УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2022 Том 22 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2022 Vol. 22 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-3-585-593
УДК 004.855.5

Метод многомодального машинного сурдоперевода для естественного человеко-машинного взаимодействия

Александр Александрович Аксёнов[1]✉, Ильдар Амирович Кагиров[2], Дмитрий Александрович Рюмин[3]

[1,2,3] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] axyonov.a@iias.spb.su✉, https://orcid.org/0000-0002-7479-2851
[2] kagirov@iias.spb.su, https://orcid.org/0000-0003-1196-1117
[3] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569

Аннотация

Предмет исследования. Исследована возможность повышения надежности автоматической системы распознавания как отдельных жестов, так и жестового языка, за счет использования наиболее информативных пространственно-временных визуальных признаков. **Метод.** Представленный метод автоматического распознавания жестовой информации основан на интегральной нейросетевой модели, которая анализирует пространственно-временные визуальные признаки: 2D и 3D расстояния от лица до руки; площадь пересечения лица и руки; конфигурацию руки; гендерную и возрастную информацию о дикторе. Для извлечения информации о конфигурации руки разработана нейросетевая модель на основе архитектуры 3DResNet-18 для получения гендерной и возрастной информации. В метод встроены нейросетевые модели из программной платформы Deepface. **Основные результаты.** Предложенный метод апробирован на данных многомодального корпуса элементов жестового языка TheRuSLan, результаты которого достигают точности распознавания жестов 91,14 %. **Практическая значимость.** Результаты исследования позволяют повысить точность и робастность не только машинного сурдоперевода, но и естественность человеко-машинного взаимодействия в целом. Полученные результаты могут найти применение в сферах социального обслуживания медицины и образования, в робототехнике и в центрах обслуживания населения.

Ключевые слова

язык тела, жестикуляция, машинный сурдоперевод, естественность коммуникации

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 21-71-00141, https://rscf.ru/project/21-71-00141/

Ссылка для цитирования: Аксёнов А.А., Кагиров И.А., Рюмин Д.А. Метод многомодального машинного сурдоперевода для естественного человеко-машинного взаимодействия // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 3. С. 585–593. doi: 10.17586/2226-1494-2022-22-3-585-593

A method of multimodal machine sign language translation for natural human-computer interaction

Alexandr A. Axyonov[1]✉, Ildar A. Kagirov[2], Dmitry A. Ryumin[3]

[1,2,3] Saint Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation

[1] axyonov.a@iias.spb.su✉, https://orcid.org/0000-0002-7479-2851
[2] kagirov@iias.spb.su, https://orcid.org/0000-0003-1196-1117
[3] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569

Abstract
This paper aims to investigate the possibility of robustness enhancement as applied to an automatic system for isolated signs and sign languages recognition, through the use of the most informative spatiotemporal visual features. The authors present a method for the automatic recognition of gestural information, based on an integrated neural network model, which analyses spatiotemporal visual features: 2D and 3D distances between the palm and the face; the area of the hand and the face intersection; hand configuration; the gender and the age of signers. A 3DResNet-18-based neural network model for hand configuration data extraction was elaborated. Deepface software platform neural network models were embedded in the method in order to extract gender and age-related data. The proposed method was tested on the data from the multimodal corpus of sign language elements TheRuSLan, with the accuracy of 91.14 %. The results of this investigation not only improve the accuracy and robustness of machine sign language translation, but also enhance the naturalness of human-machine interaction in general. Besides that, the results have application in various fields of social services, medicine, education and robotics, as well as different public service centers.

Keywords
body language, gesticulation, machine sign language translation, naturalness of a communication medium

Acknowledgements
This research is financially supported by the Russian Science Foundation (No. 21-71-00141, https://rscf.ru/en/project/21-71-00141/)

For citation: Axyonov A.A., Kagirov I.A., Ryumin D.A. A method of multimodal machine sign language translation for natural human-computer interaction. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 3, pp. 585–593 (in Russian). doi: 10.17586/2226-1494-2022-22-3-585-593

Введение

Известно, что инвалиды по слуху ограничены в возможностях при общении со слышащими людьми, а при обращении в различные государственные учреждения им иногда предоставляются сурдопереводчики, возможности которых недостаточны на практике. Согласно данным Всемирной организации здравоохранения[1], на 2021 год в мире примерно 466 млн человек (более 5 % от общего количества населения земного шара, из них 34 млн дети) страдают полной глухотой или испытывают проблемы со слухом. Кроме того, каждый третий человек в возрасте старше 65 лет сталкивается с проблемой снижения качества слуха и, согласно оценкам, к 2050 году более 2 млрд человек будут страдать глухотой или испытывать проблемы со слухом. По этой причине необходимо развитие новых интеллектуальных технологий (систем) эффективного автоматического машинного сурдоперевода для организации естественного и универсального человеко-машинного взаимодействия.

Один из основных критериев успешной организации человеко-машинного взаимодействия [1] — естественность коммуникации [2]. В идеальном случае взаимодействие человека с машиной в терминах модальности не должно отличаться от межличностной коммуникации. Главное отличие современных интеллектуальных систем — применение способов коммуникации, характерных для общения между людьми. Неотъемлемой составляющей естественной коммуникации служат невербальные средства взаимодействия (в частности, язык тела и жестикуляция) [3]. Так например, при помощи жестов можно взаимодействовать с интеллектуальной информационной системой на некотором расстоянии и в условиях сильных фоновых шумов, когда звучащая речь малоэффективна [4]. Заметим, что в настоящее время полноценных автоматических систем машинного сурдоперевода не существует. Это обусловлено рядом факторов технического характера (визуальные шумы, окклюзии, изменение освещенности), недостаточностью описания грамматики и семантики жестовых языков (ЖЯ) для задач интерпретации жестов, а также факторами, которые напрямую связаны с диктором. Так гендерные и возрастные характеристики отдельно взятого диктора могут влиять на размеры ладоней, удаленность рук от тела, расстояние между активной и пассивной рукой и скорость представления разнообразных лексических жестовых единиц или клауз. Данные влияния гендерных и возрастных аспектов на невербальное поведение широко описано в гендерной лингвистике [5, 6], невербальной семиотике [7] и психологии [8], но такие влияния практически не учтены в контексте машинного сурдоперевода и имплементации различных методов цифровой обработки изображений, компьютерного зрения и нейросетевых моделей. В результате можно сделать вывод о том, что задача машинного сурдоперевода — комплексное междисциплинарное исследование и требует принципиально новых научно-технических результатов, которые позволят максимально эффективно распознавать отдельные жесты, а также элементы ЖЯ с учетом интеллектуального анализа гендерных и возрастных характеристик диктора.

Цель работы — разработка и применение нового метода автоматического извлечения наиболее информативных пространственно-временных визуальных признаков (характеристик) из лексических жестовых единиц или клауз на основе предварительных знаний о гендерно-возрастных характеристиках диктора. Метод может найти применение для автоматического распознавания отдельных ручных жестов и элементов ЖЯ (машинный сурдоперевод).

Предмет исследования

С помощью жестов и языка тела человек может передавать свои мысли, чувства, а также эмоции. Такие способы общения определяются как невербальная коммуникация. В лингвистике ЖЯ принято описывать

[1] Глухота и потеря слуха [Электронный ресурс]. Режим доступа: http://www.who.int/mediacentre/factsheets/fs300/ru/, свободный. Яз. рус. (дата обращения: 22.03.2022).

жесты при помощи конечного набора дифференциальных признаков [9, 10]: конфигурация руки [11], место выполнения жеста (локализации) и характер исполнения [12, 13]. Конфигурация руки задает определенную ориентацию ладони, конфигурацию и направление пальцев [14, 15]. Информация о месте выполнения жеста — релевантна, локализация всех жестов обычно постоянна [4]. Можно сделать вывод, что исполнение любого жеста в полной мере связано с его статическим или динамическим характером. Отметим, что для статических жестов характерна устойчивость признака локализации, тогда как при исполнении динамических жестов локализация и конфигурация изменяются во времени [16]. Другой известный факт — во время воспроизведения жеста диктором общее понимание складывается из множества движений руки (одноручные жесты) или рук (двуручные жесты) [17, 18]. Все описанные особенности показа жестов диктором должны учитываться при машинном сурдопереводе. Настоящая работа посвящена интеллектуальному анализу влияния дополнительных факторов (гендерных и возрастных характеристик диктора) на надежность автоматических систем распознавания как отдельных жестов, так и ЖЯ.

Данное исследование представляет продолжение существующих исследований [1, 3, 4, 16, 19, 20] авторов настоящей работы в таких областях, как машинный сурдоперевод, повышение качества межличностной коммуникации и человеко-машинное взаимодействие.

Описание метода

Современные методы автоматического распознавания жестовой информации [21–24] на основе интегральных нейросетевых моделей (End-to-End, E2E) могут уступать по скорости базовым подходам [3, 16], но существенно превосходить их в точности. В связи с этим в работе [20] был представлен собственный метод многомодального видеоанализа движений рук для автоматического распознавания ручных жестов и элементов ЖЯ на основе модели E2Ev1. Основная идея метода заключалась в анализе информативных пространственно-временных визуальных характеристик жеста в определенный момент времени с помощью предварительно обученной нейросетевой модели с долгой кратковременной памятью (Long Short-Term Memory, LSTM) [25].

В настоящей работе предложены следующие улучшения метода:

— расширен список из информативных пространственно-временных визуальных характеристик жеста;
— применена нейросетевая модель E2Ev2 для распознавания конфигураций рук диктора (вместо 2D сверточной нейросети). Функциональная схема метода показана на рис. 1.

Нейросетевая модель LSTM анализирует пространственно-временные визуальные характеристики жеста: 1) нормализованных 2D и 3D расстояний от лица до руки (зона артикуляции жеста); 2) нормализованной 2D площади пересечения лица и руки (в случае отсутствия пересечения, площадь нулевая); 3) конфигурации руки (представляется числовым значением класса от 0 до 22); 4) гендерная характеристика о дикторе; 5) возрастная характеристика о дикторе. Для решения задачи машинной классификации гендерных и возрастных характеристик диктора использованы нейросетевые модели[1] из программной платформы с открытым исходным кодом Deepface [26]. Таким образом, в отличие от метода [20] к списку визуальных характеристик добавлена гендерная информация о дикторе, представленная числовым значением класса (0 — мужчина, 1 — жен-

[1] Deepface Models [Электронный ресурс]. Режим доступа: https://github.com/serengil/deepface/tree/master/deepface, свободный. Яз. англ. (дата обращения: 23.03.2022).

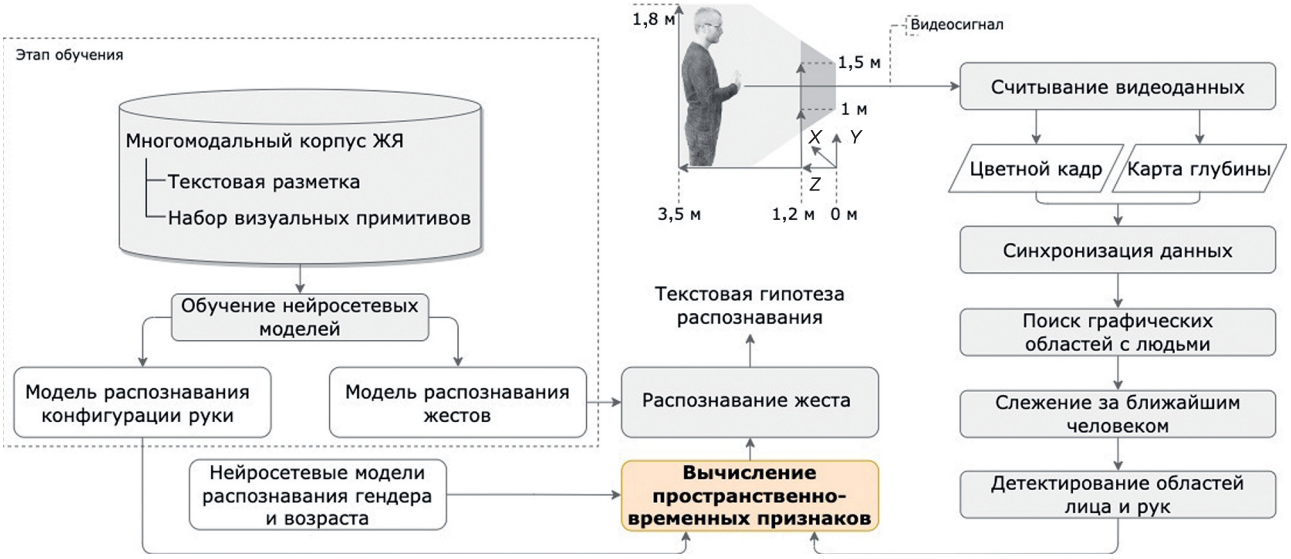

Рис. 1. Функциональная схема метода многомодального видеоанализа движений рук для распознавания изолированных жестов рук и элементов жестовых языков

Fig. 1. Functional diagram of the method of multimodal video analysis of hand movements for recognizing isolated hand gestures and sign language elements

щина), а также возрастная информация — от 0 до 100 лет. В результате новый список из информативных пространственно-временных визуальных характеристик жеста расширяется с 3 до 5 пунктов.

На рис. 2 показан наглядный пример изменения конфигурации руки в зависимости от пола и возраста для одного и того же динамического жеста. Из рис. 2 видно, что один и тот же динамический жест показывается по-разному тремя дикторами. Можно заметить, что женщины показывают схожие конфигурации руки, при этом их отличия видны в развороте руки и положении пальцев. В то время как мужчина показывает совершенно отличные от дикторов женского пола конфигурации руки. Таким образом, возникает необходимость в учете гендерной и возрастной информации при многомодальном видеоанализе движений рук для распознавания изолированных жестов рук и элементов ЖЯ.

По сравнению с исследованием в [20] количественный показатель (точность) процесса распознавания конфигураций рук диктора улучшен за счет предлагаемой архитектуры нейросетевой модели E2Ev2 (рис. 3).

Архитектура E2Ev2 получает на вход последовательности изображений (конфигурации рук диктора) с разрешением 192 × 192 пикселов и длиной в 30 кадров (Sequence_Length). Далее происходит извлечение карт признаков размерностью 512 × 6 × 6 из каждого изображения полученной последовательности с помощью 3D сверточного слоя (3D Conv) и модифицированных остаточных блоков (Residual Blocks модели ResNet-18), включающих модули внимания (Squeeze-and-Attention, SA). В свою очередь слой подвыборки (Global Average Polling) преобразует карты признаков из размерности 512 × 6 × 6 в одномерные вектора размерностью 30 × 512. Dropout — метод регуляризации нейросети, предотвращающий переобучение сети. В заключении нейросетевая модель LSTM обрабатывает полученные одномерные вектора и выдает результат в виде распознавания конфигураций рук диктора.

Рис. 2. Примеры видеокадров показа одного и того же динамического жеста тремя дикторами: женщина в возрасте 34 лет (по прогнозу модели распознавания возраста) (*a*); женщина (21 год) (*b*); мужчина (21 год) (*c*)

Fig. 2. Examples of video frames showing the same dynamic gesture by three speakers: woman, 34 y.o. (as predicted by the age recognition model) (*a*); woman, 21 y.o. (*b*); man, 21 y.o. (*c*)

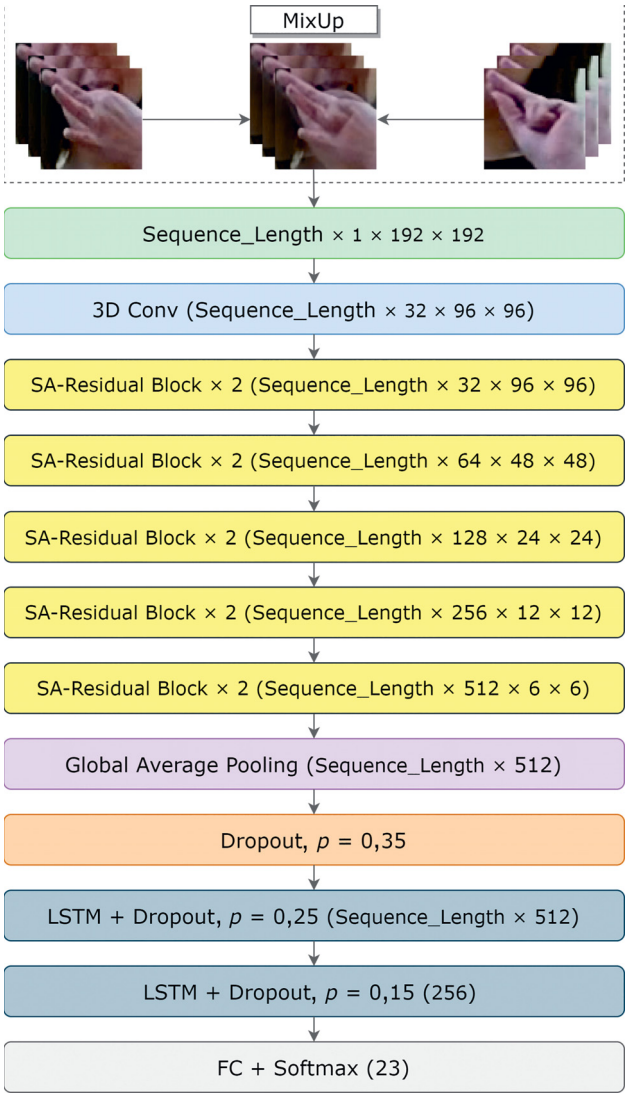

Рис. 3. Архитектура нейросетевой модели E2Ev2 для распознавания конфигураций рук диктора

Fig. 3. E2Ev2 architecture of a neural network model for recognizing speaker's hand configurations

Все видеопоследовательности, которые поступают на вход представленной архитектуры E2Ev2, разбиваются на сегменты, состоящие из 30 кадров и частичным перекрытием в 40 % (12 кадров). В случаях, когда кадров недостаточно (конец видеопоследовательности), то недостающие кадры заполняются последним кадром. Также для уменьшения вычислительных затрат все цветные изображения преобразуются в градации серого и нормализуются до 192 × 192 пикселов. Увеличение общего контраста изображений достигается путем выравнивания гистограммы яркости пикселов[1].

Возможный риск переобучения нейросетевой модели E2Ev2 минимизируется за счет процесса аугментации данных MixUp (применяется только к 60 % изображений и их меткам) [27, 28]. Параметр объединения двух изображений и бинарных векторов их меток (классов конфигураций рук) изменяется от 30 до 70 %, при соблюдении условия нулевой прозрачности для генерируемого изображения (сумма всегда равна 100 %). С математической точки зрения, процесс MixUp может быть представлен в виде:

$$\tilde{x} = \lambda \times x_1 + (1 - \lambda) \times x_2,$$
$$\mathbf{y} = \lambda \times \mathbf{y}_1 + (1 - \lambda) \times \mathbf{y}_2,$$

где $\tilde{x}$ и $\mathbf{y}$ — сгенерированное новое изображение и вектор меток; λ — параметр объединения двух изображений или бинарных векторов; x_1 и x_2 — первое и второе исходные случайные изображения; $\mathbf{y}_1$ и $\mathbf{y}_2$ — первый и второй исходные бинарные вектора меток, которые всегда соответствуют выбранным случайным изображениям x_1 и x_2.

Отметим, что для оставшихся 40 % бинарных векторов меток, которые не попали в выборку для процесса аугментации данных MixUp, применяется техника их сглаживания (Label Smoothing, LS)[2], представленная в виде:

$$\mathbf{y} = (1 - \alpha) \times \mathbf{y}' + \alpha / K,$$

где α — параметр степени сглаживания бинарного вектора меток; $\mathbf{y}'$ — исходный бинарный вектор меток; K — количество классов (конфигураций рук диктора).

В результате формируется новый вектор размерностью, равной количеству конфигураций рук диктора (в данном случае всего 23 конфигурации), в котором значение, равное единице, заменяется на 0,75, а все оставшиеся 22 нулевых значения — на 0,25.

Для непосредственного процесса извлечения визуальных признаков из изображений с конфигурациями рук диктора использована обученная с нуля модифицированная нейросетевая модель с архитектурой 3DResNet-18[3], в которую был добавлен модуль внимания SA [29] и отключен последний слой. Таким образом, каждый сегмент видеопоследовательности состоит из набора визуальных признаков, размерность которых равна 30 × 512, где 30 — это длина последовательности кадров сегмента, а 512 — количество извлеченных визуальных признаков.

Извлеченные визуальные признаки являются входными данными для нейросетевой модели LSTM. В свою очередь, полносвязный слой (Fully Connected, FC) вместе с функцией активации (Softmax) с количеством нейронов, равных 23, формирует вектор вероятностных значений, сумма которых равна 1. Индекс правильно предсказанной конфигурации руки диктора имеет наибольшее вероятностное значение.

Заметим, что все представленные значения архитектуры нейросетевой модели E2Ev2, необходимые

[1] Random Equalize [Электронный ресурс]. Режим доступа: https://pytorch.org/vision/stable/generated/torchvision.transforms.RandomEqualize.html, свободный. Яз. англ. (дата обращения: 26.03.2022).

[2] Label Smoothing [Электронный ресурс]. Режим доступа: https://paperswithcode.com/method/label-smoothing, свободный. Яз. англ. (дата обращения: 26.03.2022).

[3] 3D ResNet [Электронный ресурс]. Режим доступа: https://pytorch.org/hub/facebookresearch_pytorchvideo_resnet, свободный. Яз. англ. (дата обращения: 26.03.2022).

для ее обучения, подобраны на основе эмпирических экспериментов.

Эксперименты и результаты

Метод с представленными улучшениями был апробирован на данных (18 одноручных жестов, по аналогии с [20]) из многомодального корпуса элементов русского ЖЯ TheRuSLan [30, 31]. Процесс обучения производился на 11 дикторах, а тестирование — на мужчине и женщине. В роли планировщика скорости обучения использован Cosine Annealing с холодным перезапуском (Cosine Annealing Warm Restarts, Cosine WR)[1], значения которого варьировались от 0,0001 до 0,001. График измерение скорости обучения (*lr*, отн. ед.) нейросетевой модели относительно количества эпох, представлен на рис. 4.

Из рис. 4 видно, что максимальное количество эпох установлено в значение 100. Если на протяжении 10 эпох улучшение показателя точности не наблюдалось, то процесс машинного обучения останавливался, и тогда лучшим считался результат, полученный за все время обучения нейросетевой модели.

Программная реализация метода выполнена на языке программирования Python v.3.9. Для машинного обучения использована платформа с открытым исходным кодом PyTorch v.1.11.0 в связке с ее расширением в виде модуля TorchVision v.0.12.0.

Сравнение предложенной архитектуры E2Ev2 и обученной на ее основе нейросетевой модели для распознавания конфигураций рук диктора происходит с ранее обученными архитектурами 2D моделей сверточных нейросетей (табл. 1), которые были представлены в работе [24].

Для многомодального машинного сурдоперевода использована глубокая нейросеть LSTM, архитектура и процесс обучения которой подробно описаны в работе [20]. На вход нейросети LSTM направлен вектор признаков, состоящий из 5 пространственно-временных визуальных характеристик жестов. Объединение характеристик в один вектор осуществлен путем простой их конкатенации. В табл. 2 представлены сравнительные результаты итогового процесса распознавания 18 одноручных жестов русского ЖЯ из многомодального корпуса TheRuSLan [30, 31].

Сравнение представленного метода многомодального видеоанализа движений рук для автоматического распознавания ручных жестов и элементов ЖЯ с описанными улучшениями (E2Ev2) произведено с реализованными ранее методами [20], а именно: генетическим программированием на основе графов (RGGP); моделей эллиптического распределения (EDS); многопоточной рекуррентной нейросетью (MRNN); рекуррентной 3D сверточной нейросетью (R3DCNN); многомерным методом на основе сверточных нейросетей (MultiD-CNN) и первой версией предложенного метода на основе интегральных нейросетевых моделей E2Ev1.

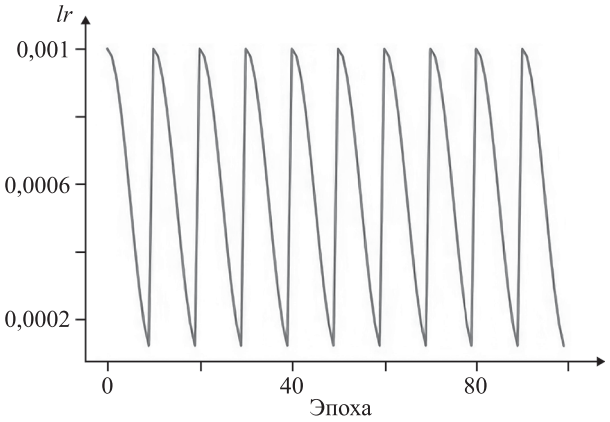

Рис. 4. График изменения скорости обучения (*lr*) нейросетевой модели относительно количества эпох

Fig. 4. Graph of the learning rate (*lr*) of a neural network model versus the number of epochs

[1] Cosine Annealing Warm Restarts [Электронный ресурс]. Режим доступа: https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.CosineAnnealingWarmRestarts.html, свободный. Яз. англ. (дата обращения: 26.03.2022).

Таблица 1. Сравнение точности распознавания конфигураций рук для различных архитектур нейросетевых моделей

Table 1. Comparison of hand configuration recognition accuracy for different architectures of neural network models

Размер входных изображений (ширина × высота × глубина)	Нейросетевая архитектура	Ссылка	Точность, %
224 × 224 × 3	EfficientNetB0	[24]	70,32
224 × 224 × 3	MobileNetV2		72,47
224 × 224 × 3	VGG19		73,19
299 × 299 × 3	InceptionV3		75,92
224 × 224 × 3	ResNet152V2		76,11
224 × 224 × 3	DenseNet169		76,54
299 × 299 × 3	Xception		80,03
299 × 299 × 3	InceptionResNetV2		81,44
331 × 331 × 3	NASNetLarge		84,44
528 × 528 × 3	EfficientNetB7		87,01
192 × 192 × 1	Предложенная архитектура E2Ev2 (рис. 3)	—	**94,78**

Таблица 2. Сравнительная таблица методов распознавания жестов
Table 2. Comparative table of gesture recognition methods

Метод распознавания	Модальности видеоданных	Ссылка	Точность, %
Restricted Graph-Based Genetic Programming (RGGP)	RGB	[20]	69,74
	Карта глубины		53,07
	RGB + карта глубины		74,28
Elliptical Density Shape Model (EDS)			77,43
Multi-Stream Recurrent Neural Network (MRNN)	RGB		68,23
	Карта глубины		73,54
	RGB + карта глубины		79,98
Recurrent 3D Convolutional Neural Network (R3DCNN)			84,67
Multi-Dimensional Convolutional Neural Networks (MultiD-CNN)			87,38
E2Ev1			88,92
Предложенный метод (E2Ev2)		—	**91,14**

Из табл. 1 и 2 следует, что за счет расширенного списка из информативных пространственно-временных визуальных характеристик жеста можно получить увеличение точности распознавания жестов.

Заключение

В работе получены результаты, которые могут найти широкое применение в областях научной и социальной деятельности, а также в сферах социального обслуживания, медицины, образования, робототехники, в центрах обслуживания населения и для взаимодействия с людьми в различных чрезвычайных ситуациях. Кроме того, за последние годы все более широкое распространение находят ассистивные и социальные роботы, ориентированные на выполнение определенных прикладных задач, а также естественного человеко-машинного взаимодействия. Для взаимодействия с такими роботами классических графических и сенсорных пользовательских интерфейсов недостаточно. Помимо них, необходимы интуитивные и естественные для человека интерфейсы (например, на основе жестовой модальности). В свою очередь, интеллектуальный анализ и оценка важности влияния гендерных и возрастных характеристик диктора на машинный сурдоперевод позволят максимально эффективно распознавать отдельные жесты рук, а также элементы жестовых языков.

Предложенная нейросетевая модель E2Ev2 имеет значительный инновационный потенциал, так как позволяет повысить точность и робастность машинного сурдоперевода и, в свою очередь, естественность человеко-машинного взаимодействия в целом.

Литература

1. Ryumin D., Kagirov I., Ivanko D., Axyonov A., Karpov A. Automatic detection and recognition of 3D manual gestures for human-machine interaction // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2019. V. 42. N 2/W12. P. 179–183. https://doi.org/10.5194/isprs-archives-XLII-2-W12-179-2019
2. Карпов А.А., Юсупов Р.М. Многомодальные интерфейсы человеко-машинного взаимодействия // Вестник Российской академии наук. 2018. Т. 88. № 2. С. 146–155. https://doi.org/10.7868/S0869587318020056
3. Ryumin D., Karpov A.A. Towards automatic recognition of sign language gestures using kinect 2.0 // Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017. V. 10278. P. 89–101. https://doi.org/10.1007/978-3-319-58703-5_7
4. Рюмин Д. Метод автоматического видеоанализа движений рук и распознавания жестов в человеко-машинных интерфейсах // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 4. С. 525–531. https://doi.org/10.17586/2226-1494-2020-20-4-525-531
5. Томская М.В., Маслова Л.Н. Гендерные исследования в отечественной лингвистике // Русский язык в современном обществе: функциональные и статусные характеристики. М., 2005. С. 102–130.
6. Carli L., LaFleur S., Loeber C. Nonverbal behavior, gender, and influence // Journal of Personality and Social Psychology. 1995. V. 68. N 6. P. 1030–1041. https://doi.org/10.1037/0022-3514.68.6.1030

References

1. Ryumin D., Kagirov I., Ivanko D., Axyonov A., Karpov A. Automatic detection and recognition of 3D manual gestures for human-machine interaction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2019, vol. 42, no. 2/W12, pp. 179–183. https://doi.org/10.5194/isprs-archives-XLII-2-W12-179-2019
2. Karpov A.A., Yusupov R.M. Multimodal interfaces of human–computer interaction. *Herald of the Russian Academy of Sciences*, 2018, vol. 88, no. 1, pp. 67–74. https://doi.org/10.1134/S1019331618010094
3. Ryumin D., Karpov A.A. Towards automatic recognition of sign language gestures using kinect 2.0. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10278, pp. 89–101. https://doi.org/10.1007/978-3-319-58703-5_7
4. Ryumin D. Automated hand detection method for tasks of gesture recognition in human-machine interfaces. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 4, pp. 525–531 (in Russian). https://doi.org/10.17586/2226-1494-2020-20-4-525-531
5. Tomskaia M.V., Maslova L.N. Gender research in national linguistics. *Russian language in the modern society: functional and status characteristics*. Moscow, 2005, pp. 102–130. (in Russian)
6. Carli L., LaFleur S., Loeber C. Nonverbal behavior, gender, and influence. *Journal of Personality and Social Psychology*, 1995, vol. 68, no. 6, pp. 1030–1041. https://doi.org/10.1037/0022-3514.68.6.1030

7. Iriskhanova O., Cienki A. The semiotics of gestures in cognitive linguistics: Contribution and challenges // Вопросы конгинивной лингвистики. 2018. Т. 4. С. 25–36. https://doi.org/10.20916/1812-3228-2018-4-25-36
8. Masson-Carro I., Goudbeek M., Krahmer E. Coming of age in gesture: A comparative study of gesturing and pantomiming in older children and adults // Proc. of the 4th Gesture and Speech in Interaction Conference (GESPIN). 2015. P. 1–7.
9. Reviewed Work: Sign language structure: An outline of the visual communication systems of the American deaf by William C. Stokoe, Jr. // Language. 1961. V. 37. N 2. P. 269–271. https://doi.org/10.2307/410856
10. Димскис Л.С. Изучаем жестовый язык. М.: Издательский центр «Академия», 2002. 128 с.
11. Sonkusare J., Chopade N., Sor R., Tade S. A review on hand gesture recognition system // Proc. of the 1st International Conference on Computing, Communication, Control and Automation. 2015. P. 790–794. https://doi.org/10.1109/ICCUBEA.2015.158
12. De Smedt Q., Wannous H., Vandeborre J. Heterogeneous hand gesture recognition using 3D dynamic skeletal data // Computer Vision and Image Understanding. 2019. V. 181. P. 60–72. https://doi.org/10.1016/j.cviu.2019.01.008
13. Grif M., Prikhodko A., Bakaev M. Recognition of signs and movement epentheses in Russian Sign Language // Communications in Computer and Information Science. 2022. V. 1503. P. 67–82. https://doi.org/10.1007/978-3-030-93715-7_5
14. Гришина Е.А. Кольцо и щепоть: семантика соединенных пальцев в русской жестикуляции // Компьютерная лингвистика и интеллектуальные технологии. 2014. № 13. С. 182–202.
15. Zhang C., Yang X., Tian Y. Histogram of 3D Facets: A characteristic descriptor for hand gesture recognition // Proc. of the 10th International Conference Automatic Face and Gesture Recognition (FG). 2013. P. 6553754. https://doi.org/10.1109/FG.2013.6553754
16. Рюмин Д.А., Кагиров И.А. Подходы к автоматическому распознаванию жестовой информации: аппаратное обеспечение и методы // Пилотируемые полеты в космос. 2021. № 3(40). С. 82–99. https://doi.org/10.34131/MSF.21.3.82-99
17. Camgoz C.N., Hadfield S., Koller O., Bowden R. SubUNets: End-to-end hand shape and continuous sign language recognition // Proc. of the 16th International Conference on Computer Vision (ICCV). 2017. P. 3075–3084. https://doi.org/10.1109/ICCV.2017.332
18. Гриф М.Г., Королькова О.О., Приходько А.Л. Распознавание жестовой речи с учетом комбинаторных изменений жестов // Информатика: проблемы, методы, технологии: Материалы XXI Международной научно-технической конференции. 2021. С. 1387–1393.
19. Ryumin D., Kagirov I., Axyonov A., Pavlyuk N., Saveliev A., Kipyatkova I., Zelezny M., Mporas I., Karpov A. A multimodal user interface for an assistive robotic shopping cart // Electronics. 2020. V. 9. N 12. P. 1–25. https://doi.org/10.3390/electronics9122093
20. Axyonov A., Ryumin D., Kagirov I. Method of multi-modal video analysis of hand movements for automatic recognition of isolated signs of Russian sign language // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2021. V. 44. N 2/W1. P. 7–13. https://doi.org/10.5194/isprs-archives-XLIV-2-W1-2021-7-2021
21. Wu J., Zhang Y., Zhao X. A prototype-based generalized zero-shot learning framework for hand gesture recognition // Proc. of the 25th International Conference on Pattern Recognition (ICPR). 2021. P. 3435–3442. https://doi.org/10.1109/ICPR48806.2021.9412548
22. Voskou A., Panousis K.P., Kosmopoulos D., Metaxas D.N., Chatzis S. Stochastic transformer networks with linear competing units: Application to end-to-end SL translation // Proc. of the 18th International Conference on Computer Vision (ICPR). 2021. P. 11926–11935. https://doi.org/10.1109/ICCV48922.2021.01173
23. Jiang S., Sun B., Wang L., Bai Y., Li K., Fu Y. Skeleton aware multi-modal sign language recognition // Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 3408–3418. https://doi.org/10.1109/CVPRW53098.2021.00380
24. Рюмин Д. Модели и методы автоматического распознавания элементов русского жестового языка для человеко-машинного взаимодействия: диссертация на соискание ученой степени кандидата технических наук / Университет ИТМО. 2020. 352 с. [Электронный ресурс]. URL: http://fppo.ifmo.ru/dissertation/?number=246869, свободный. Яз. рус. (дата обращения: 26.03.2022).

7. Iriskhanova O., Cienki A. The semiotics of gestures in cognitive linguistics: Contribution and challenges. *Voprosy Kognitivnoy Lingvistiki*, 2018, vol. 4, pp. 25–36. https://doi.org/10.20916/1812-3228-2018-4-25-36
8. Masson-Carro I., Goudbeek M., Krahmer E. Coming of age in gesture: A comparative study of gesturing and pantomiming in older children and adults. *Proc. of the 4th Gesture and Speech in Interaction Conference (GESPIN)*, 2015, pp. 1–7.
9. Reviewed Work: Sign language structure: An outline of the visual communication systems of the American deaf by William C. Stokoe, Jr. *Language*, 1961, vol. 37, no. 2, pp. 269–271. https://doi.org/10.2307/410856
10. Dimskis L.S. *Learning Sign Language*. Moscow, Akademija Publ., 2002, 128 p. (in Russian)
11. Sonkusare J., Chopade N., Sor R., Tade S. A review on hand gesture recognition system. *Proc. of the 1st International Conference on Computing, Communication, Control and Automation*, 2015, pp. 790–794. https://doi.org/10.1109/ICCUBEA.2015.158
12. De Smedt Q., Wannous H., Vandeborre J. Heterogeneous hand gesture recognition using 3D dynamic skeletal data. *Computer Vision and Image Understanding*, 2019, vol. 181, pp. 60–72. https://doi.org/10.1016/j.cviu.2019.01.008
13. Grif M., Prikhodko A., Bakaev M. Recognition of signs and movement epentheses in Russian Sign Language. *Communications in Computer and Information Science*, 2022, vol. 1503, pp. 67–82. https://doi.org/10.1007/978-3-030-93715-7_5
14. Grishina E.A. Ring and grappolo: Fingertip connections in Russian gesticulation and their meanings. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2014, no. 13, pp. 182–202. (in Russian)
15. Zhang C., Yang X., Tian Y. Histogram of 3D Facets: A characteristic descriptor for hand gesture recognition. *Proc. of the 10th International Conference Automatic Face and Gesture Recognition (FG)*, 2013, pp. 6553754. https://doi.org/10.1109/FG.2013.6553754
16. Ryumin D.A., Kagirov I.A. Approaches to automatic gesture recognition: hardware and methods overview. *Manned Spaceflight*, 2021, no. 3(40), pp. 82–99. (in Russian). https://doi.org/10.34131/MSF.21.3.82-99
17. Camgoz C.N., Hadfield S., Koller O., Bowden R. SubUNets: End-to-end hand shape and continuous sign language recognition. *Proc. of the 16th International Conference on Computer Vision (ICCV)*, 2017, pp. 3075–3084. https://doi.org/10.1109/ICCV.2017.332
18. Grif M.G., Korolkova O.O., Prikhodko A.L. Sign speech recognition taking into account combinatorial changes of gestures. *Problems, Methods, and Technologies in the Computer Science. Proceedings of the XXI International Scientific and Technical Conference*, 2021, pp. 1387–1393. (in Russian)
19. Ryumin D., Kagirov I., Axyonov A., Pavlyuk N., Saveliev A., Kipyatkova I., Zelezny M., Mporas I., Karpov A. A multimodal user interface for an assistive robotic shopping cart. *Electronics*, 2020, vol. 9, no. 12, pp. 1–25. https://doi.org/10.3390/electronics9122093
20. Axyonov A., Ryumin D., Kagirov I. Method of multi-modal video analysis of hand movements for automatic recognition of isolated signs of Russian sign language. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2021, vol. 44, no. 2/W1, pp. 7–13. https://doi.org/10.5194/isprs-archives-XLIV-2-W1-2021-7-2021
21. Wu J., Zhang Y., Zhao X. A prototype-based generalized zero-shot learning framework for hand gesture recognition. *Proc. of the 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3435–3442. https://doi.org/10.1109/ICPR48806.2021.9412548
22. Voskou A., Panousis K.P., Kosmopoulos D., Metaxas D.N., Chatzis S. Stochastic transformer networks with linear competing units: Application to end-to-end SL translation. *Proc. of the 18th International Conference on Computer Vision (ICPR)*, 2021, pp. 11926–11935. https://doi.org/10.1109/ICCV48922.2021.01173
23. Jiang S., Sun B., Wang L., Bai Y., Li K., Fu Y. Skeleton aware multi-modal sign language recognition. *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3408–3418. https://doi.org/10.1109/CVPRW53098.2021.00380
24. Ryumin D. *Models and Methods for Automatic Recognition of Russian Sign Language Elements for Human-Machine Interaction.* Academic dissertation candidate of engineering. ITMO University, 2020, 352 p. Available at: http://fppo.ifmo.ru/dissertation/?number=246869 (accessed 26.03.2022). (in Russian)
25. Winata G.I., Kampman O.P., Fung P. Attention-based LSTM for psychological stress detection from spoken language using distant

25. Winata G.I., Kampman O.P., Fung P. Attention-based LSTM for psychological stress detection from spoken language using distant supervision // Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. P. 6204–6208. https://doi.org/10.1109/ICASSP.2018.8461990
26. Serengil S.I., Ozpinar A. LightFace: A Hybrid deep face recognition framework // Proc. of the Innovations in Intelligent Systems and Applications Conference (ASYU). 2020. P. 9259802. https://doi.org/10.1109/ASYU50717.2020.9259802
27. Zhang H., Cisse M., Dauphin Y.N., Lopez-Paz D. MixUp: Beyond empirical risk minimization // Proc. of the 6th International Conference on Learning Representations (ICLR). 2018.
28. Dresvyanskiy D., Ryumina E., Kaya H., Markitantov M., Karpov A., Minker W. End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild // Multimodal Technologies and Interaction. 2022. V. 6. N 2. P. 11. https://doi.org/10.3390/mti6020011
29. Zhong Z., Lin Z.Q., Bidart R., Hu X., Daya I.B., Li Z., Zheng W., Li J., Wong A. Squeeze-and-attention networks for semantic segmentation // Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 13062–13071. https://doi.org/10.1109/cvpr42600.2020.01308
30. Kagirov I., Ivanko D., Ryumin D., Axyonov A., Karpov A. TheRuSLan: Database of Russian Sign Language // Proc. of the 12th Conference on Language Resources and Evaluation (LREC). 2020. P. 6079–6085.
31. Кагиров И.А., Рюмин Д.А., Аксёнов А.А., Карпов А.А. Мультимедийная база данных жестов русского жестового языка в трехмерном формате // Вопросы языкознания. 2020. № 1. С. 104–123. https://doi.org/10.31857/S0373658X0008302-1

supervision. *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6204–6208. https://doi.org/10.1109/ICASSP.2018.8461990
26. Serengil S.I., Ozpinar A. LightFace: A Hybrid deep face recognition framework. *Proc. of the Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 9259802. https://doi.org/10.1109/ASYU50717.2020.9259802
27. Zhang H., Cisse M., Dauphin Y.N., Lopez-Paz D. MixUp: Beyond empirical risk minimization. *Proc. of the 6th International Conference on Learning Representations (ICLR)*, 2018.
28. Dresvyanskiy D., Ryumina E., Kaya H., Markitantov M., Karpov A., Minker W. End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild. *Multimodal Technologies and Interaction*, 2022, vol. 6, no. 2, pp. 11. https://doi.org/10.3390/mti6020011
29. Zhong Z., Lin Z.Q., Bidart R., Hu X., Daya I.B., Li Z., Zheng W., Li J., Wong A. Squeeze-and-attention networks for semantic segmentation. *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13062–13071. https://doi.org/10.1109/cvpr42600.2020.01308
30. Kagirov I., Ivanko D., Ryumin D., Axyonov A., Karpov A. TheRuSLan: Database of Russian Sign Language. *Proc. of the 12th Conference on Language Resources and Evaluation (LREC)*, 2020, pp. 6079–6085.
31. Kagirov I.A., Ryumin D.A., Axyonov A.A., Karpov A.A. Multimedia database of russian sign language items in 3D. *Voprosy Jazykoznanija*, 2020, no. 1, pp. 104–123. (in Russian). https://doi.org/10.31857/S0373658X0008302-1

Авторы

Аксёнов Александр Александрович — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57203963345, https://orcid.org/0000-0002-7479-2851, axyonov.a@iias.spb.su

Кагиров Ильдар Амирович — научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 25121369400, https://orcid.org/0000-0003-1196-1117, kagirov@iias.spb.su

Рюмин Дмитрий Александрович — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57191960214, https://orcid.org/0000-0002-7935-0569, ryumin.d@iias.spb.su

Authors

Alexandr A. Axyonov — Junior Researcher, Saint Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57203963345, https://orcid.org/0000-0002-7479-2851, axyonov.a@iias.spb.su

Ildar A. Kagirov — Scientific Researcher, Saint Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 25121369400, https://orcid.org/0000-0003-1196-1117, kagirov@iias.spb.su

Dmitry A. Ryumin — PhD, Senior Researcher, Saint Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57191960214, https://orcid.org/0000-0002-7935-0569, ryumin.d@iias.spb.su

Статья поступила в редакцию 30.03.2022
Одобрена после рецензирования 19.04.2022
Принята к печати 30.05.2022

Received 30.03.2022
Approved after reviewing 19.04.2022
Accepted 30.05.2022