https://doi.org/10.25559/SITITO.021.202501.103-112
УДК 004.89

Оригинальная статья

Метод временных сегментов в задаче распознавания жестовых слов с предобученной сетью CNN-LSTM

С. Ли*, И. Н. Полякова
ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова», г. Москва, Российская Федерация
Адрес: 119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1
*nugejusko@gmail.com

Аннотация

Распознавание жестового языка играет важную роль в обеспечении доступности и инклюзивности для людей с нарушением слуха, создавая условия для более эффективной коммуникации в социальной и профессиональной среде. Однако обработка видеоданных в задачах классификации жестов требует значительных вычислительных ресурсов из-за необходимости анализа большого количества кадров, что приводит к увеличению времени обучения и потребления памяти. Кроме того, видеоданные содержат избыточные кадры, не несущие важной информации для распознавания жестов, что дополнительно усложняет процесс обработки и приводит к неэффективному использованию ресурсов.

В данной работе представлена модель автоматического распознавания слов жестового языка, в основе которой лежит метод временных сегментов (TSN). Этот метод позволяет отбирать ключевые кадры из видеопоследовательностей, сокращая избыточные данные и значительно уменьшая время обучения и использование оперативной памяти без значительного ухудшения качества классификации. В качестве базовой технологии использовалась архитектура CNN-LSTM, где сверточная сеть ResNet отвечает за извлечение пространственных признаков, а рекуррентный слой LSTM обрабатывает временные зависимости в последовательностях.

В качестве данных использовался датасет WLASL, содержащий 6 классов жестов. Оценка производительности модели проводилась с использованием метрик Accuracy и F1-мера, что позволило объективно сравнить её с альтернативными подходами. В ходе экспериментов проведён сравнительный анализ различных предварительно обученных версий ResNet (ResNet18, ResNet34, ResNet50, ResNet101, ResNet152) и выявлена оптимальная конфигурация модели.

Результаты показали, что применение TSN привело к значительному снижению вычислительных затрат: время обучения сократилось в 2.173 раза, использование GPU RAM — в 0.5514 раза, а системной памяти — в 1.027 раза. При этом модель с TSN продемонстрировала даже более высокую точность по сравнению с версией без TSN, что подтверждает эффективность метода в задаче классификации жестового языка.

Таким образом, комбинация CNN-LSTM с ResNet18 и TSN обеспечивает не только высокую точность, но и эффективное использование вычислительных ресурсов. Полученные результаты могут служить основой для дальнейшего развития систем автоматического распознавания жестового языка, включая масштабирование на более крупные датасеты и интеграцию с мультимодальными системами обработки жестов.

Ключевые слова: датасет WLASL, классификация видеоданных, метод временных сегментов, предобученная модель, распознавание жестовых слов, сверточная рекуррентная сеть

Финансирование: данная работа выполнена в рамках НИР «Математическое и программное обеспечение перспективных систем обработки символьной информации с элементами искусственного интеллекта», проводимой на кафедре алгоритмических языков факультета вычислительной математики и кибернетики Московского государственного университета имени М.В.Ломоносова.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Ли С., Полякова И. Н. Метод временных сегментов в задаче распознавания жестовых слов с предобученной сетью CNN-LSTM // Современные информационные технологии и ИТ-образование. 2025. Т. 21, № 1. С. 103-112. https://doi.org/10.25559/SITITO.021.202501.103-112

Original article

Temporal Segment Method in Sign Word Recognition Using a Pretrained CNN-LSTM Network

Seungju Lee*, I. N. Polyakova
Lomonosov Moscow State University, Moscow, Russian Federation
Address: 1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation
*nugejusko@gmail.com

Abstract

Sign language recognition plays a crucial role in enhancing accessibility and inclusion for people with hearing impairments, facilitating more effective communication in social and professional environments. However, gesture classification from video data typically demands substantial computational resources due to the large number of frames that need processing, significantly increasing training time and memory consumption. Additionally, video data often contain redundant frames without meaningful information for gesture recognition, further complicating the processing and leading to inefficient resource usage.

This paper introduces a model for automatic sign language word recognition based on the Temporal Segment Networks (TSN) method. TSN effectively selects key frames from video sequences, reducing redundant information, significantly decreasing training time and RAM usage, without substantially compromising classification accuracy. The CNN-LSTM architecture is employed as the underlying technology, with the ResNet convolutional neural network responsible for extracting spatial features and the recurrent LSTM layer handling temporal dependencies in sequences.

The model was trained and tested on the WLASL dataset containing 6 gesture classes. Model performance was evaluated using Accuracy and F1-score metrics, allowing an objective comparison with alternative approaches. Experiments included a comparative analysis of different pretrained ResNet models (ResNet18, ResNet34, ResNet50, ResNet101, ResNet152), resulting in the identification of the optimal configuration.

Results demonstrated that applying TSN significantly reduced computational costs: training time was reduced by a factor of 2.173, GPU RAM usage decreased by a factor of 0.5514, and system memory usage by a factor of 1.027. Furthermore, the TSN-based model exhibited higher accuracy compared to the non-TSN variant, confirming the method's effectiveness in gesture classification tasks.

Thus, the combination of CNN-LSTM with ResNet18 and TSN not only achieves high accuracy but also ensures efficient use of computational resources. The obtained results can serve as a foundation for further development of automatic sign language recognition systems, including scaling to larger datasets and integration with multimodal gesture-processing systems.

Keywords: WLAL Dataset, Video Data Classification, Temporal Segment Networks, Pretrained Model, Gesture Word Recognition, Convolutional Recurrent Network

Funding: This work was carried out within the framework of the research project "Mathematical methods and software for advanced symbolic information processing systems with artificial intelligence features", conducted at the Department of Algorithmic Languages of the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University.

Conflict of interests: The authors declare no conflict of interest.

For citation: Lee Seungju, Polyakova I.N. Temporal Segment Method in Sign Word Recognition Using a Pretrained CNN-LSTM Network. *Modern Information Technologies and IT-Education*. 2025;21(1):103-112. https://doi.org/10.25559/SITITO.021.202501.103-112

Введение

Язык – это ключевая система человеческого общения, которая позволяет передавать мысли, эмоции. Однако люди с нарушением слуха, используя жесты, письменную речь и чтение по губам, сталкиваются с различными барьерами в социальной среде, при культурной интеграции.

Доступность медиасредств также представляет собой актуальную проблему. Недостаточные или отсутствующие, например, субтитры в телевизионных программах и фильмах могут ограничивать возможность взаимодействия людей, имеющих нарушение слуха, с медиаконтентом, что значительно сужает их культурный и информационный доступ к ресурсам. Кроме того, факт 30 % неграмотности среди людей с нарушением слуха [1] усугубляет эти проблемы, так как некоторые из них не могут эффективно использовать письменную речь.

Во многих случаях медиасредства не предоставляют услуги сурдоперевода, хотя практика новостных программ показывает важность жестового языка для людей с нарушением слуха. Слуховые аппараты частично помогают при незначительных нарушениях, а кохлеарные импланты и пересадка барабанной перепонки остаются недоступными или неподходящими для многих пациентов[1]. Это подчеркивает острую необходимость в решениях, которые обеспечивают интерпретацию жестового языка в реальном времени, особенно для тех, кто не получил формального образования или сталкивается с другими барьерами в общении.

Жестовые языки разнообразны и уникально структурированы, что затрудняет создание универсальных моделей перевода. Для каждого из них необходимы адаптированные модели, а отсутствие данных по некоторым языкам, например африканским, ограничивает исследования. Тем не менее, существуют наборы данных по арабскому [2], индийскому [3], корейскому [4] и другим языкам [5-9]. Предлагаемое исследование стремится решить эти проблемы путем разработки модели для классификации и распознавания слов на жестовом языке.

Одним из ключевых вызовов при разработке моделей распознавания жестового языка является обработка большого объема видеоданных. Традиционные подходы требуют значительных вычислительных ресурсов, так как анализируют все кадры видео. В данном исследовании предлагается использовать метод временных сегментов *(Temporal Segment Networks, TSN)* [10], который позволяет выбирать наиболее значимые кадры, устраняя дублирующуюся информацию. Ожидается, что этот метод приведет к значительному снижению нагрузки на вычислительные мощности, особенно в части потребления оперативной памяти и времени обучения модели, без потери точности классификации.

Данная работа закладывает основу для развития доступных инструментов коммуникации и улучшения социальной инклюзии людей с нарушением слуха.

1. Подходы к решению

Проблема распознавания человеческих жестов решалась с использованием различных подходов, большинство из которых основаны на методах глубокого обучения с применением нейронных сетей [11-15]. Эти подходы можно классифицировать следующим образом.

Трехмерная сверточная сеть – 3D CNN

Видеоданные представляют собой трёхмерный массив. К двумерному массиву изображений добавляется ось времени. Классическим методом извлечения признаков из таких данных является использование свёрточных нейронных сетей *(Convolutional Neural Networks, CNN)*. 3D CNN способны обрабатывать видеоданные с учётом временной информации и продемонстрировали отличные результаты в распознавании и классификации движений [16].

Например, в работе [17] использовали 3D CNN для обнаружения движений, а в исследовании [18] применяли 3D свёртку для улучшения классификации видео. Другие исследования, такие как [19], также использовали 3D свёртки для сегментации видео и повышения точности. Эти исследования подчеркивают преимущества 3D CNN как мощной модели, учитывающей пространственные и временные характеристики видео. Однако у 3D CNN есть ограничение, связанное с необходимостью фиксированной длины по временной оси [20].

Сверточная рекуррентная сеть – CNN-LSTM

Видеоданные можно рассматривать как последовательность кадров, которые необходимо обрабатывать как временной ряд, учитывая временные зависимости. Сети долгой краткосрочной памяти *(Long short-term memory, LSTM)* подходят для выполнения таких задач. В отличие от 3D CNN, CNN-LSTM не требует фиксированной длины по временной оси, но обработка каждого кадра требует больше памяти и времени для обучения.

Работа [21] представляют гибридную модель CNN-LSTM с 1D-сверточными слоями для динамического распознавания жестового языка, повышающую точность и скорость интерпретации жестов. Авторы [22] предлагают оптимизированную с помощью Adam архитектуру CNN и LSTM для улучшения распознавания жестового языка, достигая высокой точности в задачах классификации жестов. Исследования [23] разрабатывают систему распознавания изолированных жестов на основе видео, используя гибрид CNN-LSTM с механизмами внимания для улучшения детекции жестов. Разработчики [24] сравнивают эффективность архитектур LSTM1024 и 2D CNN для непрерывного распознавания жестового языка, выявляя оптимальные модели для повышения точности. Исследователи [25] реализуют сети LSTM для перевода тайского жестового языка в текст в реальном времени, что способствует коммуникации для людей с нарушением слуха. Авторы [26] проектируют систему распознавания жестового языка на основе глубокой CNN, используя нормализацию ключевых кадров и LSTM для повышения точности предсказаний.

Также существуют временные свёрточные сети, разработанные для работы с временными рядами [27]. Однако в данном исследовании они не рассматриваются.

[1] Arslan N. Cochlear implants can help restore hearing – but not for everyone [Электронный ресурс] // Fast Company. 24 Jan 2023. URL: https://www.fastcompany.com/90838284/cochlear-implants-can-help-restore-hearing-but-not-for-everyone (дата обращения: 11.01.2025).

Трансформер для обработки видео

Трансформер для обработки видео *(Video Vision Transformer, ViViT)* расширяет концепцию Vision Transformer на временное измерение. Vision Transformer эффективно анализирует изображения, используя механизмы внимания[2] ViViT развивает эти идеи, применяя их к видеоданным. Каждый кадр рассматривается как токен, чтобы учитывать как пространственную, так и временную информацию. Эта модель способна обрабатывать пространственные и временные характеристики видеоданных и используется для задач классификации, отслеживания объектов, распознавания действий и т.д.

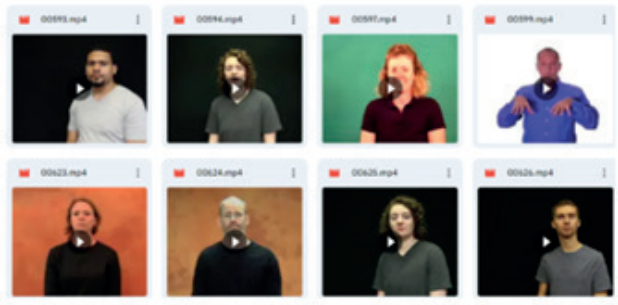
Механизм внимания позволяет ViViT учитывать зависимости на длинных временных промежутках [28-30].

Выводы

Среди этих трёх моделей CNN-LSTM можно считать наиболее подходящей для разработки системы распознавания жестового языка по следующим причинам: она эффективно учитывает временные зависимости, анализирует пространственные характеристики изображений, обладает относительно простой архитектурой с меньшими вычислительными затратами и уже широко применяется в задачах классификации видео.

2. Датасет

Язык жестов представлен более чем 300 вариантами по всему миру[3], но единого стандарта не существует. Среди наиболее известных – ASL, BSL, RSL и IS. Одним из самых распространённых является индо-пакистанский язык жестов[4], однако большинство исследований основаны на ASL из-за его доступности. В связи с этим в предлагаемом исследовании был выбран датасет *(Word-Level American Sign Language, WLASL)* [31], содержащий слова-жесты, а не только алфавит.



Р и с. 1. Пример датасета[5]

F i g. 1. Dataset example[5]

WLASL является крупнейшим видеодатасетом слов американского языка жестов *(American Sign Language, ASL)* и включает 2000 общеупотребимых слов [31]. Из-за высокой вычислительной нагрузки для данного исследования были выбраны 6 классов: «before», «book», «chair», «computer», «drink», «go», по 15 видеороликов на каждый, с уникальными идентификаторами (рис. 1).

3. Метод временных сегментов

Метод временных сегментов [10] оптимизирует обработку видеоданных, сокращая избыточные кадры и снижая нагрузку на вычислительные ресурсы. Видео V разделяется на K сегментов $\{S_1, S_2, \cdots, S_K\}$, из каждого случайно выбирается фрагмент T_K. Последовательность сегментов моделируется как:

$TSN(T_1, T_2, \cdots, T_K) = H(G(F(T_1;W), F(T_2;W), \cdots, F(T_K;W)))$

где $F(T_K;W)$ – сверточная сеть *(ConvNet)* с параметрами W, оценивающая класс каждого фрагмента.

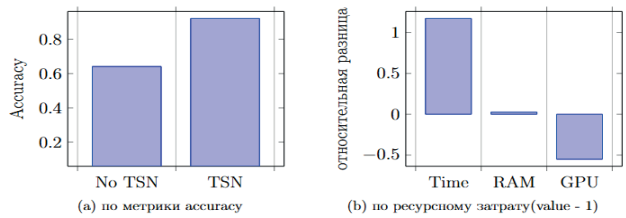
Она объединяет результаты, а (Softmax) предсказывает итоговый класс. В данной работе применяется упрощенная модель:

$$TSN(T_1, T_2, , T_K) = H(\sum_{i=0}^{n} F(T_i; W))$$

Изображения извлекаются заранее, хранятся в каталоге, а аннотации содержат пути к видео, начальный и конечный кадры и метки.

В среде PyTorch, используя Conv2d для свертки и LSTM для обработки последовательностей, проведено сравнение использования ресурсов с исходным набором данных.

Эксперимент (20 эпох) показал сокращение времени обучения в 2,173 раза, снижения загрузки GPU RAM в 0,5514 раза и системной памяти в 1,027 раза. Это говорит об эффективности метода TSN в оптимизации вычислительных ресурсов (рис. 2).



(a) по метрики accuracy (b) по ресурсному затрату(value - 1)

Рис 2: Сравнительный график с применением и без применения

Р и с. 2. Сравнительный график с применением и без применения

F i g. 2. Comparison graph with and without application

(a) by Accuracy metric (b) by resource consumption (value – 1)

Источник: здесь и далее в статье все таблицы и рисунки составлены авторами.

Source: Hereinafter in this article all tables and figures were made by the authors.

[2] Boesch G. Vision Transformers (ViT) in Image Recognition [Электронный ресурс] // viso.ai. 25 Nov. 2023. URL: https://viso.ai/deep-learning/vision-transformer-vit/ (дата обращения: 11.01.2025).

[3] There is no Universal Sign Language [Электронный ресурс] // Easier Project, 2025. URL: https://www.project-easier.eu/news/2022/09/22/there-is-no-universal-sign-language (дата обращения: 11.01.2025).

[4] What are the top 200 most spoken languages? / Ethnologue: Languages of the World ; ed by. D. M. Eberhard, G. F. Simons, C. D. Fennig. Twenty-eighth edition. Dallas, Texas: SIL International, 2025 [Электронный ресурс]. URL: https://www.ethnologue.com/insights/ethnologue200/ (дата обращения: 11.01.2025).

[5] WLASL (World Level American Sign Language) Video [Электронный ресурс] // Globose Technology Solutions : офиц. сайт. URL: https://gts.ai/dataset-download/wlasl-world-level-american-sign-language-video (дата обращения: 11.01.2025).

4. Предобученные модели CNN – LSTM

Модель представляет собой композицию CNN и LSTM. В исследованиях [32] отмечено, что увеличение глубины CNN повышает производительность, но чрезмерное усложнение приводит к исчезновению градиента, затрудняя обучение. Для решения этой проблемы используются предварительно обученные сети, такие как ResNet [33], VGG [34], EfficientNet [35], сокращая затраты на разработку модели.

В данном эксперименте были исследованы слои ResNet из библиотеки torchvision PyTorch. ResNet использует остаточные блоки $F(x) = H(x) - x$, что позволяет минимизировать разницу между входом и целевым значением, ускоряя обучение [33]. Структура модели (Таблица 1).

Таблица 1. **Таблица структур ResNet**
Table 1. **Architectures for ImageNet**

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Источник: [33].
Source: [33].

Каждая модель обучалась 50 эпох, сравнивалась максимальная train ccuracy (Таблица 2). ResNet18 и ResNet34 показали лучшие результаты, поэтому в итоговой модели используется ResNet18.

Таблица 2. **Таблица сравнения ResNet**
Table 2. **ResNet Comparison Table**

ResNet Model	Accuracy
ResNet18	94%
ResNet34	94%
ResNet50	73%
ResNet101	68%
ResNet152	81%

5. Методы оценки качества модели

Для оценки производительности модели в работе используются метрики Accuracy и F1-мера.

Accuracy: доля правильно классифицированных примеров:

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

Используется как интуитивно понятный показатель общей точности модели.

F1-мера: гармоническое среднее Precision и Recall:

$$F = \frac{(2+1)PR}{\beta^2 R + P}, \beta = 1$$

Применяется для оценки качества классификации, обеспечивая баланс между полнотой и точностью.

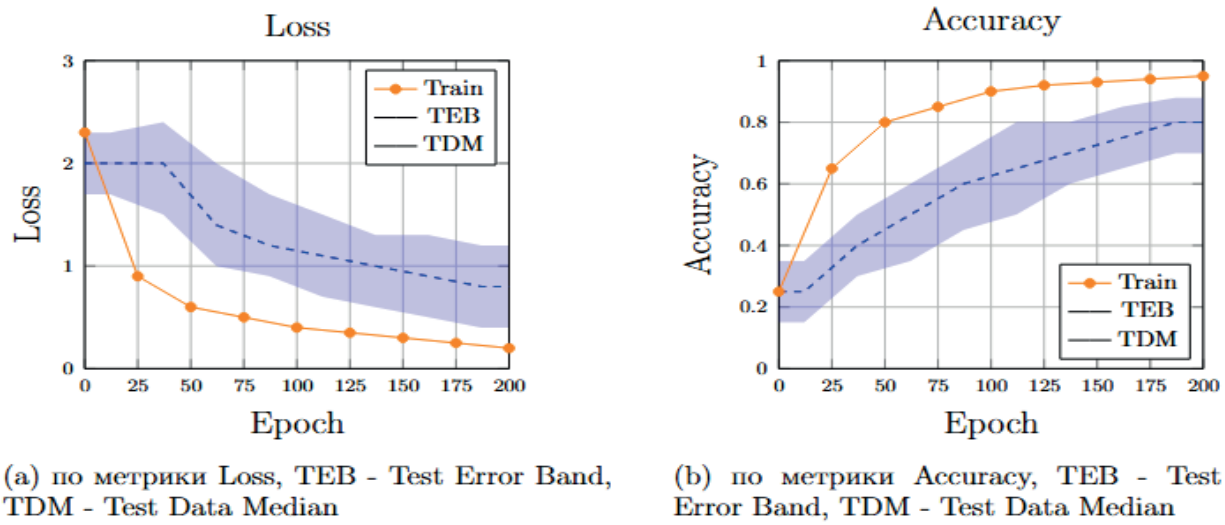
Эти метрики позволяют сравнивать результаты с другими исследованиями по классификации жестовых слов.

Для оценки стабильности модели можно использовать среднее значение метрики Accuracy по всем эпохам, что позволяет проверить последовательность работы без оптимизации количества эпох. Однако этот метод может скрывать динамику обучения. Для измерения максимальной производительности модели используется максимальное значение Accuracy, отражающее наивысший результат за все эпохи. Поскольку в данном исследовании акцент сделан на повышение эффективности модели, в качестве метрики выбрано максимальное значение Accuracy (Таблица 3).

Таблица 3. **Таблица сравнения ResNet**
Table 3. **ResNet Comparison Table**

Accuracy	F-1 Score
89%	0.83

Accuracy и F1-оценка тестового набора были получены с достаточно высокими значениями.
Ошибка на обучающем наборе *(Loss)* стабильно уменьшается, что свидетельствует об успешном обучении модели, в то время как тестовая ошибка колеблется в пределах заданного диапазона, отражая вариативность данных. Точность на обучающем наборе *(Accuracy)* уверенно растёт с увеличением числа эпох, но при этом тестовая точность сохраняет определённый разброс. Несмотря на то, что диапазон Test Median не является узким, наблюдается явная тенденция к его сходимости (рисунок 3).



(a) по метрики Loss, TEB - Test Error Band, TDM - Test Data Median

(b) по метрики Accuracy, TEB - Test Error Band, TDM - Test Data Median

Рис. 3. График процесса обучения
Fig. 3. Learning Process Graph
(a) by Loss metric, TEB – Test Error Band, TDM – Test Data Median
(b) by Accuracy metric, TEB – Test Error Band, TDM – Test Data Median

Максимальное значение Accuracy других методов, использующих этот набор данных, представлена в следующей таблице (Таблица 4) [31].

Таблица 4. **Сравнение с другими моделями**
Table 4. **Comparison with other models**

Methods	Accuracy
ST-GCN	65.45%
SignBERT	72.38%
RGB-based	81.33%
HMA	70.00%
BSL	78.47%
SignBERT (+ R)	86.93%
Ours	88.98%

Сравнительно высокое значение Accuracy теста можно наблюдать по сравнению с другими моделями на том же наборе данных. Однако данная модель использует небольшой набор данных из 6 классов, в то время как другие модели используют данные с более чем 2000 классами.

Заключение

В данной работе была разработана и реализована модель автоматического распознавания слов жестового языка, основанная на архитектуре CNN-LSTM и использующая предварительно обученную сеть ResNet. Для оптимизации вычислительных ресурсов применён метод временных сегментов (TSN), что позволило значительно сократить время обучения, уменьшить использование оперативной памяти, при этом загрузка GPU осталась практически неизменной. Основной эффект применения TSN заключается в сокращении времени обучения как ключевого фактора оптимизации, что позволяет эффективнее использовать доступные вычислительные мощности без существенной потери качества классификации. Более того, результаты показали, что нейросеть с применением TSN демонстрирует лучшую точность по сравнению с моделью без TSN, что подтверждает гипотезу о том, что выборка ключевых кадров из почти дублирующихся кадров улучшает не только эффективность использования ресурсов, но и качество обучения. Проведён сравнительный анализ предложенной модели с альтернативными методами по метрике Accuracy, подтвердивший её эффективность. В ходе экспериментов были протестированы различные предварительно обученные модели

ResNet, включая ResNet18, ResNet34, ResNet50, ResNet101 и ResNet152. Анализ показал, что увеличение глубины сети не всегда приводит к улучшению качества классификации, при этом ResNet18 обеспечил оптимальное сочетание точности и вычислительных затрат.

Достигнутые результаты формируют основу для дальнейших исследований, направленных на улучшение методов классификации и интерпретации жестового языка. В перспективе планируется расширение набора данных, тестирование модели на более сложных лексических структурах и интеграция с мультимодальными системами распознавания жестов.

References

[1] Gil Y.H., Jee H.K., Lee J.S., Nam D.S. E-Book Accessibility Technology Trends and Prospects. *Electronics and Telecommunications Trends.* 2015;30(4):59-70. Available at: https://koreascience.kr/article/JAKO201552057196138.page (accessed 11.01.2025). (In Korean)

[2] Noor T.H., Noor A., Alharbi A.F., Faisal A., Alrashidi R., Alsaedi A.S., Alharbi G., Alsanoosy T., Alsaeedi A. Real-Time Arabic Sign Language Recognition Using a Hybrid Deep Learning Model. *Sensors*. 2024;24(11):3683. https://doi.org/10.3390/s24113683

[3] Renjith S., Manazhy R. Indian Sign Language Recognition: A Comparative Analysis Using CNN and RNN Models. In: 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT). Kollam, India: IEEE Press; 2023. p. 1573-1576. https://doi.org/10.1109/ICCPCT58313.2023.10245525

[4] Shin J., Musa Miah A.S., Hasan M.A.M., Hirooka K., Suzuki K., Lee H.-S., Jang S.-W. Korean Sign Language Recognition Using Transformer-Based Deep Neural Network. *Applied Sciences*. 2023;13(5):3029. https://doi.org/10.3390/app13053029

[5] Kapitanov A., Karina K., Nagaev A., Elizaveta P. Slovo: Russian Sign Language Dataset. In: Christensen H.I., Corke P., Detry R., Weibel J.B., Vincze M. (eds.) Computer Vision Systems. ICVS 2023. *Lecture Notes in Computer Science*. Vol. 14253. Cham: Springer; 2023. p. 63-73. https://doi.org/10.1007/978-3-031-44137-0_6

[6] Hama Rawf K.M., Abdulrahman A.O., Mohammed A.A. Improved Recognition of Kurdish Sign Language Using Modified CNN. Computers. 2024;13(2):37. https://doi.org/10.3390/computers13020037

[7] Bora J., Dehingia S., Boruah A., Chetia A.A., Gogoi D. Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning. *Procedia Computer Science*. 2023;218:1384-1393. https://doi.org/10.1016/j.procs.2023.01.117

[8] Das S., Imtiaz M.S., Neom N.H., Siddique N., Wang H. A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier. *Expert Systems with Applications.* 2023;213(B):118914. https://doi.org/10.1016/j.eswa.2022.118914

[9] Abdelrazik M.A., Zekry A., Mohamed W.A. Efficient Deep Learning Algorithm for Egyptian Sign Language Recognition. In: 2023 33rd Conference of Open Innovations Association (FRUCT). Zilina, Slovakia: IEEE Press; 2023. p. 3-8. https://doi.org/10.23919/FRUCT58615.2023.10142991

[10] Wang L., et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: Leibe B., Matas J., Sebe N., Welling M. (eds.) Computer Vision – ECCV 2016. ECCV 2016. *Lecture Notes in Computer Science.* Vol. 9912. Cham: Springer; 2016. p. 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[11] Papatsimouli M., et al. Real Time Sign Language Translation Systems: A review study. In: 2022 11th International Conference on Modern Circuits and Systems Technologies (MOCAST). Bremen, Germany: IEEE Press; 2022. p. 1-4. https://doi.org/10.1109/MOCAST54814.2022.9837666

[12] Liang Z., Li H., Chai J. Sign Language Translation: A Survey of Approaches and Techniques. *Electronics*. 2023;12(12):2678. https://doi.org/10.3390/electronics12122678

[13] He S. Research of a Sign Language Translation System Based on Deep Learning. In: 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM). Dublin, Ireland: IEEE Press; 2019. p. 392-396. https://doi.org/10.1109/AIAM48774.2019.00083

[14] Grif M.G., Kozlov A.N., Manueva Yu.S. *Komp'yuternaya model' russkogo zhestovogo yazyka* [A computer model of the Russian sign language]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii* = Proceedings of the Russian higher school Academy of sciences. 2017;(1):46-57. (In Russ., abstract in Eng.) https://doi.org/10.17212/1727-2769-2017-1-46-57

[15] Ryumin D.A., Kagirov I.A., Axyonov A.A., Karpov A.A. Analytical Review of Models and Methods for Automatic Recognition of Gestures and Sign Languages. *Information and Control Systems*. 2021;(6):10-20. (In Russ., abstract in Eng.) https://doi.org/10.31799/1684-8853-2021-6-10-20

[16] Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., Fei-Fei L. Large-Scale Video Classification with Convolutional Neural Networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE Press; 2014. p. 1725-1732. https://doi.org/10.1109/CVPR.2014.223

[17] Ji S., Xu W., Yang M., Yu K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013;35(1):221-231. https://doi.org/10.1109/TPAMI.2012.59

[18] Kuehne H., Jhuang H., Garrote E., Poggio T., Serre T. HMDB: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE Press; 2011. p. 2556-2563. https://doi.org/10.1109/ICCV.2011.6126543

[19] Ding L., Xu C. TricorNet: A Hybrid Temporal Convolutional and Recurrent Network for Video Action Segmentation. *arXiv:*1705.07818. 2017. https://doi.org/10.48550/arXiv.1705.07818

[20] Sridhar V. Convolutional Neural Networks Based Sign Language. *International Journal of All Research Education and Scientific Methods.* 2024;12(9):2394-2398. Available at: https://www.ijaresm.com/uploaded_files/document_file/Varadala_Sridhar9iF9.pdf (accessed 11.01.2025).

[21] Gupta A., Sawan A., Singh S., Kumari S. Dynamic Sign Language Recognition with Hybrid CNN-LSTM and 1D Convolutional Layers. In: 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). Noida, India: IEEE Press; 2024. p. 1-6. https://doi.org/10.1109/ICRITO61523.2024.10522339

[22] Paul S.K., et al. An Adam based CNN and LSTM approach for sign language recognition in real time for deaf people. *Bulletin of Electrical Engineering and Informatics*. 2024;13(1):499-509. https://doi.org/10.11591/eei.v13i1.6059

[23] Kumari D., Anand R.S. Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework Based on Attention Mechanism. *Electronics*. 2024;13(7):1229. https://doi.org/10.3390/electronics13071229

[24] Amangeldy N., Krak I., Kurmetbek B., Gazizova N. A Comparison of the Effectiveness Architectures LSTM1024 and 2DCNN for Continuous Sign Language Recognition Process. *CEUR Workshop Proceedings.* 2024;3702:48-57. Available at: https://ceur-ws.org/Vol-3702/paper5.pdf (accessed 11.01.2025).

[25] Jintanachaiwat W., Jongsathitphaibul K., Pimsan N., et al. Using LSTM to translate Thai sign language to text in real time. *Discover Artificial Intelligence*. 2024;4:17. https://doi.org/10.1007/s44163-024-00113-8

[26] Jayanthi P., Sathia Bhama P.R.K., Madhubalasri B. Sign Language Recognition using Deep CNN with Normalised Keyframe Extraction and Prediction using LSTM. *Journal of Scientific & Industrial Research*. 2023;82(07):745-755. https://doi.org/10.56042/jsir.v82i07.2375

[27] Ahmadi S.A., Muhammad F., Dawsari H.A. CNN-TCN Deep Hybrid Model Based on Custom CNN with Temporal CNN to Recognize Sign Language. *Journal of Disability Research*. 2024;3(5):e20240034. https://doi.org/10.57197/JDR-2024-0034

[28] McGill E., Saggion H. Can True Zero-shot Methods with Large Language Models be Adopted for Sign Language Machine Translation? In: Proceedings of the First International Workshop on Knowledge-Enhanced Machine Translation. Sheffield, United Kingdom: European Association for Machine Translation; 2024. p. 40-43. Available at: https://aclanthology.org/2024.kemt-1.5/ (accessed 11.01.2025).

[29] Liu Y., Nand P., Hossain M.A., et al. Sign language recognition from digital videos using feature pyramid network with detection transformer. *Multimedia Tools and Applications*. 2023;82:21673-21685. https://doi.org/10.1007/s11042-023-14646-0

[30] Kothadiya D.R., Bhatt C.M., Saba T., Rehman A., Bahaj S.A. SIGNFORMER: DeepVision Transformer for Sign Language Recognition. *IEEE Access*. 2023;11:4730-4739. https://doi.org/10.1109/ACCESS.2022.3231130

[31] Li D., Opazo C.R., Yu X., Li H. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass, CO, USA: IEEE Press; 2020. p. 1448-1458. https://doi.org/10.1109/WACV45572.2020.9093512

[32] Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010. Vol. 9 of JMLR: W&CP 9. Chia Laguna Resort, Sardinia, Italy; 2010. p. 249-256. Available at: https://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf (accessed 11.01.2025).

[33] He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE Press; 2016. p. 770-778. https://doi.org/10.1109/CVPR.2016.90

[34] Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:*1409.1556. 2015. https://doi.org/10.48550/arXiv.1409.1556

[35] Tan M., Le Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36 th International Conference on Machine Learning. Vol. PMLR 97. Long Beach, California; 2019. p. 6105-6114. Available at: http://proceedings.mlr.press/v97/tan19a.html (accessed 11.01.2025).

Поступила 11.01.2025; одобрена после рецензирования 02.03.2025; принята к публикации 21.03.2025.
Submitted 11.01.2025; approved after reviewing 02.03.2025; accepted for publication 21.03.2025.

Об авторах:

Ли Сынгжу, студент кафедры алгоритмических языков факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), **ORCID: https://orcid.org/0009-0007-0395-946X,** nugejusko@gmail.com

Полякова Ирина Николаевна, доцент кафедры алгоритмических языков факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), кандидат физико-математических наук, доцент, **ORCID: https://orcid.org/0000-0003-1432-4906,** polyakova@cs.msu.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Seungju Lee, Bachelor's Degree student of the of the Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), **ORCID: https://orcid.org/0009-0007-0395-946X,** nugejusko@gmail.com

Irina N. Polyakova, Associate Professor of the Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), Cand. Sci. (Phys.-Math.), Associate Professor, **ORCID: https://orcid.org/0000-0003-1432-4906,** polyakova@cs.msu.ru

All authors have read and approved the final manuscript.