НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2023 Том 23 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2023 Vol. 23 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2023-23-5-980-988
УДК 004.932.72'1, 004.852

Сегментация жестов слов на видео жестового языка

Данг Хань[1]✉, Игорь Александрович Бессмертный[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] dangkhanhmta.2020@gmail.com✉, https://orcid.org/0009-0009-5882-7653
[2] bessmertny@itmo.ru, https://orcid.org/0000-0001-6711-6399

Аннотация

Введение. Несмотря на широкое распространение средств автоматического распознавания речи и сопровождения видео субтитрами, язык жестов по-прежнему является ключевым средством коммуникации для людей с нарушениями слуха. Важной задачей в процессе автоматического распознавания жестового языка является сегментация видео на фрагменты, соответствующие отдельным словам. В отличие от известных методов сегментации слов жестового языка, предложен подход, не требующий использования сенсоров (акселерометров). **Метод.** Для сегментации видео на слова использована оценка динамики изображения, а граница между словами определена с помощью порогового значения. На практике в кадре, кроме диктора, могут присутствовать сторонние движущиеся объекты, которые создают шум. В связи с этим предложено оценить динамику по среднему изменению от кадра к кадру евклидова расстояния между координатными характеристиками кисти, предплечья, глаз и рта. Вычисление координатных характеристик рук и головы осуществлено с помощью библиотеки MediaPipe. **Основные результаты.** Разработанный алгоритм апробирован для жестового вьетнамского языка на открытом наборе из 4364 видео, собранном во Вьетнамском центре обучения языку жестов. Алгоритм продемонстрировал высокую точность, сопоставимую с ручной сегментацией видео оператором, и низкую ресурсоемкость, что позволяет его использовать при автоматическом распознавании жестов в реальном времени. **Обсуждение.** Выполненные эксперименты показали, что задача сегментации языка жестов в отличие от известных методов может быть эффективно решена без использования сенсоров. Как и другие методы сегментации жестов, предложенный алгоритм неудовлетворительно работает при высокой скорости жестового языка, когда имеет место наложение слов друг на друга. Данная проблема является предметом дальнейших исследований.

Ключевые слова

язык жестов, сегментация жестов слов, MediaPipe, LSTM, метод порогового значения, распознавание языка жестов

Ссылка для цитирования: Хань Д., Бессмертный И.А. Сегментация жестов слов на видео жестового языка // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 5. С. 980–988. doi: 10.17586/2226-1494-2023-23-5-980-988

Segmentation of word gestures in sign language video

Dang Khanh[1]✉, Igor A. Bessmertny[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] dangkhanhmta.2020@gmail.com✉, https://orcid.org/0009-0009-5882-7653
[2] bessmertny@itmo.ru, https://orcid.org/0000-0001-6711-6399

Abstract

Despite the widespread use of automatic speech recognition and video subtitles, sign language is still a significant communication channel for people with hearing impairments. An important task in the process of automatic recognition of sign language is the segmentation of video into fragments corresponding to individual words. In contrast to the known methods of segmentation of sign language words, the paper proposes an approach that does not require the use of sensors (accelerometers). To segment the video into words in this study, an assessment of the dynamics of the image

is used, and the boundary between words is determined using a threshold value. Since in addition to the speaker, there may be other moving objects in the frame that create noise, the dynamics in the work is estimated by the average change from frame to frame of the Euclidean distance between the coordinate characteristics of the hand, forearm, eyes and mouth. The calculation of the coordinate characteristics of the hands and head is carried out using the MediaPipe library. The developed algorithm was tested for the Vietnamese sign language on an open set of 4364 videos collected at the Vietnamese Sign Language Training Center, and demonstrated accuracy comparable to manual segmentation of video by an operator and low resource consumption, which will allow using the algorithm for automatic gesture recognition in real time. The experiments have shown that the task of segmentation of sign language, unlike the known methods, can be effectively solved without the use of sensors. Like other methods of gesture segmentation, the proposed algorithm does not work satisfactorily at a high speed of sign language when words overlap each other. This problem is the subject of further research.

Keywords
sign language, word gesture segmentation, MediaPipe, LSTM, thresholding method, sign language recognition

For citation: Dang Khanh, Bessmertny I.A. Segmentation of word gestures in sign language video. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 5, pp. 980–988 (in Russian). doi: 10.17586/2226-1494-2023-23-5-980-988

Введение

Каждая страна и этническая группа имеют свои отличительные культурные и языковые особенности, которые также влияют на систему жестового языка каждой страны. Исследования по распознаванию жестов в языке жестов способствуют устранению коммуникативного разрыва между людьми с нарушениями слуха в повседневной жизни. Распознавание языка жестов является очень актуальной темой, привлекающей внимание исследователей, особенно в области искусственного интеллекта и машинного обучения. Исследования в области коммуникации для людей с нарушениями слуха развиваются от простого к сложному, например: проблемы распознавания жестового языка по статическим изображениям, распознавание жестового языка на видео, анализ семантики и распознавание жестового языка, перевод жестового языка на родной язык. На простом уровне распознавание языка жестов в статических изображениях и видео достигло высокой точности. Однако на более высоком уровне, приближающемся к машинному переводу жестового языка, необходимо решить множество трудных и сложных проблем, таких как неоднородности жестового языка между странами и недостаточности наборов данных, сложность при анализе семантики и грамматики жестового языка. Проблема эффективного распознавания жестов до сих пор не решена из-за серьезных различий в семантико-синтаксической структуре любых жестов, вследствие чего пока невозможно выполнять однозначный перевод с жестового языка, например текстовое представление. По этой причине полностью действующих автоматизированных моделей и методов для систем распознавания множества, статических и что важно динамических жестов на данный момент не существует. Для создания оптимальных моделей необходимо производить глубокий семантический анализ, а это на данный момент возможно только на поверхностном уровне из-за несовершенства алгоритмов анализа текстов, баз знаний и т. д. Для вьетнамского языка жестов основными факторами являются жесты рук и изменения лица, которые составляют жесты языка жестов [1]. Жесты, представляющие слова, также разделяются паузами. Грамматика вьетнамского языка жестов отличается от письменного языка наличием изменения порядка слов. Для решения проблемы машинного перевода жестового языка необходимо перейти к распознаванию жестов языка жестов на видео, а затем использовать языковую модель для семантического и грамматического анализов. Отметим, что первый и самый важный этап — решение проблемы сегментации жестов слов на видео жестового языка.

Постановка задачи исследования

Проблему распознавания жестов на видео исследователи делят на две популярные группы: в статической и в динамической формах. Исследования распознавание движения на видео в динамической форме является более сложной из-за особенностей временных рядов и разнообразия жестов.

В работе [2] выполнено распознавание жестов языка жестов на уровне слова на видео. На уровне слов каждое отдельное видео представляет собой отдельное слово. Определен временной интервал для жеста на видео с помощью определения разницы между первым и последним кадрами в отсутствие повторяющихся жестов. В случае наличия повторяющихся жестов данные паттерны вручную были отмечены и сохранены для повышения точности обучающей модели. При этом обучающий набор данных учтен в двух формах: 2D-соединения человеческой позы и Holistic визуальный подход на основе видео. Результаты модели оценены на множестве наборов данных с точностью 62,63 % в топ-10[1] точности по статистическим результатам исследования распознавания жестового языка на наборе данных WLASL-2000 (2000 слов/лоск). Этот метод оказался эффективным для небольших наборов данных. Но в случае большого набора данных и разнообразии жестов необходимо учитывать сложность и время вычислений метода.

В [3] изучено распознавание жестов языка жестов на уровне предложения с использованием метода Visual Alignment Constraint. Основная идея метода заключается в том, что движения языка жестов имеют определенную временную логику, например движение, которое

[1] [Электронный ресурс]. Режим доступа: https://paperswithcode.com/sota/sign-language-recognition-on-wlasl-2000 (дата обращения: 12.09.2023).

начинается и заканчивается в определенное время. Исходя из этого, можно использовать ограничение на выравнивание изображений, чтобы найти соответствие между предсказанными и правильными метками путем выравнивания начального и конечного времени движений языка жестов. Авторы обучили нейросетевую модель непрерывного распознавания жестов языка жестов, такую как BiLSTM, на наборах данных RWTH-PHOENIX-Weather-2014 (PHOENIX14) и китайском языке жестов (Chinese Sign Language, CSL). Результаты показали улучшение точности при комбинации с методом Visual Alignment Constraint.

В работах [4–6] исследования также основаны на распознавании языка жестов на видео и достигли обнадеживающих результатов. Однако эти исследования остановились только на распознавании языка жестов на видео в статической форме. Это связано с тем, что данные, которые должны быть протестированы, представляют собой только изображения особенностей жеста или короткое видео, описывающее жест.

Перечисленные методы не могут дать высокую эффективность при применении к модели машинного перевода языка жестов в реальном времени. Далее рассмотрим модель распознавания языка жестов на динамическом уровне.

Проблема распознавания языка жестов на динамическом уровне характеризуется следующим образом. Входные данные — видео заданной длины или неизвестной до момента окончания. На момент времени t видео имеет размер n кадров и выражает некоторое содержание на языке жестов. Задача — распознать жесты языка жестов на видео и вывести содержание

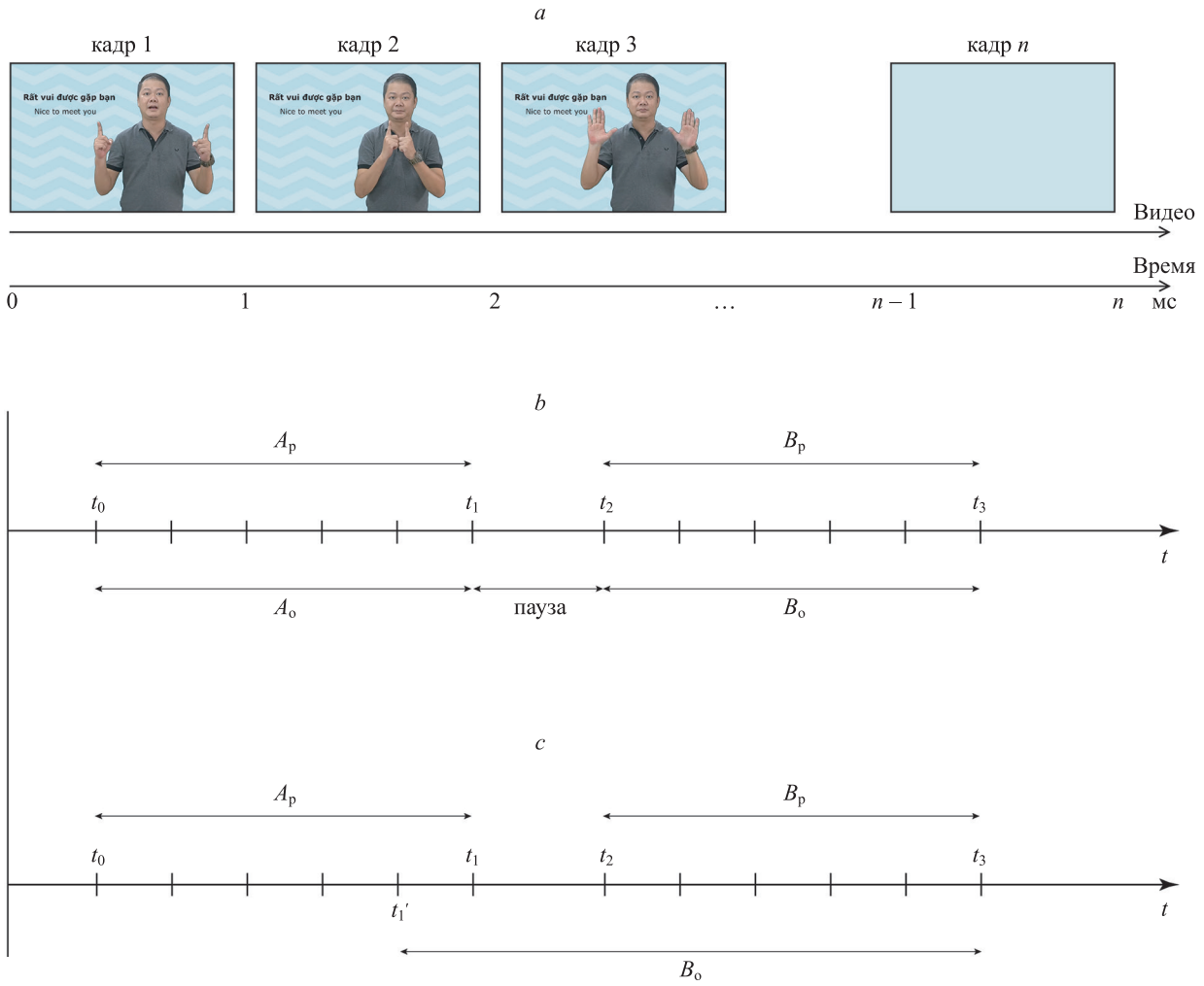

Рис. 1. Задача распознавания языка жестов во времени: пример языка жестов во временных рядах (*a*); жесты *A* и *B* сегментированы (*b*) и не сегментированы (*c*).

A_p — описание жеста *A*, который выполняется в течение периода времени с t_0 по t_1; B_p — описание жеста *B* — с t_2 по t_3; A_o, B_o — описание жестов *A* и *B*, которые программа прогнозирования может распознать и обработать, в случае явной сегментации (*b*) и без сегментации (*c*); t_0, t_1, t_1', t_2, t_3 — точки времени

Fig. 1. The task of recognizing sign language in time: an example of sign language in time series (*a*); gestures *A* and *B* are segmented (*b*) and not segmented (*c*).

A_p is a description of gesture *A*, which actually occurs during the time period from t_0 to t_1. B_p is a description of the gesture *B*, which actually occurs during the time period from t_2 to t_3. A_o, B_o is a description of gestures *A* and *B* that the program can recognize and process, in the case of explicit segmentation (*b*) and without segmentation (*c*); t_0, t_1, t_1', t_2, t_3 are time points

на родном языке во временной последовательности. Эта задача рассматривается как проблема машинного перевода языка жестов.

Основные методы распознавания языка жестов: анализ данных сенсоров и технологии компьютерного зрения. Оба метода имеют свои преимущества и недостатки. Метод распознавания языка жестов путем анализа данных сенсоров обеспечивает быстрый и точный сбор данных, но его недостатком является то, что коммуникатор должен использовать устройство с датчиком, например, перчатки или виртуальные руки. Метод распознавания языка жестов с помощью технологии компьютерного зрения изучен более тщательно из-за своей точности и легкости развертывания. Однако оба метода имеют трудности из-за разнообразия и сложности жестов в языке жестов, для построения модели распознавания с высокой точностью требуется огромный набор данных и компьютер с высокой скоростью обработки. Подчеркнем, что текущие исследования в основном достигают хороших результатов только при распознавании языка жестов в коротких видео или на статических изображениях. Традиционные модели нейронных сетей, такие как рекуррентная нейронная сеть (Recurrent Neural Network, RNN) и долгая краткосрочная память (Long Short-Term Memory, LSTM) используются в качестве популярных моделей для обработки данных временных рядов. Выполним анализ и сравнение распознавание языка жестов с нейронной сетевой моделью LSTM для двух случаев: случай 1 (рис. 1, *b*) — данные разбиты на сегменты между жестами, случай 2 (рис. 1, *c*) — данные не разбиты на сегменты.

Обозначим вектор $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ как входные данные для модели LSTM (рис. 2). $x_1, x_2, ..., x_n$ — вектора извлечения характеристик для кадров 1, 2, ..., n. На момент времени t фильтр забывания f_t получит два входных значения h_{t-1} и x_t.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f).$$

Через функцию сигмоидного преобразования фильтр забывания f_t имеет значения 0 и 1. Эти значения обозначают степень сохранения предыдущей характеристики x_{t-1}. Величина f_t пропорциональна способности сохранять информацию. Когда f_t равен 1, предыдущая характеристика полностью сохраняется, а когда он равен 0, вся старая информация забывается.

Чтобы обновить новое состояние для текущей входной характеристики, вычислим значение вектора контекста — C_t в моменте t следующим образом:

$$C_t = f_t C_{t-1} + i_t \hat{C}_t,$$

где $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$ — входной шлюз (input gate); $\hat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$ — скрытый слой с функцией tanh; C_{t-1} — вектор контекста в момент $t - 1$.

Рассчитаем выход для текущей характеристики h_t по формуле:

$$h_t = o_t \tanh(C_t),$$

где $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$ — выходной шлюз (output gate).

Так как параметры W, b обучаются для оптимизации процесса прогнозирования выхода в соответствии с входной характеристикой, теоретически сложно четко продемонстрировать эффективность распознавания жестов при разграничении между жестами по сравнению с перекрывающимися жестами. Рассмотрим некоторые проблемы более подробно.

Как показано на рис. 1, описываются жесты *A* и *B* в двух случаях.

В случае 1 существует ясное сегментирование границы между жестами *A* и *B*. Тогда кадры паузы будут игнорироваться и не включаться в обработку, поэтому уменьшается шум в модели. Более того, модель LSTM может определять начало и конец жеста. В этом случае между жестами *A* и *B* существует четкое разграничение. Затем кадры паузы не будут обрабатываться, что минимизирует шум в модели. С помощью параметров обобщения жестов во временных рядах модель может дать точные результаты предсказания.

В случае 2, когда жесты накладываются друг на друга, кадры интервала паузы подаются в модель для обработки, уменьшая или увеличивая сохранение устаревшей информации, например программа распознает язык жестов по видео с включенными жестами *A* и *B*. Когда временная граница жестов не может быть определена (рис. 1, *c*), для распознавания жеста *A* включаются кадры жеста *B*. Шум, генерируемый кадрами *B*, снижает точность задачи распознавания языка жестов. Кроме того, предположим, что некоторые кадры жеста *A* похожи на кадры жеста *B*, тогда результат модели может выводить жест *B* и вместо этого правильным результатом будет жест *A*. Исследования [7–9] также показали эффективность сегментации жестов на распознавание модели.

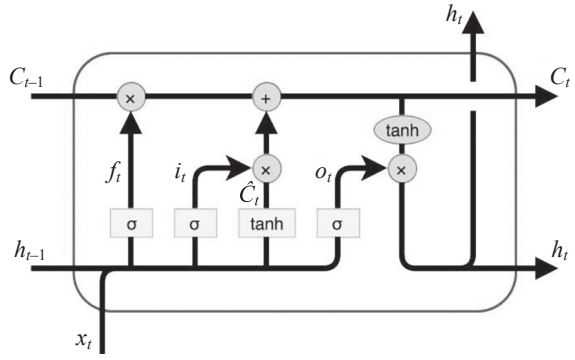

Рис. 2. Структура ячейки памяти нейронной сети LSTM

Fig. 2. The structure of the memory cell of the LSTM neural network

Релевантные работы

Сущность проблемы распознавания языка жестов в видео также является частным случаем проблемы распознавания действий в видео. Разница в том, что язык жестов в основном зависит от жестикуляции рук и некоторых частей лица, таких как рот и глаза. В работе [7] предложен унифицированный метод сегментации

действий с использованием временной сверточной сети (Temporal Convolutional Networks, TCN) для изучения временных взаимосвязей между действиями в ряду данных видео или датчика. Метод протестирован на трех общедоступных наборах данных: University of Dundee 50 Salads, JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) и Georgia Tech Egocentric Activities (GTEA). Каждый набор данных имеет свой метод сбора данных, включая данные видео и акселерометра, но все методы используются для сегментации действия с помощью TCN. Полученные результаты показали, что предложенный метод достиг наилучших результатов по всем наборам данных и превзошел другие сравниваемые методы. В [8] рассмотрен метод обнаружения непрерывного действия на видео с использованием метода динамического искажения кадра (Dynamic Frame Warping, DFW). DFW — метод согласования тестовой последовательности с последовательностью образцов, позволяющий обнаруживать действие в непрерывной видеопоследовательности. Метод включает в себя два расширения DFW: однопроходный и двухпроходный. Данные расширения позволяют распознавать одновременные действия с сегментами и основаны на методах, используемых при непрерывном распознавании речи, и ранее не применялись при непрерывном распознавании действий на видео. Авторы оценили предложенный метод на недавно выпущенном наборе данных RAVEL, а также на двух популярных наборах данных в action recognition, Hollywood-1 и Hollywood-2. Достигнутые результаты подтвердили, что метод обеспечивает высокую точность распознавания как изолированных, так и непрерывных действий и превосходит некоторые недавно опубликованные методы. В работе [7] отмечена проблема непрерывного распознавания действий, которая требует одновременного выполнения классификации и сегментации, и предложен новый метод для решения этой проблемы. Разработанный метод оценен на многих наборах данных и продемонстрировал высокую точность как для распознавания изолированных, так и непрерывных действий, что является важным вкладом в область распознавания действий. В [10] предложена новая модель трансформера под названием ASFormer для решения задачи сегментации действий на видео. В модели использована архитектура кодер-декодер с блоками кодера, состоящими из уровня самоконтроля с несколькими входами и уровня свертки во времени затяжки. ASFormer также применяет предопределенный шаблон иерархического представления для сокращения вычислений и улучшения обучения модели. ASFormer оценен на трех общедоступных наборах данных: GTEA61, 501 и DHF1K. Результаты тестирования показали, что модель ASFormer достигла лучших результатов, чем другие передовые методы, особенно в тех случаях, когда сегменты действия короткие и расположены близко друг к другу.

В случае, если рассматривается скорость движений жестов на видео подобна интенсивности звука в речи, то задача сегментации жестов на видео аналогична задаче сегментации звука при обработке речи. Результаты исследований работ [11–13] показали, что при обработке речи четкая сегментация звука обеспечивает лучшее распознавание и семантический анализ для модели машинного обучения.

В [10, 14] выявлено, что эффективность сегментации жестов на видео будет способствовать лучшему распознаванию и анализу контекста на видео.

Сегментация жестов на видео на жестовом языке методом порогового значения

Проблема. Учитывая видео на языке жестов, в момент времени t видео содержит n кадров $\mathbf{F} = \{1, 2, \ldots, f_n\}$.

Задание. Определить моменты границы между жестами языка жестов. На рис. 3 представлен алгоритм определения моментов границы между жестами на видео на языке жестов.

Исследования по распознаванию языка жестов [2, 4–6, 15, 16] показали, что движения рук и изменения лица являются основными факторами распознавания языка жестов. Такие признаки являются базами для проведения сегментации жестов на языке жестов на

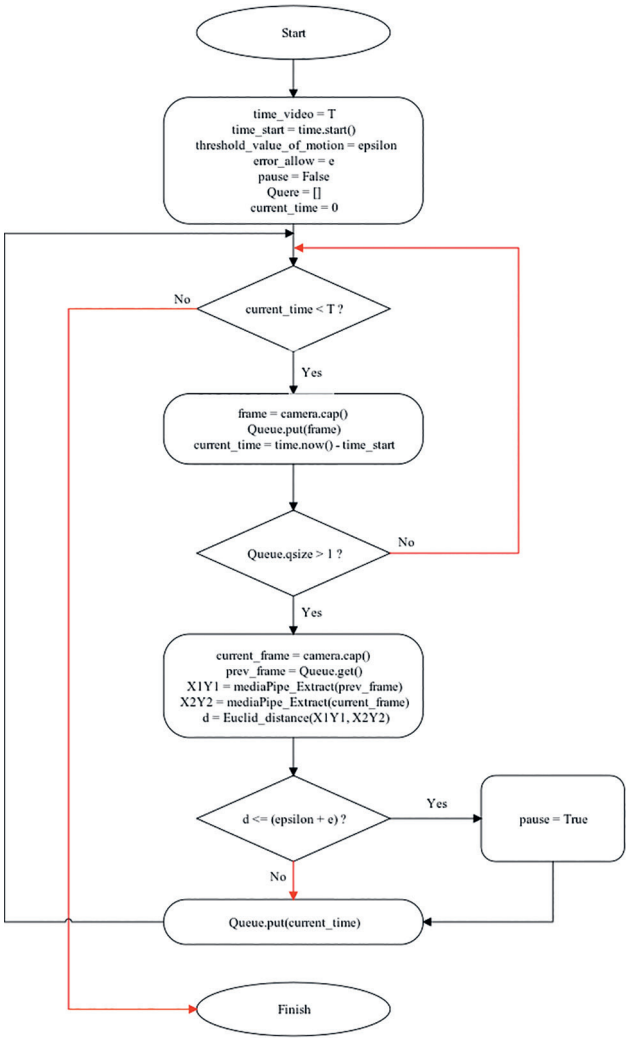

Рис. 3. Блок-схема алгоритма определения границы между жестами на видео

Fig. 3. Block diagram of the algorithm for determining the boundary between gestures in the video

видео, основанной на обнаружении движений рук и изменений в выражении лица. Для извлечения координат объектов точек на руке, предплечье и лице в настоящей работе использована библиотека MediaPipe от Google[1]. Выполненные исследования по распознаванию языка жестов с выделением функций при помощи библиотеки MediaPipe (рис. 4) также продемонстрировали положительные результаты [15–19].

Для уменьшения скачка погрешности расчета в предложенном методе выборочно извлечены координатные характеристики кисти, предплечья, глаза и рта, для учета внешнего вида жестов на языке жестов.

Пусть $M_{current} = \{(x_{c1}, y_{c1}), (x_{c2}, y_{c2}), \ldots (x_{cn}, y_{cn})\}$ — координаты извлеченных объектов кадра в момент времени t.

$M_{prev} = \{(x_{p1}, y_{p1}), (x_{p2}, y_{p2}), \ldots (x_{pn}, y_{pn})\}$ — координаты извлеченных объектов кадра в момент времени $t - 1$.

Согласно евклидовой формуле, рассчитаем длину отрезка AB с точками $A(x_1, y_1)$ и $B(x_2, y_2)$ по формуле:

$$|AB| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

Для оценки движения жеста в момент времени t по сравнению $t - 1$ усредним изменение расстояния между характерными точками позы объекта. Пусть $d = |M_{current}\, M_{prev}|$ — среднее значение перемещения объекта от времени $t - 1$ до времени t.

Тогда получим

$$d = \frac{1}{n}\left(\sqrt{(x_{c1} - x_{p1})^2 + (y_{c1} - y_{p1})^2} + \ldots + \right.$$
$$\left. + \sqrt{(x_{cn} - x_{pn})^2 + (y_{cn} - y_{pn})^2}\right).$$

Обозначим ε — в качестве максимального порога ошибки, рассчитанного при отсутствии движения: $\varepsilon = \max\{d_1, d_2, \ldots, d_n\}$, измеренный в момент отсутствия движения рук и изменения выражения лица. Пусть e — допустимый порог ошибки. При сравнении среднего значения перемещения с максимальным пороговым значением изменения, когда объект находится в состоянии покоя, определим движущийся объект, когда условие выполнено: $d > (\varepsilon + e)$.

В каждый момент времени выберем один кадр из очереди на обработку, извлечем координатные характеристики жеста для вычисления среднего смещения относительно кадра в предыдущий момент времени, чтобы определить, изменяется ли состояние жеста или объект приостановлен. В конце процесса текущий кадр поставим в очередь для обработки в следующий момент времени.

Дополнительно можно обнаружить появление движения жестов на языке жестов на видео на основе анализа разницы в спектре объекта между двумя последовательными кадрами. Отметим, что так исследуются только движения рук и изменения выражения лица, но этот подход не работает и многие элементы на видео можно рассматривать как генерируемый шум, который снижает точность решения проблемы.

[1] [Электронный ресурс]. Режим доступа: https://developers.google.com/mediapipe (дата обращения: 12.09.2023).

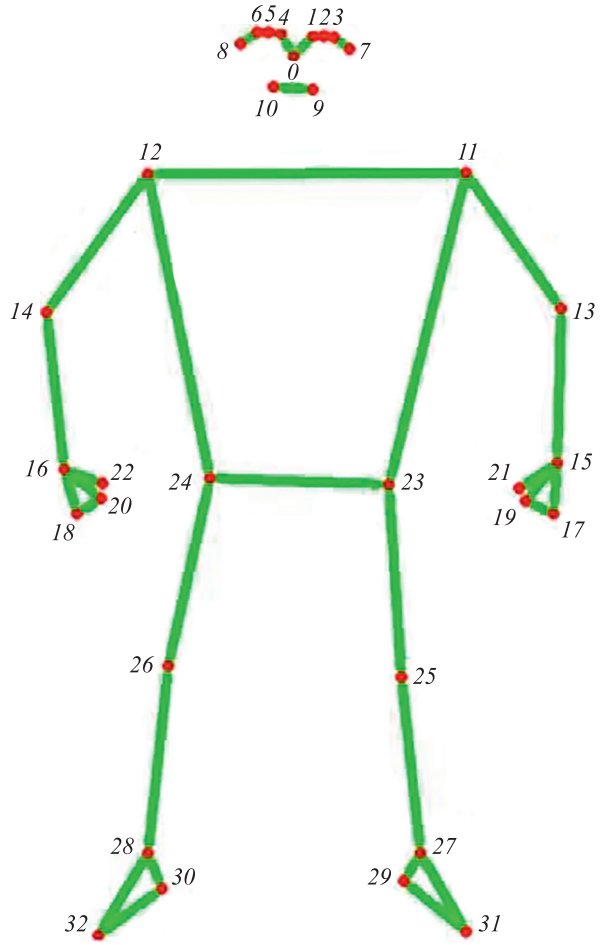

Рис. 4. Описание поз человеческого тела, распознанных библиотекой MediaPipe:

0 — нос; *1* и *4* — внутренний левый и правый глаз; *2* и *5* — левый и правый глаз; *3* и *6* — внешний левый и правый глаз; *7* и *8* — левое и правое ухо; *9* и *10* — рот слева и справа; *11* и *12* — левое и правое плечо; *13* и *14* — левый и правый локоть; *15* и *16* — левое и правое запястье; *17* и *18* — левый и правый мизинец; *19* и *20* — указательные пальцы левой и правой рук; *21* и *22* — большие пальцы левой и правой рук; *23* и *24* — левое и правое бедро; *25* и *26* — левое и правое колено; *27* и *28* — левая и правая лодыжка; *29* и *30* — левая и правая пятка; *31* и *32* — указательные пальцы левой и правой ног

Fig. 4. Description of the poses of the human body recognized by the MediaPipe library:

0 — nose; *1* and *4* — inner left and right eye; *2* and *5* — left and right eye; *3* and *6* — outer left and right eye; *7* and *8* — left and right ear; *9* and *10* — mouth left and right; *11* and *12* — left and right shoulder *13* and *14* — left and right elbow; *15* and *16* — left and right wrist; *17* and *18* — left and right little finger; *19* and *20* — index fingers of left and right hands; *21* and *22* — thumbs of left and right hands; *23* and *24* — left and right thigh; *25* and *26* — left and right knee; *27* and *28* — left and right ankles; *29* and *30* — left and right heel; *31* and *32* — index fingers of the left and right feet

Эксперименты и результаты

Набор данных вьетнамского жестового языка собран во Вьетнамском центре обучения языку жестов[2]. Размер

[2] [Электронный ресурс]. Режим доступа: https://tudienngonngukyhieu.com (дата обращения: 22.08.2023).

Таблица. Данные для тестирования

Table. Data for testing

Y_test, с	4–5	9–11	14–15	20–24	30–32	40–41	48–51	56–57	60–62
Y_pred, с	3,95–4,8	8,55–10,8	13,5–14,7	19,4–24,4	30,3–31,6	40,2–41,5	47,6–51,4	55,7–57,2	59,4–61,8

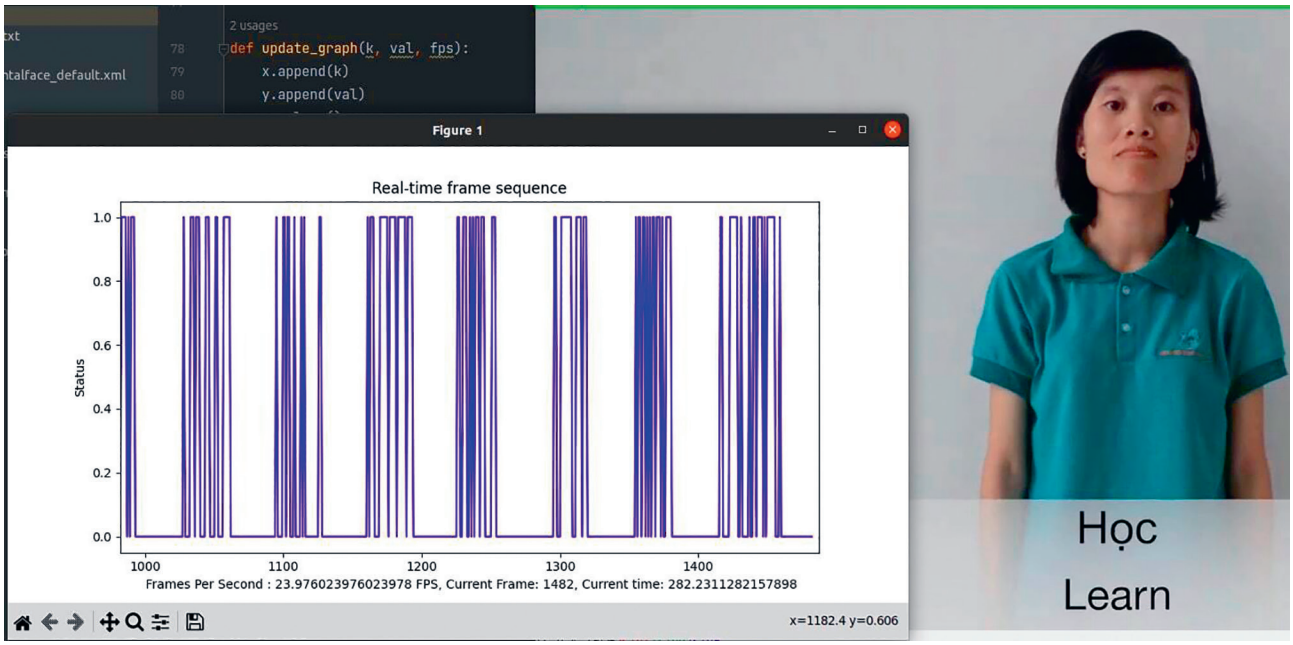

Рис. 5. График, отображающий сегментацию жестов в реальном времени

Fig. 5. A graph showing the segmentation of gestures in real time

собранного набора данных: 4364 видео, каждое видео соответствует одному слову, а названия видеороликов — значениям слов, описанных в файле vnDiSL.csv[1]. Для маркировки типа слова использована библиотека VnCoreNLP [20], далее полученные результаты проверены вручную. Чтобы обеспечить лучшее распознавание языка жестов, создан автоматический набор данных вьетнамского языка жестов на основе метода генеративно-состязательной нейросети (Generative Adversarial Network, GAN). Результаты работы данного метода являются направлением дальнейших исследований.

Для оценки результатов предложенного алгоритма выполнены тесты с несколькими видеороликами на языке жестов на вьетнамском языке с разной скоростью представления жестов. На рис. 5 показаны результаты определения времени остановки кадра в виде графика в реальном времени.

Для оценки точности (таблица) выполнено сравнение результатов, полученных с помощью программы прогнозирования и значений, определенных вручную.

На рис. 6 продемонстрированы результаты сопоставления заданного вручную граничного времени жеста с паузами (0 — пауза, 1 — движение), обнаруженными программой.

Точность. В результате анализа полученных результатов, при условии, что допустимая погрешность определения времени окончания и начала жеста составляет 0,5 с, а скорость выполнения на языке жестов

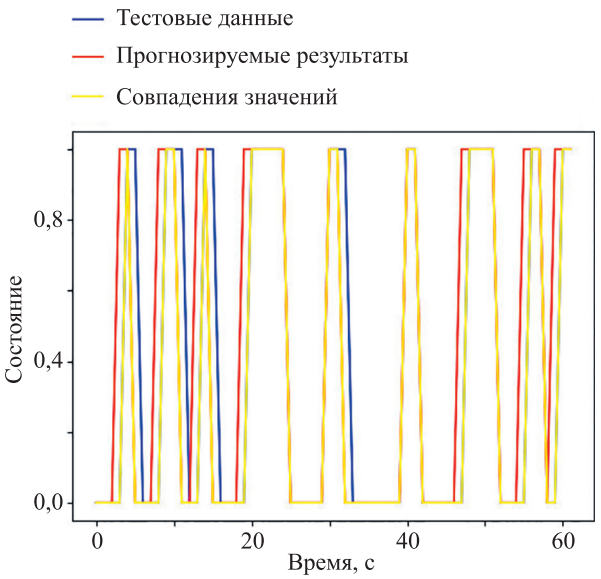

Рис. 6. График сравнения результатов прогнозирования программы со значениями, определенными вручную

Fig. 6. A graph comparing the results of the program prediction with the values determined manually

[1] [Электронный ресурс]. Режим доступа: https://github.com/DangKhanhITMO/VnSignLanguage (дата обращения: 12.09.2023).

составляет 25 слов в минуту, точность решения задачи достигла 90 %.

Обсуждение и заключение

В работе сегментированы на видео языковые жесты на основе определения времени паузы. Результаты сегментации представлены графически в режиме реального времени. Выполнено сравнение методов для выполнения поставленной задачи, таких как: методов, использующих модель глубокого обучения сверточная нейронная сеть и метода сегментации жестов языка жестов в видео на основе порогового значения, вычисляемого по евклидову расстоянию при извлечении координатных объектов с использованием библиотеки MediaPipe. В результате получены следующие преимущества выбранного метода: простота внедрения и развертывания приложений; низкие вычислительные затраты могут быть гарантированы для приложений, требующих высокой скорости и реального времени.

Для видео с умеренной и медленной скоростью жестикуляции метод сегментации видео с использованием библиотеки MediaPipe и на основе пороговых значений показал высокие результаты. Основная сложность при решении поставленной задачи возникла, когда скорость жеста очень высока, что усложнило сегментацию, так как паузы очень короткие и нечеткие. Отмечено, что настройка порога оценки степени изменения жестов в видео также зависит от скорости жеста. В будущих исследованиях будет оптимизирован процесс сегментации жестов на языке жестов, для распознавания видео с очень высокой скоростью жестикуляции. Планируется объединение данных для классификации факторов, рассматриваемых как помехи при общении на языке жестов, таких как движения, которые не являются языком жестов. Когда проблема сегментации жестов на языке жестов достигнет хороших результатов, понимание проблемы машинного перевода на язык жестов будет значительно улучшено.

Литература

1. Thoa N.T.K. Vietnamese sign language - unresolved isSUES // Proc. of the 4th Conference on Language Teaching and Learning" (LTAL). 2022. https://doi.org/10.21467/proceedings.132.23
2. Li D., Rodriguez C., Yu X., Li H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2020. P. 1459–1469. https://doi.org/10.1109/wacv45572.2020.9093512
3. Min Y., Hao A., Chai X., Chen X. Visual alignment constraint for continuous sign language recognition // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. P. 11522–11531. https://doi.org/10.1109/iccv48922.2021.01134
4. Camgoz N.C., Hadfield S., Koller O., Bowden R. SubUNets: End-to-end hand shape and continuous sign language recognition // Proc. of the IEEE International Conference on Computer Vision (ICCV). 2017. P. 3075–3084. https://doi.org/10.1109/iccv.2017.332
5. Camgoz N.C., Kindiroglu A., Karabuklu S., Kelepir M., Ozsoy A.S., Akarun L. BosphorusSign: A Turkish sign language recognition corpus in health and finance domains // Proc. of the International Conference on Language Resources and Evaluation (LREC). 2016.
6. Ko S.-K., Kim C.J., Jung H., Cho C. Neural sign language translation based on human keypoint estimation // Applied Sciences. 2019. V. 9. N 13. P. 2683. https://doi.org/10.3390/app9132683
7. Lea C., Vidal R., Reiter A., Hager G.D. Temporal convolutional networks: A unified approach to action segmentation // Lecture Notes in Computer Science. 2016. V. 9915. P. 47–54. https://doi.org/10.1007/978-3-319-49409-8_7
8. Kulkarni K., Evangelidis G., Cech J., Horaud R. Continuous action recognition based on sequence alignment // International Journal of Computer Vision. 2015. V. 112. N 1. P. 90–114. https://doi.org/10.1007/s11263-014-0758-9
9. Luc P., Neverova N., Couprie C., Verbeek J., LeCun Y. Predicting deeper into the future of semantic segmentation // Proc. of the 2017 IEEE International Conference on Computer Vision (ICCV). 2017. P. 648–657. https://doi.org/10.1109/ICCV.2017.77
10. Yi F., Wen H., Jiang T. ASFormer: Transformer for action segmentation // arXiv. 2021. arXiv:2110.08568. https://doi.org/10.48550/arXiv.2110.08568
11. Brognaux S., Drugman T. HMM-based speech segmentation: improvements of fully automatic approaches // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016. V. 24. N 1. P. 5–15. https://doi.org/10.1109/TASLP.2015.2456421
12. Atmaja B.T., Akagi M. Speech emotion recognition based on speech segment using LSTM with attention model // IEEE International Conference on Signals and Systems (ICSigSys). 2019. P. 40–44. https://doi.org/10.1109/ICSIGSYS.2019.8811080

References

1. Thoa N.T.K. Vietnamese Sign language - unresolved issues. *Proc. of the 4th Conference on Language Teaching and Learning" (LTAL)*, 2022. https://doi.org/10.21467/proceedings.132.23
2. Li D., Rodriguez C., Yu X., Li H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1459–1469. https://doi.org/10.1109/wacv45572.2020.9093512
3. Min Y., Hao A., Chai X., Chen X. Visual alignment constraint for continuous sign language recognition. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11522–11531. https://doi.org/10.1109/iccv48922.2021.01134
4. Camgoz N.C., Hadfield S., Koller O., Bowden R. SubUNets: End-to-end hand shape and continuous sign language recognition. *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3075–3084. https://doi.org/10.1109/iccv.2017.332
5. Camgoz N.C., Kindiroglu A., Karabuklu S., Kelepir M., Ozsoy A.S., Akarun L. BosphorusSign: A Turkish sign language recognition corpus in health and finance domains. *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 2016.
6. Ko S.-K., Kim C.J., Jung H., Cho C. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 2019, vol. 9, no. 13, pp. 2683. https://doi.org/10.3390/app9132683
7. Lea C., Vidal R., Reiter A., Hager G.D. Temporal convolutional networks: A unified approach to action segmentation. *Lecture Notes in Computer Science*, 2016, vol. 9915, pp. 47–54. https://doi.org/10.1007/978-3-319-49409-8_7
8. Kulkarni K., Evangelidis G., Cech J., Horaud R. Continuous action recognition based on sequence alignment. *International Journal of Computer Vision*, 2015, vol. 112, no. 1, pp. 90–114. https://doi.org/10.1007/s11263-014-0758-9
9. Luc P., Neverova N., Couprie C., Verbeek J., LeCun Y. Predicting deeper into the future of semantic segmentation. *Proc. of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 648–657. https://doi.org/10.1109/ICCV.2017.77
10. Yi F., Wen H., Jiang T. ASFormer: Transformer for action segmentation. *arXiv*, 2021, arXiv:2110.08568. https://doi.org/10.48550/arXiv.2110.08568
11. Brognaux S., Drugman T. HMM-based speech segmentation: improvements of fully automatic approaches. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, vol. 24, no. 1, pp. 5–15. https://doi.org/10.1109/TASLP.2015.2456421
12. Atmaja B.T., Akagi M. Speech emotion recognition based on speech segment using LSTM with attention model. *IEEE International Conference on Signals and Systems (ICSigSys)*, 2019, pp. 40–44. https://doi.org/10.1109/ICSIGSYS.2019.8811080

13. Gujarathi P.V., Patil S.R. Gaussian filter-based speech segmentation algorithm for Gujarati language // Smart Innovation, Systems and Technologies. 2021. V. 224. P. 747–756. https://doi.org/10.1007/978-981-16-1502-3_74
14. Chen M.-H., Li B., Bao Y., AlRegib G., Kira Z. Action segmentation with joint self-supervised temporal domain adaptation // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 9454–9463. https://doi.org/10.1109/cvpr42600.2020.00947
15. Madrid G.K.R., Villanueva R.G.R., Caya M.V.C. Recognition of dynamic Filipino Sign language using MediaPipe and long short-term memory // Proc. of the 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2022. https://doi.org/10.1109/ICCCNT54827.2022.9984599
16. Adhikary S., Talukdar A.K., Sarma K.K. A vision-based system for recognition of words used in Indian Sign Language using MediaPipe // Proc. of the 2021 Sixth International Conference on Image Information Processing (ICIIP). 2021. P. 390–394. https://doi.org/10.1109/ICIIP53038.2021.9702551
17. Zhang S., Chen W., Chen C., Liu Y. Human deep squat detection method based on MediaPipe combined with Yolov5 network // Proc. of the 2022 41st Chinese Control Conference (CCC). 2022. P. 6404–6409. https://doi.org/10.23919/CCC55666.2022.9902631
18. Quiñonez Y., Lizarraga C., Aguayo R. Machine learning solutions with MediaPipe // Proc. of the 2022 11th International Conference On Software Process Improvement (CIMPS). 2022. P. 212–215. https://doi.org/10.1109/CIMPS57786.2022.10035706
19. Ma J., Ma L., Ruan W., Chen H., Feng J. A Wushu posture recognition system based on MediaPipe // Proc. of the 2022 2nd International Conference on Information Technology and Contemporary Sports (TCS). 2022. P. 10–13. https://doi.org/10.1109/TCS56119.2022.9918744
20. Nguyen D.Q., Vu T., Nguyen D.Q., Dras M., Johnson M. 2017. From word segmentation to POS tagging for Vietnamese // Proc. of the 15th Australasian Language Technology Association Workshop. 2012. P. 108–113.

13. Gujarathi P.V., Patil S.R. Gaussian filter-based speech segmentation algorithm for Gujarati language. *Smart Innovation, Systems and Technologies*, 2021, vol. 224, pp. 747–756. https://doi.org/10.1007/978-981-16-1502-3_74
14. Chen M.-H., Li B., Bao Y., AlRegib G., Kira Z. Action segmentation with joint self-supervised temporal domain adaptation. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9454–9463. https://doi.org/10.1109/cvpr42600.2020.00947
15. Madrid G.K.R., Villanueva R.G.R., Caya M.V.C. Recognition of dynamic Filipino Sign language using MediaPipe and long short-term memory. *Proc. of the 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2022. https://doi.org/10.1109/ICCCNT54827.2022.9984599
16. Adhikary S., Talukdar A.K., Sarma K.K. A vision-based system for recognition of words used in Indian Sign Language using MediaPipe. *Proc. of the 2021 Sixth International Conference on Image Information Processing (ICIIP)*, 2021, pp. 390–394. https://doi.org/10.1109/ICIIP53038.2021.9702551
17. Zhang S., Chen W., Chen C., Liu Y. Human deep squat detection method based on MediaPipe combined with Yolov5 network. *Proc. of the 2022 41st Chinese Control Conference (CCC*), 2022, pp. 6404–6409. https://doi.org/10.23919/CCC55666.2022.9902631
18. Quiñonez Y., Lizarraga C., Aguayo R. Machine learning solutions with MediaPipe. *Proc. of the 2022 11th International Conference On Software Process Improvement (CIMPS)*, 2022, pp. 212–215. https://doi.org/10.1109/CIMPS57786.2022.10035706
19. Ma J., Ma L., Ruan W., Chen H., Feng J. A Wushu posture recognition system based on MediaPipe. *Proc. of the 2022 2nd International Conference on Information Technology and Contemporary Sports (TCS)*, 2022, pp. 10–13. https://doi.org/10.1109/TCS56119.2022.9918744
20. Nguyen D.Q., Vu T., Nguyen D.Q., Dras M., Johnson M. 2017. From word segmentation to POS tagging for Vietnamese. *Proc. of the 15th Australasian Language Technology Association Workshop*, 2012, pp. 108–113.

Авторы

Данг Хань — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0009-5882-7653, dangkhanhmta.2020@gmail.com
Бессмертный Игорь Александрович — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 36661767800, https://orcid.org/0000-0001-6711-6399, bessmertny@itmo.ru

Authors

Dang Khanh — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0009-5882-7653, dangkhanhmta.2020@gmail.com
Igor A. Bessmertny — D. Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 36661767800, https://orcid.org/0000-0001-6711-6399, bessmertny@itmo.ru

Статья поступила в редакцию 10.05.2023
Одобрена после рецензирования 27.07.2023
Принята к печати 26.09.2023

Received 10.05.2023
Approved after reviewing 27.07.2023
Accepted 26.09.2023