Ву Минь Куанг, студент
МГТУ им. Н. Э. Баумана

Быстрицкая Анна Юрьевна, к.т.н.
МГТУ им. Н. Э. Баумана

МЕТОД ПЕРЕВОДА ЖЕСТОВ РУК В ТЕКСТ С ИСПОЛЬЗОВАНИЕМ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ

Аннотация. В данной статье рассматривается актуальная задача автоматического распознавания и перевода жестов рук в текст. Предложен метод на основе рекуррентных (RNN) нейронных сетей. Данный подход позволяет эффективно обрабатывать видеопоследовательности жестов, извлекать пространственные особенности и учитывать их временную зависимость для точной классификации символов.

Ключевые слова: Распознавание жестов, рекуррентные нейронные сети (RNN), долгая краткосрочная память (LSTM), компьютерное зрение.

Введение

Дактилология – это особая форма речи, представляющая собой воспроизведение букв алфавита с помощью конфигураций пальцев и их движений. Она является жизненно важным средством коммуникации для глухих и слабослышащих людей. Однако общение между глухими и слышащими людьми часто сопряжено с трудностями из-за отсутствия массового знания жестового языка.

Разработка автоматических систем, преобразовывать дактильные жесты в текст, может значительно снизить коммуникационный барьер. Такие системы найдут применение в образовании, телефонии, общественных службах и умных домах. Задача распознавания букв из дактильного алфавита является частным случаем более общей проблемы распознавания жестов и человеческой активности, которая характеризуется несколькими вызовами:

1. **Вариабельность данных:** жесты одного и того же символа могут по-разному выполняться разными людьми (скорость, размер кисти, анатомические особенности).
2. **Временная зависимость:** дактилология – это не просто набор статических поз, а последовательность, где контекст и динамика движения играют ключевую роль.
3. **Необходимость сегментации:** в непрерывном потоке жестов необходимо точно определить начало и конец каждого знака.

Традиционные подходы, основанные на статистических методах или алгоритмах вроде Hidden Markov Models (HMM), часто не справляются с высокой вариативностью и сложными временными зависимостями. Современные достижения в области глубокого обучения, в частности, рекуррентные нейронные сети, предлагают мощный инструментарий для решения подобных задач.

Постановка задачи распознавания жестов рук

Рассмотрим задачу распознавания жестов рук. Пусть имеется множество объектов S – жесты рук. Каждый элемент из множества S имеет геометрические признаки, такие как кончики пальцев, направление пальцев, контур руки, а также негеометрические признаки – цвет кожи, форма, текстура и другие.

Также имеется I – дискретное изображение сцены, поступающей с вебкамеры пользователя, в которой могут присутствовать жесты рук. При этом сама сцена находится в идеальных условиях, то есть жесты рук различимы (отсутствует взаимное перекрытие элементов руки, имеют место хорошие условия освещения и т.п.). Изображение I представлено цветовым пространством RGB.

Задача заключается в обработке изображения I таким образом, чтобы идентифицировать и распознать находящиеся на нем объекты множества S.

В качестве входных данных программы будут использоваться изображения с устройства видеофиксации (веб-камеры). Ввод входных данных программы должен быть организован посредством захвата видеопотока.

В качестве выходных данных программы будут выступать результаты распознавания жестов, выводимые на экран в виде текстовой надписи.

Структура решения задачи распознавания жестов рук

На рисунке 1 представлена формальная постановка задачи распознавания жестов рук.

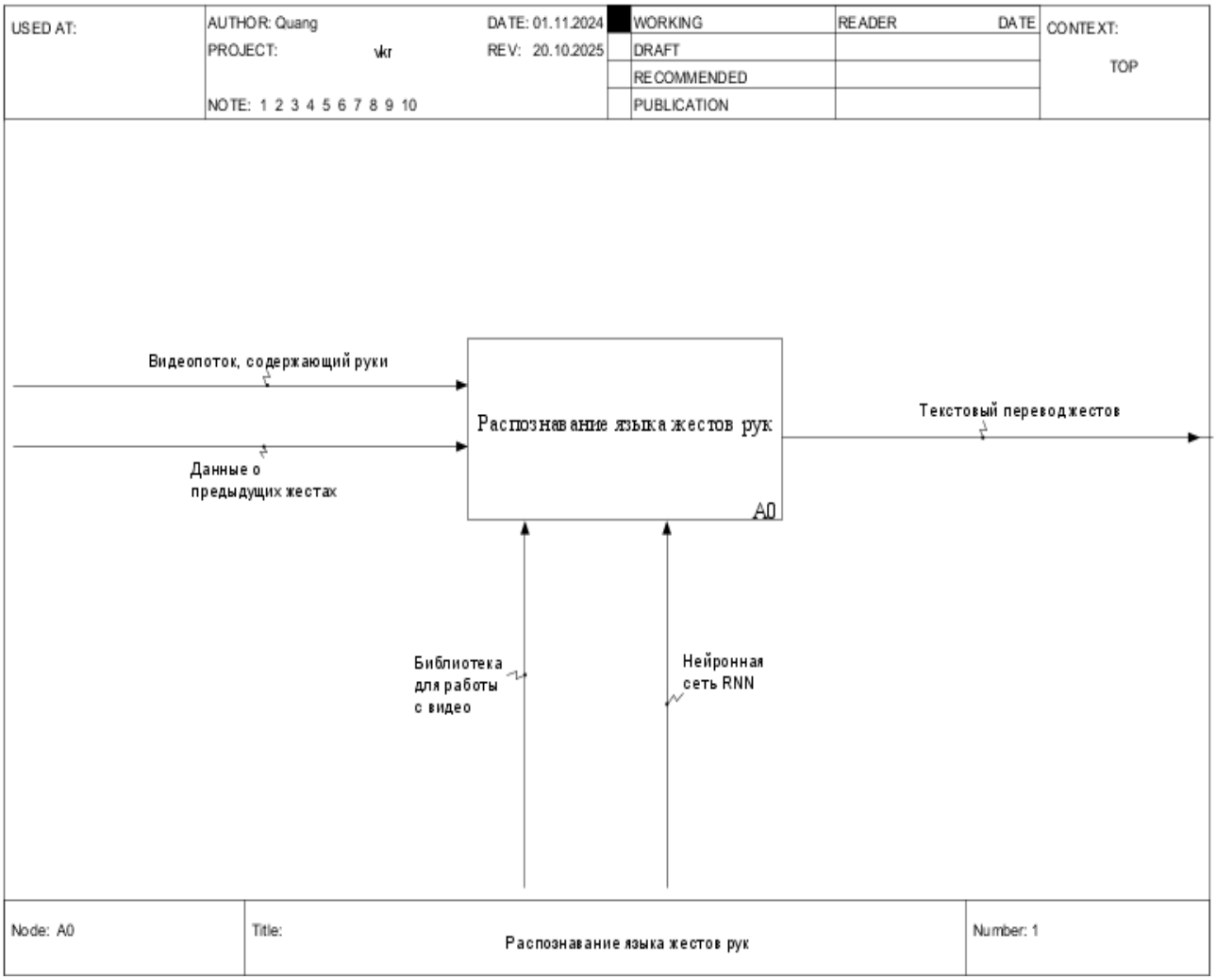

Рис. 1 – Формальная постановка задачи.

На вход подается изображение с веб-камеры пользователя. В качестве механизмов выступают алгоритмы обработки изображений и алгоритмы распознавания образов, участие пользователя и программно-технических средств. С учетом имеющейся базы данных жестов ручной азбуки глухонемых выдается результат распознавания жестов рук (в виде текстовой надписи), при условии, что таковые присутствуют на входном изображении.

Декомпозиция задачи распознавания жестов рук осуществляется в соответствии с рисунком 2.

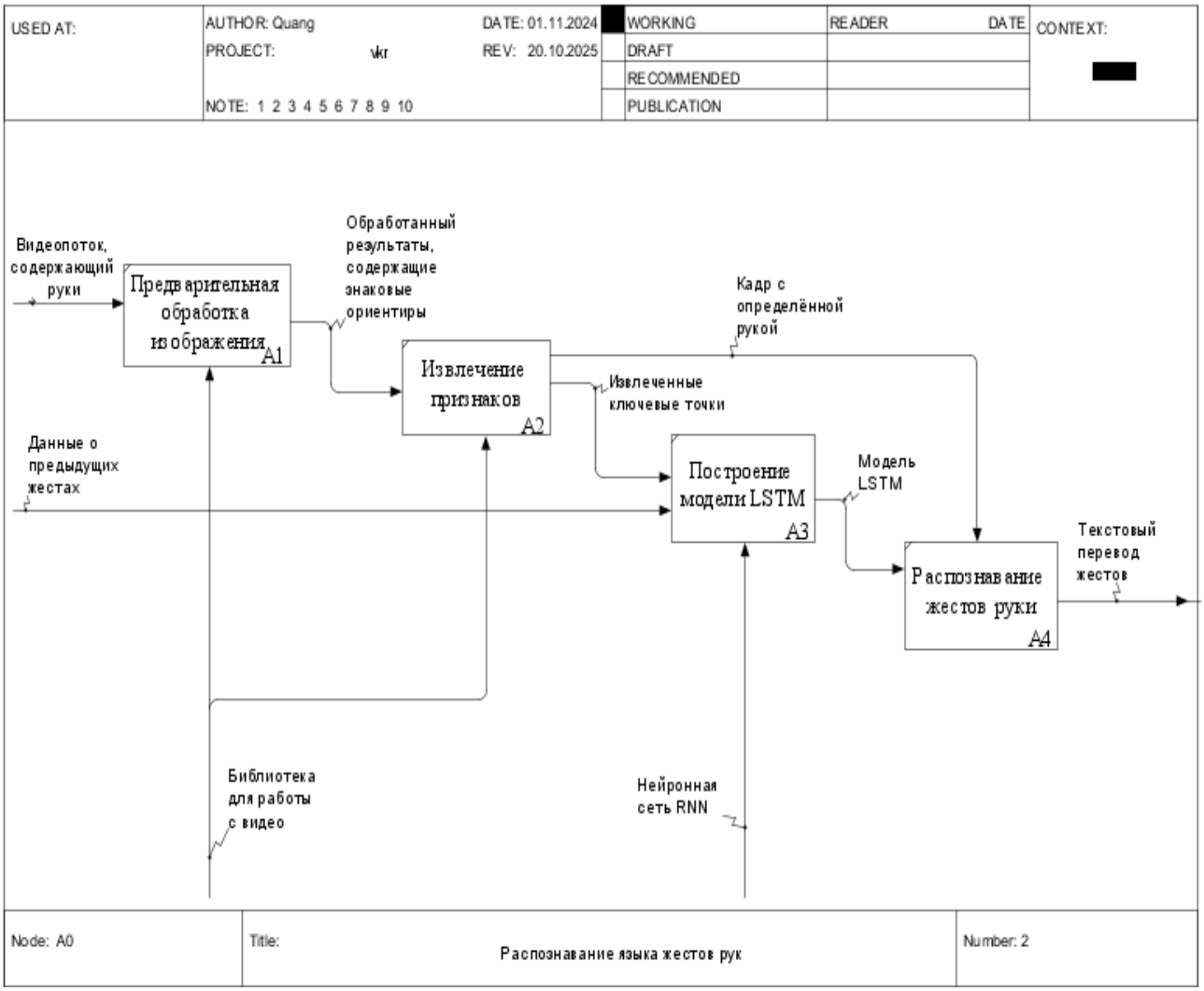

Рис. 2 – Декомпозиция задачи распознавания жестов рук.

Рассмотрим подробнее процесс распознавания жестов рук. Решение данной задачи состоит из следующих этапов: обработка входного изображения (задача обнаружения); выделение области интереса (задача отслеживания); распознавание области интереса (задача распознавания). Задача обнаружения состоит из процедур преобразования цветовых компонентов, вычисления гистограммы и обратной проекции, цифровой обработки. Задача отслеживания включает процедуры нахождения контуров и аппроксимации найденного контура. Решение задачи распознавания основано на рекуррентных нейронных сетях.

Алгоритм распознавания жестов

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

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

Чтобы классифицировать действие, нам сначала нужно найти различные части тела, в частности руки в каждом кадре, а затем проанализировать их движение с течением времени. Первый шаг достигается с помощью Mediapipe Hands [1], который выводит характерные точки ладони (21 ключевых точек) после наблюдения одного кадра в видео.

Кроссплатформенный фреймворк Mediapipe Hands для отслеживания рук и пальцев, предложенный Google, используется в качестве инструмента для извлечения признаков. Это

решение доступно для работы как с онлайн видеопотоком с веб-камеры, так и с фото- и видеофайлами. В случае Python MediaPipe доступен в виде пакета модуля Python. Модель может быть использована для обнаружения ладони и расчета 21 ориентира для каждой обнаруженной руки, где каждый ориентир соответствует определенной точке на ладони и представляет собой три координаты – x, y и z (рисунок 4).

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

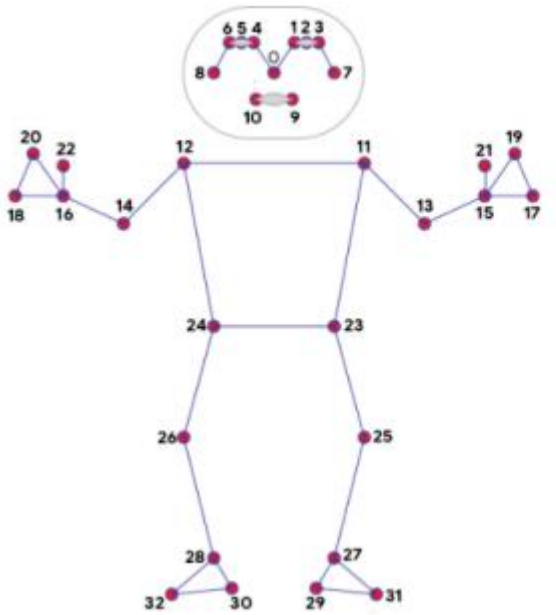

Рис. 3 – MediaPipe Pose.

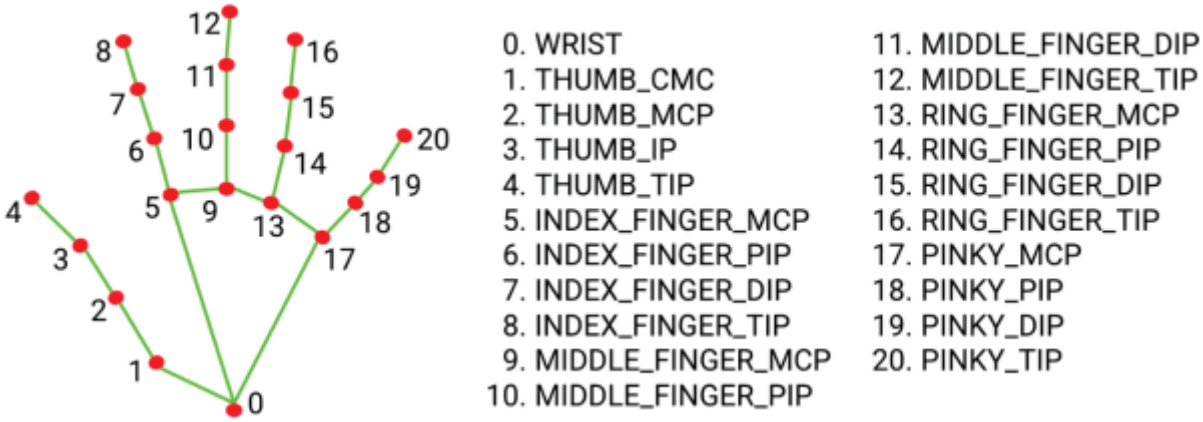

Рис. 4 – MediaPipe Hands.

Под сетью LSTM [2] подразумевается нейронная сеть с рекуррентным слоем LSTM, где ячейка LSTM имеет задачу возвратить новое скрытое состояние h_t, при этом основываясь на прошлом скрытом слое h_{t-1} и вектором представлении текущего значения x_t. В слое LSTM есть одна ячейка, определяемая количеством содержащихся в ней узлов. Ячейка LSTM поддерживает состояние C_t, которое можно рассматривать как внутреннее ее убеждение о текущем состоянии последовательности. Это состояние отличается от скрытого состояния h_t, которое в конечном итоге возвращается ячейкой после последнего временного шага. Состояние C_t имеет ту же длину, что и скрытое состояние (равную количеству узлов в ячейке). На рисунке 5 показано как происходит обновление скрытого состояния отдельной ячейки.

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

Таким образом, модуль распознавания представляет собой нейронную сеть для многоклассовой классификации. Каждому классу соответствует один жест русского жестового языка. Сеть состоит из пяти слоев – трех слоев двунаправленного LSTM для анализа временных зависимостей и два полносвязных слоя для финальной классификации (рисунок 6).

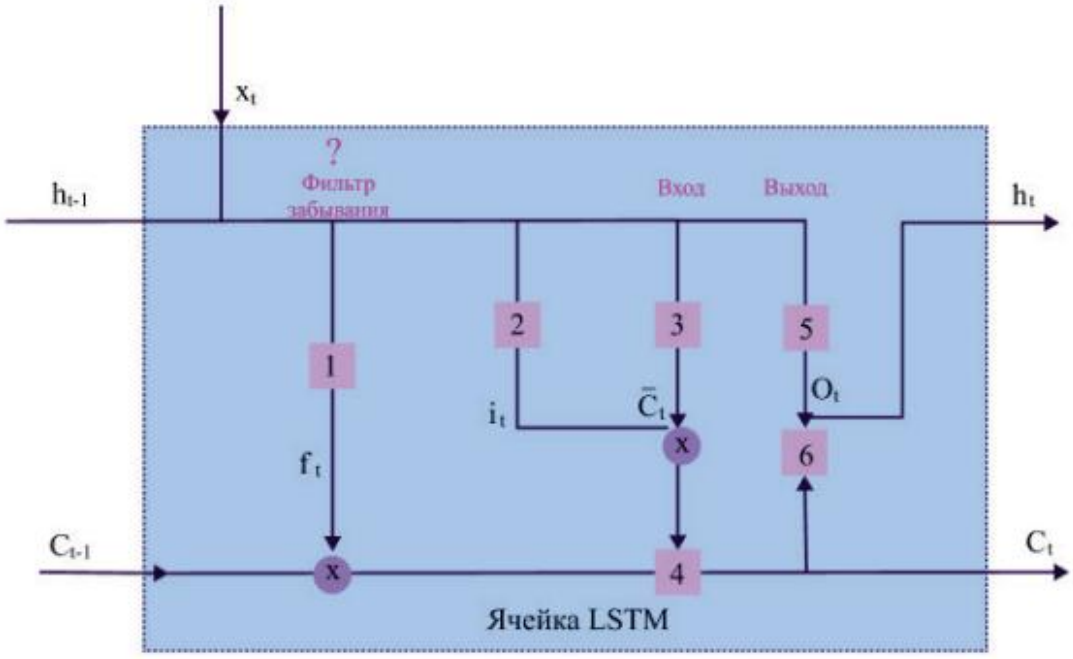

Рис. 5 – Архитектура ячейки LSTM.

Благодаря рекурсивным ссылкам слои LSTM могут эффективно классифицировать данные, продолжающимися во времени, что необходимо для классификации динамических жестов. Первый, второй и третий слои содержат 32, 64 и 32 нейронов соответственно, а полносвязные слои содержат 32 нейронов и имеют функцию активации relu. Выходной слой с 18 нейронами, соответствующими количеству распознанных жестов, использует функцию активации softmax для распределения вероятностей по классам.

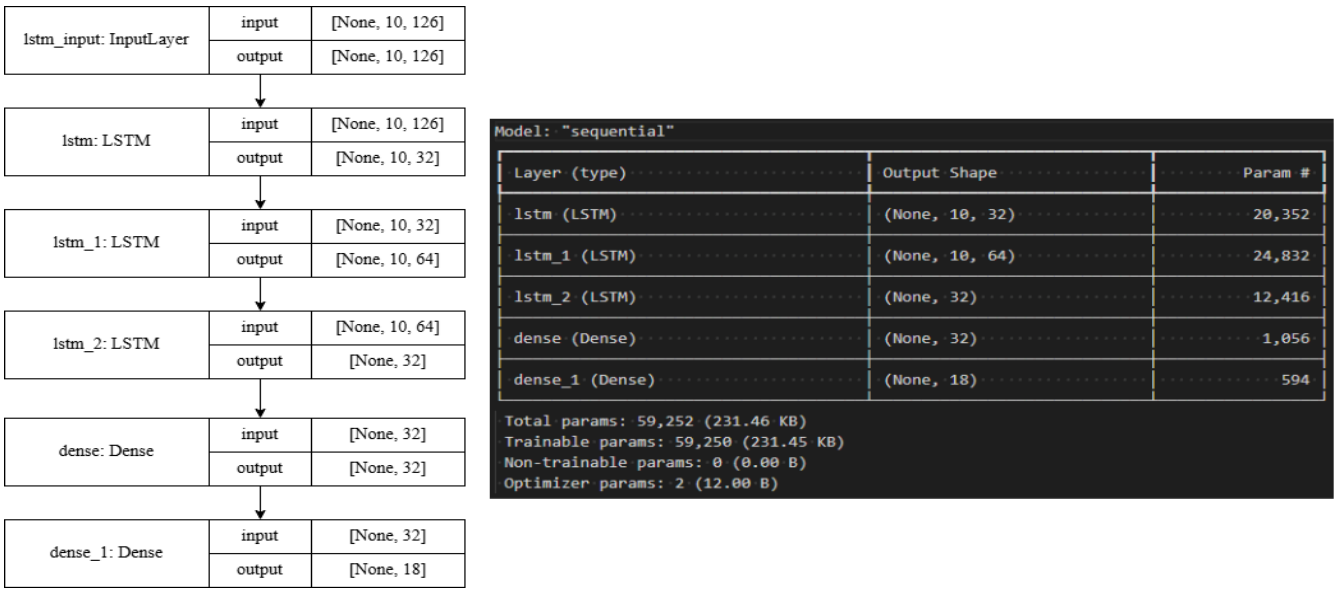

Рис. 6 – Архитектура модели LSTM.

Таким образом, ключевые точки из последовательности кадров отправляются в нейронную сеть LSTM для классификации жестов (рисунок 7).

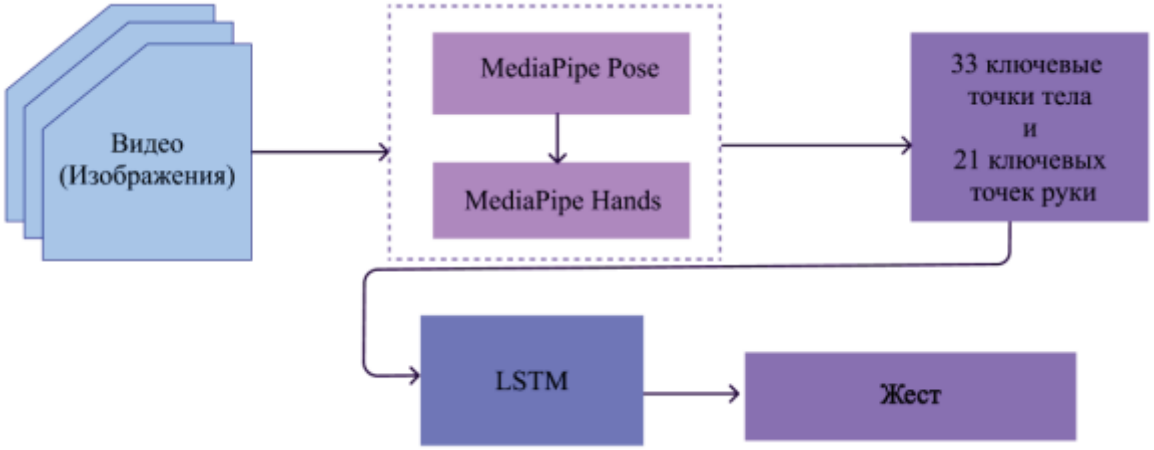

Рис. 7 – Схема работы программы.

Реализация алгоритма распознавания жестов выполнена с использованием технологий проектирования и глубокого обучения нейронных сетей Keras [3], библиотеки компьютерного зрения OpenCV [4], рекуррентной нейронной сети на базе LSTM, фреймворка машинного обучения MediaPipe [1] а также посредством других вспомогательных библиотек.

Заключение

Метод перевода дактильных жестов в текст, основанный на совместном использовании рекуррентных нейронных сетей, представляет собой перспективное направление в области компьютерного зрения и обработки естественных языков для людей с ограниченными возможностями слуха. Данный подход успешно решает ключевые задачи: вариабельность исполнения жестов и необходимость анализа временных последовательностей. Программное решение было апробировано на алфавите глухонемых и может быть использовано при обучении сурдопереводчиков.

Дальнейшее развитие связано с улучшением архитектур (например, применение трансформеров для временных рядов), сбором более обширных и разнообразных датасетов, а также с интеграцией таких систем в мобильные приложения и устройства для обеспечения доступности технологий.

Список литературы:

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

2. Hochreiter, Sepp, and Jürgen Schmidhuber, "Long short-term memory.", Neural computation 9, Vol. 8, с. 1735-1780, 1997.

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

4. OpenCV: [Электронный ресурс]. – Режим доступа URL: https://opencv.org/ (дата обращения: 06.10.2025).