Система распознавания жестов в реальном времени с использованием MediaPipe

М. В. Хнюнин, email: xnyunin.2016@stud.nstu.ru
М. Г. Гриф, email: grifmg@mail.ru

Новосибирский государственный технический университет

Аннотация. *Рассмотрены существующие разработки распознавания жестов при помощи компьютерного зрения, оценены возможности применения библиотеки MediaPipe для решения задачи распознавания жестовой речи, разработан подход, позволяющий распознавать русскую жестовую речь в реальном времени с одной камеры, без дополнительного оборудования.*

Ключевые слова: *распознавание жестов, русский жестовый язык, реальное время, MediaPipe.*

Введение

Существует множество сфер, где требуется распознавать жесты рук, показываемые человеком: управление вспомогательными роботами в медицине [1], удалённое управление станками на производствах в условиях повышенного шума, когда неприменимо голосовое управление, или в строительстве для управления автоматическими кранами, в автомобилях для реализации управления функционалом без отведения взгляда водителя от дороги [2], а также в индустрии интерактивных развлечений и управлении домашними бытовыми приборами, в том числе различными мультимедийными системами и телевизорами [3]. Однако, самым востребованным, и в то же время сложным в реализации является распознавание жестовой речи для преодоления коммуникативного барьера между людьми с ограниченными возможностями по слуху и теми, кто не знает жестовый язык. На данный момент для общения глухого и слышащего требуется сурдопереводчик, чьи услуги платные и вносят в общение дополнительные неудобства от посредничества, т.к. не может быть соблюдена конфиденциальность диалога. Существуют иные способы общения людей, такие как текстовая переписка, но это значительно медленнее общения жестовым языком и создаёт дополнительный дискомфорт для людей с ограниченными возможностями. Уже разработаны системы, способные генерировать движения виртуального аватара для преобразования текста в жестовую речь, но распознавание

непрерывной жестовой речи компьютером остаётся не решённой задачей. Основную сложность вызывает то, что нельзя придумать наиболее простые для распознавания статические жесты, а нужно распознавать уже существующий жестовый язык, который отличается в разных странах. К тому же, использование дополнительного оборудования, такого как датчики глубины или специальные перчатки не позволит добиться повсеместной применимости системы, она должна работать посредством простой камеры, например в смартфоне, и производить распознавание в реальном времени с учётом низкой ресурсоёмкости мобильных устройств. Все эти ограничения усложняют задачу распознавания жестовой речи и требуют иных подходов, нежели в задачах автоматизации управления жестами.

1. Анализ существующих разработок

Существующие разработки в области распознавания жестов можно разделить на три основные категории: использующие специальные перчатки и браслеты для считывания жестов, использующие стереокамеры, а также системы из камер и датчиков глубины, такие как Microsoft Kinect, и системы, основанные только на компьютерном зрении с использованием одной камеры [3]. При этом важнейшими параметрами данных систем является точность распознавания жестов и производительность. Также на точность значительно влияет качество изображений, освещённость в кадре, сложность фона и специфичность условий применения, под которые адаптирован алгоритм. Кроме этого, при распознавании непрерывного жестового языка возникает проблема коартикуляции в жестовых языках, которую никто до сих пор не может полностью решить [4].

Распознавание кисти руки и её положения в пространстве – это лишь первый этап. Далее следует классификация жестов. Для этого также может быть использована свёрточная нейросеть как для анализа всего изображения, так и для координат ключевых точек пальцев рук или нормализованного положения скелета руки. Но в случае с базовыми свёрточными нейронными сетями, возможно анализировать лишь каждый кадр по отдельности и для распознавания динамичных жестов следует анализировать целый временной промежуток, например при помощи сетей долгой краткосрочной памяти (англ. Long short-term memory; LSTM), как представлено в недавней работе индийских исследователей [5]. Но такой подход очень ресурсоёмкий и даже при распознавании всего 10 жестов отстаёт от реального времени при работе на базе персонального компьютера. Таким образом, остаётся лишь оптимизации применения нейронных сетей за счёт снижения числа значимых компонент при классификации, а также за счёт использования

больших объёмов данных для обучения; или алгоритмические подходы, где сложность классификации уменьшается за счёт предварительного аналитического анализа жеста и использования различных нейросетей для разных подклассов жестов или их компонент. Но на данный момент нет доступной информации о разработках, использующих такой подход т.к. он требует совершенно иного подхода к построению массива данных для обучения.

2. Возможности применения библиотеки MediaPipe

В рамках поставленной задачи приемлемо использование лишь тех подходов, которые для распознавания руки используют одну камеру, направленную на человека, показывающего жесты. Из существующих разработок, предназначенных для поиска рук в кадре и распознавания их положения наиболее развитым является программное решение от компании Google – MediaPipe [6]. Кроме этого, данная библиотека позволяет определять как само наличие человека в кадре, так и позу его тела и ключевые точки мимики лица. Это важно, т.к. в русской жестовой речи большое значение имеет то, где именно относительно тела и лица показывается жест, какие при этом происходят касания как рук относительно друг друга, так и рук с телом и лицом. А в исследовании учёных университета Оксфорда [4] особое внимание уделено мимике губ и выражению эмоций во время распознавания британской жестовой речи, что также приемлемо и для РЖЯ. Поэтому применение библиотеки MediaPipe позволяет значительно упростить предварительный сбор данных из кадра, которые требуются для последующей классификации жеста, и такие решения уже существуют, например, в упомянутой выше работе с использованием сети долгой краткосрочной памяти [5].

3. Существующие массивы данных для обучения

В задачах распознавания жестов существует множество массивов данных, однако все они предназначены для определённого типа задач. Наиболее популярные из них представлены в таблице.

Таблица

Общедоступные массивы данных с жестами рук

Массив данных	Экземпляры	Видео	Классы	Сцены
ChaLearn ConGD	47,933	22,535	249	15
nvGesture	1,532	1,532	25	1
EgoGesture	24,161	2,081	83	6
IPN Hand	4,218	200	13	28

Из представленных массивов в таблице, первый, ChaLearn ConGD достаточно обширный, но в нём данные по китайскому жестовому языку и его диалектам; второй массив из таблицы предназначен для управления функциями автомобиля и данные в этом массиве представляют собой видео в соответствующем ракурсе [2]; третий массив записан от первого лица, а последний содержит лишь 13 классов, т.е. всего 13 распознаваемых жестов. Примеры этих массивов представлены на рис. 1.

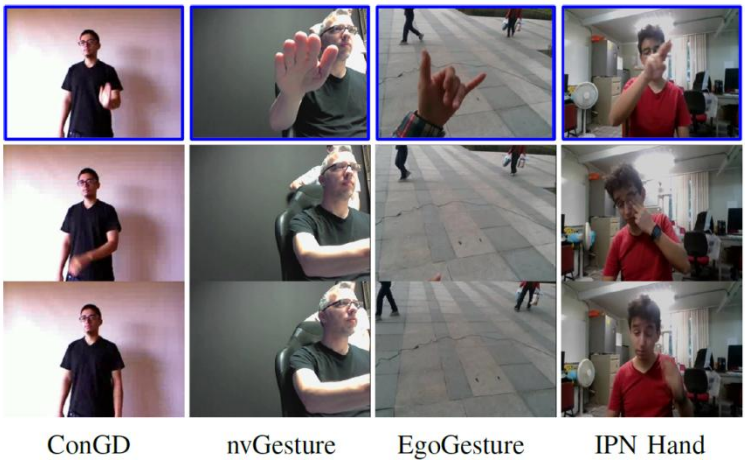

Рис. 1. Примеры общедоступных массивов данных с жестами рук

Т.к. задачей нашей работы является распознавание жестов русского жестового языка, рассмотрим отечественные разработки. Наиболее развитой на данный момент является разработка компании «SberDevices», массив HaGRID (HAnd Gesture Recognition Image Dataset). Данный массив был собран и размечен в течении нескольких месяцев с помощью краудсорсинг платформ «Яндекс Толока» и «ABC Элементари». При этом за использования этих платформ для создания массива из 552992 изображений компания заплатила более 50000$. И это для обучения системы распознавания лишь 18 жестов, используемых для управления телевизором [7].

Русский жестовый язык (РЖЯ) обширен и неизвестно точное число существующих жестов. На данный момент разрабатывается документация по стандартизации, и можно сделать вывод, что создание массива данных для обучения свёрточной нейронной сети, способной

распознавать хотя бы наиболее используемые в РЖЯ жесты, требует слишком большого числа ресурсов. Таким образом, при всех условиях задачи, реализуемым является подход с предварительным аналитическим анализом и распределением жестов на несколько подклассов для использования различных нейросетей для разных подклассов жестов.

4. Предлагаемый подход

В связи с тем, что существующие решения для распознавания некоторых динамических жестов оказываются ресурсоёмкими, а достаточных для обучения массивов данных с фотографиями и видео жестов из РЖЯ не существует, было принято решение разделить задачу на несколько основных этапов, выполняемых для распознавания:

1. Определение положения руки, тела и лица посредством библиотеки MediaPipe, а именно модулей «Holastic» и «Hands». Проверка на реалистичность распознанного положения пальцев также проводится посредством функционала библиотеки MediaPipe.

2. Оптимизация собранных данных за счёт сокращения числа переменных, а именно перехода от 21 ключевой точки с координатами в трёх плоскостях к 15 углам сгиба пальцев и двум параметрам - градусам поворота и наклона запястья. А также два дополнительных параметра – расстояния от запястий руки до груди и до лица человека. Итого 19 численных значений для одной руки и 38 значений для двух рук вместо 1629 и 1660 соответствующих численных значений, предоставляемых используемыми модулями библиотеки MediaPipe.

3. Предварительная классификация жеста в текущем кадре на основе сравнения с предыдущим распознанным жестом для выявления четырёх основных состояний и 14 соответствующих вариантов с учётом всех направлений: статичное положение, вращение по вертикали и горизонтали в каждую сторону, движение и его направление (одно из 6), а также касания (с грудью, с лицом, со второй рукой). Таким образом, в результате данного этапа определяются основные составляющие жестов для последующей классификации.

4. Классификация жеста на основе нейросети, обученной для соответствующего подкласса, определённого на предыдущем этапе за счёт выделения ключевых компонент.

Таким образом, при выполнении всех четырёх этапов может быть реализована система распознавания русской жестовой речи, которая будет достаточно ресурсоёмкой, чтобы работать на базе мобильных устройств и персональных компьютеров в реальном времени.

Заключение

Работа по распознаванию жестов РЖЯ на данный момент находится на этапе сбора видео материалов с жестами, и их классификация для распределения на подклассы, основываясь на ключевых компонентах жестов. Основной сложностью является выявление оптимального числа подклассов с минимальным пересечением для последующего обучения отдельных нейронных сетей жестам из каждого подкласса. Для этого предстоит определить значимость каждой компоненты жеста в общей классификации.

Список литературы

1. Amsterdam, B. Gesture Recognition in Robotic Surgery: a Review /B. Amsterdam, M. Clarkson, D. Stoyanov / IEEE Transactions on Biomedical Engineering. – 2021. – Volume 68. – Issue 6. – P. 2021-2035.
2. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks / P. Molchanov et al. // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 4207– 4215.
3. Oudah, M. Hand Gesture Recognition Based on Computer Vision: A Review of Techniques / M. Oudah, A. Al-Naji, J. Chahl // Journal of Imaging. – 2020. – Volume 6. – Issue 8. – P. 73-102.
4. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues [Электронный ресурс] / S. Albanie et al. // European Conference on Computer Vision. – 2020. – Режим доступа: https://arxiv.org/abs/2007.12131.
5. Real-Time Hand Gesture Recognition System Using MediaPipe and LSTM / S. Agrawal, et al. // International Journal of Research Publication and Reviews. – April 2022. – Volume 3. – No 4. – P. 2509-2515.
6. MediaPipe Hands: On-device Real-time Hand Tracking [Электронный ресурс] / Fan Zhang et al. // CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, USA. – 2020. – Режим доступа: https://arxiv.org/abs/2006.10214.
7. Kapitanov, A. HaGRID – HAnd Gesture Recognition Image Dataset [Электронный ресурс] / A. Kapitanov, A. Makhlyarchuk, K. Kvanchiani // Computer Vision and Pattern Recognition. – 2022. – Режим доступа: https://doi.org/10.48550/arXiv.2206.08219.