УДК 004.896
DOI: 10.26102/2310-6018/2024.45.2.035

Исследование эффективности моделей глубокого обучения в задаче распознавания технологических операций, как последовательности движений кистей рук

С.Е. Штехин, А.В. Стадник✉

«Отраслевой центр разработки и внедрения информационных систем» Сириус, филиал № 11, Сочи, Российская Федерация

Резюме. В работе изучаются методы распознавания на видео специфического класса технологических операций ручного труда, который представляет собой последовательности движения кистей и пальцев рук. Технологическая операция здесь определяется как последовательность новых специфических символов жестового языка. Рассмотрены различные методы распознавания жестов на видео. Исследован двухэтапный подход: на первом этапе распознаются ключевые точки рук на каждом кадре с помощью открытой библиотеки mediapipe, на втором этапе покадровая последовательность ключевых точек трансформируется в текст с помощью обученной нейросети архитектуры трансформер. Основное внимание уделено обучению модели нейросети архитектуры трансформер на базе открытого датасета американского жестового языка (ASL) для распознавания предложений жестового языка на видео. Затронут вопрос применимости данного подхода и обученной модели ASL для распознавания технологических операций ручного труда с мелкой моторикой в виде текстовой последовательности. Полученные результаты могут быть полезны при исследовании трудовых процессов с быстрыми движениями и малыми отрезками времени в алгоритмах распознавания технологических операций ручного труда на видеоданных.

Ключевые слова: видеоанализ движений рук, распознавание жестов, распознавание действий, глубокие нейронные сети, трансформер, технологические операции.

Для цитирования: Штехин С.Е., Стадник А.В. Исследование эффективности моделей глубокого обучения в задаче распознавания технологических операций, как последовательности движений кистей рук. *Моделирование, оптимизация и информационные технологии.* 2024;12(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1574 DOI: 10.26102/2310-6018/2024.45.2.035

Study of deep learning models in the task of recognizing technological operations as a sequence of hand movements

S.E. Shtekhin, A.V. Stadnik✉

"Industry center for the development and implementation of information systems" Sirius, branch No. 11, Sochi, the Russian Federation

Abstract. In this paper, we consider methods for recognizing on video a specific class of technological manual labor operations, which are a sequence of movements of the hands and fingers. The technological operation in this work is considered as a sequence of new specific symbols of the sign language. The paper considers various methods of gesture recognition on video. In this paper, a two-step approach was investigated. At the first stage, the key points of the hands on each frame are recognized by using the open mediapipe library. At the second stage, a frame-by-frame sequence of keypoints transformed into text using a trained neural network of the transformer architecture. The main attention is paid to training a neural network model of the Transformer architecture based on the open American Sign Language (ASL) dataset for recognizing sign language sentences in video. The paper considers the applicability of approach and the trained model of ASL for recognizing technological

operations of manual labor with fine-motor skills as a text sequence. The results obtained in this paper can be useful in the study of labor processes with fast movements and short time intervals in algorithms for recognizing technological operations of manual labor on video data.

Keywords: video analysis of hand movements, gesture recognition, action recognition, deep neural networks, transformer, technological operations.

For citation: Shtekhin S.E., Stadnik A.V. Study of deep learning models in the task of recognizing technological operations as a sequence of hand movements. *Modeling, Optimization and Information Technology*. 2024;12(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1574 DOI: 10.26102/2310-6018/2024.45.2.035 (In Russ.).

Введение

При работах на железной дороге стоит задача определения правильного порядка последовательности технологических операций ручного труда и определения времени каждой такой операции на видеоданных. Исходя из специфики операций, их можно разбить на технологические операции, выполняемые сотрудником с помощью различных инструментов и выполняемые сотрудником руками без инструментов (Рисунок 1). К технологическим операциям, выполняемым с инструментами, относится большинство операций, выполняемых при ремонте и обслуживании железнодорожного пути.

Рисунок 1 – Примеры работ на железнодорожных путях
Figure 1 – Examples of work on railway tracks

Распознавание таких технологических операций на видео относится к классу задач компьютерного зрения – детектирование взаимодействия человека с объектом (HOI) [1]. В последние годы был достигнут значительный прогресс в области решения такого типа задач [2–5]. Обнаружение взаимодействия человека с объектом было первоначально предложено в работе [6].

Для решения этого класса задач необходимо детектировать объекты на изображении и классифицировать тип их взаимодействия. Для класса задач HOI одним из объектов является человек и необходимо определить взаимодействие человека с объектом. При распознавании технологических операций ручного труда на видео необходимо детектировать сотрудника, выполняющего операцию, инструмент и определить взаимодействие сотрудника с инструментом. Методы распознавания таких технологических операций были исследованы [7] и реализованы в цифровой системе обработки видео.

Ко второй группе технологических операции относятся операции, связанные с мелкой моторикой, которые в основном выполняются без инструментов, например, закручивание гаек руками, включение кнопок, тумблеров и т. д. (Рисунок 2).

Рисунок 2 – Пример работы, связанной с мелкой моторикой
Figure 2 – An example of work related to fine motor skills

Каждая такая технологическая операция представляет собой последовательность движений руками и пальцами. Для распознавания таких технологических операций на видео необходимо найти руки на видео, определить координаты всех ключевых точек каждой руки и каждого пальца на изображении и распознать последовательность движений пальцев и рук, т. е. определить типовые последовательности паттернов смещения координат ключевых точек рук, соответствующие конкретным технологическим операциям второй группы.

Целью данной работы является проверка гипотезы успешного применения методов распознавания жестового языка для распознавания специфических технологических операций, связанных с мелкой моторикой.

Материалы и методы

Задача распознавания последовательностей движений рук и пальцев на видео широко исследовалась в области распознавания жестового языка, который используют слабослышащие и глухонемые люди.

Каждая фраза жестового языка состоит из букв, подобно тому, как это происходит в устной речи. Каждая буква жестового языка обозначает конкретное движение, то есть последовательность перемещений кистей рук и пальцев. Переходы между жестами также представляют собой последовательность движений кистей и пальцев. Таким образом, видеозапись фразы жестового языка (Рисунок 3) является аналогом специфической технологической операции ручного труда.

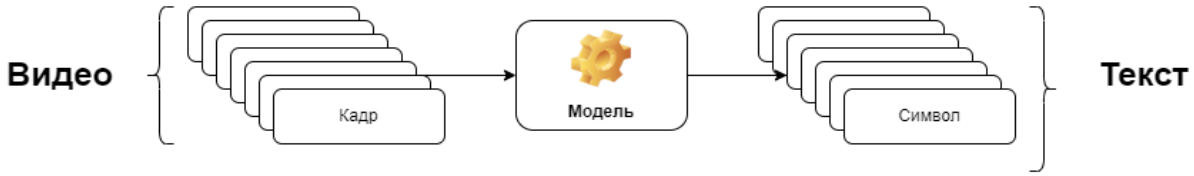

Рисунок 3 – Схема архитектуры задачи распознавания предложения на жестовом языке
Figure 3 – Architecture diagram of the task of recognizing sentences in sign language

Американский жестовый язык (ASL) является одним из подмножеств жестовых языков, которым пользуются от 350 000 до 500 000 человек [8]. Задача распознавания жестового языка заключается в анализе видеофрагмента, где фраза показана пальцами, и определении соответствующих букв. Это задача предсказания последовательности действий, аналогичная задаче распознавания слов в аудио домене, но в этой области

данных есть несколько интересных свойств, специфичных для языка жестов. Одно из них – наличие большого количества движений и отсутствие устойчивого состояния для каждой буквы. Обычно каждая буква представляется кратким «пиком артикуляции», когда движение руки сводится к минимуму, а форма жеста наиболее приближена к целевому жесту. Этот пик сопровождается более продолжительным переходом между текущей и предыдущими / следующими буквами [9]. Классификаторы выполняют функцию распознавания пиков. При демонстрации пальцами последовательности букв получается соответствующая последовательность «пиков» артикуляции. Это кадры, в которых рука достигает заданной формы для определенной буквы. Пиковый кадр и кадры вокруг него обычно характеризуются минимальным движением, поскольку переход к текущей букве завершен, а переход к следующей букве еще не начался. Между тем, переходные кадры между пиковыми значениями букв обладают большим движением. Использование «функций обнаружения пиков», специфичных для букв, позволяет обеспечить, чтобы каждый прогнозируемый сегмент букв имел один пик.

В работе [9] активно изучалась тема распознавания жестов ASL на видео. Авторы предложили использовать «монобуквенные» модели, где каждая единица является контекстно-независимой буквой. В работе описана попытка создать модели, способные учитывать информацию, связанную с языком жестов, включая динамические аспекты движений для букв, демонстрируемых пальцами. Это подразумевает использование сегментарных моделей, отдельных для каждого демонстрируемого символа. Также аспекты устойчивости модели и легкой адаптации моделей при изменении пользователя, демонстрирующего жесты, являются важной задачей для авторов исследования.

Для достижения этой цели используются независимые классификаторы глубоких нейронных сетей (DNN) [10, 11], которые могут быть адаптированы и интегрированы с различными моделями последовательностей. Обучение DNN осуществляется с использованием регуляризованной кросс-энтропийной потери L2. Входные данные представлены элементами изображения, собранными в многокадровом окне с центром в текущем кадре, которые проходят через несколько полностью связанных слоев, за которыми следует выходной слой softmax с таким количеством единиц измерения, сколько имеется меток. Нейронные сети обучаются идентификации фонем, а их результаты (сообщения на основе фонем) обрабатываются и применяются в качестве данных для стандартного распознавателя на основе скрытой марковской модели HMM [12] с использованием гауссова смешанного распределения наблюдений.

На уровне каждого кадра изображения передаются в семь классификаторов DNN. Один из них отвечает за предсказание буквенной метки кадра, а остальные шесть – за определение фонологических характеристик жеста. После получения результатов от классификаторов и обработки объектов изображения с помощью метода уменьшения размерности PCA, эти данные объединяются. Полученные таким образом объединенные объекты становятся основой для работы распознавателя на базе HMM с гауссовой плотностью наблюдений. В данном случае используется HMM с тремя состояниями для каждой буквы (независимо от контекста), а также по одному HMM для начального и конечного сегментов без подписи («тишины»).

Используется набор классификаторов фреймов, каждый из которых предназначен для классификации либо букв, либо лингвистических признаков жеста. Для каждого признака или буквы обучается классификатор, который выдает оценку по каждому значению признака для каждого видеокадра. Также обучается отдельный классификатор букв на основе глубоких нейронных сетей.

В исследовании [13] изучались подходы, основанные на использовании сверточных нейронных сетей, где распознавались отдельные жесты (отдельные буквы). В данной научной статье представлена система распознавания жестов ASL в реальном

времени, которая использует современные методы компьютерного зрения и машинного обучения. В предложенном методе применяется библиотека Mediapipe для извлечения признаков и сверточная нейронная сеть (CNN) для классификации жестов на языке ASL.

Для повышения эффективности оценки конфигурации в динамических сценах предлагается использовать архитектуру, объединяющую CNN и LSTM. Наилучшие результаты для распознавания слитной речи достигаются с помощью моделей, основанных на архитектуре типа CNN + LSTM. Основная проблема распознавания слитной речи заключается в отсутствии предварительных знаний о границах жестов. Комбинация CNN и LSTM помогает решить эту проблему. Совмещение модальностей считается наиболее перспективным подходом.

Так, исследование [14] посвящено методам работы с предложениями на жестовом языке, основанным на рекуррентных нейронных сетях типа LSTM [15, 16]. Предложены два подхода для зеркального отображения на уровне предложений, которое заключается в сопоставлении видеозаписей предложений на жестовом языке с последовательностями письменных текстов. В обеих моделях используется коннекционистская временная классификация (CTC), позволяющая избежать предварительного разделения предложений на отдельные слова. Первая модель основана на LRCN, вторая – на мультиключевой сети. LRCN – это модель, в которой CNN служит средством выделения признаков для каждого кадра перед передачей их в LSTM. В первом подходе не используются предварительные знания, сырые кадры загружаются в 18-слойный LRCN с CTC наверху. Во втором подходе с помощью Mediapipe извлекаются три основные характеристики (форма руки, положение руки и информация о движении руки), соответствующие каждому знаку. Для создания скелета кистей используются 2D-ориентиры формы кистей, которые затем переносятся в модель CONV-LSTM. Указываются координаты кистей и их относительное расстояние до головы.

В последние годы основные направления интеллектуального анализа жестов и жестовой речи включают отслеживание и определение конфигурации рук, анализ временных и пространственных характеристик жестов, а также цепочек жестов [17]. Методы оценки конфигурации руки могут быть разделены на 2D и 3D, при этом большинство современных моделей работают с 3D-данными [18]. Использование карт глубины для решения этой задачи рассматривается в ряде обзоров, например, [19, 20]. В исследовании [21] обсуждаются методы использования трехмерной реконструкции для определения позы человека на отдельных изображениях. Авторы предлагают подход, который комбинирует прямую кинематику (FK) и нейросети для быстрого и точного предсказания трехмерной позы. Поза представляется в виде иерархической структуры с узлами, соответствующими суставам человека, которые имитируют их физические ограничения. Определив ключевые точки на изображении в двух измерениях, авторы преобразуют скелет в трехмерный формат, используя нейросети для прогнозирования угловых позиций суставов и длин костей. Эти прогнозы затем объединяются со скелетными ограничениями с помощью слоя FK, реализованного в библиотеке PyTorch. Это позволяет получить быстрый и точный метод оценки положения скелета в трехмерном пространстве.

Однако в рассмотренных выше работах не исследовались такие современные методы глубокого обучения, как искусственные нейронные сети, построенные на архитектуре глубокого обучения трансформер. В последние годы данные методы показали высокую эффективность для работы с последовательностями, особенно с текстом. Основа этих методов, основанных на механизме multi-head attention, была заложена в статье [22]. Архитектура трансформера состоит из кодирующих и декодирующих слоев. Каждый слой состоит из механизма self-attention и нейронной сети прямого распространения (MLP). За счет этих механизмов, в отличие от рекуррентных

нейронных сетей [23], данная архитектура позволяет обрабатывать данные параллельно, что увеличивает ее производительность. Широкое распространение получили данные методы для работы с текстами, как с последовательностями букв, слов и т. д. Данные методы лежат в основе современных LLM – больших языковых моделей, таких как GPT 3 [24], Llama [25] и т. д.

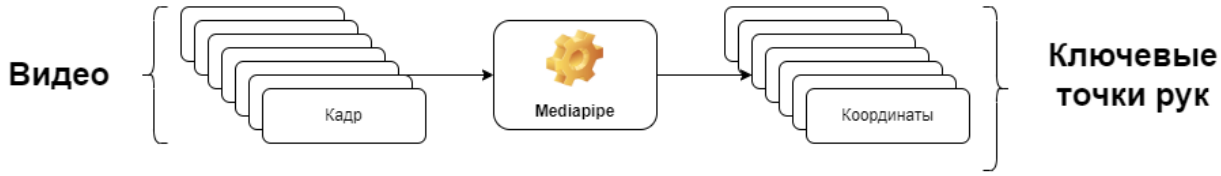

Рисунок 4 – Схема архитектуры формирования датасета с помощью библиотеки mediapipe
Figure 4 – A diagram of the architecture for generating a dataset using the mediapipe library

Для исследования был взят открытый датасет американского жестового языка от компании Google на платформе Kaggle, который представляет собой набор ключевых точек человека для более 46000 фраз, полученных с видеозаписей более 100 различных человек, с помощью открытой библиотеки mediapipe (Рисунок 4), основанной на статье [26].

Фразы состоят из символов, от 5 до 30 символов в каждой фразе. Большая часть фраз (14000) состоит из 10 символов. Каждая видеозапись фразы может быть от 30 до 600 кадров. Датасет представлен реальными данными, содержит большое количество артефактов, иногда пропущены жесты, соответствующие буквам фразы, также в фразе могут быть пропуски, когда ключевые точки не определены или первый кадр входной последовательности может начинаться с символа, который не является начальным в целевой фразе.

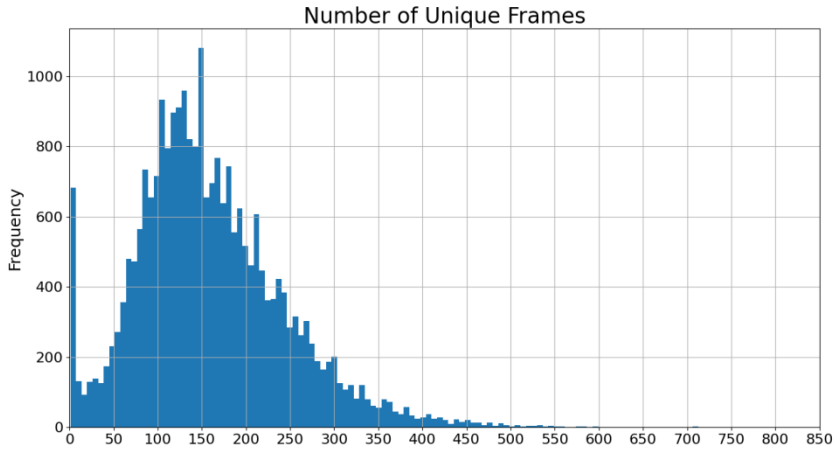

Рисунок 5 – Распределение количества кадров на фразу
Figure 5 – Distribution of the number of frames per phrase

Целевые фразы для «озвучивания» были случайным образом сгенерированы в виде почтовых адресов, имен и фамилий людей, а также номеров телефонов и интернет-адресов. Особенностью подобных данных является то, что на жестовом языке они артикулируются как последовательность символов алфавитной последовательности один за другим. Некоторая статистика по использованному датасету представлена ниже на Рисунках 5–6, где показаны распределения по длине целевой фразы, количество

уникальных фреймов во входной последовательности и среднее количество фреймов с ключевыми точками, приходящееся на один демонстрируемый символ фразы.

Для понимания структуры данных и особенностей построения фразы в выбранном датасете мы обучили классификатор на детектирование символа алфавита на одном кадре. Для этих целей небольшое подмножество входных данных, каждая последовательность в котором была разбита на интервалы в соответствии с гипотезой, что минимальное движение соответствует демонстрируемому символу – кратковременному «пику артикуляции», а максимальное движение кисти руки означает переход между символами. На Рисунках 7, 8 можно видеть покадровый результат классификации двух слов «very» и «lawanda». Из графиков видно, что символы и паузы распределены крайне неравномерно внутри фразы, что и указывает на необходимость искать решение задачи в виде постановки sequence-to-sequence, когда на вход и выход мы имеем последовательность (Рисунок 9).

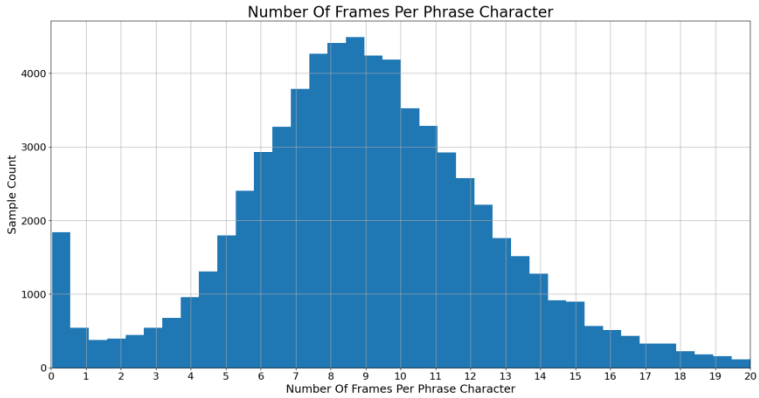

Рисунок 6 – Распределение числа кадров на один символ фразы во входных данных
Figure 6 – Distribution of the number of frames per character of the phrase in the input data

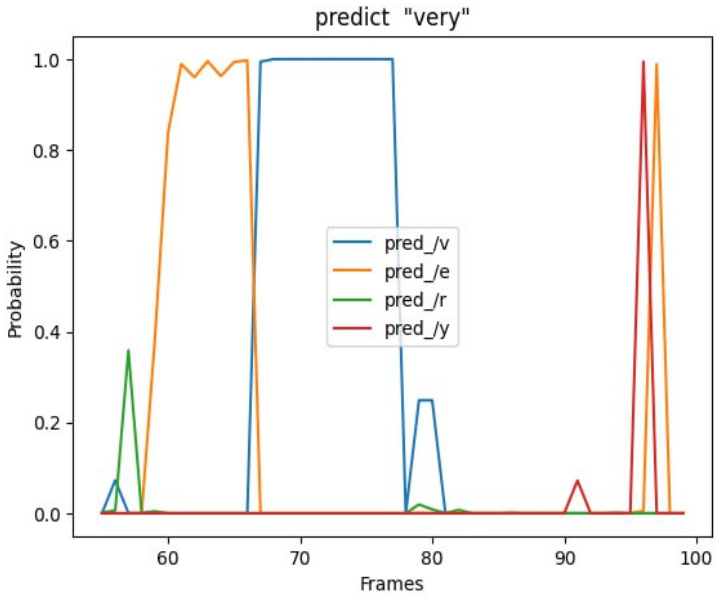

Рисунок 7 – Вероятности нахождения символов в фразе «very»
Figure 7 – The probability of finding characters in the phrase "very"

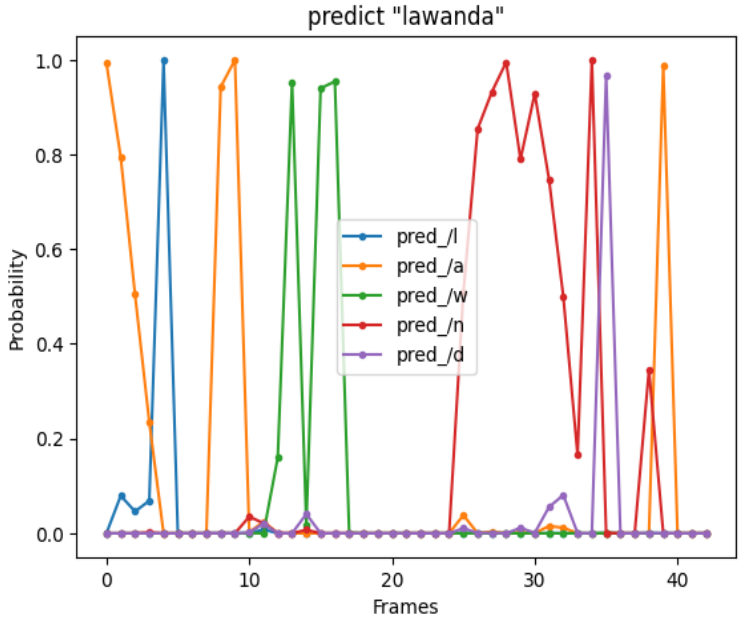

Рисунок 8 – Вероятности нахождения символов в фразе «lawanda»
Figure 8 – The probability of finding characters in the phrase "lawanda"

В качестве входных данных для трансформера использовался вектор координат ключевых точек руки для каждого кадра видеофрагмента. Число классов для символов равно 60.

Рисунок 9 – Постановка задачи распознавания предложения жестового языка из ключевых точек рук с помощью модели на базе нейросети архитектуры трансформер
Figure 9 – Task of recognizing a sign language sentence from key points of the hands using a model based on the transformer neural network architecture

Для распознавания последовательности была выбрана архитектура трансформера с энкодером и декодером с MultiHead Attention. Максимальная длина последовательности была выбрана равной 100 символам, соответственно для входа и выхода. Количество блоков энкодера равно 6, декодера равно 2. Количество голов MultiHead Attention равно 4, размерность внутреннего латентного представления составляет 128. Использовалась линейная активация уровня классификации для логитов в функции потерь. В качестве лосс-функции использовалась категориальная кросс-энтропия. Метрикой оценки релевантности фразы выступало расстояние Левенштейна. Для обучения используются все данные, тестовое множество не выделялось отдельно.

Модель, обученная на распознавании фраз жестового языка на открытом датасете, в дальнейшем с помощью переноса обучения может быть дообучена на датасете технологических операций для распознавания специфических технологических операций ручного труда с мелкой моторикой.

Результаты

Обучение проводилось на 100 обучающих эпохах, с размером батча 64 сэмпла. Для регуляризации обучения использовались 20 % dropout для выхода MultiHead Attention. Модель достигла следующих значений на валидационном датасете:

loss: 2.0904 – top1acc: 0.7667 – top5acc: 0.9174

Кривая обучения показывает асимптотический выход ошибки валидации на указанное значение (Рисунок 10).

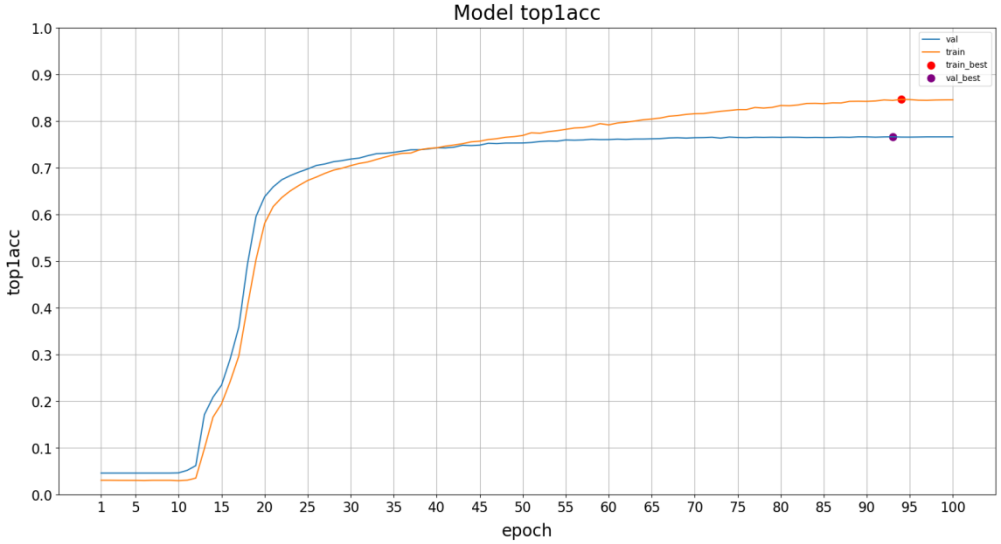

Рисунок 10 – Кривая обучения топ 1 accuracy модели для 100 эпох
Figure 10 – Learning curve top 1 accuracy models for 100 epochs

Из примеров видно, что концентрация ошибок смещается к концу фразы, что также проявляется часто и в разговорной речи, когда окончание фразы менее четко произносится в отличие от ее начала (Таблица 1).

Таким образом, можно сделать вывод, что модель способна распознавать последовательность операций достаточно хорошо в случае, если движение рук является четко видимым.

Таблица 1 – Примеры распознавания фраз
Table 1 – Examples of phrase recognition

GT фраза	Predicted фраза
S3 creekhouseE	S3 creek houseE
Sscales/kuhaylahE	Sscales/pouhapouhaillaE
S1383 william lanierE	S1385 william lanierE
S988 franklin laneE	S9888 frankan daneE
S6920 northeast 661st roadE	S6920 northeast 66th roadE
Swww.freem.ne.jpE	Swww.freem.itE
Shttps://jsi.is/hukuokaE	Shttps://jsitsi.is/hkurokaE
S239613 stolze streetE	S2396 dar stoless roadE
S271097 bayshore boulevardE	S271097 baymore boulevardE
Sfederico pearsonE	Sferico pearoonE

Проведенные исследования технологических операций с помощью обученной на датасете модели без специфических жестов показали, что для одной и той же технологической операции, выполненной различным образом на разных видео, обученная модель выдает похожие близкие текстовые последовательности.

Рисунок 11 – Различные исполнения одной технологической операции
Figure 11 – Different versions of the same technological operation

Для технологической операции закручивания гаек результаты модели на показанных на Рисунке 11 различных видеофрагментах – это строки 'carlen galen' и 'carlen cassen'. В целом, для данной технологической операции на всех видеофрагментах были достигнуты результаты, достаточно близкие по расстоянию Левенштейна = 3.

Обсуждение

Для распознавания таких технологических операций были исследованы различные методы глубокого обучения. В работе рассмотрена гипотеза, что технологическая операция является последовательностью определенных движений кисти и пальцев рук человека.

Для исследования распознавания движений кистей и пальцев рук была предложена следующая методика: детектирования сотрудника, детектирования кистей человека, распознавания ключевых точек кисти и пальцев и на основании последовательности движений, определения технологической операции. Исследовано использование двухэтапного подхода, в котором получение ключевых точек на первом этапе производится с помощью открытой библиотеки mediapipe. На втором этапе производится распознавание технологической операции как фразы жестового языка с помощью предложенной модели. Эксперименты с распознаванием некоторых технологических операций подтвердили применимость предложенной методики и подтвердили гипотезу, что методы распознавания жестового языка применимы к задаче распознавания технологических операций с мелкой моторикой. Эксперименты с формированием входных данных из набора косинусных расстояний между фалангами пальцев показали худшую точность в ходе экспериментов по обучению. Также использование нормированных на длину кисти евклидовых расстояний показало худший результат в сравнении с конкатенированным вектором координат.

В дальнейшем авторы планируют провести исследование технологических операций ручного труда мелкой моторики, для распознавания которых необходимо добавить новые специфические жесты в существующий алфавит ASL и дообучить модель с учетом новых жестов.

Заключение

В данной работе был рассмотрена задача распознавания на видеоданных специфического класса технологических операций ручного труда, который представляет собой последовательность движений кистей и пальцев рук. Такие технологические операции были рассмотрены как последовательность жестов американского жестового языка в задаче распознавания фраз жестового языка на видео.

Для распознавания предложений жестового языка на видео были рассмотрены различные методы, в результате был предложен двухэтапный подход для решения этой задачи. На первом этапе для распознавания на каждом кадре видео, ключевых точек кистей руки, пальцев применялась открытая библиотека mediapipe. На втором этапе для трансформации последовательности ключевых точек в предложение была обучена модель нейронной сети архитектуры трансформер. Модель была обучена на открытом датасете американского жестового языка.

В целом, в задаче определения соответствия жестов на видео заданному набору фраз, рассмотренной в данной работе, модель показала хорошие результаты.

Для набора технологических операций без инструментов, где операция представляет собой последовательность движений кисти, пальцев, определение конкретной операции по ключевым точкам применена представленная методика. Данные технологические операции были успешно распознаны на видео как текстовая последовательность.

Полученные результаты могут быть полезны при исследовании трудовых процессов с быстрыми движениями и малыми отрезками времени в алгоритмах распознавания технологических операций ручного труда на видеоданных.

Полученные результаты показывают, что методы и модель распознавания жестового языка может быть применима и для распознавания специфических технологических операций с мелкой моторикой.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Hou Z., Peng X., Qiao Y., Tao D. Visual Compositional Learning for Human-Object Interaction Detection. In: *Computer Vision – ECCV 2020: 16th European Conference: Proceedings: Part XV, 23-28 August 2020, Glasgow, United Kingdom.* Cham: Springer; 2020. P. 584–600. https://doi.org/10.1007/978-3-030-58555-6_35
2. Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature Pyramid Networks for Object Detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 July 2017, Honolulu, HI, USA.* IEEE; 2017. P. 936–944. https://doi.org/10.1109/CVPR.2017.106
3. Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A.C. SSD: Single Shot MultiBox Detector. In: *Computer Vision – ECCV 2016: 14th European Conference: Proceedings: Part I, 11-14 October 2016, Amsterdam, The Netherlands.* Cham: Springer; 2016. P. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
4. Nie J., Anwer R.M., Cholakkal H., Khan F.S., Pang Y., Shao L. Enriched Feature Guided Refinement Network for Object Detection. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 October 2019 – 02 November 2019, Seoul, Korea (South).* IEEE; 2019. P. 9536–9545. https://doi.org/10.1109/ICCV.2019.00963
5. Pang Y., Xie J., Khan M.H., Anwer R.M., Khan F.S., Shao L. Mask-Guided Attention Network for Occluded Pedestrian Detection. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 October 2019 – 02 November 2019, Seoul, Korea (South).* IEEE; 2019. P. 4966–4974. https://doi.org/10.1109/ICCV.2019.00507

6. Gupta J., Malik J. Visual Semantic Role Labeling. URL: https://doi.org/10.48550/arXiv.1505.04474 (Accessed 19th March 2024).
7. Штехин С.Е., Карачёв Д.К., Иванова Ю.А. Разработка алгоритма распознавания движений человека методами компьютерного зрения в задаче нормирования рабочего времени. *Труды Института системного программирования РАН.* 2020;32(1):121–136. https://doi.org/10.15514/ISPRAS-2020-32(1)-7
Shtekhin S.E., Karachev D.K., Ivanova Yu.A. Computer vision system for Working time estimation by Human Activities detection in video frames. *Trudy Instituta sistemnogo programmirovaniya RAN = Proceedings of the Institute for System Programming of the RAS.* 2020;32(1):121–136. (In Russ.). https://doi.org/10.15514/ISPRAS-2020-32(1)-7
8. Mitchell R.E., Young T.A., Bachleda B., Karchmer M.A. How Many People Use ASL in the United States? Why Estimates Need Updating. *Sign Language Studies.* 2006;6(3):306–335. https://doi.org/10.1353/sls.2006.0019
9. Kim T. American Sign Language fingerspelling recognition from video: Methods for unrestricted recognition and signer-independence. URL: https://doi.org/10.48550/arXiv.1608.08339 (Accessed 19th March 2024).
10. Suresh S., Mithun H.T.P, Supriya M.H. Sign Language Recognition System Using Deep Neural Network. In: *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 15-16 March 2019, Coimbatore, India.* IEEE; 2019. P. 614–618. https://doi.org/10.1109/ICACCS.2019.8728411
11. Kim S., Ji Y., Lee K.-B. An Effective Sign Language Learning with Object Detection Based ROI Segmentation. In: *2018 Second IEEE International Conference on Robotic Computing (IRC), 31 January 2018 – 02 February 2018, Laguna Hills, CA, USA.* IEEE; 2018. P. 330–333. https://doi.org/10.1109/IRC.2018.00069
12. Shivashankara S., Srinath S. A Review on Vision Based American Sign Language Recognition, its Techniques, and Outcomes. In: *2017 7th International Conference on Communication Systems and Network Technologies (CSNT), 11-13 November 2017, Nagpur, India.* IEEE; 2017. P. 293–299. https://doi.org/10.1109/CSNT.2017.8418554
13. Kumar R., Bajpai A., Sinha A. Mediapipe and CNNs for Real-Time ASL Gesture Recognition. URL: https://doi.org/10.48550/arXiv.2305.05296 (Accessed 19th March 2024).
14. Akandeh A. Sentence-Level Sign Language Recognition Framework. URL: https://doi.org/10.48550/arXiv.2211.14447 (Accessed 23rd March 2024).
15. Lee C.K.M. et al. American sign language recognition and training method with recurrent neural network. *Expert Systems with Applications.* 2021;167. https://doi.org/10.1016/j.eswa.2020.114403
16. Jayanthi P., Ponsy R.K., Bhama S.P.R., Madhubalasri B. Sign Language Recognition using Deep CNN with Normalised Keyframe Extraction and Prediction using LSTM. *Journal of Scientific and Industrial Research.* 2023;82(7):745–755.
17. Рюмин Д. Метод автоматического видеоанализа движений рук и распознавания жестов в человеко-машинных интерфейсах. *Научно-технический вестник информационных технологий, механики и оптики.* 2020;20(4):525–531. https://doi.org/10.17586/2226-1494-2020-20-4-525-531
Ryumin D. Automated hand detection method for tasks of gesture recognition in human-machine interfaces. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki = Scientific and Technical Journal of Information Technologies, Mechanics and Optics.* 2020;20(4):525–531. (In Russ.). https://doi.org/10.17586/2226-1494-2020-20-4-525-531
18. Tekin B., Bogo F., Pollefeys M. H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions. In: *2019 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR), 15-20 June 2019, Long Beach, CA, USA. IEEE; 2019. P. 4506–4515. https://doi.org/10.1109/CVPR.2019.00464

19. 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), 01-05 March 2020, Snowmass, CO, USA.* IEEE; 2020. P. 1448–1458. https://doi.org/10.1109/WACV45572.2020.9093512
20. Supančič Ja.S., Rogez G., Yang Yi., Shotton Ja., Ramanan D. Depth-Based Hand Pose Estimation: Methods, Data, and Challenges. *International Journal of Computer Vision.* 2018;126(11):1180–1198. https://doi.org/10.1007/s11263-018-1081-7
21. Ivashechkin M., Mendez O., Bowden R. Improving 3D Pose Estimation for Sign Language. URL: https://doi.org/10.48550/arXiv.2308.09525 (Accessed 25th March 2024).
22. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł. Polosukhin I. Attention Is All You Need. In: *NIPS'17: 31st International Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 30 (NIPS 2017), 4-9 December 2017, Long Beach, CA, USA.* Montreal: Curran Associates; 2017. P. 5998–6008.
23. Goodfellow I., Bengio Y., Courville A. *Deep Learning.* Cambridge: MIT Press; 2016. 800 p.
24. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A. et al. Language Models are Few-Shot Learners. In: *NeurIPS 2020: 34th Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 06-12 December 2020, Vancouver, Canada.* Curran Associates; 2020. P. 1877–1901.
25. Touvron H. et al. LLaMA: Open and Efficient Foundation Language Models. URL: https://doi.org/10.48550/arXiv.2302.13971 (Accessed 5th April 2024).
26. Lugaresi C., Tang J., Nash H. et al. MediaPipe: A Framework for Building Perception Pipelines. URL: https://doi.org/10.48550/arXiv.1906.08172 (Accessed 5th April 2024).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Штехин Сергей Евгеньевич, руководитель группы компьютерного зрения, «Отраслевой центр разработки и внедрения информационных систем» Сириус, филиал № 11, Сочи, Российская Федерация.
e-mail: shs77@bk.ru
ORCID: 0000-0003-2866-4864

Sergey E. Shtekhin, Head of the Computer Vision Group, "Industry center for the development and implementation of information systems" Sirius, branch No. 11, Sochi, the Russian Federation.

Стадник Алексей Викторович, кандидат физико-математических наук, старший специалист по анализу данных группы компьютерного зрения, «Отраслевой центр разработки и внедрения информационных систем» Сириус, филиал № 11, Сочи, Российская Федерация.
e-mail: i@lxstd.ru

Alexey V. Stadnik, Candidate of Physical and Mathematical Sciences, Senior Data Analysis Specialist of the Computer Vision Group, "Industry center for the development and implementation of information systems" Sirius, branch No. 11, Sochi, the Russian Federation.

Статья поступила в редакцию 06.06.2024; одобрена после рецензирования 06.06.2024; принята к публикации 19.06.2024.

The article was submitted 06.06.2024; approved after reviewing 06.06.2024; accepted for publication 19.06.2024.