# Рекуррентная нейронная сеть для распознавания жестов русского языка с учетом языкового диалекта Сибирского региона

В.А. Жуковская[1], А.В. Пятаева[1]

[1]*Сибирский федеральный университет, Академика Киренского, 26Б, г. Красноярск, 660074, Российская Федерация*

**Аннотация**

Распознавание жестов является важной задачей, в частности для общения глухого и слабослышащего населения с людьми, не владеющими жестовым языком. Русский жестовый язык изучен слабо, русский жестовый язык Сибирского региона обладает значительными отличиями от других в рамках русской языковой группы. Для русского жестового языка отсутствует общепризнанный набор данных. В работе представлен алгоритм распознавания жестов по видеоданным. В основе алгоритма распознавания жестов лежит выделение ключевых признаков рук и позы человека. Классификация жестов выполнена с помощью рекуррентной нейронной сети LSTM. Для обучения и тестирования результатов распознавания жестов самостоятельно разработан набор данных, состоящий из 10 жестовых слов. Отбор слов для набора данных выполнен среди наиболее популярных слов русского языка, а также с учетом максимальной разницы в произношении жестов языкового диалекта Сибирского региона. Реализация алгоритма распознавания жестов выполнена с использованием технологий проектирования и глубокого обучения нейронных сетей Keras, библиотеки компьютерного зрения OpenCV, фреймворка машинного обучения MediaPipe, а также посредством других вспомогательных библиотек. Экспериментальные исследования, проведенные на 300 видеопоследовательностях, подтверждают эффективность предложенного алгоритма.

**Ключевые слова**

Распознавание жестов; рекуррентная нейронная сеть; жесты Сибирского региона; глубокое обучение.

# Recurrent Neural Network for Recognition of Gestures of the Russian Language, Taking into Account the Language Dialect of the Siberian Region

V.A. Zhukovskaya [1], A.V. Pyataeva [1]

[1] *Siberian Federal University, 26B, Academician Kirensky, Krasnoyarsk, 660074, Russian Federation*

**Abstract**

Sign recognition is an important task, in particular for the communication of the deaf and hard of hearing population with people who do not know sign language. Russian sign language is poorly studied, Russian sign language of the Siberian region has significant differences from others within the Russian language group. There is no generally accepted data set for Russian Sign Language. The paper presents a gesture recognition algorithm based on video data. The gesture recognition algorithm is based on the identification of key features of the hands and posture of a person. Gestures were classified using the LSTM recurrent neural network. To train and test the results of

*ГрафиКон 2022: 32-я Международная конференция по компьютерной графике и машинному зрению, 19-22 сентября 2022 г., Рязанский государственный радиотехнический университет им. В.Ф. Уткина, Рязань, Россия*
EMAIL: zhukovskaya.vlada00@mail.com (В.А. Жуковская); anna4u@list.ru (А.В. Пятаева)
ORCID: 0000-0002-6113-3128 (В.А. Жуковская); 0000-0002-0140-263X (А.В. Пятаева)
© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

gesture recognition, we independently developed a data set consisting of 10 sign words. The selection of words for the data set was made among the most popular words of the Russian language, as well as taking into account the maximum difference in the pronunciation of gestures of the language dialect of the Siberian region. The implementation of the gesture recognition algorithm was carried out using Keras neural network design and deep learning technologies, the OpenCV computer vision library, the MediaPipe machine learning framework, and other auxiliary libraries. Experimental studies conducted on 300 video sequences confirm the effectiveness of the proposed algorithm.

**Keywords**
Gesture recognition; recurrent neural network; gestures of the Siberian region; deep learning.

## 1. Введение

По данным Всемирной организации здравоохранения по состоянию на 2020 год более 5% всех людей в мире (360 млн человек) страдают от серьёзных нарушений слуха, ожидается, что к 2050 году их число вырастет до 900 миллионов [1]. Обеспечение коммуникативного взаимодействия таких людей в обществе является важнейшей социальной задачей, обеспечивающей устранение мотивационных, психологических, идеологических или межкультурных барьеров. Языки жестов - это языки, используемые для общения людей с нарушениями слуха. Это естественные языки, в которых общение достигается не за счет артикуляции звуков, а за счет жестов, изменения положения тела и мимики для создания знаков. Плоскость выражения таких языков базируется на жестово-мимической основе. Функциональные и коммуникативные возможности этих языков не уступают разговорным языкам. Язык жестов является основным языком глухого и слабослышащего населения, их основным средством общения и он так же разнообразен, как и речь. Жестовые языки даже в рамках одной лингвистической группы обладают существенной вариативностью, поэтому разработка системы преобразования жестового языка в речь для улучшения взаимодействия людей с нарушениями слуха и без таких нарушений с учетом лингвистических особенностей жестов является актуальной задачей. Трудность коммуникации может также возникать при общении глухого человека со слышащим, когда слышащий человек не владеет жестовым языком. У каждой страны существует свой язык жестов. Русский жестовый язык изучен мало, использование жестов в разных регионах различно. Кроме того, наблюдается нехватка общедоступных наборов данных русского жестового языка, для жестового языка Сибирского региона такой набор данных отсутствует.

С развитием компьютерного зрения многие исследователи придумали новые подходы, чтобы помочь людям с ограниченными физическими возможностями. Распознавание жестов в видеорядах является сложной задачей компьютерного зрения из-за схожести визуального содержания, изменения угла обзора на одни и те же действия, масштаба и различных условий освещения. Для полноценного распознавания одного слова из языка жестов требуется классифицировать эмоции на лице и позиции всего тела, выявлять скорость жестов и учитывать много других факторов. Системы машинного перевода на русском жестовом языке играют важную роль в решении проблемы обеспечения коммуникации для людей с проблемами со слухом. Это необходимо из-за недостаточного количества персонала в области жестового перевода, а также из-за не всегда желательной медиации в общении глухих и слышащих граждан, например, в вопросах медицины и личных отношений.

## 2. Анализ существующих технологий

Под жестами будем понимать различные телодвижения, являющиеся невербальным способом передачи информации [2]. Жесты могут быть разделены на статические и динамические. Движение пальцев, рук, головы, плеч, мимика лица - все перечисленные действия являются динамическими жестами. Распознавание динамических жестов является сложной

задачей, потому что каждый человек выполняет один и тот же жест по-разному. Несмотря на существенные различия в произношении жестов для разных языковых групп и регионов внутри одной группы, технологии и подходы для распознавания жестов едины: как правило, для распознавания жестов производится пространственно-временной анализ видеопоследовательностей для выделения признаков жестов с применением технологий глубокого обучения.

Так, для распознавания жестов американского языка в работе [3] рука выделяется как связный компонент на основе цветовой сегментации и оконтуривания, затем на выделенной области выполняется построение дескриптора жеста. Схожесть жестов оценивается расстоянием Евклида между векторами дескрипторов жестов, заявлено качество распознавания около 90%. В работе [4] для распознавания статических жестов руки азбуки глухонемых Японии в качестве входного устройства применяется перчатка со встроенными сенсорами, которая возвращает характеристики показанного жеста (10 точек и 3 угла). Характеристики жеста нормализуются и передаются на вход нейронной сети. На выходе сеть выдает результат в виде массива из 42 элементов, которые соответствуют 42 жестам руки. Профессор машинного обучения Ричард Боуден [5] предложил в работе по распознаванию конфигурации руки для множества статических жестов из ручной азбуки ASL алгоритм случайных лесов. В работе [6] для распознавания жестов применено устройство ввода с сенсором Kinect, которое на основе дальностного и цветного изображение человека принимает решение о том или ином воспроизводимом жесте. В работе [7] на основе платформ OpenNI и Nite происходит поиск рук и сегментируется как связанный объект в дальностном изображении. После применения фильтра Габора изображения преобразуются в матрицы – дескрипторы показанного жеста. Разработанная система позволяет в реальном времени распознавать 24 жеста ручной азбуки ASL и среднее качество распознавания предложенным алгоритмом при тестировании системы на жестах пяти разных людей 75%.

Кроме того, распознавание жестов находит применение при осуществлении человеко-машинного взаимодействия [8-12], а также в биомедицине, например, для управления протезами конечностей на основе нервных импульсов предплечья [13]. Дополнительно, для повышения точности работы алгоритмов распознавания жестов применяют специализированное оборудование, такое как перчатки для распознавания жестов [14], другие носимые на руках датчики (например, акселерометр, гироскоп и магнитометр [15] или приборы, регистрирующие биосигналы тела человека.

Технологии глубоких нейронных сетей в группе работ по распознаванию жестов выступают в роли строительных блоков для решения различных проблем. На вход глубокие нейронные сети принимают пространственно-временные характеристики жеста, а на выходе выдают распознанный жест. Чаще всего используемыми моделями нейронных сетей в задачах распознавания жестов являются сети прямого распространения и рекуррентные нейронные сети.

## 3. Языковой диалект Сибирского региона

Жестовые языки обладают существенными различиями в рамках одной языковой группы. Так, в рамках Сибирского региона даже для близкорасположенных городов Красноярск и Абакан, расстояние между которыми по автомобильной дороге составляет всего 430 км и шесть часов пути жесты для обозначения одного слова разные. На рисунке 1 можно увидеть, что информанты демонстрируют жест «близко» по-разному. Такие различия требуется учитывать при разработке алгоритма распознавания жестов по видеоданным для того, чтобы обеспечить понимание людей друг другом.

Жест «близко» для Московской области получен с Youtube-канала «Город жестов» [16], жесты для городов Красноярск и Абакан получены при проведении исследований жестовых языков учеными института Филологии и языковой коммуникации Сибирского федерального университета [17].

**Рисунок 1 –** Вариативность в показе жеста «близко»: (а) – жест для Московской области; (б) – жест для Красноярска; (в) – жест для Абакана.

## 4. Алгоритм распознавания жестов

Алгоритм распознавания жестов содержат два основных этапа. Сначала выполняется обнаружение жестикулирующего человека и выделение ключевых точек для анализа жестов на лице, кистях и пальцев рук. Затем производится классификация жестов.

Оценка позы человека по видео играет важную роль в распознавании языка жестов. MediaPipe Pose – решение машинного обучения для отслеживания позы тела с высокой точностью, выводящее 33 3D-ориентира и маску сегментации фона на всем теле из видеокадров RGB (рисунок 2).

Чтобы классифицировать действие, нам сначала нужно найти различные части тела, в частности руки в каждом кадре, а затем проанализировать их движение с течением времени. Первый шаг достигается с помощью Mediapipe Hands [18], который выводит характерные точки ладони (21 ключевых точек) после наблюдения одного кадра в видео.

Кроссплатформенный фреймворк Mediapipe Hands для отслеживания рук и пальцев, предложенный Google, используется в качестве инструмента для извлечения признаков. Это решение доступно для работы как с онлайн видеопотоком с веб-камеры, так и с фото- и видеофайлами. В случае Python MediaPipe доступен в виде пакета модуля Python. Модель может быть использована для обнаружения ладони и расчета 21 ориентира для каждой обнаруженной руки, где каждый ориентир соответствует определенной точке на ладони и представляет собой три координаты - $x$, $y$ и $z$ (рисунок 3).

Для классификации жестов использована рекуррентная нейронная сеть LSTM. Среди нейронных сетей глубокого обучения высокое качество работы алгоритмов распознавания жестов обеспечивает архитектура LSTM - разновидность особой архитектуры рекуррентных нейронных сетей, способная к обучению на долговременную зависимость.

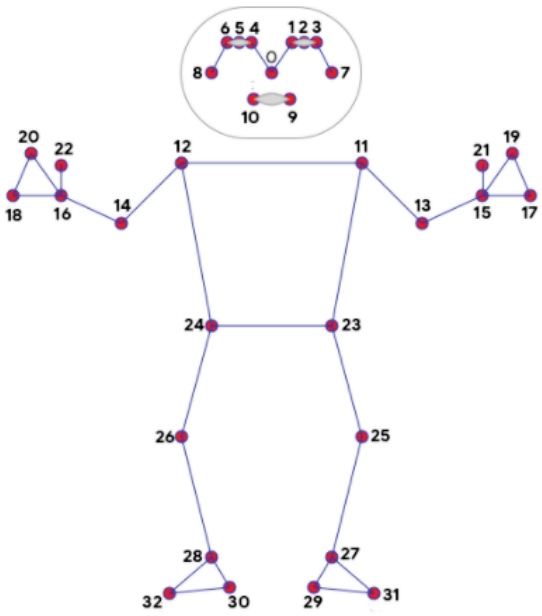


**Рисунок 2 –** MediaPipe Pose

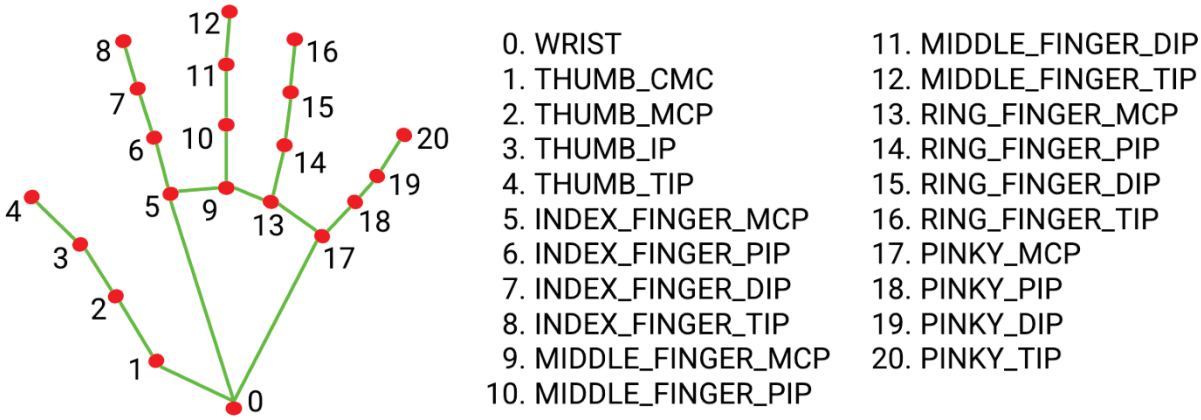


**Рисунок 3 –** MediaPipe Hands

Под сетью LSTM [19] подразумевается нейронная сеть с рекуррентным слоем LSTM, где ячейка LSTM имеет задачу возвратить новое скрытое состояние $h_t$, при этом основываясь на прошлом скрытом слое $h_{t-1}$ и вектором представлении текущего значения $x_t$. В слое LSTM есть одна ячейка, определяемая количеством содержащихся в ней узлов. Ячейка LSTM поддерживает состояние *Ct*, которое можно рассматривать как внутреннее ее убеждение о текущем состоянии последовательности. Это состояние отличается от скрытого состояния *ht*, которое в конечном итоге возвращается ячейкой после последнего временного шага. Состояние *Ct* имеет ту же длину, что и скрытое состояние (равную количеству узлов в ячейке). На рисунке 4 показано как происходит обновление скрытого состояния отдельной ячейки.

Она была специально разработана для задач, где необходимо распознавание продолжающихся во времени действий. Именно к таким действиям относятся динамические жесты. Мы используем LSTM для классификации действий по последовательности обнаружений ключевых точек из видео.

Таким образом, модуль распознавания представляет собой нейронную сеть для многоклассовой классификации. Каждому классу соответствует один жест русского жестового языка. Сеть состоит из шести слоев — трех слоев двунаправленного LSTM и три полносвязных слоя (рисунок 5).

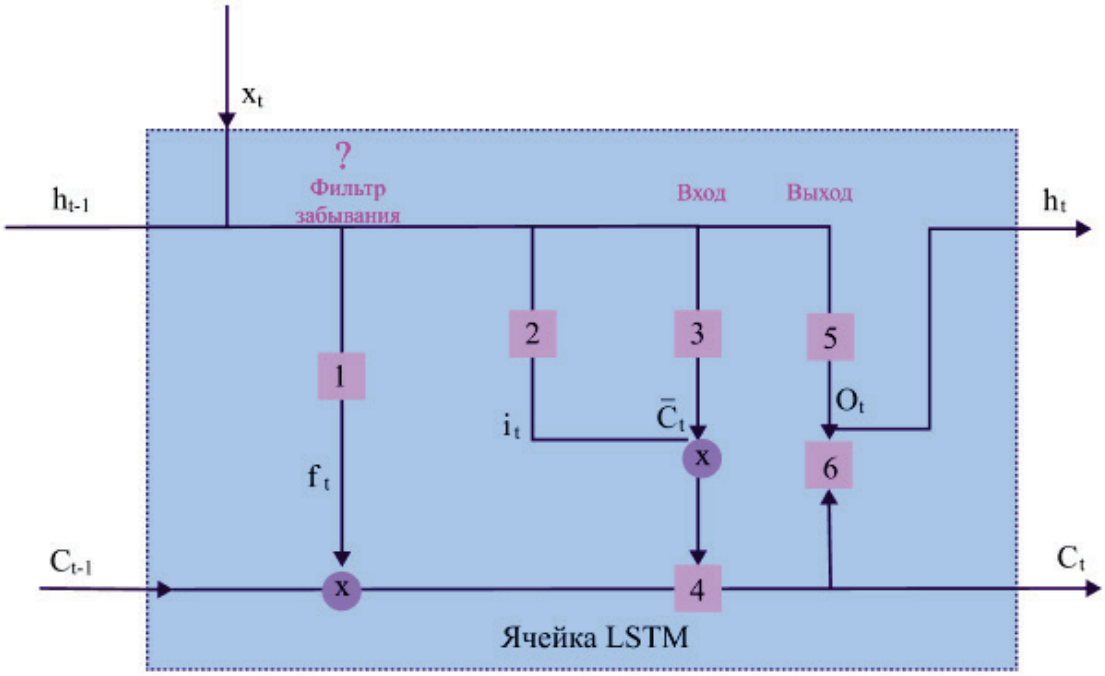


**Рисунок 4** – Архитектура ячейки LSTM

Благодаря рекурсивным ссылкам слои LSTM могут эффективно классифицировать данные, продолжающимися во времени, что необходимо для классификации динамических жестов. Первый, второй и третий слои содержат 65, 128 и 64 нейронов соответственно, а полносвязные слои содержат 64 и 32 нейронов и имеют функцию активации relu. Последний полносвязный слой активации с функцией softmax содержит 10 нейронов, соответствующее количеству распознаваемых жестов.

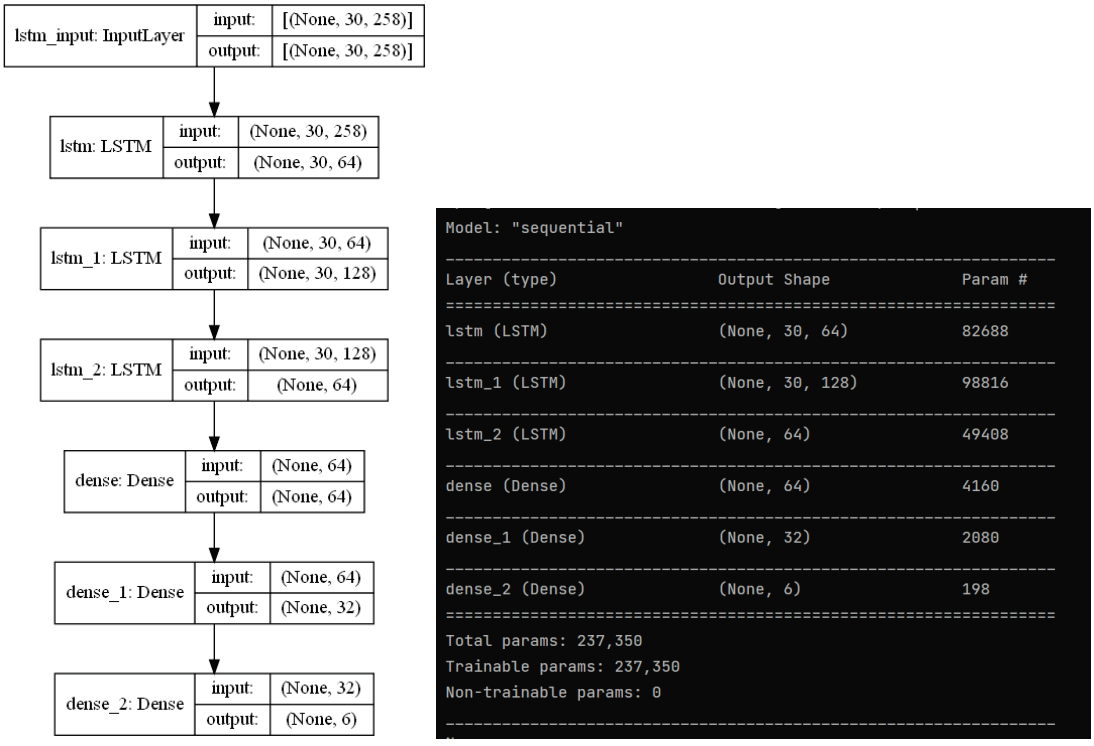


**Рисунок 5** – Архитектура модели LSTM

Таким образом, ключевые точки из последовательности кадров отправляются в нейронную сеть LSTM для классификации жестов (рисунок 6).

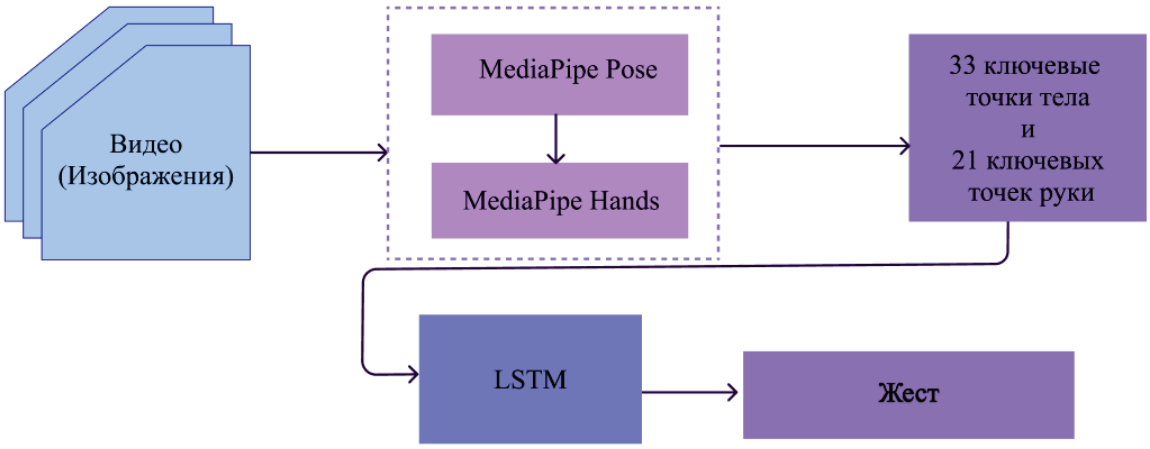


**Рисунок 6** – Схема работы программы

Реализация алгоритма распознавания жестов выполнена с использованием технологий проектирования и глубокого обучения нейронных сетей Keras [20], библиотеки компьютерного зрения OpenCV [21], рекуррентной нейронной сети на базе LSTM, фреймворка машинного обучения MediaPipe [22] а также посредством других вспомогательных библиотек.

## 5. Экспериментальные исследования

Для проведения экспериментальных исследований с целью проверки качества работы алгоритма распознавания жестов выбраны несколько наиболее популярных слов жестового русского языка, такие как: до свидания, здравствуйте, еда, мой, человек, боль, время, близко, гордость, дом. Тестовые слова для проверки качества работы алгоритма отбирались с учетом максимальной разницы в произношении одного и того же жеста. Для каждого жеста было самостоятельно записано по 30 видео, длиной в 30 кадров из которых мы получаем ориентиры точек для дальнейшей классификации воспроизводимых жестов. Общий объем сформированного набора данных составил 300 видеороликов или 9000 кадров. Объем обучающей выборки составил 80% от общего числа видеороликов, контрольной выборки – 20%. В таблице 1 показаны результаты экспериментальных исследований для распознавания, выбранных среди наиболее популярных, жестов русского языка с учетом диалекта Сибирского региона по метрике Accuracy.

**Таблица 1** – Таблица точности распознавания жестов

| Жест | Accuracy | Жест | Accuracy |
|---|---|---|---|
| до свидания | 0.93 | боль | 0.95 |
| здравствуйте | 0.90 | время | 0.95 |
| еда | 0.91 | близко | 0.96 |
| мой | 0.94 | гордость | 0.91 |
| человек | 0.92 | дом | 0.94 |

Метрика Accuracy предназначена для задач классификации и рассчитывается классическим способом согласно формуле:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

где $TP$ – True Positive, число истинно положительных кадров, то есть кадров, в которых жест распознан верно; $TN$ – True Negative, число истинно негативных кадров; $FP$ – False Positive, число ложно положительных кадров; $FN$ – False Negative, число ложно негативных кадров.

Как показывают результаты экспериментальных исследований, точность распознавания наиболее популярных жестов русского языка по видеоданным является достаточно высокой. При наличии сложного динамического фона или при нахождении в кадре нескольких жестикулирующих персон, благодаря работе функции MediaPipe Pose, обнаруживающей и далее отслеживающей движения только одной обнаруженной персоны качество работы алгоритма не снижается. В другом сложном случае, например, если кисть руки при воспроизведении жеста пересекается с уровнем лица, что является распространённым при общении глухонемых, алгоритм позволяет избежать наложения лица и кисти. Так, в жесте «время» рука находится на уровне лица и этот жест распознается достоверно.

Из графиков точности модели и кривых потерь модели на рисунке 7 можно увидеть, что на 100 итерации, нейронная сеть достигает максимальной точности при минимальных потерях, это говорит о том, что процесс обучения на этой итерации стоит остановить для избежания проблем с переобучением нейронной сети.

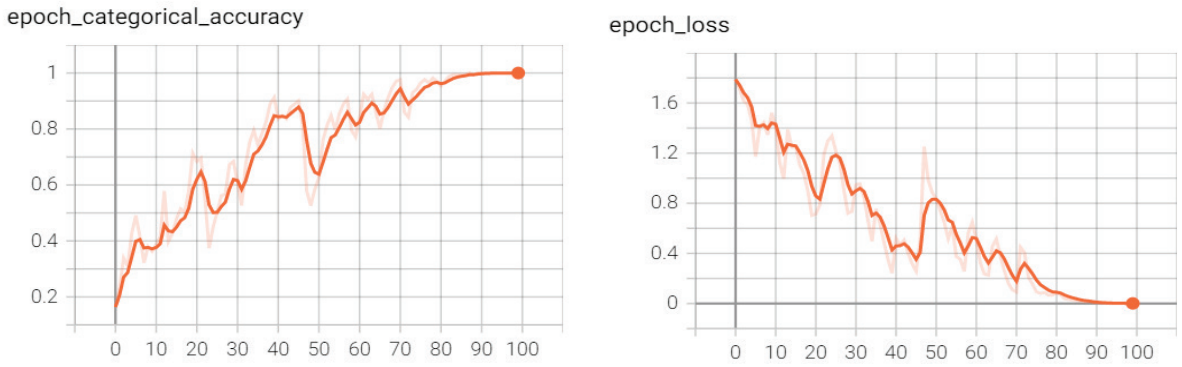


**Рисунок 7** – Графики точности и кривых потерь модели

Программная реализация разработанного алгоритма распознавания жестов выполнена на ноутбуке Asus с характеристиками Intel Core i7-1165G7 CPU 4, 2.80GHz, ОЗУ 16,00 ГБ и установленной на нем Windows 11 Home Single Language. Алгоритм позволяет распознавать жесты в режиме реального времени.

## 6. Заключение

Распознавание жестов является важной задачей для обеспечения коммуникативного взаимодействия людей с нарушениями слуха. Русские жестовые языки изучены мало, для них отсутствует общедоступный набор данных. Так как даже в рамках одного региона жестовые языки обладают существенной вариативностью для одного жеста разработан собственный набор данных, состоящий из 10 наиболее популярных слов русского жестового языка. Для распознавания жестов использовано выделение признаков, таких как ключевые точек руки, лица и тела человека. В качестве классификатора выделенных признаков использована рекуррентная нейронная сеть. Экспериментальные исследования подтверждают эффективность предложенного алгоритма распознавания жестов по видеоданным.

## 7. Список источников

[1] Всемирная организация здравоохранения [Электронный ресурс] URL: https://www.who.int/ru (дата обращения: 10.08.2022).

[2] Введение в лингвистику жестовых языков. Русский жестовый язык: учебник / ред.: С.И. Буркова, В.И. Киммельман. – Новосибирск: Изд-во НГТУ, 2019. – 356 с.

[3] Pansare J.R., Gawande S.H., Ingle M. Real-Time Static Hand Gesture Recognition for American Sign Language (ASL) in Complex Background // Journal of Signal and Information Processing, Volume 3, Number 3, 2012. — pp. 364-367.

[4] Murakami K., Taguchi H. Gesture Recognition using Recurrent Neural Networks // ACM Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology (CHI '91), 1999. — pp. 237-242.
[5] Pugeault N., Bowden R. Spelling It Out: Real-Time ASL Fingerspelling Recognition // In Proceedings of the 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, jointly with ICCV'2011, 2011. — pp. 1114-1119 112.
[6] Kinect для Windows [Электронный ресурс] // Microsoft Kinect for Windows URL: https://docs.microsoft.com/ru-ru/windows/apps/design/devices/kinect-for-windows (дата обращения: 01.08.2022).
[7] OpenNI: The standard framework for 3D sensing [Электронный ресурс] URL: http://www.openni.org/openni-sdk (дата обращения:01.08.2022).
[8] Graziano Fronteddua, Simone Porcuab, Alessandro Florisab, Luigi Atzoriab A dynamic hand gesture recognition dataset for human-computer interfaces [Электронный ресурс] // Computer Networks. - 2022. - №Volume 205. URL: https://www.sciencedirect.com/science/article/abs/pii/S1389128622000172 (дата обращения 11.08.2022).
[9] R. Elakkiyaa, K. Selvamanib, S.Kanimozhic, Rao. Velumadhavac, A. Kannand Intelligent System for Human Computer Interface Using Hand Gesture Recognition [Электронный ресурс] // Procedia Engineering. - 2012. - №Volume 38. URL: https://www.sciencedirect.com/science/article/pii/S1877705812022825 (дата обращения 11.08.2022).
[10] Byong K. K, Hyun S. Yang Finger mouse and gesture recognition system as a new human computer interface [Электронный ресурс] // Computers & Graphics. - 1997. - №Volume 21. - С. 555-561. URL: https://www.sciencedirect.com/science/article/pii/S0097849397000344 (дата обращения 11.08.2022).
[11] Katherina A. Jurewicza, David M. Neyensb Redefining the human factors approach to 3D gestural HCI by exploring the usability-accuracy tradeoff in gestural computer systems [Электронный ресурс] // Applied Ergonomics. - 2022. - №Volume 105. URL: https://www.sciencedirect.com/science/article/pii/S0003687022001569 (дата обращения 11.08.2022).
[12] Chengshuo Xia, Ayane Saito, Yuta Sugiura Using the virtual data-driven measurement to support the prototyping of hand gesture recognition interface with distance sensor [Электронный ресурс] // Sensors and Actuators A: Physical. - 2022. - №Volume 338. URL: https://www.sciencedirect.com/science/article/pii/S0924424722001017 (дата обращения 11.08.2022).
[13] Baao Xieab, James Mengc, Baihua Lid, Andy Harlande Biosignal-based transferable attention Bi-ConvGRU deep network for hand-gesture recognition towards online upper-limb prosthesis control [Электронный ресурс] // Computer Methods and Programs in Biomedicine. - 2022. - №Volume 224. URL: https://www.sciencedirect.com/science/article/abs/pii/S0169260722003819 (дата обращения 11.08.2022).
[14] M. Neela Harisha, S. Poonguzhalib Review of gesture recognition gloves and its applications in thermal insulation and heat resistant [Электронный ресурс] // Materials Today: Proceedings. - 2022. - №Volume 21. URL: https://www.sciencedirect.com/science/article/pii/S221478532202956X (дата обращения 11.08.2022).
[15] Shu Wang, Aiguo Wang, Mengyuan Ran, Li Liu, Yuxin Peng, Ming Liu, Guoxin Su, Adi Alhudhaif, Fayadh Alenezi, Norah Alnaim Hand gesture recognition framework using a lie group based spatio-temporal recurrent network with multiple hand-worn motion sensors [Электронный ресурс] // Information Sciences. - 2022. - №Volume 606. URL: https://www.sciencedirect.com/science/article/pii/S0020025522005230 (дата обращения 11.08.2022).
[16] Близко (Русский Язык Жестов) [Электронный ресурс] // Город жестов URL: https://www.youtube.com/watch?v=345cw8vSilg (дата обращения: 15.07.2022).
[17] Куликова Людмила Викторовна, Магировская Оксана Валериевна, Шатохина Софья Александровна, Срмикян Виолетта Сименовна, Привалихина Екатерина Сергеевна,

Генвальд Андрей Сергеевич Региональные жестовые языки: мультимодальный электронный корпус (на материале коммуникативного пространства Восточной Сибири) // 2022.

[18] Mediapipe: a framework for building perception pipelines / C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M.G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, M. Grundmann. – arXiv preprint arXiv:1906.08172. – 2019.

[19] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780.

[20] Антонио Джулли, Суджит Пал Библиотека Keras – инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow. - ДМК Пресс, 2017. 296 с.

[21] OpenCV [Электронный ресурс]. URL: http://opencv.org/about.html

[22] Mediapipe: a framework for building perception pipelines / C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M.G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, M. Grundmann. – arXiv preprint arXiv:1906.08172. – 2019.