**МИНОБРНАУКИ РОССИИ**

**Федеральное государственное бюджетное образовательное**

**учреждение высшего образования**

**«Ярославский государственный университет им. П.Г. Демидова»**

Кафедра теоретической информатики

Сдано на кафедру

« » 2023 г.

Заведующий кафедрой,

д. ф.-м. н., профессор

Е. В. Кузьмин

Выпускная квалификационная работа

**Алгоритм распознавания жестов рук на видео**

по направлению подготовки 02.04.02 Фундаментальная информатика и

информационные технологии

Научный руководитель

к. ф.-м. н., доцент

Н. С. Лагутина

« » 2023 г.

Студент группы ИТ-21МО

Э. В. Куликова

« » 2023 г.

Ярославль 2023 г.

**Реферат**

**Введение**

Распознавание жестов — сложная задача, которая уже давно находится в центре внимания исследований компьютерного зрения и машинного обучения. Распознавание жестов рук, в частности, становится все более важной темой в последние годы, поскольку оно имеет широкий спектр приложений в таких областях, как виртуальная и дополненная реальность и взаимодействие человека и робота.

Ещё важным применением распознавания жестов является помощь глухим, слабослышащим или немым людям. Для этих людей язык жестов часто является основным средством общения. Точно распознавая и интерпретируя жесты рук, технологии могут помочь преодолеть разрыв в общении между этими людьми и остальным обществом. Это может улучшить качество их жизни и повысить их способность участвовать в самых разных видах деятельности.

Поэтому одной из ключевых задач сейчас является возможность точного отслеживания и интерпретации жестов рук в режиме реального времени, особенно на видео.

Существует несколько ключевых проблем связанных с распознаванием жестов рук в видеорядах: руки постоянно находятся в движении, меняют углы наклона и перекрывают друг друга, изменчивость формы и размеров рук. Всё это может затруднить создание универсального алгоритма.

Кроме того, жесты рук сильно зависят от контекста, а это означает, что один и тот же жест может иметь разное значение в зависимости от ситуации. Например, жест рукой, означающий «стоп» в одном контексте, может означать «подойти» в другом.

В этой работе проводятся исследования некоторых ключевых методов и алгоритмов, используемых для распознавания жестов рук на видео, а также будет рассмотрено их потенциальное применение в области русского жестового языка.

**Постановка задачи**

В ходе данной выпускной квалификационной работы было поставлено несколько задач. Во-первых, исследовать существующие методы обнаружения рук на видеопотоке. Во-вторых, разработать и реализовать алгоритм распознавания и классификации цифр и букв русского жестового языка на видео в реальном времени. В-третьих, применить алгоритм на цифрах и буквах американского жестового языка и сравнить полученные результаты с результатами аналогичного алгоритма на данных русского жестового языка.

Поставленные задачи включают себя следующие подзадачи:

* + Самостоятельное создание корпуса данных с цифрами и буквами жестового русского языка;
  + Подбор корпуса данных с цифрами и буквами жестового американского языка;
  + Обработка наборов данных для получения дополнительных характеристик, которые могут быть использованы для классификации цифр и букв;
  + Проведение экспериментов по классификации цифр и букв русского и американского жестового языка.

1. **Описание предметной области**
   1. **Язык жестов**

В настоящее время более 5% населения мира (около 430 миллионов человек) нуждаются в реабилитации для решения проблемы инвалидизирующей потери слуха. Потеря слуха более широко распространена среди людей в возрасте старше 60 лет. Однако, и около 1,1 миллиарда молодых людей подвергаются риску потери слуха из-за прослушивания музыки на слишком высоком, опасном для слуха уровне громкости. Согласно прогнозам, к 2050 г. почти 2,5 миллиарда человек будут страдать от проблем со слухом в той или иной степени []. Обеспечение возможности общения этих людей в обществе является важной социальной задачей, чтобы устранить психологические, идеологические и межкультурные барьеры.

Язык жестов является основным средством коммуникации с глухими или слабослышащими людьми. Это естественный язык, на котором общаются с помощью жестов, а не звуков. Он используется для передачи информации и выражения эмоций и чувств. Жестовый язык включает в себя различные жесты, каждый из которых имеет свое значение. Эти жесты могут использоваться для передачи слов и фраз, а также для выражения концепций и идей. Даже в пределах одной языковой группы существуют значительные различия, поэтому разработка системы преобразования языка жестов в речь для улучшения взаимодействия между людьми является острой необходимостью. К тому же, в разных регионах и странах есть свои языки жестов, из-за этого могут возникнуть коммуникативные трудности, в особенности, у тех, кто не знает местного языка жестов.

Русский жестовый язык изучен недостаточно хорошо, использование знаков варьируется от региона к региону. Помимо этого, не хватает общедоступного набора жестового русского языка для исследований в области компьютерного зрения и машинного обучения, поэтому его необходимо собирать самостоятельно.

* 1. **Методологии распознавания языка жестов**

В связи с развитием компьютерных технологий многие исследователи разработали новые методы для помощи людям, имеющим ограниченные физические возможности. Необходимость систем машинного перевода обусловлена нехваткой человеческих ресурсов в области сурдоперевода и в ситуациях, когда посредничество в общении между глухими или слабослышащими и слышащими людьми не всегда желательно, например, в сфере медицины (конфиденциальность данных или врачебная тайна).

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 для распознавания жестов, а также нестабилен в некоторых средах. Второй метод основан на обработке изображений, где конвейер этой методологии выполняется следующим образом: захват изображения с помощью веб-камеры, сегментация, извлечение признаков и классификация жестов.

Сообщества с нарушениями слуха часто используют язык жестов, который представляет собой систему, использующую визуально-мануальную модальность для передачи смысла. Язык жестов зависит в основном от жестов рук, движений тела и выражения лица. Распознавание языка жестов (SLR) — сложная задача, особенно распознавание динамических знаков, зависящих от движения. Вот почему многие исследователи заинтересованы в разработке приложения SLR с целью уменьшения барьера между сообществом с нарушениями слуха и обществом.

Методологии распознавания языка жестов на основе компьютерного зрения принято делить на две категории: статические и динамические. Статические признаки — это те, которые требуют обработки только одного изображения на входе классификатора, то есть его можно рассматривать как картинку формы руки. Динамические знаки можно рассматривать как видео, содержащее ряд последовательных кадров для построения знака. Как правило, в языке жестов знаки строятся из серии быстрых движений рук и выражений тела. Следовательно, статическое распознавание не является хорошим решением проблем языка жестов, поскольку оно не может справиться с вариациями знаков. Значит, динамическое решение является более эффективным и действенным.

Есть некоторые проблемы, с которыми сталкивается SLR, которые можно классифицировать как основные и вторичные факторы.

1. Основные факторы:

* Форма руки: разница в форме руки меняет знак;
* Расположение руки: расположение руки относительно тела может изменить значение знака, даже если форма руки такая же;
* Движение руки: самый сложный параметр, так как знак может содержать набор движения с разными направлениями и формами.

2. Второстепенные факторы:

* Выражение лица: выражение лица играет жизненно важную роль в иллюстрации знака; повышает смысл и силу смысла в процессе общения;
* Ориентация ладони: направление ладони при выполнении знака — вверх или вниз, вправо или влево.

Многие исследователи использовали методы извлечения признаков вручную с алгоритмами машинного обучения для классификации жестов рук, но в последнее время в большинстве разработок использовались методы глубокого обучения. Изначально для распознавания жестов рук применялись свёрточные нейронные сети (CNN), но с помощью них было трудно распознать динамические жесты рук, содержащие пространственно-временную информацию. Некоторые исследователи использовали рекуррентные нейронные сети (RNN), которые в основном похожи на CNN, хотя свёрточные нейронные сети оказались более успешными. Недавно разработчики использовали долговременную кратковременную память (LSTM) для извлечения долговременной зависимости. Комбинация CNN и LSTM использовалась для достижения высокой точности распознавания жестов рук. Долгосрочная зависимость требует высокой сложности вычислений, что является основной проблемой LSTM. Нейронные сети, основанные на внимании, напротив, создают кратковременную зависимость, которая требует меньшей вычислительной сложности.

Таким образом, учитывая все особенности и сложности распознавания и классификации жестов рук на видео в реальном времени, было решено провести несколько экспериментов с разными моделями нейронных сетей для извлечения признаков и сравнить их результаты.

1. **Сбор корпусов данных**
   1. **Методология**

Задача сбора корпуса данных, содержащего видео и фотографии, является важной задачей для многих исследований в области компьютерного зрения и машинного обучения. Ниже будет представлена одна из методологий, которая использовалась при самостоятельном создании набора данных цифр и букв русского жестового языка.

Методология включает в себя следующие шаги:

* Определение цели и задач. Первым шагом при сборе корпуса данных является определение цели и задач исследования. Это поможет определить, какие типы видео и фотографий должны быть включены в корпус данных. Например, если исследование связано с распознаванием объектов на фотографиях, то необходимо включить фотографии, на которых присутствуют различные объекты.
* Определение источников данных. Источники могут быть различными, такими как интернет, социальные сети, базы данных и т.д. Необходимо определить, какие источники будут использоваться для сбора данных, и какие ограничения могут быть наложены на использование этих источников.
* Сбор данных. Третьим шагом является непосредственный сбор данных. Для сбора фотографий может быть использовано программное обеспечение для сканирования сайтов или социальных сетей, а для сбора видео может быть использовано программное обеспечение для загрузки видео с любых видеохостингов. Кроме того, можно найти добровольцев и с помощью них заснять видео или фото самостоятельно. При сборе данных необходимо учитывать правовые и этические аспекты, такие как защита личных данных и авторские права.
* Обработка данных. Это может включать в себя удаление дубликатов, удаление нежелательных изображений или видео, а также ручную проверку данных на наличие ошибок.
* Аннотация данных. Это может включать в себя маркировку изображений или видео, а также создание метаданных для каждого элемента данных.
* Разбиение данных на обучающую и тестовую выборку. Обучающая выборка используется для обучения модели, а тестовая выборка - для проверки точности модели.
* Оценка качества данных. Это может включать в себя проверку точности разметки данных, а также оценку покрытия данных для каждого класса объектов.
  1. **Корпус данных для русского жестового языка**

Поскольку не существует общедоступного и общепринятого набора данных для цифр и букв русского жестового языка, то он был создан самостоятельно с помощью добровольцев. При сборе корпуса данных использовались принципы методологии, описанной выше.

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 девушек и 9 мужчин в возрасте от 20 до 55 лет. Добровольцы на видео показывали цифры от 1 до 10 на правой и левой руке с помощью жестов. Инструкции о том, как показывать цифры жестового русского языка были взяты из проекта «Словарь. Русский жестовый язык» []. В итоге, было записано 38 видео и взято одно видео с цифрами от 1 до 10 из проекта «Словарь. Русский жестовый язык». Кроме того, дополнительно было снято 2 видео с правой и левой рукой для проверки на реальных видео качества распознавания и классификации жестов цифр.
2. Буквы. На втором этапе записывались на видео буквы русского жестового языка. Участвовало 11 добровольцев, из которых 7 девушек и 4 мужчин в возрасте от 20 до 55 лет. Добровольцы показывали 25 букв русского алфавита (а, б, в, г, е, ж, и, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ы, э, ю, я) с помощью жестов. Инструкции о том, как показывать буквы жестового русского языка были взяты из проекта «Словарь. Русский жестовый язык» []. В итоге, было записано 8 видео и взято одно видео из проекта «Словарь. Русский жестовый язык». Кроме того, дополнительно было снято видео для проверки на реальном видео качества распознавания и классификации жестов букв.
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.

В результате всех вышеуказанных действий, получилось 10 классов с жестами цифр от 1 до 10 по 420 цветных фотографий на каждый жест и 25 классов с жестами букв с N цветных фотографий на каждый жест (показ букв более динамичный, поэтому каждый участник показывал жест N количество секунд).

* 1. **Корпус данных для американского жестового языка**

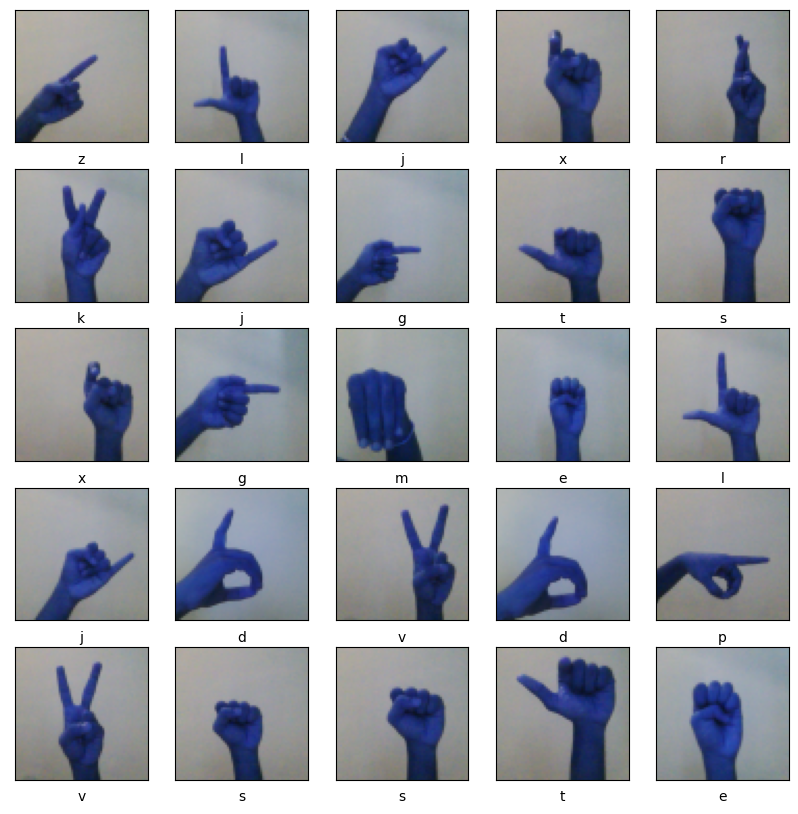
Для американского жестового языка (ASL) существует достаточное количество общедоступных наборов данных. Поэтому было выбрано два набора данных, один для классификации цифр, другой для – букв, которые соответствовали целям и задачам исследования. Оба корпуса данных были найдены на «Kaggle» – система организации конкурсов по исследованию данных, а также социальная сеть для специалистов по обработке данных и машинному обучению [].

Первый набор данных [] – это набор цифр от 0 до 9 американского языка жестов. Содержит 10 классов по 500 цветных изображений рук на каждый жест с тёмным фоном размером 400 × 400 (пример на *Рис. 1*).

Второй набор данных [] – это набор букв от a до z американского алфавита. Из корпуса данных было выбрано 3000 цветных фотографий жестов размером 400 × 400 для каждой буквы (пример на *Рис. 2*).



**Рис. 1** - Пример набора данных цифр ASL



**Рис. 2** - Пример набора данных букв ASL

К тому же, поскольку найденные корпуса данных не содержали видео, было дополнительно снято два видео, на которых можно проверить качество классификации цифр и букв на реальном видео.

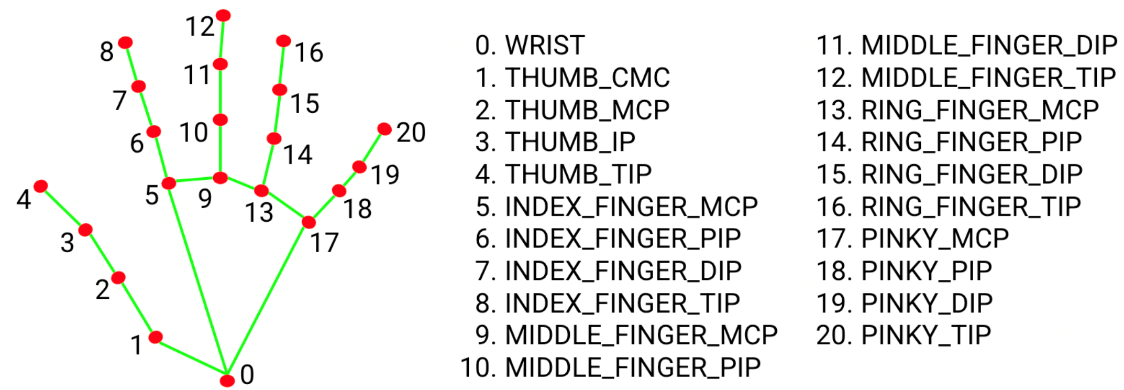
1. **Обработка корпусов данных**
   1. **MediaPipe**

Для обработки собранных наборов данных использовалось решение «MediaPipe Hand Landmarker». Оно является частью проекта «MediaPipe» с открытым исходным кодом, код решения которого можно дополнительно настроить в соответствии с потребностями разработчика. «MediaPipe», в свою очередь, – 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. Это решение работает с данными изображения с помощью модели машинного обучения в виде статических данных или непрерывного потока и выводит ориентиры рук в координатах изображения, ориентиры рук в мировых координатах и принадлежность (левая/правая рука) нескольких обнаруженных рук [].

«Hand Landmarker» использует пакет моделей с двумя упакованными моделями: моделью обнаружения ладони и моделью обнаружения ориентиров руки. Модель обнаружения ладони определяет местоположение рук на входном изображении, а модель обнаружения ориентиров для рук определяет конкретные ориентиры для рук на обрезанном изображении руки, определенном моделью обнаружения ладони [].

Пучок моделей ориентиров руки определяет локализацию ключевых точек 21 координаты костяшек кисти в пределах обнаруженных областей руки. Модель была обучена примерно на 30 тыс. реальных изображений, а также на нескольких синтетических моделях рук, наложенных на различные фоны []. Определение 21 ориентира приведено на Рис. 3.



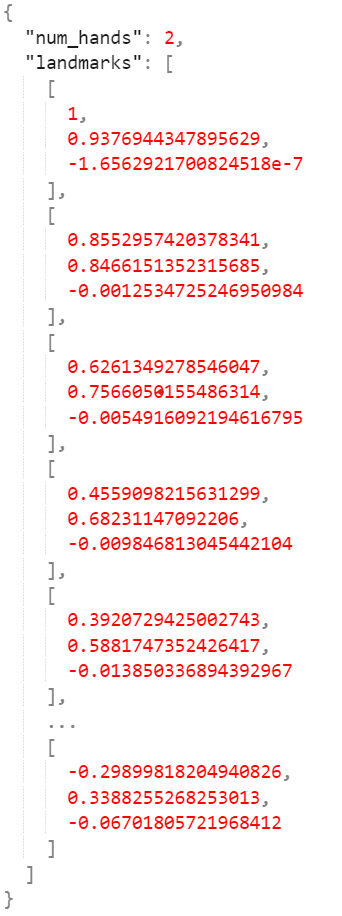
**Рис. 3** - 21 ключевая точка руки

В данной работе решение «Hand Landmarker» использовалась как для обнаружения рук на видео, так и для получения ключевых точек обнаруженных рук. Ключевые точки рук будут использоваться для классификации жестов цифр и букв русского и американского жестового языка.

* 1. **Корпуса данных для русского жестового языка**

Оба корпуса данных с цифрами и буквами русского жестового языка (RSL) были переработаны решением «MediaPipe Hand Landmarker». 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 длинной 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 точки также были нормализованы от 0 до 1 относительно координат изображения.



**Рис. 4** - Пример экземпляра набора данных с точками

Второй набор содержит фотографии обнаруженных обрезанных рук для каждого жеста размером 64 × 64.



**Рис. 5** - Пример набора данных цифр RSL



**Рис. 6** - Пример набора данных букв RSL

В итоге, 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 размером 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 (если жест показывался меньше чем за 50 кадров, то кадры интерполировались), то есть массивы размером 50 × 43 ориентиров рук. И 25 классов по 8 папок для каждого человека с 50 цветных фотографий для каждого жеста только обнаруженных рук размером 64 × 64.

* 1. **Корпуса данных для американского жестового языка**

Оба корпуса данных с цифрами и буквами американского жестового языка также были переработаны решением «MediaPipe Hand Landmarker». Поскольку оба набора и так содержали фотографии только рук, то в данном случае, решение MediaPipe использовалось для извлечения ключевых точек. Таким образом, появилось два дополнительных корпуса данных. Для цифр - 10 классов по 430 документов в формате ‘.json, для букв - 26 классов по 3000 документов в формате ‘.json’ для каждого жеста соответственно. Все массивы данных также были нормализованы от 0 до 1 относительно координат изображения.

1. **Алгоритм распознавания языка жестов**
   1. **Модели нейронных сетей**

Нейронные сети - это математические модели, состоящие из множества связанных между собой нейронов, которые обрабатывают информацию и передают ее дальше по сети. Они, в основном, используются для анализа и обработки больших объёмов данных, что позволяет получать более точные прогнозы и принимать более обоснованные решения [].

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

Многослойный персептрон (MLP) – это тип нейронной сети, которая состоит из одного входного слоя, одного или нескольких скрытых слоев и одного выходного слоя. Сигнал проходит от входного слоя через скрытые слои к выходному слою, при этом связь допускается только между соседними слоями. Модель используется для решения различных задач, таких как прогнозирование, классификация и распознавание образов [].

Одномерная свёрточная нейронная сеть (1D CNN) – это тип нейронной сети, которая обычно используется для задач обработки сигналов, таких как классификация изображений или аудио. Она предназначена для извлечения признаков из одномерных последовательностей данных, таких как временные ряды или сигналы [].

Нейронная сеть долгой краткосрочной памяти (LSTM) – это тип рекуррентной нейронной сети (RNN), которая особенно полезна для обработки и прогнозирования последовательных данных, таких как временные ряды или текст на естественном языке. Во время обучения используется последовательная информация, которая проходит через нейронную сеть от входного вектора к выходным нейронам, а ошибка вычисляется и распространяется обратно через сеть для обновления параметров сети. Информация в этих сетях включает петли в скрытый слой, позволяя информации течь разнонаправленно, так что скрытое состояние означает прошлую информацию, имеющуюся на данном временном шаге [].

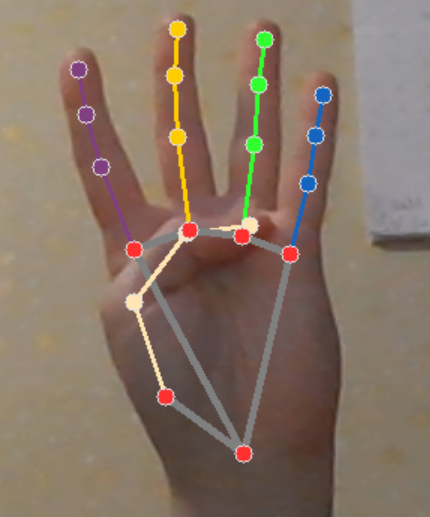
Двумерная свёрточная нейронная сеть (2D CNN) - это тип архитектуры глубокого обучения, который наиболее широко используется для обработки изображений и временных рядов. Она подходит для анализа и обработки двумерных данных, поскольку признаки извлекаются из входных данных посредством свёртки. Модель состоит из повторяющегося свёрточного слоя, объединяющего слоя и полносвязанного слоя [].

Многослойная полносвязная сеть была выбрана, поскольку является классической сетью, которая годится для многих задач классификации. Последние три модели – исходя из применения их исследователями в похожих областях.

* 1. **Алгоритм**

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 визуализирует их (пример на Рис. 7).



**Рис. 7** - Пример работы "MediaPipe Hand Landmarker"

Алгоритм распознавания языка жестов включает в себя следующие шаги:

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);
2. Захват живого потока с веб-камеры;
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 прогнать его через 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 видеокадре;

1. 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 сеть пытается предсказать какой жест изображён на кадре;
* Если выбрана одна из моделей, обученная на ключевых точках рук (MLP/1D CNN/LSTM). Превратить кадр в 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 точки 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;

1. На кадре визуализируются ориентиры обнаруженных рук и символами отображается предсказанный нейронной сетью жест (пример на Рис. 8);
2. Шаги 3-5 повторяются до тех пор, пока пользователь не выйдет из режима камеры.

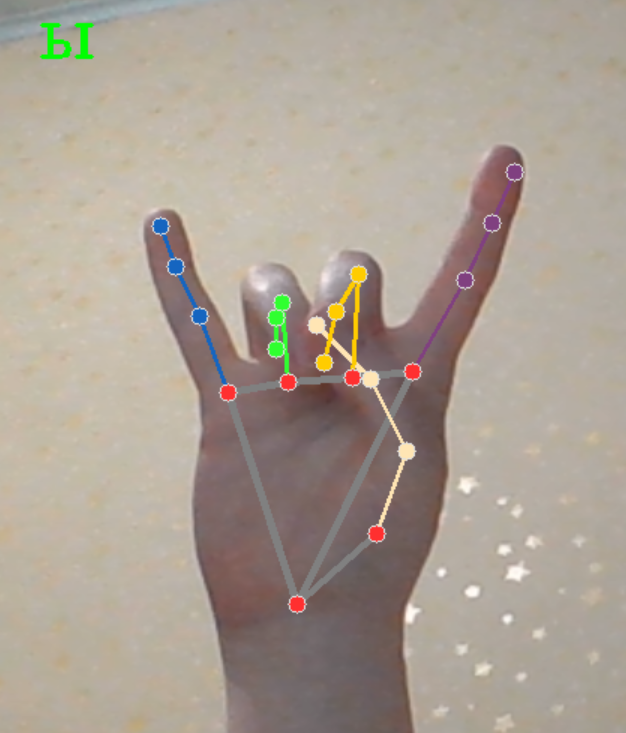


Рис. 8 - Пример распознавания руки и классификации жеста