

УДК 004.93

Алгоритм распознавания жестов русского языка на видео

Э. В. Куликова

E-mail: elina.kulikova.2018@inbox.ru

Ярославский государственный университет им. П.Г. Демидова

Аннотация

В работе предложен алгоритм распознавания жестов рук на видео, который позволяет определять и классифицировать 10 цифр и 25 букв русского жестового языка. Основой решения являются цифровые характеристики изображений, полученные инструментом «MediaPipe Hand Landmarker», и классификация с помощью нейронных сетей. Алгоритм реализован на языке программирования Python с использованием библиотек OpenCV, MediaPipe и TensorFlow. Для проведения экспериментов собран собственный корпус данных для русского жестового языка. Лучшие результаты по распознаванию жестов на видео показали многослойный персептрон и одномерная свёрточная нейронная сеть для цифр с F-мерой 0.84 и 0.80 соответственно.

Ключевые слова: Распознавание жестов на видео, классификация жестов, язык жестов, русский жестовый язык.

Введение

Распознавание языка жестов (SLR) — сложная задача, особенно распознавание динамических знаков, зависящих от движения [1]. Русский жестовый язык имеет свои особенности, поэтому требует отдельных создания собственных алгоритмов распознавания. Кроме того, одной из трудностей при разработке алгоритма является сложность сбора достаточного количества данных для обучения модели. Поэтому автор поставил перед собой задачу собрать собственный набор данных и разработать алгоритм распознавания цифр и букв русского жестового языка на основе актуальных методов глубокого обучения.

Существует два типа распознавания жестов рук: на основе носимых перчаток и на основе машинного зрения. Минус первого метода в том, что он дорог и требует ношения на руке специального устройства для распознавания жестов. Второй метод основан на обработке изображений, где последовательность операций выполняется следующим

образом: захват изображения с помощью веб-камеры, сегментация, извлечение признаков и классификация жестов [2]. Именно он является предметом данного исследования.

Сначала исследователи использовали методы извлечения признаков вручную с алгоритмами машинного обучения для классификации жестов рук, но в последнее время в большинстве разработок применяют методы глубокого обучения [2]. В статье [3] представлена реализация модели рекуррентной нейронной сети (RNN) с использованием блоков долговременной кратковременной памяти (LSTM) и плотных слоев для разработки классификатора жестов для управления протезами кисти. В качестве набора данных взяты электромиографические сигналы (разность потенциалов, возникающая в мышцах человека в покое и при их активации) пяти жестов. Каждый жест записывался в течение 20 секунд с помощью специальной ЭМГ-повязки. Для каждого жеста было записано 20000 образцов. Результаты, полученные с помощью предложенной модели, были протестированы на реальных видео, где средняя F-мера (мера точности теста) для 5 классов составила 0.8509.

В работе [4] предлагается модель множественного параллельного потока: двумерная свёрточная нейронная сеть для распознавания поз рук. Предлагаемая модель включает в себя несколько этапов и слоев для определения положения рук по картам изображений. В качестве наборов данных берутся три общедоступных эталонных набора Kaggle (10 классов), First Person (9 классов) и Dexter (7 классов), где количество данных для обучения составляет — 13375, 98842 и 19519 кадров. Средняя F-мера предлагаемого метода составляет 1, 1 и 0.92 при использовании набора данных о положении рук Kaggle, First Person и Dexter соответственно.

В статье [5] рассматриваются подходы к распознаванию жестовых языков глухих на примере русского жестового языка (РЖЯ). Авторами был собран собственный корпус данных для РЖЯ, который включает 35000 вариантов жестов (изображения и 10000 видеофрагментов), построена модель рекуррентной сети (LSTM). Средняя точность правильного распознавания жестов проверялась на реальных видео, где добровольцы показывали предложения с помощью жестов РЖЯ, и она составила 0.95. Однако сам корпус данных не находится в открытом доступе.

Все вышеописанные модели довольно хорошо распознают жесты рук, как на статичных изображениях, так и на видео. Однако каждая

модель распознает небольшое число классов жестов (от 5 до 10) и обучается на больших наборах данных. В данной работе была поставлена задача для большего количества классов как букв, так и цифр.

Сбор корпуса данных

Поскольку не существует общедоступного набора данных для цифр и букв русского жестового языка, то он был создан самостоятельно с помощью добровольцев. Инструкции о том, как показывать жесты были взяты из проекта «Словарь. Русский жестовый язык» (<https://surdo.me>).

Сбор корпуса данных для русского жестового языка включал в себя несколько этапов.

1. Цифры. На первом этапе записывались на видео цифры русского жестового языка. Участвовало 19 добровольцев, из которых 10 женщин и 9 мужчин в возрасте от 20 до 55 лет. Добровольцы показывали цифры от 1 до 10 на правой и левой руке с помощью жестов. Было записано 38 видео и взято одно видео с цифрами от 1 до 10 из проекта «Словарь. Русский жестовый язык».
2. Буквы. На втором этапе записывались на видео буквы русского жестового языка. Участвовало 11 добровольцев, из которых 7 женщин и 4 мужчин в возрасте от 20 до 55 лет. Добровольцы показывали буквы русского алфавита с помощью жестов. Было записано 10 видео и взято одно видео из проекта «Словарь. Русский жестовый язык».
3. На третьем этапе все видео были подвергнуты раскадровке и вручную отобраны лучшие фотографии, где чётко видно жест и он показан достоверно. Для последующей классификации жесты были разбиты на классы. Для цифр классы с наименованием от 1 до 10, которые соответствуют цифрам, и 25 классов для букв с их наименованием соответственно.

Дополнительно были сняты 2 видео для проверки качества распознавания и классификации жестов в реальных условиях эксплуатации.

Обработка корпусов данных

Оба корпуса данных с цифрами и буквами русского жестового языка (RSL) были переработаны решением «MediaPipe Hand Landmarker». Оно является частью проекта «MediaPipe» — это фреймворк с открытым исходным кодом, представленный Google, который помогает

создавать мультимодальные конвейеры машинного обучения. Задача «MediaPipe Hand Landmarker» позволяет обнаружить ориентиры рук на изображении. Его можно использовать для локализации ключевых точек рук и визуализации ориентиров.

Было решено разделить каждый набор данных на два. Один содержит документы в формате JSON, где каждая фотография набора представляет собой следующую информацию: сколько рук обнаружено на фото и массив данных длиной 43 элемента. Массив содержит ключевые точки обнаруженных рук в формате x , y и z : 21 ориентир для правой руки, 21 ориентир для левой руки и точка соотношения положений двух рук относительно друг друга. Если на фотографии обнаружена одна рука, то элементы с индексами после 21 будут нулевыми. Все координаты также были нормализованы от 0 до 1 относительно координат изображения.

Второй набор — изображения, содержащие кисти рук, для каждого жеста с разрешением 64×64 .

Для цифр русского жестового языка теперь существует два корпуса данных: 10 классов по 300 документов в формате JSON для каждого жеста и 10 классов по 300 цветных фотографий для каждого жеста только обнаруженных рук разрешением 64×64 . А для букв русского жестового языка — 25 классов по 8 документов в формате JSON для каждого жеста, которые содержат информацию о 50 кадрах, в течение которых, показывался жест, то есть массивы размером 50×43 ориентиров рук. И 25 классов по 8 папок для каждого человека с 50 цветными фотографиями для каждого жеста только обнаруженных рук разрешением 64×64 . Если жест показывался меньше чем за 50 кадров, то в случае точек недостающие данные дополнялись методом интерполяции, а в случае изображений дополнялись копией предыдущего.

Алгоритм распознавания жестов

Для работы с видео используется библиотека с открытым кодом алгоритмов компьютерного зрения, обработки изображений и численных алгоритмов общего назначения — OpenCV. Она предоставляет очень простой интерфейс для захвата видео с камеры, в данном случае, встроенной веб-камеры на ноутбуке.

Для распознавания рук на видео используется решение «MediaPipe Hand Landmarker», которое локализует на кадре ключевые точки правой и/или левой руки и может визуализировать их (пример на рис. 1).

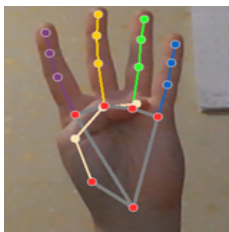


Рис. 1: Пример работы «MediaPipe Hand Landmarker»

Собранные корпуса данных для русского жестового языка будут использоваться для обучения моделей нейронных сетей, главной задачей которых является научиться правильно классифицировать цифры и буквы языка жеста. Были выбраны четыре модели: для распознавания жеста по ключевым точкам руки — многослойный перцептрон (MLP), одномерная свёрточная нейронная сеть (1D CNN) и сеть долгой краткосрочной памяти (LSTM), а также по фото — двумерная свёрточная сеть (2D CNN).

Поскольку корпус данных для русского жестового языка мал, то было решено также провести эксперименты с применением процесса аугментации. Использование такого инструмента может улучшить устойчивость моделей к изменениям входных данных и улучшить их общую производительность.

Алгоритм распознавания языка жестов включает в себя следующие шаги:

Шаг 1. Подключение решения «MediaPipe Hand Landmarker» в режиме одиночного обнаружения (каждый кадр видео рассматриваем отдельно) и минимальным значением достоверности — 0.6 (то есть при оценке ≥ 0.6 обнаружение рук считается успешным);

Шаг 2. Захват живого потока с веб-камеры;

Шаг 3. Первичная обработка каждого кадра видео для распознавания руки:

- Прочитать кадр;
- Перевернуть кадр вокруг оси Y для правильного вывода рук;
- Преобразовать изображение BGR в RGB и передать его на обработку в конвейер решения «MediaPipe Hand Landmarker», где кадр пройдёт через модель обнаружения ладони, а потом пройдёт через модель ориентира руки. В результате, получим 21 ориентир для каждой обнаруженной руки на данном видеокadre;

Шаг 4. Вторичная обработка каждого кадра видео для классификации жеста:

- Если выбрана модель двумерной свёрточной сети. Кадр обрезается с учётом найденных на прошлом шаге ориентиров обнаруженных рук, размер изображения меняется на 64×64 , кадр нормализуется и отправляется на вход обученной 2D CNN, где происходит предсказание с использованием обученной нейронной сети;
- Если выбрана одна из моделей, обученная на ключевых точках рук (MLP /1D CNN /LSTM). Представить кадр как массив из 43 точек, где 21 ориентир для правой руки, 21 ориентир для левой руки, вычисленные на прошлом шаге, и посчитанная точка соотношения положений двух рук относительно друг друга с помощью алгоритма Евклида. Если на кадре была обнаружена одна рука, то элементы с индексами после 21 будут нулевыми. Далее координаты нормализуются от 0 до 1 и массив данных отправляется на вход обученной сети, где она по ключевым точкам рук пытается предсказать какой жест изображён на кадре;

Шаг 5. На кадре визуализируются ориентиры обнаруженных рук и отображается символ, соответствующий жесту (пример на рис. 2);

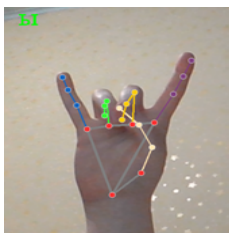


Рис. 2: Пример распознавания руки и классификации жеста

Шаг 6. Шаги 3–5 повторяются до тех пор, пока пользователь не прекратит использовать режим распознавания;

Шаг 7. Отпустить захват видеопотока и освободить ресурсы.

На основе данного алгоритма было создано приложение, распознающее жестовый язык. Входными данными для приложения является видеопоток, захваченный веб-камерой. В качестве выходных данных пользователь может получить видеоизображение с расшифровкой в виде текстового предсказания распознанного жеста. Прило-

жение реализовано на языке программирования Python с использованием библиотек OpenCV, MediaPipe, TensorFlow и разработанного алгоритма распознавания жестов на основе обученной ранее нейронной сети.

Экспериментальные исследования

Во время экспериментов проводились исследования по влиянию архитектуры и параметров обучения на качество работы нейронных сетей. Были протестированы различные комбинации архитектур и параметров, включая количество слоев, количество нейронов в каждом слое, функции активации, скорость обучения и размер пакета данных.

Нейронные сети обучались на корпусах данных, описанных в разделах «Сбор корпуса данных» и «Обработка корпусов данных». Каждый корпус данных был разделён на три выборки: обучающую (60%), валидационную (20%) и тестовую (20%). Обучение проводилось на обучающей выборке, а результаты оценивались на валидационной и тестовой выборках.

Эксперименты проводились в три этапа. Первый этап включал в себя создание таких моделей сетей, которые давали наилучший результат классификации цифр русского жестового языка, второй этап - букв, а третий одновременно цифр и букв. В итоге, было разработано 4 основных моделей для трёх разных случаев. Подобранные эмпирическим путём параметры сетей представлены в таблице 1.

| Этап | Модель | Кол-во слоёв | Кол-во нейронов на каждом слое | Гиперпараметры | |
|---------------|--------|--------------|--------------------------------|----------------|---------------|
| | | | | Кол-во эпох | Размер пакета |
| Цифры | MLP | 4 | 129, 128, 256, 10 | 50 | 32 |
| | 1D CNN | 3 | 64, 16, 10 | 120 | 128 |
| | LSTM | 6 | 64, 128, 64, 64, 32, 10 | 150 | 64 |
| | 2D CNN | 5 | 32, 64, 64, 512, 10 | 15 | 64 |
| Буквы | MLP | 4 | 129, 128, 64, 25 | 30 | 64 |
| | 1D CNN | 3 | 64, 16, 25 | 50 | 128 |
| | LSTM | 6 | 64, 128, 64, 64, 32, 25 | 170 | 128 |
| | 2D CNN | 5 | 32, 64, 64, 512, 25 | 15 | 64 |
| Цифры и буквы | MLP | 4 | 129, 256, 128, 35 | 50 | 248 |
| | 1D CNN | 4 | 64, 32, 64, 35 | 70 | 248 |
| | LSTM | 5 | 64, 64, 64, 32, 35 | 170 | 512 |
| | 2D CNN | 5 | 32, 64, 64, 512, 35 | 30 | 128 |

Таблица 1: Параметры нейронных сетей

Проверка качества обучения каждой нейронной сети проводилась на соответствующей тестовой выборке на каждом этапе экспериментов. В качестве метрик были выбраны метрики точности, полноты и

F-меры, так как они часто применяются в задачах классификации. Данная оценка модели показывает, насколько хорошо сеть обобщает данные и способна предсказывать результаты для новых данных. Кроме того, были произведены испытания на реальных видеоданных. Такое опробование является критически важным для обеспечения высокой производительности каждой модели в реальных условиях эксплуатации.

Ниже на таблицах 2 – 4 представлены вычисленные метрики качества до аугментации и после для каждой модели нейронной сети на каждом этапе исследования:

1. Классификация жестов цифр:

| Модель | До аугментации | | | После аугментации | | |
|--------|----------------|---------|--------|-------------------|---------|--------|
| | Точность | Полнота | F-мера | Точность | Полнота | F-мера |
| MLP | 0.81 | 0.79 | 0.78 | 0.84 | 0.86 | 0.84 |
| 1D CNN | 0.78 | 0.78 | 0.77 | 0.80 | 0.82 | 0.80 |
| LSTM | 0.46 | 0.50 | 0.43 | 0.77 | 0.72 | 0.70 |
| 2D CNN | 0.60 | 0.60 | 0.59 | 0.65 | 0.65 | 0.63 |

Таблица 2: Оценка классификации русских жестов цифр на видео

Модели MLP и 1D CNN лучше всего показали себя в классификации цифр на видео в реальном времени со средней F-мерой 0.84 и 0.80.

2. Классификация жестов букв:

| Модель | До аугментации | | | После аугментации | | |
|--------|----------------|---------|--------|-------------------|---------|--------|
| | Точность | Полнота | F-мера | Точность | Полнота | F-мера |
| MLP | 0.12 | 0.24 | 0.11 | 0.41 | 0.50 | 0.42 |
| 1D CNN | 0.43 | 0.52 | 0.44 | 0.52 | 0.61 | 0.53 |
| LSTM | 0.00 | 0.07 | 0.01 | 0.09 | 0.19 | 0.09 |
| 2D CNN | 0.04 | 0.09 | 0.05 | 0.17 | 0.24 | 0.17 |

Таблица 3: Оценка классификации русских жестов букв на видео

Модель 1D CNN и MLP лучше всего показали себя в классификации букв на видео в реальном времени со средней F-мерой 0.53 и 0.42 при использовании инструментов аугментации. Хуже всего качество модели LSTM и 2D CNN, с помощью них сложно

классифицировать жесты букв на реальном видео. Причина может быть в малом количестве данных и плохо подобранных параметров сети.

3. Классификация жестов цифр и букв:

| Модель | До аугментации | | | После аугментации | | |
|--------|----------------|---------|--------|-------------------|---------|--------|
| | Точность | Полнота | F-мера | Точность | Полнота | F-мера |
| MLP | 0.27 | 0.26 | 0.25 | 0.29 | 0.36 | 0.31 |
| 1D CNN | 0.24 | 0.30 | 0.23 | 0.22 | 0.28 | 0.23 |
| LSTM | 0.12 | 0.14 | 0.11 | 0.15 | 0.14 | 0.13 |
| 2D CNN | 0.17 | 0.19 | 0.16 | 0.26 | 0.20 | 0.17 |

Таблица 4: Оценка классификации русских жестов цифр и букв на видео

Все модели нейронных сетей недостаточно хорошо классифицируют и цифры, и буквы на реальных видео. Это может объясняться тем, что было малое количество данных для обучения и взято мало ракурсов фотографий. Хуже всего на видео показали себя модели LSTM и 2D CNN. Причиной может быть недостаточное количество данных и неправильный выбор параметров обучения.

Заключение

В результате работы был самостоятельно собран корпус данных русского жестового языка и разработан алгоритм с использованием четырёх созданных моделей нейронных сетей: MLP, 1D CNN, LSTM и 2D CNN, который способен определять жесты 10 цифр и 25 букв русского жестового языка на видео. Лучшие результаты по распознаванию на видео показали модели MLP и 1D CNN для распознавания жестов цифр с F-мерой 0.84 и 0.80 соответственно.

Список литературы

1. Abdallah M., Gerges Samaan A. W., Fazliddin Makhmudov Y.-I. C. Light-Weight Deep Learning Techniques with Advanced Processing for Real-Time Hand Gesture Recognition // Sensors. 2023. Vol. 23, no. 2. P. 34–64.

2. *Miah A. S. M., Al Mehedi Hasan J. S., Yuichi Okuyama Y. T.* Multistage Spatial Attention-Based Neural Network for Hand Gesture Recognition // *Computers*. 2023. Vol. 12, no. 13. P. 20–31.
3. *A. T.-O., Jaramillo-Tigreros J. T. J., Peña A. L.-G. A. C. R.* Hand Gesture Recognition Using EMG Signals // *Applied Sciences*. 2022. Vol. 12, no. 19. P. 45–66.
4. *Noreen I. and Hamid M. A. U., Malik S. S. M.* Hand Pose Recognition Using Parallel Multi Stream CNN // *Sensors*. 2021. Vol. 21, no. 24. P. 69–84.
5. *М. Г. Гриф и Р. Элаккия А. Л. Приходько и М. А. Бакаев Е. Р.* Распознавание русского и индийского жестовых языков на основе машинного обучения // *Системы анализа и обработки данных*. 2021. Т. 83, № 3. С. 53—74.