Ассоциация сцен в эндоскопических видео

Д.А. Коваленко[1], В.С. Гнатюк[2]
dmitri.a.kovalenko@gmail.com|alfredfid@gmail.com
[1] Computer Science Center, Санкт-Петербург, Россия;
[2]Москва, Россия

Эндоскопическое наблюдение является распространенным методом раннего выявления ряда заболеваний пищевода. Видеопоток с эндоскопа подвергаются ручному анализу, в ходе которого эксперт соотносит изображения участков ткани с их более ранними снимками для установления динамики протекания болезни. Предложен алгоритм ассоциации сцен для автоматизации этого процесса. Подход основан на глобальном анализе изображения и построении системы соседства визуально схожих кадров. Исследование применимости метода проведено на видеоданных реальных эндоскопических операций. Менее специфическая проблема обнаружения границ сцен в видеопотоке рассмотрена; определена ее взаимосвязь с выбранной задачей.

***Ключевые слова*:** *ассоциация сцен, обнаружение границ сцен, анализ эндоскопических видео.*

Scene association in endoscopic videos

D.A. Kovalenko[1], V.S. Gnatyuk[2]
[1]Computer Science Center, Saint-Petersburg, Russia;
[2]Moscow, Russia

Endoscopic surveillance is a common method of monitoring several diseases of the esophagus. Such a surveillance includes the manual matching of images of tissue regions with their earlier observations from previous surveillance operations to control disease dynamics. A scene association algorithm is proposed to automate this process. The algorithm is based on a global approach to image analysis and utilizes construction of neighborhood system for visually resemblant frames. The experimental evaluation of method has been conducted on a set of in vivo endoscopic videos. The article explores the relationship between the current problem and the problem of scene boundary detection.

***Keywords*:** *scene association, scene boundary detection, surgical vision.*

1. Введение

Обнаружение границ сцен в видеопотоке является темой активных исследований в течении последних двух десятилетий. Технология важна для ряда практических приложений: индексации и навигации по базам видеоданных, оптимизации методов сжатия в продвинутых видеокодеках, реализации шага предобработки для высокоуровнего семантического анализа видеопотока. Решения, основанные на обнаружении границ сцен в видеопотоке, применяются хостингами (пример: Youtube) для построения информативной раскадровки – визульной подсказки при перемотке. Кроме того, реализации технологии встраиваются в популярные средства редактирования видео*. Национальный институт стандартов и технологий США в рамках конференции TREC проводит соревнование TRECVID [5], в рамках которого с 2001 по 2007 год одной из задач являлось обнаружение границ сцен в видеопотоке. Настоящая статья рассматривает ее обобщение до задачи ассоциации сцен и ее применение в сфере обработки эндоскопических изображений.

Гастроэнтерологическая эндоскопия является одним из методов предупреждения развития аномальных образований в пищеводе. Пример такого заболевания – синдром Барретта, проявляющийся в метаплазии слизистой и считающийся фактором, повышающим риск аденокарценомы пищевода. Эндоскопический мониторинг состояния аномальных образований включает в себя задачу обработки и сопоставления видеофрагментов участков ткани, снятых в разное время. Задача трудоемка и подвержена ошибкам разметки, вызванным человеческим фактором, поэтому частичная автоматизация процесса является мотивацией исследования.

Определение 1. ***Обнаружение границ сцен в видеопотоке*** *заключается в разделении видеопотока на связные последовательности кадров, объединенных по их семантическим или морфологическим признакам. Для потока кадров $F_0 \dots F_i$, где индексы являются их порядковыми номерами, нужно дать ответ, принадлежит ли кадр F_i той же сцене, что и кадр F_{i-1}.*

Определение 2. ***Ассоциация сцен в видеопотоке*** *заключается в разделении видеопотока на подмножества кадров (не обязательно последовательных), сгруппированных по их семантическим или морфологическим признакам.*

В протоколе тестирования TRECVID, в соответствии с определением 1, идеальная разметка представляет собой последовательность индексов кадров, расположенных на границах сцен. Метрики качества вычисляются при сравнении индексов «пограничных» кадров, предсказанных алгоритмом, с индексами, содержащимися в идеальной разметке.

*FFmpeg Issue Tracker – https://trac.ffmpeg.org/ticket/442

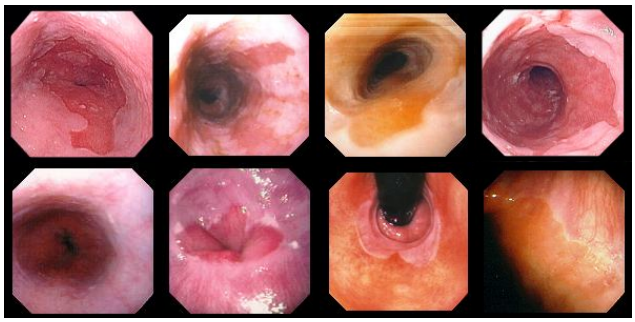

Рис. 1. Пример видеоданных. Первый ряд – метаплазировавшая ткань; Второй ряд – здоровая ткань.

Данный протокол не подходит для проблемы ассоциации сцен, так как принадлежность кадра определенной сцене не следует из множества границ, обнаруженных в видеопотоке. Например, сцена часто состоит из нескольких непоследовательных фрагментов видеопотока, что отражено на рис. 2. Такое разбиение на сцены следует из того, что специалист при проведении исследования перемещает эндоскоп по сложной траектории, многократно возвращаясь к уже знакомым участкам ткани.

Вкладом статьи являются:

— метод ассоциации сцен, основанный на глобальном анализе изображения и построении системы соседства визуально схожих кадров
— протокол тестирования решений для ассоциации сцен

2. Обзор предшествующих результатов

Методы, решающие проблему обнаружения границ сцен в видеопотоке сводятся к отысканию критерия, который обеспечит устойчивую двухклассовую классификацию (см. опр. 1). Обрабатываемые кадры видеопотока подвергаются операции снижения размерности (часто – построение RGB-гистограммы). Эти компактные представления информации о кадре служат входом для критерия, в роли которого может выступать метод Отцу [3] или иной метод вычисления адаптивного порога [6].

В [2] решение принимается с помощью ансамбля таких критериев, оперирующих сегментами изображения, используя инфраструктуру и модель обработки h264 кодировщика. Результаты, полученные в [4] с использованием более сложного признакового вектора и SVM в качестве критерия демонстрируют высокие результаты на TRECVID2006. Однако авторы отмечают, что подход плохо масштабируется на данные, в которых мгновенная граница сцен отсутствует, уступая место плавному переходу между сценами в течении нескольких кадров (alpha-blending, transition, cross-fade, zoom in/out).

В эндоскопических видео переход наделен теми же сложностями, которые можно встретить в задачах общего обнаружения границ сцен:

— изменение яркости и контрастности
— размытие, вызванное движением
— появление бестекстурных участков на большей части изображения

Кроме того, добавляются проблемы, специфические для приложения:

— малая вариативность и слабая уникальность сцен (см. рис. 1)
— нарушение предположения о твердотельности мира: ткани неузнаваемо изменяются в моменты перистальтики
— перекрытие объектива потоками жидкости, инструментами хирурга

В [1] авторы обращаются к каждой из этих проблем, и предлагают решение, основанное на синергии трекера по методу оптического потока, каскадного визуального классификатора по типу Виолы-Джонса и графической вероятностной модели, использующей геометрические свойства ключевых точек для удаления статистических выбросов, пропущенных классификатором. В данной модели создание новой сцены возможно только после срыва трекера.

3. Ассоциация сцен

Настоящая работа предлагает альтернативный подход к проблеме, основанный на глобальном анализе кадра. Отказ от обработки ключевых точек и их отслеживания вызван характером данных: многие участки видео лишены сильно текстурированных и контрастных элементов, оставаясь в то же время узнаваемыми. Подход, успешно применяемый в смежных областях (пример: биометрия [7]) основан именно на построении статистической модели, описывающей общий облик кадра.

Постановка задачи Дадим формальное определение задачи ассоциации сцен в статическом сценарии использования: анализ видеопотока производится после завершения операции. Входом алгоритма является последовательность RGB кадров $F_1, \dots F_n$ эндоскопического видеофайла. Целью выступает объединение в группы кадров, являющихся изображениями одного и того же физического участка ткани пациента. Примеры кадров видеопотока на рис. 1, примеры сцен на фрагменте видеопотока – на рис. 2.

Определение 3. ***Выход алгоритма*** *Последовательность 3-кортежей длины n^2 вида $\{(i, j, c_i^j)\}_{i,j=1}^{n,n}$, где c_i^j - мера уверенности в том, что кадры i и j принадлежат одной сцене, $c_i^j \in [0 \dots 1]$, n – количество кадров в видеопотоке.*

Опишем основные шаги работы предложенного алгоритма:

1. Обучение промежуточных представлений
2. Индексация входного видеопотока
3. Попарное сопоставление кадров входного видеопотока полным перебором (см. опр. 3)

Обучение выполняется единожды и не относится к процедурам статистического вывода, выполняемым

алгоритмом в штатном режиме работы. Оно вынесено в отдельный пункт в силу того, что ряд операций, введенных для обучения, будет переиспользоваться при выводе. Таким образом, предложенный алгоритм является двупроходным оффлайн-решением для ассоциации сцен (пункты 2 и 3).

4. Обучение промежуточных представлений

В предложенном методе в качестве средства снижения размерности входных данных выступает распространенный метод - сумка слов (BoW), где визуальным «словом» является дескриптор ключевой точки на изображении. Тогда промежуточным представлением, требующем обучения (без учителя) является словарь сумки слов – совокупность самых характерных дескрипторов (центроидов), встречающихся на ключевых точках обучающей выборки. Используемый в решении словарь обучен на постороннем наборе данных общего назначения Oxford 102 flowers dataset [8].

Алгоритм 1. Построение словаря сумки слов

Вход: $F_{1\dots m}$
Выход: v (словарь)
1: $D := \{\varnothing\}$
2: **для** $i \in \{1 \dots m\}$
3: плотное вычисление ключевых точек на F_i
4: вычисление дескрипторов полученных точек d_i
5: сохраниение полученных дескрипоторов: $D := D \cup d_i$
6: обучение словаря методом k-средних $v := \text{kmeans\_train}(D, k)$

Процесс обучения описан в алгоритме 1.. Данная процедура осуществляют плотную и равномерную выборку ключевых точек на изображении и вычисление дескрипторов их окрестности [9]. Далее, множество дескрипторов изображений обучающей выборки использовано для их кластеризации по методу k-средних.

Тогда, множество дескрипторов нового входного изображения и словарь из k центроидов можно использовать для постороения гистограммы на носителе $[1 \dots k]$, которая описывает изображение в терминах частоты встречаемости характерных ключевых точек k классов. Переход от изображения к гистограмме использован в шаге 2 алгоритма 2. и в шаге 3. алгоритма 3.. Вычисление этого перехода идентично шагам 2 - 3 алгоритма 1..

5. Индексация входного видеопотока

Первый этап обработки входного видеофайла заключается в его индексации, а именно построении системы соседства визуально схожих кадров. Мерой визуальной схожести выступает расстояние между гистограммами. Последовательно вычислив гистограмы всех изображений входного видеофайла, можно выстроить систему соседства по методу k-ближайших соседей (в его постановке без учителя). Эта процедура описана в алгоритме 2..

Алгоритм 2. Индексация видеофайла

Вход: $F_{1\dots m}$, v
Выход: n (система соседства)
1: $H := \{\varnothing\}$
2: **для** $i \in \{1 \dots m\}$
3: вычисление гистограммы: $h_i := \text{hist}(v, F_i)$
4: сохранение гистограммы: $H := H \cup h_i$
5: обучение knn классификатора $n := \text{knn\_train}(H, k)$

6. Попарное сопоставление кадров видеопотока

Этап работы алгоритма, следующий за индексацией, заключается в полном попарном сравнении кадров и построении списка этих пар с соответствующими мерами (см. опр. 1). Алгоритм 3. описывает вычисления, проводимые для отдельно взятой пары (F_i, F_j), в результате которых получается c_i^j. Процедуру сопоставления можно разделить на два шага:

— определение, является ли пара (F_i, F_j) кадрами-кандидатами
— вычисление c_i^j, меры уверенности в принадлежности кадров i, j общей сцене

Первым шагом вычисляются k-ближайших соседей кадра F_i, используя словарь сумки слов и построенную прежде систему соседства. Процедура фильтрации полученных соседей удаляет тех, для которых расстояние на гистограмме до F_i больше порога. Если F_j находится в отфильтрованном множестве соседей F_i, считаем их кадрами-кандидатами и переходим к вычислению меры уверенности (см. строку 4 алгоритма 3.).

Сопоставление разделено на два шага с целью оптимизации времени работы и повышения точности. Поэтому вычислительноемкая процедура получения c_i^j вызывается только для кадров-кандидатов, а не для любой пары. Этот факт важен ввиду того, что ручная разметка видеоданных показала, что большинство кадров не представляют интереса (см. рис. 2). Фильтрация кадров и многократное сокращение числа обращений к процедуре score понижает количество ложноположительных срабатываний.

Критерии вычисления мер уверенности В работе предложен нейросетевой критерий вычисления мер уверенности в принадлежности пары кадров эндоскопического видео одному участку ткани. Для сравнения, в качестве базового алгоритма, применен алгоритм на основе проективного преобразования плоскости [11]. Для базового алгоритма, далее именуемого гомографическим, мерой является доля ключевых точек, согласующихся с проектив-

Алгоритм 3. Построение меры уверенности в принадлежности кадров сцене

Вход: $F_i,\ F_j,\ v,\ n$
Выход: c;
1: вычисление гистограмы
$h_i := \mathrm{hist}(v, F_i)$
2: вычисление k-ближайших соседей для кадра F_i
$\mathrm{neighborhood}_i := \mathrm{knn}(n, h_i)$
3: **для** $m \in \mathrm{neighborhood}_i$: $\mathrm{dst}(\mathrm{h_m}, \mathrm{h_i}) > \mathrm{t_h}$
$\mathrm{neighborhood}_i := \mathrm{neighborhood}_i \setminus \{m\}$
4: **если** $j \in \mathrm{neighborhood}_i$ **то**
выход $c := \mathrm{score}(F_i, F_j)$
5: **иначе**
выход $c := 0$

ным преобразованием, вычисленным в ходе процедуры RANSAC.

Нейросетевой критерий использует архитектуру сети Крыжевского [14]. Входом сети является разница между кадрами F_i и F_j. Выходной слой которой производит двухклассовую классификацию, где $0 : (F_i, F_j) \notin S$, $1 : (F_i, F_j) \in S$. Тогда 1-ая компонента вектора весов может быть использована как мера уверенности c_i^j. По причине ограниченности доступного набора данных, обучение сети с нуля для решения задачи не представлялось возможным. Был использован распространный подход дообучения сети, описанный в[†]. Процесс обучения сошелся менее чем за 20000 эпох на обучающей выборке из 15000 изображений. Качественный и количественный анализ вышеописанных критериев приведен в последующих разделах.

7. Эксперименты и результаты

Данные для проведения экспериментов предоставлены Имперским колледжем Лондона [1] и являются записями реальных эндоскопических операций. Набор данных содержит 6 видео, от 500 до 1400 кадров в каждом. Разметка истинной принадлежности кадров сценам была выполнена не экспертом. Следующие параграфы описывают проведенные эксперименты и их отличия от протокола TRACVID.

Протокол тестирования Структура, являющаяся выходом алгоритма (см. опр. 3), выступает в роли входа для процедуры тестирования. Каждая задача информационного поиска зависит от компромисса между требованиями к точности результата и к полноте выдачи. Введем эти определения.

Определение 4. ***Точность*** *– доля верно ассоциированных кадров среди тех, которым алгоритм присвоил сцену. Формально:* $P = \frac{TP}{TP+FP}$*, где* TP *- число верно ассоциированных кадров,* FP *- число кадров, отнесенных к сцене, которой они не принадлежат.*

[†]Fine-tuning CaffeNet for Style Recognition on Flickr Style Data – `http://caffe.berkeleyvision.org/gathered/examples/finetune_flickr_style.html`

Определение 5. ***Полнота*** *– доля верно ассоциированных кадров среди тех, которым присвоена сцена на истинной разметке. Формально:* $R = \frac{TP}{TP+FN}$*, где* TP *- число верно ассоциированных кадров,* FN *- число кадров, имеющих присвоенную им сцену в истинной разметке, но отброшенных алгоритмом.*

Для выбранной проблемы пользователи заинтересованы в максимальных значениях полноты при фиксированной точности, близкой к 100% [1]. Однако, было отмечено, что разметка проведена не экспертами, поэтому результату со 100% показателем точности, вычисленному на неверной разметке, доверять нельзя. В силу этого оценка алгоритма производится путем построения кривых точности/полноты.

Суть предложенного протокола тестирования заключается в том, что ни в ручной разметке, ни в выходных данных алгоритма сцены не определены явно. При этом, тест позволяет сказать для каждого кадра видеопотока был ли он отнесен алгоритмом к той же сцене, что и в ручной разметке. Такая постановка желательна в силу:

— упрощения работы по разметке видеопотока
— возможности строить более простые модели

Процедура тестирования обрабатывает две выходные последовательности (см. опр. 3) для получения кривых точности/полноты. Первая выдана алгоритмом, вторая – идеальная ручная разметка. При этом, ручная разметка будет содержать только меры уверенности $c_i^j = 1$. Далее описан метод вывода множества сцен из выходной последовательности. Ей в соответствие можно поставить следующий объект: ненаправленный граф с индесами кадров в вершинах и мерами уверенности c_i^j на ребрах. В графе удаляются все ребра, меньшие текущего порога на кривой. Тогда любая пара кадров, имеющая путь в графе, считается принадлежащей одной сцене. Множество сцен видеопотока – список непересекающихся множеств индексов кадров, получаемый из графа по [12].

Так восстановлены два множества сцен, идеальных и выведенных алгоритмом. Счетчик положительных срабатываний инкрементируется покадрово, если пересечение между сценами, к которым кадр отнесен в алгоритмической и ручной разметке, больше порога. Подсчет ложноположительных и ложноотрицательных срабатываний производится аналогично.

Исследователи, решающие задачи ассоциации, нуждаются в средствах качественной оценки результата на данном видеофайле. Предложено средство для сверки алгоритмически восстановленных и истинных сцен, изображающее на графике последовательные ассоциированные кадры в виде столбцов соответсвующего цвета. Данная визуализация приведена на рис. 2. Подход вдохновлен решением [13] другой проблемы сопоставления визуальных данных.

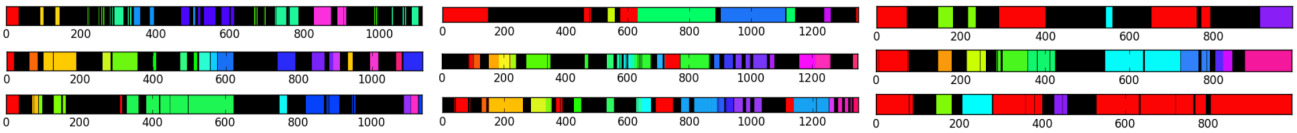

Рис. 2. Сравнение автоматической ассоциации сцен с ручной разметкой. Столбцы: результаты для 2, 4, 10 видеофайлов из набора данных; Строки: результаты разметки: идеальная, гомографическим методом, нейросетевым методом. Цветные блоки – обнаруженные сцены, черные блоки –кадры, не отнесенные ни к одной из сцен.

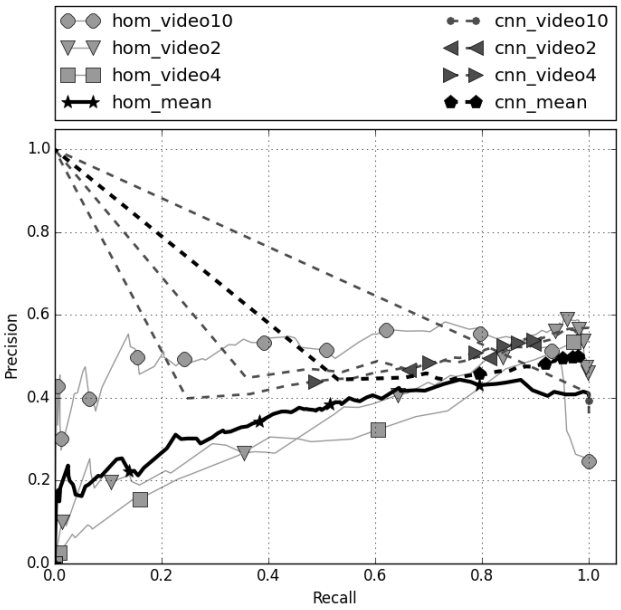

Рис. 3. Кривые точности и полноты. Кривые с префиксами *hom* и *cnn* – результаты для гомографического и нейросетевого критериев на соответствующих видеофайлах. Кривые с суффиксом *mean* – усредненный по набору тестируемых данных результат.

Эксперименты Ввиду недоступности авторских реализаций алгоритма [1], предложенный метод протестирован в сравнении с решением, обеспечивающим базовый результат. Их показатели производительности, усредненные по набору данных, и точечные для трех видеофайлов приведены на рис. 3.

Траектории изображают изменение параметров точности и полноты в зависимости от выбранного значения порога, применяемого к c_i^j. На рис. 3 значение порога изменяется слева направо от 1 до 0. Закономерно, в правой части графика прямые сходятся к примерно одним и тем же значениям для разных критериев, так как классификация перестает работать (порог стремится к нулю) и решения об ассоциации начинают в большей степени зависеть от предыдущей фазы алгоритма – выбора кадров-кандидатов. Как было упомянуто ранее, с практической точки зрения более интересны результаты из области графика, соответствующей $[0 \ldots 0.5]$ по оси абцисс, предпочитающие большую точность при меньшей полноте. Можно заметить, что в этой области преимущество нейросетевого критерия над гомографическим видно явно, хотя первый и превосходит второй на любом значении порога. Результаты замеров точности и полноты коррелируют с качественной оценкой, приведенной на рис. 2. На рис. 2 можно пронаблюдать, что нейросетевой критерий чаще чем базовый метод отнесит кадр к верной сцене. Это особенно характерно для случаев, когда эндоскоп вернулся к исходному участку ткани спустя некоторое время (кадры отнесены к той же сцене, которую пронаблюдали несколькими сотнями кадров ранее). Недостатком базового метода является недостаточно выразительный критерий, который допускает больше ложноположительных срабатываний при больших значениях порога. Это приводит к падению точности одновременно с падением полноты на интервале $[0 \ldots 0.4]$. Средняя точность ассоциации при средней полноте 90% равна 57% для предложенного алгоритма. В [1] докладывают большую точность, однако, важно отметить некорректность прямого сравнения этих результатов в силу использования разных протоколов тестирования. Решение [1] использует концепцию времени и априорные предположения о длительности и характере смены сцен, тем самым снижая размерность информационного поиска и делая разбиения на сцены, подобные изображенным на рис. 2 невозможными.

8. Выводы

Предложенный метод ассоциации сцен для эндоскопических видео позволяет выделить ключевые сцены с помощью построения системы соседства похожих кадров и глобального анализа похожих кадров путем применения нейросетевого статистического критерия. Работа формализует и описывает протокол тестирования алгоритма. Приведены результаты сравнения производительности базового и предложенного алгоритмов на наборе данных реальных эндоскопических операций.

Программная реализация алгоритмов ассоциации сцен и средств тестирования и визуализации открыта для использования исследователями[‡].

[‡] `https://github.com/wf34/scene_ass`

9. Благодарности

Работа выполнена при финансовой поддержке Фонда Содействия Инновациям, грант 10092ГУ2/2015.

10. Литература

[1] *Ye M., Johns E., Giannarou S., Yang G.* Online Scene Association for Endoscopic Navigation //International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI, 2014

[2] *Dimou A., Nemethova O., Rupp M.* Scene change detection for H. 264 using dynamic threshold techniques //Proceedings of 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Service 2005

[3] *Sahoo P. K., Soltani S., Wong A. K.C., Chen Y.C.* A survey of thresholding techniques //Computer vision, graphics, and image processing. – 1988. – Т. 41. – №. 2. – С. 233-260.

[4] *Chávez G. C., Cord M., Philipp-Foliguet S., Precioso F., Araujo A.* Robust scene cut detection by supervised learning //Signal Processing Conference, 2006 14th European. – IEEE, 2006. – С. 1-5.

[5] *Smeaton A. F., Over P., Kraaij W.* Evaluation campaigns and TRECVid //MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval //Proceedings of the 8th ACM international workshop on Multimedia information retrieval. – ACM, 2006. – С. 321-330.

[6] *Kovalenko D.A., Potapev I.A.* Increasing Performance of the Scene Boundary Detection System with use of OpenCL // XIX International Conference Modern Technique and Technologies, 2013 – Vol.2 – C.426-428. `portal.tpu.ru/files/conferences/ctt/proceedings/ctt-2013-2-Tom.pdf`.

[7] *Bharadwaj S., Bhatt H., Vatsa M., Singh R.* Periocular biometrics: When iris recognition fails //Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on. – IEEE, 2010. – С. 1-6.

[8] *Nilsback M. E., Zisserman A.* Automated flower classification over a large number of classes //Computer Vision, Graphics Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on. – IEEE, 2008. – С. 722-729.

[9] *Tola E., Lepetit V., Fua P.* Daisy: An efficient dense descriptor applied to wide-baseline stereo //IEEE transactions on pattern analysis and machine intelligence. – 2010. – Т. 32. – №. 5. – С. 815-830.

[10] *Lowe D. G.* Object recognition from local scale-invariant features //Computer vision, 1999. The proceedings of the seventh IEEE international conference on. – Ieee, 1999. – Т. 2. – С. 1150-1157.

[11] *Hartley R., Zisserman A.* Multiple view geometry in computer vision – Cambridge university press, 2003. – С. 325-340

[12] *Tarjan R.* Depth-first search and linear graph algorithms //SIAM journal on computing. – 1972. – Т. 1. – №. 2. – С. 146-160.

[13] *Zhu Y. et al.* Aligning books and movies: Towards story-like visual explanations by watching movies and reading books //Proceedings of the IEEE International Conference on Computer Vision. – 2015. – С. 19-27.

[14] *Krizhevsky A., Sutskever I., Hinton G. E.* Imagenet classification with deep convolutional neural networks //Advances in neural information processing systems. – 2012. – С. 1097-1105.