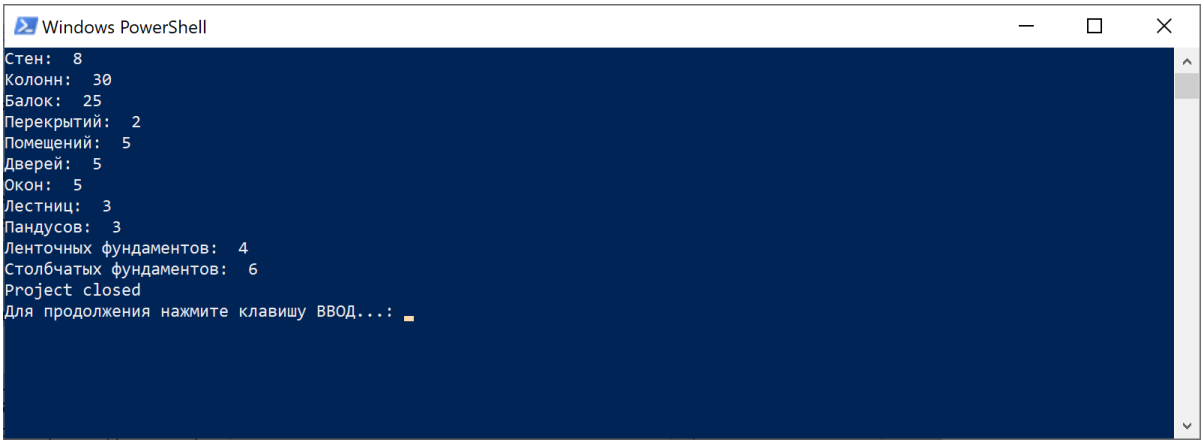


Рис.7. Окно выполнения скрипта подсчета

3. Получение свойств объектов

Свойства объекта получаются вызовом метода *GetProperties()*, объекта *object*. Для определения существования объекта, сравниваем полученные свойства с нулевым объектом: *if ($objproperties -ne $null)*.

Убедившись в существование объекта, можно получать список свойств объекта: *$ids = $objproperties.GetIds();*[10,11].

После используем цикт *for ($j = 0; $j -lt $ids.Count; $j++)* для получения списка свойства объекта по очереди: *$property = $objproperties.GetS($ids.GetS($j));*.

При помощи этого метода появляется доступ к именам свойств и их значениям, пример кода представлен на рисунке 8.

```
switch ($property.Type)
{
        0 {$property_value  = "тип не определен"}
        1 {$property_value  = $property.GetDoubleValue()}
        2 {$property_value  = $property.GetStringValue()}
        3 {$property_value  = $property.GetAngleValue(1)} #1 - градусы- degrees
        4 {$property_value  = $property. GetAreaValue(1)} #1 -мм2, 2 - см2, 3 - м2
        5 {$property_value  = $property. GetDoubleValue()}
        6 {$property_value  = $property.GetEnumerationValue()}
        7 {$property_value  = $property.GetIntegerValue()}
        8 {$property_value  = $property.GetLengthValue(1)}      #1 - мм, 2 - см, 3 - дюймы, 4 - м
        9 {$property_value  = $property.GetLogicalValue()}
        10 {$property_value = $property.GetMassValue(1)}        #1- г, 2-кг, 3-т
        11 {$property_value = $property.GetVolumeValue(3)}      #1 - мм3, 2- см3, 3 - м3
}
```

Рис. 8. Фрагмент кода получения значений свойств объекта

Далее значения выводятся пользователю на экран, пример работы кода представлен на рисунке 9.

```
80.  Окно - Одностворчатое (белый ПВХ): 900,00 мм x 1 600,00 мм
Сопротивление теплопередаче: 0
Обозначение:
Наименование:
Масса: 0
Примечание:
81.  Окно - Одностворчатое (белый ПВХ): 900,00 мм x 1 600,00 мм
Сопротивление теплопередаче: 0
Обозначение:
Наименование:
Масса: 0
Примечание:
82.  Окно - Одностворчатое (белый ПВХ): 900,00 мм x 1 600,00 мм
Сопротивление теплопередаче: 0
Обозначение:
Наименование:
Масса: 0
Примечание:
83.  Окно - Одностворчатое (белый ПВХ): 900,00 мм x 1 600,00 мм
Сопротивление теплопередаче: 0
Обозначение:
Наименование:
Масса: 0
Примечание:
84.  Окно - Одностворчатое (белый ПВХ): 900,00 мм x 1 600,00 мм
Сопротивление теплопередаче: 0
Обозначение:
Наименование:
Масса: 0
Примечание:
Project closed
Для продолжения нажмите клавишу ВВОД...:
```

Рис. 9. Фрагмент кода получения значений свойств объекта

IV. Выводы и заключение

Рассмотрев описанные в предыдущих разделах результаты, можно прийти к выводу о том, что, взаимодействуя с API программного продукта Renga, можно взаимодействовать с внушительным функционалом, а также получать доступ к большому количеству данных, хранящихся в информационной модели.

Оболочка PowerShell отлично подходит для реализации консольного приложения, но стоит отметить тот факт, что приложение такого вида не может быть полностью удобным в работе. Возможности взаимодействия с интерфейсом прикладного программирования Renga возможно посредствам различных языков программирования, таких как C++ и C#, а также динамически типизируемых языков(Python, VBScript и т.д), которые в свою очередь предоставляют разработчику возможность создания GUI (Graphical User Interface – графический интерфейс пользователя).

Разработка приложения с графическим пользовательским интерфейсом выводит взаимодействие с Renga API на качественно новый уровень, так как позволит более гибко настроить сценарии взаимодействия пользователя с информационной моделью и объединить функционал нескольких консольных приложений в одно, более мощное по функционалу.

Список литературы

1. Becerik-Gerber, B. Building information modeling in architecture, engineering, and construction: Emerging research directions and trends / B. Becerik-Gerbe, K. Kensek // J. Prof. Issues Eng. Educ. Pract. 2010. 136, 3. Pp. 139–147.

2. Kensek K.M. Advancing BIM in academia: Explorations in curricular integration. // Computational Design Methods and Technologies: Applications in CAD, CAM and CAE Education. 2012. Pp.101– 121.

3. Smith D.K. Building Information Modeling: A Strategic Implementation Guide for Architects, Engineers, Constructors, and Real Estate Asset Managers. / D.K. Smith, M. Tardiff // J Building Information Modeling: A Strategic Implementation Guide for Architects, Engineers, Constructors, and Real Estate Asset Managers. 2009. 186.

4. Stacks R. et al. BIM Handbook: A Guide to Building Information Modeling for Owners, Designers, Engineers, Contractors, and Facility Managers, 3rd Edition. 3rd ed. Wiley, 2018. 688 p.

5. Ford S. et al. An information engineering approach to modelling building design. Autom. Constr. Elsevier,

1995. 4(1). Pp. 5–15.

6. Kensek K.M. Building information modeling. Building Information Modeling. 2014. 1285 p.

7. Kensek K. BIM guidelines inform facilities management databases: A Case Study over Time // Buildings. 2015. 5(3). Pp. 899–916.

8. Соленый С. В., Воропаев И. А., Давиденко М. Н.. Об API и архитектурных методах защиты

При его разработке//Волновая электроника и инфокоммуникационные системы. 2022. Том Часть 1. С. 251-256.

9. Divin, N.V. BIM by using Revit API and Dynamo. A review// AlfaBuild; 2020; Volume 14 C. 1404 [Электронный ресурс] // Режим доступа: https://www.elibrary.ru/download/elibrary_44656415_27094399.pdf (дата обращения: 30.06.2024).

10. Renga API Documentation [Электронный ресурс] // Режим доступа: https://help.rengabim.com/api/index.html (дата обращения: 30.06.2024).

11. PowerShell Documentation [Электронный ресурс] // Режим доступа: https://learn.microsoft.com/en-us/powershell/ (дата обращения: 30.06.2024).

12. Львов Н.С. Использование функциональных возможностей PowerShell для выполнения процедур резервного копирования данных// CLOUD OF SCIENCE; Том 2 №4 С. 683-690 [Электронный ресурс] // Режим доступа: https://www.elibrary.ru/download/elibrary_25623464_16160256.pdf (дата обращения: 30.07.2024).

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-911-920

Моделирование движения автотранспорта в городской среде со сложным рельефом для определения безопасных режимов движения при различных погодных условиях
Modeling of vehicle traffic in an urban environment with complex terrain to determine safe traffic modes under various weather conditions

М.Е.Елисеев[1], В.С. Макаров[1], Д.П. Смирнов[1], Т.Н. Томчинская[1]
[1] Нижегородский государственный технический университет им. Р.Е. Алексеева, Нижний Новгород, Россия

M.E. Eliseev[1], V.S. Makarov[1], D.P. Smirnov[1], T.N. Tomchinskaya[1]
[1]Nizhny Novgorod State Technical University, n.a. R.E. Alekseev, Nizhny Novgorod, Russia

Аннотация. В целях обеспечения безопасности дорожного движения исследуется один из наиболее ДТП опасных участков Нижнего Новгорода. Участок характеризуется большим перепадом высот и малыми радиусами поворотов. Построена трёхмерная модель данного участка и разработан VR-тренажёр для обучения водителей. Реализована возможность тренировки водителей при разных погодных условиях и при разном состоянии дорожного покрытия. Для определения безопасных режимов движения построена математическая модель, позволяющая оценить момент наступления заноса. На основании математической модели определяется оптимальная для данных метеорологических условий и состояния дорожного полотна скорость. Данные математической модели планируется передавать на динамическое табло переменной информации с целью информирования водителей о безопасных режимах движения, а также применить в автосимуляторе для подготовки водителей к вождению в сложных дорожных и метеорологических условиях.

***Ключевые слова*:** безопасность дорожного движения, VR-тренажёр, трёхмерная модель, математическая модель, погодные условия, информирование водителей.

Annotation. In order to ensure road safety, one of the most accident-prone areas of Nizhny Novgorod is being investigated. The site is characterized by large elevation differences and small turning radii. A three-dimensional model of this section was built and a VR simulator was developed for driver training. The ability to train drivers under different weather conditions and different road surface conditions has been implemented. To determine safe driving modes, a mathematical model was built that allows one to estimate the moment of skidding. Based on the mathematical model, the optimal speed for given meteorological conditions and the state of the roadway is determined. The data from the mathematical model is planned to be transferred to a dynamic display of variable information in order to inform drivers about safe driving modes, and also be used in a car simulator to prepare drivers for driving in difficult road and meteorological conditions.

***Keywords*:** road safety, VR simulator, three-dimensional model, mathematical model, weather conditions, driver information.

I. Введение

Многочисленные исследования свидетельствуют о повышенном риске столкновений транспортных средств в неблагоприятных погодных условиях, таких как дождь, туман и снег [1]. Учёные из университета Айовы Джон Хилл и Линда Бойл [2] обнаружили, что неблагоприятные погодные условия и сопутствующее им ухудшение видимости повышают уровень стресса водителя, повышая риск аварии.

Например, в Финляндии в университете Хельсинки, на факультете психологии, в отделе исследования дорожного движения исследователь Микко Мальмивуо и эколог Харри Пелтола [3] подсчитали, что зимой риск аварий, приводящих к травмам или смертельному исходу, в девять раз больше на заснеженных дорогах и в двадцать четыре раза на обледенелых, по сравнению с чистой дорогой. Авторы вашингтонского университета Шанкар, Маннеринг и Барфилд сообщили [4], что максимальное количество осадков, дождливых дней в месяце и максимальное количество снега способствуют повышению уровня аварийности на северо-западе США. В том же районе, в работе [5] пришли к выводу, что количество ДТП значительно возрастает во время снегопадов, а продолжительность метели, интенсивность снегопада, более высокая скорость ветра и большая интенсивность движения еще больше увеличивают риск. Канадские исследователи Жан Андрей, Брайан Миллс, Майк Лихи и Джеф Саггетт представили подробный обзор погодных условий и риска ДТП [6]. Они сделали четыре основных вывода, которые очень важны для данного исследования:

- риск столкновений обычно возрастает во время осадков, причем типичная оценка в более тщательных исследованиях составляет 50-100%;
- значительное количество данных свидетельствует о том, что снегопад оказывает большее влияние, чем дождь, хотя и с меньшим числом погибших в каждом столкновении;
- риск наиболее высок при ледяном дожде/мороси и первых снегопадах в сезоне;
- возможно, существуют различия в типах дорожно-транспортных происшествий в период осадков и в другое время.

Учёные США Айзенберг и Уорнер проанализировали более 30 миллионов дорожно-транспортных происшествий и сообщили о меньшем количестве погибших, но большем количестве травм и повреждений транспортных средств в дни снегопадов по сравнению с сухими днями [7]. Первые дни снегопада в году были особенно склонны к увеличению числа аварий, в том числе со смертельным исходом.

Хотя оценки риска различны, общая тенденция очевидна: риск дорожно-транспортных происшествий повышается при неблагоприятных погодных и дорожных условиях. Дорожные службы пытаются уменьшить влияние погодных условий с помощью таких операций по техническому обслуживанию, как посыпка солью или песком. Однако работы по техническому обслуживанию часто задерживаются и не позволяют предотвратить значительное увеличение риска, а компенсационные действия требуются и от водителей [8].

Можно утверждать, что скользкость является самым значительным из рисков, связанных с погодными условиями. Более того, даже когда дорога очень скользкая (трение $\leqslant 0{,}20$), водители в среднем все равно едут со скоростью, несколько превышающей установленное ограничение скорости, и считают это безопасным [9]. Наконец, анализ ежедневных аварий в зимнее время выявил пики, которые отражают, что плохая погода и скользкие дороги застают водителей врасплох [10-12].

Особая сложность возникает в городах, расположенных на разных берегах крупных рек. В данной работе, на примере Нижнего Новгорода, рассматривается тип городов, располагающихся на двух берегах реки или на слиянии рек. Многие крупные города относятся к этому типу, приведем некоторые из них: Канзас-Сити, Варшава, Ростов-на-Дону, Миннеаполис, Гамбург, Нью-Йорк и многие другие. В таких населенных пунктах важное транспортное значение имеют мосты и съезды к ним. Как правило, один из берегов реки значительно выше другого, чем обусловлен большой перепад высот на съездах, так, в Нижнем Новгороде он составляет около 140 м. Нередко также участки улично-дорожной сети вблизи мостов (съезды) имеют сложную конфигурацию.

Погодные условия на мостах и съездах могут сильно отличаться от других частей города, добавляя дополнительную опасность для безопасности движения. Большой объем транспортного потока через мосты и съезды делает перекрытие даже одной полосы причиной серьезных пробок и заторов. На крутых подъемах или спусках автомобили чаще подвержены заносам, переворотам и выездам на встречную полосу, особенно при гололеде или дожде.

В зимнее время и при неблагоприятных погодных условиях, таких как снегопады и дождь, участки дороги становятся особенно опасными из-за ухудшения сцепления шин с дорожным покрытием. Это может привести к ДТП, даже при соблюдении установленных правил. В таких условиях безопасная скорость движения на этом участке оказывается ниже разрешенной в 60 км/ч.

Например, на Окском съезде в Нижнем Новгороде при сильных снегопадах безопасная скорость движения составляет от 40 до 50 км/ч в зависимости от условий и опыта водителя. Хотя правила позволяют двигаться со скоростью до 60 км/ч, рекомендуемая скорость для безопасного прохождения данного участка значительно меньше.

В исследованиях [13-17] рассматривается влияние геометрии дороги на аварийность. В статье [13] устанавливаются критические факторы, наличие которых обязательно приводит к повышенной аварийности, отмечается, что малый радиус поворота и большая глубина колеи приводят к повышенной аварийности. В работах [14-16] рассматриваются модели прогнозирования числа аварий в зависимости от геометрии дороги.

В качестве решения этой проблемы в работах [18,19] рассматривается вопрос влияния дорожных информационных табло (ДИТ) на поведение водителей, в частности влияние информирования о необходимости снижения скорости. Обычно ДИТ интегрируется с знаками переменной информации (ЗПИ). Отмечается, что даже с учетом того, что часть водителей игнорируют информацию ДИТ и ЗПИ, в целом фиксируется снижение скорости транспортного потока при соответствующей информации на табло, что доказывает эффективность их применения.

Таким образом, водители действительно воспринимают риски, связанные, как со сложными дорожными, так и с погодными условиями. Они корректируют свое поведение на дороге в определенной степени, но недостаточно [20]. В результате, стратегия предотвращения аварий, полагающаяся на компенсирующее поведение водителей на дороге, вряд ли будет успешной.

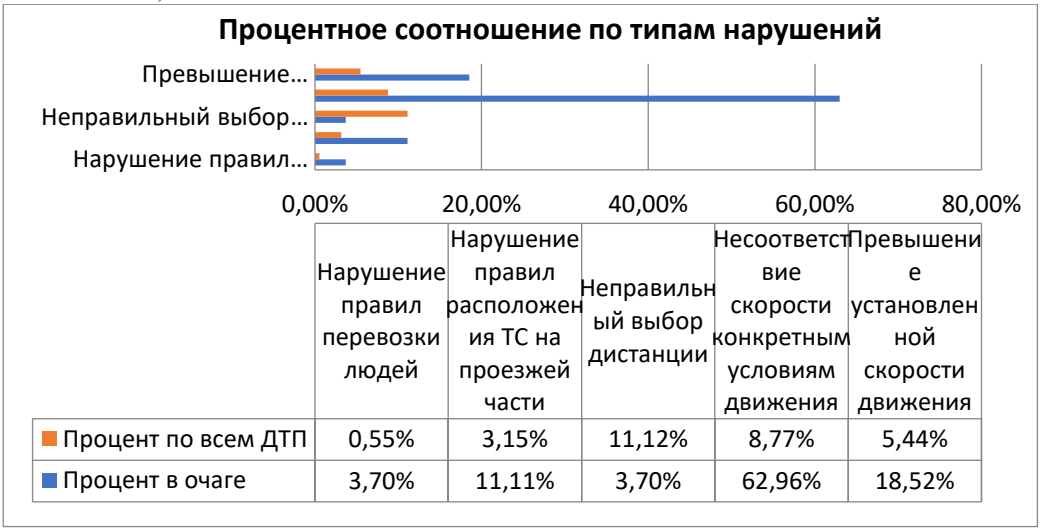
Исследования, приведённые в работе [21] показывают, что на поведение вождения на дороге преимущественно влияют преобладающие наблюдаемые условия, а не прогноз погоды. Поэтому, если администраторы хотят помочь водителям скорректировать свое поведение на дороге в неблагоприятных погодных и сложных дорожных условиях, то методы должны носить более локальный и технический характер.

II. Аварийность на исследуемом участке Нижнего Новгорода

Всего за 2023-й год в Нижегородской области произошло 6,4 тысячи ДТП. В 4 645 из них получили ранения 5 923 человека. В 2022 году произошло 4 367 ДТП с 5 615 пострадавшими. Таким образом, количество пострадавших в ДТП выросло за год на 5,9%. В качестве основных причин аварий были названы выезд водителей на встречную полосу и нарушения скоростного режима [22].

В данной работе исследуется один из наиболее опасных участков Нижнего Новгорода – Окский съезд к Молитовскому мосту через Оку. За последние пять лет на участке дороги Окского съезда произошло множество аварий с участием легковых, грузовых автомобилей и общественного транспорта. Геометрия исследуемого участка отличается переменным радиусом поворота, причём величина его значительно меняется, встречаются как правый, так и левый повороты. Перепад высот на участке длиной 150 м составляет 5 м, радиус поворота – от 65 до 90 м в зависимости от выбранной полосы движения.

Характерно, что аварии, случившиеся на данном участке за последние пять лет располагаются неравномерно, и очаги аварий совпадают с участками значительного перепада высот и одновременно наименьшими радиусами поворота. Для исследования особенностей данного участка дороги рассматривались характеристики «Тип нарушения» и «Количество транспортных средств (ТС), участников ДТП». Сравнительный анализ по данным характеристикам, приводится на рис. 1 и рис. 2 (указываются только типы аварий, встречавшиеся в период 2018 – 2022 на Окском съезде).



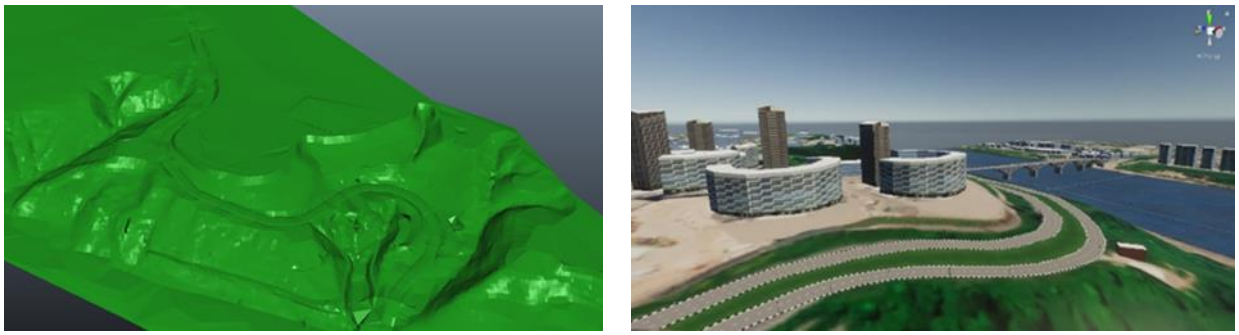
	Нарушение правил перевозки людей	Нарушение правил расположения ТС на проезжей части	Неправильный выбор дистанции	Несоответствие скорости конкретным условиям движения	Превышение установленной скорости движения
Процент по всем ДТП	0,55%	3,15%	11,12%	8,77%	5,44%
Процент в очаге	3,70%	11,11%	3,70%	62,96%	18,52%

Рис. 1. Сравнение процента нарушений на Окском съезде и по всем ДТП в Нижнем Новгороде (за 2018-2022 гг).

Проведенный анализ факторов аварийности для данного участка улично-дорожной сети позволяет установить механизм аварий. Характерным для Окского съезда является следующий механизм аварийности: водитель неправильно выбрал скорость, которая не соответствовала конкретным условиям движения, дорожное покрытие было мокрым или скользким.

Рис. 2. Количество ТС, участвовавших в ДТП на Окском съезде

Далее водитель не справился с управлением и в результате заноса произошло столкновение с отбойником или другим препятствием. Более подробно описание геометрии исследуемого участка и анализ аварийности приведён в статье [23].



а *б*

Рис. 3. Трёхмерная модель Окского съезда: *а* – выполненная в программе Civil 3D; *б* – эта же модель, достроенная в программе Blender

На этом участке планируется установить автоматическую дорожную метеостанцию. Будут сниматься показания с дорожного покрытия и на динамическом табло переменной информации выводиться рекомендуемая скорость. Выявлено, что предупреждение водителей на динамическом табло переменной информации способствует снижению скорости на 5-7 км, что снижает аварийность.

III. Математическая модель автомобиля криволинейного движения по дороге с уклоном

Для выяснения безопасных режимов движения применяется математическая модель автомобиля криволинейного движения по дороге с уклоном. В рассматриваемой модели используется ряд допущений:

- математическая модель транспортного средства плоская;
- учет движения по уклону происходит за счет перераспределения реакций на соответствующих колесах машины;
- учет инерционных составляющих происходит за счет перераспределения реакций на соответствующих колесах машины;
- дорожное покрытие на моделируемом участке полагаем ровным (отсутствуют ямы, колеи и т.п.);
- масса автомобиля без учета движения по уклону и инерционных составляющих, приходящаяся на колеса автомобиля постоянна;
- тягово-сцепные свойства автомобиля определяются через коэффициенты сцепления и сопротивления, которые постоянны (либо дискретно постоянны) для всего моделируемого участка дороги;
- пренебрегаем реакциями подвески автомобиля;
- положение центра тяжести автомобиля считаем неизменным.

При движении автомобиля по спуску с определённой кривизной могут возникать различные варианты режимов движения, которые рассматриваются далее.

Рассмотрим сначала криволинейное движение автомобиля с постоянной (наперед заданной, фиксированной) скоростью. Полагаем, что отсутствует боковое скольжение. В качестве критерия устойчивого движения принимаем отсутствие бокового скольжения колес каждой из осей автомобиля.

Общая схема движения автомобиля в повороте (рис.4) будет следующей [24-26].

На рис.4 показаны следующие обозначения:

$i = \overline{1,2}$,передняя и задняя ось машины;

$j = \overline{в,н}$, внутренний и наружный борт машины;

θ – угол поворота колес машины по продольной оси автомобиля;

θ_{ij} – угол поворота колеса машины;

δ_i – среднее значение угла увода колес автомобиля, обусловленное деформацией шин;

δ_{ij} – угол увода колес автомобиля, обусловленный деформацией шин;

L – база автомобиля;

a, b – расстояния от передней и задней осей до центра масс автомобиля;

R_δ – радиус поворота автомобиля.

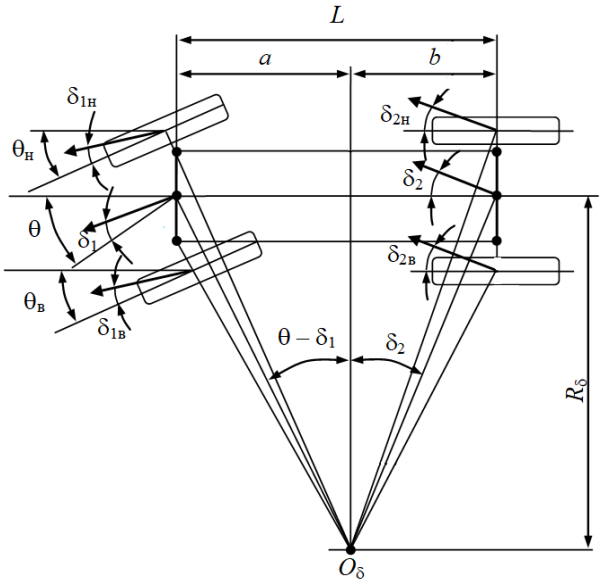


Рис. 4. Кинематика поворота автомобиля с эластичными колесами

Для исследуемой задачи, а именно расчета радиуса поворота, достаточно использовать велосипедную модель. Обоснованием этого допущения является то, что на высокой скорости радиус поворота не может быть маленьким и разница углах увода между левым и правым колесом незначительна. Это упрощает задачу и позволяет получить достаточно точные значения относительно решения задачи, где машина рассматривается со всеми четырьмя колесами. Разница в углах составляет не более 0,5-2%.

Радиус поворота рассчитывается по зависимости [24, 25]:

$$R_\delta = L / [tg(\theta - \delta_1) + tg\delta_2] \approx L / (\theta + \delta_2 - \delta_1) \quad (1)$$

В общем случае поперечная сила инерции рассчитывается по следующей формуле:

$$F_{ay} = -m_a a_y, \quad (2)$$

где m_a – масса автомобиля, a_y – поперечное ускорение.

Выражая a_y, с учетом стандартных формул теоретической механики, получаем:

$$F_{ay} = -m_a \left\{ \frac{V^2}{R_\delta} + \frac{V[b(\dot{\Theta} - \dot{\delta_1}) - a\dot{\delta_2}]}{L} + \dot{V}\left(\frac{b}{R_\delta} - \delta_2\right) \right\} \quad (3)$$

Расчеты показывают [24, 25], что до 90 % поперечной силы инерции составляет первое слагаемое в формуле (3), второе слагаемое имеет существенное значение при резких поворотах управляемых колес, третье при резких разгонах и торможениях. Поэтому в дальнейших расчетах принимаем, что сила инерции будет рассчитываться по формуле:

$$F_{ay} = -m_a \frac{V^2}{R_\delta}. \quad (4)$$

Поперечная устойчивость движения в повороте в теории автомобиля оценивается по условию отсутствия опрокидывания (при высоких скоростях) и по условию отсутствия бокового скольжения (при недостаточном сцеплении с дорогой). Рассмотрим зависимости для этих случаев. На практике второй вариант возникает чаще, первый вариант как правило возникает при недостаточной квалификации водителя.

Далее рассматривается математическая модель, криволинейного движения автомобиля с определенной скоростью при недостаточном сцеплении колес с дорогой, которая позволяет определить момент начала бокового скольжения колес с дорогой.

На рис. 5 приведена схема движения автомобиля при повороте [24, 25].

На рис. 5 показаны следующие обозначения: G_a– вес автомобиля, приходящийся на ось машины; С–точка центра тяжести (в расчетах принимаем, что она находится над осью машины и на высоте центра тяжести автомобиля); h_g - высота центра тяжести автомобиля; В – колея машины; $R_{зл}$, $R_{зп}$ – вертикальные реакции на колесах автомобиля; $R_{ул}$, $R_{уп}$ – продольные реакции на колесах автомобиля.

F_{ay}– определена по зависимости (4). Критическим условием по скольжению будет равенство поперечных сил.

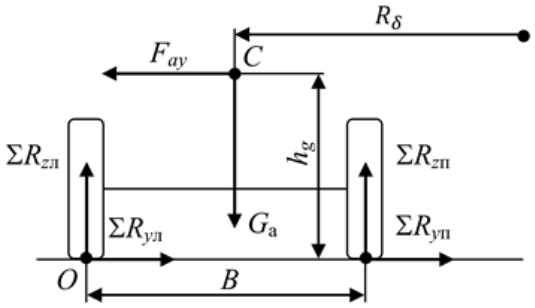


Рис. 5. Силы, действующие на автомобиль при круговом движении

$$R_{ул} + R_{уп} = F_{ay};$$

$$R_{ул} + R_{уп} = \varphi_y(R_{зл} + R_{зп});$$

$$R_{ул} + R_{уп} = \varphi_y m_a g;$$

$$m_a \frac{V^2}{R_\delta} = \varphi_y m_a g;$$

$$V = \sqrt{\varphi_y g R_\delta}, \quad (5)$$

где φ_y – коэффициент поперечного сцепления шин с дорогой, g – ускорение свободного падения.

Фактически с учетом всех допущений критическим условием, определяющим критическую скорость движения в повороте машины, является коэффициент сцепления с дорогой. Причем по разным литературным источникам φ_y=φ_x(0,5-0,85), таким как В.В. Беляков и А.А. Куркин [27], В.Н. Кравец [24], В.И. Кнороз [28], А.П. Васильев [29]. При движении по льду или обледенелой дороге большую роль играет состав материала шины и температура дорожного полотна [30,31].

ТАБЛИЦА 1
КОЭФФИЦИЕНТЫ СЦЕПЛЕНИЯ СОПРОТИВЛЕНИЯ ДОРОГИ (KRAVETS, 2013)

Тип и состояние дороги	φ_x	φ_y	f
Сухой асфальт и бетон	0,8…0,9	0,4...0,765	0,007 … 0,015
Мокрый асфальт	0,5…0,7	0,25...0,595	0,015 … 0,020
Мокрый бетон	0,75…0,8	0,375...0,68	0,015 … 0,020
Грунтовая дорога			
сухая	0,65…0,7	0,325...0,595	0,025 … 0,030
мокрая	0,5…0,55	0,25...0,4675	0,050 … 0,150
Уплотненный снег	0,15…0,2	0,075...0,17	0,030 … 0,050
Лед	0,1	0,05...0,085	0,015 … 0,030

Поэтому в соответствии с формулой (5) критическими скоростями для разных условий движения будут следующие их значения, отраженные в таблице 2 и на рис.6.

ТАБЛИЦА 2
КРИТИЧЕСКИЕ СКОРОСТИ ДВИЖЕНИЯ ПО ЗАНОСУ ПРИ РАЗНЫХ РАДИУСАХ ПОВОРОТА И КОЭФФИЦИЕНТАХ СЦЕПЛЕНИЯ С ДОРОГОЙ

φ_x	R_δ, м									
	10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*
0.05	7.97	11.28	13.81	15.95	17.83	19.53	21.09	22.55	23.92	25.21
0.1	11.28	15.95	19.53	22.55	25.21	27.62	29.83	31.89	33.83	35.66
0.2	15.95	22.55	27.62	31.89	35.66	39.06	42.19	45.10	47.84	50.43
0.3	19.53	27.62	33.83	39.06	43.67	47.84	51.67	55.24	58.59	61.76
0.4	22.55	31.89	39.06	45.10	50.43	55.24	59.66	63.78	67.65	71.31
0.5	25.21	35.66	43.67	50.43	56.38	61.76	66.71	71.31	75.64	79.73
0.6	27.62	39.06	47.84	55.24	61.76	67.65	73.07	78.12	82.86	87.34
0.7	29.83	42.19	51.67	59.66	66.71	73.07	78.93	84.38	89.50	94.34
0.8	31.89	45.10	55.24	63.78	71.31	78.12	84.38	90.20	95.68	100.85

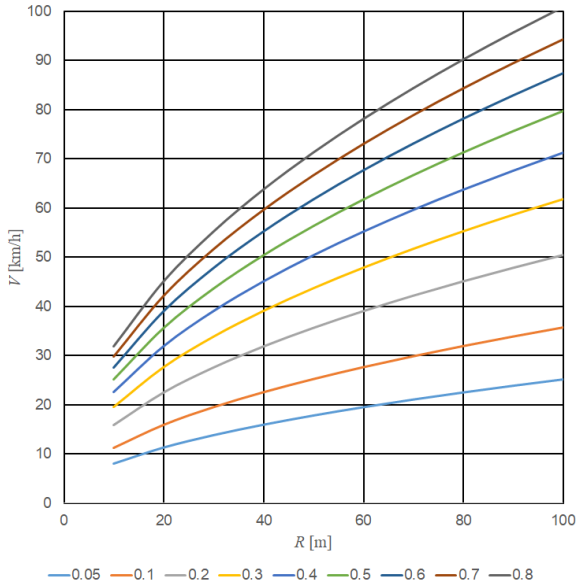


Рис. 6. Предельные скорости движения автомобиля по условию отсутствия бокового скольжения в зависимости от радиуса поворота для разных коэффициентов сцепления с дорогой

Таким образом, приведенные математические модели, позволяют определить:
- критические скорости движения по заносу при разных радиусах поворота и коэффициентах сцепления с дорогой;
- предельные замедления при разных углах уклона и коэффициентах сцепления с дорогой;
- значения остановочного пути при разных скоростях, углах уклона и коэффициентах сцепления с дорогой.

Это позволяет выбрать безопасные параметры криволинейного движения автомобилей по дороге с уклоном. Предлагается вычисленные параметры выводить на динамическом табло переменной информации, а также использовать в виртуальном тренажёре для водителей.

IV. Разработка VR-тренажера

Для тренировки начинающих и опытных водителей вождению на сложных участках при различных погодных условиях предлагается автосимулятор. Построена трёхмерная модель Окского съезда Нижнего Новгорода. Подробнее построение трёхмерной модели сцены и функции VR-тренажера рассмотрены в работе [23]. В данной работе смоделированы различные погодные сценарии и дорожное покрытие при гололёде, снегопаде и ливне.

Управление в симуляторах вождения способствует повышению реализма обучения, поскольку позволяет курсантам получать опыт вождения в различных реалистичных ситуациях на дороге, включая дождь, снег, гололед и туман, без реальной угрозы для их жизни или здоровью. Курсанты могут научиться правильно реагировать на изменяющиеся дорожные условия и принимать соответствующие меры предосторожности, что повышает их уверенность и подготовленность к реальному вождению.

В качестве инструментальной среды использовался игровой движок Unreal Engine5, разрабатываемый и поддерживаемый компанией Epic Game. Моделирование адаптивного управления в симуляторах вождения на базе Unreal Engine5 значительно повышает реалистичность симуляций. Адаптивное управление позволяет автомобилю реагировать на изменяющиеся условия на дороге так же, как это делали бы реальные автомобили.

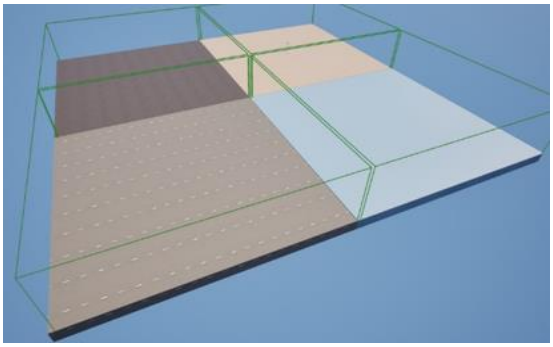
Реалистичные эффекты погоды и динамическое управление автомобилем создают более захватывающий и интересный опыт обучения для курсантов. Возможность управлять автомобилем в различных погодных условиях и наблюдать за изменениями в его поведении увеличивает уровень вовлеченности курсантов и делает процесс обучения более привлекательным.

Рис.7. Вид снежной сцены

Рис. 8. Занос автомобиля на заснеженной дороге

Для реализации адаптивного управления в зависимости от поверхности дороги и атмосферных условий использовались средства программирования в рамках среды Unreal Engine 5, в частности, Blueprint - визуальный скриптовый язык, который позволяет создавать логику поведения объектов в игре. С помощью Blueprint настраивались действия, которые должны происходить при взаимодействии автомобиля с триггерной зоной.

Например, при пересечении триггера автомобиль может активировать соответствующие сценарии поведения, которые изменяют параметры управления (рис. 9, а).



а

б

Рис. 9. Моделирование движения с различным дорожным покрытием. а - тестовый уровень с Trigger Box в симуляторе вождения; б - вид “дождливой сцены”

В частности, коэффициенты сцепления, момент инерции, сопротивление вращению колес и другие. Эти параметры настраиваются таким образом, чтобы автомобиль мог адаптироваться к различным дорожным условиям, обеспечивая управление, приближенное к реальным условиям (рис. 9, б).

Также были использованы физические материалы. Заданный материал прикрепляется к основному материалу, ответственному за текстуры и визуальное представление поверхности, что позволяет точно настраивать поведение автомобиля в зависимости от типа дорожного покрытия и текущих атмосферных условий, обеспечивая точное взаимодействие с окружающей средой.

С помощью такого материала можно настроить коэффициент трения для каждого типа поверхности, а также другие параметры, что позволяет имитировать поездку с разнообразными дорожными условиями и обеспечивать более реалистичное управление.

V. Заключение

Исследовано поведение водителей при различных погодных условиях. Выявлено, что неблагоприятные погодные и дорожные условия, например такие как, дождь, снегопад и резкие колебания температуры, увеличивают риск ДТП из-за снижения видимости, ухудшения сцепления с дорогой и увеличения тормозного пути. Особенно опасными становятся неблагоприятные погодные факторы, когда они сочетаются с крутыми поворотами и значительным уклоном.

Анализируется один из наиболее аварийных участков дороги Нижнего Новгорода – Окский съезд, включающий в себя опасные повороты и перепады высот. Разработана трёхмерная модель этого участка и VR-тренажёр. Предлагается на точно воссозданном рельефе и состоянии дорожного покрытия тренировать начинающих и опытных водителей.

Представлена математическая модель, которая позволяет определить оптимальную скорость для заданных метеорологических условий и состояния дорожного полотна.

Планируется использование результатов моделирования в интеллектуальных транспортных системах для повышения безопасности на дорогах путем информирования водителей о безопасных режимах движения. Исследования [18,19] показывают эффективность информирования при косвенном управлении транспортным потоком. Результаты показали, что эти знаки с переменным сообщением действительно имеют другие эффекты, такие как перефокусировка внимания на поиск сигналов о потенциальных опасностях, проверка скользкости дороги и более осторожное поведение при обгоне. За счет снижения аварийности и уменьшения дисперсии скоростей ТС в транспортном потоке увеличивается пропускная способность данного участка дороги и повышается мобильность ТС.

Предлагается данные математической модели ввести в автосимулятор для более эффективной тренировки водителей.

Адаптация водителя к различным погодным условиям, таким как дождь, снег, гололед или туман, играет ключевую роль в обеспечении безопасности на дорогах. Поэтому своевременное оповещение и тренировка водителей, способных эффективно адаптироваться к таким условиям, имеет огромное значение для обеспечения безопасности дорожного движения.

Список литературы

1. Andrey J., Mills B., Leahy M., Suggett, J. Weather as a chronic hazard for road transportation in Canadian cities // 2003; 17(3): 34-36.
2. John D. Hill, Linda Ng Boyle.,Driver stress as influenced by driving maneuvers and roadway conditions //Transportation Research Part F: Traffic Psychology and Behaviour, 2007,V. 10, P. 177-186, USA
3. Malmivuo M., Peltola H. Talviajanliikenneturvallisuus, tilastollinentarkastelu // 1991–1995 (6/1997); Helsinki, Finland: Tielaitos, Tiehallinto.
4. Shankar V., Mannering F., Barfield W. Effect of roadway geometrics and environmental-factors on rural freeway accident frequencies. Accident Analysis and Prevention // 1995; 27(3): 371–389
5. Khattak A. J., Knapp K. K. Snow event effects on interstate highway crashes. Journal of Cold Regions Engineering // 2001; 15(4): 219–229.
6. Andrey J., Mills B., Leahy M., Suggett, J. Weather as a chronic hazard for road transportation in Canadian cities // 2003; 17(3): 34-36.
7. Eisenberg D., Warner K. E. Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. American Journal of Public Health // 2005; 95(1): 120–124.
8. Rumar K., Berggrund U., Jernberg P., Ytterbom U. Driver reaction to a technical safety measure studded tires. Human Factors // 1976; 18(5): 443–454.
9. Heinijoki H. Kelin kokemisen, rengaskunnon ja rengastyypinvaikutusnopeuskayttaytymiseen // (19/1994); Helsinki, Finland: 52-78
10. Anttila V., Nygard M., Rama P. Liikennesa a tiedotuksentoteutuminen ja arviointitalvikaudella. Tiehallinnonselvityksia // (41/2001); Helsinki, Finland: Tiehallinto, Liikenteenpalvelu
11. Malin Fanny, NorrosIlkka, InnamaaSatu, Accident risk of road and weather conditions on different road types,Accident Analysis &Prevention,Volume 122,2019,Pages 181-188,ISSN 0001-4575, https://doi.org/10.1016/j.aap.2018.10.014.
12. Norros Ilkka, KuuselaPirkko, InnamaaSatu, Pilli-SihvolaEetu, RajamäkiRiikka, The Palm distribution of traffic conditions and its application to accident risk assessment, Analytic Methods in Accident Research, Volume 12, 2016, Pages 48-65, ISSN 2213-6657, https://doi.org/10.1016/j.amar.2016.10.002.
13. Othman S, Thomson R, Lannér G. Identifying critical road geometry parameters affecting crash rate and crash type. Ann AdvAutomot Med. 2009 Oct;53:155-65. PMID: 20184841; PMCID: PMC3256788.

14. VayalamkuzhiPraveen, AmirthalingamVeeraragavan, Influence of geometric design characteristics on safety under heterogeneous traffic flow, Journal of Traffic and Transportation Engineering (English Edition), Volume 3, Issue 6, 2016, Pages 559-570, ISSN 2095-7564, https://doi.org/10.1016/j.jtte.2016.05.006.
15. Choudhary, A., Garg, R.D., Jain, S.S. (2023). Analysis of Pavement and Geometric Factors of Selected Highways for Reduction in Road Accidents. Proceedings of the Sixth International Conference of Transportation Research Group of India , vol 273. Springer, Singapore. https://doi.org/10.1007/978-981-19-4204-4_16
16. Milton, J., Mannering, F. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, 395–413 (1998). https://doi.org/10.1023/A:1005095725001
17. Park, S.J.; Kho, S.Y.; Park, H.C. The effects of road geometry on the injury severity of expressway traffic accident depending on weather conditions. J. Korea Inst. Intell. Transp. Syst. 2019,18, 12–28.
18. Glendon A. Ian, Lewis Ioni, Field testing anti-speeding messages, Transportation Research Part F: Traffic Psychology and Behaviour, Volume 91, 2022, Pages 431-450, ISSN 1369-8478, https://doi.org/10.1016/j.trf.2022.10.022.
19. Juha Luoma, Pirkko Rama, Merja Penttinen,Virpi Anttila. Effects of variable message signs for slippery road conditions on reported driver behaviour. Transportation Research Part F: Traffic Psychology and Behaviour// 2000, V.3, P. 75-84 URL: https://doi.org/10.1016/S1369-8478(00)00017-6
20. Edwards J. B. Speed adjustment of motorway commuter traffic to inclement weather. Transportation Research Part F: Traffic Psychology and Behaviour // 1999; 2(1): 1–14.
21. Markku Kilpelainen, Heikki Summala. Effects of weather and weather forecasts on driver behaviour.Transportation Research Part F: Traffic Psychology and Behaviour// 2007; 2(1): 1–14. V. 10, P. 288-299
22. nn-now.ru. [Электронный ресурс] // 2024 URL: https://nn-now.ru/kolichestvo-dtp-v-nizhegorodskoj-oblasti-vyroslo-na-64-v-2023-godu/ (дата обращения 22.03.2024).
23. Ермолаева А.Д, Елисеев М.Е., Томчинская Т.Н. «VR-тренажер для обучения водителей в городской среде со сложным рельефом и метеоусловиями» Труды Международной конференции по компьютерной графике и зрению "ГрафиКон 2023", № 33, pp. 923-932, 2023.
24. Кравец В.Н. Теория движения автомобиля. Издательство: Нижегородский государственный технический университет им. Р.Е. Алексеева, 2014, 697 с.
25. Кравец В.Н., Селифонов В.В. Теория автомобиля. Издательство: Гринлайт, 2011, 884 с.
26. Pacejka, Hans B. (2002) Tyre Mechanics and Vehicle Dynamics. Oxford, UK: Butterworth-Heinemann. A member of the Reed Elsevier plc group. ISBN: 0-7506-5141-5. 621p
27. Беляков В.В., Куркин А.А. (2014) Полотно пути транспортно-технологических машин (справочные материалы к теории «машина-местность») Н. Новгород: НГТУ им. Р.Е. Алексеева, 447 с
28. Кнороз, В.И. Работа автомобильной шины. – М.: Транспорт, 1976. – 238 с.
29. Васильев А.П. (2004) Справочная энциклопедия дорожника II том Ремонт и содержание автомобильных дорог, МОСКВА, 1129 с
30. Shenvi Mohit Nitin, Verulkar Adwait, Sandu Corina, Mousavi Hoda. Tread rubber compound effect in winter tires: An experimental study, Journal of Terramechanics, Volume 99, 2022, Pages 57-73, ISSN 0022-4898, https://doi.org/10.1016/j.jterra.2021.11.004.
31. Mousavi Hoda, Sandu Corina, Tire-ice model development for the simulation of rubber compounds effect on tire performance, Journal of Terramechanics, Volume 91, 2020, Pages 97-115, ISSN 0022-4898, https://doi.org/10.1016/j.jterra.2020.06.001.

УДК 519.876.5
DOI: 10.25206/978-5-8149-3873-2-2024-921-928

Разработка 3D-моделей подводного аппарата с помощью технологии реверс-инжиниринг
Designing 3D models of the underwater submersible by using reverse-engineering technology

Е. А. Бажин[1], Е. С. Глумова[1], А. Д. Филинских[1]
[1]Нижегородский государственный технический университет им. Р.Е. Алексеева

E. A. Bazhin[1], E. S. Glumova[1], A. D. Filinskikh[1]
[1]Nizhny Novgorod State Technical University n. a. R.E. Alekseev

Аннотация. Реверсивный инжиниринг используется в различных сферах. Её можно условно разделить на три основные направления: создание прототипов новых деталей, модификация уже существующих, а также разработка полностью новых компонентов на базе существующих объектов и концепта. Одним из самых востребованных вариантов использования данной технологии является восстановление изношенных или поврежденных компонентов, которые невозможно приобрести у оригинального производителя, либо на которые отсутствует нужная документация. Такая ситуация часто возникает при эксплуатации импортного или старого оборудования. В статье рассматривается процесс получения геометрической информации об объекте с помощью технологий обратного проектирования. На основе данных лазерного сканирования были созданы трёхмерные твердотельные модели отдельных частей батискафа, такие как корпус, турбины, купол, чаша пилота, люк, технический отсек. Исследуется возможность автоматизированного создания геометрически параметризированных моделей корпуса батискафа и других элементов. Полученные результаты демонстрируют потенциал реверсивного проектирования и быстрого создания прототипов в судостроительной отрасли.

***Ключевые слова*:** реверс-инжиниринг, трехмерное моделирование, 3D-сканирование, твердотельная модель, облако точек, САПР, батискаф, Geomagix Design X, Geomagix Control.

Аннотация. Reverse engineering is used in a variety of applications. It can be roughly divided into three main areas: prototyping of new parts, modification of existing parts, and development of completely new components based on existing objects and concept. One of the most popular uses of this technology is the restoration of worn or damaged components that cannot be purchased from the original manufacturer or that lack the necessary documentation. This situation often arises when operating imported or old equipment. This paper discusses the process of obtaining geometric information about an object using reverse engineering techniques. Three-dimensional solid models of individual parts of the bathyscaphe, such as the hull, turbines, dome, pilot bowl, hatch, and technical compartment, were created based on laser scanning data. The possibility of automated creation of geometrically parameterised models of the bathyscaphe hull and other elements is being investigated. The obtained results demonstrate the potential of reverse engineering and rapid prototyping in the shipbuilding industry.

***Ключевые слова*:** reverse engineering, three-dimensional modelling, 3D scanning, solid model, point cloud, CAD, bathyscaphe, Geomagix Design X, Geomagix Control.

I. Введение

В настоящее время различные отрасли промышленности сталкиваются с проблемой разработки технической документации для производства как отдельных компонентов, так и сборочных единиц. Это связано с уходом многих зарубежных компаний с российского рынка, что обострило потребность в импортозамещении определённых видов продукции.

Реверс-инжиниринг представляет собой современную методику, которая позволяет преобразовывать физические объекты в цифровые модели. Одним из наиболее распространённых способов использования данной технологии является восстановление изношенных или повреждённых деталей, которые невозможно приобрести у оригинальных производителей или для которых отсутствует необходимая документация. Это даёт возможность значительно сократить как временные, так и финансовые ресурсы при разработке деталей и сборочных единиц [1].

В рамках данной статьи технология реверс-инжиниринга применяется для восстановления геометрических характеристик батискафа.

II. Постановка задачи

Целью исследования является получение геометрической информации об элементах батискафа.

Для восстановления геометрических размеров объекта необходимо сначала проанализировать данные на предмет возможности автоматического или полуавтоматического восстановления формы. В данном случае воссоздание формы будет происходить по данным лазерного сканирования, поэтому необходимо убедиться в целостность сканов. Затем требует обработать сканы для моделирования конкретных частей батискафа. После моделирования необходимо произвести анализ точности разработанных трехмерных моделей.

Перед моделированием элементов батискафа необходимо разработать требования к конечным твердотельным моделям. К таким требованиям можно отнести допустимую погрешность построения в ±3 мм. Это максимальное отклонение построенной трехмерной модели от облака точек. Поскольку дальнейшие модели будут использоваться для построения чертежей и изготовления проектной документации, необходимо предусмотреть корректный импорт в среды геометрического моделирования [2]. Было принято решение использовать универсальный формат передачи трехмерных данных .STEP.

III. Теория

1. Исходные данные

Объектом исследования является подводный аппарат «Марина» [25]— это двухместный обитаемый аппарат сухого типа для подводных погружений и подводного плавания в прибрежных районах с глубинами до 100 метров. Подводный аппарат приспособлен для быстрой буксировки любым самоходным судном.

В качестве исходных данных для восстановления использовались результаты лазерного сканирования подводного судна в разобранном виде. Всего было предоставлено 7 сканов в формате .stl и 6 фотографий (рис. 1).

Рис. 1. Исходные данные для моделирования.

2. Общая схема создания моделей

Процесс реинжиниринга можно разбить на следующие этапы:

1) сбор данных;
2) цифровая реконструкция;
3) анализ и оптимизация;
4) анализ материалов.

В данной статье подробно рассмотрены этапы цифровой реконструкции и анализа полученных данных. Общая схема создания моделей приведена на рис. 2.

[25] https://nd-marine.com/

Рис. 2. Общая схема создания моделей.

Процесс цифровой реконструкции – довольно трудоемкий и также включает в себя несколько этапов: импорт данных лазерного сканирования, чистка облака, выбор способа построения модели, наконец, само построение.

Обработка данных может осуществляться непосредственно в программных продуктах от производителя сканера или в сторонних приложениях. Данные о модели сканера могут быть полезными в процессе обработки, так как дают представление о возможной плотности получаемых облаков точек. Если таких данных нет, производить сшивку облаков все равно представляется возможным. Программные продукты по обработке облаков точек от различных вендоров предполагают идентичный функционал. Важным этапом является перевод сетчатой модели в твердотельную, поэтому при выборе программы для обработки данных трехмерного сканирования необходимо ориентироваться на методы построения, используемые в САПР. Таким решением в контексте данной задачи стал Geomagic Design X [3].

3. Подготовка сканов

Как уже говорилось выше, цифровая реконструкция начинается с импорта облака точек в специализированное ПО и его обработки. Каждый из семи предоставленных сканов подвергся очистке и исправлению дефектов. Хотя лазерное сканирование является высокоточным методом получения информации об объекте, в облаке точек могут присутствовать дубликаты, шумы и выбросы [4]. Кроме того, некоторые части объекта могут быть отсканированы, но не относиться к моделируемому в данный момент объекту (рис. 3).

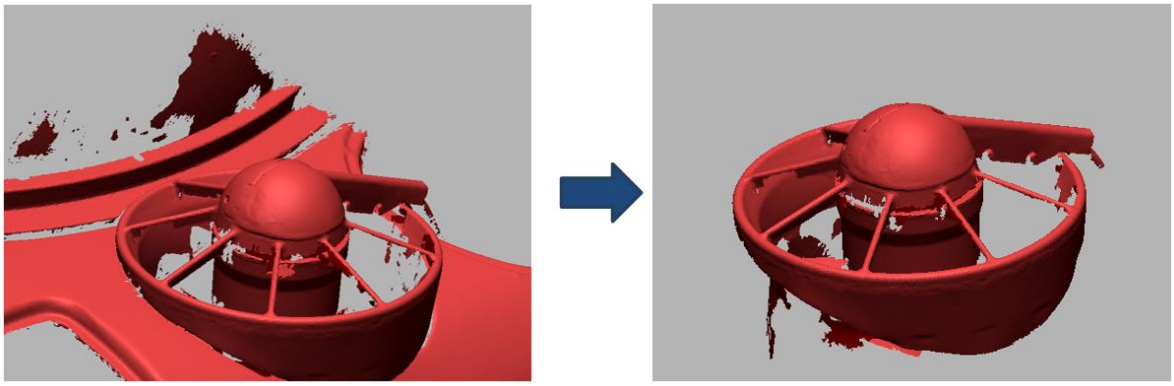


Рис. 3. Пример чистки скана: удаление поверхности верхней части корпуса со скана турбины.

Работа ведется в одном проекте, поэтому важно загрузить и выровнять сканы относительно друг друга. Для сшивки сканов можно воспользоваться инструментами ручного или автоматического приближения. Результат совмещения сканов продемонстрирован на рисунке 4.

Если говорить об исправлении дефектов, то в Geomagic Design X есть множество инструментов для автоматического исправления ошибок в сканах. С помощью них можно заполнить отверстия, искусственно создав сетку поверхности, сгладить или, наоборот, повысить детализацию некоторых областей (рис. 5).

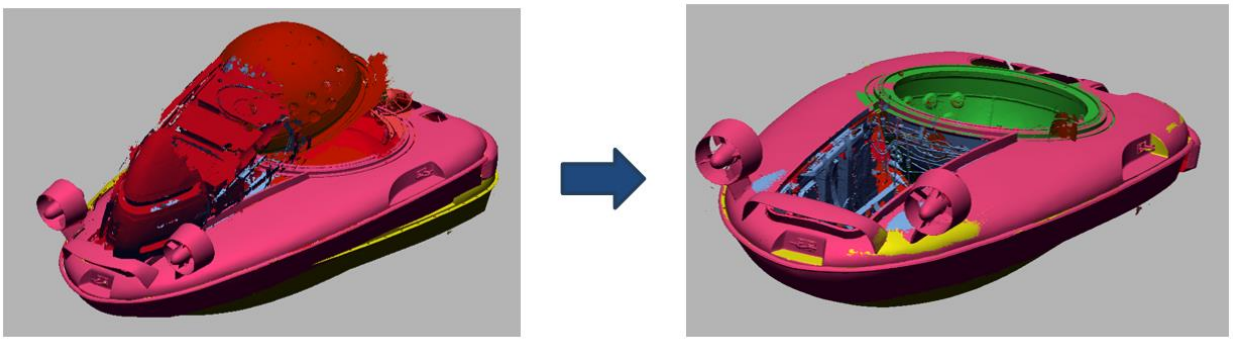
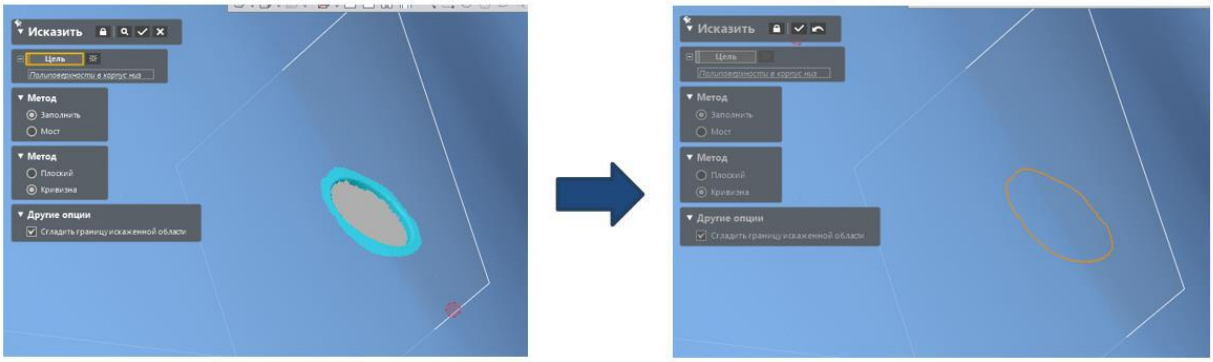
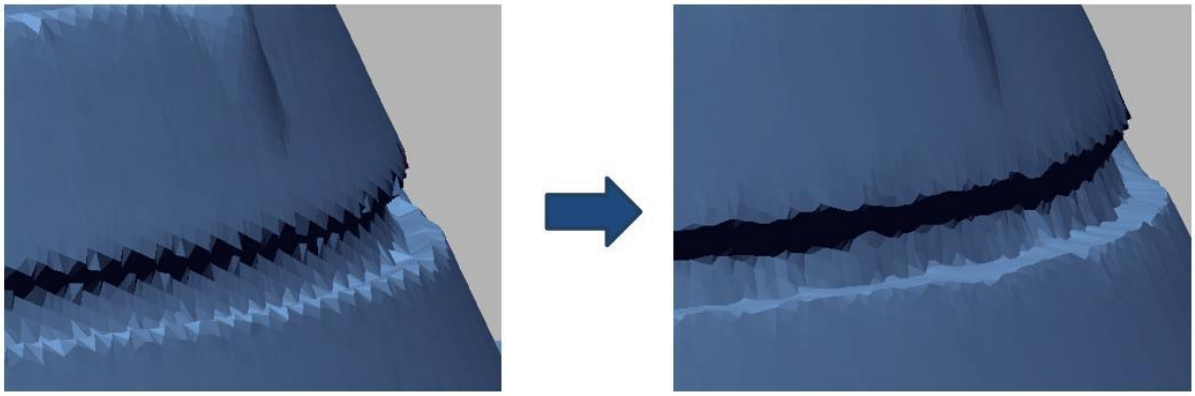


Рис. 4. Совмещение сканов.



а)



б)

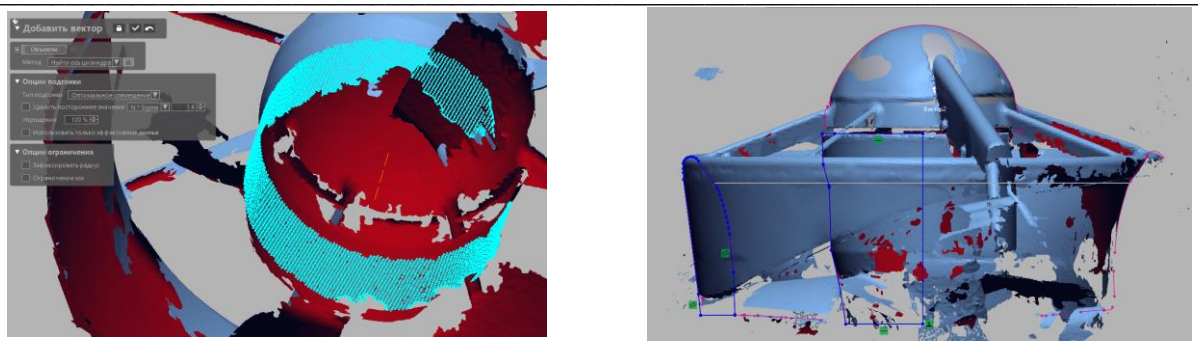
Рис. 5. Пример исправления дефектов: а — заполнение отверстий;
б — улучшение формы.

Моделирования в Geomagic Design X строится на принципах эскизирования. Для деталей с простой геометрией оптимальным выбором будет использование 2D-эскизов с последующим вытягиванием и аналогичными операциями, для построения поверхностей лучше всего использовать 3D-эскизы. В Geomagic Design X существуют три метода «Авто поверхность», «Подогнать участки поверхности» и «Вытягивание».

4. Цифровая реконструкция модели

Ниже приведено описание создания турбины подводного судна.

Для создания модели турбины использовался метод вращения, так как основная часть конструкции имеет цилиндрическую форму. Были созданы ось и профиль вращения, после чего с помощью инструмента вращения сформированы центральная и внешняя части турбины (рис. 6).



а) *б)*

Рис. 6. Пример построения модели: а — ось вращения турбины; б — построение профиля вращения.

Далее выполнено моделирование креплений между частями турбины, используя 2D-эскизы и методы вытягивания (рис. 7). Особое внимание уделено правильной симметрии модели при создании креплений с помощью инструмента «Круговой массив».

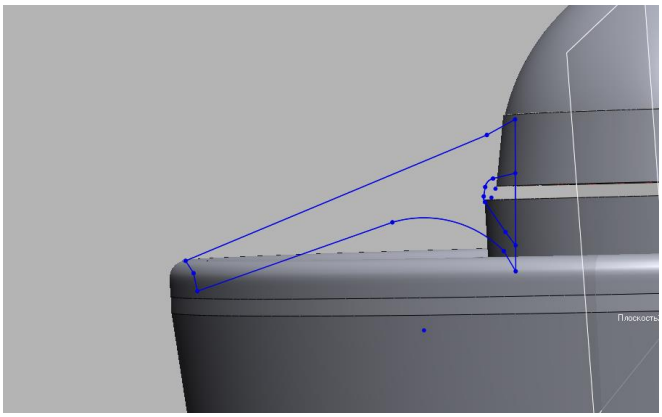


Рис. 7. Пример построения модели: эскиз крепления.

Моделирование лопастей турбины включало построение поверхности с точным повторением формы исходной детали и последующим добавлением толщины. Для завершения лопастей использовалось создание копий с помощью кругового массива.

Заключительный этап включал моделирование специфического крепления и вырезания углублений в центральной части турбины для точной установки всех компонентов.

Результат моделирования турбины представлен на рис. 8.

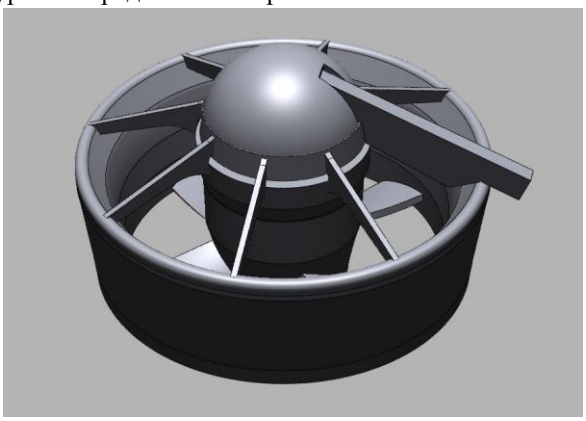


Рис. 8. Пример построения модели: модель турбины.

По такому же алгоритму были простроены другие модели.

Разработанные модели представлены на рис. 9.

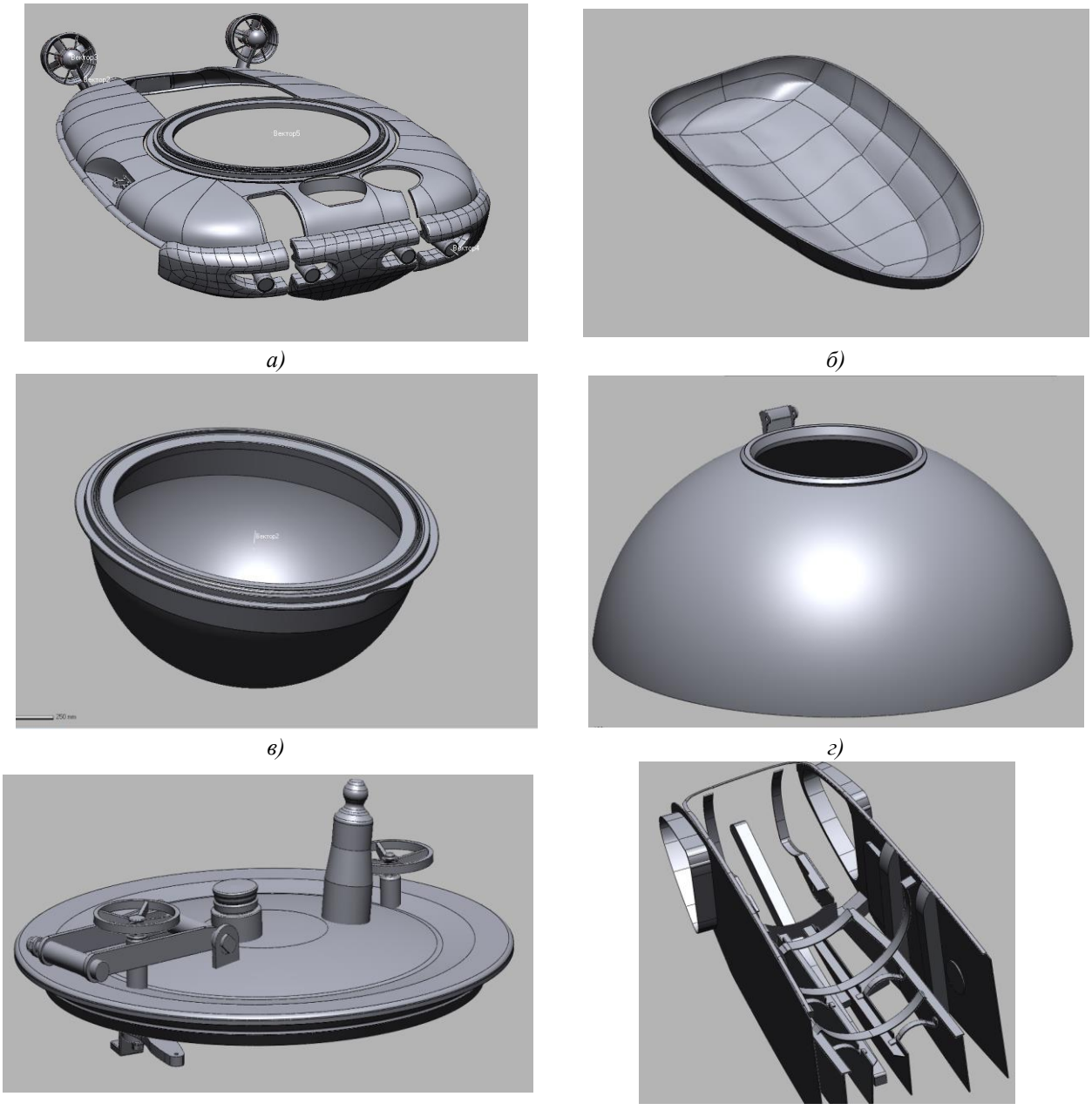


Рис. 9. Разработанные модели: а — верхняя часть корпуса; б — нижняя часть корпуса; в — чаша пилота; г — купол; д — люк; е — технический отсек.

IV. Результаты экспериментов: анализ точности моделей

Анализ точности построенных моделей проводился на основе сравнения исходного облака точек. Результат сканирования принимался за эталонный результат, а трехмерная модель сравнивалась с облаком. Проверка построения геометрии модели, вычисление отклонений между облаком и моделью позволяет оценить износ продукции, а также проверить, насколько созданный объект удовлетворяет заданным требованиям [5].

Для обеспечения корректного переноса данных для сравнения модели и облака точек использовалось программное обеспечение Geomagic Control X.

Geomagic Control X позволяет в автоматизированной форме получить подробную информацию об ошибках в виде числовых значений отклонений, а также в виде тепловой карты. Цветовое отображение расхождений между облаком и моделью может помочь идентифицировать области максимальных отклонений. Пример такой карты приведен на рис. 10.

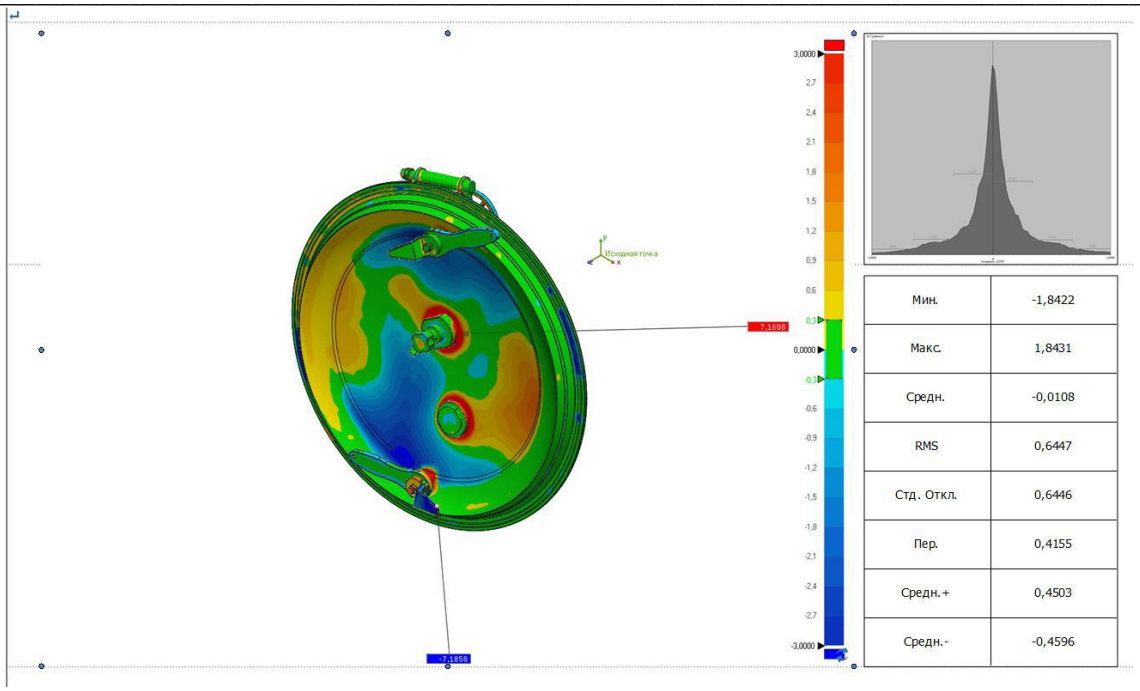


Рис. 10. Пример отчета анализа точности: анализ точности люка.

Визуализация отклонения модели от облака точек помогает принять решение о корректности восстановления объекта. Определяющим фактором может служить среднеквадратичное отклонение RMS.

Важно учитывать, что в качестве эталонного объекта для сравнения выбирался уже оптимизированный и доработанный скан. Обработка облака точек, закрытие отверстий, сглаживание или другая манипуляция вносит изменения в сами данные и в их погрешность. Поэтому на рис. 10 можно видеть, например, желтые области. Также важно понимать, что модель строился по геометрических закономерностям и не учитывает износ оборудования, который отражается при сканировании. Ко всему этому добавляется особенность восстановления объектов, в ходе которой принималось решение об упрощении модели. Так, например, в моделях полностью отсутствуют сварочные швы и другие элементы креплений. Все это также вносит долю в погрешность.

VI. Выводы и заключение

В результате были разработаны семь твердотельных моделей батискафа: верхняя и нижняя части корпуса, турбина, чаша пилота, купол, люк и технический отсек. Процесс создания трехмерной модели был автоматизирован с помощью инструментов Geomagic Design X, что существенно ускорило и упростило работу. После завершения моделирования точность каждой модели была проанализирована в Geomagic Control X, что позволило выявить проблемные участки и определить точность моделей.

Рассмотренная технология создания твердотельных моделей по результатам сканирования позволяет в короткие сроки разработать конструкторскую документацию с помощью инструментов систем автоматического проектирования, что способствует снижению как экономических, так и временных затрат.

Перспективы использования реверс-инжиниринга в судостроении включают возможность создания точных цифровых копий сложных деталей и узлов для их дальнейшего воспроизведения, анализа и модернизации. Эта технология открывает новые возможности для реконструкции, ремонта и разработки инновационных решений в судостроении.

Список литературы

(исключаются учебники, учебно-методические пособия, рукописи (диссертации, авторефераты))

1. Рубанова К.А. Стратегии применения обратного инжиниринга в условиях цифровой трансформации экономики // Human Progress. 2022. Том 9, Вып. 1. С. 13. URL: http://progresshuman.com/images/2023/Tom9_1/Rubanova.pdf. DOI 10.34709/IM.191.13. EDN AKEOLL.

2. Filinskikh, A. D. Component approach to the translation of geometric models / A. D. Filinskikh // Physics and Technology Proceedings (CPT2020) : Conference Proceedings The 8th International Scientific Conference on Computing, 09–13 ноября 2020 года. – Nizhny Novgorod: Автономная некоммерческая организация в области

информационных технологий "Научно-исследовательский центр физико-технической информатики", 2020. – P. 220-225. – DOI 10.30987/conferencearticle_5fce27710c7721.55039399. – EDN VAFVBW.

3. Устинов, А. В. Сравнительный анализ результатов трехмерного моделирования поверхности рельефа и искусственных сооружений (на примере строящейся Рогунской ГЭС) / А. В. Устинов, Д. В. Болодурин // Известия Всероссийского научно-исследовательского института гидротехники им. Б.Е. Веденеева. – 2017. – Т. 285. – С. 20-28. – EDN ZRCSSJ.

4. Глумова, Е. С. Исследование алгоритмов фильтрации облака точек для помехоустойчивой реконструкции поверхности / Е. С. Глумова, М. А. Серова // Труды Международной конференции по компьютерной графике и зрению "Графикон". – 2023. – № 33. – С. 988-998. – DOI 10.20948/graphicon-2023-988-998. – EDN WHVWNR.

5. Петров Н.В., Нозирзода Ш.С., Петрова Е.Д. Применение технологий 3D-сканирования в реверсивном инжиниринге корпусных деталей // Вестник Кузбасского государственного технического университета. 2023. № 1 (155). С. 34-41. doi: 10.26730/1999-4125-2023-1-34-41.

УДК 004:921
DOI: 10.25206/978-5-8149-3873-2-2024-929-933

Анализ методов оптимизации трехмерных сцен в webgl
Analysis of 3d scene optimization methods in webgl

И.В. Хвалина [1], В.В. Беляков [1], А.Д. Филинских [1], В.П. Хранилов [1]
[1] Нижегородский государственный технический университет им. Р.Е. Алексеева Нижний Новгород

I.V. Hvalina [1], V.V. Belyakov [1], A.D. Filinskikh [1], V.P. Khranilov [1]
[1]Nizhny Novgorod State Technical University named after R.E. Alekseev Nizhny Novgorod, Russia

Аннотация. Было проведено исследование, направленное на выявление и анализ эффективных подходов к оптимизации 3D-графики в контексте улучшения производительности веб-приложений. В частности, рассмотрены такие аспекты, как использование инстансов, LOD, PBR, батчинга и других методов. Проведение экспериментов и применение метода анализа иерархий позволило количественно оценить и выявить приоритетные методы оптимизации среди исследуемых.

Результаты проведенного эксперимента на предварительно созданных 3D-сценах показывают влияние разнообразных методик оптимизации на время загрузки страницы и общую эффективность веб-приложений. Исследование подчеркивает значимость комплексного подхода, охватывающего как подготовку 3D-моделей и текстур, так и использование возможностей современных веб-технологий для создания высокопроизводительных, визуально привлекательных трехмерных сцен.

Ключевые слова: оптимизация веб-приложений, Babylon.js, LOD, PBR, инстансы, 3D-сцена, октодеревья, WebGL, трехмерное моделирование, draw call, Нижний Новгород

Abstract. *A study was conducted to identify and analyze effective approaches to optimizing 3D graphics within the context of improving the performance of web applications. Specifically, aspects such as the use of instances, Level of Detail (LOD), Physically Based Rendering (PBR), batching, and other methods were examined. Conducting experiments and applying the Analytic Hierarchy Process allowed for the quantitative assessment and identification of priority optimization methods among those studied. The results of the experiments conducted on pre-created 3D scenes demonstrate the impact of various optimization techniques on page loading time and overall web application efficiency. The study underscores the importance of a comprehensive approach that encompasses both the preparation of 3D models and textures, as well as leveraging modern web technologies to create high-performance, visually appealing three-dimensional scenes.*

Keywords: *web application optimization, Babylon.js, LOD, PBR, instances, 3D scenes, WebGL, 3D modeling, draw call, Nizhny Novgorod.*

I. Введение

В современном мире, где скорость развития интернет-технологий стремительно растет, эффективное использование трехмерной графики в веб-приложениях приобретает особую важность. Технология WebGL, которая не требует установки дополнительных плагинов, открывает перед разработчиками широкие горизонты для создания сложных и интерактивных 3D-сцен. [1]. Несмотря на это, каждый этап разработки должен сопровождаться оптимизацией. Такой подход не только повышает общую производительность, но и существенно улучшает визуальное восприятие, оказывая тем самым положительное влияние на пользовательский опыт.

Разработка высокопроизводительных 3D-сцен в веб-ресурсах требует комплексного подхода в оптимизации, который учитывает не только визуальные характеристики, но и ресурсоемкость.

Современные технологии предоставляют различные методы оптимизации, такие как уровни детализации (LOD), использование инстансов, выбор типа материалов, батчинг и другие [3]. Эти методы позволяют значительно повысить эффективность работы веб-приложений, однако требуют тщательного анализа для их правильного применения и настройки.

Оптимизация трехмерных сцен в веб-приложениях остается сложной и многогранной задачей, требующей комплексного подхода. В этой связи особенно важно изучение существующих методов, способных обеспечить высокое качество визуализации при минимальных затратах ресурсов.

II. Постановка задачи

Целью данной статьи является анализ и применение методов оптимизации трехмерных сцен с использованием фреймворка Babylon.js, который является одним из ведущих инструментов для работы с WebGL. Производительность трехмерных сцен в веб-приложениях зависит не только от качества подготовки 3D моделей и текстур [2], но и от эффективности методов оптимизации сцен в среде WebGL.

При выборе стратегий оптимизации в веб-приложениях особенно важно учитывать:

- Оптимизация должна способствовать уменьшению времени загрузки веб-страницы, что является важным для удержания внимания и комфорта пользователей. Кроме того, скорость загрузки страницы – это важный фактор ранжирования сайта.
- Важно обеспечить высокую частоту кадров (FPS) для плавной анимации и своевременного отклика на действия пользователя.
- Оптимизация не должна существенно ухудшать визуальные характеристики, сохраняя привлекательность для пользователей и функциональное значение.

Эти критерии диктуют выбор методов оптимизации, которые должны обеспечить баланс между производительностью и визуальным качеством, минимизируя при этом ресурсоёмкость процессов.

III. Теория

Были рассмотрены некоторые из методов, применимых в WebGL, на примере фреймворка Babylon.js [5]:

- LOD (Level Of Detail) является проверенным временем способом оптимизации сцен в веб-приложениях и играх. Этот метод позволяет выбирать уровень детализации моделей в зависимости от расстояния до зрителя. В Babylon.js можно создать несколько версий объекта с разной детализацией используя метод addLODLevel [4].
- Использование инстансов позволяет создать множество копий одного объекта, используя только один вызов отрисовки. Это позволяет значительно снизить нагрузку на систему [6]. При этом копиям можно задать различный цвет, положение в пространстве и другие параметры.
- Метод MergeMeshes в Babylon.js (батчинг) объединяет несколько объектов в один меш, уменьшая таким образом количество вызовов отрисовки.
- Система Octrees (октодеревья) эффективно структурирует пространственные данные, улучшая производительность за счёт иерархического разделения пространства. Например, это позволяет более эффективно определить, какие объекты необходимо рендерить, а какие можно временно исключить [7].
- Так же одним из эффективных методов оптимизации является атлас текстур – объединение множества текстур в одну. Это снижает количество draw calls и улучшает производительность за счет уменьшения количества операций загрузки текстур.
- Такой метод, как Incremental Loading System в Babylon.js, позволяет загружать ресурсы только тогда, когда активная камера направлена на данные объекты. Что может быть незаменимо для оптимизации больших и сложных сцен [8].
- Изменение типа материалов с PBR на стандартные материалы может улучшить производительность, но PBR материалы предоставляют более высокую степень реализма [5].

В дополнение к вышеупомянутым методам Babylon.js предлагает дополнительные инструменты для детальной оптимизации и анализа 3D проектов. Одним из таких инструментов является SceneOptimizer, который адаптирует сцену под возможности и ограничения используемого оборудования, обеспечивая тем самым гибкую настройку производительности в зависимости от целевой платформы. Другой полезный инструмент — Inspector, который предоставляет разработчикам дополнительную возможность изучения и отладки сцены, выявления потенциальных проблем и узких мест производительности. Эти инструменты могут сыграть ключевую роль в процессе разработки, позволяя добиться высокой оптимизации и качества визуализации 3D сцен в веб-приложениях.

IV. Результаты экспериментов

Для оценки влияния различных методов оптимизации и типов материалов на производительность была проведена серия экспериментов. С использованием предварительно созданной сцены, представляющей Нижегородскую набережную и символ города — оленя [10]. В сцену была добавлена трава, состоящая из 4500 3D-объектов, для создания более насыщенного окружения и увеличения данных для возможности проведения экспериментов. Были интегрированы инструменты для измерения производительности, включая счётчики кадров

в секунду (FPS), времени загрузки страницы и количества вызовов отрисовки (draw calls). Эксперименты проводились на ноутбуке MSI GS63 7RD-065RU Stealth, что позволило оценить производительность сцены в реальных условиях использования.

Рис. 5. Созданная сцена при помощи Babylon.js.

После этого были применены поочередно различные настройки:

- Использован стандартный материал без использования Instances.
- С использованием Instances.
- Заменен стандартный материал на PBR.
- Применено уменьшение детализации объектов.

Усредненные результаты влияния методов оптимизации представлены в таблице (см. таб. 1.).

ТАБЛИЦА 1
ВЛИЯНИЕ МЕТОДОВ ОПТИМИЗАЦИИ

	Стандартный Материал	**Стандарт + Instances**	**PBR + Instances**	**PBR + Instances + LOD**
FPS	29	53	48	60
Скорость загрузки (мс)	6623	2459	2800	2240
Draw call	4529	29	29	29

В итоге эксперимент показал значительное влияние инстансов на ускорение загрузки страницы и снижение количества вызовов отрисовки с 4529 до всего 29 на сцене.

Результаты также демонстрируют, что применение материалов PBR оказывает отрицательное влияние на производительность, однако это влияние не является критичным, что позволяет использовать PBR-материалы с определёнными ограничениями. Заметное увеличение fps было достигнуто за счёт снижения количества детализации, что подчёркивает эффективность применения LOD-технологии и стратегии минимизации полигонов в сцене.

В рамках данной работы был рассмотрен один из распространённых сценариев для более детальной оценки методов оптимизации и определения приоритетного подхода.

Для анализа был выбран тип проекта, для которого ключевыми оценочными критериями являются: скорость загрузки страницы, количество кадров в секунду (FPS) и качество визуализации.

Анализ был сосредоточен на трех популярных методах оптимизации: использование Instances, выбор типа материалов и детализация 3D-геометрии.

Основываясь на собранных данных (таблица 2), было проведено комплексное сравнение основных альтернатив, применяя метод анализа иерархий с использованием аддитивной свертки критериев.

Было выполнено попарное сравнение критериев и альтернатив с использованием шкалы Томаса Л. Саати (см. таб. 3,4).

ТАБЛИЦА 3
СРАВНЕНИЕ АЛЬТЕРНАТИВ

Качество визуализации					**FPS**				**Скорость загрузки страницы**			
Метод	**Inst.**	**Stand.**	**PBR**	**LOD**	**Inst.**	**Stand.**	**PBR**	**LOD**	**Inst.**	**Stand.**	**PBR**	**LOD**
Inst.	1	1	0,2	3	1	0,33	0,33	0,11	1	7	9	5
Stand.	1	1	0,2	3	3	1	1	0,14	0,14	1	3	1
PBR	5	5	1	7	3	1	1	0,14	0,11	0,33	1	0,33
LOD	0,33	0,33	0,14	1	9	7	7	1	0,2	1	3	1

ТАБЛИЦА 4
СРАВНЕНИЕ КРИТЕРИЕВ

Критерии			
Альтер.	**Скорость**	**FPS**	**Визуал.**
Скорость	1	3	5
FPS	0,33	1	3
Визуал.	0,2	0,33	1

Определены веса критериев и альтернатив (таблица 5.) при помощи формулы 1.

$$X_i = \frac{b_i}{Z}$$
$$b_i = \sqrt[n]{a_1 \cdot a_2 \cdot \ldots \cdot a_n} \qquad (1)$$
$$Z = b_1 + b_2 + \cdots + b_n$$

ТАБЛИЦА 5
ВЕСА КРИТЕРИЕВ И АЛЬТЕРНАТИВ

Веса критериев	0,634	0,258	0,104
	Скорость загрузки страницы	**FPS**	**Качество визуализации**
Instances	0,153	0,051	0,676
Материал Standard	0,153	0,123	0,129
Материал PBR	0,633	0,123	0,053
LOD	0,061	0,702	0,141

В рамках расчета была выполнена проверка непротиворечивости мнения эксперта для каждой из полученных таблиц с применением отношения согласованности (формула 2).

$$\text{ОС} = \frac{\text{ИС}}{\text{СИ}} \qquad \text{ОС} > 0{,}1 \qquad \text{ИС} = \frac{|(\lambda_{max} - 4)|}{(4-1)} \qquad (2)$$

По итогам проверки согласованность соблюдена. Итоговые веса альтернатив (таблица 6) по совокупности критериев были определены с помощью аддитивной свертки критериев (формула 3).

$$W_{A_i} = \sum_{j=1}^{n} W_{A_i}(K_j) \cdot W_{k_j} \quad , \text{i} = 1..\text{k} \qquad (3)$$

ТАБЛИЦА 5
ВЕСА АЛЬТЕРНАТИВ ПО СОВОКУПНОСТИ КРИТЕРИЕВ

W_{A1} - **Instances**	0,460
W_{A2} - **Материал Standard**	0,130
W_{A3} - **Материал PBR**	0,132
W_{A4} - **LOD**	0,278

V. Выводы и заключение

По результатам можно сделать вывод. Для проектов, в которых важны все три критерия: скорость загрузки страницы, FPS, качество визуализации, применение инстансов и подобных методов оказывается особенно приоритетным методом оптимизации при наличии большого количества объектов на сцене.

В случаях, когда на сцене отсутствует большое количество объектов, рекомендуется в первую очередь минимизировать детализацию 3D-модели и окружения, сохраняя при этом баланс между визуальной привлекательностью и производительностью. Адаптируя уровень детализации и сложности эффектов к возможностям оборудования целевой аудитории.

В итоге количественные оценки основных методов оптимизации на популярном сценарии позволили выявить приоритетные методы оптимизации, а результаты, полученные при подготовке 3D-модели и окружения, позволили дополнить исследование.

Исследование подтвердило, что использование комплексного подхода, включающего различные методы оптимизации, позволяет создавать высокопроизводительные веб-приложения, сохраняя качество визуализации и улучшая пользовательский опыт.

Результаты данного исследования могут быть полезны для разработчиков, стремящихся к повышению качества и доступности своих веб-проектов с использованием WebGL.

Список литературы

1. Филинских, А. Д. Интерактивная трехмерная модель Нижегородского Кремля / А. Д. Филинских, К. С. Корсаков // Информационные технологии. – 2018. – Т. 24, № 9. – С. 582-585. – DOI 10.17587/it.24.582-585. – EDN XZOOXR.
2. Сивожелезова, А. А. Основные принципы создания 3D-моделей. Понятия и методы оптимизации в трёхмерной графике / А. А. Сивожелезова // Молодой ученый. – 2020. – № 10(300). – С. 10-15. – EDN VBPEUP.
3. Optimize Your Scene - Babylon.js Documentation. URL: https://doc.babylonjs.com/features/featuresDeepDive/scene/optimize_your_scene (дата обращения: [14.03.2024]).
4. Levels of Detail (LOD) - Babylon.js Documentation. URL: https://doc.babylonjs.com/features/featuresDeepDive/mesh/LOD (дата обращения: [14.03.2024]).
5. Using PBR Materials - Babylon.js Documentation. URL: https://doc.babylonjs.com/features/featuresDeepDive/materials/using/masterPBR (дата обращения: [14.03.2024]).
6. Instances - Babylon.js Documentation. URL: https://doc.babylonjs.com/features/featuresDeepDive/mesh/copies/instances (дата обращения: [14.03.2024]).
7. Optimize with Octrees - Babylon.js Documentation. URL: https://doc.babylonjs.com/features/featuresDeepDive/scene/optimizeOctrees (дата обращения: [14.03.2024]).
8. Incremental Loading - Babylon.js Documentation. URL: https://doc.babylonjs.com/features/featuresDeepDive/importers/incrementalLoading (дата обращения: [14.03.2024]).
9. Созданная сцена. URL: https://nn.artgl.ru/

УДК 004:921
DOI: 10.25206/978-5-8149-3873-2-2024-934-940

Оптимизация текстур в 3D-сценах с учетом расстояния визуализации
Texture optimization in 3D scenes based on visualization distance

И.В. Хвалина [1], В.В. Беляков [1], А.Д. Филинских [1], В.П. Хранилов [1]
[1] Нижегородский государственный технический университет им. Р.Е. Алексеева Нижний Новгород

I.V. Hvalina [1], V.V. Belyakov [1], A.D. Filinskikh [1], V.P. Khranilov [1]
[1]Nizhny Novgorod State Technical University named after R.E. Alekseev Nizhny Novgorod, Russia

Аннотация. В данной статье рассматривается проблема оптимизации текстур в 3D-сценах с учётом расстояния визуализации. Проведено экспериментальное исследование влияния различных разрешений текстур на качество визуализации с использованием фреймворка Babylon.js. Для исследования были созданы модели и текстуры с использованием методов ручного моделирования, текстурирования и фотограмметрии, которые затем были интегрированы в интерактивную 3D-сцену. Оценка качества отображения текстур проводилась как в статическом, так и в динамическом режимах, что позволило выявить ключевые факторы, влияющие на визуальное восприятие и производительность. Полученные результаты могут быть применены для улучшения оптимизации текстур в веб-среде и других средах с ограниченными ресурсами, что особенно актуально для разработки интерактивных веб-приложений и игр.

***Ключевые слова*:** трехмерное моделирование, оптимизация текстуры, визуализация, оптимизация 3D, текстурирование, UV-развертка, оптимизация веб-приложений, Babylon.js, LOD, WebGL, Нижний Новгород

Abstract: *This article addresses the issue of texture optimization in 3D scenes considering the visualization distance in web applications. An experimental study was conducted to examine the impact of different texture resolutions on visualization quality using the Babylon.js framework. Models and textures were created using methods of manual modeling, texturing, and photogrammetry, which were then integrated into an interactive 3D scene. The quality of texture rendering was evaluated in both static and dynamic modes, allowing for the identification of key factors affecting visual perception and performance. The findings can be applied to improve texture optimization in web environments and other resource-constrained settings, which is particularly relevant for the development of interactive web applications and games.*

Keywords: *3D modeling, texture optimization, visualization, 3D optimization, texturing, UV mapping, web application optimization, Babylon.js, LOD, WebGL, Nizhny Novgorod.*

I. Введение

Оптимизация текстур в 3D-сценах является ключевым аспектом в разработке интерактивных веб-приложений, игр и виртуальных сред. Текстуры играют важную роль в обеспечении реалистичности и эстетической привлекательности моделей, однако их чрезмерное использование может негативно сказаться на производительности, особенно в условиях ограниченных ресурсов веб-браузеров и устройств. В связи с этим, важной задачей является поиск баланса между качеством визуализации и эффективным использованием ресурсов.

Технологии, такие как LOD (Level of Detail) и Mipmapping, широко применяются для оптимизации текстур в зависимости от расстояния до объекта и уровня детализации, требуемого для текущего визуального контекста [1]. Тем не менее вопросы, связанные с выбором оптимального разрешения текстур и методов их применения в зависимости от условий просмотра пользователем, остаются актуальными и требуют дальнейших исследований.

В данной работе рассматривается оптимизация текстур для веб-сцен с использованием фреймворка Babylon.js и анализируются подходы к их адаптивному изменению или выбору определённого разрешения текстуры в статичной сцене в зависимости от расстояния визуализации.

II. Постановка задачи

Основной задачей данного исследования является разработка и экспериментальная проверка методов оптимизации текстур в 3D-сценах, отображаемых в веб-браузере, с учетом расстояния визуализации и специфики взаимодействия с моделями.

Для достижения этой цели необходимо решить следующие подзадачи:

- Создать 3D-сцену с использованием различных моделей и текстур, которые могут быть адаптированы к изменениям в расстоянии визуализации.
- Исследовать влияние различных методов создания текстур и их качества на производительность и визуальное восприятие в условиях реального времени.
- Провести экспериментальное тестирование на веб-платформе с использованием WebGL, чтобы определить оптимальные настройки текстур для различных условий просмотра и взаимодействия.
- Сформулировать рекомендации по выбору текстур и уровней детализации для веб-сцен, основываясь на результатах эксперимента.

III. Подготовка к эксперименту

1. Создание текстур и моделей

Для проведения экспериментов была создана сцена со скульптурой посвящённой сказке уроженца Нижнего Новгорода Максима Горького под названием "Воробьишко".
Для создания скульптуры в 3D было применено два метода: ручной и с применением фотограмметрии.
Для анализа текстуры в динамике персонажи скульптуры были оживлены.

В рамках исследовательской работы была разработана трехмерная сцена, посвященная скульптуре, вдохновленной сказкой "Воробьишко", автором которой является знаменитый уроженец Нижнего Новгорода, Максим Горький. Для воссоздания скульптуры в трехмерном формате были использованы два основных метода: ручное моделирование и фотограмметрия.

2. Применение фотограмметрии

Для воспроизведения скульптуры методом фотограмметрии было создано 103 фотографии, которые были загружены в программное обеспечение Agisoft Metashape Professional. Этот процесс позволил создать высокополигональную трёхмерную модель с текстурой высокого качества, что обеспечивает высокую детализацию для будущих проектов, гибкость в адаптации к необходимому разрешению и реалистичность визуального восприятия.

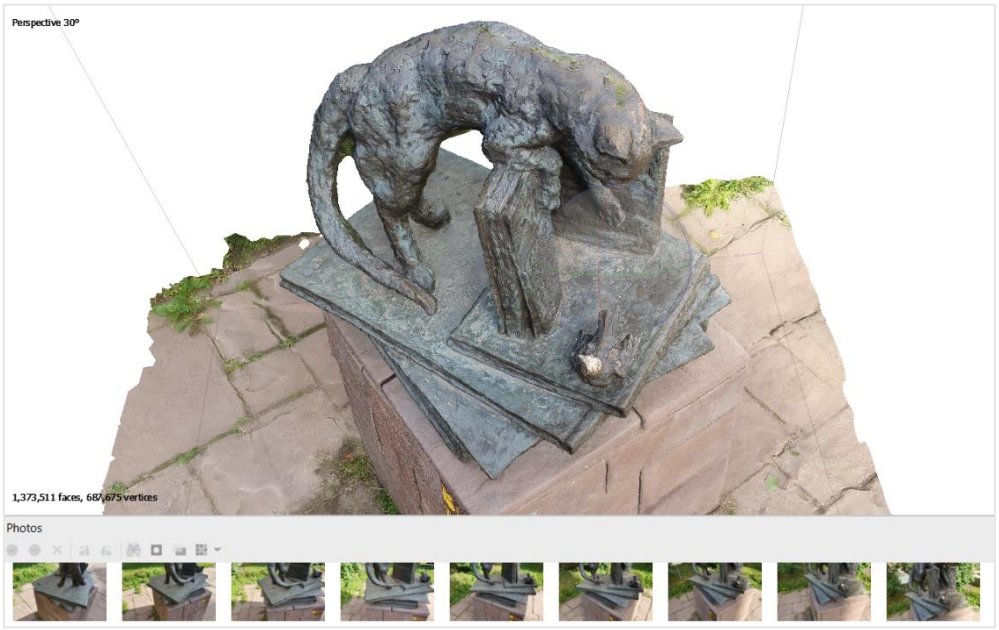


Рис. 1. Создание модели в Agisoft Metashape Professional

Далее, для оптимизации модели под будущую веб-сцену, модель была импортирована в Blender, где произведена серия работ:

- Удаление лишних объектов и персонажей, что подготовило скульптуру для новых объектов и будущей загрузки в сцену.
- Преобразование высокополигональной модели в низкополигональную при помощи ретопологии, что важно для уменьшения нагрузки и быстрой скорости загрузки страницы [3].

- Создание UV-развертки.
- Было выполнено запекание текстур. Это процесс, при котором детали и текстуры с высокополигональной модели переносятся на низкополигональную [4].

Эти шаги обеспечили подготовку модели к загрузке в веб-сцену, обеспечивая при этом баланс между качеством визуализации и требованиями к производительности. Результат отображен на рисунке 2.

Рис. 2. 3D-модель скульптуры после модификаций при помощи Blender

3. Создание персонажей

Для в дальнейшей интеграции в веб-сцену и анализа были созданы два персонажа, которые являются частью скульптуры — кошка и воробей. Разработка персонажей была выполнена в программном обеспечении Blender, с учётом требований к последующей анимации и отображению в веб-среде.

Наиболее значимые и детализированные элементы персонажей на UV-развертке были размещены с использованием большего пространства, что обеспечило более высокое разрешение текстур в наиболее важных частях модели.

Для текстурирования было применено программное обеспечение Adobe Substance 3D Painter, в котором текстуры были нарисованы вручную, что позволило достичь уникальности, соответствия концепции и высокого качества визуализации (рис. 3-4).

Рис. 3. 3D-модель кошки с текстурой и UV-развёртка.

Рис. 4. 3D-модель воробья с текстурой и UV-развёртка.

Все текстуры были экспортированы в несколько разрешений — 128x128, 256x256, 512x512, 1024x1024 и 2048x2048 пикселей — для проведения тестирования на различных уровнях детализации и оптимизации производительности при отображении в веб-браузерах.

Далее созданные модели персонажей были поставлены в позы, соответствующие композиции оригинальной скульптуры, чтобы максимально приблизить их к исходному варианту (рис. 5). После позиционирования, модели были анимированы для исследования динамического поведения текстур и проверки устойчивости визуальных качеств в движении.

Рис. 5. Созданная скульптура и персонажи

После завершения работы над моделями и их анимацией они были подготовлены для интеграции в 3D-сцену. Сцена была разработана с использованием фреймворка Babylon.js, что позволило реализовать её отображение в веб-браузере [5]. Для повышения реалистичности и создания полноценного окружения сцена была дополнена сферическим скайбоксом, созданным вместе с фотографиями, примененных при создании модели методом фотограмметрии. Это позволило реалистично вписать модель в окружение.

Рис. 6. Созданная сцена

III. Проведение эксперимента и результаты

В рамках эксперимента для каждой модели в сцене была установлена одинаковая плотность текселей для соответствия их друг другу. А также анализ проводился отдельно для каждой модели в статичном и динамичном режиме, чтобы оценить влияние анимации на визуальное восприятие текстур.

В ходе анализа были выделены критерии приемлемого и неприемлемого качества отображения текстур. Неприемлемый результат (рис. 7) включал в себя следующие аспекты:

- Значительное ухудшение качества визуализации;
- Видимые швы на текстурах;
- Присутствие других искажений, влияющих на восприятие модели.

Рис. 7. Неприемлемый результат

Приемлемый результат (рис. 5) определялся следующими характеристиками:

- Высокое качество визуализации при заданном расстоянии и размере модели;
- Отсутствие видимых швов и искажений.

В ходе эксперимента были получены значения:

1. Расстояние в процентах обеспечивающее достаточное качество визуализации.
2. Размер модели по длинной стороне в сантиметрах и пикселях при выбранном разрешении сохраняющее качество визуализации.

Для оценки качества текстур на различных расстояниях визуализации 3D-моделей был проведен экспертный анализ. В рамках эксперимента для каждой текстуры было создано пять уровней детализации LOD (Level of Detail), и расстояние до модели уменьшалось на 50% с каждым следующим уровнем детализации (см. Табл.1).

ТАБЛИЦА 1

УРОВНИ ДЕТАЛИЗАЦИИ ТЕКСТУРЫ

Разрешение текстуры	**Процент приближения расстояния визуализации**
2048x2048	150%
1024x1024	100%
512x512	50%
256x256	25%
128x128	12.5%

После достижения каждого заданного расстояния визуализации оценивалось состояние текстур для определения приемлемости визуального качества. Дополнительно, были исследованы промежуточные расстояния, на которых текстуры продолжали отображаться корректно, и полученные значения были внесены в таблицу 2.

Дополнительно в ходе измерений был проведен замер модели по длинной стороне в сантиметрах и полученные данные были преобразованы с использованием формулы 1.

$$Pix = \frac{L \cdot D}{2.54} \quad (1)$$

, где *Pix* — количество пикселей; *L* — длина в сантиметрах; *D* — плотность пикселей на дюйм (DPI); *2.54* — коэффициент преобразования из дюймов в сантиметры.

120 DPI было выбрано в качестве базового значения для расчётов. Это значение отражает современные тенденции в использовании дисплеев и соответствует характеристикам оборудования, на котором проводились экспериментальные тестирования [2].

Этот подход позволил рассчитать различные типы метрик для анализа качества текстур и их будущего применения в зависимости от условий визуализации.

ТАБЛИЦА 2
РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Модель	Расстояние (%)	Размер модели (см)	Размер модели (пиксели)
Разрешение 128x128 px			
Кошка - исходная	12.5	4.5	213
Кошка - анимация	8	2.9	137
Воробей - исходный	12.5	4	189
Воробей - анимация	10	3.2	151
Скульптура - основа	10	2.6	123
Скульптура - сборка	14	4.7	222
Разрешение 256x256 px			
Кошка - исходная	25	8.9	420
Кошка - анимация	21	7.5	354
Воробей - исходный	25	8	378
Воробей - анимация	21	6.7	316
Скульптура - основа	23	6	283
Скульптура - сборка	27	9.1	428
Разрешение 512x512 px			
Кошка - исходная	50	17.8	841
Кошка - анимация	38	13.6	642
Воробей - исходный	50	16	755
Воробей - анимация	41	13.12	620
Скульптура - основа	46	12	566
Скульптура - сборка	55	18.5	874
Разрешение 1024x1024 px			
Кошка - исходная	100	35.7	1687
Кошка - анимация	64	22.9	1081
Воробей - исходный	90	28.8	1359
Воробей - анимация	82	26.3	1242
Скульптура - основа	87	29.6	1397
Скульптура - сборка	110	36.9	1742
Разрешение 2048x2048 px			
Кошка - исходная	150	53.6	2532
Кошка - анимация	125	44.6	2108
Воробей - исходный	150	48	2267
Воробей - анимация	138	44.2	2089
Скульптура - основа	146	38	1791
Скульптура - сборка	168	56.4	2667

IV. Выводы и заключение

По результатам проведённого эксперимента было установлено, что наилучшие результаты при масштабировании и визуализации в 3D-сценах продемонстрировали текстуры, напрямую экспортированные из программы Adobe Substance 3D Painter с необходимыми настройками без дальнейших модификаций в других

программных средах. Это указывает на важность сохранения исходного качества текстур для обеспечения высокого уровня визуализации.

Исследование также показало, что постановка моделей в сложные позы и их анимация значительно увеличивают требования к разрешению текстур. В частности, более сложная поза модели кошки привела к ухудшению качества отображения текстуры по сравнению с моделью воробья, несмотря на более высокую детализацию. Это позволяет сделать вывод, что для моделей со сложными позами предпочтительнее создавать текстуры непосредственно для заданной позы, что позволит избежать искажений и излишних растяжений, улучшив таким образом общее качество визуализации и производительность.

Кроме того, для моделей, требующих высокой динамики и частых изменений поз, рекомендуется использовать текстуры с запасом по разрешению и качеству. Это позволит сохранить стабильное качество визуализации и избежать заметных артефактов при изменении масштаба или движения модели.

Текстуры, созданные методом фотограмметрии, также продемонстрировали хорошие результаты, однако уступили по качеству текстурам, созданным вручную. Это указывает на возможность использования фотограмметрии для создания высококачественных текстур, но с учётом необходимости последующей доработки для достижения наилучших визуальных характеристик.

Данное исследование предоставило данные и рекомендации по оптимизации текстур в 3D-сценах для веб-приложений. Полученные результаты позволяют более точно и быстро определять необходимый размер текстур для проектов, выбирая точное разрешение для статичной сцены или устанавливая уровни LOD для динамичной камеры и объектов в сцене.

Полученные результаты эксперимента и выводы могут быть использованы для оптимизации 3D веб-приложений и игр, обеспечивая более улучшенный пользовательский опыт.

Список литературы

1. Сивожелезова, А. А. Основные принципы создания 3D-моделей. Понятия и методы оптимизации в трёхмерной графике / А. А. Сивожелезова // Молодой ученый. – 2020. – № 10(300). – С. 10-15. – EDN VBPEUP.
2. DPI и независимые от устройства пиксели. URL: https://learn.microsoft.com/ru-ru/windows/win32/learnwin32/dpi-and-device-independent-pixels (дата обращения: [12.08.2024]).
3. Соснина, О. А. Сравнение способов создания низкополигональной 3D-модели / О. А. Соснина, А. Д. Филинских, А. С. Коротаева // Информационные технологии. – 2017. – Т. 23, № 8. – С. 564-568. – EDN ZCRLLN.
4. Вехтева, Н. А. Моделирование и анимация персонажей в рамках реконструкции исторических событий для виртуальной реальности / Н. А. Вехтева, А. А. Волков, Д. Л. Дедов // Виртуальное моделирование, прототипирование и промышленный дизайн : Материалы VI Международной научно-практической конференции, Тамбов, 16–18 октября 2019 года. Том 1, Выпуск 6. – Тамбов: Тамбовский государственный технический университет, 2020. – С. 287-291. – EDN WCUALK.
5. Созданная сцена. URL: https://nn.artgl.ru/

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-941-952

Создание виртуального стенда естественной циркуляции ФТ-101 для проведения лабораторных работ
Creation of a virtual FT-101 natural circulation stand for laboratory work

О. М. Столяров[1], Я. Д. Степанов[1], Р. М. Гунькова[1], А. А. Шутов[1]
[1]Нижегородский государственный технический университет им. Р.Е. Алексеева

O. M. Stolyarov[1], Y. D. Stepanov[1], R. M. Gunkova[1], A. A. Shutov[1]
[1]Nizhny Novgorod State Technical University n. a. R.E. Alekseev

Аннотация. В статье описаны особенности поэтапной разработки виртуального аналога стенда естественной циркуляции ФТ-101. К этапам относится: разработка трехмерных моделей, экспорт и настройка моделей и окружения, создание обучающего контента для неподготовленных пользователей, настройка логики работы проекта. Реализованный проект способствует изучению обучающих материалов, помогает обучающимся глубже изучать предмет с сохранением безопасной обстановки. Несмотря на рост затрат во время разработки такого проекта для виртуальной реальности, в будущем это может сэкономить не только время на обучение персонала, но и деньги на повсеместную установку сложного и дорогостоящего оборудования.

***Ключевые слова*:** виртуальная реальность, лабораторные работы, стенд естественной циркуляции ФТ-101, Unity, XR Interaction Toolkit

Аннотация. The article describes the step-by-step results of the development of a virtual analogue of the FT-101 natural circulation stand. The stages include: developing three-dimensional models, exporting and configuring models and environments, creating training content for untrained users, and configuring the logic of the project. As a result, the resulting project contributes to the study of training materials, helps students to learn the subject more deeply while maintaining a safe environment, despite the increase in costs during the development of such a project for virtual reality, in the future this can save not only time for staff training, but also money for the widespread installation of complex and expensive equipment.

***Ключевые слова*:** virtual reality, laboratory work, natural circulation stand FT-101, Unity, XR Interaction Toolkit

I. Введение

Применение виртуальных аналогов лабораторных установок во время обучения студентов представляет собой перспективное направление в организации образовательной деятельности.

С помощью виртуальных лабораторных стендов появляется возможность прохождения лабораторной работы большим количеством студентов одновременно, чем при работе с физическими установками, при этом каждый студент будет выполнять все необходимые взаимодействия со стендом. Благодаря этому, время прохождения лабораторных работ может сократиться.

Ещё одним из важных плюсов использования виртуальных аналогов является экономия места. Габаритные лабораторные установки занимают не только много времени на их установку, но и много пространства в аудиториях. С помощью переноса проведения таких лабораторных работ в виртуальную реальность для учебных организаций открывается доступ к созданию многофункциональных аудиторий для проведения разнообразных лабораторных работ с использованием только шлемов виртуальной реальности и возможность проведения лабораторных работ в удаленном формате.

Однако, реализация лабораторных работ в виртуальной реальности также имеет и свои особенности. Например, такой способ организации учебного процесса требует большого количества специализированного оборудования. Также, у некоторых студентов могут быть индивидуальные особенности, такие как слабая координация или плохое зрение, которые способны затруднить процесс обучения.

Таким образом, несмотря на возможные недостатки, использование виртуальных аналогов лабораторных стендов может способствовать как экономии места в аудиториях, так и увеличение скорости проведения лабораторных работ, без потери реального взаимодействия студентов с установкой.

Как раз одна из главных задач на данный момент Передовой Инженерной Школы в НГТУ им. Р. Е. Алексеева является создание вышеописанных лабораторных работ, перенос реального опыта взаимодействия с лабораторными стендами на контроллеры и шлемы виртуальной реальности

II. Постановка задачи

Исходя из вышеописанного можно сформулировать задачу: разработать лабораторный стенд, для доступности и безопасности изучения оборудования. Разрабатываемый стенд – контур естественной циркуляции ФТ-101, который может использоваться для охлаждения теплоносителя в условиях работы ядерного реактора. Так как работа напрямую с таким оборудованием может быть опасна для жизни, а реальный лабораторный стенд занимает большое пространство, необходимо повторить его в полном объеме, чтобы пользователи смогли получить полноценный опыт с такого рода установками.

III. Теория

1. Общее описание контура естественной циркуляции ФТ-101

Лабораторная работа по исследованию развития естественной циркуляции осуществляется на лабораторном стенде ФТ-101, конструкция которого изображена на рис. 1. и представляет собой замкнутый контур. Контур состоит из подъемного – обогреваемого с помощью пропускания электрического тока – и опускного – охлаждаемого водопроводной водой участков. На выходе из подъемного участка установлен прямоугольный бак, содержащий иллюминаторы.

Стенд состоит из:

1. Системы компенсации изменения объема теплоносителя в воде бака, который снабжен уровнемером и арматурой;
2. Системы заполнения и осушения стенда;
3. Системы температурного контроля, которая состоит из термопар и блока приема данных;
4. Системы энергоснабжения и управления.

Стенд ФТ-101 предназначен для исследования естественной циркуляции, возникающей при изменении плотности, а также благодаря кипению теплоносителя. Давление стенда ограничено 1 атм., но возможна модернизация до давления 6 атм.

Также возможно проведение исследования двухкомпонентной среды, которое осуществляется за счет движения жидкости с добавлением газа. Иллюминатор позволяет визуализировать процессы кипения.

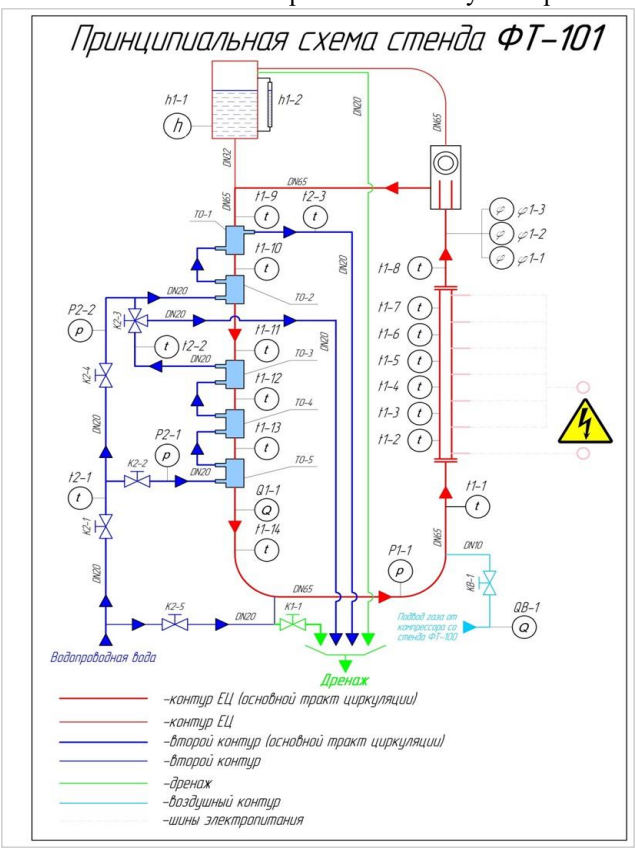


Рис. 1. Схема лабораторного стенда ФТ-101.

2. Сбор информации о лабораторном стенде

Трехмерные модели создавались в программе Blender. Перед этапом моделирования, было необходимо провести исследование установки, и её конструктивные особенности. Полученная информация позволила понять, какие компоненты должны быть использованы в виртуальном аналоге. Было сделано 195 фотографий и 2 видео в высоком качестве, чтобы получить первую тестовую трехмерную модель при помощи нейросети Luma NerF [1].

Рис. 2. Тестовая трехмерная модель.

3. Создание трехмерной модели

Были созданы в соответствии с лабораторной установкой трехмерные модели чехловых термопар, поверхностная термопара, окно визуализации, измерить РМТ490, трансформатор РОТ-25/5, холодильник труба в трубе, теплоизоляции, охлаждаемый участок и прочие трубопроводы основного тракта циркуляции, бак компенсации объема, сигнализатор уровня и другие части установки. Так же были созданы другие трехмерные модели, которые дополняли окружение лабораторного стенда (рис.3).

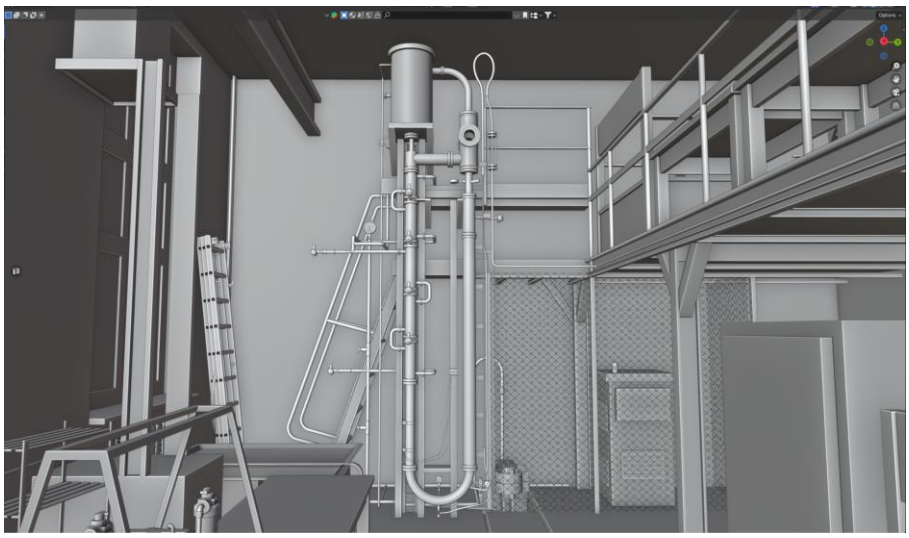


Рис. 3. Созданная трехмерная модель

4. Описание математической модели работы лабораторного стенда

Стенд естественной циркуляции ФТ-101 имеет два режима работы: стационарный и динамический. Стационарный режим работы достигается при работе установки на протяжении 5400 секунд. После этого времени температуры на всех чехловых термопарах становятся статичными. Таким образом стационарный режим работы можно описать таблицей для конкретных режимов мощностей трансформатора и расхода охлаждающей воды.

ТАБЛИЦА 1
НАБОР ДИСКРЕТНЫХ СОСТОЯНИЙ СТЕНДА ФТ-101

Мощность, кВт	Расход воды 2 контура, кг/сек	t1	t2	t3	t4	t5	t6	M
10	0.25	-	-	-	-	-	-	-
	0.5	75.9	81.6	82.1	75.6	10	14.8	0.427
	1	66.6	72.4	72.9	66	10	12.4	0.417
9.5	0.25	92.11	97.3	98.2	91.11	10	19.1	0.435
	0.5	73.4	78.8	79.8	73.3	10	14.6	0.422
	1	64.7	70.22	71.12	64.6	10	12.3	0.413
9	0.25	87.7	92.7	93.4	87.6	10	18.6	0.43
	0.5	70.9	76.1	76.7	70.5	10	14.3	0.417
	1	62.8	68	68.1	62.1	10	12.2	0.408
8.5	0.25	83.6	88.4	89.4	83.1	10	18.2	0.425
	0.5	68.4	73.3	74.1	68.1	10	14.1	0.412
	1	60.1	65.8	66.5	59.9	10	12	0.404

ПРОДОЛЖЕНИЕ ТАБЛИЦЫ 1

8	0.25	79.7	84.2	84.7	79	10	17.7	0.42
	0.5	65.9	70.6	71.1	65.2	10	13.8	0.407
	1	58.8	63.6	64.1	58.6	10	11.9	0.399
7.5	0.25	76	80.4	81.1	75.2	10	17.2	0.415
	0.5	63.3	67.8	68.4	62.6	10	13.6	0.402
	1	56.7	61.3	61.9	55.8	10	11.8	0.395
7	0.25	72.4	76.5	76.6	72.3	10	16.7	0.409
	0.5	60.8	65	65.7	60	10	13.4	0.397
	1	54.6	58.9	59.6	54	10	11.7	0.39
6.5	0.25	69	72.8	73.7	68.4	10	16.2	0.403
	0.5	58.2	62.2	62.4	57.7	10	13.1	0.392
	1	52.4	56.5	57	52	10	11.6	0.385
5.5	0.25	61.8	65.2	66	61	10	15.3	0.391
	0.5	52.8	56.2	56.9	52.1	10	12.6	0.381
	1	47.8	51.3	52.2	47.7	10	11.3	0.371
5	0.25	58.1	61.3	61.5	57.3	10	14.8	0.385
	0.5	49.9	53.1	53.4	49.5	10	12.4	0.376
	1	45.4	48.6	49.3	44.6	10	11.2	0.371

Динамический режим работы описывает изменение температур на всех термопарах в каждую единицу времени. Если строго апроксимировать изменение всех температур, то получится экспоненциальная функция вида:

$$T(\tau) = T_0 + (T_{\text{стационарное}} - T_0)(1 - e^{-\frac{\tau}{K_\tau}})$$

где T_0 – начальная температура теплоносителя, °C; $T_{\text{стационарное}}$ – температура теплоносителя, соответствующая стационарному состоянию (см.стационарное таблицу 1), °C; K_τ – постоянная времени, равная 1000 сек.

IV. Результаты экспериментов

1. Создание трехмерной модели

В процессе моделирования виртуального стенда использовался mid-poly пайплайн с модификатором weighted normals (взвешенные нормали). Face weighted normals [2] (FWN) – это метод исправления затенений 3D-моделей. Затенение можно улучшить, с помощью добавления скосов и изменения нормалей вершин так, чтобы они были перпендикулярны большим плоским многоугольникам. Это заставляет растушевывать тени по меньшим граням скоса, а не по всей модели На всех моделях есть одинарная фаска и за счет усреднения нормалей вершин получается сглаженная поверхность. Данный способ позволяет сократить время затрачиваемое на запекание текстурных карт с высокополигональной модели на низкополигональную Важный этап для сохранения

реалистичности внешнего вида стенда – текстурирование. Оно включало создание текстур для различных поверхностей виртуального аналога, а также для окружения. Для сцены были подготовлены PBR[3] текстуры. PBR — это метод затенения и рендеринга, который позволяет более точно показать, как свет взаимодействует с материалами в зависимости от их свойств. В проекте были использованы тайловые текстуры – текстуры, которые могут бесшовно соединяться между собой с любой стороны. Было создано несколько видов металлических текстур, а также текстуры для теплоизоляции, дерева, резины и пластика. Для некоторых моделей были созданы уникальные текстуры с помощью программы Substance Painter. Для создания виртуального аналога использовались текстуры albedo, normal, roughness и height, а также текстура opacity для стекла. С помощью UV-развертки текстуры наносились на соответствующие поверхности модели. Для каждой поверхности задавались материалы, определяющие их внешний вид и свойства, такие как отражение, прозрачность и блеск. Все отражения работают при помощи текстуры Roughness - она содержит один черно-белый канал, в котором чем темнее цвет, тем больше света она отражает. В Substance Painter это делается с помощью слоев материалов, которые могут быть наложены друг на друга через различные режимы наложения. Перед экспортом в игровой движок была произведена триангуляция для всех моделей в сцене. Также были выровнены масштабы и pivot point у всех объектов.

Рис. 4. Итоговый вид модели в Unity

2. Создание Unity-приложения

Создание Unity-приложения всегда начинается с установки зависимостей и нужных для работы плагинов. Одно из самых основных дополнений является API Open XR[4]. С помощью Open XR появляется возможность настроить основные компоненты для пользователя приложения: положение головы, рук. Дополнение дает продвинутый функционал для настройки кнопок контроллеров. Для добавления базового функционала взаимодействия использовался набор инструментов XR Interaction Toolkit[5]. С помощью этого инструментария можно легко добавить интерактивные предметы окружения, кнопки взаимодействия. Так же этот набор инструментов позволял экономить время на тестирование внедренного функционала и предоставлял возможность запуска проекта без наличия шлема виртуальной реальности. Одной из задачей во время создания проекта было сохранения реалистичности происходящего, чтобы пользователь получал полноценный опыт во время прохождения лабораторной работы. Для выполнения этой задачи была использована система рендеринга Universal Render Pipeline (URP)[6]. URP – это оптимизированная мультиплатформенная система отображения и рендеринга света на сценах в Unity-приложения. Она позволяет создавать мягкие и жесткие тени, блики, динамические тени, гибко настраивать свет для различных текстур и при этом сохранять оптимальную нагрузку на систему.

3. Внедрение математической модели

Во время разработки алгоритма работы контура естественной циркуляции было создано два режима работы: стационарный и динамический. Оба режима работы были записаны в скрипт для установки, который запускается после достижения пользователем определенного прогресса в выполнении лабораторной работы. Сначала была создан алгоритм, который реализует стационарный режим работы. Для этого была создана матрица значений, которая размещает данные на соответствующих термопарах при определенных значениях расхода воды и

мощности трансформатора. Значения устанавливаются только после прошествия нужного времени, но для ускорения ожидания была добавлена кнопка, которая пропускает это.

Динамический режим работы был внедрен как режим по умолчанию. Когда запускается установка, трансформатор и устанавливается мощность трансформатора, внутри установки происходит нагрев теплоносителя и начинается работа второго охлаждающего контура.

Рис. 5. Сцена приложения с добавлением URP

4. Создание обучающего контента

Для повышения степени погружения пользователя в процесс лабораторной работы и обучения новых пользователей правилам работы со шлемом и контроллерами виртуальной реальности было принято решение добавить интерактивные подсказки по ходу прохождения лабораторной работы и обучающую сцену для ознакомления пользователя с интерактивными объектами с главной сцены. Подсказки для выполнения лабораторной работы были прикреплены к соответствующим интерактивным объектам и направляющей линией с подсказкой.

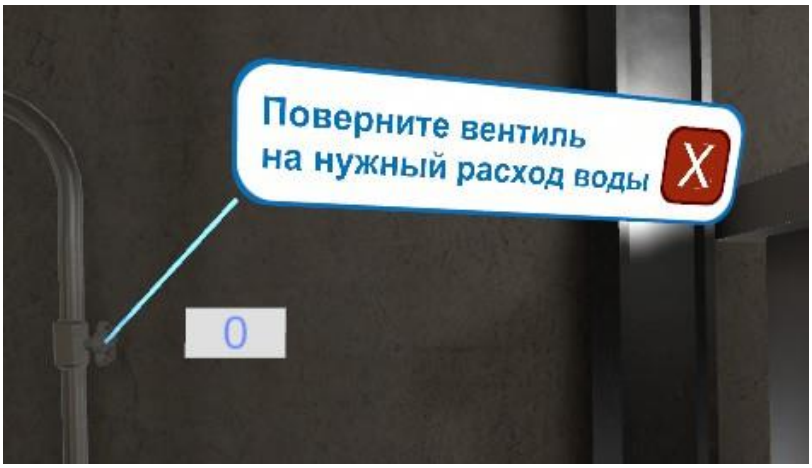


Рис. 6. Пример интерактивной подсказки

Для обучения была создана отдельная сцена с элементами из основной сцены и обучающие интерактивные подсказки, которые дают отклик пользователю о правильности его действий. Для отслеживания действий пользователя размещены триггеры на протяжении сцены. Триггеры – это невидимые геометрические фигуры, которые считывают, когда объект входит в их область и дают отклик. Таким образом получилась сцена с демонстрацией всех объектов и правил работы с лабораторным стендом и его окружением.

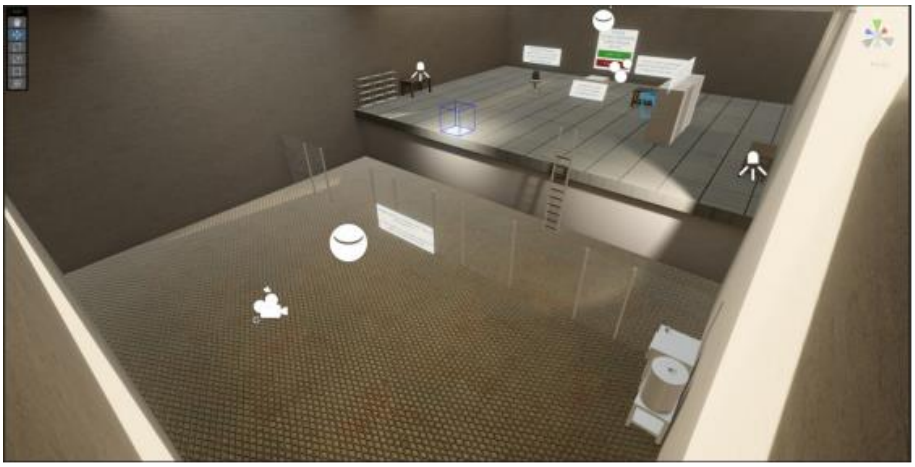


Рис. 7. Общий вид обучающей сцены

В качестве обучающего контента были добавлены интерактивные описания элементов сцены. Информационная модель лабораторной установки была сделана с помощью скрипта, который выводит полную информацию для изучения элементов лабораторного стенда. Во время исследования выбранной части установки, она подсвечивается, сосредотачивая взгляд пользователя на том, какой элемент был выбран.

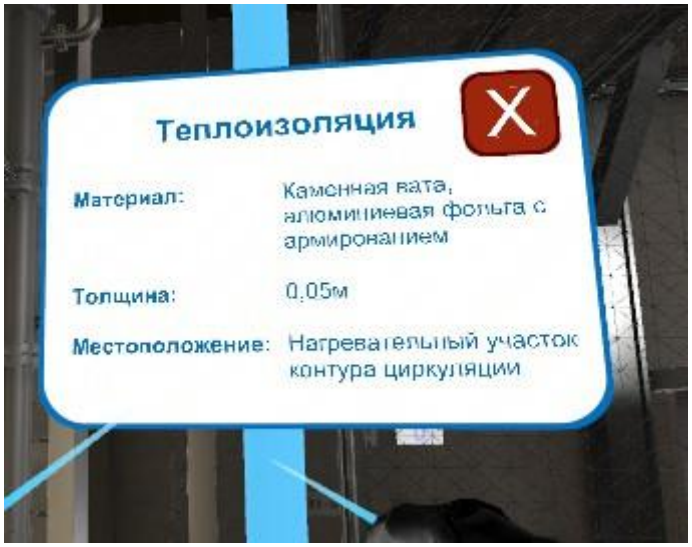


Рис. 8. Пример подсказки информационной модели

V. Обсуждение результатов

Ниже представлен процесс выполнения созданной лабораторной работы. На рис. 9, 10 продемонстрирована работа интерактивного табло для отслеживания хода лабораторной работы. На рис. 11, 12 продемонстрированы интерактивные объекты и подсказки для них, которые можно встретить по ходу выполнения работы. Ниже, на рис. 12, 13 продемонстрирован результат работы лабораторного стенда, а именно динамическое изменение температуры на термопаре под номером 3 с течением времени. Все результаты нагрева термопар отображаются на экране, на котором можно переключать каналы для отслеживания всех показателей температур.

VI. Выводы и заключение

Реализация виртуальных лабораторных стендов требует знаний как в области разработки приложений для виртуальной реальности, так и в области той лабораторной работы, которую необходимо реализовать.

Несмотря на то, что в некоторых случаях процесс создания виртуального аналога, а также первоначальная настройка оборудования для проведения лабораторных работ в виртуальной реальности может быть дороже, чем

установка реального стенда, в будущем такое приложение может помочь в экономии места в аудиториях и времени на проведение лабораторных работ.

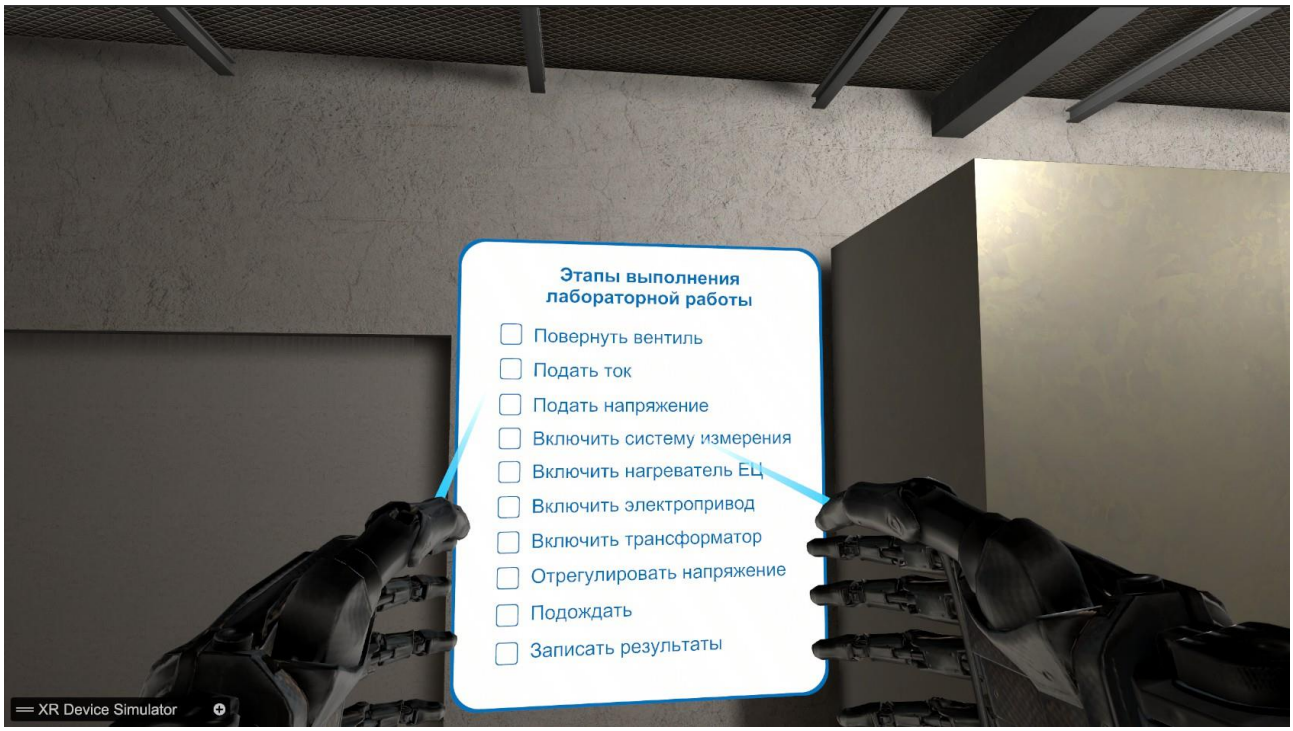


Рис. 9. Интерактивное табло без выполненных действий

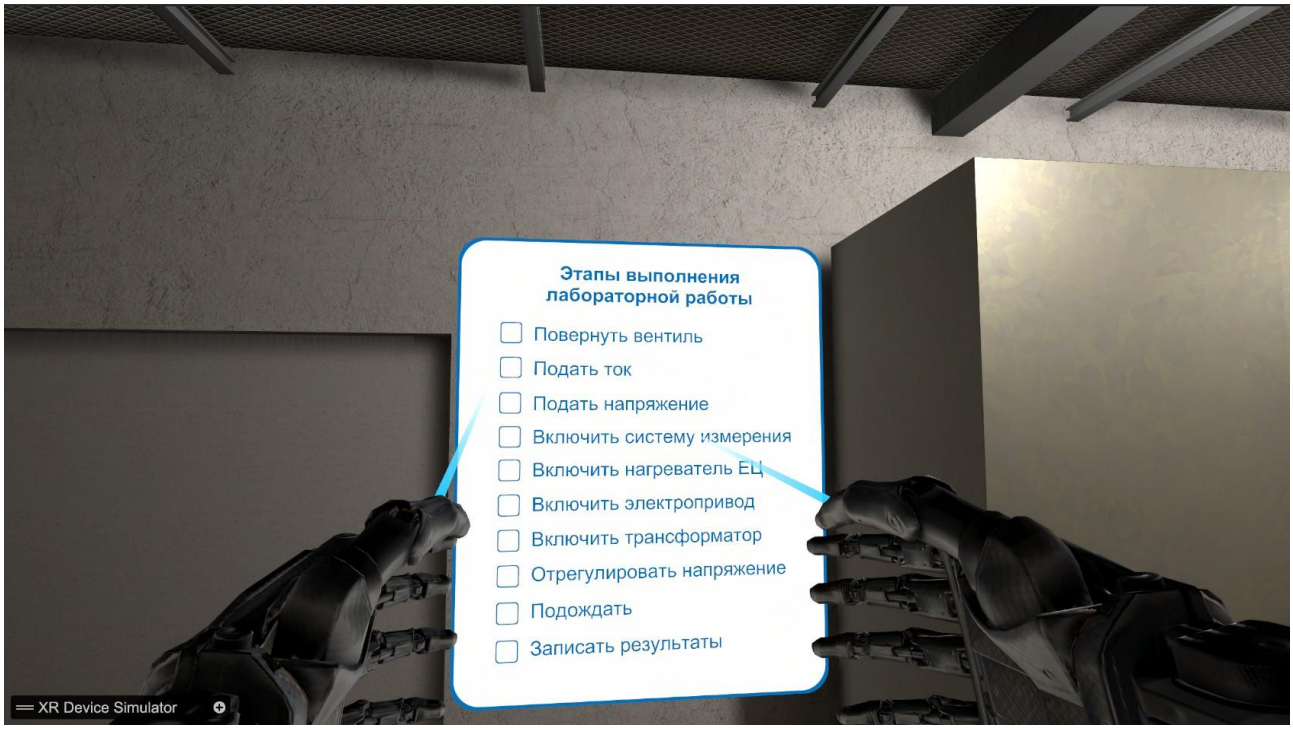


Рис. 10. Интерактивное табло с выполненным действием

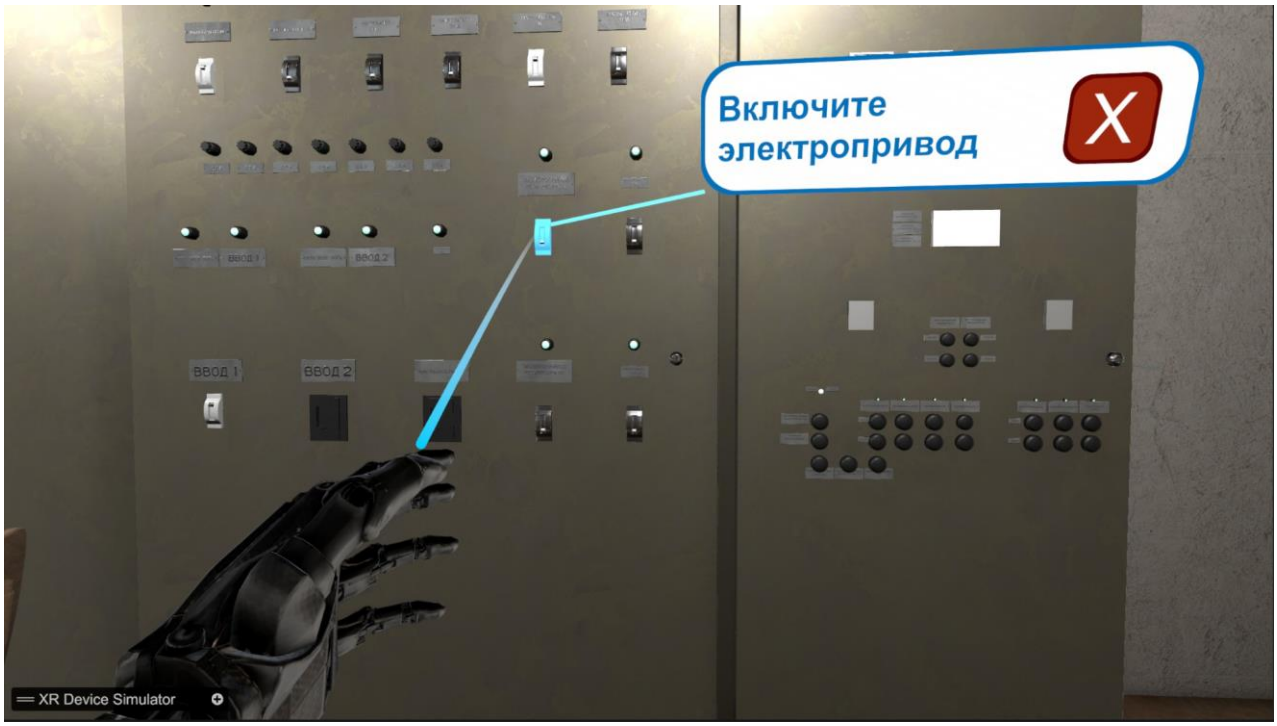


Рис. 11. Интерактивный объект в виде переключателя

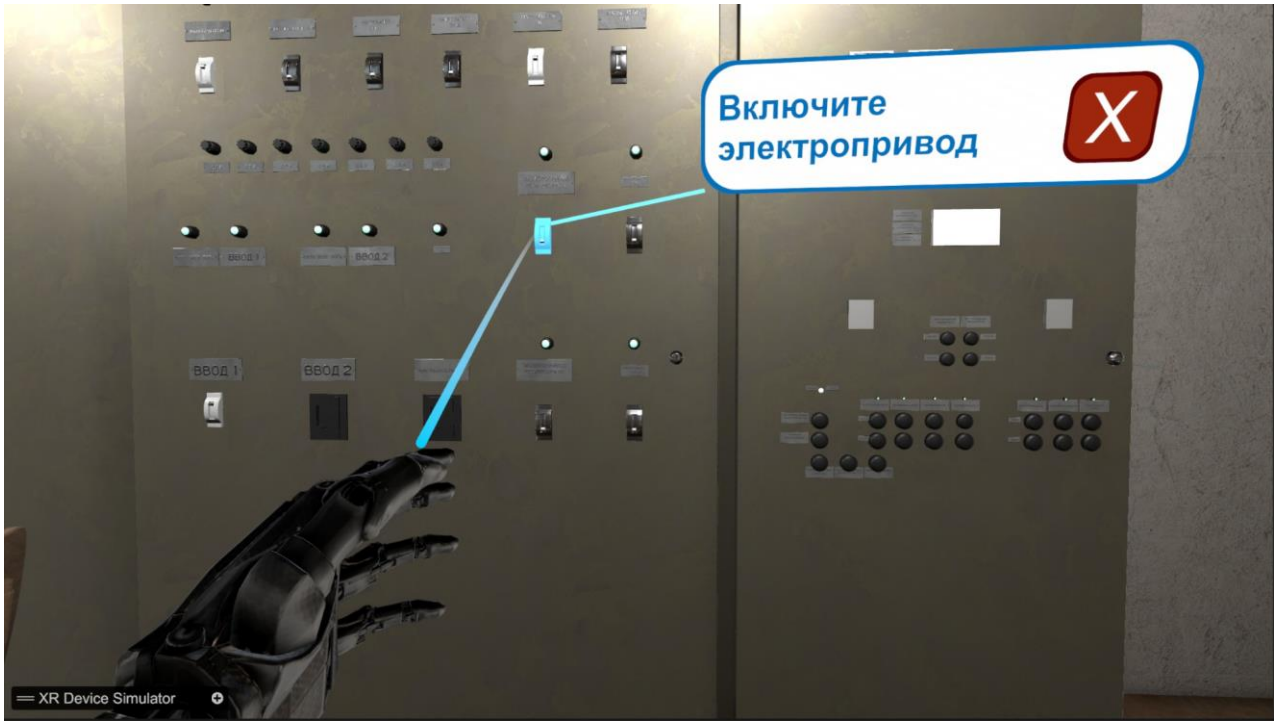


Рис. 11. Интерактивный объект в виде кнопки

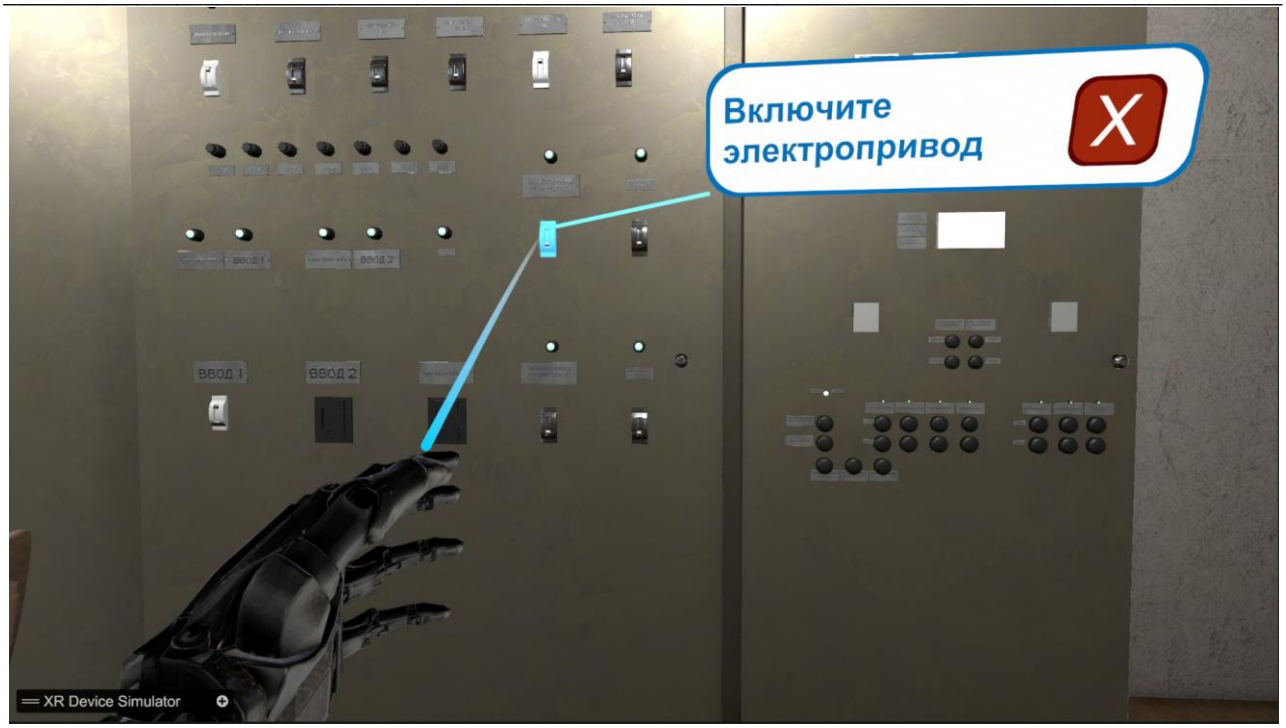


Рис. 12. Температура третьей термопары через короткий промежуток времени от начала работы

Рис. 13. Температура третьей термопары после прошествия времени от предыдущего замера

Наиболее сложный этап создания подобного проекта – это перенос математической модели лабораторной установки. Если реальный мир работает по физическим законам, то правила работы виртуального мира создаются программистами. Так, разработчики могут столкнуться с ограниченностью ресурсов оборудования во время моделирования работы сложных процессов.

Также, важным аспектом разработки является учет индивидуальных особенностей пользователей, которые могут усложнить процесс прохождения лабораторных работ.

С ростом спроса на подобные проекты неизбежно будут появляться стандарты, которые обеспечат стандартизацию всех лабораторных работ, повышая их доступность для студентов по всему миру. развивая концепцию геймификации образования.

Источник финансирования. Благодарности

Авторы благодарят Глумову Екатерину Сергеевну за помощь в редактировании статьи!

Список литературы

1. Пумарола, А., Корона, Э., Понс-Моль, Г., Морено-Ногер, Ф. Neural Radiance Fields for Dynamic Scenes // Труды конференции IEEE/CVF по компьютерному зрению и распознаванию образов (CVPR), 2021. – С. 10318-10327.
2. Меженин А.В., Извозчикова В.В. Методы построения векторов нормалей в задачах идентификации объектов // Кибернетика и программирование. 2013. № 4. С. 51-58. DOI: 10.7256/2306-4196.2013.4.9358 URL: https://nbpublish.com/library_read_article.php?id=9358
3. Фирулёва, О.Л., Филинских, А.Д. Текстурирование: PBR метод и его концепции // Труды 30-й Всероссийской научно-практической конференции по графическим информационным технологиям и системам (КОГРАФ-2020). Н.Новгород: НГТУ им. Р.Е. Алексеева, 2020. – С. 165-169.
4. Вилленав, С., Кабесас, Дж., Баерт, П., Дюпон, Ф., Лавуэ, Г. XREcho: a unity plug-in to record and visualize user behavior during XR sessions // Труды 13-й конференции ACM Multimedia Systems (MMSys '22). – С. 341-346. – DOI: https://doi.org/10.1145/3524273.3532909.
5. Гусар, Й., Хрехова, С., Кнапчиковая, Л., Трояновская, Ю. Concept of Mixed Reality Application Design for Technical Solutions // Труды 8-й Международной конференции по управлению производственными системами (MMS 2023). – С. 137-149. – DOI: https://doi.org/10.1007/978-3-030-63447-2_13.
6. Хёглунд, М., Норберг, Н. Visual and performance comparison: Physically based atmosphere method against Unity Universal Render Pipeline procedural skybox // Дипломная работа (уровень бакалавра), 2022. – Блекинге Технический Институт (BTH)
7. Самойлов, А.М., Сатаев, А.А., Блохин, А.А., Иванов, В.В. Разработка модели контура с естественной циркуляцией теплоносителя в условиях качки // Вестник ИГЭУ. – 2021. – Вып. 6. – С. 19-26.
8. Соснило, А.И. Применение технологий виртуальной реальности (VR) в менеджменте и образовании // Текст научной статьи по специальности «Компьютерные и информационные науки». – С. 158-163.
9. Андреева, О.В., Орехова, Е.Е. Верификация CFD кодов для расчета процессов естественной циркуляции // Современные проблемы науки и образования. – 2015. – № 1-1. – С. 252.
10. Гараева, Э.Р., Бикмуллина, И.И., Барков, И.А. ОСОБЕННОСТИ ПОДГОТОВКИ 3D-ОБЪЕКТОВ, СМОДЕЛИРОВАННЫХ В BLENDER, ДЛЯ ИМПОРТА В UNITY 3D // «Компьютерные и информационные науки», с. 66-74.
11. Филинских, А.Д., Глумова, Е.С., Шутов, А.А., Дудаков, Н.Ю. Научно-техническая лаборатория иммерсивных технологий как компонент программы развития передовой инженерной школы // Инновационные технологии в образовательной деятельности: материалы XXV Международной научно-методической конференции. – Н. Новгород: НГТУ им. Р.Е. Алексеева, 2023. – С. 306-312.

Секция 9

ПОДГОТОВКА КАДРОВ ВЫСШЕЙ КВАЛИФИКАЦИИ В ОБЛАСТИ ИНЖЕНЕРНОЙ ГЕОМЕТРИИ, КОМПЬЮТЕРНОЙ ГРАФИКИ И ЦИФРОВОЙ ПОДДЕРЖКИ ЖИЗНЕННОГО ЦИКЛА ИЗДЕЛИЙ

Training of Highly Qualified Personnel in the Field of Engineering Geometry, Computer Graphics and Digital Support of the Product Life Cycle

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 004:94
DOI: 10.25206/978-5-8149-3873-2-2024-954-967

Задачи цифровой поддержки проектирования аэрогидродинамики скоростных амфибийных судов
Problems of digital support for design of high-speed amphibious vessels aerohydrodynamics

А.В. Февральских
Нижегородский государственный технический университет им. Р.Е. Алексеева, Нижний Новгород, Российская Федерация

A.V. Fevralskikh
[1]Nizhny Novgorod State Technical University n.a. R.E. Alekseev, Nizhny Novgorod, Russian Federation

Аннотация. В работе обсуждаются задачи цифровой поддержки проектирования аэрогидродинамической компоновки скоростного амфибийного судна, предназначенного для всесезонной эксплуатации. Приведено описание основных задач проектирования аэрогидродинамической компоновки и подходов к их решению с использованием методов цифровой поддержки на основе численного моделирования аэрогидродинамики. Приведена математическая модель численного моделирования аэрогидродинамики, позволяющая учесть влияние экранного эффекта, деформацию водной поверхности, брызгообразовнаия и обледенения на аэрогидродинамические характеристики проектируемой компоновки судна. Даны результаты апробации и верификации элементов разрабатываемой математической модели на примерах решениях задач проектирования компоновок пассажирских экранопланов. Предложен алгоритм проектирования аэрогидродинамической компоновки скоростного амфибийного судна с использованием методов цифровой поддержки на основе численного моделирования.

***Ключевые слова*:** цифровая поддержка проектирования, численное моделирование аэрогидродинамики, скоростное амфибийное судно

Аннотация. The paper discusses the problems of digital support for the design of the aerohydrodynamic layout of a high-speed amphibious vessel intended for all-season operation. The main problems of designing the aerohydrodynamic layout and approaches to solving them using digital support methods based on numerical modeling of aerohydrodynamics are described. A mathematical model of numerical modeling of aerohydrodynamics is given, which allows taking into account the influence of the screen effect, deformation of the water surface, splash formation and icing on the aerohydrodynamic characteristics of the designed vessel layout. The results of testing and verification of the elements of the developed mathematical model are given using examples of solutions to problems of designing passenger ekranoplan layouts. An algorithm for designing the aerohydrodynamic layout of a high-speed amphibious vessel using digital support methods based on numerical modeling is proposed.

***Ключевые слова*:** design numerical support, aerohydrodynamics numerical simulation, highspeed amphibious craft.

I. Введение

Развитие удаленных территорий Российской Федерации, располагающихся в регионах Севера, Сибири, Дальнего Востока, а также острая потребность в увеличении грузооборота портов Северного морского пути обуславливает необходимость создания транспортных средств, обеспечивающих связанность промышленных и логистических объектов этих регионов с крупными городами. В качестве одного из наиболее перспективных направлений развития транспортной инфраструктуры в последние годы рассматривается внедрение скоростного флота: в некоторых регионах России уже восстановлено регулярное сообщение с использованием судов на подводных крыльях. Потребность во всесезонном транспорте в ряде регионов частично удовлетворяет эксплуатация амфибийных судов на воздушной подушке. Одно из ключевых преимуществ амфибий заключается в возможности их применения в условиях необорудованных причалов, отмелей, возможности движения по льду. Однако в качестве наиболее перспективного типа водного транспорта рассматриваются амфибийные суда на динамической воздушной подушке, или экранопланы. Преимущества экранопланов заключаются в возможности сочетания высоких значений скорости крейсерского хода (150 км/ч и более), высоких показателей аэродинамического качества и, следовательно, дальности хода, высокого коэффициента утилизации по взлетной массе, за счет движения в воздушной среде под действием экранного эффекта. Однако к настоящему времени известны лишь отдельные примеры разработки проектов экранопланов всесезонной эксплуатации (экранопланы

А.Н. Панченкова, «Буревестник-24», «Орион», «Тунгус»). Современные методы проектирования судов опираются на применение методов цифровой поддержки (МЦП) жизненного цикла, однако в практике проектирования скоростных амфибий в целом и экранопланов в частности применение современных цифровых технологий, технологий автоматизированного формирования обводов аэрогидродинамической компоновки, технологий численного моделирования аэрогидродинамики, очевидно, развито недостаточно, ввиду разрозненности сведений об основных принципах проектирования аэрогидродинамической компоновки и подходов к обеспечению противоречивых требований по эксплуатации, а также недостаточной методической проработанности методов компьютерного моделирования аэрогидродинамики с учетом экранного эффекта, деформации водной поверхности, брызгообразования и обледенения.

Некоторые основы математического моделирования аэрогидродинамики судов на воздушной подушке приведены в работах Faltinsen [1], Демешко [2], Yun [3]. Опыт математического моделирования экранной аэродинамики систематизирован в работах Маскалика [4], Rozhdestvensky [5], Панченкова [6], Белоцерковского [7]. Однако для интеграции разработанных вышеперечисленными авторами математических моделей в современный технологический базис автоматизированного проектирования, необходимо создание соответствующего комплекса алгоритмических и программных решений, что представляется нерациональным, с учетом множества ограничений этих математических моделей.

Одна из основных задач проектирования компоновки скоростного амфибийного судна заключается в определении формы обводов судна, обеспечивающих требуемые аэрогидродинамические характеристики по тяговым и несущим свойствам, показателям устойчивости и управляемости крейсерского хода. Среди известных математических подходов к определению аэрогидродинамических характеристик скоростного амфибийного судна подход на основе численного моделирования аэрогидродинамики по методу контрольных объемов представляется наиболее эффективным, поскольку результаты его применения позволяют не только определить интегральные значения сил и моментов, действующих на компоновку в целом, но и детализовать картину течения, определить вклад отдельных компоновочных элементов с учетом их взаимного влияния, а также учесть влияние деформации водной поверхности, брызгообразования, обледенения посредством прямого моделирования этих процессов. В настоящее время, компьютерные системы автоматизированного проектирования, применяемые в судостроении, имеют возможность интеграции с пакетами программ численного моделирования, однако, зачастую эта интеграция не проработана на методическом уровне, в том числе в решении задач оптимизации, оценки характеристик устойчивости и управляемости крейсерского хода, влиянии движительных устройств, брызгообразования, обледенения. Таким образом, проблемы формирования методологических основ цифровой поддержки проектирования аэрогидродинамики скоростных амфибийных судов в общем сводятся к выбору и определению математической модели аэрогидродинамики, разработке алгоритма обеспечения цифровой поддержки, разработке методик математического моделирования, оптимизации и рекомендаций по применению данных решений в проектировании.

Целью данной работы является формирование методологических основ цифровой поддержки проектирования аэрогидродинамики скоростных амфибийных судов. Для достижения цели исследования решались следующие задачи:

- определение задач цифровой поддержки проектирования аэрогидродинамики скоростного амфибийного судна с использованием методов численного моделирования;
- формирование математической модели аэрогидродинамики скоростного амфибийного судна с учетом брызгообразования и обледенения для ее применения в решении задач автоматизированного проектирования и разработка алгоритма обеспечения цифровой поддержки проектирования на ее основе;
- разработка и верификация методов численного моделирования аэрогидродинамики экраноплана с учетом работы винтокольцевого движителя, влияния деформации водной поверхности и брызгообразования;
- разработка методики численного моделирования нарастания льда на поверхности скоростного амфибийного судна в процессе движения, а также численное исследование влияния обледенения на аэродинамические характеристики крыла под действием экранного эффекта, а также на аэродинамические характеристики и характеристики продольной статической устойчивости крейсерского полета экраноплана;
- разработка и совершенствование методов оценки устойчивости и управляемости крейсерского движения скоростного амфибийного судна на основе численного моделирования;

- разработка методики автоматизированного построения компьютерной модели обводов аэрогидродинамической компоновки скоростного амфибийного судна по заданным основным размерениям компоновочных элементов;
- разработка рекомендаций по автоматизированному проектированию аэрогидродинамической компоновки скоростного амфибийного судна в виде зависимостей аэрогидродинамических характеристик элементов компоновки и компоновочных решений от относительных размерений судна и высоты движения над экраном, полученных по результатам численного моделирования, формализации протокола хранения и передачи данных об аэродинамических характеристиках проектируемых компоновочных элементов, а также методов автоматизированной оптимизации аэродинамической компоновки экраноплана.

В докладе приводится описание перечня задач цифровой поддержки проектирования аэрогидрогидинамики скоростного амфибийного судна на основе методов численного моделирования, блок-схема решения и структуры коллекции данных по результатам решения, основные положения математической модели, лежащей в основе методов решения задач, результаты верификации и апробации разрабатываемых методов.

II. Постановка задачи

Аэрогидродинамические характеристики скоростного амфибийного судна – коэффициенты сопротивления, подъемной силы, момента, их вращательные производные, а также характеристики продольной статической устойчивости – в основном определяются обтеканием обводов, разрабатываемых в процессе проектирования. Процесс формирования и обоснования (проектирования) аэрогидродинамической компоновки скоростного пассажирского амфибийного судна с сопутствующей цифровой поддержкой можно представить в виде перечня задач, представленных в таблице 1.

ТАБЛИЦА 1
ЗАДАЧИ ЦИФРОВОЙ ПОДДЕРЖКИ ПРОЕКТИРОВАНИЯ АЭРОГИДРОДИНАМИЧЕСКОЙ КОМПОНОВКИ (АГДК) С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ЦИФРОВОЙ ПОДДЕРЖКИ

№	**Задачи проектирования АГДК**	**Геометрические (компоновочные) и позиционные характеристики**	**Задачи цифровой поддержки аэрогидродинамики**	**Аэрогидродинамические характеристики**
1.	Разработка обводов несущего крыла	Определение аэродинамического профиля, формы в плане (хорды, удлинения, стреловидности по передней и задней кромке и т.д.), угла установки	Численное моделирование аэродинамики профиля и вариантов исполнения крыла	Зависимости коэффициентов сопротивления, подъемной силы и момента тангажа от угла атаки и высоты движения над экраном
2.	Разработка обводов рубки	Определение размерений рубки и профилей ее сечений, абсциссы установки крыла относительно рубки	Численное моделирование аэродинамики рубки и компоновки рубки с крылом	Зависимости коэффициентов сопротивления, подъемной силы, момента тангажа от угла тангажа и высоты движения над экраном отдельно для рубки и для рубки с крылом
3.	Разработка обводов горизонтального оперения (ГО)	Определение профиля и формы в плане оперения, определение плеча, высоты и угла установки ГО,	Численное моделирование аэродинамики горизонтального оперения, аэродинамики	Зависимости коэффициентов сопротивления, подъемной силы, момента тангажа от угла тангажа и высоты движения над

		определение хорды, размаха, углов отклонения руля высоты	компоновки крыла, рубки и горизонтального оперения, в том числе, с учетом отклоненных рулей управления на крейсерском режиме движения и в переходном процессе с вращением по тангажу	экраном отдельно для ГО и для рубки с крылом и ГО, определение скосов потока на место установки ГО, определение вращательных производных по тангажу, определение эффективности руля высоты, расчет характеристик продольной статической устойчивости, расчет вращательных производных аэродинамических характеристик по тангажу, определение характеристик переходных
4.	Разработка обводов вертикального оперения	Определение профиля, компоновки, хорды и высоты оперения, угла и плеча установки, хорды и размаха руля направления, углов отклонения руля направления	Численное моделирование аэродинамики компоновки с вертикальным оперением и отклоненным рулем направления при заданных значениях угла скольжения в поступательном и вращательном движении	Зависимости коэффициентов боковой силы, моментов крена и рыскания для заданных значений угла скольжения, угла перекладки руля направления, скорости вращения по курсу и крену
5.	Разработка обводов скегов	Определение продольных и поперечных размерений скегов, угла установки по курсу (заклинки)	Численное моделирование аэрогидродинамики скега и компоновки со скегами, в том числе, с учетом деформации (в случае исполнения скегов в виде пневмооболочек)	Зависимости коэффициентов сопротивления скегов от угла дифферента и заклинки при различной осадке и скорости движения
6.	Разработка компоновки движительно-поддувного устройства (ДПУ)	Определение обводов лопасти, количества лопастей, диаметра винта, обводов втулки и несущих пилонов, кольцевого насадка (при наличии)	Численное моделирование аэродинамики изолированного ДПУ, аэрогидродинамики ДПУ в компоновке с крылом и скегами	Зависимости коэффициентов тяги ДПУ от угла установки и высоты движения над экраном, а также от дифферента и скольжения при различных значениях угла установки и уровня осадки, определение характеристик брызгообразования в виде

				зависимостей диаметра капель от водности (массы воды капель данного размера в единице объема воздуха) при различных скоростях движения
7.	Определение интегральных аэродинамических характеристик крейсерского полета	Обводы компоновки с учетом характеристик общего расположения компоновочных элементов в продольной плоскости.	Численное моделирование аэродинамики компоновки на крейсерском режиме движения без учета и с учетом работы ДПУ. Расчет балансировочных режимов движения, крейсерских значений углов перекладки рулей	Зависимости коэффициентов сопротивления, подъемной силы и момента тангажа от угла тангажа и высоты движения над экраном
8.	Определение характеристик продольной статической устойчивости крейсерского полета	Обводы компоновки с учетом характеристик общего расположения компоновочных элементов в продольной плоскости по результатам решения задач оптимизации.	Определение оптимальных режимов движения в соответствии с заданными условиями эксплуатации, разработка траектории движения. Численное решение задачи оптимизации на пространстве множества значений углов тангажа и высоты движения над экраном по критерию максимального аэродинамического качества и критериям Иродова	Рекомендуемые значения тяги, скорости движения, угла тангажа и высоты движения над экраном для крейсерских режимов движения, значения коэффициентов сопротивления, подъемной силы и момента тангажа от угла тангажа, высоты движения над экраном и угла перекладки руля высоты.
9.	Определение характеристик боковой устойчивости крейсерского полета	Обводы компоновки с учетом характеристик общего расположения компоновочных элементов в продольной плоскости по результатам решения задач оптимизации.	Определение характеристик переходных процессов бокового возмущенного движения. Поиск оптимальных значений размерений вертикального оперения по результатам численного моделирования.	Зависимости коэффициентов боковой силы, моментов рыскания и крена компоновки, а также их вращательных производных от высоты движения над экраном при заданных величинах углов перекладки руля направления, значения величин периода и коэффициента затухания переходных процессов

10.	Определение аэрогидродинамических характеристик на режиме взлета	Обводы компоновки с учетом угла и плеча установки движительно-поддувного устройства	Численное моделирование аэрогидродинамики компоновки с учетом работы движительно поддувного устройства.	Зависимости коэффициентов аэрогидродинамического сопротивления, тяги, продольного момента, подъемной силы, а также характеристик брызгообразования при заданной взлетной массе от положения центра масс судна
11.	Исследование влияния обледенения на эксплуатационные характеристики	Обводы аэродинамической компоновки с учетом наростов льда.	Численное моделирование обледенения и аэродинамики с учетом обледенения	Зависимости аэродинамических характеристик от времени при заданной температуре, водности и распределении размеров капель, характеристики продольной статической устойчивости крейсерского полета с учетом влияния наростов льда.

Каждая из строк таблицы 1 включает итерационный процесс, блок-схема которого представлена на рисунке 1.

Рис. 1. Блок-схема решения отдельной задачи проектирования АГДК с использованием методов численного моделирования

Таким образом, в результате решения каждой из задач, перечисленных в таблице 1, формируется коллекция данных, структура которой приведена на рис.2.

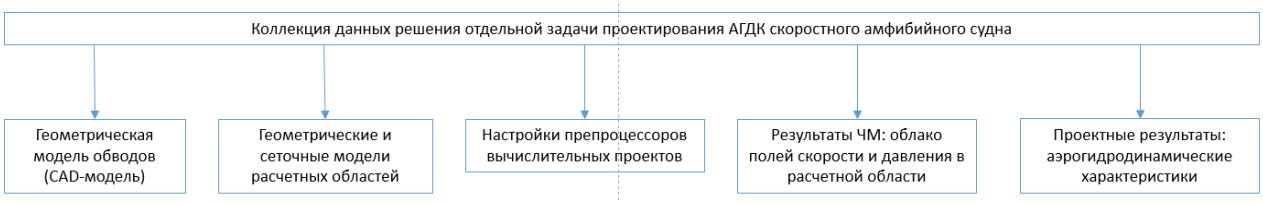


Рис. 2. Структура коллекции данных решения отдельной задачи проектирования АГДК

Однако, для достоверного определения набора данных аэродинамических характеристик необходимо в первую очередь разработать методы их расчета, то есть определить, апробировать и верифицировать математическую модель. Математическая модель аэрогидродинамики скоростного амфибийного судна с учетом брызгообразования и обледенения рассматривается в смысле задачи Коши. Характеристики течения воздуха и воды рассчитываются в некоторой окрестности обводов судна и описывается уравнениями вязкого турбулентного течения несжимаемой многофазной среды, с учетом деформации свободной поверхности, возможного перехода воды из неразрывного в дисперсное состояние и твердую фазу (лед), с учетом влияния обледенения на изменения аэродинамических характеристик. В основе математического моделирования обледенения используется модель Мессингера. Исследование характеристик устойчивости судна под действием экранного эффекта выполняется с использованием критериев Иродова. При определении аэрогидродинамических характеристик судна и элементов его компоновки принимается принцип обращения движения. Внешние граничные условия выбираются из соображений сходимости численного решения системы дифференциальных уравнений. Внутренние граничные условия – обводы аэрогидродинамической компоновки судна – принимаются в формате твердотельной компьютерной модели, которая при построении сеточной модели для моделирования по методу контрольных объемов приводится к триангулированному формату. Для построения твердотельной компьютерной модели рекомендуется использование возможностей языка Python, поскольку данный подход открывает возможности для автоматизированной оптимизации компоновки по требованиям несущих свойств и характеристик устойчивости крейсерского хода.

III. Теория

С связанное течение двухкомпонентной несжимаемой среды (воды и воздуха) с учетом свободной границы раздела сред описывается системой уравнения Навье-Стокса и уравнения неразрывности (1).

$$\begin{cases} \dfrac{\partial \vec{V}}{\partial t} = -(\vec{V}\cdot\nabla)\vec{V} + \nu_k \Delta \vec{V} - \dfrac{1}{\rho_k}\nabla p + \vec{f} \\ \dfrac{\partial}{\partial t}\alpha_k + \nabla \alpha_k \vec{V} = 0 \\ \sum\limits_{k=1}^{n} \alpha_k = 1 \end{cases} \quad (1)$$

В системе уравнений (1) $\vec{V}$ – вектор скорости, ν_k – кинематическая вязкость, ρ – плотность, p – давление, $\vec{f}$ – сумма внешних (массовых) сил, α_k – объемная доля вещества, обозначенного индексом k (воздуха или воды)

Система уравнений (1) дополняется набором граничных условий, описывающих изменение поля скорости и давления на границе раздела сред (2).

$$\begin{cases} p_2 - p_1 = \sigma\left(\dfrac{1}{R_1} + \dfrac{1}{R_2}\right) \\ \nabla \alpha_1 = -\nabla \alpha_2 \\ \rho = \alpha_1\rho_1 + \alpha_2\rho_2 \\ k_1 = -k_2 \end{cases} \quad (2)$$

В системе уравнений (2) σ – коэффициент поверхностного натяжения; p_2 и p_1, R_2 и R_1 – соответственно, величины давления и радиуса кривизны по разные стороны границы раздела водной и воздушной среды;

$k = \nabla \cdot \left(\frac{\vec{n}}{|\vec{n}|} \right)$ – кривизна границы раздела. Таким образом, задача определения аэрогидродинамических сил, действующих на компоновку экраноплана, в общем случае сводится к решению краевой задачи многофазного стратифицированного течения в заданной окрестности поверхности судна, с граничными условиями торможения на поверхности судна.

Динамика, устойчивость и управляемость крейсерского движения экраноплана описывается системой уравнения, в общем случае записанная для трех поступательных и трех вращательных степеней свободы.

$$\begin{cases} m\vec{a} = \vec{P} + \vec{D} + m\vec{g} \\ J\vec{\varepsilon} = \sum \vec{M} \end{cases} \quad (3)$$

В системе уравнений (3): m – масса судна, $\vec{a}$ – вектор ускорения, $\vec{P}$ – вектор силы тяги, $\vec{D}$ – вектор аэрогидродинамических сил (силы сопротивления, подъемной силы, силы скольжения), $\vec{g}$ – вектор ускорения свободного падения, J – момент инерции, $\vec{\varepsilon}$ – вектор углового ускорения, $\vec{M}$ – момент силы.

Конкретная форма записи системы уравнений (3) в скалярных величинах зависит от выбора системы координат и компоновки экраноплана. В частности, для экранопланов самолетной схемы с Т-образным оперением уместно раздельно рассматривать продольное и боковое движение, в то время как при исследовании устойчивости экраноплана с V-образным оперением необходимо учитывать взаимное влияние движения в трех плоскостях.

Поскольку при расчетах движения натурного скоростного амфибийного судна на режимах глиссирования, воздушной подушки и крейсерского полета выполняется условие $We \gg 1$ (число Вебера $We = \frac{\rho_a b V^2}{\sigma}$, ρ_a – плотность воздуха, b – средняя аэродинамическая хорда крыла экраноплана, или длина воздушной подушки, V – скорость движения судна), а размер брызг существенно меньше линейных размеров судна, то для расчета дисперсного течения водной среды вблизи поверхности экраноплана целесообразно использовать метод Лагранжа, в то время, как метод Эйлера на основе системы уравнений (1) целесообразно использовать для расчета течения сплошной водной среды и движения ее крупных фрагментов в воздухе.

В соответствии с подходом Лагранжа, скорость движения капли в воздушной среде $\vec{u}_p$ может быть найдена в результате решения уравнения (4) .

$$m_p \frac{d\vec{u}_p}{dt} = m_p \frac{\vec{u}_a - \vec{u}_p}{\tau_r} + m_p \frac{\vec{g}(\rho_w - \rho_a)}{\rho_w} + \vec{F}_{other} \quad (4)$$

В уравнении (4): m_p - масса капли, $\vec{u}_a$ - скорость воздушной среды, ρ_w и ρ_a - плотность воды и воздуха соответственно, $\vec{F}_{other}$ - сумма прочих сил, учитывающих давление, столкновение и вращение частиц и ряд других эффектов, τ_r - время релаксации движения капли относительно воздуха, которое показывает, как быстро скорость движения капли уравнивается со скоростью движения воздушной среды при изменении скорости воздушной среды (5).

$$\tau_r = \frac{\rho_w d_p^2}{18\mu_a} \frac{24}{C_D \operatorname{Re}_p} \quad (5)$$

В формуле (5): d_p - диаметр капли, μ_a - динамическая вязкость воздуха, C_D - коэффициент сопротивления капли, Re_p - число Рейнольдса относительного движения (6).

$$\operatorname{Re}_p = \frac{\rho_a d_p \left| \vec{u}_p - \vec{u}_a \right|}{\mu_a} \quad (6)$$

Коэффициент сопротивления капли C_D определяется с использованием соотношений (7) [8].

$$C_D = \begin{cases} 0.424 & \operatorname{Re}_p > 1000 \\ \frac{24}{\operatorname{Re}_p}\left(1 + \frac{1}{6}\operatorname{Re}^{2/3}\right) & \operatorname{Re}_p \le 1000 \end{cases} \quad (7)$$

Принимается, что процессы массопереноса в фрагменте водной пленки на поверхности экраноплана характеризуются уравнением (8) [9].

$$\rho_f \left(\frac{\partial h_f}{\partial t} + \vec{\nabla} \cdot (V_f h_f) \right) = V_\infty \cdot \alpha \cdot \beta - \dot{m}_{evap} - \dot{m}_{ice} \tag{8}$$

В уравнении (8): ρ_f - плотность воды в жидком агрегатном состоянии; h_f - высота водной пленки; V_f - местная скорость пленки; V_∞ - скорость потока на отдалении от поверхности экраноплана; $\beta = -\frac{\alpha(\vec{u}_d \cdot \vec{n})}{\alpha_\infty \|\vec{u}_{d,\infty}\|}$ - коэффициент осаждения капель на поверхности экраноплана; α и α_∞ - масса капель воды, содержащаяся в единице объема воздуха вблизи поверхности осаждения и на отдалении от нее, соответственно; $\vec{u}_d$ и $\vec{u}_{d,\infty}$ - скорость капли вблизи поверхности осаждения и на отдалении от нее, сооветственно; $\vec{n}$ - вектор нормали к поверхности осаждения; $\dot{m}_{evap}$ - скорость «испарения» жидкости с поверхности экраноплана (масса воды, потерянной с единицы площади в единицу времени); $\dot{m}_{ice}$ - скорость кристаллизации (масса воды, совершившей фазовый переход в единицу времени на единице площади).

Моделирование процесса образования льда на поверхности экраноплана основывается на применении термодинамической модели Мессингера [9], в соответствии с которой изменение тепловой энергии в водной пленке описывается уравнением (9).

$$\rho_f \left(\frac{\partial h_f c_w \tilde{T}}{\partial t} + \vec{\nabla} \cdot (V_f h_f c_w \tilde{T}) \right) = \left(c_w \tilde{T}_d + \frac{\|\vec{V_d}\|^2}{2} \right) \cdot V_\infty \cdot \alpha \cdot \beta - \tag{9}$$

$$-0.5(L_{evap} + L_{subl})\dot{m}_{evap} + (L_{fusion} - c_{ice}\tilde{T})\dot{m}_{ice}$$

В уравнении (9): $\tilde{T}$ - температура фрагмента водной пленки (определенная в градусах Цельсия относительно опорного значения - 0); c_w -удельная теплоемкость воды; $\tilde{T}_d$- температура капли перед столкновением; $\|\vec{V_d}\|$ - скорость капель вблизи поверхности; L_{evap} - удельная теплота парообразования; L_{subl} - удельная теплота сублимации; L_{fusion} - удельная теплота плавления; c_{ice} - удельная теплоемкость льда.

Для замыкания системы уравнений используются условия (10).

$$\begin{cases} h_f \geq 0 \\ \dot{m}_{ice} \geq 0 \\ h_f \tilde{T}_f \geq 0 \\ \dot{m}_{ice} \tilde{T}_f \leq 0 \end{cases} \tag{10}$$

Семейство условий (10) определяет два основных допущения теории формирования ледяных наростов, используемой в численном моделировании движения экраноплана:

принимается, что часть воды в пленке, находящейся при температуре ниже температуры кристаллизации, мгновенно переходит в твердое агрегатное состояние;

лед на обтекаемой поверхности не может существовать при температуре, выше температуры кристаллизации.

Кроме того, принимается, что в каждый отдельный момент времени система фаз «воздух» - «дисперсная вода» - «водная пленка» - «лед» и поверхность экраноплана пребывают в термодинамическом равновесии, то есть справедлива гипотеза квазистационарности.

Приведенные в настоящем разделе основные уравнения теории аэрогидродинамики экраноплана использовались в разработке методов решения задач автоматизированного проектирования экраноплана с применением технологий численного моделирования.

IV. Результаты экспериментов

Верификация результатов численного моделирования экранной аэродинамики выполнялась с использованием результатов экспериментов с моделями экранопланов в аэродинамической трубе Крыловского государственного научного центра (г. Санкт-Петербург). Результаты сопоставления аэродинамических характеристик компоновочного варианта экраноплана типа Б (рис. 1) приведены на рис. 2.

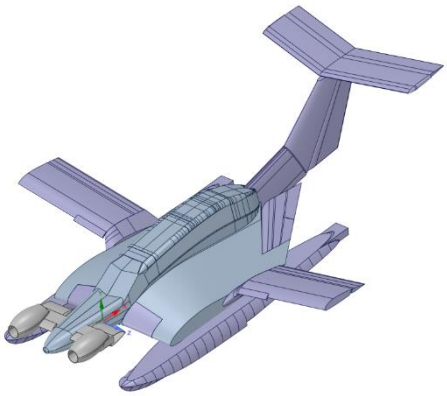


Рис. 1. Аэродинамическая компоновка экраноплана типа Б

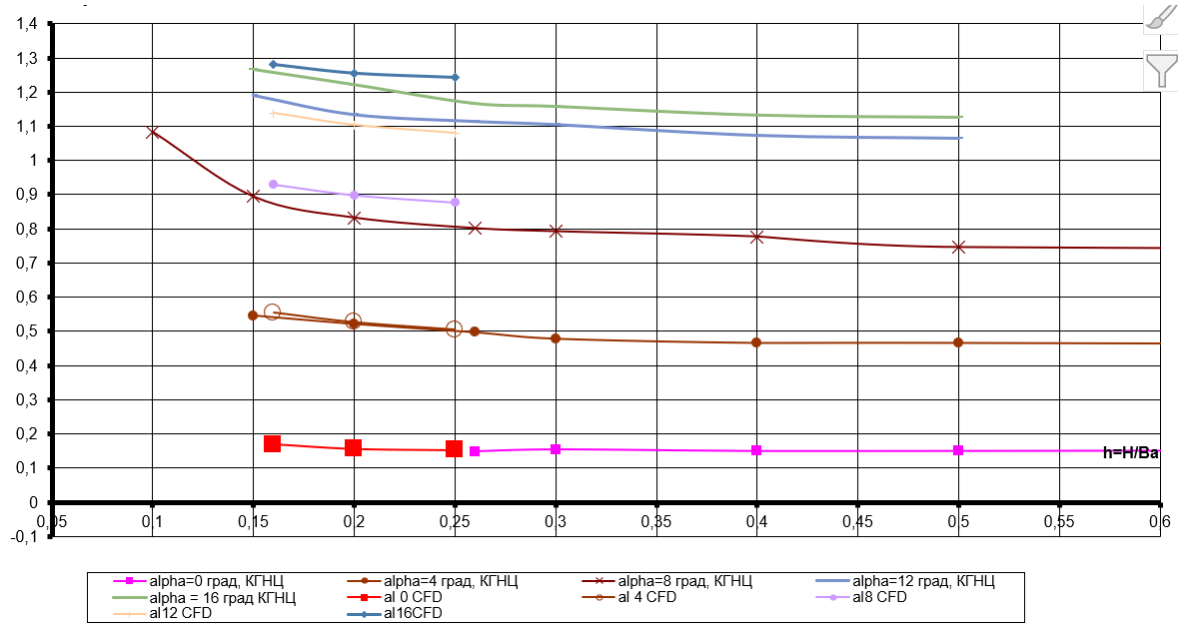


Рис. 2. Сравнение результатов определения коэффициента подъемной силы компоновочного варианта экраноплана в ходе численного решения уравнений течения (CFD) с результатами экспериментов (КГНЦ)

Верификация результатов численного моделирования брызгообразования воздушной подушки приведена на рис. 3 в сравнении с результатами экспериментов, выполненных в Крыловском государственном научном центре [10]

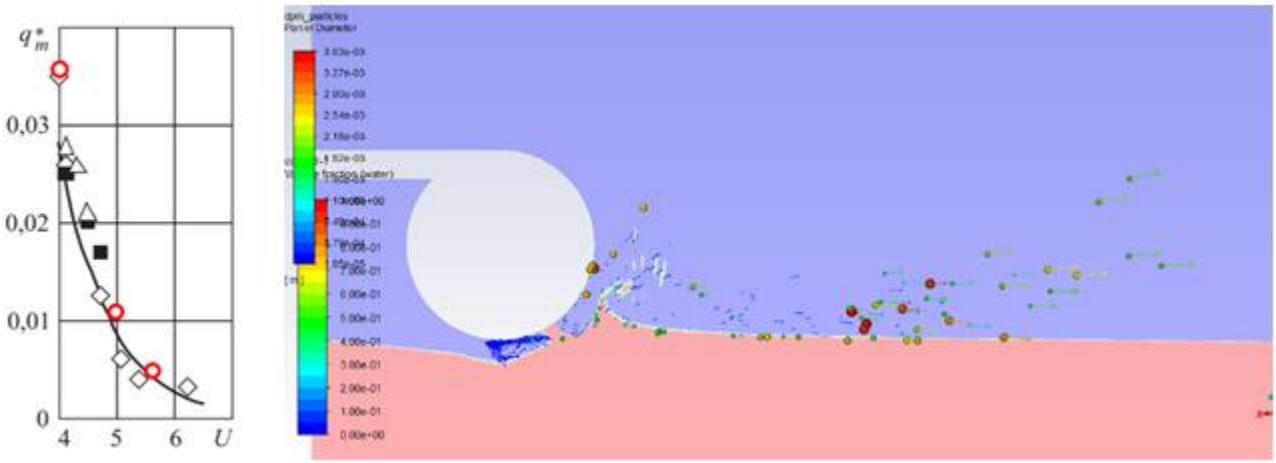


Рис. 3 Сравнение результатов численного моделирования брызгообразования воздушной подушки (красные точки на графике) с результатами экспериментов (черная линия и точки)

Результаты численного моделирования брызгообразования в виде распределения размеров капель используются в качестве исходных данных при моделировании обледенения. Результаты численного моделирования обледенения профиля в сравнении с результатами экспериментов показан на рис. 4.

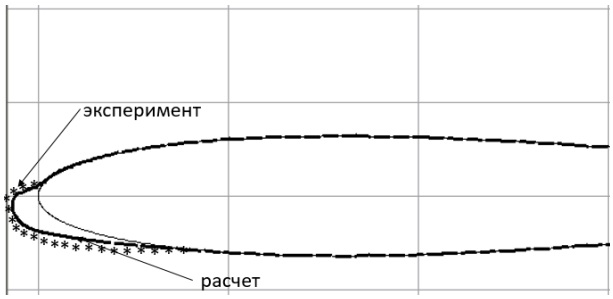


Рис. 4 – Сравнение форм ледяных наростов, полученных в ходе эксперимента и в результаты численного моделирования обледенения сегмента крыла с профилем NACA 0012

Изменение аэродинамических характеристик экраноплана вследствие обледенения, полученное по результатам численного моделирования, приведено на рис. 5 и рис. 6.

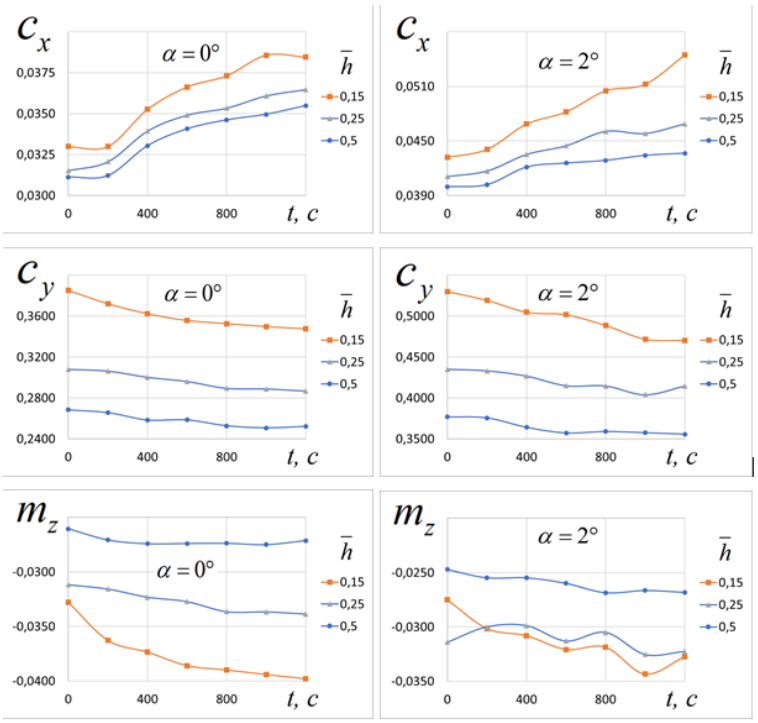


Рис. 5 – Изменение вследствие формирования ледяных наростов на поверхности экраноплана его аэродинамических характеристик: коэффициента сопротивления cx, коэффициента подъемной силы cy и коэффициента момента тангажа mz при различных значениях угла атаки α и высоты движения над экраном $\bar{h}$

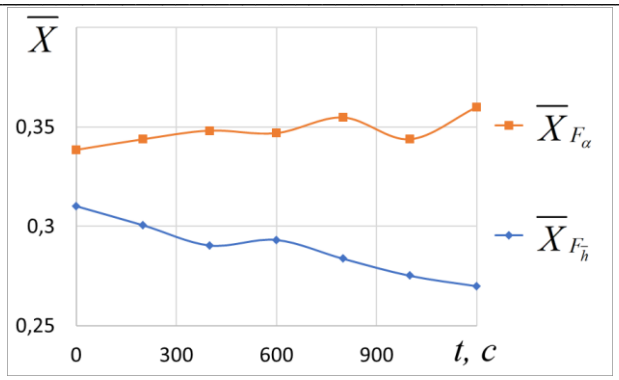


Рис. 6 – Изменение координат фокуса по высоте $\overline{X}_{F_{\bar{h}}}$ и фокуса по углу атаки $\overline{X}_{F_\alpha}$ компоновки легкого пассажирского экраноплана вследствие нарастания льда при значении угла атаки α=1° и высоты движения над экраном $\bar{h}$=0,2

Визуализация результатов численного моделирования обледенения экраноплана приведена на рис.7.

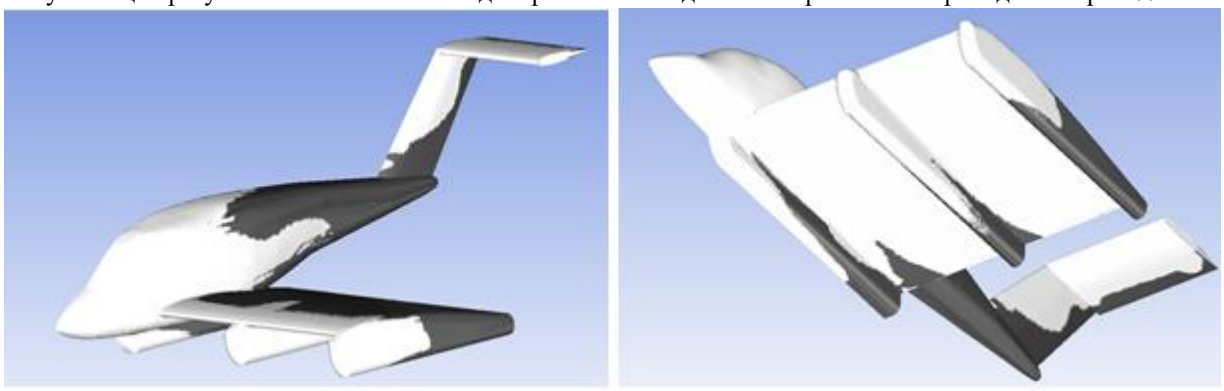


Рис. 7 – Распределение льда по поверхности экраноплана при величине угла атаки α=1° и высоте движения над экраном $\bar{h}$=0,2

V. Обсуждение результатов

Результаты численного моделирования аэрогидродинамических процессов, сопровождающих движение экраноплана, находятся в хорошем согласовании с результатами экспериментов. Изменение аэродинамических характеристик экраноплана с учетом обледенения указывает на снижение коэффициента подъемной силы, увеличение сопротивления и увеличение межфокусного расстояния, что, в соответствии с критериями Иродова, указывает на возможность потери динамической устойчивости крейсерского хода судна. В качестве мер по профилактике обледенения экраноплана может рассматриваться обеспечение запаса устойчивости за счет компоновочных решений (например, уменьшения плеча установки оперения), так и за счет установки противообледенительной системы – в зависимости массы судна и дальности его хода с учетом обледенения.

С использованием представленной математической модели и ее применения в решении задач цифровой поддержки проектирования аэрогидродинамической компоновки скоростного амфибийного судна разработан и предложен алгоритм проектирования аэрогидродинамической компоновки, представленной на рис. 8.

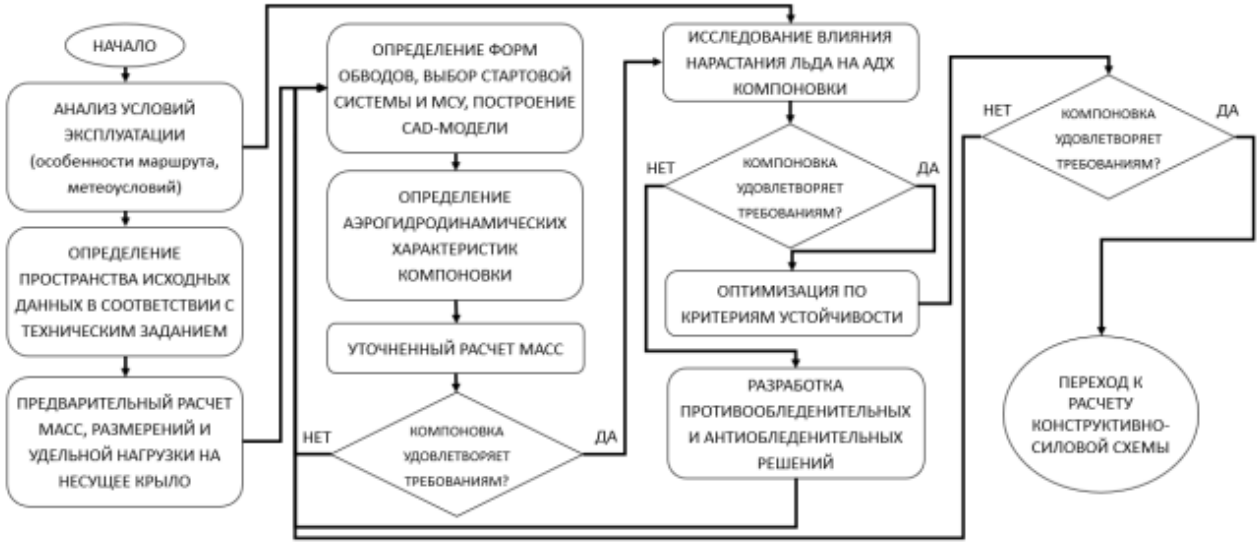


Рис. 8 – Алгоритм проектирования аэрогидродинамической компоновки скоростного амфибийного судна с использованием методов цифровой поддержки на основе численного моделирования

VI. Выводы и заключение

Сформирован перечень задач цифровой поддержки проектирования аэрогидродинамической компоновки скоростного амфибийного судна, представлен алгоритм решения с использованием методов численного моделирования аэрогидродинамики.

Разработана математическая модель аэрогидродинамики экраноплана с учетом брызгообразования и обледенения для ее применения в решении задач автоматизированного проектирования аэрогидродинамической компоновки экраноплана. В основе математической модели лежит система дифференциальных уравнений связанного турбулентного движения воздушной и водной среды - система осредненных по Рейнольдсу уравнений Навье-Стокса, замкнутых моделью турбулентности k-ω SST, дополненных уравнениями движения капель для учета процессов брызгообразования и переноса брызг, а также уравнениями массо и теплопереноса для учета процессов намерзания льда.

Выполнена разработка и апробация методики численного моделирования аэродинамики экраноплана вблизи плоского подвижного экрана: хорошее согласование результатов численного моделирования экранной аэродинамики экранопланов различных проектов с результатами испытаний моделей экранопланов в аэродинамических трубах подтверждает эффективность использования разработанной методики для определения значений коэффициентов аэродинамических сил и моментов, действующих на экраноплан на режиме крейсерского полета.

Разработана методика численного моделирования аэрогидродинамики экраноплана на основе метода контрольных объемов с учетом деформации водной поверхности, брызгообразования и брызгопереноса. Выполнена апробация элементов методики на примерах решения задачи об определении характеристик брызгообразования амфибийного судна на воздушной подушке с аэродинамической разгрузкой. Особенностью разработанного методического подхода является связанное применение метода контрольных объемов, демпфирования турбулентности связанного течения воды и воздуха со свободной границей раздела (VoF-модель), метода дискретных частиц (DPM) для моделирования движения брызг и метода пересчета массы при переходе от модели VoF к модели DPM и обратно, с учетом аэродинамики набегающего потока и возмущений со стороны воздушных винтов. Применение разработанного методического подхода позволяет определять аэрогидродинамическое сопротивление экранопланов, форму водной поверхности в области воздушной подушки и величину зазора между задней кромкой крыла и экранирующей поверхностью с учетом работы движительно-поддувного устройства, а также характеристики образования и движения брызг в единой математической модели с учетом обтекания полноразмерных обводов судна.

Разработана методика численного моделирования процесса нарастания льда на несущей поверхности, движущейся вблизи экрана. Методика основывается на автоматизированном многостадийном моделировании, на каждой из стадий выполняется расчет аэродинамики с учетом теплопереноса, расчете характеристик

осаждения капель из воздушной среды на несущей поверхности и расчете форм и массы льда, нарастающего на несущей поверхности. Перестроение сеточной модели между стадиями моделирования осуществляется с использованием инструментов программы Fluent Meshing и авторских программ.

Выполнена верификация методики численного моделирования нарастания льда на примере обтекания крыла: наблюдается удовлетворительное согласование с результатами экспериментов.

Выполнена апробация методики численного моделирования нарастания льда на примере аэродинамической компоновки легкого пассажирского экраноплана. В качестве начальных метеорологических условий моделирования приняты результаты расчета аэрогидродинамики экраноплана с учетом брызгообразования по методике, изложенной в главе 4 настоящей работы. Впервые определены зависимости коэффициента сопротивления, коэффициента подъемной силы, коэффициента момента тангажа, а также координат аэродинамических фокусов по высоте и по углу атаки экраноплана самолетной схемы с Т-образным оперением от времени в условиях нарастания льда. Показано, что скорость изменения коэффициентов аэродинамических сил и момента тангажа в условиях нарастания льда на поверхности экраноплана зависит от высоты движения над экраном. Также показано, что нарастание льда на поверхности экраноплана может способствовать увеличению межфокусного расстояния – в основном, за счет перемещения аэродинамического фокуса по высоте в направлении носа судна, что, в свою очередь, может приводить к развитию динамической неустойчивости крейсерского движения.

Предложен алгоритм проектирования аэрогидродинамической компоновки скоростного амфибийного судна с использованием разработанных методов цифровой поддержки на основе численного моделирования.

Список литературы

1. Faltinsen, O.M. Hydrodynamics of High-Speed Marine Vehicles / O.M. Faltinsen // Cambridge University Press, 2006. – 474 p.

2. Демешко, Г.В. Проектирование судов. Амфибийные СВП: Учебник. В 2-х кн. Кн. 1. / Г.В. Демешко. – СПб: Судостроение, 1992. – 269 с.

3. Yun, L. Theory and Design of Air Cushion Craft / Yun, L., Bliault, A. – London: Arnold,2000. – 632 p.

4. Маскалик, А.И. Экранопланы – транспортные суда будущего / А.И. Маскалик, Р.А. Нагапетян, А.И. Лукьянов. – Л.: Судостроение, 2013. – 352 с.–ISBN 978-5-906135-01-8.

5. Rozhdestvensky, K.V., “Aerodynamics of a lifting system in extreme ground effect” Heidelberg: Springer, 2000, p. 352.

6. Панченков, А.Н. Экспертиза экранопланов / А.Н. Панченков, П.Т. Драчев, В.И. Любимов. – Н. Новгород: Издательство «Поволжье», 2006. – 655 с.

7. Белоцерковский, С.М. Нелинейная теория крыла и ее приложения / С.М. Белоцерковский, Т.О. Аубакиров, А.И. Желанников, М.И. Ништ // Алматы: изд-во Гылым, 1997. – 448 с.

8. Liu, A.B. Modeling the Effects of Drop Drag and Breakup on Fuel Sprays / A.B. Liu, D. Mather, and R. D. Reitz // SAE Technical Paper 930072. – SAE. – 1993.

9. Messinger, B.L. Equilibrium Temperature of an Unheated Icing Surface as a Function of Air Speed / B.L. Messinger // Journal of the Aeronautical Sciences. – 1953. – № 20(1). – pp. 29-42.

10. Аносов В.Н., Дьяченко Н.В. Расчет количества воды, выносимой в атмосферу струей воздуха из воздушной подушки. Труды ЦНИИ им. акад. А.Н. Крылова, №49 (333), 2009, с.: 123-134.

УДК 514.85:004.942:621.548
DOI: 10.25206/978-5-8149-3873-2-2024-968-972

Оптимизация геометрической формы башен высотного комплекса при аэродинамическом обтекании
Optimization of the geometric shape of the towers of a high-rise complex with aerodynamic flow

П. А. Хазов
ФГБОУ ВО «Нижегородский государственный архитектурно-строительный университет», Нижний Новгород, Россия

P.A. Khazov
[1]Nizhny Novgorod State University of Architecture and Civil Engineering, Nizhny Novgorod, Russia

Аннотация. Проектирование зданий с применением альтернативных источников энергии в последние десятилетия набирает все большую популярность, в связи с чем возникает необходимость углубленного изучения факторов, влияющих на энергоэффективность проектируемых объектов. В статье приводятся результаты аэродинамических исследований комплексов высотных зданий с интегрированными ветрогенерирующими установками, основанные на компьютерном моделировании. Проведен анализ обтекания ветровым потоком трех форм башен и выполнена оценка коэффициента увеличения скорости потока при прохождении через ветрогенератор, а также сравнивались между собой возникающие в основании здания реакции от аэродинамических сил. Показано, что применение плохообтекаемых тел нецелесообразно по причине больших значений суммарных реакций и низкой вовлеченности воздуха в работу ветрогенератора при небольших изменениях направленности ветра.

***Ключевые слова*:** геометрическая оптимизация, высотный комплекс, компьютерное моделирование, аэродинамическое обтекание, ветровое воздействие, ветрогенерирующие установки.

Аннотация. The design of buildings using alternative energy sources has been gaining increasing popularity in recent decades, which necessitates an in-depth study of the factors affecting the energy efficiency of the designed facilities. The article presents the results of aerodynamic studies of complexes of high-rise buildings with integrated wind generating plants based on computer modeling. The analysis of wind flow around three forms of towers was carried out and the coefficient of increase in flow velocity when passing through a wind generator was estimated, and the reactions arising from aerodynamic forces at the base of the building were compared. It is shown that the use of poorly streamlined bodies is impractical due to the large values of total reactions and low air involvement in the operation of the wind generator with small changes in wind direction.

***Ключевые слова*:** geometric optimization, high-rise complex, computer modeling, aerodynamic flow, wind impact, wind generating plants.

I. Введение

Геометрическая оптимизация и анализ формообразования систем, взаимодействующих с аэродинамическими потоками, является первостепенной задачей при проектировании объектов. Применение геометрически невыгодных с аэродинамической точки зрения форм может приводить к колоссальным ветровым нагрузкам, увеличению массы конструкций и излишнему расходу материалов, а также к возникновению резонансных эффектов в связи с возрастанием инерционных сил в системе.

Перед современным градостроительством стоит не только задача проектирования и строительства прочных и экономически эффективных зданий, но и решение экологической проблемы, в связи с чем, в последние десятилетия популярными становятся альтернативные источники энергии – энергии ветра, солнца, недр земли и т.д. [1-4]. Одним из наиболее активно изучаемых решений по использованию энергии ветровых процессов, которые являются непрерывными и неиссякаемыми, являются специальные установки, преобразующие кинетическую энергию ветра в кинетическую энергию вращения ротора для последующей переработки в электричество – ветрогенераторы [5-7].

При проектировании объектов с интегрированными ветрогенераторами задачи поиска рациональной формы и взаимного расположения выходят на первый план, поскольку именно грамотно подобранная форма позволяет выполнять наибольший «захват» воздушных потоков, что, в свою очередь, приводит к выработке больших объемов электроэнергиии [7, 8]. Таким образом геометрическая оптимизация является ключевым инструментом для обеспечения рационального и экономически эффективного решения.

II. Постановка задачи

При оценке аэродинамических особенностей комплексов объектов сложной геометрической формы, наблюдается изменение статических характеристик каждого объекта ввиду непосредственного влияния близкорасположенных зданий на воздушные потоки. Для поиска оптимальной формы рационально принять башни с линейно-уменьшающимся по высоте профилем (рис. 1), что будет приводить к минимизации статической составляющей ветровой нагрузки, а также к уменьшению инерционных сил от ее динамической составляющей. При этом рассматриваются 3 формы башен в плане: прямоугольного очертания (форма №1), треугольного очертания (форма №2) и треугольного очертания со скругленными углами (форма №3). Ожидается, что переход между формами будет приводить к улучшению обтекаемости объектов и уменьшению ветрового воздействия, по при этом скорость потоков в пространстве вблизи окрестности ветрогенераторов между башнями не будет получать значительные изменения.

Таким образом, целью настоящего исследования является определение такой формы, которая будет менее чувствительна к небольшим изменениям направленности ветрового потока. Наиболее рациональная форма должна, в первую очередь, обеспечивать стабильный поток в области ветрогенераторов и приводить к максимизации вырабатываемой электроэнергии за весьма продолжительный промежуток времени.

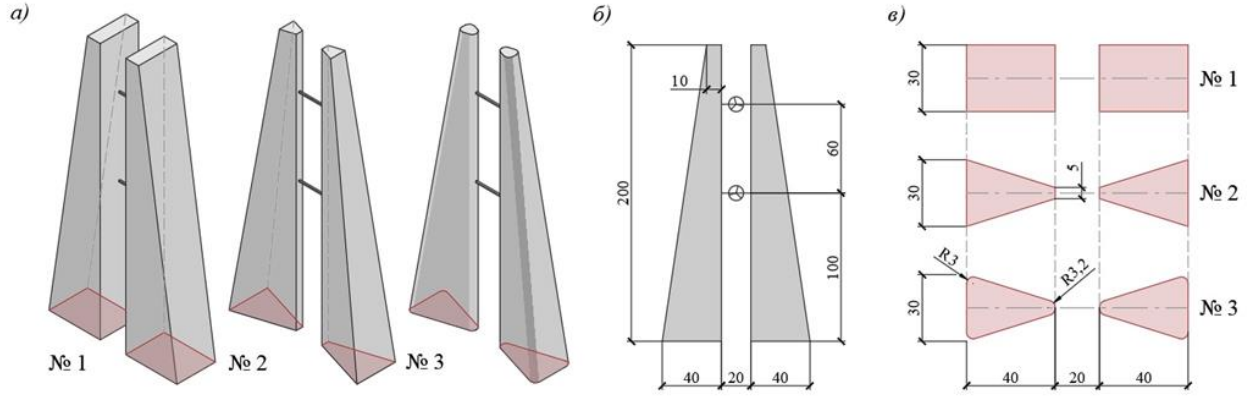


Рис. 1. Изучаемые формы (а), профиль (б) и планы башен (в) комплекса высотных зданий с интегрированными ветрогенераторами. Размеры приведены в метрах.

III. Теория

Моделирование ветрового воздействия на проектируемый комплекс зданий заключается в решении систем дифференциальных уравнений Навье-Стокса, которые позволяют описывать обтекание твердых объектов потоками жидкости или газа. При этом выполняется совместное использование уравнений движения [9, 10]:

$$\begin{cases} \dfrac{\partial u}{\partial t}+u\dfrac{\partial u}{\partial x}+v\dfrac{\partial u}{\partial y}+w\dfrac{\partial u}{\partial z}=-\dfrac{1}{\rho}\dfrac{\partial p}{\partial x}+\dfrac{\mu}{\rho}\left[\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}+\dfrac{\partial^2 u}{\partial z^2}\right] \\ \dfrac{\partial v}{\partial t}+u\dfrac{\partial v}{\partial x}+v\dfrac{\partial v}{\partial y}+w\dfrac{\partial v}{\partial z}=-\dfrac{1}{\rho}\dfrac{\partial p}{\partial y}+\dfrac{\mu}{\rho}\left[\dfrac{\partial^2 v}{\partial x^2}+\dfrac{\partial^2 v}{\partial y^2}+\dfrac{\partial^2 v}{\partial z^2}\right] \\ \dfrac{\partial w}{\partial t}+u\dfrac{\partial w}{\partial x}+v\dfrac{\partial w}{\partial y}+w\dfrac{\partial w}{\partial z}=-\dfrac{1}{\rho}\dfrac{\partial p}{\partial z}+\dfrac{\mu}{\rho}\left[\dfrac{\partial^2 w}{\partial x^2}+\dfrac{\partial^2 w}{\partial y^2}+\dfrac{\partial^2 w}{\partial z^2}\right] \end{cases},$$

и уравнения неразрывности:

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}=0,$$

в которых: u, v, w (м/с) – проекции скоростей движения частиц в элементарных потоках на оси принятой общей системы координат XYZ; t (с) – время; ρ (кг/м^3) – плотность аэродинамической среды, в условиях решаемой задачи принимается плотность атмосферы; p (Па) – давление в изучаемой точке пространства; μ/ρ (м^2/с) – коэффициент кинематической вязкости среды; μ (кг/(м·с))– коэффициент динамической вязкости [9, 10].

При этом система уравнений движения может быть представлена в матричном виде:

$$\frac{\partial \mathbf{V}}{\partial t}=-(\mathbf{V}\cdot\nabla)\mathbf{V}+\frac{\mu}{\rho}\Delta\mathbf{V}-\frac{1}{\rho}\nabla p+f,$$

где: ∇ – оператор набла; Δ – векторный оператор Лапласа; $\mathbf{V}$ – поле векторов скоростей; f – поле векторов массовых сил.

При учете сжимаемости среды ($\rho \neq const$) система уравнений движения существенно усложняется, однако данный эффект имеет место при значительно больших скоростях в задачах сверхзвукового и гиперзвукового обтекания, а в условиях аэродинамического взаимодействия статичных объектов с ветровыми потоками эффектом сжимаемости можно пренебречь [11].

Вышеупомянутые уравнения не имеют аналитического решения, в связи с чем возникает необходимость применения численных методов, при которых расчетная область представляется в виде совокупности большого числа объемов малого размера – конечных элементов, законы распределения искомых величин внутри которых (давлений, скоростей и пр.) считаются известными. В этом случае система дифференциальных уравнений упрощается до системы с большим числом алгебраических уравнений, решение которой однозначно определяется с помощью программных пакетов.

Компьютерное моделирование аэродинамического обтекания каждой модели производилось в программном продукте ANSYS SFX. Для моделирования были созданы геометрические схемы зданий в натуральную величину, из которой может быть получено большое количество изучаемых физических величин любой интересующей точке модели, также было создано окружающее пространство для последующего задания граничных условий [12]. Аэродинамические потоки моделировались с учетом нарастания скорости ветра по мере увеличения высоты. В качестве основной модели турбулентности была принята модель SST, которая наиболее подходит для описания турбулентного течения во всей расчетной области при условии моделирования расчетной сетки на вблизи поверхности здания. Для уменьшения ресурсов оперативной памяти ПК было принято допущение об отсутствии возмущений в точках, удаленных от концентраторов, которыми являются изучаемые геометрические объекты.

Моделирование аэродинамического обтекания производилось для углов атаки 0^0 и 45^0 к главному фасаду. Увеличение угла атаки выше 45^0 является нерациональным, поскольку при данных углах практически полностью прекращается вращение турбины ветрогенератора, а проектирование ветрогенераторов со следящей флюгерной системой в данной постановке задачи невозможно.

IV. Результаты экспериментов

На рис. 2 представлены результаты компьютерного моделирования при угле атаки 0^0 (перпендикулярно главному фасаду). Аналогичные исследования были проведены для всех трех моделей.

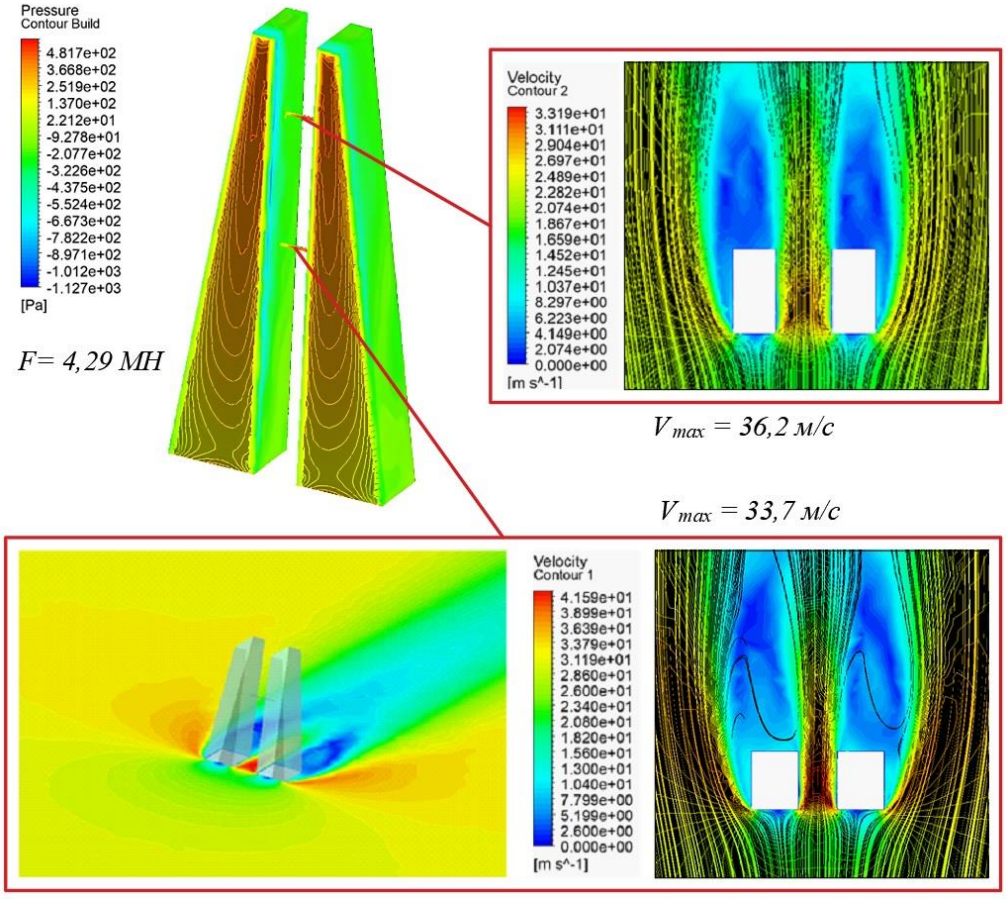


Рис. 2. Компьютерная модель высотного комплекса в ANSYS SFX, распределение давлений и скоростей в уровнях ветрогенераторов для модели №1 при угле ветровой атаки 0 градусов.

Скорости движения воздушных потоков в уровне обеих установок для каждой из рассмотренных моделей имеют близкие значения в рамках одной модели. Это достигается путем наложения двух противоположных эффектов: с одной стороны, скорости воздушного невозмущенного потока нарастают по высоте, но, в то же время, уменьшающаяся площадь здания на высоких отметках меньше способствует захвату воздушных масс. Таким образом при взаимодействии с аэродинамической средой ветрогенераторы будут задействованы в равной степени, а их детали будут подвержены одинаковому износу, вырабатывая сопоставимые объемы энергии. При анализе эффективности формы каждой из моделей важным показателем является не абсолютное значение скорости потоков, а коэффициент увеличения скорости, равный отношению скорости в окрестности ветрогенераторов к скорости на подходе к зданию – $k = v_к/v_{вх}$.

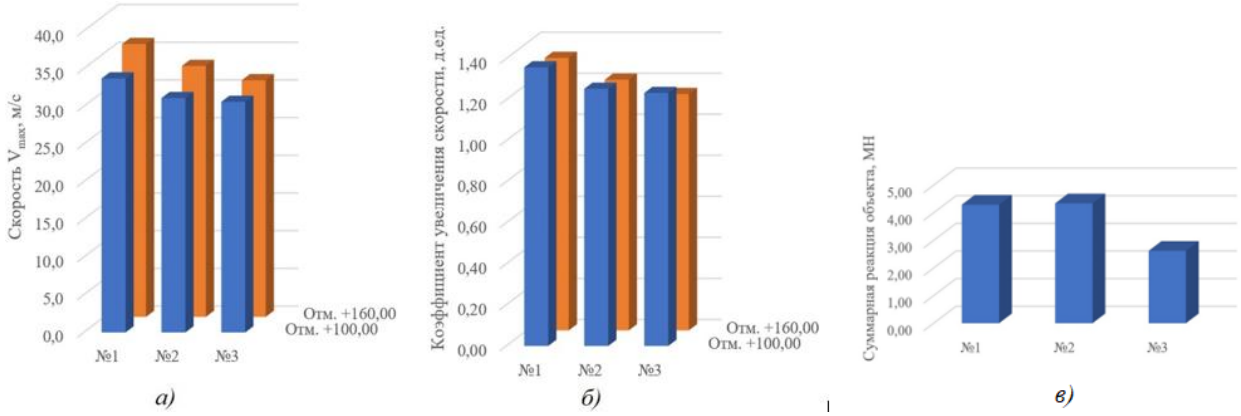


Рис. 3. Диаграммы распределения скоростей (а), коэффициентов увеличения скорости k (б) и суммарных реакций (в) для моделей при атаке потоков перпендикулярно главному фасаду.

V. Обсуждение результатов

Согласно представленным на рис. 3 диаграммам видно, что все модели имеют сравнимые показатели аэродинамических эффектов. Уровень влияния формы, критерием которого является коэффициент увеличения скорости, ожидаемо оказывается более значим на нижних уровнях в силу низких значений скорости набегающего потока.

Важным критерием для принятия той или иной формы поперечного сечения башни также является динамика коэффициента увеличения скорости движения потоков в рабочей зоне ветрогенератора в зависимости от изменения угла ветровой атаки. Так, при изменении угла ветровой атаки до 45^0 абсолютные значения скоростей потоков для модели №1 понизились в среднем на 31%, для подели №2 – на 4%, для модели №3 – на 15%. Диаграмма сравнения скоростей для каждой из трех моделей в уровнях установки ветрогенераторов для угла ветровой атаки равного 45^0 демонстрируется на рис. 4, а. Аналогичную динамику показывает и изменение коэффициентов увеличения скорости ветровых потоков при изменении угла ветровой атаки (рис. 4, б).

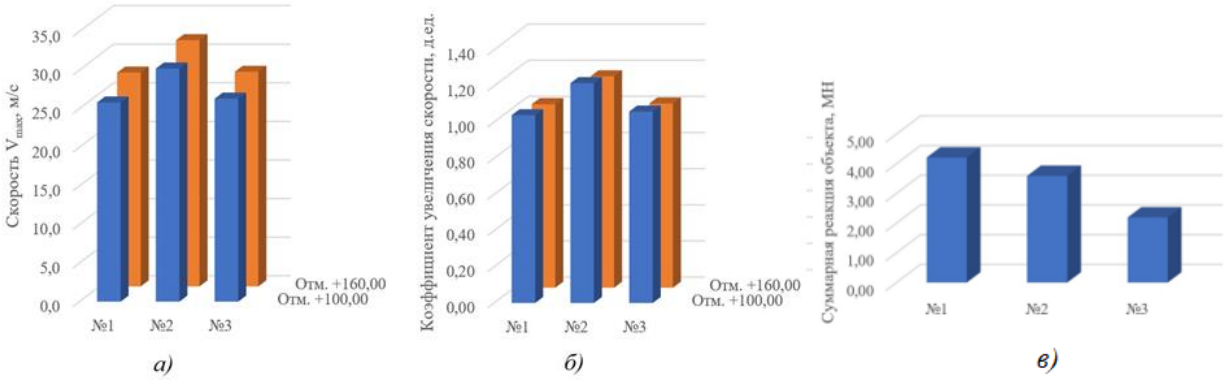


Рис. 4. Диаграммы распределения скоростей (а), коэффициентов увеличения скорости k (б) и суммарных реакций (в) для моделей при ветровой атаке 45 градусов.

Модель №1 является наименее обтекаемой и создает зону вакуума в рабочей области ветрогенератора, что уменьшает ее энергоэффективность над моделями №2 и №3 при меньших углах атаки. Модель №3 показала наиболее благоприятную аэродинамическую картину, где при изменении угла ветровой атаки практически полностью отсутствует турбулентный след.

Помимо эффективной работы ветрогенераторов необходимым условием оптимизации также является обеспечение экономической эффективности строительства объекта, критерием чего может быть принято значение равнодействующей всех аэродинамических сил, распределенных по поверхности объекта (рис. 3, в и рис. 4, в). Этот фактор также позволяет говорить о том, что именно модель №3 является наиболее рациональной при проектировании подобных процессов.

VI. Выводы и заключение

При проектировании зданий с интегрированными ветрогенераторами процесс геометрической оптимизации формы поперечного сечения башни (башен) должен быть направлен на смягчение характера аэродинамического взаимодействия воздушных масс и объекта путем рассмотрения более обтекаемых форм. Данная форма является компромиссной при решении противоположных задач – уменьшения контакта с воздушными массами для снижения общей реакции объекта, с одной стороны, и вовлечения как можно больших объемов воздуха в рабочую зону ветрогенератора с другой. Дальнейшее увеличение степени обтекаемости не приводит к экономии затрат на применяемых конструктивных решениях, а вследствие усложнения геометрической формы здания может привести к избыточным затратам.

Список литературы

1. Stankovic S., Campbell N., Harries A. Urban Wind Energy // London: Taylor&Francis, 2009.

2. Smith RF, Killa S. Bahrain world trade Center (BWTC): The first large-scale integration of wind turbines in a building. // The Structural Design of Talland Special Buildings. 2007; Vol. 16(4). P. 429-439.

3. Karadag I. / Wind Turbine Integration to Tall Buildings / Ilker Karadag, Izzet Yüksek // Intech Open. 2020. 15 p. DOI: 10.5772/intechopen.91650.

4. Клюзко В.М. Приемы объемно-планировочных решений формирования энергоэффективных высотных полифункциональных зданий, использующих энергию ветра. Архитектура и современные информационные технологии. Московский архитектурный институт. Москва, 2015. №2(31). 13 с.

5. Елистратов В.В. Ветроэнергетические установки – архитектурный элемент здания / В.В. Елистратов, Д.М. Боброва // Архитектура и современные информационные технологии. 2013. № 2 (23). 9 с.

6. Li Q.S., Shu Z.R., Chen F.B. Performance assessment of tall building-integrated wind turbines for power generation. Applied Energy.2016;165:777-788. DOI: 10.1016/j.apenergy.2015.12.114.

7. Хазов П.А., Поздеев М.Л. Оптимизация форм энергоэффективных зданий с ветрогенераторами // Приволжский научный журнал. 2021. № 4 (60). С. 55-63.

8. Ерофеев В.И., Сатанов А.А., Хазов П.А., Поздеев М.Л., Симонов А.В. Рациональная ориентация объектов с интегрированными ветрогенераторами по критерию максимизации вырабатываемой электроэнергии // Машиностроение и инженерное образование. 2023. № 4(73). С. 39-46.

9. Седов Л.И. Механика сплошной среды. Том 1. М.: Наука, 1973. 536 с.;

10. Поддаева О. И. Архитектурно-строительная аэродинамика : учебное пособие / О. И. Поддаева, А. С. Кубенин, П. С. Чурин / НИУ МГСУ. Москва: НИУ МГСУ, 2015. 88 с.

11. Герасимов С.И. Оценка погрешности оптико-физического метода измерения параметров воздушной ударной волны / С.И. Герасимов, А.Г. Сироткина, Н.А. Трепалов, Р.В. Герасимова // Измерительная техника, 2019. № 4. С. 38-41.

12. Шилов С.С. Численное моделирование и оптимизация ориентации высотного здания по розе ветров / С.С. Шилов, Е.С. Кашкина, П.А. Хазов // Приволжский научный журнал. 2023. № 3 (67). С. 42-50.

Секция 10

ЦИФРОВАЯ ТРАНСФОРМАЦИЯ ОБРАЗОВАНИЯ

Digital Transformation of Education

GraphiCon 2024
17 – 19 сентября 2024
Омск, Россия

УДК 378:004
DOI: 10.25206/978-5-8149-3873-2-2024-974-980

Преобразующая сила конвергентных технологий
The transformative power of convergent technologies

С.Г. Сорокина[1], О.Е. Баксанский[2]
[1]МГУ им. М.В.Ломоносова
[2]НИЦ «Курчатовский институт»

Svetlana G. Sorokina[1], Oleg Ye. Baksanskiy[2]
[1]Moscow University, Moscow, Russian Federation
[2]National Research Centre "Kurchatov Institute", Moscow, Russian Federation

Аннотация. Сфера образования постоянно развивается, и педагоги постоянно ищут инновационные методы для улучшения преподавания и обучения. В последние годы конвергентный подход к дидактике стал многообещающей парадигмой. Этот подход включает в себя интеграцию различных педагогических методов, стратегий и передовых технологий для создания сплоченной и высокоэффективной учебной среды. В статье рассматриваются основные проблемы, возникающие при переходе к этой новой образовательной парадигме, обусловленной конвергенцией сетевых технологий. Особое внимание уделено интеграции искусственного интеллекта (ИИ) в рамках образовательной сферы. Он проникает в таксономию целей и методологий обучения, присущих этому трансформационному сдвигу. Особое внимание уделяется вкладу ChatGPT в сферу языкового образования, проливая свет на его уникальные возможности и потенциальное влияние. Основываясь на существующем анализе литературы, в работе освещаются критические аспекты, которые необходимо учитывать при включении технологий на основе искусственного интеллекта в образовательный ландшафт.

Ключевые слова: конвергентная дидактика, конвергентная модель обучения, искусственный интеллект (ИИ), изучение языка, образовательная парадигма, генеративные предварительно обученные модели, ChatGPT.

Abstract. The field of education is constantly evolving, and educators continually seek innovative methods to enhance teaching and learning experience. In recent years, the convergent approach to didactics has emerged as a promising paradigm. This approach involves the seamless integration of diverse pedagogical techniques, strategies, and cutting-edge technologies to forge a cohesive and highly effective learning environment. This paper explores the the central challenges that arise during the transition to this new educational paradigm, driven by the convergence of network technologies. Particular attention is devoted to the integration of artificial intelligence (AI) within the educational domain. It delves into the taxonomy of goals and teaching methodologies intrinsic to this transformative shift. A special emphasis is placed on the noteworthy contributions of ChatGPT to the realm of language education, shedding light on its unique capabilities and potential impact. Based on the existing literature analysis, this paper highlights the critical facets to consider when incorporating AI-powered technologies into the educational landscape.

Keywords: convergent didactics, convergent teaching model, artificial intelligence (AI), language learning, educational paradigm, generative pre-trained models, ChatGPT

I. Introduction

The science of teaching and learning has a rich history. From the earliest days of instructing knowledge and skills to the contemporary era of advanced educational theories and technologies, this discipline has evolved and flourished, leaving an indelible mark on the progress of society. Despite the changing epochs, the foundational principles of the pedagogical system, originally articulated by Jan Amos Comenius (1592-1670), have persevered, remaining largely unaltered to this day [1]. Nevertheless, they have witnessed various educational paradigms and approaches throughout the years, with the most recent shift being towards digitalization and information abundance.

Researchers argue that the technological transformation of education can pursue two distinct pathways. The first route involves instrumental transformation and underscores the pragmatic aspects of digital tools with their capacity to address practical challenges such as offering online courses, saving cost by substituting traditional paper textbooks with electronic ones, and ensuring the accessibility of education during periods of self-isolation. This approach centers on socio-economic rather than socio-pedagogical objectives maintaining traditional teaching and learning methodologies. [2]

The second approach to the digital transformation of education is a systemic or deep modernization based on aligning the educational values, objectives, principles and content with the demands of the digital economy and digitally drenched society. The transformation in this case represents a radical overhaul of educational processes and redefines the

roles and functions of educators within a networked educational environment. The result of this transformation is expected to produce an education model characterised by its learner-centered, personalized, project-oriented, open, and participatory nature [Ibid.].

This innovative educational model combines traditional methodological foundations with the insights from past achievements in pedagogical science, while also firmly grounding itself in the contemporary digital realities. Consequently, the conventional educational paradigm within didactic science necessitates a profound reevaluation to adapt to the context of a network environment. Both Russian [3], [4] and foreign scholars [5], [6] are actively discussing diverse theoretical and practical facets of digital transformation in education. They collectively recognize convergence as the pivotal hallmark of the modern world.

It is widely acknowledged that the digital environment serves as the habitat and sphere of activity for modern individuals, seamlessly intertwining the physical and virtual domains, where they mutually converge, influence, and enrich one another. According to researchers, in today's age, this convergent network environment has become a novel socio-cultural space for socialization and personal development of the new generation [7].

The key attributes of this environment include the interconnectivity of all its components, fusion of real-world and virtual communication spaces, a seamless blend between the tangible and digital aspects of society as well as personalized services creative endeavors, and various social activities. Essentially, the convergent network environment represents a profound shift in how individuals interact, communicate, and engage with society. This underscores the crucial need for educational systems to adapt and prepare learners for this ever-evolving landscape that can be achieved through the adoption of a convergent approach to didactics.

This paper seeks to emphasize the primary challenges associated with transitioning to a new educational paradigm driven by the convergence of network technologies, with a particular focus on the integration of artificial intelligence within the realm of education. The primary conclusions have been drawn by synthesizing insights gathered from the existing body of literature.

II. Problem statement

Fundamental Aspects of the Convergent Teaching Model

Let us consider the most distinct features of the convergent approach in teaching comparing it to the traditional teaching methods.

In the conventional teaching model, educators typically follow a prescribed curriculum, disseminating information to passive learners in a one-size-fits-all manner. Undoubtedly, this approach has its merits; however, it tends to neglect the varied learning styles, abilities, and interests of students, as well as the untapped potential of the multitude of resources accessible in today's digital age.

The convergent approach to didactics serves as a response to these limitations. It represents a departure from traditional, linear teaching methods pursuing a more comprehensive and dynamic approach. This approach focuses on the amalgamation of various teaching strategies and resources, inviting educators to create dynamic, engaging, and adaptive learning environments for knowledge acquisition.

The convergent didactics capitalizes on the following principles:

Seeing learning as a multifaceted process. It acknowledges that no single teaching method can effectively address the diverse needs of all learners.

Practicing interdisciplinary learning. It recognizes that the real world is not neatly divided into subject compartments; instead, it is a complex tapestry of interconnected knowledge. By breaking down the traditional barriers between subjects, students can develop a more comprehensive understanding of complex problems and better prepare for the challenges they will face in their future careers.

Tailoring learning experiences to individual students' needs and interests. This principle acknowledges that students may have diverse backgrounds, abilities, and learning profiles that require different approaches to learning. It can also be challenging for educators since teachers need to strike a balance between catering to individual needs and maintaining a structured curriculum.

Promoting active learning strategies that encouraging students to take responsibility for their education. Through problem-based learning, group discussions, and hands-on activities, learners are given the opportunity to apply their knowledge, think critically, and develop essential problem-solving skills. This active engagement not only enhances comprehension but also fosters a deeper appreciation for the subject matter.

Pursuing integration of multimedia and technology into the learning process. In today's digital age, learners are accustomed to accessing information through a variety of mediums, from videos and interactive simulations to podcasts and online forums. By harnessing the power of multimedia, educators can create a more immersive and captivating learning environment, catering to the visual, auditory, and kinesthetic preferences of their students.

The present stage of technological advancement in society is characterized by the dynamic penetration of artificial intelligence into various facets of human life, with a notable focus on the domains of science and education. One of the most prominent manifestations of artificial intelligence is the Generative Pre-training Model (GPT), a neural language model that learns from extensive text datasets and possesses the capability to generate human-like text [8].

III. Theory

Controversy Surrounding the Integration of Artificial Intelligence into Education

In recent years, there has been a burgeoning body of literature in pedagogy concerning the utilization of chatbots, not only within the broader context of education [9], [10], but also in the specific realm of subject-specific teaching, particularly foreign languages [11], 12], 13], [14]. Among these chatbots, OpenAir ChatGPT stands out as the most technologically advanced, possessing significant didactic and linguo-didactic potential. It empowers students to engage in the creation of written content and the execution of research projects.

There was a well-known case where a Russian student used ChatGPT as a resource to aid in the Final Year Paper, which was positively assessed. Other cases include the artificial intelligence-driven chatbot GPT-3's managing to pass the final exam for the school's Master of Business Administration (MBA) program, generating an essay for philosophy class and writing research articles [15].

Thus, the linguistic and cognitive capabilities of ChatGPT, which expedite the preparation of coursework and research papers, have become a hot topic and divided scholars and educators into two camps, one opposing and even prohibiting the use of ChatGPT in teaching [16], while the other advocating for its judicious integration into the student research process [17]. The latter argue that it is no longer feasible to universally prohibit students from enjoying the capabilities of artificial intelligence technologies, including ChatGPT [4] as soon as the use of AI is properly regulated by university policies [18].

According to the advocates of integrating AI-powered models into the education process, it would be more prudent and ethical, to involve students in collaborative work with chatbots, teaching them to recognize the potential risks associated with the use of neural network utilization encouraging them to take responsibility for their coursework and dissertation results. Moscow City University in Russia, for example, has responded to the challenge presented by neural networks and recently adopted generative technologies. This permits students to incorporate these technologies into their research and studies, though while using them, they must be mindful of the importance of a critical approach. Students are mandated to assess the ideas and texts generated by these technologies and recognize that not all information is inherently reliable and requires validation from more credible sources.

IY. Experimental results

Concerns of Using Generative Models

Using generative models, while powerful and versatile, raises several concerns and challenges that need to be carefully addressed. Here are some of the main concerns expressed by individuals demonstrating caution when applying generative models in research.

Lack of transparency. Many generative models, especially complex deep learning models, are often considered as "black boxes". The inner codes and workings of such models may be accessible and comprehensible to only a select group of experts in the field.

Ethical concerns. The generation of highly realistic synthetic content raises ethical concerns, particularly when it comes to creating deepfake videos or generating fake news articles.

Data privacy and security. Generative models often require large datasets for training, and the use of personal or sensitive data can pose privacy and security risks. There are concerns about data breaches or misuse when handling sensitive information.

Data quality and bias. Generative models, trained by individuals, can be influenced by biased training data, potentially leading to the generation of biased outputs. This can perpetuate or amplify existing societal biases, particularly in fields like natural language processing and image generation.

Legal and copyright issues. The use of generative models to create content, such as art or music, can raise legal and copyright issues. It may be challenging to determine who owns the rights to generated content and whether it infringes on existing copyrights.

Data privacy and security. Generative models often require large datasets for training, and the use of personal or sensitive data can pose privacy and security risks. There are concerns about data breaches or misuse when handling sensitive information

Unintended consequences. As artificial intelligence advances and gains autonomy, its increasing complexity raises questions about whether it could eventually surpass human intellect through emulation. Prominent scientists, including Elon Musk, have expressed growing concerns about the potential for a technological singularity—a scenario in which technological development becomes uncontrollable and leads to profound changes in human civilization. While many experts believe that humanity does not currently face an imminent threat from such a scenario, there is a pressing need to conduct comprehensive research, gain a deep understanding, and responsibly manage the evolution of self-learning artificial intelligence.

Y. Discussion of results

ChatGPT capacities aligned with key didactic principles

While these concerns are evident, it is important not to underestimate the educational potential of generative models, particularly given their capacity to align with key didactic principles. [19]

The use of GPT within Bloom's Taxonomy

By categorizing educational objectives according to Bloom's Taxonomy, we can seamlessly align the use of ChatGPT with various cognitive domains, spanning from lower-order thinking skills to higher-order thinking skills (Table 1).

While ChatGPT is primarily an information tool, it can indirectly contribute to emotional and value-based experiences by providing content that may evoke emotional responses and trigger discussions on these matters.

TABLE 1. ALIGNING THE ROLE OF CHATGPT WITH THE EDUCATIONAL GOALS OUTLINED BY BLOOM'S TAXONOMY

Levels of education goals	ChatGPT role in achieving education goals
Remember (Knowledge)	- serves as a valuable source of information and knowledge - provides factual information, explanations, and insights on a wide range of topics, thus expanding their cognitive knowledge base. - assists learners in recalling facts, definitions, or specific details related to a topic by providing concise answers to factual questions
Understand (Comprehension)	- helps users understand complex concepts by explaining them in simpler terms - provides additional context and explanations - summarizes information to enhance comprehension
Apply (Application)	- can demonstrate how theoretical knowledge is applied in real-world scenarios by offering practical solutions to problems or challenges. - can support the application of knowledge by providing practical solutions to problems or scenarios
Analyze (Analysis)	- assists in breaking down information into components, - assists in comparing and contrasting ideas - provides a critical analysis of texts or data by offering insights and explanations
Evaluate (Synthesis)	- helps users evaluate arguments, theories, or proposals by providing alternative viewpoints, pros and cons, or criteria for assessment - involves users in discussions to evaluate different perspectives
Create (Evaluation)	- can assist users in brainstorming sessions, generating creative content, or exploring innovative concepts - stimulates creative thinking by providing prompts, ideas, or alternative perspectives

To maximize the alignment with Taxonomy, it's important to use ChatGPT as a facilitator of learning rather than a passive source of information. Students need to be encouraged to actively engage with the content, ask thought-provoking questions, and apply the knowledge and insights gained from ChatGPT in practical and creative ways.

Aligning the role of ChatGPT with the teaching approaches

The educational potential of the ChatGPT model readily aligns with teaching methods when students formulate appropriate questions to make inquiries (Table 2).

TABLE 2. ALIGNING THE ROLE OF CHATGPT WITH THE TEACHING APPROACHES

Teaching approaches	The role of ChatGPT in educational process
Explanatory-Illustrative	- can provide clear and detailed explanations on various topics. - can offer visual aids, examples, or diagrams to illustrate explanations effectively - can supplement explanations provided by teachers or textbooks, helping students gain a deeper understanding of complex concepts
Reproductive	- encourages users to ask questions on a particular topic this way reproducing and practicing what they've learned - can make students formulate questions or answers, reinforcing their understanding of the material
Problem-Based	- can be integrated into problem-based learning scenarios - can help students analyze and solve complex problems by providing relevant information, alternative approaches, or suggestions - can encourage students to brainstorm solutions or gather data for their problem-solving tasks

Heuristic	- can serve as a heuristic tool by encouraging critical thinking and exploration. - can encourage students to ask open-ended questions, explore different viewpoints, and engage in discussions that promote problem-solving and discovery - can challenge students to think beyond the obvious and consider various possibilities
Inquiry-Based	- can support inquiry-based learning by helping students gather information, formulate research questions, and explore topics of interest - can assist in the initial stages of research projects by suggesting keywords, or providing background information

Leveraging ChatGPT Responsibly and Effectively in the Classroom

Incorporating ChatGPT into these teaching methods can make the learning process more interactive, engaging, and dynamic. However, it's important to guide students on how to use ChatGPT or any AI-powered tool properly, honestly, fairly, and reasonably as a learning tool explaining what ChatGPT is, how it works, and its capabilities and limitations clarifying that ChatGPT is a tool and not a source of absolute truth or authority. Here are several key considerations when integrating AI-powered technologies into the education process

Students should be encouraged to critically evaluate the information obtained and question AI-generated content for bias, and take responsibility for verifying information obtained from AI sources before accepting it as fact.

When incorporating AI-generated content into assignments or projects, students need to understand the importance of providing proper citation and attribution to uphold originality and prevent plagiarism. It's important to note that citations provided by ChatGPT are not always accurate. In addition, students need to make sure AI-generated content meets any ethical or legal requirements and align with academic standards.

Students should receive instruction in cybersecurity best practices to guard against malicious AI-generated content. Additionally, they need to understand the significance of privacy when engaging with AI and the importance of not sharing sensitive personal information.

Students should recognize the significance of practicing polite and constructive communication, employing respectful and considerate language when interacting with AI systems, as these systems are designed to self-learn and emulate human behavior and can inherit communication patterns and biases from their training data.

By providing guidance and fostering a sense of responsibility and ethics, educators can help students use AI tools like ChatGPT in ways that benefit them and the education process as a whole. However, educators should ensure that ChatGPT complements classroom instruction and learning objectives being a component of the teaching methodology rather than entirely replacing traditional teaching methods. Tokyo University even adopted the Policy on the use of AI tools in classes [18].

ChatGPT in language learning

Using ChatGPT in language learning has shown promising potential. It can provide students with opportunities for natural language practice and conversation, which is crucial for improving language skills. ChatGPT can generate prompts for conversation, correct grammar and vocabulary, and offer instant feedback on pronunciation and fluency.

Moreover, ChatGPT can adapt to the learner's level, making it suitable for both beginners and advanced language learners. It can also expose students to colloquial language and various accents, enhancing their comprehension skills in real-world situations. Table 3 provides a summary of the primary activities that students can engage in through the use of ChatGPT.

TABLE 3. CHATGPT CONTRIBUTION TO THE LANGUAGE EDUCATION

Activities	Comments
Conversational practice	ChatGPT engages in text-based conversations. Students practice their verbal and written communication skills. This practice helps students become more confident and fluent in expressing their ideas.
Personalized Learning Assistance	ChatGPT can provide personalized learning assistance to students. It can answer questions, offer explanations, and provide additional resources based on individual student needs. This level of personalized support enhances the adaptive nature of convergent education, ensuring that students receive the help they require in real-time.
Articulating thoughts	When students interact with ChatGPT, they are required to articulate their thoughts and ideas clearly. This process of turning their thoughts into coherent sentences and responses enhances their ability to communicate effectively.
Asking questions	ChatGPT encourages students to formulate questions, whether they are seeking clarification on a topic or delving deeper into a subject. The practice of asking questions fosters critical thinking and curiosity.

Feedback and correction Assessment and Feedback	ChatGPT can provide immediate feedback on grammar, vocabulary, and sentence structure. This real-time correction helps students identify and rectify language errors, thereby improving their writing and verbal skills. ChatGPT can facilitate formative assessments and provide feedback on student work. It can evaluate assignments, quizzes, or essays, offering constructive comments and suggestions for improvement. This assists educators in monitoring student progress and tailoring their teaching methods accordingly.
Vocabulary expansion	Engaging in diverse conversations with ChatGPT exposes students to a wide range of vocabulary. They can learn new words in context and apply them in their own language usage, enriching their vocabulary
Support for language learners	For students learning a new language, ChatGPT can provide a supportive environment for practicing and improving language skills. It can offer translations, explanations of idiomatic expressions, and language-specific cultural insights.
Communication in different formats	ChatGPT can simulate different modes of communication, such as casual conversation, formal correspondence, academic writing, and professional communication. This exposure helps students adapt their language and communication style to different contexts.
Encouragement of reflective communication	Through interactions with ChatGPT, students have the opportunity to reflect on their own communication patterns and identify areas for improvement. This self-awareness is a fundamental step in becoming a more effective communicator.
Promotion of active engagement	Engaging in meaningful conversations with ChatGPT promotes active engagement with content. Students are encouraged to think critically, respond thoughtfully, and connect ideas, all of which contribute to their language and communication development
Enhanced Accessibility	ChatGPT can improve accessibility for students with diverse needs. It can provide text-to-speech support, assist with reading comprehension, or offer explanations in multiple languages. This inclusivity aligns with the convergent model's goal of accommodating the needs and preferences of a diverse student population.

As we can observe, AI-powered models such as ChatGPT can serve as valuable tools for language practice and reinforcement. However, they should complement rather than replace human interaction and personalized instruction. It is essential to use ChatGPT alongside traditional language learning methods to ensure a well-rounded learning experience.

YI. Conclusions

ChatGPT plays a multifaceted role in convergent education, offering personalized learning assistance, a wide range of educational resources, assessment and feedback facilitation, improved accessibility, and support for interactive learning experiences.

Its versatility and availability make it a valuable tool for both students and educators as they embrace the convergent approach to teaching and learning in the modern era. Interdisciplinary learning transcends traditional subject boundaries, encouraging educators and students to explore complex, real-world problems from diverse perspectives..

As an AI language model, ChatGPT can be a pivotal facilitator of interdisciplinary learning, promoting collaboration among educators, fostering integrated curricula, and aiding in the development of a holistic understanding of multifaceted issues. In interdisciplinary learning, ChatGPT can guide students in their research, helping them locate relevant sources, synthesize information from different fields, and develop well-rounded arguments for their projects.

Furthermore, ChatGPT offers insights from a wide range of academic disciplines. It can explain concepts across different fields and provide diverse perspectives on complex issues. This multidisciplinary knowledge can be a valuable resource for educators as they design interdisciplinary curricula.

References

[1]. Comenius J. A. The Great Didactic. Translated into English and edited with biographical, historical and critical introductions by M. W. Keatinge, M.A. New York: Russell & Russell. 1967. 316.

[2]. Kondakov A.M., Sergeev I.S. The subject of digital didactics is education in a convergent network environment. Problems of didactics development in the context of digital transformation of education: Collection of scientific works / Compiled by: V.G. Martynov, V.M. Zhurakovsky. – Moscow: Russian State University of Oil and Gas (National Research University) named after I.M. Gubkina, 2022. – P. 98-121. – EDN YFLEXP.

[3]. Garkusha N. S., Gorodova Yu. S. Pedagogical capabilities of ChatGPT for the development of cognitive activity of students // Professional education and the labor market. 2023. No. 1. P. 6–23. https://doi.org/10.52944/PORT.2023.52.1.001

[4]. Sysoyev P.V., Filatov E.M. Chatbots in teaching a foreign language: advantages and controversial issues. Bulletin of Tambov University. Series: Humanities. – 2023. – T. 28, No. 1. – P. 66-72. – DOI 10.20310/1810-0201-2023-28-1-66-72. – EDN PXGZTJ.

[5]. Firaina, R., & Sulisworo, D. Exploring the Usage of ChatGPT in Higher Education: Frequency and Impact on Productivity. Buletin Edukasi Indonesia, (2023). 2(01), 39–46. https://doi.org/10.56741/bei.v2i01.310

[6]. Firat M. What ChatGPT means for universities: Perceptions of scholars and students. Journal of Applied Learning and Teaching. 2023. 6(1).

[7]. Baksanskiy, O. E. Convergent paradigm and modern education. New developments in psychological, physiological and medical neuroresearch: All-Russian interdisciplinary scientific School with international participation, in preparation for the XXIII Congress of the Russian Physiological Society named after. I.P. Pavlova (Voronezh, 2017), dedicated to the 100th anniversary of the creation of this society by Ivan Petrovich Pavlov, Sudak, May 30 - 03, 2017. – Sudak: MAKS Press LLC, 2017. – pp. 20-22. – EDN ZEDSLX.

[8]. Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding with unsupervised learning.

[9]. Stokel-Walker C, VanNoorden R.What ChatGPT and generative AI mean for science. Nature. 2023;614(7947):214-216. https://doi.org/10.1038/d41586-023-00340-6

[10]. Sorokina S.G. Constructing the phenomenon of self-concept: semantics and functions of the self lexeme Modern pedagogical education. – 2023. – No. 5. – P. 266-270. – EDN FXHCAK.

[11]. Han D. The Effects of voice-based AI chatbots on Korean EFL Middle School students' speaking competence and affective domains. Asia-pacific Journal of Convergent Research Interchange. 2020. Vol. 6. № 7. P. 71-80. http://dx.doi.org/10.47116/apjcri.2020.07.07

[12]. Çakmak F. Chatbot-human interaction and its effects on EFL students' L2 speaking performance and speaking anxiety. Novitas-ROYAL (Research on Youth and Language). 2022. № 16 (2). P. 113-131. URL: https://novitasroyal.org/download/chatbot-human-interaction-and-its-effects-on-efl-students-l2-speaking-performance-and-anxiety/

[13]. Kim H.S., Cha Y., Kim N.Y. Effects of AI chatbots on EFL students' communication skills // Korean Journal of English Language and Linguistics. 2021. № 21. P. 712-734. URL: https://doi.org/10.15738/kjell.21..202108.712

[14]. Haryanto E., Ali R. Students' attitudes towards the use of Artificial Intelligence SIRI in EFL learning at one public university // International Seminar and Annual Meeting BKS-PTN Wilayah Barat. 2018. Vol.1. №1. P.190-195. URL: http://conference.unsri.ac.id/index.php/semirata/article/view/1102

[15]. Biswas, S. ChatGPT and the Future of Medical Writing. Radiology. 2023. 307:2. https://pubs.rsna.org/doi/full/10.1148/radiol.223312

[16]. Yang M. New York City schools ban AI chatbot that writes essays and answers prompts. The Guardian. 2023. URL: https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt (accessed 27 Mar 2023).

[17]. Mollick E. All my classes suddenly became AI classes. One Useful Thing Also Some Things. 2023. URL: https://oneusefulthing.substack.com/p/all-my-classes-suddenly-became-ai (accessed 25 Aug 2023).

[18]. Policy on the use of AI tools in classes. April 28, 2023. https://utelecon.adm.u-tokyo.ac.jp/en/docs/ai-tools-in-classes

[19]. Baksansky O.E. Convergence of natural scientific and human knowledge. Noospheric research. – 2021. – No. 1. – P. 88-98. – DOI 10.46724/NOOS.2021.1.88-98. – EDN REFYCV.

УДК: 378:147
DOI: 10.25206/978-5-8149-3873-2-2024-981-987

Управление качеством образования в области графики в технических университетах
Quality management of education in the field of graphics at technical universities

Кучкарова Диларом Файзуллаевна[1],
Ачилова Дилноза Ахматовна[2]

[1]Национальный исследовательский университет «Ташкентский институт инженеров ирригации и механизации сельского хозяйства», Ташкент, Узбекистан
[2]Совместный Белорусско-Узбекский межотраслевой институт прикладных технических квалификаций в городе Ташкенте, Ташкент, Узбекистан

Dilarom F. Kuchkarova[1],
Dilnoza A. Achilova[2]

[1]National Research University "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers", Tashken,t Uzbekistan

[2]Join Belarusian-Uzbek Interdisciplinary Institute of Applied Technical Qualifications in Tashkent, Tashken,t Uzbekistan

Аннотация. В настоящей статье рассматривается вопрос управления качеством обучения графическим дисциплинам. Основной идеей настоящей статьи является необходимость внедрения непрерывного контроля в процессе обучения. Рассматриваются следующие виды контроля – письменная работа, устный контроль, компьютерное тестирование и защита самостоятельного проекта. Все виды контроля обладают как достоинствами, так и недостатками. Сама технология применения видов контроля опирается на определенную очередность. Предлагаемая концепция непрерывного контроля предполагает технологию разделения академической группы студентов на подгруппы с разной степенью мотивации и подготовки. Различные корреляционные связи между видами контроля позволяют дать квалиметрическую оценку каждому конкретному студенту и в целом академической группе. Индекс качества обучения определяется средней арифметической оценкой группы. Анализ взаимосвязей между видами контроля позволяет выявить недостатки процесса обучения и оперативно управлять этим процессом. В статье также приводятся определения качества образования и пути достижения необходимого уровня качества. В результате внедрения непрерывного контроля в учебный процесс вуза были выявлены как положительные стороны, так и недостатки предлагаемой технологии.

Ключевые слова: образовательный процесс, квалиметрия, качество образования, управление образовательным процессом, непрерывный мониторинг.

Abstract. This article discusses the issue of managing the quality of teaching graphic disciplines. The main idea of this article is the need to introduce continuous monitoring in the learning process. The following types of control are considered: written work, oral questioning, computer testing and defense of an independent project. All types of control have both advantages and disadvantages. The technology for applying types of control is based on a certain sequence. The proposed concept of continuous monitoring involves the technology of dividing an academic group of students into subgroups with varying degrees of motivation and preparation. Various correlations between types of control make it possible to give a qualimetric assessment to each individual student and to the academic group as a whole. The learning quality index is determined by the arithmetic average of the group. Analysis of the relationships between types of control allows us to identify shortcomings in the learning process and quickly manage this process. The article also provides definitions of the quality of education and ways to achieve the required level of quality. As a result of the introduction of continuous monitoring into the educational process of the university, both positive aspects and disadvantages of the proposed technology were identified.

Key words: educational process, qualimetry, quality of education, management of the educational process, continuous monitoring

Введение

Качество образования одно из ключевых в современной науке трактуется по-разному разными исследователями, но сводится в общем случае к определению, данному в ИСО 9000-2001 «Качество – степень соответствия присущих объекту характеристик установленным требованиям». [1]

Понятие качества, являясь многомерным и мультифакторным феноменом, включает такие составляющие как качество результатов деятельности (процесса), качество самих процессов и качество системы или организации деятельности. К ключевым факторам качества образования относят: качество содержания образования, качество абитуриентов и студентов, качество методического и материально-технического обеспечения образовательного процесса, качество профессорско-преподавательского состава, качество технологий тестирования и проверки знаний, умений и навыков студентов, качество общего менеджмента ВУЗа. [2]

Каждый временной период в развитии национальных систем образования выдвигает свою квалиметрическую парадигму, наполняя её соответствующими представлениями о содержании, методах и технологиях, этической и процессуальных сторонах контрольно-оценочной деятельности в области образования.[1]

Конечным результатом работы высшего учебного заведения является специалист, который способен занять определённую социальную позицию, а его качество определяется:

1. Знаниями в области теоретических и фундаментальных основ будущей профессии, а также в сфере социальных коммуникаций.
2. Способностью к самостоятельным действиям в области профессиональных умений и навыков в реальных условиях трудовой деятельности.
3. Готовностью к положительным взаимоотношениям с участниками образовательного процесса, что способствует дальнейшему обучению и самообразованию.

В настоящее время основной акцент при оценке профессиональной подготовленности выпускника ВУЗа переносится на умение в условиях неопределённости продуцировать и использовать знания, умения и навыки для выполнения нестандартной задачи и разрешения сложной ситуации. Сопряженным «качеству образовательного процесса» выступает «качество выпускника вуза», определяемого «моделью качества выпускника вуза» как его «потенциальным качеством». [3, 4]

Как известно, высшее профессиональное образование – главный механизм кадровой политики государства на долгосрочную перспективу, что нашло свое отражение во многих исследованиях.

Феномен «качества образования» трактуется как изменения в учебно-познавательном процессе вуза, которые способствуют повышению ценностей формируемых студентами по завершению определенного уровня обучения. [5]

Критерии выбора и структурирования набора показателей качества могут быть разнообразными. В качестве основы для выделения такого набора иногда используются модели образовательной системы, структурированные цели этой системы, критерии её эффективного функционирования и актуальные проблемы управления и т.д.

Методы

«Для принятия обоснованного решения о достижении либо не достижении определенного качества необходимо по каждому показателю выбрать некоторый критерий или набор критериев, характерный для уровневого подхода к оценке качества результатов образования.» [6]

Эти количественные показатели отражают динамику изменений результатов обучения конкретного студента. Однако для общей оценки результатов обучения графическим дисциплинам необходимы интегральные показатели, которые демонстрируют динамику развития отдельной академической группы.

В настоящее время в ряде зарубежных стран активно применяется динамическая модель улучшения качества образования, в которой качество рассматривается как положительные изменения в процессах и результатах обучения, вызванные совершенствованием образовательной системы и отражающие новые требования общества. Для реализации этой модели необходимо собирать данные о познавательной и творческой активности, а также других учебных достижениях студентов на протяжении всего обучения и анализировать их прирост с использованием дескриптивной статистики. Таким образом, в рамках динамического подхода оценка

качества учебных достижений основывается на выявлении изменений в подготовке обучаемых, которые воспринимаются как улучшение знаний и умений или формирование компетенций.

Одним из подходов к установлению количественных показателей учебного процесса является постоянный контроль знаний, умений и навыков студентов. В настоящее время существующие формы контроля, такие как письменные работы, устные экзамены, компьютерное тестирование и защита самостоятельных проектов, в основном применяются в соответствии с планами учебно-методического отдела конкретного вуза.

Рассмотрим достоинства и недостатки каждой формы контроля.

В устном контроле можно выделить несколько достоинств: -непосредственный контакт между педагогом и студентом, -возможность оценить кругозор студента в определенной области, -приобретение опыта формулирования ответов техническим языком, а также -развитие самостоятельного мышления и умения аргументировать свои ответы с помощью примеров.

Однако существуют и недостатки: -значительное количество времени, которое преподаватель тратит на устное общение, -определенная субъективность в оценке знаний, а также -низкая степень достоверности результатов.

В письменном контроле можно выделить несколько преимуществ: -выявление творческих способностей студентов в ходе самостоятельной письменной работы, -объективная оценка письменных работ при наличии четких критериев, а также -приобретение студентами навыков выражения своих мыслей в письменной форме.

Среди недостатков можно отметить: -значительные временные затраты на проверку работ, -невозможность автоматизации процесса проверки и -отсутствие непосредственного контакта со студентом.

В тестовых испытаниях достоинства: -средство прозрачного педагогического оценивания. Результаты тестирования не зависят от преподавателя; это позволяет охватить весь учебный материал; процесс тестирования автоматизирован.

Однако есть и недостатки: -студент не имеет возможности аргументировать и расширять свой ответ; -невозможно оценить знания, умения и навыки, выходящие за рамки программы; -создание качественных тестов требует значительного времени.

Достоинства защиты самостоятельного проекта включают: -возможность самостоятельного изучения новой темы, -развитие навыков публичных выступлений у студентов и -умение логически отвечать на вопросы.

К недостаткам можно отнести: -невозможность проверки знаний по всем разделам программы и -отсутствие возможности оценить навыки письменного изложения..

В Национальном исследовательском университете «Ташкентском институте инженеров ирригации и механизации сельского хозяйства» в течении многих лет внедрялась концепция непрерывного контроля на кафедре «Начертательная геометрия и инженерная графика»

В течение семестра студенты проходят непрерывную аттестацию, которая включает сдачу письменных заданий, устный контроль, решение задач, защиту самостоятельного проекта и компьютерное тестирование. На основе оценок за каждую аттестацию формируется рейтинг студента.

Отметим интересный факт - независимо от вида контроля распределение оценок студентов неизменно подчиняется нормальному закону.

Анкетирование позволяет классифицировать студентов на три условные группы: I — студенты с высокой подготовкой и сильной мотивацией к учебе; II — студенты со средней подготовкой и умеренной мотивацией; III — студенты с низкой подготовкой и слабой мотивацией. Анкетирование разработано авторами статьи специально для студентов первых курсов для определения уровня мотивации к учёбе и графической подготовки. [7]

Мониторинг процесса обучения осуществляется на протяжении учебного года. Условные три группы студентов не являются статичными — в течение семестра наблюдается постоянный переход студентов из одной группы в другую, однако ядро каждой группы остается неизменным..

Все наблюдения обрабатываются как при помощи классических статистических методов измерений – ранговой корреляции, дисперсионного анализа, так и специально разработанных- Q- критерия Розенбаума, U-критерия Манна-Уитни и др. [8]

В течение учебного семестра студентам предлагались компьютерные тесты первой степени сложности и задания на решение задач. Выяснилось, что высокие результаты по компьютерным тестам показывают студенты с хорошей зрительной памятью и высокой концентрацией внимания.

Рассмотрим конкретный пример результатов непрерывного контроля академической группы направления ИЭВИС 1-8. (*использование энергии воды в ирригационных системах*), 23 студента, 1 семестр, 30 лек., 60 прак.

ТАБЛИЦА 1

РАНГОВАЯ КОРРЕЛЯЦИЯ МЕЖДУ РАЗЛИЧНЫМИ ВИДАМИ КОНТРОЛЯ.

№	Студент	Балл по письменной работе	Ранг	Балл по устному контролю	Ранг	Баллы за компьютерное тестирование	Ранг	Баллы за защиту реферата	Ранг	d(A-B) (разница между рангами письменной работы и устному контролю)	d^2 (A-B) (квадрат разницы между рангами письменной работы и устного контроля)	d(A-C) (разница между рангами письменной работы и компьютерного тестирования)	d^2(A-C) (квадрат разницы между рангами письменной работы и компьютерного тестирования)	d(A-D) (разница между рангами письменной работы и защиты реферата)	d^2(A-D) (квадрат разницы между письменной работе и защите реферата)
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		A		B		C		D							
1	Б.А.	56	6	52	8	72	9	74	10	-2	4	-3	9	-4	16
2	Э.Ж.	60	8	56	11	76	10	72	9	-3	9	-2	4	-1	1
3	А.А.	65	11	54	9	80	11	68	7	2	4	0	0	4	16
4	А.А.	72	13	62	13	82	12	78	12	0	0	1	1	1	1
5	К.Ж.	80	14	70	14	86	14	82	13	0	0	0	0	1	1
6	А.Т.	48	4	50	7	84	13	86	14	-3	9	-9	81	-10	100
7	З.Д.	50	5	40	4	70	8	60	6	-1	1	-3	9	-1	1
8	И.Ж.	58	7	46	5	68	7	56	4	-2	4	0	0	3	9
9	Н.А.	62	9	58	12	66	6	58	5	-3	9	3	9	4	16
10	Х.Х.	40	2	30	1	60	4	54	3	1	1	-2	4	-1	1
11	А.А	42	3	32	2	52	2	50	1	1	1	1	1	2	4
12	Т.М.	36	1	34	3	50	1	52	2	-2	4	0	0	-1	1
13	О.Б.	64	10	55	10	62	5	76	11	0	0	5	25	-1	1
14	И.Т.	68	12	49	6	56	3	70	8	6	36	9	81	4	16
											$\sum = 8$		$\sum = 224$		$\sum = 184$

В таблице приведены баллы за виды контроля. Баллы приведены по шкале от 0 до 100 баллов.

Применим формулу коэффициента ранговой корреляции Спирмена для определения взаимосвязей между оценками по различным видам контроля.

Вычислим коэффициент

$$r_s = 1 - \frac{6 * \sum(d^2)}{N * (N^2 - 1)}$$

где d-разность между рангами; N-количество ранжируемых значений (студентов)

$$r_s = 1 - \frac{6 * 82}{14 * (14^2 - 1)} = \frac{492}{2730} = 0{,}18$$

Вычислим коэффициент r_s для двух видов контроля – письменная работа и компьютерное тестирование:

$$r_S = 1 - \frac{6 * 224}{14 * (14^2 - 1)} = 0{,}51$$

Вычислим коэффициент r_s для двух видов контроля – письменная работа и защита реферата:

$$r_S = 1 - \frac{6 * 184}{14 * (14^2 - 1)} = 0{,}59$$

Количественные показатели ранговой корреляции между различными видами контроля указывают на наличие слабой положительной корреляции между ними. Обработка результатов аттестации более 2300 студентов с 2011 года с использованием методов математической статистики позволила сделать ряд выводов. Рассмотрим пример вычисления интегрального показателя для группы направления «Ирригационные системы». В таблице представлены средние арифметические баллы по четырем видам контроля.

Табл.

Средние баллы по четырем видам контроля

№	Количество студентов	Среднее арифметическое баллов по письменному контролю	Среднее арифметическое баллов по устному контролю	Среднее арифметическое баллов по компьютерному тестированию	Среднее арифметическое баллов по защите рефератов
		A	**B**	**C**	**D**
	14	57,2	49,1	68,9	66,9

Вычислим среднее арифметическое для всех видов контроля

$$M(x_i) = \frac{57{,}2 + 49{,}1 + 68{,}9 + 66{,}9}{4} = 60{,}53$$

Для данной академической группы студентов средний уровень усвоения учебного материала оказался выше порогового значения, равного 55 баллам. Анализ представленной таблицы показывает, что средний результат по устному контролю ниже установленного порога, что указывает на недостаточную подготовку студентов по определённым темам. Низкие оценки по устному контролю ясно демонстрируют, что некоторые студенты не понимают ключевые понятия и испытывают трудности с пространственным мышлением. Это требует дополнительных усилий и занятий со стороны преподавателя.

В то же время высокие баллы за письменные работы не всегда свидетельствуют о хорошем усвоении студентами основных тем. Общее развитие и умение работать с учебной литературой также могут влиять на получение высоких оценок за письменные задания. Компьютерное тестирование выявляет такие качества, как хорошая зрительная память и скорость реакции.

Опыт показывает, что только одновременное и технологически обоснованное использование всех видов контроля позволяет, во-первых, объективно оценить знания, умения и навыки студентов, а во-вторых, эффективно управлять процессом обучения.

Рассмотренные выше квалиметрические оценки студентов, математические ожидания оценок и показатели ранговой корреляции между оценками по всем видам контроля в совокупности формируют интегральную оценку учебного процесса. Анализ этих показателей предоставляет ясное представление о том, как студенты усваивают учебный материал.

Преимуществом данного подхода является возможность оперативно реагировать на результаты контроля как для отдельных студентов, так и для академической группы в целом. Интегральные оценки по каждому виду контроля в академической группе позволяют сделать выводы о общем уровне знаний студентов, их достижениях и недостатках в изучении конкретных тем учебного материала.

Концепция непрерывного контроля показана на схеме Рис. 3.3

Рис. 3.3. Схема концепции непрерывного контроля

Резюмируя вышеизложенное можно сказать, что предложенный подход к управлению учебным процессом при обучении «Начертательной геометрии и инженерной графики» даёт количественные оценки результатов обучения конкретного студента и конкретной академической группы. Каждый студент получает итоговую квалиметрическую оценку, а академическая группа – интегральный показатель по разным видам контроля.

Вывод Анализ интегральных показателей по разным видам контроля позволяет констатировать, что индекс качества обучения удовлетворительный и далёк от максимального. Интегральный показатель по письменным

работам демонстрирует удовлетворительное умение студентов работать с литературой, умение письменно выражать свои мысли. Интегральный показатель по устному контролю демонстрирует неумение части студентов выражать устно мысли, формулировать правильный ответ,

находить за отведённое время правильное решение задач. Интегральный показатель компьютерного тестирования демонстрирует хорошую зрительную память, хорошую скорость реакции, умение работать на компьютере. Интегральный показатель презентации рефератов демонстрирует умение самостоятельно изучить незнакомую тему, хороший навык публичных выступлений, умение логично излагать материал, убеждать аудиторию в правильности своих выводов.

Литература

1. Субетто А.И. Качественность образования. СПб. – М.: Научно-исследовательский проблемный центр качества подготовки специалистов, 2000. – 220 с.

2. Гарри Азгалдов, Альваро Падилья Омисте Азбука квалиметрии. Инструментарий для измерения ImmeaSurable 2016. 120 стр

3. Engchun, Rudjanee; Sungtong, Ekkarin; Haruthaithanasan, Theera (6 September 2017). "Homeschooling in Southern Thailand: Status and proposed guidelines for learning process management". Kasetsart Journal of Social Sciences. 39 (3): 502–508. doi:10.1016/j.kjss.2017.08.003

4. Гузенков В. Н., Якунин В. И. Геометрико-графическое обучение как интегрирующий фактор образовательного процесса // Образование и общество. 2014 № 2 С. 26–28.

5. Issues and trends in Education for Sustainable Development. France: UNESCO. 2018. p. 96. ISBN 9789231002441

6. Бородачев С. М. Математические методы статистики - учебник / Екатеринбург: УрФУ, 2012- 129 с. ISBN 978–5–321–02218–4

7. Ачилова Д.А. «Моделирование учебного процесса в ВУЗе. (На примере Начертательной геометрии и инженерной графики)», диссертация канд. пед. наук., 05.01.01./Ташкент, 2020. – 106 с.

8. Kuchkarova D.F., Achilova D.A. «Quality management of engineering graphics teaching», The 18th International Conference on Geometry and Graphics (ICGG2018), Politecnico di Milano, 3-7 August 2018 y, Milano, Italy #034 (E)

УДК 378:514
DOI: 10.25206/978-5-8149-3873-2-2024-988-996

Об опыте проведения занятий в рамках проекта «Цифровая кафедра» для студентов инженерного профиля
About the experience of conducting classes within the «Digital Department» project for engineering students

Е. В. Любчинов, Ю. Е. Ионова, Т. М. Мясоедова
Омский государственный технический университет, Омск, Россия

E. V. Lyubchinov, Yu. E. Ionova, T. M. Myasoedova
Omsk State Technical University, Omsk, Russia

Аннотация. В настоящей работе авторы делятся опытом проведения занятий для студентов в рамках курса дополнительного образования «Инженерный дизайн и программирование в САПР». Курс был нацелен в первую очередь на освоение навыков работы с цифровыми инструментами проектирования в области машиностроения в условиях нехватки часов на изучение основной дисциплины «Инженерная и компьютерная графика». В ходе обучения студенты подробно изучали инструментарий САПР, а также обучались элементам программирования и работе с API САПР. По завершению курса были выявлены некоторые «пробелы» в базовых знаниях геометро-графической подготовки студентов, что затрудняло освоение функционала САПР и дальнейшее изучение инструментов ее API. В работе также приводятся фрагменты выполненных студентами заданий и подготовленных преподавателями методических вспомогательных материалов.

***Ключевые слова*:** цифровая кафедра, высшее образования, геометро-графическая подготовка, анализ проблем, программирование, САПР

Аннотация. In this paper, the authors share their experience of conducting classes for students in the additional education course "Engineering Design and Programming in CAD". The course was primarily aimed at mastering the skills of working with digital design tools in the field of mechanical engineering in the context of a shortage of hours for studying the main discipline "Engineering and Computer Graphics". During the training, students studied CAD tools in detail, and were also taught programming elements and working with the CAD API. Upon completion of the course, some "gaps" in the basic knowledge of geometric and graphic training of students were identified, which made it difficult to master the functionality of CAD and further study of its API tools. The work also provides fragments of assignments completed by students and methodological auxiliary materials prepared by teachers.

***Ключевые слова*:** digital department, higher education, geometric-graphic training, problem analysis, programming, CAD.

I. Введение

Выпускникам технических специальностей по роду своей профессиональной деятельности необходимо понимать и самостоятельно уметь работать с чертежной документацией, создавать 3D-модели машин, механизмов, зданий и сооружений, при этом используя современные цифровые возможности. Всему этому учит дисциплина «Инженерная и компьютерная графика». И чтобы выпускники были конкурентоспособными на рынке труда, они должны иметь качественную геометро-графическую подготовку.

В научной литературе в последнее время часто поднимается вопрос о недостаточном уровне геометро-графической подготовки выпускников технических специальностей высших учебных заведений [1-5]. Например, одной из причин этой проблемы в работах [1,2] авторы отмечают сокращение часов на изучение дисциплины «Инженерная и компьютерная графика». В некоторых вузах этой дисциплине отводится всего три или даже две зачетные единицы [1,5] и таким образом, со слов авторов, дисциплина переходит в разряд изучаемых «на ознакомление» и многое из того, что ранее давалась в рамках аудиторных занятий сейчас уходит на исключительно самостоятельное изучение, либо не дается вовсе. Не стоит забывать и том, что преподаватели высшей школы уже давно сталкиваются с проблемой отсутствия у поступающих базовых знаний по школьному курсу черчения, что еще больше усугубляет проблему [1,2,5].

Еще ряд причин низкой геометро-графической подготовки выпускников подняли авторы в работе [6]. К этим причинам авторы относят сокращение кафедр геометро-графической подготовки в вузах и нехватку высококвалифицированных научно-педагогических кадров по данному профилю. Поглощение кафедр

начертательной геометрии и инженерной графики выпускающими кафедрами, слияние их с другими кафедрами, очевидно, не способствуют развитию кадрового потенциала в области научной специальности 2.5.1. «Инженерная геометрия и компьютерная графика. Цифровая поддержка жизненного цикла изделий». А именно эта специальность готовит научно-педагогические кадры непосредственно для кафедр геометро-графической подготовки.

В совокупности эти факторы дают низкий уровень приобретенных знаний выпускников. И если с отсутствием черчения в школах и сокращением кафедр инженерной графики рядовой преподаватель ничего сделать не может, то нехватка аудиторного времени заставляет преподавателей искать новые подходы, методики и способы обучения, которые позволили бы сохранить должное качество образования в рамках дисциплины «Инженерная и компьютерная графика» в текущих ограниченных условиях.

Стоит отметить, что в ходе освоения дисциплины «Инженерная и компьютерная графика» студент должен овладеть навыками работы с как минимум одной системой автоматизированного проектирования (САПР). С учетом того, что как правило курс дисциплины «Инженерная и компьютерная графика» стоит из трех разделов, а именно начертательной геометрии, инженерной графики и компьютерной графики, в условиях ограниченности часов чем-то приходится жертвовать. И зачастую такие жертвы осуществляются за счет часов на обучение инструментарию САПР, что безусловно является негативным фактором в век внедрения цифровизации и «бесчертежного» производства на отечественные промышленные предприятия [7-9].

Нехватка часов, как уже отмечалось выше, помимо ряда других вышеназванных факторов, ставит преподавателей в затруднительное положение. В настоящей работе авторы хотят поделиться опытом проведения дополнительных занятий для студентов ряда технических специальностей, которые добровольно записались на курс дополнительного профессионального образования «Инженерный дизайн и программирование в САПР», организованный кафедрой «Инженерная геометрия и САПР» Омского государственного технического университета. Данный курс был направлен на повышение уровня геометрической и графической грамотности студента, развитие его навыков работы с инструментами САПР, а также обучение студента основам программирования и его применения в проектно-конструкторской деятельности.

II. Постановка задачи

Курс «Инженерный дизайн и программирование в САПР» был открыт в рамках проекта «Цифровые кафедры», который входит в государственную программу «Приоритет-2030». Авторами курса выступили преподаватели кафедры «Инженерная геометрия и САПР» совместно с представителями индустриальных партнеров. При составлении программы курса ставились следующие основные задачи:

1) выявить и должным образом ликвидировать «пробелы» в геометро-графической подготовке студентов;
2) помочь студентам достичь высокого уровня владения инструментами современных российских САПР;
3) дать возможность студентам инженерных специальностей приобщиться к IT-сфере путем обучения их языку программирования и его применению при решении инженерно-конструкторских задач.

В настоящей работе авторы ставят перед собой целью не только поделиться опытом проведения дополнительных занятий, но и показать, насколько могут быть полезны дополнительные занятия для качественной геометро-графической подготовки студента и повышению его навыка работы с САПР до действительно высокого уровня. Интерес также представляет и изучение программирования студентами инженерного профиля и его последующее применение в оптимизации их будущей проектно-конструкторской деятельности.

III. Теория

На курс «Инженерный дизайн и программирование в САПР» в 2023 году поступило 30 студентов старших курсов (начиная с третьего курса и заканчивая пятым). Большая часть из них обучалась по следующим направлениям: 15.03.01 «Машиностроение», 15.03.02 «Технологические машины и оборудование», 15.03.05 «Конструкторско-технологическое обеспечение машиностроительных производств», 24.03.01 «Ракетные комплексы и космонавтика», 24.05.01 «Проектирование, производство и эксплуатация ракет и ракетно-космических комплексов», 24.05.02 «Проектирование авиационных и ракетных двигателей» и др. Как видно из перечня направлений – все поступившие студенты являются представителями технических специальностей. Все они на первом курсе успешно окончили изучение дисциплины «Инженерная и компьютерная графика». Данная дисциплина в Омском государственном техническом университете входит в дисциплины обязательной части учебных планов 38 направлений подготовки. В большинстве случаев она изучается 2 семестра, за которые студенты посетят 36 часов лекций и 54 часа практических занятий. В ходе изучения дисциплины студенты, как уже отмечалось выше, изучают следующие разделы: начертательную геометрию, инженерную и компьютерную

графику. Содержание разделов является по большей части классическим и приведение в рамках данной работы подробного описания изучаемых тем будет излишним.

Сам же курс «Инженерный дизайн и программирование в САПР» состоял из шести модулей и длился один учебный год (с октября 2023 г. по июнь 2024 г.). Общая продолжительность курса составила 250 часов. Занятия проходили в очном формате и по итогу прохождения курса студенты должны были выполнить выпускную итоговую работу.

Первый модуль назывался «Инженерный дизайн и электронная модель изделия» и был направлен на совершенствование навыков работы студентов в САПР. Здесь слушатели курса создавали 3D-модели (а также сборочные единицы) не только с помощью инструментов твердотельного моделирования, но и с помощью каркасного и листового моделирования, а также обучались работе с различными встроенными библиотеками. Помимо прочего студенты в рамках лекционных и практических занятиях осваивали работу с инструментами оформления электронной модели изделия согласно ГОСТ 2.052-2006 «Электронная модель изделия» и ГОСТ 2.053-2013 «Электронная структура изделия». Этому однозначно поспособствовал тот факт, что система КОМПАС-3D имеет развитые инструменты для оформления электронной модели. Второй модуль – «Разработка проектно-конструкторской документации в CAD-системах», был логическим продолжением первого. Здесь слушатели курса создавали ассоциативные чертежи по полученным ранее моделям и оформляли их в соответствие с требованиями ЕСКД. Третий модуль был посвящен автоматизированным прочностным расчетам в ходе цифрового инженерного анализа деталей и сборочных единиц средствами САПР.

Предыдущие три модуля для студентов были так или иначе знакомы и на этих занятиях они помимо получения новых знаний и навыков совершенствовали уже приобретённые ранее. Тем не менее, следующие три модуля, связанные напрямую с программированием, для большинства были абсолютно новыми. Тема о том, нужны ли навыки программирования инженеру-проектировщику или инженеру-конструктору, давно поднимается в научной литературе [9,10]. Авторы говорят о том, что этот навык как минимум будет полезен и сделает такого специалиста более конкурентоспособным на рынке труда. Также авторы отмечают, что обучение основам программирование важно еще и потому, что помимо создания цифровых моделей различных машин, механизмов или зданий и сооружений, задачей для инженера является и оптимизация самого процесса проектирования, прежде всего за счет повышения эффективности выполнения однотипных и многократно повторяющихся операций. Для четкого понимания надобности изучения программирования, студентами в течение курса был прослушан ряд лекций, где была продемонстрирована наглядная польза программирования в их будущей конструкторской деятельности на нескольких примерах, а все последующие практические занятия были нацелены на изучение основ программирования и его внедрения в инженерную деятельность.

Четвертым модулем стал курс основ программирования на языке Python. Как понятно из его названия, модуль ставил перед собой целью обучить студентов основам программирования. Язык программирования Python здесь был выбран не случайно. Во-первых, Python среди языков программирования по праву считается наиболее дружелюбным для новичка и позволяет освоить базовые аспекты программирования в короткие сроки. Во-вторых, API (application programming interface – программный интерфейс приложения) КОМПАС-3D поддерживает несколько языков программирования, одним из которых является Python. И третья причина – это наличие встроенной утилиты «КОМПАС-МАКРО» непосредственно в сам КОМПАС-3D, которая сопровождается приложением «PyScripter» (идущим в комплекте вместе с КОМПАС-3D). Эти приложение и утилита позволяют создавать и редактировать макросы непосредственно в КОМПАС-3D на языке Python.

Пятый модуль был направлен на изучение основ работы с API КОМПАС-3D. В ходе этого модуля слушатели курса изучали синтаксис свойств и методов API, учились создавать и редактировать макросы, а также работать с SDK (software development kit — комплект для разработки программного обеспечения). Шестой модуль был сконцентрирован на применение полученных знаний и навыков работы с API для решения несложных прикладных задач в работе инженера. Основной целью здесь было научить студентов автоматизировать некоторые рутинные задачи во время работы с САПР, чем в последствие повысить скорость и эффективность разработки 3D-моделей и чертежей. В дальнейшем решение подобных задач должно было помочь ребятам выйти на выпускную итоговую работу.

IV. Результаты экспериментов

В рамках изучения первых модулей, напрямую посвященных работе с 3D-моделями, акцент был сделан на максимально полное раскрытие потенциала инструментария САПР. Задания были построены таким образом, чтобы при создании 3D-моделей и сборочных единиц задействовать как можно большее число различных команд системы КОМПАС-3D. Например, в ходе изучения раздела «Твердотельное моделирование» помимо классических команд «Элемент выдавливания» и «Вырезать выдавливанием» активно использовались и другие команды для работы с телами: «Булева операция», «Оболочка» и т.п.

Зачастую в рамках стандартного курса студенты изучают только инструментарий твердотельного моделирования и работу с библиотекой стандартных крепежных изделий. Здесь же большой акцент был сделан и на другие способы моделирования, а именно на поверхностное и листовое. Следует отметить также, что при работе с поверхностным моделированием использовались не только команды для создания и редактирование поверхностных форм, но также и активно велась работа с пространственными кривыми (сплайнами).

Еще одной немаловажной составляющей при создании 3D-моделей был акцент на работу с параметризацией, а именно с геометрическими и размерными ограничениями в эскизах модели. Это было сделано потому, что в дальнейшем навыки работы с параметризацией применялись слушателями на третьем модуле курса в ходе инженерного анализа в цифровой среде для определения оптимальной конструкции изделия, а также в ходе последующей работе с API системы КОМПАС-3D в заключительной части курса.

Помимо прочего студенты повысили свои навыки в работе с автоматизированным получением технической чертежной документации за счет активного использования инструментов оформления 3D-модели как электронной модели изделия. Студенты, как продемонстрировано на примере на рис. 1, заполняли атрибуты (например, шифр, наименование, материал и т.п.) создаваемых моделей, учились грамотно выводить технические требования на поле модели (в том числе это касается и наиболее наглядной простановки размеров на самой 3D-модели), создавать и сохранять сечения модели и многое другое. В дальнейшем при получении технических чертежей с автоматическим переносом всех атрибутов модели, размеров, позиций, технических требований и пр., необходимость в редактирование чертежа была минимальна.

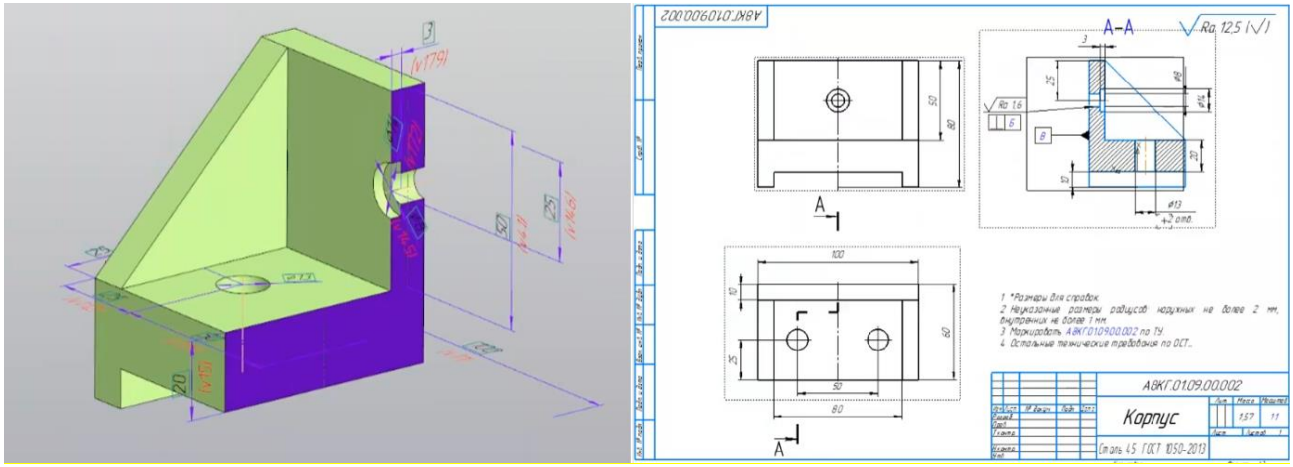


Рис. 1. Пример оформления студентами электронной модели изделия и полученного с нее ассоциативного чертежа

Преподаватели курса старались строить все задания по принципу отработки определённых инструментов САПР начиная с простых упражнений (см. рис 2), а по окончанию темы завершать заданиями для самостоятельной работы уже более высокого (зачастую олимпиадного) уровня сложности (см. рис.3). И стоит отметить. что слушатели успешно с этими заданиями справлялись.

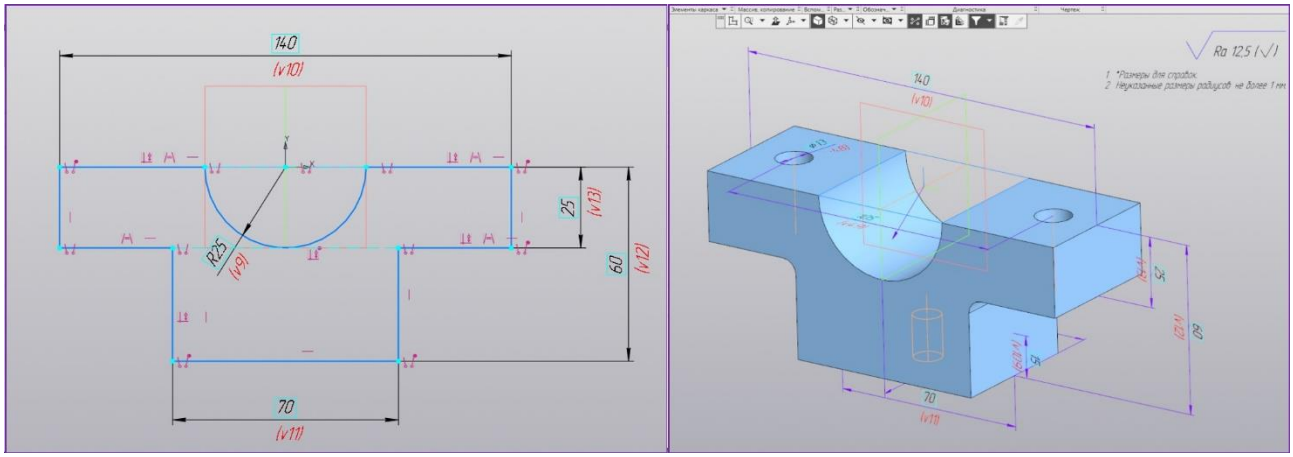


Рис. 2. Пример простого задания на тему «Твердотельного моделирования»

Подводя промежуточные итоги по прошествию первой половиной курса, у студентов был выявлен ряд «пробелов» в знаниях. Поскольку практически каждая команда в САПР так или иначе имеет под собой какую-либо геометрическую основу, наиболее частой выявленной проблемой стало незнание и непонимание некоторых базовых геометрических понятий, например:

– при работе с геометрическими ограничениями многим студентам было не понятно значения термина «коллинеарность»;

– при попытке создания кривой линии в эскизе слушатели через окно поиска команд часто пытались найти команду «Кривая», совершенно не понимая значения термина «Сплайн». Также для многих было не ясно почему же существует команда «Коническая кривая» и в чем ее особенность;

– само понятие кривизны (а также касательной, нормали), гладкости стыковки в местах соединения сегментов сплайна, также было студентами либо благополучно забыто, либо не усвоено в ходе изучения дисциплины ранее в рамках основного курса обучения;

– при работе с каркасным и поверхностным моделированием некоторые студенты не смогли дать определение понятию «Линейчатая поверхность», а именно с таким названием в КОМПАС-3D существует команда, которой в ходе обучения на курсе студентам необходимо был пользоваться.

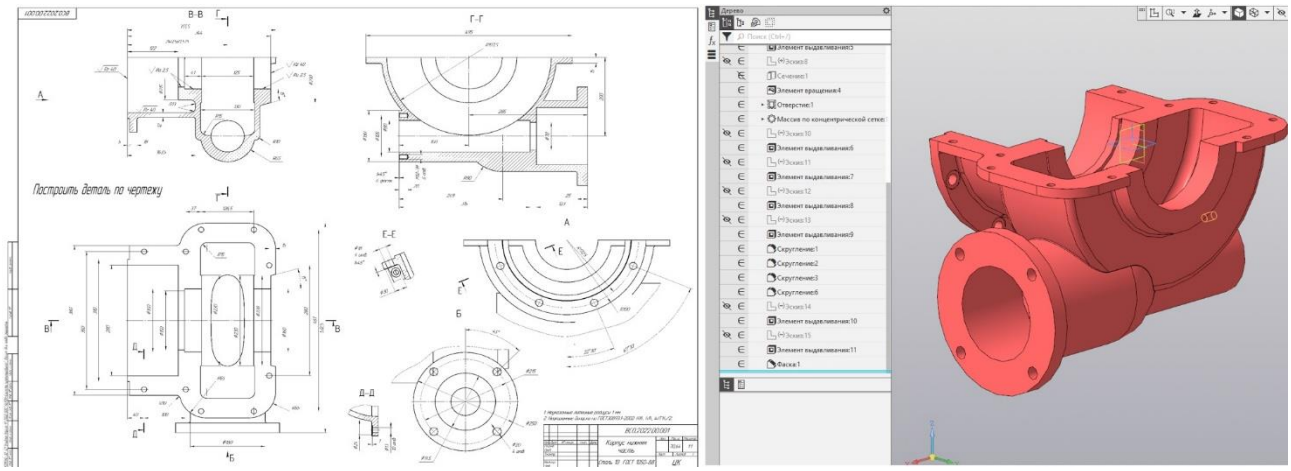


Рис. 3. Пример заключительного задания на тему «Твердотельного моделирования»

Естественно, это не все «пробелы» в знаниях, выявленные у студентов, а лишь наиболее часто встречаемые. Однако это уже заставляет задуматься о том, что многие базовые и фундаментальные геометрические понятия, смысл которых заложен непосредственно в функционал САПР, студентами к третьему курсу обучения забываются или попросту незнакомы.

Чтобы максимально возможно ликвидировать «пробелы» в знаниях и повышать геометрическую грамотность студентов в ходе обучения основам языка программирования Python, помимо изучения базовых тем, таких как переменные, функции, операторы, циклы и т.п., преподавателями кафедры были разработаны задания, которые так или иначе касались математической и, в частности, геометрической составляющей подготовки студентов. Например, в ходе работы с библиотекой «matplotlib» студенты по параметрическим уравнениям делали визуализацию различных кривых второго порядка, сплайнов, поверхностей. В качестве примера выполнения подобного задания на рис. 4 представлен вывод поверхности сферы по ее параметрическим уравнениям в оконном режиме приложения, созданного непосредственно студентами. Интерфейс модуля создавался с помощью встроенной библиотеки «Tkinter», позволяющей на языке Python создавать приложения с графическим интерфейсом.

Завершив обучение базовым аспектам программирования, студенты приступали к изучению API системы КОМПАС-3D. Работа с API для только что начавших изучать программирование студентов безусловно была крайне сложным испытанием. Для того, чтобы помочь студентам, преподавателями был подготовлен ряд электронных методических материалов, разбитых на уроки, где максимально детально рассматривался один из вариантов задания. Для лучшего понимания и упрощения поиска ошибок был приведен полный листинг кода, а также проиллюстрированы все необходимые подсказки для ориентации в иерархии SDK. Примеры фрагментов такого электронного методического материала представлены на рис. 5. По этим материалам студенты сначала с преподавателем полностью выполняли представленный там пример, разбирая каждую строчку кода, а уже затем получали индивидуальное задание, где самостоятельно реализовывали поставленную перед ними задачу.

В результате работы с API системы КОМПАС-3D студенты выполняли ряд интересных работ по созданию собственных:

– библиотек фрагментов (эскизов) или моделей на основе базовых возможностей системы;
– библиотек шаблонов с помощью Менеджера шаблонов;
– модулей построения чертежей, оформления чертежей, деталей, сборочных единиц и др.

Отчасти сложность в освоении работы с API системы КОМПАС-3D для студентов, а именно в ходе работы по созданию макросов, была вызвана непониманием геометрических основ (как отмечалось выше), а также сложностью самой иерархической структуры SDK системы.

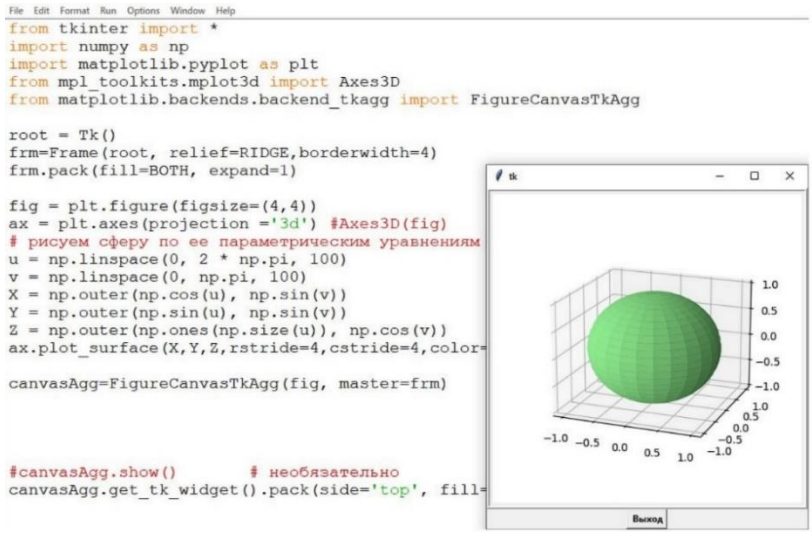


Рис. 4. Визуализация поверхности сферы по ее параметрическим уравнениям в ходе изучения языка программирования Python

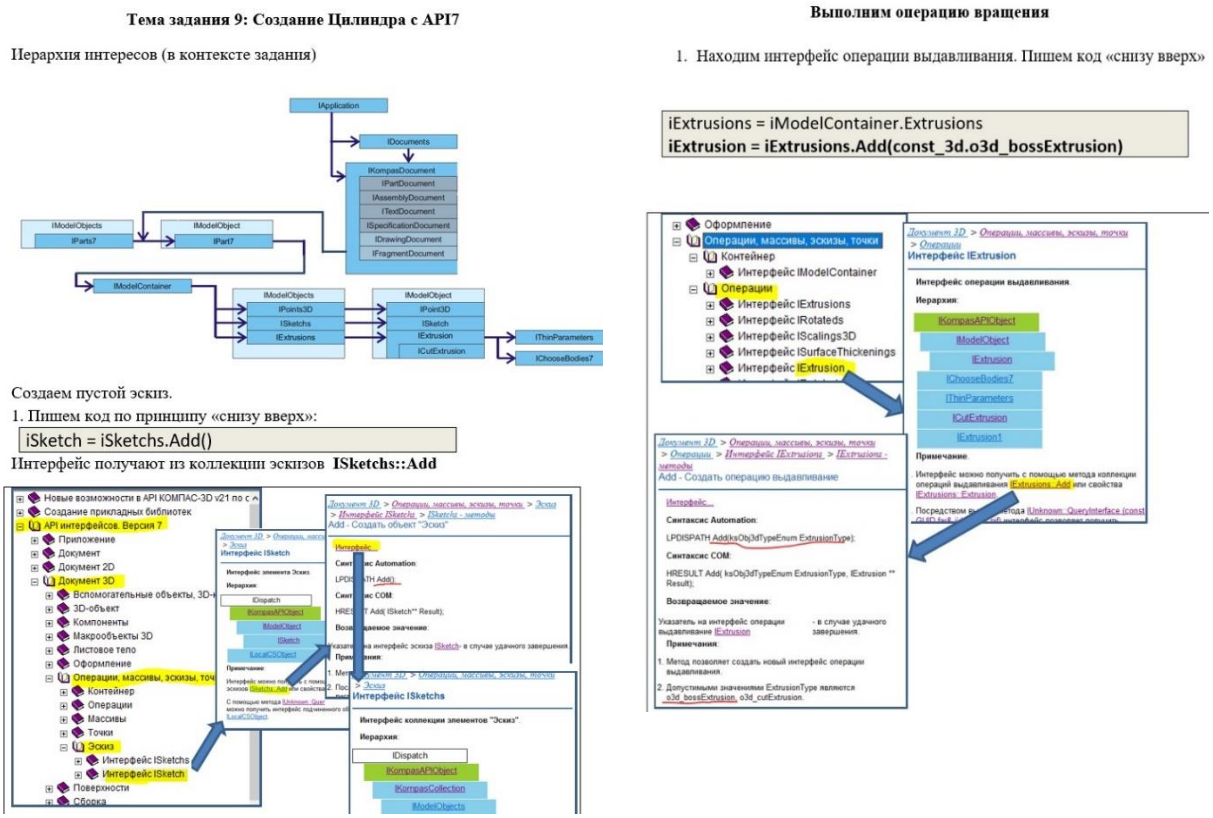


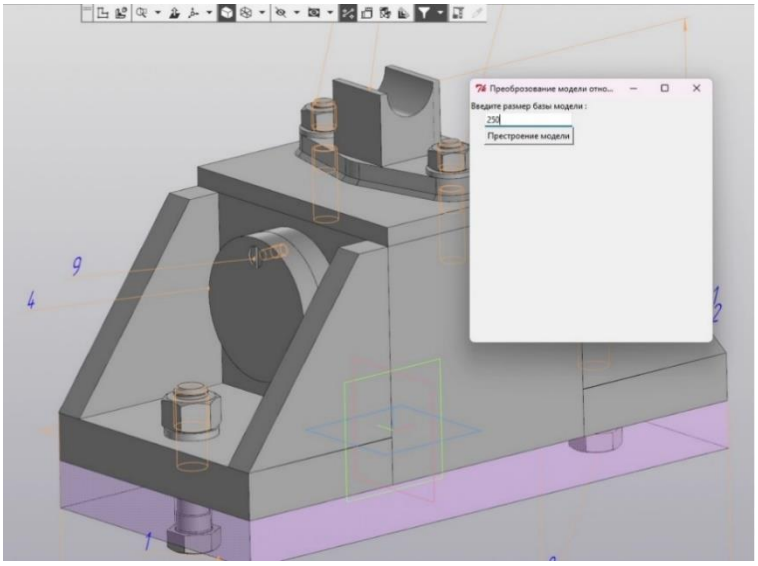
Рис. 5. Примеры выкладок из методических материалов для студентов, изучающих работу с API системы КОМПАС-3D

В конце курса каждый студент должен был продемонстрировать полученные навыки на практике, выполнив выпускную итоговую работу. Суть работы была для всех студентов идентична, отличия были только в исходных данных и творческой составляющей. В ходе задания требовалось по данному сборочному чертежу:

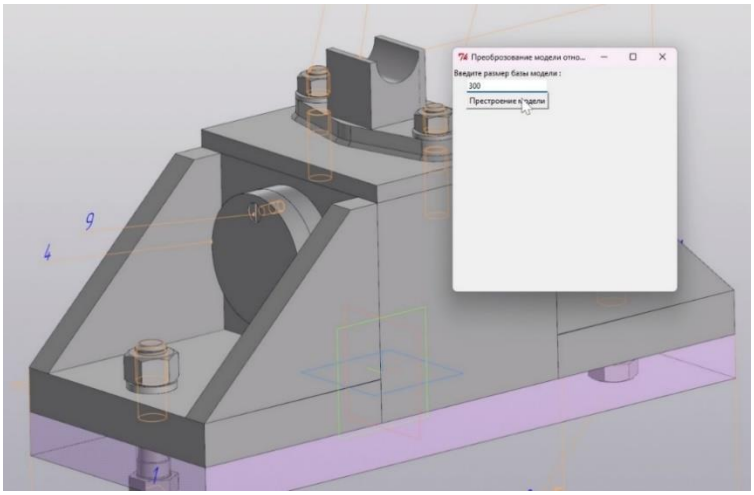
1) разработать модели деталей, сборочных единиц, и оформить их как электронную модель изделия;
2) выполнить все необходимые технические чертежи этих деталей и сборочных единиц;
3) выполнить по указанным преподавателем исходным данным расчет на прочность либо отдельной детали, либо сборочной единицы в целом с помощью цифровых инструментов инженерного анализа;
4) на выбор студента создать макрос или интерфейс для автоматизации какой-либо операции с помощью языка программирования Python и API системы автоматизированного проектирования.

В качестве примера на рис. 6 представлен фрагмент выпускной работы одного из студентов. Студент сделал внешний интерфейс и с помощью параметризации создал связь нескольких размеров сборочной единицы от одного единственного размера габаритной длины. При вводе в окно интерфейса значения длины сборочной единицы – модель успешно перестраивалась и изменяла остальные размеры: при значении габаритного размера 250 мм результат выполнения модели представлен на рис. 6(а), а при значении 300 мм – на рис. 6(б).

Очевидно, что приведенный пример хоть и относительно несложен в своей реализации, но он уже показывает, что студентами получена базовая подготовка в работе с API систем САПР, что в дальнейшем может быть использовано ими непосредственно в будущей проектно-конструкторской деятельности.



(а)



(б)

Рис. 6. Интерфейс параметризации размеров детали, выполненный студентом в ходе выпускной итоговой работы

V. Обсуждение результатов

Пройденный студентами курс «Инженерный дизайн и программирование в САПР» показал, что как минимум для освоения инструментария САПР на хорошем уровне классического курса дисциплины «Инженерная и компьютерная графика» недостаточно. Помимо того, что в ходе основного курса студенты осваивают лишь узкий набор навыков работы с САПР (и как правило это инструменты «Твердотельного моделирования»), они также имеют «пробелы» и в базовой теоретической подготовке. Однако, стоит отметить, что при регулярных дополнительных практических занятиях в рамках курса ребята в конечном итоге выходили на уровень решения задач по 3D-моделированию олимпиадного уровня, что безусловно говорит о положительном результате проведенной работы.

Нельзя не отметить тот факт, что и обучение программированию, а также работа с API системы КОМПАС-3D были, хоть и не без труда, но все же освоены студентами. По результатам обучения программированию и работе с API можно сделать вывод, что студентами был освоен базовый уровень программирования в САПР и в дальнейшем они могут развивать полученные навыки, расширяя возможности стандартного модуля систем автоматизированного проектирования.

В качестве обратной связи был проведен опрос окончивших курс студентов. Большинство ребят отметило следующие плюсы:

– освоили большое количество новых команд и способов для создания 3D-моделей различных форм и конструкций;

– узнали о возможностях оформления 3D-моделей как электронных моделей изделия с последующим автоматизированным переносом атрибутивной и технической информации на чертеж;

– навыки, полученные в ходе цифрового инженерного анализа в САПР, пригодились на других предметах и курсовых проектах, проходивших параллельно с обучением на курсе;

– для большинства стал открытием тот факт, что при работе с САПР можно создавать макросы и автоматизировать работу с помощью программирования и API;

– нескольким студентам понравилось обучаться программированию, и они задумались о дальнейшем продолжении обучения в магистратуре на ИТ-специальности.

VI. Выводы и заключение

В работе авторы хотели поделиться опытом проведения дополнительных занятий по работе студентов с современными отечественными САПР. Были выявлены «пробелы» в геометро-графических знаниях студентов, которые при нынешнем объеме часов, выделяемых на освоение дисциплины «Инженерная и компьютерная графика» решить представляется крайне затруднительным. При этом дополнительные занятия в рамках курса «Инженерный дизайн и программирование в САПР» предоставили возможность не только ликвидировать эти «пробелы», но значительно улучшили навыки работы студентов с инструментарием САПР. Студенты помимо изучения широкого набора команд САПР с точки зрения пользователя, изучали также работу с API САПР и тем самым проявляли свои знания уже непосредственно в автоматизации и оптимизации (создании макросов, плагинов и пр.) своей будущей проектно-конструкторской деятельности. Обратная связь, полученная от студентов, показала, что данные занятия были для них не только полезны, но и выявили желание более подробно погрузиться в возможности современных отечественных САПР.

Список литературы

1. Борисова О. А. Проблемы при преподавании дисциплины «Инженерная и компьютерная графика» в современных условиях в вузе: сборник трудов конференции. // Актуальные вопросы гуманитарных и социальных наук: от теории к практике : материалы III Всерос. науч.-практ. конф. с междунар. участ. 2023. С. 129-133.

2. Бойков А. А., Егиазарян К. Т., Ефремов А. В., Кадыкова Н. С. Проблемы геометро-графической подготовки студентов вузов // Геометрия и графика. 2023. Т. 11, № 1. С. 4-22.

3. Таланова М. Б. Особенности преподавания инженерной и компьютерной графики // Балтийский гуманитарный журнал. 2015. № 1(10). С. 146-148.

4. Кирпичникова Н. Н., Сулина О. В. Интенсификация и преемственность обучения инженерно-графическим дисциплинам в техническом вузе // Современные проблемы науки и образования. 2015. № 1-1. С. 277.

5. Горбатовская Т. А. Основные проблемы преподавания графических дисциплин в вузе и возможные пути их решения // Инновационное развитие. 2018. № 4(21). С. 205-207.

6. Ротков С. И., Конопацкий Е. В., Панчук К. Л. Проблема сокращения кафедр геометро-графической подготовки в университетах России // Омский научный вестник. 2023. № 4(188). С. 29-37.

7. Демин А. А., Хабаров Р. А. Анализ функций "Бесчертежного проектирования" в концепции "Бережливого производства" // Информационные технологии в проектировании и производстве. 2021. № 3(183). С. 16-22.

8. Березанский Д. П. Проблемы внедрения безбумажного обмена данными на предприятиях ОПК // Известия Тульского государственного университета. Технические науки. 2022. № 1. С. 23-34.

9. Самаркина Е. И., Самаркин А. И., Дмитриев С. И., Евгеньева Е. А. Электронная модель изделия как носитель конструкторско-технологической информации // Сборка в машиностроении, приборостроении. 2019. № 6. С. 284-288.

10. Чегулова А. С., Филинских А. Д. Обзор систем автоматизированного проектирования: возможности и применение в области API // КОГРАФ - 2020 : сборник материалов 30-й Всероссийской научно-практической конференции по графическим информационным технологиям и системам, Нижний Новгород, 13–16 апреля 2020 года. 2020. С. 277-282.

11. Альшакова Е. Л., Альшакова Е. А. Программная реализация параметрических компонентов САПР // Информатика и технологии. Инновационные технологии в промышленности и информатике : сборник научных трудов международной научно-технической конференции, Москва, 04–06 апреля 2016 года. 2016. С. 218-221.

УДК 378.145
DOI: 10.25206/978-5-8149-3873-2-2024-997-1002

Стратегическое планирование как инструмент методической работы при проектировании курсов графической подготовки в вузах
Strategical planning as a tool for metodical work while planning enginering drawing courses in universities

Е. О. Подчасов1, Е.Н. Сломинская2, Н.Г. Суркова1
1МГТУ им. Н.Э. Баумана, Москва, Россия
2Калужский филиал МГТУ им. Н.Э. Баумана, Калуга, Россия

E. O. Podchasov1, E. N. Slominskaya2, N.G. Surkova1
[1]Bauman Moscow State Technical University, Moscow, Russian Federation
[2]Kaluga branch of Bauman Moscow State Technical University, Kaluga, Russian Federation

Аннотация. Описан опыт применения подходов бизнес-планирования к поиску перспективных направлений трансформации преподавания фундаментальных, естественнонаучных и общеинженерных дисциплин. На примере стратегической сессии в формате быстрого Форсайта, проведенной 24.04.2024 на базе кафедры МК4 Калужского филиала МГТУ им. Н.Э. Баумана, показан набор первичных материалов, получаемых при адаптации методов бизнес-планирования к решению методических задач. Показан пример анализа получаемых материалов и выбора критериев значимости факторов, оказывающих влияние на определение направлений трансформации образовательного процесса. Приводится типовая схема проекта трансформации образовательного процесса по дисциплине «Инженерная и компьютерная графика» с учетом унификации подходов в головном вузе и филиале. Обоснована необходимость применимости данного метода к планированию трансформации иных фундаментальных, естественнонаучных и общеинженерных дисциплин.

***Ключевые слова*:** Быстрый Форсайт, образовательный опыт, рабочие программы дисциплины, учебные планы, цифровая трансформация образования.

Аннотация. The experience of usage of business-planning methods to determine the perspective ways of improvement of teaching fundamental, natural science and engineering courses is considered. Strategic session in form of Rapid Foresight that has been held at 24.02.2024 in Kaluga branch of BMSTU is shown as example with set of raw session materials being obtained when using a business methods for methodical cases. The example of material analysis and selection of valuable criteria is shown. The typical plan of educational transformation of Engineering drawing is given considering unification of teaching in main campus and in branch. The necessity of usage such method for planning transformation of other courses is estimated/

***Ключевые слова*:** Rapid Foresight, Educational experience, Course program, Educational plans, Digitalizing of education.

I. Введение

Методы стратегического планирования широко применяются в управленческой деятельности разных отраслей экономики. Они позволяют прогнозировать события и принимать управленческие решения в условии недостатка информации о протекающих процессах и переменчивости действия внешних факторов. Проектирование образовательной среды ввиду имеющихся организационных и экономических факторов изменчивости так же требует при своем проектировании учет нечетко сформулированных критериев в условиях неизвестности воздействия факторов. Вследствие этого применение методов проектного управления становится уместно при установлении целей, задач и траекторий развития как отдельных дисциплин и краткосрочных программ дополнительного образования, так и основных образовательных программ и целых образовательных стандартов [1, 2].

Среди различных инструментов стратегического планирования обыкновенно выделяют такие методы, как SWOT-анализ, STEP-анализ, анализ пяти сил Портера, VRIO-анализ и прочие. Отдельно среди таких методов следует выделить метод Быстрого Форсайта (Rapid Foresight), получивший распространение в разных областях экономической деятельности, но изначально разработанный для трансформации образовательной среды. Разработчики характеризуют эту методику как «инструмент доя прогнозирования и формирования будущего,

позволяющий получать за короткий срок прогнозы высокой точности относительно будущего, путей их достижения и объединять людей для реализации их представлений о будущем»[3].

В данной статье приводится анализ результатов стратегической сессии в формате быстрого Форсайта, проведенной в калужском филиале МГТУ им. Н.Э. Баумана 25.04.2024.

II. Постановка задачи

Стратегическая сессия состоялась 14.02.2024 на базе кафедры МК4 Калужского филиала МГТУ им. Н.Э. Баумана. К участию в сессии привлекались работники Научно-учебного комплекса «Робототехника и комплексная автоматизация» МГТУ им. Н.Э. Баумана, работники кафедры МК4 «Инженерная графика» калужского филиала МГТУ им. Н.Э. Баумана, Представители ООО «АСКОН». Проведение стратегической сессии показано на рис. 1.

Рис. 1. Проведение стратегической сессии в формате Rapid Foresigt

Информация о составе участников приведена в таблице 1.

ТАБЛИЦА 1
УЧАСТНИКИ СТРАТЕГИЧЕСКОЙ СЕССИИ

Категория участников		**Количество**
Всего участников сессии		13
Из них от организаций		-
	МГТУ им. Н.Э. Баумана	6
	КФ МГТУ им. Н.Э. Баумана	6
	ООО «АСКОН»	1
Из них по трудовым функциям		-
	Научно-педагогические работники	12
	Руководители структурных подразделений	3
	Исполняющие административные функции, не являющиеся руководителями подразделений	4
Из них по профилю преподаваемых дисциплин		-
	Преподают графические дисциплины	10
	Преподают родственные механические дисциплины	2
	Преподают родственные информационные дисциплины	4

При проведении сессии участник были разделены на три группы [4, 5], таким образом, чтобы каждая из групп включала в состав и представителей НУК РК, так и КФ МГТУ им. Баумана. Каждая группа включала одного

руководителя структурного подразделения. Формирование групп только из работников, исполняющих административно-управленческие функции, не допускалось.

Общая тематика Форсайт-сессии была объявлена как «Цифровая трансформация общеинженерной подготовки в МГТУ им. Н.Э. Баумана».

III. Теория

При проведении Форсайт-сессии предполагалось формирование стратегического плана в виде трех т.н. карт – графических документов, содержащих единицы предполагаемого наполнения плана развития и взаимосвязи между ними.

Карта будущего показывает предполагаемые варианты трансформация образовательной среды и ее окружения совместно с факторами, влияющими на эту трансформацию. В рамках Форсайт-сессии традиционно предполагается разделение факторов на группы по этапам сессии с формированием отдельных элементов в виде так называемых трендов, технологий, форматов, нормативных актов, трансформируемых по итогам сессии в угрозы и возможности. Все предложенные участниками элементы записываются и оформляются на бланках соответствующих карт.

Карта стейкхолдеров отображает потенциальных участников формирования будущего образовательного процесса. В зависимости от степени вовлеченности стейкхолдеров в процесс трансформации образования и степени их влияния и типа позиции (консервативная/прогрессивная) на графике выставляются точки, соответствующие каждому из типов участников. Их положение на диаграмме определяет их вовлечение в процесс стратегического планирования и определяет этапы, на которых это вовлечение является необходимым.

Карта проекта визуализирует действия, необходимые для достижения планируемых результатов трансформации образовательной среды с учетом факторов, выявленных при составлении карты будущего и степени участия акторов, выявленных в карте стейкхолдеров. На данном этапе определяются предполагаемый порядок выполнения проектных действий, выбранные этапы ранжируются по степени значимости и приоритетности, в зависимости от чего распределяются трудовые и материальные ресурсы.

IV. Результаты экспериментов

При формировании карты будущего участниками было совокупно предложено 46 карт. Итоговая карта имеет вид, приведенный на рис 2.

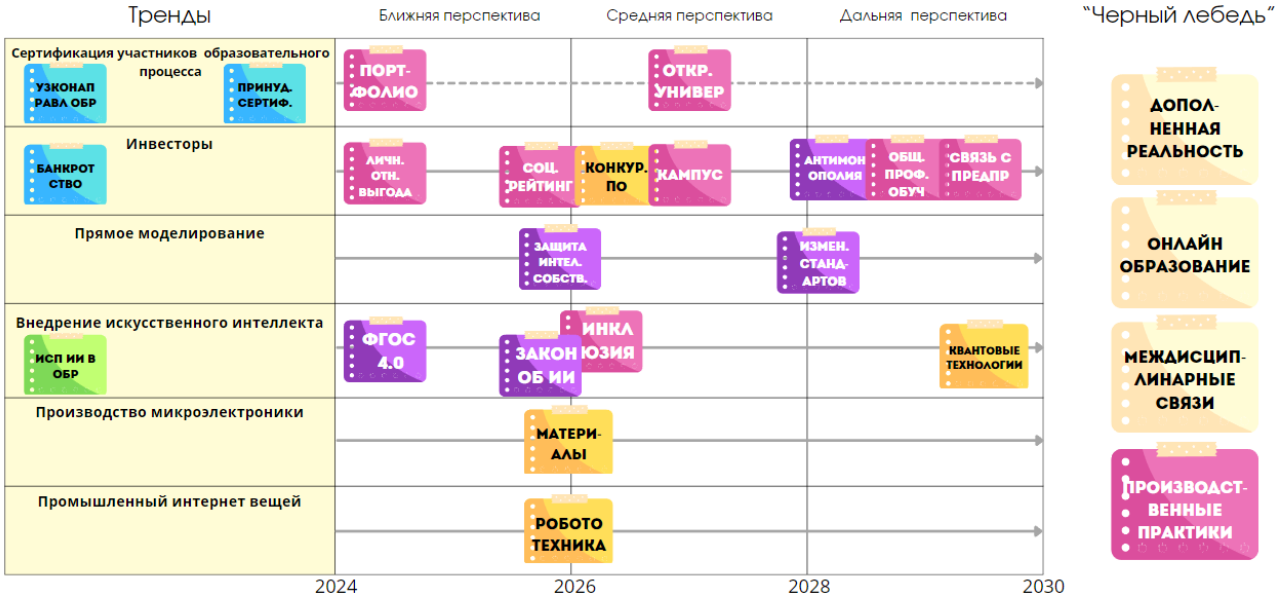


Рис. 2. Карта будущего, полученная по результатам стратегической сессии.

Карта стейкхолдеров, полученная в результате сессии, приведена на рис. 3.

По результатам проведения сессии был сформирован проект трансформации образовательных курсов графической подготовки в МГТУ им. Н.Э. Баумана и его калужском филиале. Особенностью данного проекта является опора на систему ДПО, как способ «выращивания» независимых модулей, из которых, в зависимости от требований образовательных стандартов, могут быть построено обучение графическим дисциплинам студентов разных направлений подготовки.

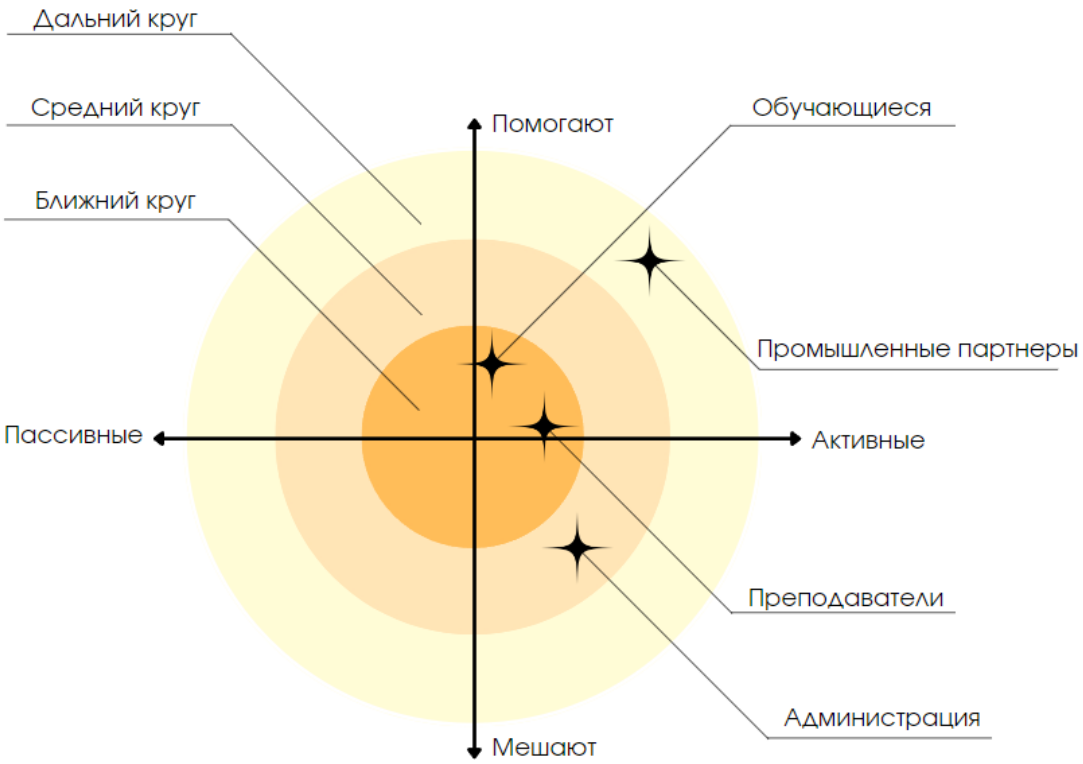


Рис. 3. Карта стейкхолдеров, полученная по результатам стратегической сессии.

План реализации проекта представлен на рис. 4, при этом пункты плана, обведенные сплошной линией, должны обладать приоритетной важностью при распределении ресурсов, в то время как пункты без обводки могут быть реализованы за счет остаточных ресурсов, или отложены на более позднее время и реализовываться после достижения горизонта планирования.

Рис. 4. Проект трансформации курса инженерная графика

V. Обсуждение результатов

Полученные при формировании карты будущего отдельные карты можно тематически разделить на укрупненные группы, объединенные тематически или организационно [6]. Распределение карт по группам приведено на основе текстового анализа полученных карт при проведении Форсайт-сессии материалов. Группировка осуществлялась на основе содержания в ответах схожих ключевых слов. В группу прочее объединены предложенные варианты, ключевые слова которых встречались по одному разу. Графическое распределение категорий изображено на рис. 5

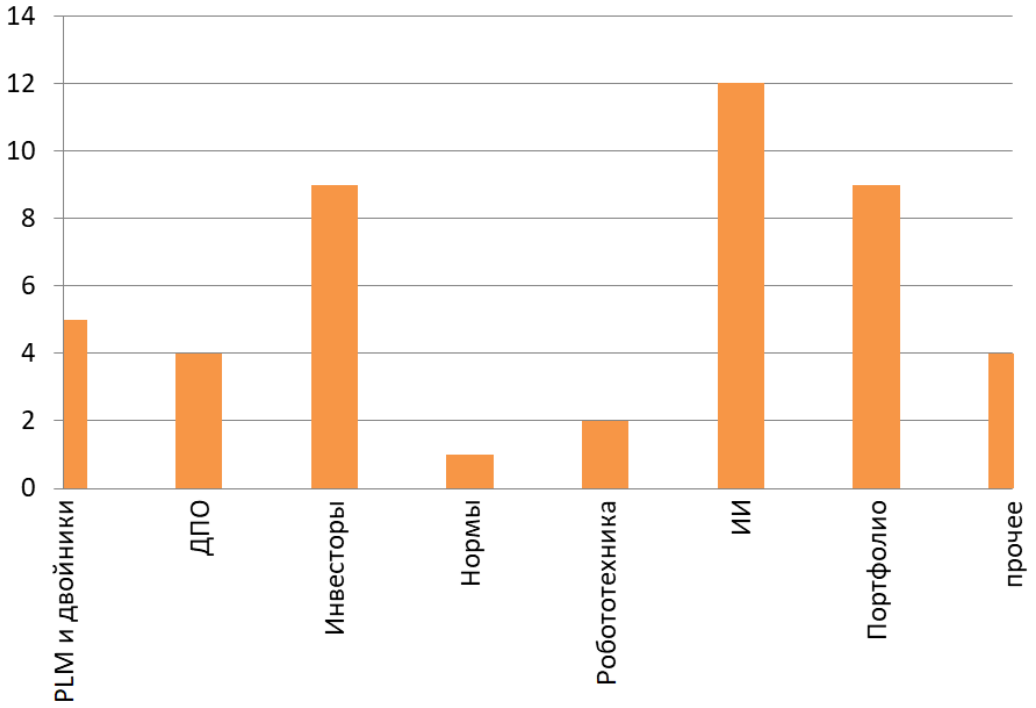


Рис. 5. Распределение карт сессии по категориям

В ходе обсуждения итогов стратегической сессии после подготовки карты проекта, участниками сессии предложено реструктуризировать группы карт таким образом, чтобы их состав больше соответствовал разработанному проекту. Для этого столбцы «Портфолио» и «PLM и двойники» были объединены в одну категорию трансформации обучения. Предполагается, что в рамках обучения студент должен разрабатывать не просто портфолио, а собственную MDM-библиотеку деталей, узлов и конструкций, которая после выпуска будет являться его рабочим инструментом, и сможет быть легко интегрирована в цифровую среду его предприятия. Наличие такой библиотеки, по мнению участников стратегической сессии, повысит востребованность выпускника на рынке труда, а так же темпы научно-технического развития и скорость конструкторско-технологической подготовки производств.

Категорию «Нормы» участниками сессии предложено исключить, входящие в нее карты включены в категорию прочее.

Проблемой проведенной Форсайт-сессии является неоднородность состава участников и неполная репрезентативность (превышение доли работников, исполняющих административно-управленческие обязанности в сравнении с составом кафедр) для доказательства справедливости выводов, полученных в данной сессии, проведен статистический анализ результатов.

Чтобы показать возможность использования полученных результатов в планировании трансформации учебного процесса, проведена статистическая обработка качественных результатов сессии[7].

Каждым участником сессии предложено хотя бы по одному понятию, группирующемуся в одну из именованных категорий, однако все карты, отнесенные в категорию прочего, выработаны группой с долей участи административных работников ниже средней по сессии. Предполагая фактор участия управленцев незначимым, первая гипотеза сформулирована следующим образом: «Исполнение участником административных функций не влияет на частоту предложения понятий, не выделяемых в именованную категорию». Для проведения построена таблица 2 совместимости критериев.

ТАБЛИЦА 2
ТАБЛИЦА СОВМЕСТИМОСТИ КРИТЕРИЕВ

	Нет ответов прочее	**Есть ответы прочее**	**Итого**
Администраторы	5	3	8
Не администраторы	3	1	4
Итого	8	4	12

Для первой статистической гипотезы вычислен критерий χ^2 с поправкой Йетса. Значение рассчитанного критерия равно 0,0469. Критическое значение критерия при уровне значимости 0,05 и числу степеней свободы p = 0,05 составляет 9,488. Значение критерия меньше критического, следовательно, гипотеза принимается, и исполнение административной работы в данной выборке не влияет на классифицируемость предложений.

VI. Выводы и заключение

Результаты стратегической сессии показывают готовность преподавателей МГТУ им. Н.Э. Баумана и его Калужского филиала к трансформации программ графической подготовки как в рамках дополнительного профессионального образования, так и в длительной перспективе – в рамках основных образовательных программ высшего образования. Анализ полученных результатов показывает важность по мнению преподавателей МГТУ им. Н.Э. Баумана перехода к работе в рамках экосистем проектирования и создания персональных проектных библиотек-портфолио. Среди проблем трансформации образования выделено ограниченное количество ресурсов, что приводит к сложностям разработки комплектов оценочных средств больших объемов; в рамках потенциального решения данной проблемы видится использование искусственного интеллекта, применению которого в учебных целях предложено более четверти всех влияющих факторов.

Статистическая обработка показывает допустимость решения задач планирования трансформации программ обучения на Форсайт-сессиях с небольшим числом участников из преподавательского актива, даже при сравнительно высоком представительстве администраторов.

С учетом перечисленных фактов, видится рациональным проведение аналогичных сессий с участием иных вузов с целью определения схожести видения направлений развития и определения возможности частичного объединения ресурсов.

Источник финансирования. Благодарности

Авторы благодарят руководителя научно-учебного комплекса «Робототехника и комплексная автоматизация» Г.В. Щащурина и центр дополнительного образования МГТУ им. Н.Э. Баумана в лице директора М.В. Стояновой за создание условий для организации стратегической сессии. Авторы выражают глубокую благодарность сотрудникам кафедры «Инженерная графика» Калужского филиала МГТУ им. Н.Э. Баумана и кафедр «Инженерная графика», «Теория механизмов и машин», «Системы автоматизированного проектирования», «Компьютерные системы автоматизации производства» и руководителю образовательной программы АСКОН О.Я. Чернядьевой за участие в Форсайт-сессии.

Список литературы

1. Фадель Ч., Бялик М., Триллинг Б. Четырехмерное образование: Компетенции, необходимые для успеха / Чарльз Фадель, Майя Бялик, Берни Триллинг : Пер. с англ. — М.: Издательская группа «Точка», 2018. 240 с.

2. Ван Мариенбор Й. Дж. Г., Киршнер П.А. Десять шагов комплексного обучения. Четырехкомпонентная модель дизайна обучения: М.: Альпина ПРО, 2023. 613 с.

3. Руководство по применению технологического форсайта для определения будущих потребностей в компетенциях/Судаков Д, Лукша П., Стриецка-Ильина О., Конн Г., Хофман К., Хачатрян Л.: Международное бюро труда. – Женева: МОТ, 2016 г.

4. Алпысбаев С.А., Алпысбаев К.С. Форсайт-технологии и стратегическое планирование: эффект взаимодействия // Education. Quality. Assurance. - 2023. - № 1 (30). - С. 33-40.

5. S. K. Dossayeva, I. Y. Aksarina, A. V. Kosov, G. A. Stepanova, I. Y. Akentyeva, S. N. Brovkina, A. I. Kozhedyorov, M. R. Arpentieva, R. I. Khoteeva, and K. G. Kassymova "Foresight innovations in educational systems in the BRICS countries". inBulletin of National Academy of Sciences of the Republic of Kazakhstan. 2019

6. Носкова Т. Н. Цифровизация профессиональной подготовки в высшей школе в контексте форсайта образования 2035. Отчет о НИР № 19-18-00108. Российский научный фонд. 2020.

7. Бакаева О. А., Щенников В. Н. Использование критерия x2 для выявления связи между качественными переменными на основе «Идеальных» таблиц сопряженности // Ярославский педагогический вестник. 2011. №4.

УДК 519
DOI: 10.25206/978-5-8149-3873-2-2024-1003-1009

Саморегуляция учебной деятельности: когнитивный подход
Self-regulation of learning activities: cognitive approach

А. Е. Ложников[1], В. А. Маренко[2]
[1]Омский государственный технический университет, Омск, Россия
[2]Институт математики им. С.Л. Соболева СО РАН, Омск, Россия

A. E. Lozhnikov[1], V. A. Marenko[2]
[1]Omsk State Technical University, Omsk, Russia
[2]Sobolev Institute of Mathematics Siberian Branch of the Russian Academy of Sciences, Omsk, Russia

Аннотация. Саморегуляция учебной деятельности – это набор навыков, который можно развивать, чтобы помочь учащимся управлять собой в процессе обучения. Цель работы – построение моделей, которые характеризуют саморегуляцию познавательной деятельности студентов. Для достижения цели реализуются следующие задачи: выявление факторов, влияющих на исследуемый феномен; построение и анализ графических и когнитивных моделей проблемы; формулирование выводов. Формирование моделей осуществлено с применением ресурсного подхода, в рамках которого саморегуляция рассматривается как управляющий метаресурс, включающий универсальные и специальные регуляторные компетенции. В результате работы установлено, что факторы «мотивация» учебной деятельности и «гибкость» в действиях являются определяющими для процесса саморегуляции учебной деятельности. Выявлены дополнительные управляющие факторы: «практическая реализация намерений», «инициативность» студента и «личность педагога». Новизна исследования состоит в том, что большинство специалистов изучают процессы саморегуляции с применением статистики. Нами использован когнитивный подход, который позволяет получить дополнительную информацию об управляющих факторах и взаимосвязях между ними.

***Ключевые слова*:** когнитивная модель, симплициальный анализ, имитационное моделирование, образование, саморегуляция, мотивация.

Аннотация. Self-regulation of learning is a set of skills that can be developed to help students manage themselves during the learning process. The purpose of the work is to build models that characterize the self-regulation of a student's cognitive activity. To achieve the goal, the following tasks are implemented: identifying factors influencing the phenomenon under study; construction and analysis of graphical and cognitive models of the problem, formulation of conclusions. As a result of the work, it was established that the factors "motivation" of educational activities and "flexibility" in actions are decisive for the process of self-regulation of educational activities. Additional control factors were identified: "practical implementation of intentions", "initiative" of the student and "personality of the teacher". The listed factors are recommended to be used by administrative structures to make management decisions in the field of education. The novelty of the study lies in the fact that most specialists study the processes of self-regulation using statistics. We used a cognitive approach, which allows us to obtain additional information about control factors and the relationships between them.

***Ключевые слова*:** cognitive model, simplicial analysis, simulation experiment, education, self-regulation, motivation.

I. Введение

В современном мире, где электронные образовательные технологии и цифровые платформы становятся все более распространенными, самостоятельное обучение становится неотъемлемой частью процесса образования [1, 2]. Это требует от учащихся высокой степени самодисциплины и саморегуляции, чтобы успешно усваивать новые знания и навыки. Онлайн формат обучения предполагает самостоятельное упорядочение учебной деятельности. Такая организация учебы относится к регулятивным функциям индивида. Уникальность каждого ученика и его потребности в знаниях делают вопрос о самоорганизации учебной деятельности особенно актуальным [3].

Саморегуляция играет ключевую роль в управлении поведением людей для достижения целей, так как включает множество когнитивных, эмоциональных, мотивационных и поведенческих аспектов [4, 5]. Этот процесс способен адаптировать человека к различным формам стремления к целям, зависящим от индивидуальных особенностей и окружающих условий.

Исследования в области процессов саморегуляции представлены такими признанными отечественными учеными, как В.И. Моросанова, О.Л. Конопкин, А.Г. Дикая, А.К. Осницкий и другими исследователями, рассматривающими саморегуляцию как комплексную характеристику, способствующую выдвижению цели

деятельности, следованию плану по ее достижению и варьированию средств деятельности [6, 7]. Чем выше у обучающихся уровень развития саморегуляции, тем лучше их образовательные возможности. Уровень саморегуляции влияет на устойчивость к стрессовым факторам и адаптацию к новой образовательной среде [8]. В российской педагогической практике принято подразделять саморегуляцию на основании того, что выступает в качестве объекта: состояние, поведение или деятельность человека. Зарубежные исследования связаны с именами таких ученых, как: R. Baumeister, M. Gailliot, M. Oaten, P. Karoly, D. Carver и других [9, 10, 11]. В работах известных зарубежных авторов саморегуляция определяется как качество личности, обращение к которому происходит при возникновении трудностей по достижению цели, как правило, на начальных этапах деятельности, и имеет ситуационный характер.

II. Постановка задачи

Цель данного исследования заключается в разработке моделей, описывающих факторы саморегуляции процесса обучения. Для достижения цели определены и реализованы следующие задачи: обзор научной литературы для выявления ключевых факторов, влияющих на саморегуляцию; создание и анализ моделей, включая когнитивные; выводы по результатам исследования. Формирование моделей осуществлено с применением ресурсного подхода. Это означает, что саморегуляция, охватывая как специфические регулирующие компетенции, так и универсальные, воспринимается как управляемый метаресурс.

Методология исследования включает сбор данных через анкетирование студентов, используя инструменты доктора педагогических наук Моросановой В.И. и кандидата психологических наук Галузо П.Р., что позволило оценить различные аспекты саморегуляции, такие как моделирование, планирование, оценка результатов и другие аспекты. Опросники могут работать как единые шкалы: «Общий уровень саморегуляции поведения» и «Общий уровень саморегуляции учебной деятельности», которые позволяют выявлять универсальные и специальные регуляторные компетенции, влияющие на эффективность индивидуального учебного процесса.

III. Построение моделей саморегуляции

18. Диаграммы

Побудителем деятельности, в том числе учебной, является установка, готовность индивида на основе мотивации к соответствующему виду активности. Термин «саморегуляция», включающий совокупность психологических ресурсов человека, используется как средство решения задач жизнедеятельности и реализуется, в том числе в поведении человека.

В рамках реализации поставленных нами задач проведено экспериментальное исследование по опросникам Моросановой В.И. и Галузо П.Р. [12, 13]. Первичные данные получены с применением анкетирования 40 студентов IT-специальности в возрасте от 19 до 22 лет в годах Омск и Бишкек. Анкеты содержат несколько утверждений и работают как комплекс шкал. Утверждения из опросника выбирались студентом. После этого экспериментатор проводил сравнение утверждений с таблицей «ключей». При совпадении, например, со шкалой «моделирование», студенту выставлялся 1 балл. Полученные баллы суммировались и использовались в дальнейшем для построения диаграмм (рис. 1).

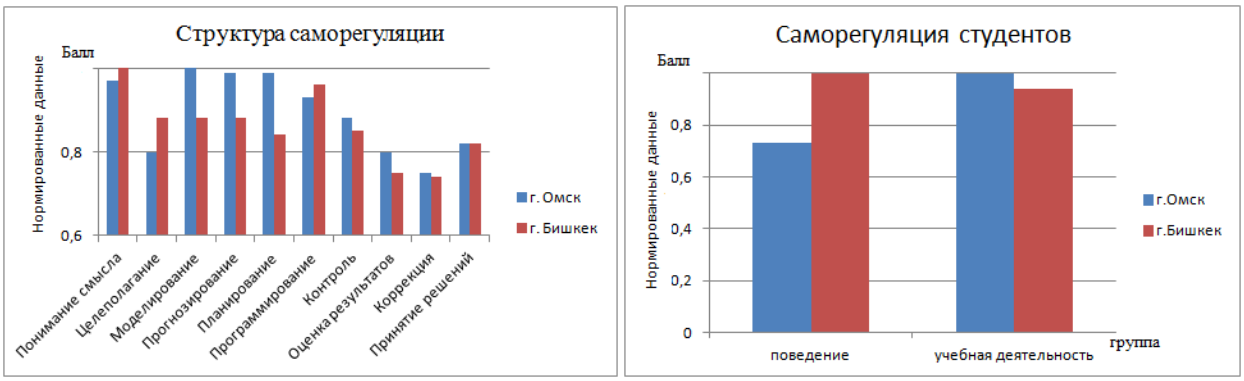


Рис. 1. Диаграммы саморегуляции учебной деятельности и поведения

На рисунке 1 слева представлены диаграммы, сформированные с применением нормированных первичных данных, которые показывают значения элементов структуры саморегуляции (понимание смысла, целеполагание, моделирование и др.). Как видно из рисунка, они разные, что указывает на динамику значений элементов. На рисунке 1 справа даны диаграммы уровней саморегуляции поведения и учебной деятельности студентов. Корреляционный анализ показал, что для показателей «поведение» и «учебная деятельность» коэффициент корреляции составил 0,48 в группе IT-студентов г. Бишкек и 0,39 в группе IT-студентов г. Омска. Наличие слабой

корреляции можно объяснить тем, что в опросе участвовали студенты 1-го и 2-го курсов, которые только начали «погружаться» в профессиональную деятельность. Из рисунка видно, что уровень саморегуляции поведения у студентов г. Бишкек больше, чем у студентов г. Омска. А уровень саморегуляции учебной деятельности у студентов г. Омска больше, чем у студентов г. Бишкек. Несовпадения показателей обусловлены различными факторами, в том числе менталитетом, традициями, особенностями образовательной системы, условиями среды жизнедеятельности и т.д.

19. Обобщенная когнитивная карта «Саморегуляция» учебной деятельности

Обзор научных публикаций показал, что для изучения саморегуляции учебной деятельности многие исследователи используют концептуальные модели и статистические средства. Наше предложение заключается в применении когнитивного подхода, что позволит включить в модели субъективный экспертный взгляд на проблему для углубления ее понимания. Мы рассматриваем саморегуляцию учебной активности студентов как систему, состоящую из различных подсистем (таких как функциональные характеристики и личные качества студентов и др.). Для группировки и организации элементов (например, оценка результатов, инициативность, ответственность и др.) в подсистемы необходимо построение иерархии. Далее подсистемы и их элементы позволяют создание когнитивных карт различной степени детализации.

Состояние системы «Саморегуляция» учебной деятельности обучающегося можно проследить с применением имитационного моделирования, которое является частным случаем математического моделирования и позволяет наблюдать изменение состояния системы в искусственных условиях. Обобщенная когнитивная карта формируется в виде орграфа, вершинами которого являются подсистемы, а дугами – взаимосвязи между ними (рис. 2). Целевой фактор на обобщенной когнитивной карте обозначен цифрой «один». К управляющим относятся факторы с «двойки» по «четверку». Взаимосвязи между факторами обозначены стрелками, направление и величины которых определяются с применением экспертных методов оценки. Причинно-следственные отношения между факторами формируются в ходе экспертных рассуждений. Например, «если динамические свойства обучающегося низкие, то саморегуляция учебной деятельности у него средняя»: «если саморегуляция учебной деятельности высокая, то личностно-стилевые качества значительные» и т.д. В приведенных примерах факторы имеют прямую взаимосвязь, с положительной экспертной оценкой в интервале от -1 до 1.

Рис 2. Обобщенная когнитивная карта «Саморегуляция» учебной деятельности

При имитационном моделировании используется авторское программное средство, сформированное на основе теории английского профессора Робертса Ф.С. [14]. При внесении импульса в вершины орграфа можно обнаружить изменения в системе. Вначале вносится импульс по 1 усл.ед. во все подсистемы одномоментно. Целевой фактор, обозначенный на рисунках пунктирной линией, в данном случае расположен на уровне 1,2 усл.ед. (рис. 3). Имитации поочередного улучшения значений подсистем в 5 раз приводит к улучшению показателей целевого фактора. При этом улучшение значений подсистем «функциональные» и «динамические» свойства, дает меньший результат, чем улучшение значений подсистемы «личностно-стилевые качества» студента. Значение целевого фактора в таком случае регистрируются на уровне 3,9 усл.ед. (рис. 4). То есть фактор «личностно-стилевые качества» студента является наиболее значимым среди рассматриваемых подсистем.

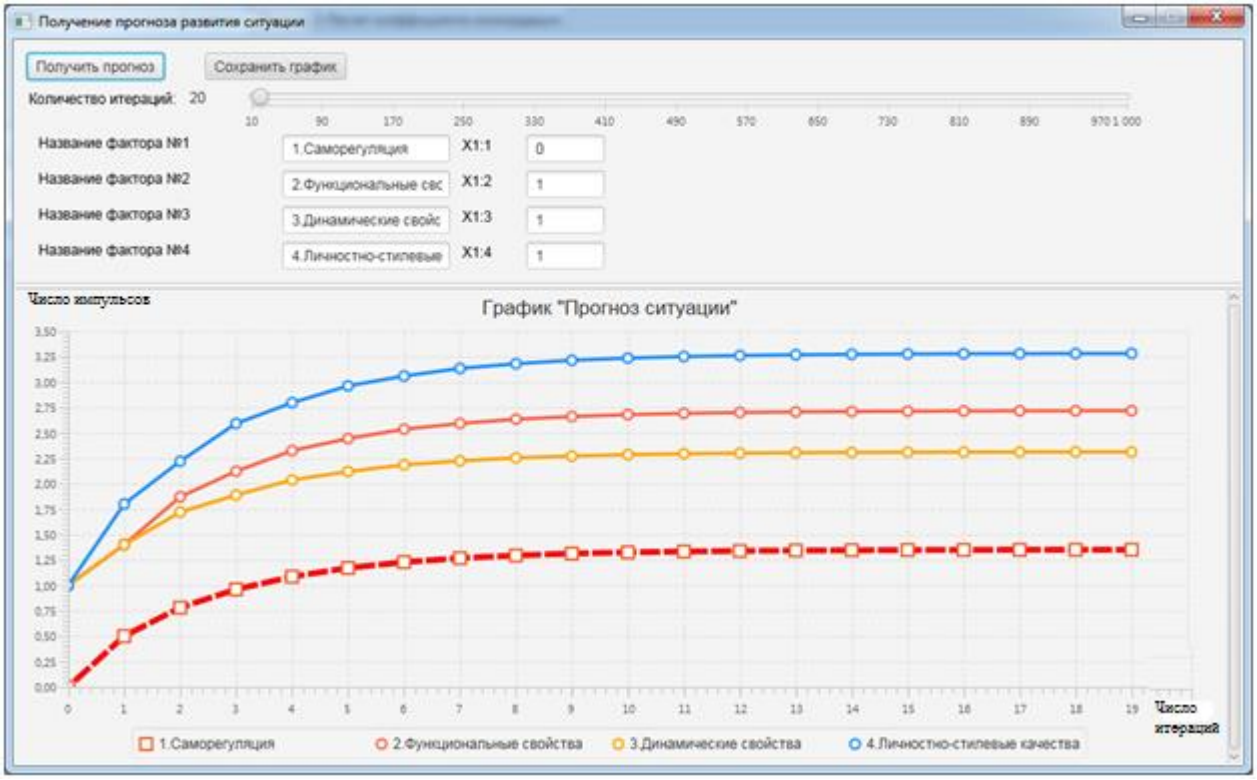


Рис. 3. Результаты первого имитационного эксперимента

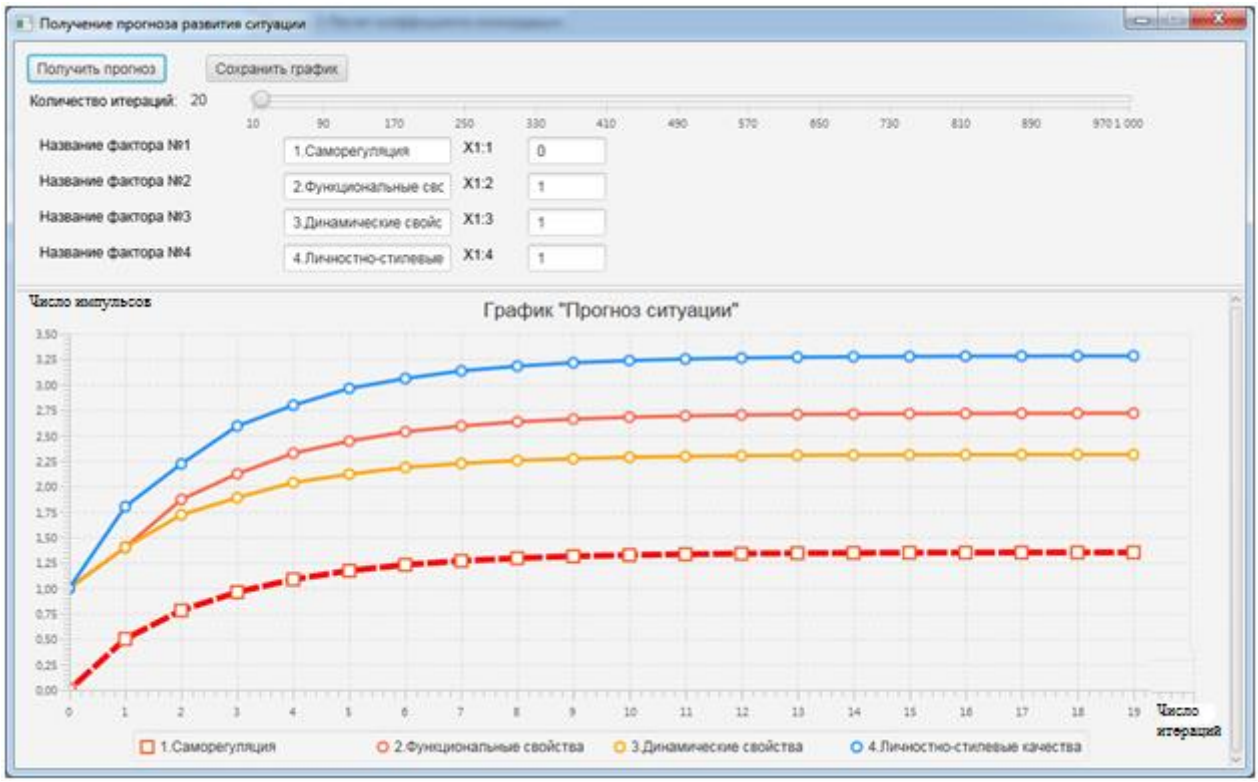


Рис. 4. Результаты второго имитационного эксперимента

Далее проведен симплициальный анализ для установления функционально значимых связей между факторами обобщенной когнитивной модели.

Результаты *q*-связности для комплекса $K_X(Y, R)$:

$q=2\ Q_2=1\ \{x_4\}$

$q=1\ Q_1=3\ \{x_4\}\{x_1\}\{x_2\}$

q=0 Q_0=1 {все}.

Результаты *q*-связности для комплекса $K_Y(X, R)$:

q=2 Q_2=1 {x_4 }

q=1 Q_1=3 {x_4}{ x_1}{ x_3}

q=0 Q_0=1 {все}.

Симплициальный анализ показал, что вершина орграфа с симплексом наибольшей размерности – это фактор x_4 «личностно-стилевые качества» студента. По нашему мнению его можно отнести к универсальной регуляторной компетенции. Таким образом, симплициальный анализ подтвердил результаты имитационного моделирования.

20. Детализированная когнитивная карта «Саморегуляция» учебной деятельности

Детализированная когнитивная карта «Саморегуляция» учебной деятельности студента построена с использованием факторов, входящих в состав опросника Осницкого А.К. [15]. Построение проводится также в виде орграфа. При этом компоненты опросника являются вершинами, а дуги – взаимосвязями между ними и представлены в формате причинно-следственных отношений (рис. 5). В качестве примеров рассуждений экспертов используются следующие выражения: «если фактор ответственности сильный, то вовлеченность полезных привычек в регуляцию действий средняя»; «если фактор практической реализации намерений небольшой, то фактор упорядочения деятельности невысокий»; «если уровень образовательных возможностей высокий, то адекватность регуляции действий лучше» и т.д.

Такой компонент когнитивной карты, как «мотивация», является стержневым для активности человека, обеспечивая ее качественность и результативность. Особую роль в адаптации учащегося к внешним условиям играет такой компонент когнитивной карты, как «гибкость в действиях» [16].

Рис. 5. Детализированная когнитивная карта «саморегуляция» учебной деятельности

Такой компонент системы, как «мотивация» учебной деятельности, можно представить в виде когнитивной карты, состоящей из 33-х элементов и связей между ними. Новая когнитивная карта «саморегуляция» учебной деятельности включает детализированную когнитивную карту «мотивация» учебной деятельности. Таким образом, получается новая расширенная когнитивная карта «саморегуляция» учебной деятельности, содержащая 51 элемент.

С новой когнитивной картой проводится симплициальный анализ с целью выявления дополнительных значимых факторов. Девятая, одиннадцатая, двенадцатая, шестнадцатая и сорок девятая строки комплекса $K_Y(X, R)$ содержат наибольшее число элементов – три. Наибольшая связность комплекса *q*=2.

Результаты вычислительной процедуры для комплекса $K_Y(X, R)$:

$q=2$ $Q_2=5$ $\{x_9\}$ $\{x_{11}\}$ $\{x_{12}\}$ $\{x_{16}\}$ $\{x_{49}\}$

$q=1$ $Q_1=12$ $\{x_9\}$ $\{x_{11}\}$ $\{x_{12}\}$ $\{x_{16}\}\{x_{49}\}$ $\{x_1\}$ $\{x_2\}$ $\{x_7\}\{x_{13}\}\{x_{15}\}\{x_{19}\}\{x_{27}\}$

$q=0$ $Q_0=1$ {все}.

Связных компонентов нет. Симплексы с наибольшим числом элементов – это факторы x_9 «мотивация» учебной деятельности, x_{11} – «практическая реализация намерений», x_{12} – «инициативность», x_{16} – «гибкость в действиях» и x_{49} – «личность педагога».

Девятый столбец комплекса $K_X(Y, R)$ содержат наибольшее число элементов – шестнадцать. Наибольший уровень связности комплекса равен пятнадцати.

Результаты вычислительной процедуры для комплекса $K_X(Y, R)$:

$q=15$ $Q_{15}=1$ $\{x_9\}$

…

$q=3$ $Q_3=3$ $\{x_9\}\{x_1\}\{x_{38}\}$

$q=2$ $Q_2=6$ $\{x_9\}\{x_1\}\{x_{38}\}\{x_{11}\}\{x_{16}\}\{x_{42}\}$

$q=1$ $Q_1=11$ $\{x_9\}\{x_1\}\{x_{38}\}\{x_{11}\}\{x_{16}\}$ $\{x_{42}\}\{x_{23}\}\{x_{24}\}\{x_{25}\}$ $\{x_{30}\}\{x_{32}\}$

$q=0$ $Q_0=1$ {все}.

Связных компонентов так же нет. Симплекс с наибольшим числом элементов – это фактор x_9 – «мотивация» учебной деятельности. Таким образом, симплициальный анализ подтвердил экспертные умозаключения о большой значимости фактора «мотивация», который, по нашему мнению, является универсальной регуляторной компетенцией.

V. Обсуждение результатов

Увеличение числа элементов когнитивной карты «саморегуляция» учебной деятельности и связей между ними привело к подтверждению факта, что значимыми управляющими факторами когнитивной модели «Саморегуляция» являются факторы x_9 – «мотивация» учебной деятельности и x_{16} – «гибкость в действиях» студента, которые мы рассматриваем как универсальные регуляторные компетенции. Выявлены так же дополнительные управляющие факторы: x_{49} – «личность педагога», x_{11} – «практическая реализация намерений» и x_{12} – «инициативность» студента.

VI. Выводы и заключение

Центральной задачей в области когнитивных наук является создание процедур, повышающих результативность деятельности человека. Результаты исследования показали, что значимыми факторами «саморегуляции» учебной деятельности являются факторы «личность педагога», «мотивация» учебной деятельности, «гибкость» в действиях, «практическая реализация намерений» и «инициативность» студента. Их рекомендуется использовать педагогическим административным структурам для управления образовательной системой в целом.

Новизна исследования состоит в том, что большинство специалистов изучают процессы саморегуляции с применением статистики. Нами использовано когнитивное моделирование, которое позволяет улучшить анализ проблемы в целом и получить дополнительную информацию о взаимосвязях компонентов ее структуры.

Результаты исследования могут быть использованы в рамках психолого-коррекционной работы с учащимися для стабилизации их психоэмоционального состояния.

Источник финансирования.

Работа выполнена в рамках государственного задания ИМ СО РАН, проект FWNF-2022-0016.

Список литературы

1. Егоров С.Ю., Шилко Р.С., Ковалев А.И., Зинченко Ю.П. Перспектива развития цифрового образования: системно-деятельностный подход // Фундаментальные проблемы в исследованиях психического здоровья человека и общества. 2019. № 4 (104). С. 120-127.

2. Шилко Р.С., Егоров С.Ю., Зинченко Ю.П. Проблемы саморегуляции учебной деятельности в массовых открытых онлайн курсах // Герценовские чтения: психологические исследования в образовании. 2020. Вып. 3. С. 632-639.

3. Саморегуляция и прогнозирование социального поведения личности: Диспозиционная концепция / Под ред. В. А. Ядова. 2-е изд. – М.: ЦСПиМ, 2013. 376 с..

4. Ложников А.Е., Маренко В.А. Моделирование форм когнитивной активности индивида на примере системы «учебная мотивация» студента / GraphiCon 2023: труды 33-й Междунар. конф. по компьютерной графике и машинному зрению (Москва, 19–21 сент. 2023 г.). – М. : Институт прикладной математики им. М.В. Келдыша РАН, 2023. С. 771-779.

5. Маренко В.А., Лучко О.Н., Рыспаева Ч.К. Моделирование структуры мотивационного комплекса студента на примере педагогической деятельности // Наука о человеке: гуманитарные исследования. Т.17. № 3. 2023. С.143-150. DOI: 10.57015/issn1998-5320.2023.17.3.15.

6. Моросанова В.И. Осознанная саморегуляция как метаресурс достижения целей и разрешения проблем жизнедеятельности // Вестник Московского Университета. Серия 14. Психология. 2021. № 1. С.4–37.

7. Моросанова В.И. Психология осознанной саморегуляции: от истоков к современным исследованиям // Теоретическая и экспериментальная психология 2022. Т. 15. №. 3.С. 57–82.

8. Альмешкина А.А., Абрамов В.А. Саморегуляция учебной деятельности студентов в зависимости от их отношения к дистанционному образованию // Журнал психиатрии и медицинской психологии. 2021. № 2 (54). С. 70-79.

9. Baumeister R.F., Gailliot M., DeWall C.N., Oaten M.. Self-regulation and personality: How interventions increase regulatory success, and how depletion moderates the effects of traits on behavior. J. of Personality, 2006. 74 (6), pp. 1773-1802.

10. Hoyle R.H., Davisson E.K. Self-regulation and personality. In O.P. John, R.W. Robins (Eds.). Handbook of personality: Theory and research. The Guilford Press, 2021, (4th ed., pp. 608–624).

11. Hoyle R.H., Gallagher P. The interplay of personality and self-regulation. In M. Mikulincer, P.R. Shaver, M.L. Cooper, & R.J. Larsen (Eds.). Handbook of personality and social psychology. Personality processes and individual differences, 2015, Vol. 4, pp. 189–207.

12. Моросанова В.И. Опросник «Стиль саморегуляции поведения – ССП-98» https://kbmk.org/uploads/kolledj/vosp_rabota/20170830_izuch_lich_osob.pdf. (Дата обращения 18.03.23).

13. Галузо Р.П. Опросник осознанной регуляции учебной деятельности – ОРУДС. https://onlinetestpad.com/ru/test/2693222-oprosnik-osoznannoj-regulyacii-uchebnoj-deyatelnosti-oruds. (Дата обращения 18.03.23).

14. Roberts F.S. Discrete Mathematical Models, with Applications to Social, Biological and Environmental Problems. Prentice-Hall, 1976, ISBN 978-0-13-214171-0.

15. Осницкий А.К. Деятельность, сознание, саморегуляция / В сб. Психология человека как субъекта познания, общения и деятельности. Отв.ред. В.В. Знаков, А.Л. Журавлев. Москва, 2018. С. 1996-2005.

16. Неверов В.Н. Особенности изменения структуры саморегуляции в студенческом возрасте // Экономическая психология: прошлое, настоящее, будущее. 2016. № 3-1. С. 223-233.

УДК 378:147
DOI: 10.25206/978-5-8149-3873-2-2024-1010-1015

Обучение будущих инженеров-строителей в пиринговогом взаимодействии в ходе выполнения расчетно-графических работ
Training of future civil engineers in peer-to-peer interaction during the execution of calculation and graphic works

М.В. Лагунова, В.А. Тюрина, А.В. Назаровская, М.Н. Пятницына
Нижегородский государственный архитектурно-строительный университет, Нижний Новгород, Россия
M.V. Lagunova, V.A. Tyurina, A.V. Nazarovskaya, M.N. Pyatnitsyna
Nizhny Novgorod State University of Architecture and Civil Engineering, Nizhny Novgorod, Russia

Аннотация. Статья посвящена выявлению потенциала коллаборации как актуального способа формирования гибких компетенций будущего специалиста строительного профиля в области технологий информационного моделирования. Предложена инновационная коммуникативная практика обучения – пиринговое взаимодействие. Охарактеризована геометро-графическая подготовка будущих инженеров-строителей в ФГБОУ ВО ННГАСУ, демонстрирующая пиринговое взаимодействие в ходе выполнения расчетно-графической работы в рамках освоения дисциплины «Начертательная геометрия и компьютерная графика». Предложена схема формата пиринга, где в качестве педагогического сопровождения используется в том числе система дистанционного обучения lms.nngasu.ru. Показано, что форма обучения носит характер совместного, согласованного и конструктивного сотрудничества, опосредованного межличностными отношениями. Описаны результаты включенного наблюдения за работой команд, что позволило выявить и обобщить особенности пирингового взаимодействия. Проанализировано мнение студентов о степени готовности работать в подобных командах. Качественные результаты показывают, что геометро-графическая подготовка, осуществляемая в сотрудничестве по схеме субъект-субъектного взаимодействия с использованием lms.nngasu.ru, имеет результативные преимущества перед индивидуальной, может рекомендоваться как эффективный путь нахождения технических, интеллектуальных и психологических ресурсов для подготовки к решению профессиональных задач в новой информационной реальности.

***Ключевые слова*:** гибкие компетенции, ТИМ-специалисты, коллаборация, пиринговое взаимодействие, геометро-графическая подготовка.

Abstract. The article is devoted to identifying the potential of collaboration as a relevant way of developing flexible competencies of a future construction specialist in the field of information modeling technologies. An innovative communicative teaching practice is proposed - peer-to-peer interaction. The geometric and graphic training of future civil engineers at the Federal State Budgetary Educational Institution of Higher Education NNGASU is characterized, demonstrating peer-to-peer interaction in the course of performing computational and graphic work within the framework of mastering the discipline “Descriptive Geometry and Computer Graphics”. A peering format scheme is proposed, where the distance learning system lms.nngasu.ru is used as pedagogical support. It is shown that the form of training is in the nature of joint, coordinated and constructive cooperation, mediated by interpersonal relationships. The results of participant observation of the work of teams are described, which made it possible to identify and generalize the features of peer-to-peer interaction. The students' opinions on the degree of readiness to work in such teams were analyzed. Qualitative results show that geometric-graphic training, carried out in collaboration according to the scheme of subject-subject interaction using lms.nngasu.ru, has effective advantages over individual training and can be recommended as an effective way to find technical, intellectual and psychological resources to prepare for solving professional problems. tasks in the new information reality.

Keywords: flexible competencies, BIM-specialists, collaboration, peer-to-peer interaction, geometric and graphic training.

I. Введение

Глобальные задачи развития, провозглашенные в «Стратегии развития строительной отрасли и жилищно-коммунального хозяйства РФ на период до 2030 г. с прогнозом до 2035 г.» [1] требуют комплексных изменений, нацеленных на цифровую трансформацию, обеспечение технологического суверенитета, развитие инновационной деятельности и трансформации подготовки специалистов, ориентированных на решение профессиональных задач в связи с новыми технологиями информационного моделирования (ТИМ).

Вузовская практика демонстрирует, как правило, направленность на обучение студентов в рамках своего профиля с приоритетом на освоение специализированных программных продуктов, дополняется практико-

ориентированным обучением с участием работодателей [2]. Авторами апробирована концепция организации учебного процесса в ФГБОУ ВО «Нижегородский государственный архитектурно-строительный университет» для сквозной подготовки строителей технологиям информационного моделирования, реализующая комплексно-дифференцированный подход в обучении с учетом трудовых функций профильных профессиональных стандартов, ориентированная на междисциплинарность [3]. Значимым является возможность студентов «достроить» свой уникальный профиль компетенций, получив дополнительную квалификацию по информационному моделированию в строительстве на программах «Цифровой кафедры».

В то же время эксперты утверждают, что при развитии системы обучения для всех категорий ТИМ-специалистов «речь должна идти о развитии личности участника процесса информационного моделирования вместо наработки ремесленных навыков» [4], актуальны коммуникативная компетентность, психологическая гибкость, ориентированные на инновационную деятельность в области ТИМ [5].

2. Постановка задачи

Несмотря на широкий интерес к формированию актуальных гибких компетенций в целом, в меньшей степени исследованы методические аспекты геометро-графической подготовки, ориентированные на развитие коммуникативной компетентности, что делает актуальным рассмотрение потенциала коллаборации, партнерства как актуального и эффективного способа формирования гибких компетенций будущего ТИМ-специалиста.

3. Теоретическая база формирования гибких навыков в условиях коллаборации

Термины «гибкие», «не когнитивные», «мягкие» навыки, пришедшие из экономических исследований, известны в мировом опыте как «Навыки 21-го века» [6], включают совокупность навыков, необходимых для успешной адаптации в современном профессиональном пространстве. Наиболее известен операциональный подход, выделяющий группы актуальных характеристик: компетенции (критическое мышление, креативность, коммуникация, работа в команде) и личные качества (инициативность, адаптивность и др.) [7].

Одним из эффективных способов формирования означенных компетенций может стать обучение в коллаборации, сотрудничестве. Понятие «коллаборация» (от фр. *collaboration* – сотрудничество) позволяет выделить следующие составляющие: направленность деятельности (ориентирует на равноправное сотрудничество); характер деятельности (равноправность, баланс интересов, достижение консенсуса, компромисс); способы, которыми оно осуществляется (совместные проекты) [8].

Как организационная система активности взаимодействующих субъектов сотрудничество характеризуется пространственным и временным соприсутствием. Организацию совместной деятельности необходимо строить на психологических принципах [9] (рис. 1).

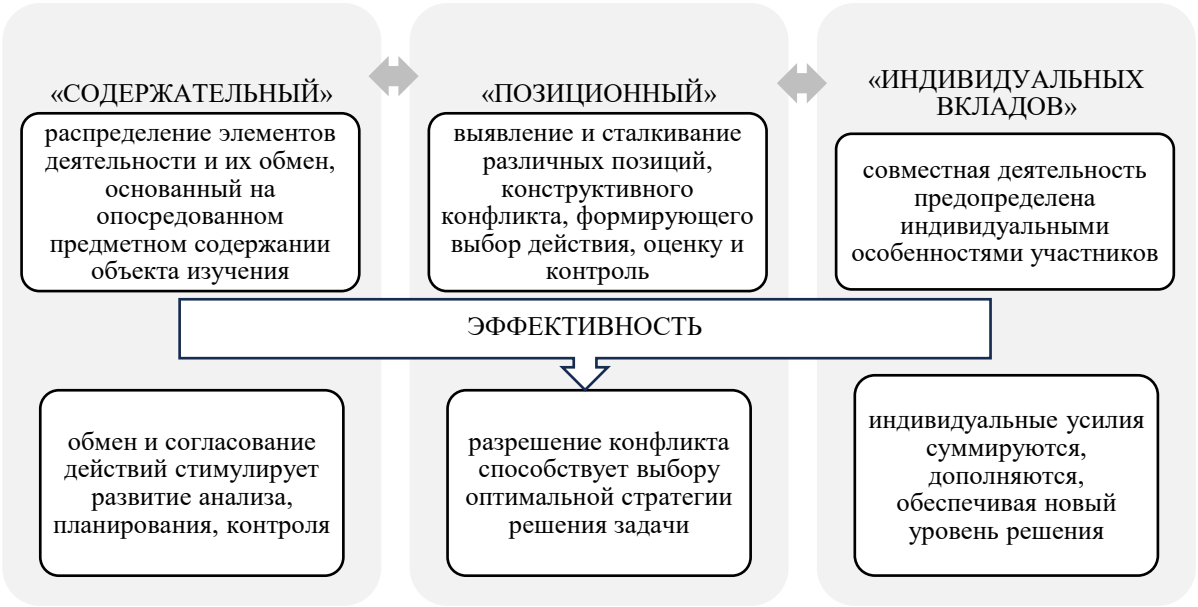


Рис. 1. Принципы организации совместной деятельности

Обучение в коллаборации последние годы является одним из мировых трендов [10], в рамках которого перспективной коммуникативной практикой выступает пиринговое взаимодействие (*peering*- равноправный информационный обмен), особенностью которого являются условия равноправия, когда участники образовательного процесса находятся на одном уровне учебной иерархии. Практиками отмечается эффективность применения пиринга в системе повышения квалификации. Пиринг подразумевает установление межличностных отношений не только в аудиторных условиях, но обеспечивает гибкость, отсутствие временных

и пространственных ограничений, опирается на электронные средства (образовательные платформы, тестирующие системы, электронные тренажеры, мультимедийные ресурсы и т.д.) [11], что актуально и прогнозирует ее эффективность при подготовке бакалавров.

4. Результаты экспериментального обучения

На этапе развивающего этапа эксперимента осуществлялось исследование образовательной ситуации обучения будущих инженеров-строителей в пиринговом взаимодействии в ходе выполнения РГР при изучении дисциплины «Начертательная геометрия и компьютерная графика». Эксперимент проводился в ФГБОУ ВО «Нижегородский государственный архитектурно-строительный университет» в рамках сквозной подготовки технологиям информационного моделирования [3].

Поскольку новый стандарт образования смещает фокус на развитие способностей к поиску, анализу, синтезу информации, применения системного подхода к решению поставленных задач, формированию мягких навыков, ведущими среди которых являются коммуникативные, то в рамках освоения дисциплины «Начертательная геометрия и компьютерная графика» наряду с общепрофессиональными компетенциями уметь применять методы начертательной геометрии в профессиональной деятельности; иметь навыки автоматизированного геометрического моделирования, как основы для проектирования, производства и эксплуатации с использованием современных ТИМ, ставится задача формирования способностей определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений, осуществлять социальное взаимодействие и реализовывать свою роль в команде, управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни.

С целью формирования общепрофессиональных компетенций студентам предлагается выполнение исследовательских РГР, связанных с исследованием наиболее общих понятий геометрического описания предметного мира с помощью процедуры проецирования на различных геометрических моделях, приобретения навыков анализа параметров положения точек, прямых, плоскостей, их отношений и свойств, навыков оформления чертежей в соответствии с нормативами оформления графических работ. С целью формирования универсальных компетенций задание представляет собой специфическую учебную ситуацию, требующую совместных учебных действий.

Поток делится на группы из двух человек случайным образом. Учебная ситуация отвечает требованиям: общность цели, выполнение индивидуальных действий каждым студентом, координированность и неаддитивность, т.е. не простое сложение индивидуальных действий, а получение общего согласованного результата, оцениваемого по установленным критериям.

Исследование проводилось в два этапа. Первый этап включал выполнение двух чертежей (рис.2 и рис.3) Так, исследование прямых или плоскостей, изначально заданных координатами, включает ряд локальных операций: изображение проекций геометрических образов с учетом перезадания ее различными определителями; построение следов и определение координат точек их схода; построение главных линий плоскости и их следов; решение разнообразных метрических задач. Форма презентации решений включает двумерные геометрические модели: эпюр Монжа и аксонометрию. На эпюре построения могут требоваться на двух или трёх проекциях. Работа на аксонометрическом чертеже выполняется по выбору в любом стандартном виде. Построения сопровождаются символьными пояснениями и определением численных значений найденных параметров. На втором этапе эти же операции выполняются на трехмерных геометрических моделях: выполняется 3D-модель данной плоскости в системе геометрического моделирования Компас-3D, а также (по желанию) разборный физический макет плоскости, повторяющий расчетные параметры модели.

Вторым примером является параметрический анализ плоских контуров, вначале проводимый в ручной технологии, затем выполняемый в параметрическом режиме в системе геометрического моделирования.

Задача команды разделить работы на отдельные подпроцессы и выполнить их через кросс-функциональные взаимодействия. Студенты распределяли роли, позволяющие обеспечить координацию сбора и анализа теоретического материала, решить и оформить локальные задачи с максимальной эффективностью, в том числе, обеспечив визуальное качество чертежа в соответствии с государственными стандартами, выдержав установленные временные рамки.

Роль педагогического сопровождения выполнял электронный курс в системе дистанционного обучения lms.nngasu.ru с высоким уровнем детализации учебного контента, учитывающего мозаичность восприятия информации современным студентом, включающем цель задания, перечень формируемых компетенций, разноплановые теоретические материалы, стандарты ЕСКД, рекомендуемый порядок решения, примеры, список рекомендуемых источников, контрольно-измерительные материалы по каждому блоку информации. Возможность обсуждения приемов решения задачи, общих и частных случаев могла проводится в различных формах. Кроме того, были сохранены прямые управляющие функции педагога с высоким удельным весом *on-* и *off-line* консультаций, преследовавших наряду с дидактическими целями поддержку эмоционально-волевой регуляции, помощь в планировании, предметно-содержательным распределении действий, согласовании и контроле.

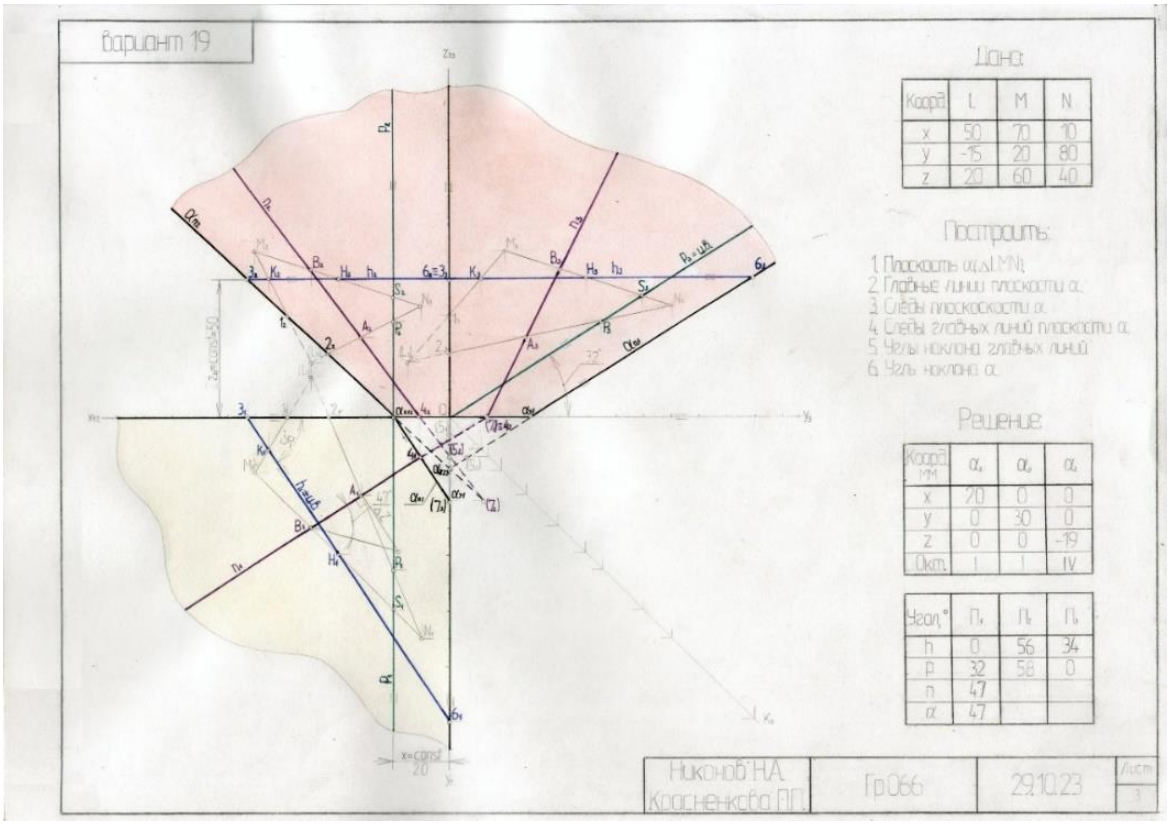


Рис. 2. Ортогональный чертеж, выполненный в пиринговом взаимодействии

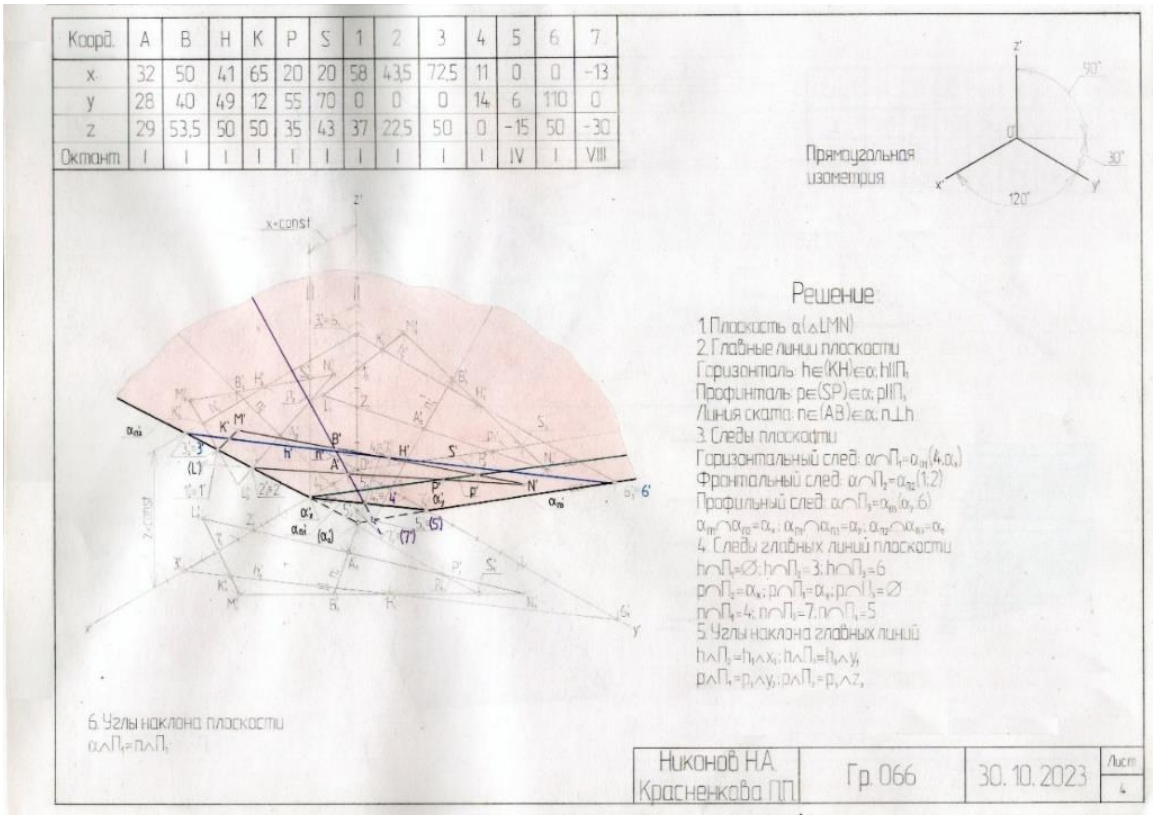


Рис. 3. Аксонометрический, выполненный в пиринговом взаимодействии

Каждый результат работы оценивался по ранее опубликованным на lms.nngasu.ru компонентным критериям, являющимся своеобразной планом выполнения задания.

Данное обучение носило выраженный субъект-субъектный характер – процесс совместного, согласованного и конструктивного сотрудничества, опосредованный межличностными отношениями. При переходе к выполнению следующей РГР проводилась смена партнера.

5. Обсуждение результатов

В ходе эксперимента нами использовался метод включенного наблюдения за работой команд, что позволило выявить и обобщить особенности пирингового взаимодействия в ходе решения геометро-графических задач.

Работа в творческих группах позволила выйти на более высокий уровень РГР, сделать их более сложными и интересными, с большим объемом прорабатываемого материала. Например, при выполнении РГР, посвященной исследованию основных геометрических образов, основным фактором эффективности явилось предметно-содержательное распределение операций в группе, применение различных графических моделей в качестве средств организации групповой работы, способствовавшей генерализации знаний о геометрических образах и общему способу решения данного класса задач участниками. Рационализация образовательного процесса проявилась в повышении эффективности механизмов обмена информацией между студентами, обусловленной многозадачностью и критериальной определенностью, в повышении темпов генерации и трансляции геометро-графических знаний, концентрации и оперативности графической деятельности студентов, возможности тиражирования успешных практик. Вышли на новый уровень приемы само- и взаимооценивания, способствующие самоорганизации в целом.

Опишем выявленные недостатки. Ряд студентов не готовы были принять условия выполнения заданий, например 3% потока отказались участвовать в пиринговом взаимодействии, настояв на индивидуальном выполнении полного объема задания или индивидуальном выполнении части задания.

Часть студентов чувствовала неуверенность и беспокойство за итоговый результат, поскольку испытывала недоверие к ответственности и качеству работы партнёра.

Выявлены (единичные) случаи отсутствия индивидуальной ответственности в группе, так называемой «социальной лености», приводящей к бездействию одного из партнеров, сотрудничество терпело неудачу. Решение этой проблемы следует искать в уточнении критериев, по которому оценивается вклад каждого в общий результат.

Около 15% выразили негативное отношение к использованию технологии при случайном выборе партнера, без учета их когнитивного уровня и личностных качеств, приводящем к непредсказуемости и вероятностном характере образовательного процесса. Однако комплектация групп рандомно была предложена нами исходя из целей эксперимента и краткосрочного формата взаимодействия. На следующем этапе подготовки студентом была предложена возможность выбора партнера.

6. Заключение

В связи с современными требованиями к уровню подготовки будущих инженеров-строителей, способных реализовывать новую концепцию использования информационной среды, существует актуальная необходимость совершенствования учебного взаимодействия субъектов учебной деятельности в ходе пирингового взаимодействия. Проведенный эксперимент показал, что геометро-графической подготовка, осуществляемая в сотрудничестве по схеме субъект-субъектного взаимодействия с использованием электронного курса, размещенного в электронной образовательной среде вуза, имеет результативные преимущества перед индивидуальной, поскольку показала значительное повышение качества обучения, выраженного повышением академической успеваемости, способствовала развитию у студентов личностных, предметных, метапредметных компетенций, дала эффективность с точки зрения временных затрат как студентов, так и преподавателей.

Список литературы

1. Распоряжение Правительства РФ от 31 октября 2022 г. № 3268-р Об утверждении Стратегии развития строительной отрасли и жилищно-коммунального хозяйства РФ на период до 2030 г. с прогнозом до 2035 г. – URL: https://www.garant.ru/products/ipo/prime/doc/405560559/(дата обращения: 25.10.2023).

2. Фомин Н. И., Миронова Н. И. Образовательный кластер инженерного творчества для обеспечения инновационной деятельности выпускников архитектурно-строительных специальностей // Педагогическое образование в России. 2023. № 2. С. 146–154.

3. Ротков С. И., Конопацкий Е. В., Лагунова М. В. Организация учебного процесса в ННГАСУ для сквозной подготовки строителей технологиям информационного моделирования // Приволжский научный журнал / ННГАСУ. Н. Новгород, 2023. № 3. С. 10-21.

4. Технология BIM: подготовка новых кадров – URL: https://isicad.ru/ru/articles.php?article_num=16712. (дата обращения: 25.10.2023).

5. Привалова М. А., Симченко И. В. Социокультурная компетенция будущих BIM-инженеров // Человек. Знак. Техника: Сборник статей I Междисциплинарного молодежного форума с международным участием. Самара: Самарский научный центр РАН, 2021. С. 289-293.

6. К обществам знания: Всемирный доклад ЮНЕСКО. Париж: ЮНЕСКО, 2005. – URL: http://www.ifap.ru/library/book042.pdf. (дата обращения: 25.10.2023).

7. Davies L. M. Creativity as a twenty-first-century competence : An exploratory study of provision and reality // Education 3-13. 2018. Vol. 46, № 7. P. 879–891.

8. Жмакина М. В. Пиринговое обучение как результат трансформации системы взаимодействий в школе // Современная наука: актуальные проблемы теории и практики. Серия: Гуманитарные науки. 2017. № 3. С. 49-53.

9. Поливанова Н., Ривина И. В. Принципы и формы организации совместной учебной деятельности // Психологическая наука и образование. 1996. № 2. С. 43-53.

10. Candy P., Crebert G., O'Leary J. Developing Lifelong Learners Through Undergraduate Education // NBEET, Commissioned Report. №28. Canberra: Australian Government Publishing Service, 1994. – URL: https://www.voced.edu.au/content/ngv%3A22704 (дата обращения: 25.10.2023).

11. Рекешева Л. Н. Модель использования технологии взаимообучения в дистанционном повышении квалификации педагогических кадров // Вестник МГПУ. Серия: Педагогика и психология. 2022. № 1(59). С. 197-211.

УДК 378 : 514.18 : 004.9
DOI: 10.25206/978-5-8149-3873-2-2024-1016-1024

Об организации автоматической проверки инженерно-геометрических заданий на образовательных интернет-платформах
On the organization of automatic verification of engineering and geometric tasks on educational Internet platforms

А. А. Бойков
РТУ МИРЭА, г. Москва, Российская Федерация

A. A. Boykov
MIREA — Russian Technological University, Moscow, Russian Federation

Аннотация. Анализируются возможности образовательных платформ, поддерживающих автоматическую проверку решений задач по программированию, на примере платформы «Stepik». Предлагается подход к организации проверки решений инженерно-геометрических задач, который может быть реализован при помощи имеющихся в настоящее время технических возможностей. Он заключается в формировании текстового представления чертежа или трехмерной модели в окне редактора CAD-системы с последующей вставкой в окне образовательной платформы и проверкой встроенной Python-подпрограммой или удаленным сервером в режиме External grader. Создан Python-сценарий, формирующий описание чертежа для одного класса задач начертательной геометрии, и показана принципиальная возможность проверки решения сравнением с эталоном в виде последовательности строк текста.

***Ключевые слова*:** автоматизированная проверка чертежей, автоматизированная проверка электронных моделей, CAD-системы в образовании, образовательные платформы, Stepik, Компас-3D

Аннотация. The capabilities of educational platforms that support automatic verification of solutions to programming problems are analyzed using the «Stepik» as an example. An approach to organizing verification of solutions to engineering and geometric problems is proposed, which can be implemented using currently available capabilities. It consists of forming a text representation of a drawing or a 3D-model in the CAD system editor window, followed by insertion into the educational platform window and verification by a built-in Python subroutine or a remote server in the External grader mode. A Python script has been created that generates a description of a drawing for one class of descriptive geometry problems, and the fundamental possibility of verifying a solution by comparing it with a standard in the form of a sequence of text lines has been demonstrated.

***Ключевые слова*:** computer aided verifying of drawings, computer aided verifying of 3D models, CAD-systems in education, educational platforms, Stepik, Kompas-3D.

I. Введение

Анализ числа публикаций, которые выдает поиск в российской наукометрической базе *eLIBRARY* за 2014–2023 гг. по названиям известных образовательных интернет-платформ, показывает (табл. 1), что популярность компьютерных образовательных систем с удаленным доступом – дистанционных – растет.

ТАБЛИЦА 1
ЧИСЛО ПУБЛИКАЦИЙ В РЕЗУЛЬТАТАХ ПОИСКА НАУКОМЕТРИЧЕСКОЙ БАЗЫ ELIBRARY ПО НАЗВАНИЮ ОБРАЗОВАТЕЛЬНЫХ ПЛАТФОРМ

Год	**2014 – 2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**Всего**
Учи.ру	2	15	29	82	154	698	955	880	655	**3470**
Stepik	2	7	65	109	183	346	503	565	567	**2347**
Skillbox			2	3	16	120	277	467	548	**1433**
Инфоурок	25	55	99	81	114	162	213	192	167	**1108**
Geekbrains	1	6	14	34	33	108	172	198	219	**785**

Уже сейчас образовательные платформы предлагают курсы, посвященные изучению *CAD*-систем, геометрических и графических редакторов, хотя проверка в них все еще реализуется или в виде тестов, или вручную преподавателем.

Проверка графических заданий при традиционной форме обучения является трудоемким процессом [1, 2, 3] и не может быть полностью заменена тестированием (это отмечается, например, в рецензии [4]), хотя такие попытки и предпринимаются [5]. При переходе к дистанционному обучению проблема проверки решений практических заданий приобретает особую остроту, поскольку трудоемкость проверки сама по себе усугубляется ([6]) издержками из-за необходимости сканирования и загрузки чертежей, или файлов, выполненных в CAD-системе, необходимости разметки ошибок, теряется оперативность проверки, повышается вероятность плагиата и т.д.

Очевидно, что появление средств автоматической проверки инженерно-геометрических заданий существенно расширило бы возможности таких платформ, и это обуславливает актуальность создания инструментов, выполняющих автоматическую проверку решений инженерно-геометрических заданий применительно к образовательным-интернет платформам.

II. Постановка задачи

Целью работы является анализ возможностей внедрения/создания инструментов автоматической проверки решений инженерно-геометрических заданий в образовательные интернет-платформы. Для этого требуется решить следующие задачи:

- Рассмотреть возможности образовательных платформ.
- Рассмотреть возможности имеющихся систем и известные подходы к проверке решений инженерно-геометрических заданий.
- Разработать способ внедрения инструментов автоматической проверки в курсы на базе образовательных платформ.

III. Теория

1. Возможности образовательных интернет-платформ

Рассмотрим на примере платформы *Stepik*, поскольку она:

a) российская;
b) популярная;
c) постоянно расширяет возможности, в частности, имеет средства автоматической проверки решений задач по программированию.

Платформа *Stepik* позволяет создавать бесплатные и платные курсы. Курсы состоят из модулей (3–7), модули, рассчитанные на комфортное прохождение в течение недели, – из уроков (5–7), уроки – из шагов (до 16). Шаг является информационным (теоретическим блоком) или задачей (практическим блоком). Система автоматически поддерживает создание информационных блоков с текстовым содержанием или видеоуроков. Задачи могут быть следующих видов [7]:

1) тесты;
2) задачи со вводом ответа (в частности, для проверки введенного текста можно дополнительно описать подпрограмму на языке *Python*);
3) сложные задачи (в частности, задачи на программирование, на написание *SQL*-запроса, на составление *HTML* и *CSS*);
4) экспериментальные задания (в частности, с внешним оцениванием – *External grader*).

Анализ возможностей платформы *Stepik* показывает, что, кроме проверки введенных результатов измерений, как показано в [8], или взаимного рецензирования, как показано в [9], принципиально имеющиеся технические возможности платформы позволяют организовать проверку решений инженерно-геометрических заданий одним из следующих способов:

A. Решение подается в виде текста программы, проверку выполняет соответствующий транслятор в составе платформы *Stepik*.
B. Решение подается в виде текста, проверку выполняет сценарий на языке Python.
C. Решение подается в виде текста или файла, проверку выполняет внешняя система проверки при помощи интерфейса *External grader*.

Рассмотрим подробнее, как осуществляется проверка практических заданий в программировании. Подробный обзор можно найти в [10]. Наиболее распространенным способом проверки является запуск проверяемой программы с ограничением по времени для заданного набора тестов. Каждый тест включает текстовое

представление исходных данных и текстовое представление эталонного результата. В процессе тестирования программа «считывает» исходные данные теста и «печатает» результат, который затем сравнивается с эталонным. Проверка основывается на гипотезе, что только правильный с точки зрения задачи код, способен генерировать требуемый результат, или что любой код, генерирующий требуемый результат, – правильный.

Это позволяет: во-первых, обеспечить исполняемость кода-решения (транслятор сообщит обо всех синтаксических ошибках), во-вторых, обучаемый получает возможность неограниченно запускать варианты своей программы до тех пор, пока не найдет правильное решение.

Недостатки подхода:

- не проверяется содержание решения, оно может быть получено иным способом или на другом языке;
- составитель задания может пропустить тесты для выявления некоторых частных ситуаций;
- проверка в настоящее время ориентирована на ввод-вывод текстовой информации;
- проверка используется в отношении сравнительно простых программ, реализующих одну какую-либо точно сформулированную функцию; проверка программных комплексов, библиотек подпрограмм с большим числом функций, графическим интерфейсом пользователя и др. не выполняется.

2. Анализ имеющихся подходов и систем проверки решений инженерно-геометрических заданий

В [11, 12] был дан подробный обзор имеющихся подходов и систем проверки решений геометрических, графических и инженерно-графических задач. Инженерно-геометрические задачи, связанные с созданием чертежей и моделей в *CAD*-системах или геометрических редакторах, естественным образом обобщают все вышеперечисленные частные виды задач. Наиболее проработан т.н. метод проверки наложением (совмещением), состоит в распознавании в решении совокупности геометрических объектов, соответствующих эталонной грамматике [13], показано его применение к чертежам задач по начертательной геометрии, в том числе содержащим изображения кривых поверхностей, к одному классу чертежей деталей [14] и др.

В [15] были предложены понятия ***верификации*** (организация самостоятельной проверки решений задач при поддержке компьютерной системы) и ***верифицируемости*** (практическая доступность их верификации). Выделены три уровня верификации инженерно-графических заданий:

- Верификация упражнений (элементарных навыков).
- Верификация задач (многошаговых заданий).
- Верификация оформления решений.

Автоматическая проверка задач в программировании при помощи контрольных тестов соответствует в этом случае верификации упражнений и, частично, верификации (простых) задач. Верификация оформления в программировании осуществляется, до некоторой степени, автоматически в ходе синтаксической проверки кода транслятором.

По аналогии с задачами программирования, имеющиеся технологии в составе образовательных платформ можно использовать для организации верификации упражнений и некоторых задач с некоторыми элементами оформления. Проверка решений в виде программного кода (А) может применяться в том случае, если решение инженерно-геометрической задачи представлено в виде некоторой программы, однако, трансляторов подобных языков в составе системы *Stepik* и др. в настоящее время нет.

В [16] проводился методологический анализ способов организации проверки решений геометро-графических задач. Приведем некоторые из них:

2.1. Решение формируется в геометрическом редакторе (например, *CAD*-системе):

2.1.1. Геометрический редактор содержит функции для передачи решения на проверку удаленной системе.

2.1.2. Удаленная система проверяет файлы, созданные в геометрическом редакторе.

Использование их показано в [12] для 2.1.1 и в [17] для 2.1.2.

Рассмотрим подробнее взаимодействие *CAD*-системы и обучающей платформы в этом случае.

3. Организация автоматической проверки решений инженерно-геометрических заданий на базе Stepik

Чтобы в составе курсов на платформе *Stepik* автоматически проверить решение инженерно-геометрического задания (упражнения, простой задачи и т.д.), необходимо:

1. Передать данные решения (электронный чертеж или 3*D*-модель) в *Stepik*.
2. Организовать проверку встроенными средствами (В) или при помощи внешней системы (С).

Подзадача 1. Передача данных для проверки

Платформа *Stepik* ориентирована на ввод и проверку решений задач в виде текста, поэтому на начальных этапах можно рассмотреть следующий способ взаимодействия:

1. Задание решается в *CAD*-системе, где имеется инструмент отправки решения на проверку (см. [17]).
2. Инструмент отправки, в отличие от [17], преобразует данные решения (элементы чертежа или модели) в текстовое представление, которое помещается автоматически, например, в буфер обмена.
3. Учащийся переходит в окно *Stepik*, где вставляет подготовленное решение в поле ввода текста.
4. Проверка решения может затем проверяться как *Python*-сценарием (В), так и внешней системой (С).

Такой вариант не противоречит традиционным приемам работы в *Stepik*. Например, при решении сложных задач по программированию часто удобно перейти в какую-либо среду разработки, написать и отладить код, скопировать и вставить в *Stepik* для проверки.

В дальнейшем при поддержке со стороны разработчиков Stepik возможна более полная интеграция *CAD*-системы, ядра *Stepik*, подсистем проверки и т.п.

Подзадача 2. Автоматическая проверка геометрических данных решения

В большинстве случаев проверка решения инженерно-геометрической задачи – процесс трудоемкий, поскольку грамматический разбор с последующей проверкой и отсечением ошибочных вариантов интерпретации решения сопровождается процедурами перебора как объектов эталона, так и объектов решения [11–13]. Ранее также был показан ряд приемов (нормализация и предварительная обработка (свертка) решения) для ускорения проверки, выделены четыре типа задач с типовой структурой грамматики правильного решения – инвариантные задачи, задачи с альтернативными решениями, задачи со свободным размещением и комбинированные [12].

С учетом особенной проверяющей среды *Stepik* рассмотрим прием, который позволяет выполнить проверку решений перечисленных типов используя возможности символьного сравнения.

Проверяемое решение, в соответствии с [17], представляет совокупность строк вида:

ключ = вид фигуры (имя параметра1=значение1, имя параметра2=значение2, …)

Пример описания фигур из [17]:

T4294967331 = Text(x=25.19512007419307, y=267.23920 236796960, width=101.70019531250000, height=35.67906951904297, angle=0, mirror=0, content="ASCODE#001\r\nЗадание. По данным размерам построить\r\nчертеж треугольника...")

S4294967297 = Segment(ax=151.76421064830959, ay=115. 24344890420329, bx=129.76421064830959, by=85.243448904 20335, style=solid)

Перед преобразованием в текст выполним сортировку объектов решения и их параметров в некотором порядке, например, вначале геометрические объекты (точки, отрезки, окружности и т.д.), затем аннотирующие (тексты, линии-выноски, линейные размеры, диаметры и т. д.), объекты одного вида – в порядке расположения (слева направо, сверху вниз) и т.д. Такая сортировка упрощает привязку обозначений к объектам, выделение компонент и пр. [12, 13].

1. В результате такой сортировки два правильных решения задачи типа (1) (инвариантной задачи) должны различаться лишь ключами, символическими именами объектов и несущественными параметрами.
2. В отношении задач типа (2) (задача со свободным размещением) выполним дополнительную операцию: после сортировки объектов возьмем координаты первой геометрической фигуры (точки или отрезка) и выполним сдвиг всех объектов. В результате задача типа (2), если она решена правильно, становится задачей типа (1).
3. В отношении задач типа (3) (комбинированная задача) объекты добавляются в текст решения покомпонентно (сами компоненты – виды – также вводятся в виде сущностей), компоненты сортируются – главный вид, проекционные виды (сверху, слева, снизу, справа сзади), дополнительные виды и т.д.
4. Задачи типа (4) (с альтернативными эталонами), как и в [12], проверяются последовательным сравнением с эталонами принципиально разных решений.

Этот подход не исключает и этапа выделения из чертежа пространственных фигур, когда более сложные задачи оказывается возможным свести к типам (1)–(4) [12, 13].

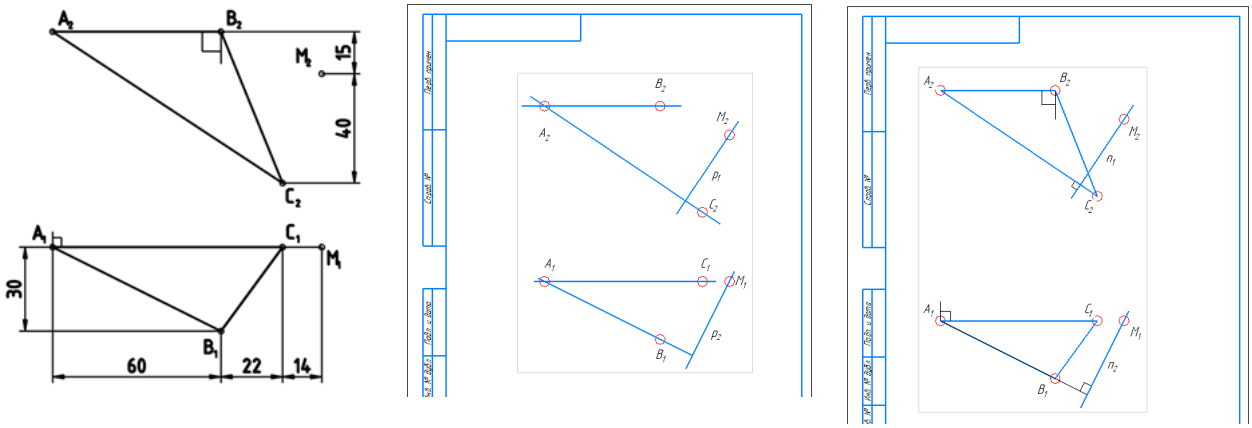
IV. Результаты экспериментов

Был создан сценарий на языке *Python*, который запускается в редакторе «Компас-3*D*» в режиме чертежа или эскиза и формирует текстовое описание, подобно тому, как ранее было показано в [17].

В качестве примера рассмотрим проверку решения задачи из [12]: по заданным размерам требуется построить чертеж плоскости *ABC*, точки *M* и прямой, перпендикулярной плоскости *ABC* через точку *M*.

На рис. 1,а показан чертеж к заданию, на рис. 1,б – эталон (содержит только обязательные объекты), на рис. 1,в – вариант решения. В таблице 2 показано текстовое представление эталонного чертежа до описанной в предыдущем разделе обработки.

В таблице 3 показано текстовое представление эталонного чертежа и проверяемого решения после обработки (для одной компоненты – горизонтальной проекции). Из таблицы видно, что именованные в тексте задания объекты (точки *A*, *B*, *C* и *M*) различаются только в пределах точности числовых значений (строки 1–3), не именованные – также могут различаться именами, но не индексами (строка 5), прямые – могут различаться значениями координат (*ax*, *ay*, *bx*, *by*), но совпадают в пределах точности числовых значений коэффициентов *a*, *b* и *c*; и т.д. Отрезки в проверяемом решении, изображенные тонкой линией, при проверке можно опускать (это – линии связи, значки перпендикулярности, продолжения прямых), если они не ставятся в сопоставление к требуемым (строка 14).



а) б) в)

Рис. 1. Пример задачи типа (3): а – эталонный чертеж, б – вариант решения

ТАБЛИЦА 2

ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ ЭТАЛОННОГО ЧЕРТЕЖА, ПОКАЗАННОГО НА РИСУНКЕ

№ п/п	Эталон
1	P4294967297 = Point (x=71.31400717100284, y=153.5429359257013, style=6, angle=0.0)
2	P4294967298 = Point (x=131.3140071710028, y=123.5429359257013, style=6, angle=0.0)
3	P4294967299 = Point (x=153.31400717100286, y=153.5429359257013, style=6, angle=0.0)
4	P4294967303 = Point (x=167.31400717100286, y=153.5429359257013, style=6, angle=0.0)
5	P4294967304 = Point (x=153.31400717100286, y=189.40866921607744, style=6, angle=0.0)
6	P4294967307 = Point (x=131.31400717100286, y=244.40866921607744, style=6, angle=0.0)
7	P4294967309 = Point (x=71.31400717100286, y=244.40866921607744, style=6, angle=0.0)
8	P4294967311 = Point (x=167.31400717100286, y=229.40866921607744, style=6, angle=0.0)
9	S4294967312 = Segment (a=0, b=1.0, c=-244.4086692160774, ax=59.31400717100286, ay=244.40866921607744, bx=142.31400717100286, by=244.40866921607744, style=solid)
10	S4294967314 = Segment (a=1.0, b=1.4909090909090905, c=-435.7051140022455, ax=64.0307871733507, ay=249.29375579986853, bx=162.57655836060306, by=183.1959824425651, style=solid)
11	S4294967315 = Segment (a=0, b=1.0, c=-153.5429359257013, ax=65.63839888872633, ay=153.5429359257013, bx=160.4322508509606, by=153.5429359257013, style=solid)
12	S4294967317 = Segment (a=1.0, b=1.9999999999999996, c=-378.39987902240534, ax=68.13105775019316, ay=155.1344106361061, bx=148.1140071710028, by=115.14293592570128, style=solid)
13	S4294967318 = Segment (a=1.0, b=-0.6707317073170735, c=-13.442338794365485, ax=139.79376857617547, ay=188.37849531106207, bx=172.19410636373118, by=236.6844534670542, style=solid)
14	S4294967320 = Segment (a=1.0, b=-0.49999999999999944, c=-95.99999999999997, ax=73.20325813324179, ay=-45.59348373351642, bx=98.89894625252937, by=5.7978925050587975, style=solid)
15	T4294967300 = Text (x=71.09723714133814, y=160.5474977390035, width=6.862985610961914, height=8.358438491821289, angle=0.0, content=A 1)
16	T4294967301 = Text (x=129.09723714133813, y=113.54749773900349, width=6.4661102294921875, height=8.358438491821289, angle=0.0, content=B 1)
17	T4294967302 = Text (x=152.09723714133816, y=160.5474977390035, width=6.069235801696777, height=8.358438491821289, angle=0.0, content=C 1)
18	T4294967305 = Text (x=170.09723714133816, y=151.5474977390035, width=7.259860515594482, height=8.358438491821289, angle=0.0, content=M 1)
19	T4294967306 = Text (x=156.09723714133816, y=191.41323102937963, width=6.7747907638549805, height=8.358438491821289, angle=0.0, content=C 2)
20	T4294967308 = Text (x=128.6754503911547, y=253.69602500228095, width=7.171666145324707, height=8.358438491821289, angle=0.0, content=B 2)
21	T4294967310 = Text (x=67.99475468829431, y=228.41323102937966, width=7.568541049957275, height=8.358438491821289, angle=0.0, content=A 2)

22	T4294967331 = Text (x=160.28862475448955, y=236.66907095823447, width=7.965415954589844, height=8.358438491821289, angle=0.0, content=M 2)
23	T4294967332 = Text (x=157.93160560916283, y=207.1986414450923, width=6.069235801696777, height=8.358438491821289, angle=0.0, content=p 1)
24	T4294967333 = Text (x=157.7739728626443, y=124.91163726172283, width=6.7747907638549805, height=8.358438491821289, angle=0.0, content=p 2)

Сравнение эталона и решения в этом случае можно выполнить при помощи простого лексического анализатора или даже регулярных выражений. Достаточно в эталоне предварительно для каждого объекта задать параметры проверки (обязательный или нет, прямая или отрезок, проверять имя или только индекс и т.п.)

Каждый обязательный объект следует дополнить сообщением, которое отображается в случае его отсутствия – «Горизонтальная проекция перпендикуляра не построена или построена с ошибками». Можно задать несколько сообщений для разных флагов проверки. Более подробное рассмотрение данного вопроса выходит за рамки настоящей работы.

V. Обсуждение результатов

Выше был показан способ, который позволяет организовать проверку решений некоторых классов инженерно-геометрических задач, в частности, из разделов начертательной геометрии, инженерной графики, используя имеющиеся на сегодняшний момент возможности платформы *Stepik*. Однако, не были рассмотрены два существенных вопроса:

- Проверка общих требований к решению, показанная в [12, 13], которая является универсальной для широкого круга задач из некоторого класса, например, правила именования и размещения проекций на двух- и трех- видовых чертежах в задачах начертательной геометрии, а также контекстное сообщение о найденных ошибках. В рассмотренной модели проверки учащийся (студент) получит лишь набор сухих текстовых сообщений в общих чертах указывающих на ошибку – выделить каким-либо образом неверные объекты в окне редактора невозможно, кроме того, вспомогательные построения не проверяются, сколько бы ни было допущено в них ошибок, так как соответствующие объекты не включены в эталон.

ТАБЛИЦА 3
ТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ ЭТАЛОНА И ПРОВЕРЯЕМОГО ЧЕРТЕЖА ПОСЛЕ ОБРАБОТКИ

№ п/п	**Эталон**	**Проверяемое решение**
1	P4294967297 = PointObj (0, x=0.0, y=0.0, style=6, angle=0.0, name=A 1)	P4294967346 = PointObj (0, x=0.0, y=0.0, style=6, angle=0.0, name=A 1)
2	P4294967298 = PointObj (0, x=59.99999999999996, y=-30.0, style=6, angle=0.0, name=B 1)	P4294967347 = PointObj (0, x=59.99999999999997, y=-30.0, style=6, angle=0.0, name=B 1)
3	P4294967299 = PointObj (0, x=82.00000000000001, y=0.0, style=6, angle=0.0, name=C 1)	P4294967348 = PointObj (0, x=82.0, y=0.0, style=6, angle=0.0, name=C 1)
4	P4294967303 = PointObj (0, x=96.00000000000001, y=0.0, style=6, angle=0.0, name=M 1)	P4294967352 = PointObj (0, x=96.0, y=0.0, style=6, angle=0.0, name=M 1)
5	S4294967320 = SLineObj (0, a=1.0, b=-0.49999999999999944, c=-95.99999999999997, ax=73.20325813324179, ay=-45.59348373351642, bx=98.89894625252937, by=5.797892505058797 5, style=solid, name=p 1)	S4294967369 = SLineObj (0, a=1.0, b=-0.5, c=-96.0, ax=73.36688014395321, ay=-45.26623971209358, bx=98.61187898537142, by=5.223757970742838, style=solid, name=n 1)
6		S4294967378 = Segment (0, a=0, b=1.0, c=-5.0, ax=0.0, ay=5.0, bx=5.0, by=5.0, style=thin)
7	S4294967315 = Segment (0, a=0, b=1.0, c=-0.0, ax=-5.675608282276514, ay=0.0, bx=89.11824367995776, by=0.0, style=solid)	S4294967364 = Segment (0, a=0, b=1.0, c=-0.0, ax=0.0, ay=0.0, bx=82.0, by=0.0, style=solid)
8		S4294967365 = Segment (0, a=1.0, b=-0.7333333333333343, c=-82.0, ax=59.99999999999997, ay=-30.0, bx=82.0, by=0.0, style=solid)
9		S4294967373 = Segment (0, a=1.0, b=-0.5000000000000032, c=-91.17403177262399, ax=72.93922541809906, ay=-36.46961270904956, bx=75.14486329081242, by=-32.05833696362285, style=thin)

10		S4294967379 = Segment (0, a=1.0, b=0.0, c=-5.0, ax=5.0, ay=0.0, bx=5.0, by=5.0, style=thin)
11		S4294967377 = Segment (0, a=1.0, b=0.0, c=0.0, ax=0.0, ay=0.0, bx=0.0, by=10.0, style=thin)
12		S4294967372 = Segment (0, a=1.0, b=2.0000000000000147, c=-11.028189363566277, ax=75.14486329081242, ay=-32.05833696362285, bx=79.00563787271335, by=-33.9887242545733, style=thin)
13	S4294967317 = Segment (0, a=1.0, b=1.9999999999999996, c=5.721088371247918e-14, ax=-3.1829494208096776, ay=1.5914747104048104, bx=76.79999999999997, by=-38.40000000000002, style=solid)	S4294967366 = Segment (0, a=1.0, b=1.9999999999999991, c=-0.0, ax=0.0, ay=0.0, bx=59.99999999999997, by=-30.0, style=solid)
14		S4294967368 = Segment (0, a=1.0, b=1.999999999999993, c=-1.9350048067787428e-13, ax=48.59868395406251, ay=-24.299341977031247, bx=76.79999999999998, by=-38.400000000000034, style=thin)

- Не рассмотрены вопросы привязки решения к конкретному студенту – свое ли решение прислал учащийся или скопировал его из чата поддержки *Stepik*, где прошедшие урок учащиеся с удовольствием делятся своими правильными вариантами решений.

Подробное их рассмотрение выходит за рамки настоящей работы. Отметим, однако, направление в котором можно найти приемлемое решение обеих задач. Согласно [17], решение из редактора *CAD*-системы передавалось удаленному серверу для проверки, в ответ в редактор передавалась информация о результатах проверки и найденных ошибках. Эту модель можно частично перенести и в случае использования образовательной платформы типа *Stepik* (рис. 2).

Предлагается использовать идею *External grader* наоборот:

- Решение задачи передается из редактора при помощи сценария с открытым кодом удаленному серверу для предварительной обработки, в ходе которой, например, выявляются общие ошибки оформления, о чем сразу сообщается редактору *CAD*-системы с возможностью пометки ошибок.

Рис. 2. Схема формирования текстового представления решения задачи для проверки

- Удаленный сервер, в том числе, для некоторого класса задач может выполнить преобразование проекций в пространственную модель [12, 13]. В любом случае в редактор возвращается текст, подготовленный для копирования в качестве решения в урок *Stepik*. Таким образом, минимизируется возможность ручного изменения текста решения со стороны пользователя и добавления каких-либо собственных элементов.

- Текст, возвращаемый с удаленного сервера, может, помимо прочего, содержать элементы точной идентификации автора решения (вариант такой проверки показан, например, в [18]), например, текст является зашифрованным и затем расшифровывается уже на стороне проверяющего сценария *Stepik* и решение не засчитывается, если обнаруживаются следы подлога.

VI. Выводы и заключение

Ввиду трудоемкости проверки решений инженерно-геометрических заданий преподавателем, при переходе к цифровому образованию на базе современных интернет-платформ создание средств автоматической проверки инженерно-геометрических заданий приобретает особую актуальность.

В работе был проведен анализ возможностей внедрения/создания инструментов автоматической проверки решений инженерно-геометрических заданий в образовательные интернет-платформы:

- Были рассмотрены возможности образовательных платформ на примере *Stepik*.
- На основе анализа известных систем и подходов к организации проверки геометрических и графических задач предложен способ, который может быть реализован при помощи имеющихся в настоящее время технических возможностей *Stepik*. Он заключается в формировании текстового представления чертежа или трехмерной модели в окне редактора *CAD*-системы с последующей вставкой в окне образовательной платформы *Stepik* и проверкой встроенным *Python*-сценарием или удаленным сервером в режиме External grader.

- Создан *Python*-сценарий, формирующий описание чертежа для одного класса задач начертательной геометрии, и показана принципиальная возможность проверки решения сравнением с эталоном в виде последовательности строк текста.

Дальнейшие исследования в данном направлении позволят создать методику и реализовать проверку различных классов инженерно-геометрических задач на образовательных платформах типа *Stepik*:

- Задачи начертательной геометрии.
- Задачи в проекциях с числовыми отметками.
- Построение аксонометрических изображений.
- Построение теней и перспективных изображений.
- Задачи инженерной и компьютерной графики: построение чертежей деталей, сборочных чертежей и спецификаций, создание трехмерных моделей деталей и сборок.
- Профилирование инструмента.
- Построение номограмм и др.

Список литературы

1. Голощапов В.Г., Урманов В.Г. К совершенствованию методики проверки графических работ студентов // Современное вузовское образование: теория, методология, практика : мат. Межд. уч.-мет. конф. 21–22 марта 2013. Уфа: Башкирский ГАУ, 2013. С. 123–125.

2. Степанова М.В., Степанов В.В. Критериальные параметры при проверке расчетно-графических работ по дисциплине «Начертательная геометрия и инженерная графика» // Научные чтения имени профессора Н.Е. Жуковского : Сб. науч. ст. XIII Межд. науч.-практ. конф. Краснодар, 2023. С. 164–169.

3. Басс Н.В. Методика проверки, приема и оценки графических работ студентов вузов при изучении графических дисциплин // Тенденции развития науки и образования. 2024. №107-1. С. 24–27. DOI: 10.18411/trnio-03-2024-06

4. Кайгородцева Н.В. Рецензия на открытый онлайн курс «Начертательная геометрия и инженерная графика» // Электронный научно-методический журнал Омского ГАУ. 2017. S4. С. 22.

5. Нестерова Т.В., Понетаева Н.Х., Конакова И.П. Автоматизированная проверка чертежа детали при использовании открытых образовательных ресурсов // Инновации в науке и практике : Сб. ст. по мат. III межд. науч.-практ. конф. Часть 2. Уфа: Дендра, 2017. С. 162–167.

6. Ерцкина Е.Б., Орешкова С.П. Практические аспекты дистанционного обучения графическим дисциплинам в техническом вузе в период пандемии // Сибирский педагогический журнал. 2021. №2. С. 80–89. DOI: 10.15293/1813-4718.2102.08

7. Документация типов шагов на Stepik. — Текст : электронный // Stepik [сайт]. Режим доступа: требуется авторизация. URL: https://stepik.org/course/4087/syllabus (дата обращения: 02.06.2024)

8. Петухова А.В. Использование систем электронного тестирования для оценки знаний при обучении студентов вузов CAD- и BIM- комплексам // Инновационные технологии в инженерной графике: проблемы и перспективы : сб. тр. Межд. науч.-практ. конф. 20 апреля 2018 года Брест, Республика Беларусь Новосибирск, Российская Федерация. Брест: БрГТУ, 2018. С. 237–241.

9. Кайгородцева Н.В., Лузгина В.Б. Приобретение практических компетенций посредством открытого онлайн-курса // Омский научный вестник. Серия общество. история. Современность. 2018. №3С. 79–84. DOI: 10.25206/2542-0488-2018-3-79-84

10. Иртегов Д.В. Системы автоматизированной оценки заданий по программированию: разработка, использование и перспективы / Д.В. Иртегов, Т.В. Нестеренко, Т.Г. Чурина // Вестник НГУ. Серия: Информационные технологии. 2019. Т. 17, №2. С. 61–73. DOI: 10.25205/1818-7900-2019-17-2-61-73.

11. Бойков А.А. Автоматизация проверки инженерно–графических заданий // Проблемы качества графической подготовки студентов в техническом вузе: традиции и инновации. Пермь: ПНИПУ, 2016. Т. 1. С. 99–120.

12. Бойков А.А. Компьютерная проверка решений задач начертательной геометрии для инженерно-графического образования // Геометрия и графика. 2020. Т. 8, № 2.С. 66–81. DOI: 10.12737/2308-4898-2020-66-81.

13. Бойков А.А., Федотов А.М. Автоматическая проверка решений задач инженерной геометрии // GraphiCon 2016: Труды 26–й Международной научной конференции. Москва–Протвино, 2016. С. 352–356.

14. Бойков А.А. О реализации автоматического нормоконтроля графической части конструкторских документов для одного класса чертежей деталей // СРТ2020. 8-я Международная конференция «Физико-техническая информатика». Часть 2. Нижний Новгород, 2020. С. 339–346. DOI: 10.30987/conferencearticle_5fd755c0cb0533.83347349

15. Бойков А.А. Верифицируемость инженерно–графических задач как необходимое условие эффективной самостоятельной работы // Проблемы качества графической подготовки студентов в техническом вузе: традиции и инновации. Пермь: ПНИПУ, 2016. Т. 1. С. 177–190

16. Бойков А.А. Анализ подходов к созданию системы компьютерной проверки решений конструктивных задач инженерной геометрии / А.А. Бойков // Надежность и долговечность машин и механизмов. Сб. мат. IX Всерос. науч.–практ. конф. Иваново, 2018. С. 400–403.

17. Бойков А.А., Федотов А.М. Внедрение средств автоматической проверки решений конструктивных задач инженерной геометрии в CAD–систему // GraphiCon 2019 : тр. 29–й Межд. конф. по компьютерной графике и машинному зрению (Брянск, 23–26 сент. 2019 г.). Брянск, 2019. С. 172–175. DOI: 10.30987/graphicon-2019-1-172-175

18. Ермошкин Э.В. Автоматизация контроля работ студентов, выполненных в Компас // Цифровые трансформации в образовании (E-Digital Siberia 2022) : мат–лы VI Межд. науч.-практ. конф. Новосибирск, 2022. С. 115–123.

УДК 378.14:004.9
DOI: 10.25206/978-5-8149-3873-2-2024-1025-1029

Разработка специализированных модулей САПР бакалаврами Омского Государственного технического университета
Development of specialized CAD modules by bachelors of Omsk State Technical University

А.Г. Янишевская, А.О. Звонов
Омский государственный технический университет, г. Омск, Россия

A.G. Yanishevskaya, A.O. Zvonov
Omsk state technical university, Omsk, Russia

Аннотация. Рассмотрен более чем десятилетний опыт преподавателей факультета информационных технологий ОмГТУ в области руководства выполнением выпускных квалификационных работ (ВКР) по запросам промышленных предприятий города Омска в области специализированных САПР. Подчеркнута актуальность разработки специализированных САПР, приведена структура обсуждаемых ВКР. Подробно рассмотрен пример соответствующей бакалаврской работы, оценены результаты данного формата взаимоотношений высшей школы с предприятиями. На основании положительного опыта разработок сделан вывод о высоком уровне подготовки студентов в ОмГТУ.

***Ключевые слова*:** разработка алгоритмов, маршрутные карты, операционные карты, внедрение на производстве.

Аннотация. The work reviews more than ten years of teaching experience of the Information Technologies Faculty from the Omsk State Technical University in the field of management of the final qualification works (FQW) implementation at the request of industrial factories of the city of Omsk in the field of specialized CAD systems. The relevance of the development of specialized CAD is emphasized, the structure of the FQW is given under discussion. An example of a corresponding bachelor's work is considered in detail, the results of this format of relations between higher education and industry are assessed. Based on the positive experience of such developments, a conclusion about the high level of students training at Omsk State Technical University is made.

***Ключевые слова*:** *development of algorithms, route maps, operational maps, implementation in production*.

I. Введение

Системы автоматизированного проектирования (САПР) преимущественно служат для автоматизации рутинного труда конструкторов и других технических специалистов. На рынке существуют как универсальные САПР, предоставляющие широкий набор общих функций, так и специализированные, которые предлагают автоматизацию каких-либо специфических операций.

Расширению применения специализированных САПР в настоящее время способствует целый ряд факторов [1]:

- возможность радикального повышения уровня автоматизации отдельных операций, в т.ч. и в сравнении с САПР общего назначения;
- возрастающая заинтересованность потенциальных заказчиков;
- наличие в САПР общего назначения программных интерфейсов, позволяющих реализовать собственные разработки в виде подключаемых программных модулей;
- затрудненный доступ к ряду иностранных продуктов;
- применение специализированных САПР является проверенным методом повышения эффективности бизнеса [2, 3];
- приобретение специализированных программных решений зачастую является более экономически выгодным [4], особенно учитывая высокие цены на «средние» и «тяжелые» САПР общего машиностроения.

Как отдельный важный фактор необходимо отметить наблюдающийся прогресс технологий и сред программирования приложений. Современные среды программирования включают современные языки, гибкие средства разработки и отладки, высокий уровень автоматизации, что обеспечивает значительное ускорение программирования специализированных САПР и повышение их качества.

Факультет информационных технологий и компьютерных систем Омского Государственного технического университета 16 декабря 2021 года отметил 55 лет со дня основания. Факультет включает в себя кафедру

«Инженерная геометрия и САПР» с не менее богатой историей. Более 10 лет студенты и аспиранты факультета активно участвуют в разработке специализированных САПР для нужд промышленных предприятий города. Сложившиеся у ВУЗа и преподавателей связи с предприятиями позволяют осуществлять подготовку студентов по схеме «обучение – практика – диплом – трудоустройство». На сегодняшний день такие проекты выполнены для АО «ОНИИП», ПАО «Омскнефтехимпроект», НПО «Волна» (филиал в г. Омске), ПО «Иртыш». В настоящей работе рассмотрен пример такого дипломного проекта: практическая ценность, структура, основные методы и результаты.

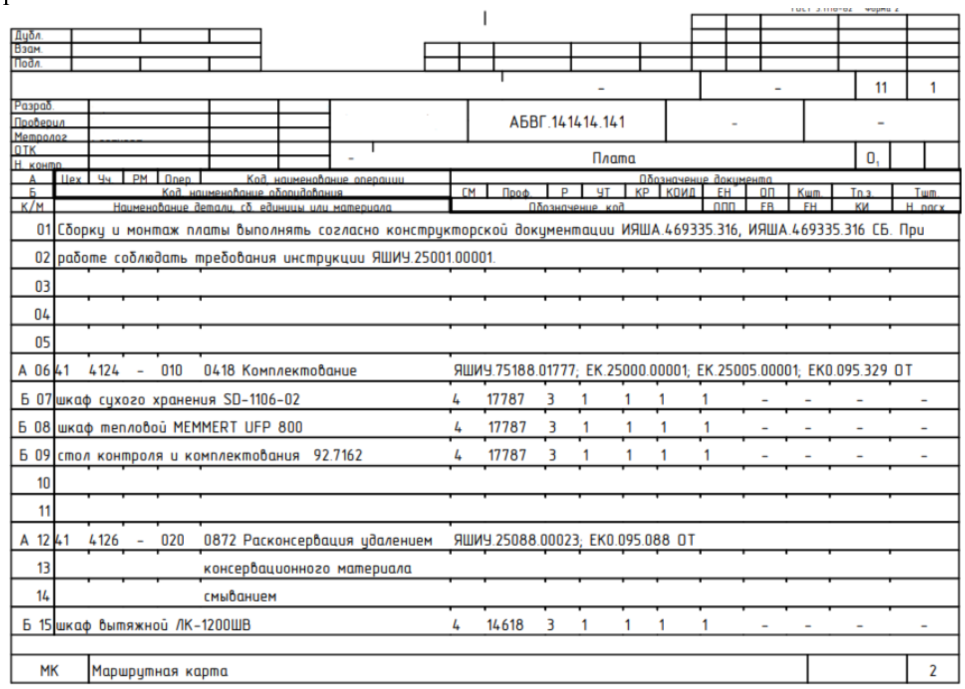
II. Постановка задачи

Как пример выпускной квалификационной работы по программированию специализированного модуля САПР рассмотрена разработка приложения для автоматизации заполнения технологической документации при изготовлении печатной платы. Работа выполнена в Омском государственном техническом университете и защищена в 2024 году.

В качестве заполняемых форм в работе рассматривались маршрутные и операционные карты.

Маршрутная карта в документации (см. рис. 1) – это документ, который описывает очередность выполнения операций и переходов в процессе производства платы. В ней описывается каждое технологическое действие, необходимое для изготовления детали.

Операционная карта (см. рис. 2) – это важный документ технологической документации, который подробно описывает выполнение конкретных операций при изготовлении детали. Она служит руководством для рабочих и технического персонала, обеспечивая точное выполнение каждой операции в соответствии с установленными стандартами и требованиями.

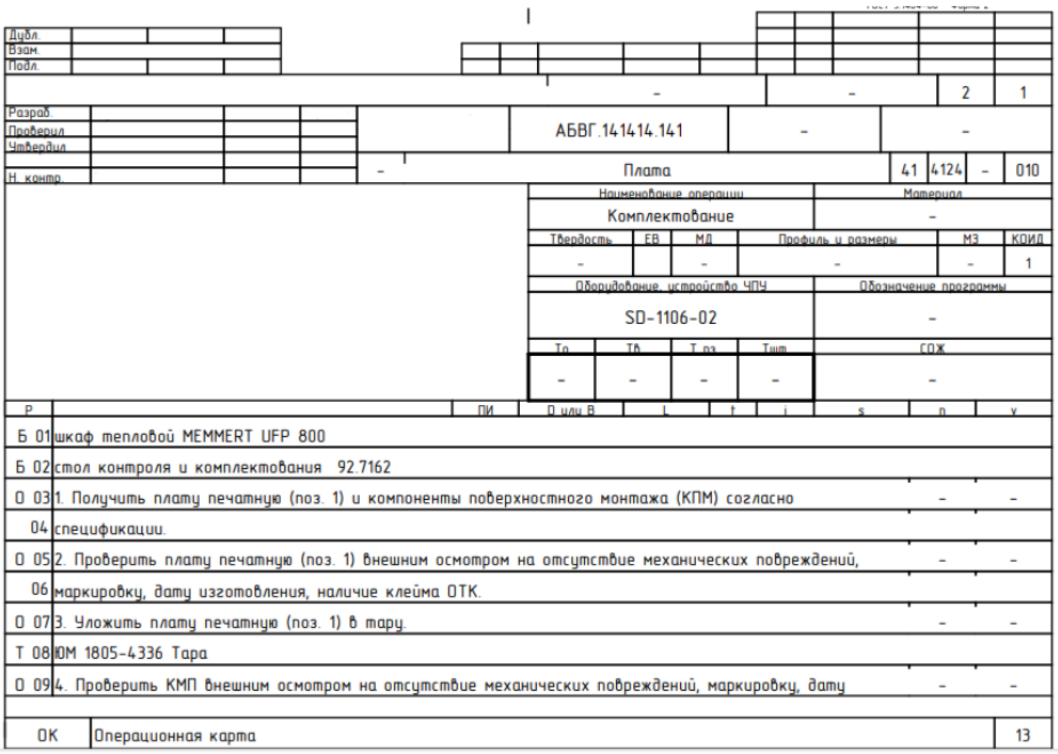


Дубл.
Взам.
Подл.
– – 11 1
Разраб.
Проверил
Метролог
ОТК
Н. контр.
АБВГ.141414.141 – –
– Плата 0,

	Цех	Уч	РМ	Опер	Код, наименование операции	СМ	Проф.	Р	УТ	КР	КОИД	ЕН	ОП	Кшт.	Тп.з.	Тшт.
01	Сборку и монтаж платы выполнять согласно конструкторской документации ИЯША.469335.316, ИЯША.469335.316 СБ. При															
02	работе соблюдать требования инструкции ЯШИУ.25001.00001.															
03																
04																
05																
А 06	41	4124	–	010	0418 Комплектование	ЯШИУ.75188.01777; ЕК.25000.00001; ЕК.25005.00001; ЕК0.095.329 ОТ										
Б 07	шкаф сухого хранения SD-1106-02					4	17787	3	1	1	1	1	–	–	–	–
Б 08	шкаф тепловой MEMMERT UFP 800					4	17787	3	1	1	1	1	–	–	–	–
Б 09	стол контроля и комплектования 92.7162					4	17787	3	1	1	1	1	–	–	–	–
10																
11																
А 12	41	4126	–	020	0872 Расконсервация удалением	ЯШИУ.25088.00023; ЕК0.095.088 ОТ										
13					консервационного материала											
14					смыванием											
Б 15	шкаф вытяжной ЛК-1200ШВ					4	14618	3	1	1	1	1	–	–	–	–

МК Маршрутная карта 2

Рис. 1. Пример маршрутной карты



Дубл.					
Взам.					
Подл.					
	-	-	2	1	
Разраб.					
Проверил	АБВГ.141414.141	-	-		
Утвердил					
Н. контр.	-	Плата	41	4124	-	010

Наименование операции	Материал
Комплектование	-

Твердость	ЕВ	МД	Профиль и размеры	МЗ	КОИД
-		-	-	-	1

Оборудование, устройство ЧПУ	Обозначение программы
SD-1106-02	-

То	Тв	Т п.з.	Тшт	СОЖ
-	-	-	-	-

Р		ПИ	D или В	L	t	i	S	n	V
Б 01	шкаф тепловой MEMMERT UFP 800								
Б 02	стол контроля и комплектования 92.7162								
О 03	1. Получить плату печатную (поз. 1) и компоненты поверхностного монтажа (КПМ) согласно							-	-
04	спецификации.								
О 05	2. Проверить плату печатную (поз. 1) внешним осмотром на отсутствие механических повреждений,							-	-
06	маркировку, дату изготовления, наличие клейма ОТК.								
О 07	3. Уложить плату печатную (поз. 1) в тару.							-	-
Т 08	ЮМ 1805-4336 Тара								
О 09	4. Проверить КМП внешним осмотром на отсутствие механических повреждений, маркировку, дату							-	-

ОК	Операционная карта	13

Рис. 2. Пример операционной карты

III. Теория

Структура обсуждаемых ВКР по разработке специализированных модулей САПР включает следующие основные этапы:

1. Изучение предметной области, анализ существующего опыта и программных решений, выбор языка программирования, инструментальных средств, постановка целей и задач разработки.

2. Теоретическая проработка программного решения: разработка алгоритмов, схем баз данных и т.д.

3. Разработка пользовательских интерфейсов, программная реализация предложенных решений.

4. Тестирование полученного модуля САПР как разработчиком, так и потенциальными пользователями; практическое внедрение модуля на производстве, анализ полученных результатов, получение акта внедрения.

Постановка целей и задач разработки является этапом, на котором необходима наиболее активная помощь преподавателей поскольку с одной стороны, студенты не имеют опыта разработки больших приложений. С другой стороны, специалистам предприятия-заказчика сложно объяснить свои требования студентам.

Учитывая, что длительность выполнения ВКР в среднем составляет 4 месяца, студенты могут уделить достаточное внимание каждому из аспектов работы и получить необходимые консультации, в т.ч. в режиме онлайн. Соблюдению сроков каждого этапа работы необходимо уделять большое внимание. Консультации же могут иметь различный формат, но приоритетными являются практические примеры.

Разработка сравнительно сложного программного решения одним разработчиком имеет свою специфику. Преподаватели ОмГТУ рекомендуют студентам вначале разработать каркас приложения, тщательно провести анализ его работоспособности и удобства для пользователя, а затем увеличивать его функционал.

В рассматриваемой работе в качестве одной из предпосылок разработки рассмотрено распределение времени специалистов по разработке платы (см. таблицу 1). Данные таблицы подтверждают, что автоматизация проектирования схемы и подготовки макета может существенно ускорить разработку плат на предприятии. Мотивация студентов также повышается когда они видят, что могут создать конкурентный и полезный программный продукт.

Далее на рисунках 3 и 4 приведены схемы БД и алгоритма, разработанные в рамках ВКР. Указанное время разработки порядка 1 месяца повышает качество создаваемых БД и алгоритмов, позволяет разработать и сравнить различные варианты решения поставленных заказчиком задач.

ТАБЛИЦА 1
ВРЕМЯ ВЫПОЛНЕНИЯ ЭТАПОВ РАЗРАБОТКИ ПЛАТЫ НА ПРЕДПРИЯТИИ

Этап проектирования	**Время выполнения, ч.**
Проектирование схемы платы	20
Подготовка макета платы	10
Изготовление платы	30
Нанесение паяльной маски и травление	15

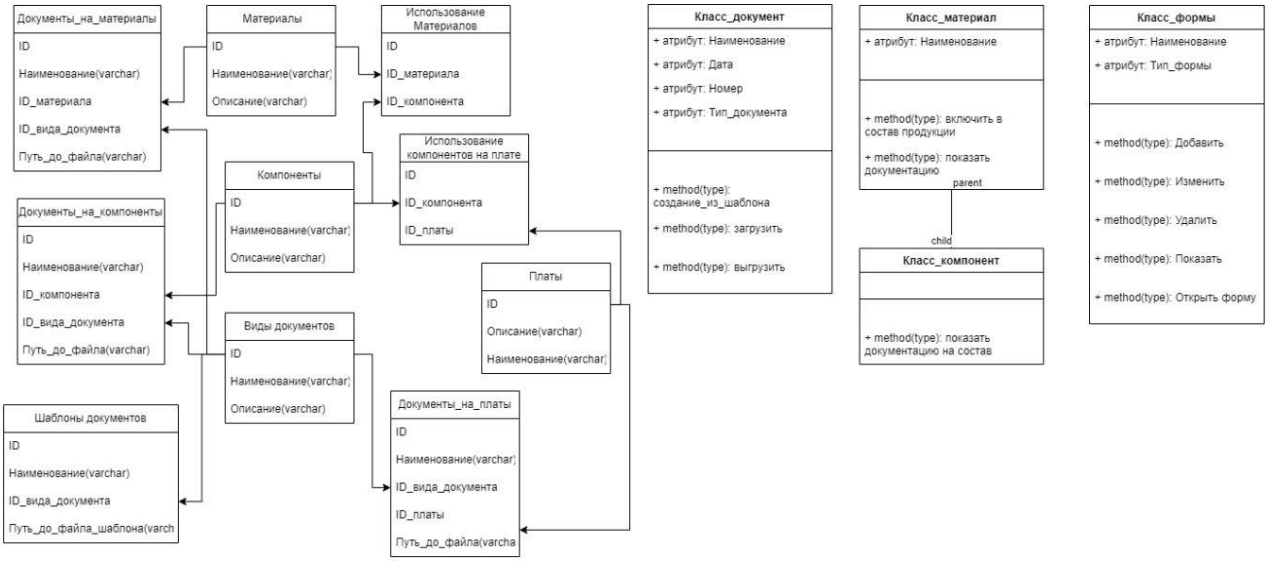


Рис. 3. Разработанная схема БД

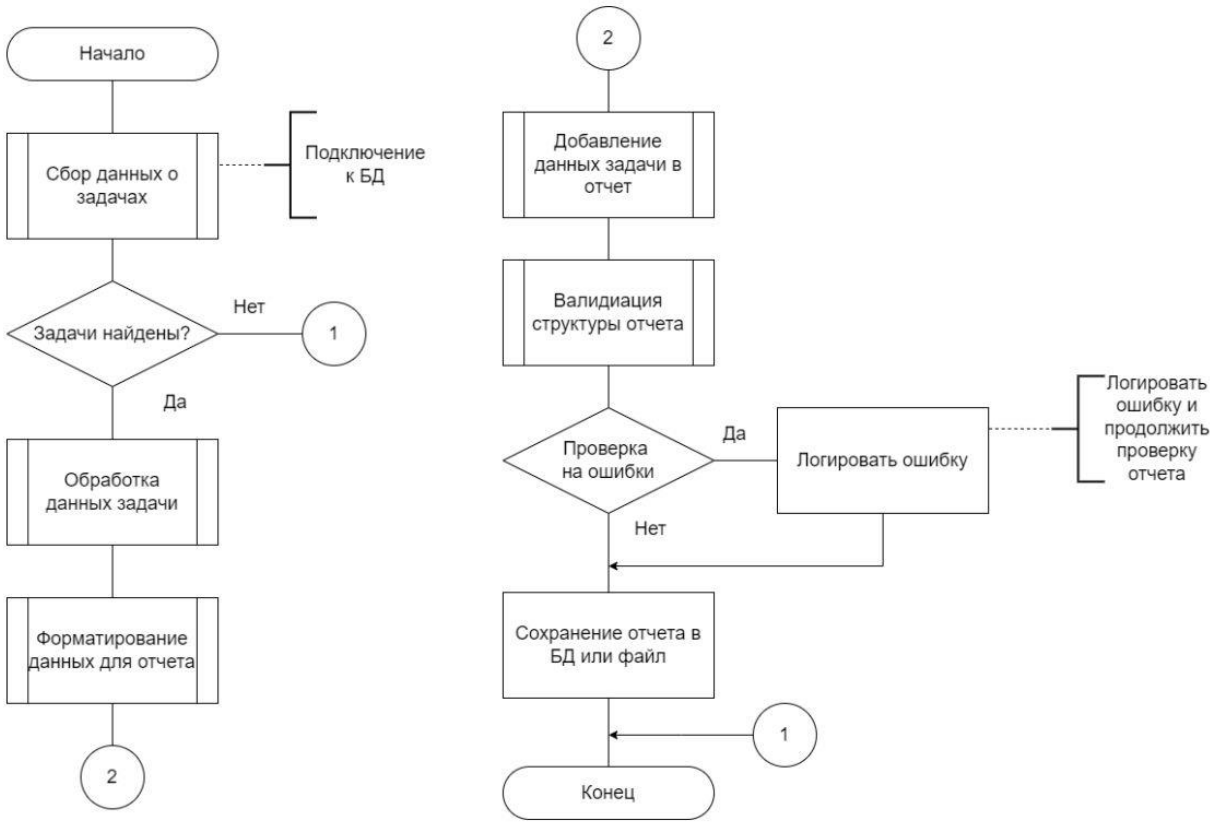


Рис. 4. Алгоритм «создание отчета по задаче»

IV. Результаты

Разработанные алгоритмы и БД реализованы на языке высокого уровня C# (см. рис. 5). Руководители ВКР от предприятия активно участвуют в тестировании полученных разработок и предоставляют свои рекомендации, что обеспечивает обратную связь, необходимую при разработке подобных систем [5].

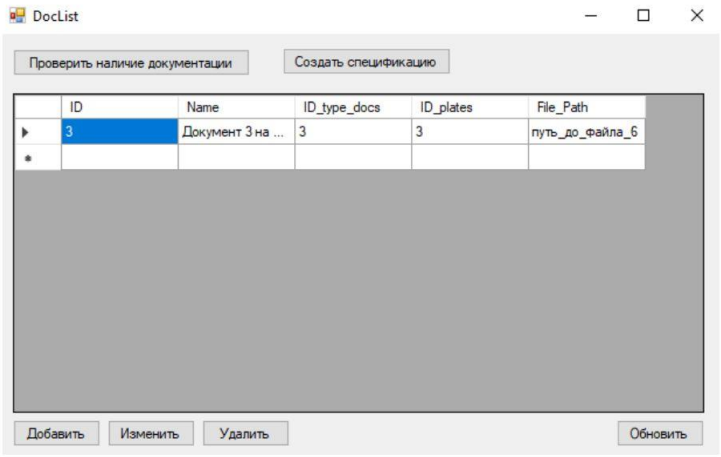


Рис. 5. Основное окно приложения

V. Обсуждение результатов

В качестве логичного завершения практической разработки, студены получают акт внедрения разработки на производстве. Учитывая положительный опыт подобных разработок, описанный подход к организации выполнения ВКР стоит признать успешным, а схему взаимодействия ВУЗа с предприятиями города Омска – взаимовыгодной [6].

VI. Выводы и заключение

Для студента успешная разработка и внедрение программного решения на этапе выпускной квалификационной работы является серьезной объективной проверкой его знаний и навыков. На кафедрах ОмГТУ «Информатика и вычислительная техника», «Инженерная геометрия и САПР» все работы, связанные с разработкой САПР, выполняются студентами самостоятельно, помощь руководителей сводится преимущественно к теоретическим консультациям и проверке оформления текстов работ. Это, несомненно, свидетельствует о том, что выпускники ОмГТУ обладают хорошо сформированными компетенциями в рамках своих специальностей и будут востребованы на рынке труда.

Список литературы

1. Yao, JF., Yang, Y., Wang, XC. et al. Systematic review of digital twin technology and applications // Visual Computing for Industry, Biomedicine, and Art. 2023. Vol. 6, no. 10. https://doi.org/10.1186/s42492-023-00137-4.

2. Фидельман Г.Н., Дедиков С.В., Адлер Ю.П. Альтернативный менеджмент: Путь к глобальной конкурентоспособности. М.: Альпина Бизнес Букс, 2005. 186 с.

3. Хейвуд Дж. Б. Аутсорсинг: в поисках конкурентных преимуществ // Компас промышленной реструктуризации. 2004. № 2. С. 26-28.

4. Груничев Ю.А. Разработка методики расчета эксплуатационных затрат предприятия на ИТ-сервисы. // Управление персоналом. 2009. № 6. С.50-52.

5. Каменев А.В., Ишакова Е.Н. Специфика инженерного образования в области разработки программного обеспечения // Интеллект. Инновации. Инвестиции. Специальный выпуск. 2011. Спец.выпуск. С. 78-81.

6. Уайт Т. Чего хочет бизнес от ИТ: Стратегия эффективного сотрудничества руководителей бизнеса и ИТ-директоров. Пер. с англ. А.Н. Поплавская. Минск: Гревцов Паблишер, 2007. 256 с.

УДК 004.421.6+510.635
DOI: 10.25206/978-5-8149-3873-2-2024-1030-1035

Использование протокола WebSocket для технологии “Internet of things” как элемент электронного образования

Using the WebSocket protocol for the “Internet of things” technology as an element of electronic education

Железняков Р.А.[1], Хачатрян А.А.[2], Берберова М.А.[3], Курнасов Е.В.[4]
[1, 2]Студент РТУ МИРЭА, Москва, Россия
[3]к.т.н., доцент кафедры РТУ МИРЭА, Москва, Россия, заместитель директора АНО «НИЦФТИ», Нижний Новгород, Россия
[4]к.т.н., доцент кафедры РТУ МИРЭА, Москва, Россия

Zheleznyakov R.A.[1], Khachatryan A.A.[2], Berberova M.A.[3], Kurnasov E.V.[4]
[1, 2]Student of RTU MIREA, Moscow, Russia
[3]candidate of technical sciences, Associate Professor of the Department of RTU MIREA, Moscow, Russia, Deputy Director of ANO "SRCIPT", Nizhny Novgorod, Russia
[4]candidate of technical sciences, Associate Professor of the Department of RTU MIREA, Moscow, Russia

Аннотация: В работе представлено исследование протокола WebSocket для изучения в высших учебных заведениях технологии Internet of things. В ходе работы было рассмотрено устройство технологии IoT, а также изучено устройство протокола WebSocket. В работе реализованы несколько различных вариаций использования данного протокола в учебных целях. В дальнейшем был проведен анализ преимуществ и недостатков данного метода, в ходе которого появляется очевидное решение о использовании протокола в учебных целях для построения имитации IoT или формирования Web-платформы.

Ключевые слова: Web-Socket, c++, c#, Python, Internet of things, http,Modbus, Bluetooth, Zigbee,TCP/IP.

Abstract: The paper presents a study of the WebSocket protocol for studying Internet of things technology in higher education institutions. During the work, the device of IoT technology was examined, and the device of the WebSocket protocol was also studied. The work implements several different variations of using this protocol for educational purposes. Subsequently, an analysis of the advantages and disadvantages of this method was carried out, during which an obvious decision appeared to use the protocol for educational purposes to build an IoT simulation or form a Web platform.

Keywords: Web-Socket, c++, c#, Python, Internet of things, http,Modbus, Bluetooth, Zigbee,TCP/IP.

I. Введение

В данный отрезок времени в электронном образовании в сфере производственных систем набирает популярность технология Internet of things – Интернет вещей. Но полная физическая реализация этой технологии достаточно трудна и имеет множество недостатков, вызванных физической ограниченностью и количеством студентов одновременно работающих с этой технологией

О технологиях: Концепция Интернета вещей (IoT) представляет собой сеть, где различные предметы, оснащённые специальными устройствами и технологиями, способны обмениваться информацией между собой и с внешней средой. Эта идея предполагает, что создание таких сетей может кардинально изменить экономическую и социальную жизнь, уменьшая человеческий вклад в некоторые процессы и операции.

В 1999 году эта концепция была заложена как развитие возможностей использования радиочастотной идентификации для связи между объектами и их взаимодействия с окружающей средой. За последние десятилетия, начиная с 2010-х, концепция Интернета вещей наполнилась разнообразием технологических решений, что стало заметной тенденцией в области информационных технологий. Этому способствовало распространение беспроводных сетей, развитие облачных вычислений, прогресс в области взаимодействия между устройствами, массовый переход на IPv6 и внедрение программно-определяемых сетей.

Структура системы Интернета вещей может быть представлена в простейшей форме как трехуровневая: Первый уровень – это само оборудование, Второй уровень – это шлюзы на границе, а Третий уровень – это облачный сервис. На первом этапе работают различные сетевые компоненты, включая датчики и элементы управления, которые являются основой Интернета вещей. Они обычно поддерживают такие технологии как Modbus, Bluetooth, Zigbee или специализированные протоколы для взаимодействия с шлюзами на границе.

Шлюзы на границе представляют собой системы для сбора информации от датчиков, которые также отвечают за предварительную обработку данных, соединение с облачным сервисом.

о технологии Websocket: WebSocket был создан для интеграции в web сервисы и серверные системы, однако его применение не ограничено только ими. Этот протокол, который функционирует на базе TCP, обеспечивает гибкость и универсальность, позволяя использовать его в различных клиентских и серверных решениях. WebSocket открывает новые горизонты для взаимодействия между интернет-браузером и web-ресурсом, стимулируя развитие интерактивного контента и поддержку приложений, работающих в режиме реального времени.

II. Постановка Задачи

Основной целью работы является изучение протокола WebSocket как эмуляцию физической передачи данных, так и для использования в системах IoT мелкопроизводственного или учебного характера. Изучение протокола WebSocket позволяет использовать его в целях упрощения обучения новых специалистов так как позволяет эмулировать процесс работы технологии Internet of things.

III. Реализация протокола WebSocket с помощью языков программирования

Ниже представлены 3 различные реализации протокола WebSocket, позволяющие понять строение и концепцию данного метода.

Часть 1- Python

Реализация такого типа позволяет использовать протокол WebSocket без глубокой настройки. Это самый простейший способ соединить два устройства используя данный протокол. В данном примере (рисунок 1) используется реализация библиотеки FastAPI – достаточного мощного инструмента разработки web приложений.

Socket.py

```python
from fastapi import FastAPI, WebSocket

app = FastAPI()

@app.websocket("/ws")
async def websocket_endpoint(websocket: WebSocket):
    await websocket.accept()
    while True:
        data = await websocket.receive_text()
        await websocket.send_text(f"said: {data}")
```

Рисунок 1 – WebSocket on python

Часть 2- C#

Данный способ реализации протокола использует библиотеку TCP LISTENNER (рисунок 2) и предлагает гибкую настройку опытному пользователю. Таким образом пользователь может использовать эти возможности для реализации метода "handshake" и прочих других методов – формирующих дополнительные настройки, в том числе и безопасности.

```csharp
using System.Net.Sockets;
using System.Net;
using System;

Ссылок: 0
class Server
{
    Ссылок: 0
    public static void Main()
    {
        TcpListener server = new TcpListener(IPAddress.Parse("127.0.0.1"), 80);

        server.Start();
        Console.WriteLine("Server has started on 127.0.0.1:80.{0}Waiting for a connection…", Environment.NewLine);

        TcpClient client = server.AcceptTcpClient();

        Console.WriteLine("A client connected.");
    }
}
```

Рисунок 2 – WebSocket on C#

Часть 3- C++

Последний представленный тип реализации (рисунки 3.1 и 3.2) использует язык программирования c++ который в свою очередь представляет большие возможности контроля памяти, что позволяет создать самую тонкую конфигурацию данного протокола. Использование работы с памятью позволяет ускорить передачу и прием данных и позволяет вручную создать необходимые настройки безопасности.

```cpp
#include "easywsclient.hpp"
#include <iostream>
#include <string>
#include <memory>
#include <mutex>
#include <deque>
#include <thread>
#include <chrono>
#include <atomic>

// a simple, thread-safe queue with (mostly) non-blocking reads and writes
namespace non_blocking {
template <class T>
class Queue {
    mutable std::mutex m;
    std::deque<T> data;
public:
    void push(T const &input) {
        std::lock_guard<std::mutex> L(m);
        data.push_back(input);
    }

    bool pop(T &output) {
        std::lock_guard<std::mutex> L(m);
        if (data.empty())
            return false;
        output = data.front();
        data.pop_front();
        return true;
    }
};
}

// eastwsclient isn't thread safe, so this is a really simple
// thread-safe wrapper for it.
class Ws {
    std::thread runner;
    non_blocking::Queue<std::string> outgoing;
    non_blocking::Queue<std::string> incoming;
    std::atomic<bool> running { true };

public:
    void send(std::string const &s) { outgoing.push(s); }
    bool recv(std::string &s) { return incoming.pop(s); }

    Ws(std::string url) {
        using easywsclient::WebSocket;

        runner = std::thread([=] {
            std::unique_ptr<WebSocket> ws(WebSocket::from_url(url));
            while (running) {
```

Рисунок 3.1 – WebSocket on c++

```
    std::atomic<bool> running { true };

public:
    void send(std::string const &s) { outgoing.push(s); }
    bool recv(std::string &s) { return incoming.pop(s); }

    Ws(std::string url) {
        using easywsclient::WebSocket;

        runner = std::thread([=] {
            std::unique_ptr<WebSocket> ws(WebSocket::from_url(url));
            while (running) {
                if (ws->getReadyState() == WebSocket::CLOSED)
                    break;
                std::string data;
                if (outgoing.pop(data))
                    ws->send(data);
                ws->poll();
                ws->dispatch([&](const std::string & message) {
                    incoming.push(message);
                });
                std::this_thread::sleep_for(std::chrono::milliseconds(10))
            }
            ws->close();
            ws->poll();
        });
    }
    void close() { running = false; }
    ~Ws() { if (runner.joinable()) runner.join(); }
};

int main() {
    Ws socket("ws://localhost:40800");

    std::atomic<bool> run{true};
    auto receiver = std::thread([&] {
        std::string s;
        while (run) {
            if (socket.recv(s))
                std::cout << s << '\n';
            std::this_thread::sleep_for(std::chrono::milliseconds(10));
        }
    });

    std::string line;
    while (std::getline(std::cin, line))
        socket.send(line);
    run = false;
    receiver.join();
    socket.close();
}
```

Рисунок 3.2 – WebSocket on c++

IV. Использование протокола WebSocket в технологии IOT

Использование данного протокола представляет собой шлюз, направленный на передачу данных от полевого уровня к экрану так и от экрана к конечному уровню (Оператор, пункт диспетчеризации). Конфигурации

технологии WebSocket верхнего уровня могут использоваться для передачи данных от шлюза к устройству контроля системы IOT. Конфигурации нижнего уровня в свою очередь способны запускаться на полевом уровне – при поддержке устройства полевого уровня сетевых протоколов. В данном случае наблюдается огромное количество возможных способов использования WebSocket, ведь полевой уровень способен мгновенно передавать данные через физическую шину на устройство, у которого больше возможностей работы с web-пространством (рисунок 4). На рисунке 4 представлен учебный макет IoT позволяющий использовать для передач данных протокол WebSocket.

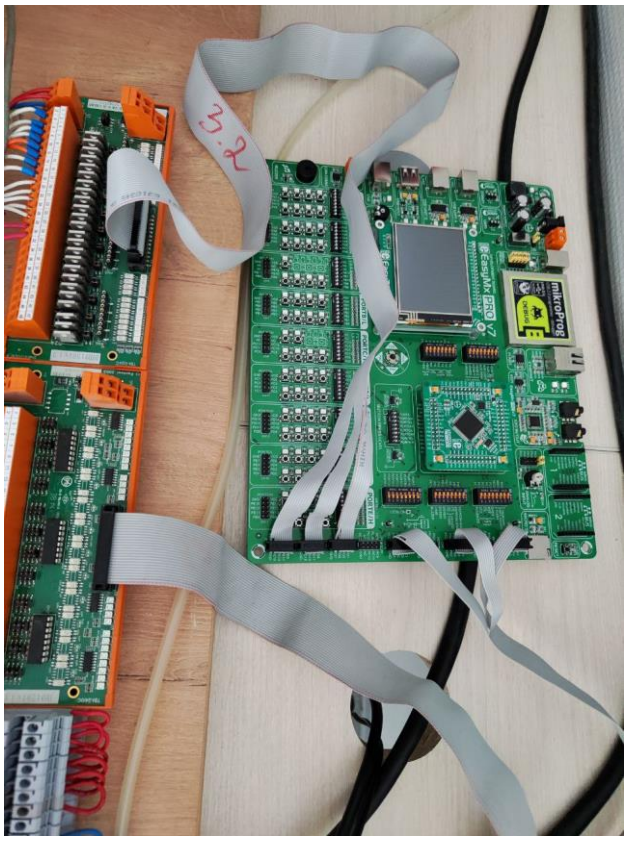


Рисунок 4 – Устройство stm32 (программируемый контроллер) с поддержкой физической передачи данных

V. Преимущества и недостатки метода

К числу основных преимуществ использования websocket для IoT в образовании можно отнести:

- Низкую стоимость реализации – использование протокола WebSocket является бесплатным на соответствующем оборудовании (к примеру, с поддержкой TCP/IP)

- Простота в использовании – использование языков программирования высокого уровня понижает порог вхождения в использование данной технологии

- Высокая скорость передачи данных – передача данных с помощью вышеупомянутого протокола позволяет передавать данные максимально быстро так как соединение происходит всего один раз, при подключении устройств между собой (в отличии от стандартного метода Http запросов в случае реализации web контроля)

- Асинхронность – поддержка протоколом использования многопоточных вычислений позволяет выдерживать высокую нагрузку на устройство, а также позволяет сохранять высокую скорость ответа для большого количества устройств

- Двунаправленные запросы – основное преимущество технологии WebSocket, оно позволяет посылать сигналы управления, не дожидаясь запроса от полевого уровня что позволяет корректировать работу с помощью web интерфейса.

- Минимизация коллизии сигналов – последние версии протокола поддерживают отправку данных фреймами, что позволяет реализовать разные способы минимизации коллизии, например создание

"депо" для принятых сигналов с их последующим решением асинхронной системой, что в свою очередь формирует систему гибкого реального времени

Таким образом использование WebSocket выглядит очень привлекательным для использования в IoT в образовании, но стоит выделить несколько недостатков:

- В случае глубокой настройки профиля передачи данных требуется высокий уровень квалификации инженера-программиста, так как требуется обеспечить бесперебойную передачу данных.
- Формирование надежной защиты – как и любое web соединение, WebSocket подвержен угрозе проникновения извне. Задача инженера состоит в формировании правильной аутентификации.
- Маленький радиус применения – данный протокол обеспечивает систему лишь приближенную к системе жесткого реального времени что не подходит для крупного производства и производства в условиях повышенной опасности, где потеря соединения приведет к фатальной ошибке и поломке оборудования.

VI. Заключение

В заключении стоит отметить, что данная методика может быть использована в любых IoT-системах, начиная от повседневного использования, заканчивая эксплуатацию в профессиональных целях, но использование должно сопровождаться высококвалифицированными специалистами, что делает данный протокол хорошей основой для изучения IoT в вузах. В рассмотренных реализациях работы можно выбрать степень нагрузки в зависимости от языка программирования, что делает WebSocket широко применимым в разных направлениях обучения как для имитации физических протоколов, так и для реализации IoT с Web интерфейсом.

Список литературы

1. Баланов А.Л. IoT-решения: принципы, примеры, перспективы // 2024 г. С. 3 – 6.
2. Что такое IoT и что о нем следует знать [Электронный ресурс] // URL: https://habr.com/ru/companies/otus/articles/549550/
3. All Things Considered: An Analysis of IoT Devices on Home Networks // Deepak Kumar, University of Illinois at Urbana-Champaign 2019 г. С. 1169 – 1170.
4. Груздева М.Л., Груздева К.Е. ПЛЮСЫ И МИНУСЫ «ИНТЕРНЕТА ВЕЩЕЙ» В ОБРАЗОВАНИИ // ШКОЛА БУДУЩЕГО №3 2020 г. С. 292 – 299.
5. Хабаров С.П., Шилкина М.Л.: "Построение распределенных систем на базе WebSocket. Учебное пособие для вузов" // Лань 2022г.

УДК: 004.75
DOI: 10.25206/978-5-8149-3873-2-2024-1036-1041

Средства безопасности и предохранения баз данных в IoT современного вуза
Security tools and database protection in the IoT of a modern university

Железняков Р.А.[1], Хачатрян А.А.[2], Берберова М.А.[3], Курнасов Е.В.[4]
[1, 2]Студент РТУ МИРЭА, Москва, Россия
[3]к.т.н., доцент кафедры РТУ МИРЭА, Москва, Россия, заместитель директора АНО «НИЦФТИ», Нижний Новгород, Россия
[4]к.т.н., доцент кафедры РТУ МИРЭА, Москва, Россия

Zheleznyakov R.A.[1], Khachatryan A.A.[2], Berberova M.A.[3], Kurnasov E.V.[4]
[1, 2]Student of RTU MIREA, Moscow, Russia
[3]candidate of technical sciences, Associate Professor of the Department of RTU MIREA, Moscow, Russia, Deputy Director of ANO "SRCIPT", Nizhny Novgorod, Russia
[4]candidate of technical sciences, Associate Professor of the Department of RTU MIREA, Moscow, Russia

Аннотация: В работе проведены теоретическое описание концепции безопасности и предохранения баз данных современного вуза. В качестве объекта изучений была взята IoT-концепция. В ходе исследования были изучены главные угрозы безопасности, с которыми могут столкнуться системы Интернета вещей (IoT), такие как атаки на сетевом уровне, уязвимости в программном обеспечении и другие риски. Были рассмотрены способы шифрования данных, аутентификации пользователей и управления доступом, которые были внедрены в приложение с целью повышения уровня безопасности. Также был разработан метод решения данных угроз безопасности, в число которых является настройки шифрования, резервное копирование данных, настройка подключений баз данных, и другие способы. Работа будет полезна разработчикам, системным администраторам и исследователям в области IoT, стремящимся обеспечить надежную защиту данных и безопасность своих систем в условиях современного образовательного процесса.

Ключевые слова: IoT-системы, образование, информационная безопасность, современный вуз, базы данных, шифрование данных, SQL, PostgreSQL, протоколы связи, конфигурационные файлы, DDoS-атаки, резервное копирование.

Abstract: The paper provides a theoretical description of the concept of security and protection of databases of a modern university. The IoT concept was taken as the object of study. The study examined the main security threats that Internet of Things (IoT) systems may face, such as network-level attacks, software vulnerabilities, and other risks. The methods of data encryption, user authentication, and access control that were implemented in the application to improve security were considered. A method for solving these security threats was also developed, including encryption settings, data backup, database connection settings, and other methods. The work will be useful for developers, system administrators, and researchers in the field of IoT seeking to ensure reliable data protection and the security of their systems in the context of the modern educational process.

Keywords: IoT systems, education, information security, modern university, databases, data encryption, SQL, PostgreSQL, communication protocols, configuration files, DDoS attacks, backup.

I. Введение

Интернет вещей (Internet of Things, IoT) – это современная концепция, которая включает в себя множество физических объектов, подключённых к сети Интернет и способных обмениваться между собой данными [1]. Данная концепция существенно улучшает сферы, в которых она находится. Она помогает пользователям в создании более умных, безопасных и удобных в эксплуатации систем. Примеры Интернета вещей варьируются от маленьких приспособлений, который человек может переносить с собой, до концепции Умного дома [2]. В нём существуют различные устройства. Системы подключены и общаются друг с другом, чтобы автоматизировать и оптимизировать домашнюю среду. Для этого используются такие компоненты, как: умные датчики, которые собирают информацию об окружающей среде, умные устройства, которые могут управляться удалённо, голосовые помощники, благодаря которым пользователи могут взаимодействовать с домом с помощью голосовых команд и другие средства.

Данные устройства, как и любые другие, подвержены атакам злоумышленников. По данным компании Avast, совместно со Стэндфордским университетом, в 2019-м г. из 83 000 000 IoT-устройств около 6 000 000 устройств подвержено атакам [3]. Все эти инциденты связаны с использованием устаревших протоколов связи FTP и Telnet. В случаях, если к устройствам был получен доступ, это может обернуться для владельца IoT-устройства

следующими последствиям. Контроль над устройством: хакеры способны захватить управление над устройством, осуществляя его активацию и деактивацию, корректируя настройки или извлекая информацию. Доступ к данным: протоколы, не обновленные до современных стандартов, часто имеют уязвимости, которые злоумышленники могут использовать для доступа к приватной информации, включая личные сведения, коды доступа и финансовые данные. Распространение вредоносных программ: используя устаревшие протоколы, киберпреступники могут загрузить и активировать вредоносные программы на устройстве, что может привести к его дальнейшему взлому и воздействию на сеть. Формирование бот-сетей: злоумышленники могут задействовать взломанные IoT-устройства для создания бот-сетей, которые применяются для осуществления распределенных атак "отказ в обслуживании" (DDoS) и других видов киберпреступлений. Шпионская деятельность: Устройства IoT часто служат инструментом для сбора сведений о пользователях и их окружении. Хакеры могут воспользоваться устаревшими протоколами для получения доступа к этим данным с целью шпионажа или вымогательства.

Широко распространенные датчики и возможность соединить материальный и цифровой мир формируют фундаментальные принципы новой методики обучения. Этот переход к новой парадигме опирается на идею интеграции датчиков во все объекты и использование сетевых связей для интеграции миллиардов устройств в глобальную сеть Интернет [4]. По этой причине образовательных учреждениях наблюдается рост использования IoT-устройств, таких как умные доски, системы управления освещением и климатом, а также устройства для мониторинга температуры студентов. Также вузы хранят и обрабатывают большое количество личной информации о студентах и сотрудниках, включая данные о успеваемости, тьюторах, студентах и другие отчёты внутри учебного заведения [5]. В случае, если злоумышленник получит доступ к IoT устройству в рамках вуза, необходимо обеспечить дополнительную защиту каждого элемента. Утечка этих данных может привести к серьезным последствиям, включая нарушение конфиденциальности и репутационные потери. В данном случае, будет подробно рассмотрен сегмент с данными, который является ключевым элементом в обеспечении безопасности и стабильности работы IoT устройств.

II. Постановка задачи и материалы исследования

Задачей данного исследования является анализ IoT-систем, связанных с хранением и обработкой данных, а также разработке и оценке методов и средств защиты баз данных от потенциальных угроз. Цель исследования – предложить комплексный подход, ориентирующийся на устойчивость защиты баз данных от злоумышленников.

На данный момент распространён следующий способ сбора, передачи и хранения данных. В первую очередь происходит сбор данных IoT на уровне прошивки. На этапе разработки устройства реализуется механизм сбора необходимых данных с периферийных устройств, включая датчики, исполнительные механизмы и встроенные модули. Данные IoT могут храниться следующими способами. Первым методом являются локальные или облачные хранилища, в которые данные поступают с устройств непосредственно после сбора. Данный метод требует стабильного сетевого подключения, для передачи данных в режиме реального времени. Второй метод уровня прошивки представляет из себя хранение во внутренней памяти устройства. Данный способ подходит для устройств, для которых невозможно настроить стабильное сетевое подключение. В таких случаях данные сохраняются локально и передаются на сервер впоследствии. Также существует способ сбора данных IoT на стороне сервера. Для обеспечения бесперебойного сбора и хранения данных, необходимо выполнить следующие задачи: разработать серверный API для приема данных со всех подключенных устройств, организовать хранение полученных данных для последующей обработки и обеспечить масштабируемости серверной части платформы для обработки больших объемов данных.

В случае, если выбран способ сбора данных на отдельное хранилище или облако, то это в большинстве случаев будет удовлетворять безопасности устройств. Тем более, если говорим об использовании MQTT протокола для передачи данных. Данный метод сбора информации IoT-устройств упрощает реализацию прошивки. Им нужно только подключиться к брокеру сообщений и обновить свои значения датчиков и телеметрии, не сохраняя их в памяти. Сервер, в свою очередь, отвечает за сбор и сохранение данных. Однако, использование устаревших протоколов, таких как FTP, маловероятно обеспечит сохранность данных. FTP-протокол не обеспечивает шифрования, что делает его уязвимым для перехвата данных и атак типа "злоумышленник посередине".

Когда в качестве основного метода сбора данных используется метод прошивки устройства, то он работает по следующему принципу. Устройство интернета вещей функционирует независимо и накапливает определенное количество информации в процессе своей работы, при этом оно может не поддерживать связь с сервером из-за особенностей его работы или по другим причинам. Тем не менее, благодаря настройке точного времени, прошивка устройства фиксирует все происходящие события с соответствующими временными метками в своей памяти. Когда устройство в итоге подключается к сети, оно отправляет собранные данные на сервер, который, в

свою очередь, интегрирует их в свою базу данных. Благодаря привязке каждого события к конкретному времени, информация становится доступной для пользователя или администратора с учетом времени ее возникновения. В данном случае возникают риски получения данных непосредственно через IoT устройство.

В результате анализа данного метода сбора информации из IoT-систем можно заключить, что надежность системы в значительной степени определяется ее архитектурой. Однако для уже существующих систем необходимо разработать подход, который позволит повысить уровень безопасности баз данных. Это позволит минимизировать уязвимости и обеспечить защиту данных в условиях постоянно меняющихся угроз.

III. Разработанная методика дополнительной защиты баз данных IoT-систем

Чтобы обеспечить безопасность баз данных, в первую очередь необходимо настроить конфигурационные файлы выбранной базы данных. Например, в PostgreSQL основным конфигурационный файлом является postgresql.conf [6]. В нём можно увидеть прослушиваемый IP-адрес, порт, через который происходит общение с сервером, метод шифрования паролей пользователей, параметр шифрования соединения и другие параметры. Устройства IoT часто имеют слабую защиту, например, используют пароли по умолчанию, что делает их легкодоступными для хакеров. Поэтому важно всегда вмешиваться в конфигурационный файл, для стандартного его изменения.

Значительная часть сетевого трафика от устройств IoT по-прежнему не зашифрована, что означает, что угрозы, такие как программы-вымогатели, могут получить доступ к личным данным. Для того, чтобы свести это к минимуму, нужно также заняться шифрованием данных. На данный момент существует спектр способов шифрования данных. Самые популярные из них это MD5, SHA-256, SCRAM-SHA-256, AES и другие. В случае, когда необходимо зашифровать пароль пользователя, лучше использовать SCRAM-SHA-256, поскольку он предназначен для хеширования паролей. А это значит, что пользователю не имеет никакого смысла заново получать информацию о паролях, ему необходимо только вводить его. И система будет сравнивать хеш полученного значения с внесённым в базу данных. Тем не менее, для конфиденциальных данных, хранящихся в базе данных, требуется дополнительный уровень безопасности для предотвращения несанкционированного доступа. Для этой цели используется симметричное шифрование, такое как стандарт шифрования AES. AES шифрует данные с помощью сгенерированного ключа, а для расшифровки требуется тот же ключ.

Создание резервных копий данных – это ключевая мера для предотвращения их утраты или повреждения [7]. Существует несколько подходов к резервированию: Полное резервное копирование: захватывает все данные на носителе, включая системные файлы и метаданные. Такой метод обеспечивает максимальную защиту, однако требует значительных временных и ресурсных затрат. Инкрементное резервное копирование: сохраняет только те файлы, которые были изменены или добавлены с момента последнего резервирования. Это позволяет сэкономить пространство и время, но для полного восстановления системы потребуется исходная полная копия и все последующие инкрементные. Дифференциальное резервное копирование: фокусируется на файлах, измененных после последнего полного резервирования. Для восстановления системы достаточно иметь полную копию и последнюю дифференциальную. Клонирование: создает полную копию всего диска или его раздела, что позволяет быстро восстановить систему. Однако этот метод также требует значительного объема памяти и ресурсов.

Ограничение количества соединений с базой данных играет важную роль в структуре систем, особенно в сфере Интернета вещей (IoT), где устройства активно обмениваются данными с сервером. Это ограничение выполняет несколько функций: во-первых, оно предотвращает перегрузку сервера, обеспечивая стабильную работу системы даже при высокой нагрузке; во-вторых, это важный элемент безопасности, который уменьшает риск атак типа "отказ в обслуживании" (DDoS), когда злоумышленник пытается исчерпать ресурсы базы данных, создавая множество соединений. В контексте IoT, где множество устройств может одновременно обращаться к базе данных, правильная настройка ограничений на количество соединений помогает обеспечить надежность и безопасность системы, снижая вероятность сбоев и утечек данных. Это особенно важно для защиты конфиденциальной информации и сохранения целостности данных в динамичной среде.

Общая схема работы метода обеспечения безопасности баз данных в IoT-системах представлена на рис. 1. Использование перечисленных критериев позволяет обеспечить надежную защиту данных, соответствие нормативным требованиям и удобство управления системами шифрования баз данных.

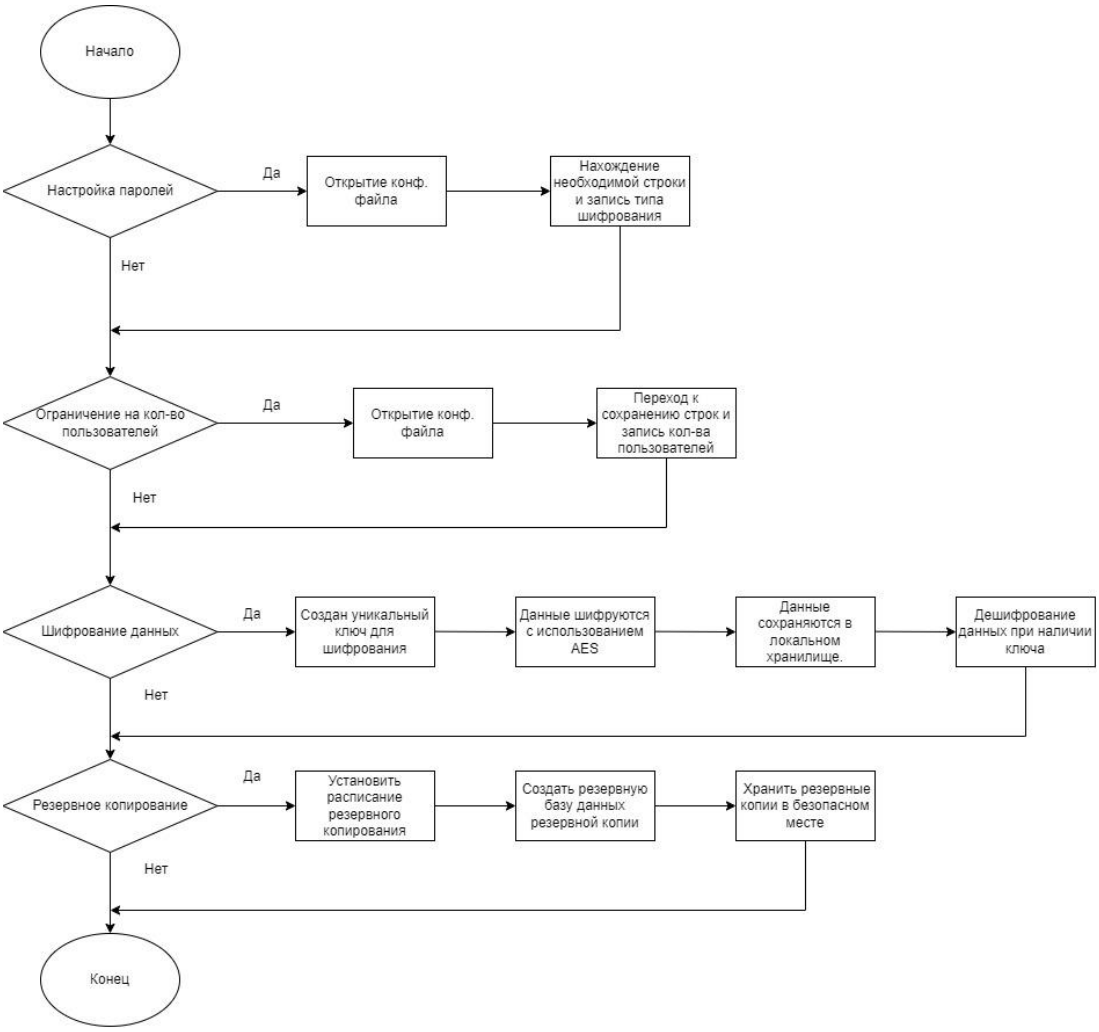


Рис.1 - Схема процесса настройки защиты IoT-системы

Базовыми критериями оценки качества работы системы шифрования баз данных, настройки шифрования паролей, настройки количества подключений к базе данных и частоты резервного копирования являются: Конфиденциальность: Данные должны быть защищены от несанкционированного доступа и раскрытия. Целостность: Данные должны быть защищены от несанкционированного изменения или уничтожения. Доступность: Данные должны быть доступны для авторизованных пользователей и приложений своевременно и надежно. Производительность: Система шифрования не должна существенно влиять на производительность базы данных. Удобство использования: Система шифрования должна быть простой в использовании и управлении для системных администраторов.

IV. Демонстрация решения

На рис.2 представлен пример того, как может выглядеть данное приложение. На нём представлен базовый функционал, который был описан ранее.

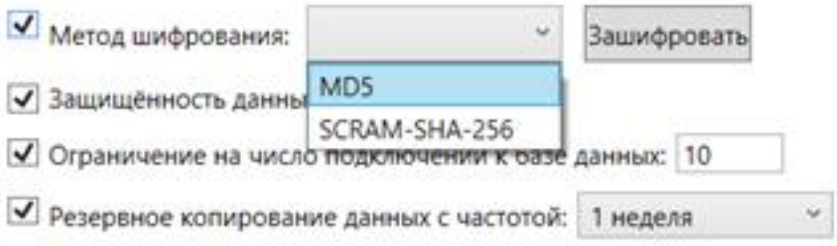


Рис.2 - Приложение для настройки базы данных IoT-системы

Процесс защиты данных происходит по следующему принципу. Сперва создаётся ключ шифрования. Этот ключа может быть разной длинны, но для минимизации шансов подбора ключа третьими лицами, функция создаёт 256-битный ключ. Ключ должен храниться в безопасности, так как его утечка может привести к компрометации зашифрованных данных.

После этапа создания объекта AES начинается формирование. Этот объект представляет собой конкретную реализацию алгоритма шифрования, который предназначен для зашифровки и расшифровки информации. AES оперирует данными в блоках фиксированного размера (128 бит или 16 байт) [8]. Это означает, что информация разбивается на отдельные блоки, каждый из которых обрабатывается независимо. Процесс происходит в режиме CBC. [9] При этом каждый блок шифруется с использованием результатов шифрования предыдущего блока. Такой подход обеспечивает повышенную безопасность, поскольку одни и те же входные данные при разном положении будут иметь различные зашифрованные результаты. Для работы в режиме CBC необходим инициализационный вектор (IV). IV представляет собой случайный набор байтов, который гарантирует уникальность процесса шифрования. IV не обязан быть секретным, однако он должен быть уникальным для каждой сессии шифрования.

Процесс шифрования начинается после этапа подготовки данных. Если объем информации, который требуется зашифровать, не соответствует стандартному размеру блока AES (16 байт), необходимо добавить дополнительные байты (паддинг), чтобы общая длина данных стала кратной 16. После этого данные передаются в объект AES для шифрования. Каждый блок данных обрабатывается с использованием ключа и инициализирующего вектора (IV). В итоге получается зашифрованный текст, который представляет собой произвольный набор байтов. IV сохраняется для последующей дешифрации сообщений, обычно в начале зашифрованных данных. Зашифрованные данные (включая IV) представляют собой бинарные данные, которые могут содержать символы, несовместимые с текстовыми форматами. Преобразование в Base64 позволяет перевести эти бинарные данные в строку ASCII, которая безопасно передается через текстовые интерфейсы. В результате выполнения функции возвращается строка с закодированными в Base64 данными, готовыми к сохранению или передаче.

V. Выводы и заключение

В ходе работы были проведены теоретические и экспериментальные исследования средств безопасности и предохранения баз данных в IoT-системах современного вуза. Были получены следующие результаты.

1. Основные проблемы безопасности IoT-устройств в вузах заключаются в использовании устаревших технологий, такие как: устаревшие протоколы, такие как FTP и Telnet, слабые пароли по умолчанию, которые пользователи не изменяют и по другим причинам.
2. Были разобраны базовые принципы безопасной эксплуатации IoT-системы. Совмещение нескольких базовых методов в случае, если злоумышленник получит доступ к устройству, позволяет обеспечить конфиденциальность, устойчивость к атакам распределенного типа "отказ в обслуживании" (DDoS), и восстановить систему в случае потери данных.
3. Разработано приложение, которое позволяет пользователю обеспечить предохранение базы данных от третьих лиц. Данное приложение выполняет следующие функции:

3.1. Позволяет настроить шифрование паролей пользователя предпочтительным способом, используя встроенный функционал баз данных. А именно, благодаря внесению изменений в конфигурационный файл используемой базы данных.

3.2. Производит шифрование информации, генерируя шифр случайным образом, основываясь на принципах AES шифрования – симметричный алгоритм блочного шифрования.

3.3. Ограничивает число подключений к базе данных, благодаря чему система добивается улучшение производительности, стабилизирует систему работы, оптимизирует запросы, и уменьшает риски атаки.

3.4. Использует резервное копирования, благодаря чему защищает от потери данных, обеспечивает восстановление данных после сбоев, а также защищает данные от вредоносных программ.

В заключение можно обозначить, что поставленная цель исследования была достигнута и аналогичная методика может быть использована в любых IoT-системах, начиная от повседневного использования, заканчивая эксплуатацию в профессиональных целях.

Список литературы

1. Джейд Картер Интернет вещей (IoT): Разработка, Интеграция и Управление Устройствами // 2024 г. С. 3 – 6.

2. Что такое IoT и что о нем следует знать [Электронный ресурс] // URL: https://habr.com/ru/companies/otus/articles/549550/

3. All Things Considered: An Analysis of IoT Devices on Home Networks // Deepak Kumar, University of Illinois at Urbana-Champaign 2019 г. С. 1169 – 1170.

4. Груздева М.Л., Груздева К.Е. Плюсы и минусы «Интернета вещей» в образовании // Школа будущего № 3 2020 г. С. 292 – 299.

5. Ромашкова О.Н., Каптерев А.И. Анализ угроз и рисков информационной безопасности в вузе // Вестник Московского городского педагогического университета. Серия: Информатика и информатизация образования, 2023 г. С. 37–47.

6. Документация PostgreSQL 16 [Электронный ресурс] // URL: https://www.postgresql.org/docs/current/config-setting.html

7. Основные методы резервного копирования для обеспечения безопасности информации // Волгоградский государственный университет 2023 г. Чумбуридзе Я.А.

8. Шифрование данных // Научно-образовательный журнал для студентов и преподавателей «StudNet», Зацепа А. И. студент бакалавр, Донской государственный технический университет, г. Ростов-на-Дону 2022, С. 346-348.

9. Кузьминых Е.С., Ильина С.П., Маслова М.А. Анализ непробиваемых алгоритмов шифрования // Научный результат. Информационные технологии. - Т.9, №1, 2024. - С. 10-18. DOI: 10.18413/2518-1092-2024-9-1-0-2

УДК 004.421.6+510.635
DOI: 10.25206/978-5-8149-3873-2-2024-1042-1047

Программный комплекс HiConf для проведения конференций в высших учебных заведениях
HiConf software package for holding conferences in higher education institutions

Бедарев И.А[1], Хачатрян А.А.[2], Железняков Р.А.[3], Берберова М.А.[4], Курнасов Е.В.[5]
[1, 2,3]Студент РТУ МИРЭА, Москва, Россия
[4] к.т.н., доцент кафедры РТУ МИРЭА, Москва, Россия, заместитель директора АНО «НИЦФТИ», Нижний Новгород, Россия
[5]к.т.н., доцент кафедры РТУ МИРЭА, Москва, Россия
Bedarev I.A.[1], Khachatryan A.A.[2], Zheleznyakov R.A.[3], Berberova M.A.[4], Kurnasov E.V.[5]
[1, 2,3]Student of RTU MIREA, Moscow, Russia
[4]candidate of technical sciences, Associate Professor of the Department of RTU MIREA, Moscow, Russia, Deputy Director of ANO "SRCIPT", Nizhny Novgorod, Russia
[5]candidate of technical sciences, associate professor of the department of RTU MIREA, Moscow, Russia

Аннотация: В работе представлены теоретические сведения о сложности формирования конференции без автоматизированного сервиса. В работе поставлена задача разработки программного обеспечения призванного упростить создание конференций в высших учебных заведениях. В дальнейшем были рассмотрены инструменты с помощью которых производилась разработка сервиса, был наглядно описан интерфейс работы пользователей сервиса.В заключении подведен итог о получившемся сервисе HiConf.

Ключевые слова: образование, конференция, современный вуз, базы данных, шифрование данных, SQL, PostgreSQL, TypeScript, JavaScript, ReactJS, Pythnon, HTML, CSS.

Abstract: The paper presents theoretical information about the complexity of creating a conference without an automated service. The work sets the task of developing software designed to simplify the creation of conferences in higher educational institutions. Subsequently, the tools with which the service was developed were examined, the interface of the service users was clearly described. In conclusion, the resulting HiConf service was summarized.

Keywords: education, conference, modern university, databases, data encryption, SQL, PostgreSQL, TypeScript, JavaScript, ReactJS, Python, HTML, CSS.

I. Введение

Изучение проблемы проведения конференций в высших учебных заведениях приводит к пониманию того, что программного обеспечения для проведения такого рода мероприятий достаточно мало или оно находится под санкционными ограничениями.

II. Постановка проблемы

Обратная связь в обучении – это ответная реакция обучающегося на поступающую информацию от преподавателя или другого источника знаний. Под поступающей информацией подразумеваются любые действия, которые происходят в процессе образования и воспитания. Проблематика эффективного создания механизмов обратной связи в образовательном процессе продолжает оставаться актуальной для исследований. Способность организовать продуктивную двустороннюю коммуникацию является ключевым навыком педагога, который развивается на всех этапах их профессионального роста [1]. Почти 75% обучающихся утверждают о важности обратной связи для их успеха в учебе. Несмотря на то, что некоторые студенты полагают, что обратная связь ограничивается лишь комментариями к их работе и оценками, даже такая минималистичная форма связи уже представляет собой правильный шаг в направлении улучшения коммуникации и учебного процесса. Следовательно, необходимо обеспечить в сервисе обратную связь между студентом и преподавателем.

Одной из серьезных проблем в современном образовании является нехватка рецензентов в высших учебных заведениях, что негативно сказывается на качестве обучения и оценках студентов. Одной из основных причин этой проблемы является перегрузка преподавателей, которые в среднем проводят 900 академических часов в год [2]. При всём этом преподавателям, которые вынуждены совмещать множество обязанностей, мешает давать

детальные отзывы на работы студентов. Кроме того, наблюдается дефицит квалифицированных специалистов в некоторых областях знаний, что затрудняет привлечение экспертов с необходимыми навыками. Недостаток рецензентов приводит к ухудшению обратной связи для студентов, увеличению времени ожидания результатов проверки и потере мотивации учащихся, что в итоге может повлиять на их успеваемость. Поэтому решение проблемы нехватки рецензентов требует комплексного подхода, включая оптимизацию процессов проверки и внедрение систем стимулирования для преподавателей. Так, сервис может включать в себя не только преподавателей, но и специально отобранных людей, которые могли бы заниматься оценкой работ студентов.

Проблема, связанная с доступом студентов к научным материалам, является серьезным препятствием для их учебного и профессионального роста. Ограниченный доступ к предыдущим практическим работам и их оценкам затрудняет усвоение навыков, необходимых для развития исследовательских навыков и избегания ошибок. Кроме того, студенты оказываются в ситуации, когда им недоступны свежие исследования, методики и результаты, которые могли бы послужить основой для их собственных проектов. Это приводит к недостаточному освоению современных тенденций в научной сфере и ухудшает качество их научных работ. В результате студенты могут испытывать затруднения при применении теоретических знаний на практике и формировании собственной научной позиции. Это отрицательно сказывается на их подготовке к будущей карьере и снижает их конкурентоспособность на рынке труда. Однако, внедрение новых методов обучения и активное применение современных образовательных технологий в учебном процессе стимулируют студентов к самостоятельной работе и помогают улучшить качество обучения [3]. Поэтому сохранение и анализ предыдущих практик с акцентом на ошибки способствует развитию студентов. Наличие такой возможности в сервисе HiConf может положительно повлиять на качество образования.

III. Разработка Программного обеспечения HiConf

Разработка программного обеспечения HiConf велась в трех направлениях. Эти направления имели четкое разделение по конструкции сервиса. Первым направлением являлось использование серверной части приложения – полного механизма внутренней работы, эту часть принято называть бэкэнд (от англ. backend). Для этой части приложения была использована монолитная архитектура, которая позволяет с помощью единой программы обрабатывать запросы в приложении. Для этой части был также использован язык программирования Python и библиотеки надстройки Flask для правильного построения Web-процессов (Рис. 1).

```
pip3 install Flask
```

Рисунок 1 – Установка библиотеки (фреймворка) Flask

Для более правильного понимания работы ниже приведен простой пример программы с использованием фреймворка Flask (Рис. 2).

```
1  from flask import Flask
2  from flask import render_template
3
4  app = Flask(name)
5
6  @app.route('/hello/<name>')
7  def hello(name):
8    return render_template( template_name_or_list: 'hello.html', name=name)
9
10 if name == 'main':
11   app.run(debug=True)
```

Рисунок 2 – Пример использования Фреймворка Flask

С помощью данного фреймворка удалось спроектировать приложение, отвечающее изначально заданным требованиям по работе внутреннего механизма. Это значит что сервис поддерживает многопользовательскую работу, отвечая на запросы одновременно.Также с помощью данного метода были реализованы средства обмена информацией с базой данных и возможности загрузки файлов в локальное хранилище.

Следующим аспектом разработанного приложения, которое необходимо рассмотреть является его визуальная часть, которую обычно называют "фронтенд"(от англ. Frontend). В этой области были применены множество технологий такие как: HTML, CSS, JavaScript, TypeScript, ReactJS. Подробнее будет рассмотрено направление TypeScript и React. TypeScript является самостоятельной надстройкой для использования JavaScript. React выступает в качестве многофункциональной библиотеки для работы всего сервиса.React.js — это мощная

инструментальная библиотека на JavaScript для создания интуитивно понятных и функциональных пользовательских интерфейсов, то есть той части веб-сайтов и мобильных приложений, с которой непосредственно взаимодействует пользователь.

Ключевые особенности React.js заключаются в использовании компонентов и управлении состояниями.

Компоненты представляют собой отдельные блоки кода, отвечающие за дизайн и функциональность отдельных элементов веб-страницы или приложения. Эти компоненты могут быть организованы в иерархические структуры, включая вложенные компоненты. Состояние, в свою очередь, является набором данных, определяющих характеристики элемента, включая его визуальное представление. К примеру, состояние термометра может включать в себя параметры, такие как текущая температура, минимальная и максимальная температуры (Рис. 3).

```
function getGreeting(user) {
  if (user) {
   return <h1>Здравствуй, {formatName(user)}!</h1>;
 }
 return <h1>Здравствуй, незнакомец.</h1>;
}
```

Рисунок 3 – Функция, использующая React

Последней частью разработанного программного обеспечения стала база данных. Ее прообраз был спроектирован почти в самом начале разработки на этапе понимания механики дальнейшей работы сервиса. Для этого был использован мощный инструмент СУБД – Postgres. PostgreSQL — это мощная и надежная система управления базами данных с открытым исходным кодом, которая активно развивается уже более тридцати пяти лет. За это время она заслужила признание пользователей благодаря своей надежности, стабильности и выдающейся производительности.

PostgreSQL является одним из самых популярных решений в области управления данными, и это не удивительно. Ее функциональность и гибкость позволяют решать самые сложные задачи в области хранения и обработки информации. Благодаря активному сообществу разработчиков, PostgreSQL постоянно обновляется и совершенствуется, что делает ее еще более привлекательной для пользователей.

Одним из ключевых преимуществ PostgreSQL является ее открытый исходный код, который позволяет пользователям создавать индивидуальные решения под свои потребности. Это открывает широкие возможности для разработчиков и администраторов баз данных, позволяя им оптимизировать работу с данными и повысить эффективность бизнес-процессов.

Благодаря своей долгой истории развития PostgreSQL стала надежным партнером для многих компаний и организаций по всему миру. Ее стабильность и высокая производительность делают ее идеальным выбором для крупных проектов и критически важных приложений, где надежность и безопасность данных играют решающую роль.

В итоге, PostgreSQL продолжает оставаться одной из лучших систем управления базами данных на современном рынке, обеспечивая пользователям надежность, производительность и функциональность, необходимые для успешной работы в сфере информационных технологий.

В данной работе была использована библиотека для Python под нзванием psycopg2, которая выступает связующим для базы данных и внутренней части сервиса (Рис. 4).

```
import psycopg2

try:
    connect_str = "dbname='yourdbname' user='yourname' host='localhost' " + \
                  "password='yourpassword'"
    conn = psycopg2.connect(connect_str)
    cursor = conn.cursor()
    cursor.execute("""CREATE TABLE postgresqlconn (name char(40));""")
    cursor.execute("""SELECT * from postgresqlconn""")
    conn.commit()
    rows = cursor.fetchall()
    print(rows)
    cursor.close()
    conn.close()
except Exception as e:
    print("Can't connect. Invalid dbname, user or password")
    print(e)
```

Рисунок 4 – Использование Psycopg2 для работы с Postgres

IV. Использование Web приложения HIconf

При входе на сервис пользователя встречает главное меню программы (Рис. 5).

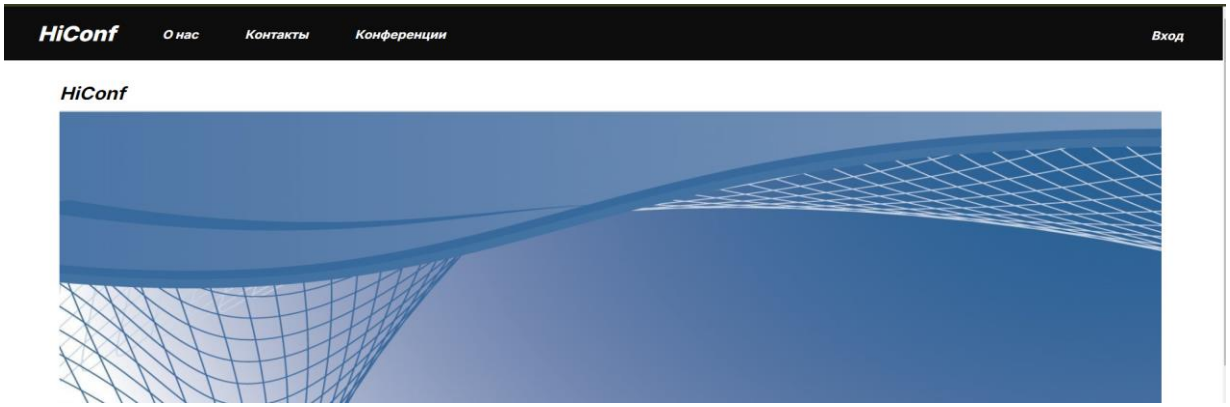


Рисунок 5 – Главное меню программы

Здесь может быть осуществлен выбор необходимого пункта, но прежде сервис потребует авторизации, что приведет пользователя к сервису регистрации и авторизации (Рис. 6).

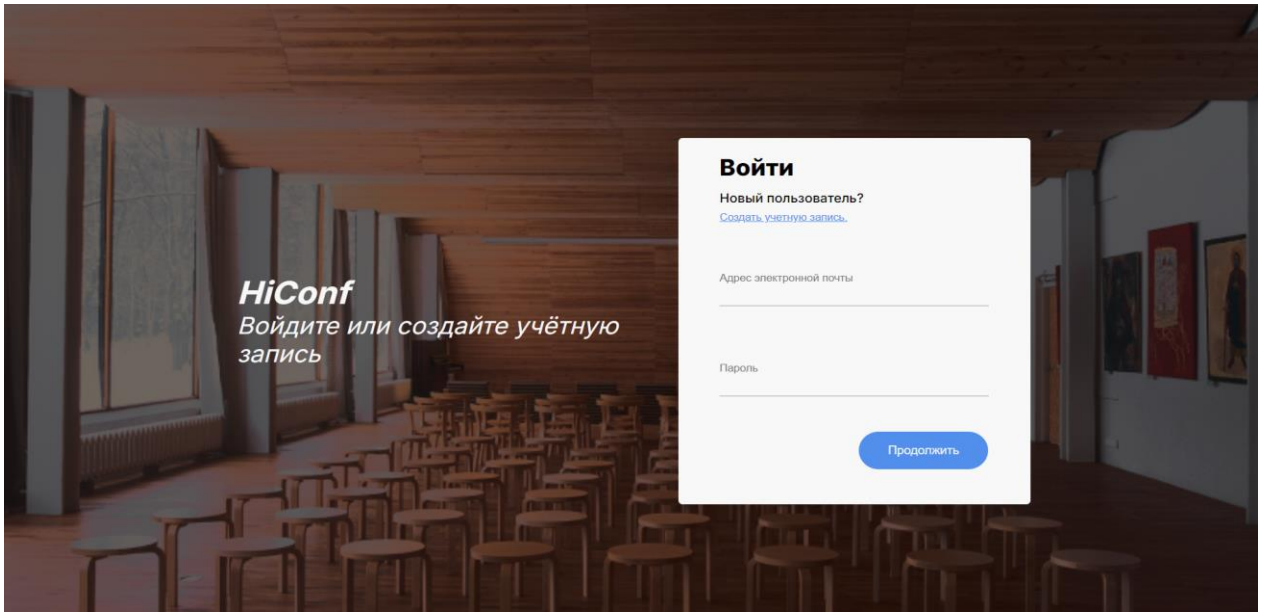


Рисунок 6 – Окно регистрации сервиса HiConf

Далее формат управления сервисом разделяется на несколько путей, администратору разрешено создать конференцию (Рис. 7.1), рецензенту представлено право просматривать поданные на секцию статьи, а также оставлять рецензии (Рис. 7.2), а участнику, в свою очередь – загрузить статью (Рис 7.3).

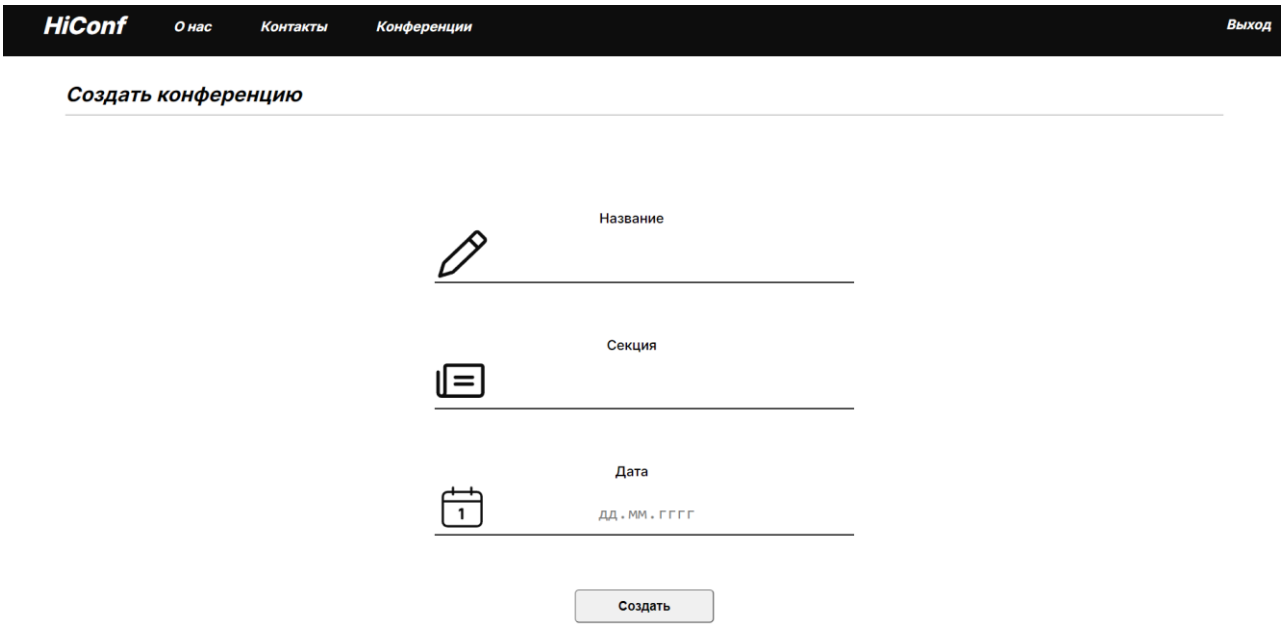


Рисунок 7.1 – Создание конференции администратором

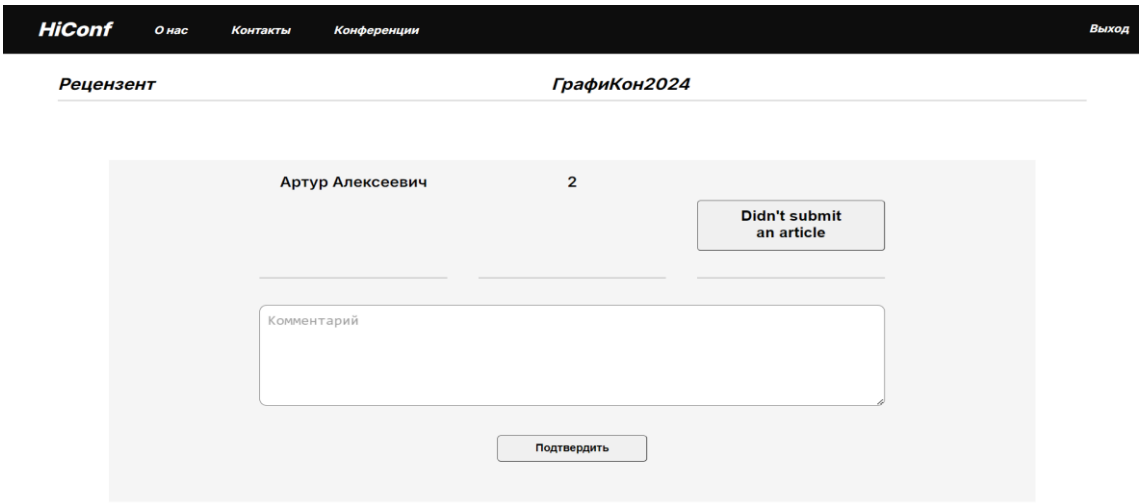


Рисунок 7.2 – Просмотр и рецензирование

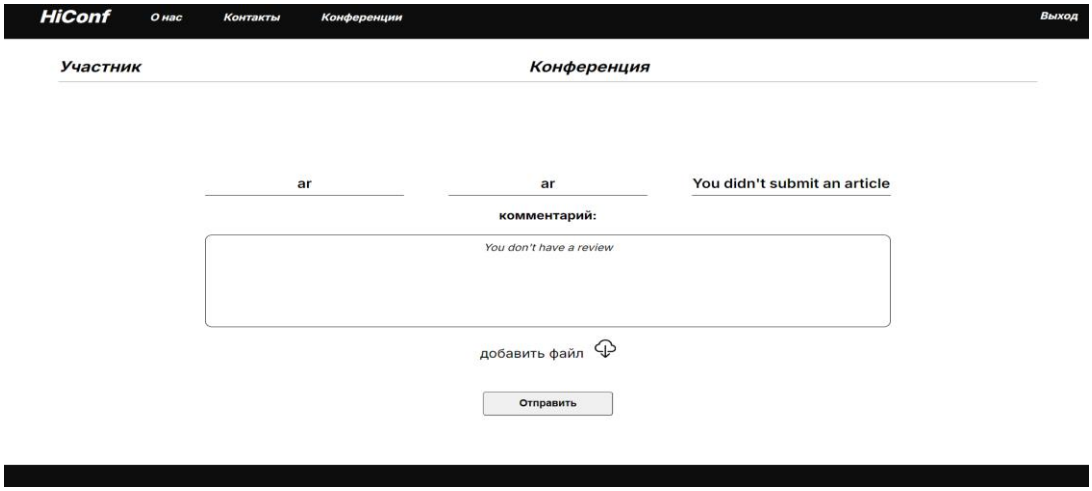


Рисунок 7.3 – Функции участника конференции

V. Заключение

В заключение необходимо отметить пользу данного программного обеспечения, это в первую очередь упрощение создания конференций в высших учебных заведениях. HiConf позволяет преподавателям снизить их

нагрузку и передать функции организации на автоматизированную систему. Данное программное обеспечение продолжает находится в стадии конечной разработки – происходит поиск и устранение неисправностей. В дальнейшем планируется создать крупную инфраструктуру данного приложения, которое бы максимально упростило создание конференций и работу с ними. Предполагается, что сервис поможет сделать конференции максимально популярными не только для студентов разных вузов, но и многих других желающих.

Список литературы:

1. Тарханова И.Ю., Морозов А.С. Обратная связь в дистанционном обучении глазами студентов и преподавателей педагогического университета // Преподаватель XXI век. 2022. № 2. Часть 1. С. 78–90. DOI: 10.31862/2073-9613-2022-2-78-90

2. Эрштейн Л.Б. Чрезмерная нагрузка преподавателей вузов как фактор разрушения высшего образования в России // Вестник ПНИПУ. Социально-экономические науки. 2021. № 2 DOI: 10.15593/2224-9354/2021.2.

3. Михалёва С. Н., Ушникова О. В. Анализ проблем и особенностей организации самостоятельной работы в вузе // Ученые записки ЗабГУ. Серия Педагогические науки. 2014. №5 (58).