

## Исследование методов классификации движений человека

*Токарев Андрей Сергеевич*

После достижения результатов в задаче сегментации человека и отслеживании его на видео стало интересна возможность оценки занимаемой им позы и ее классификация. Для этого было необходимо научиться распознавать выбранные ключевые точки на теле человека, которые несут информацию о его положении. Некоторые группы исследователей уже добились хороших результатов в данном направлении и двигаются дальше в задачах анализа взаимодействия между людьми и предсказания следующих движений человека.

В данной работе представлен обзор задачи классификации движений на основе информации о ключевых точках на теле человека, а также приведены технический обзор и исследование систем глубокого обучения для оценки позы на основе публичных наборов данных.

# Содержание

<b>1 Введение</b>	<b>4</b>
<b>2 Постановка задачи</b>	<b>9</b>
2.1 Задача распознавания ключевых точек на теле человека . . . . .	9
2.2 Задача классификации движений человека . . . . .	11
<b>3 Обзор существующих моделей</b>	<b>13</b>
3.1 Модели для распознавания ключевых точек на теле человека	13
3.1.1 DeepPose . . . . .	13
3.1.2 AlphaPose . . . . .	14
3.1.3 BlazePose . . . . .	15
3.1.4 MoveNet.SinglePose . . . . .	17
3.1.5 OpenPose . . . . .	19
3.1.6 MMPose . . . . .	20
3.2 Модели для классификации позы человека . . . . .	22
3.2.1 Классификатор от MediaPipe . . . . .	22
3.2.2 mmakos . . . . .	23
3.2.3 MMaction2 by OpenMMLab . . . . .	24
<b>4 Исследование моделей</b>	<b>26</b>
4.1 Описание эксперимента . . . . .	26
4.2 Поиск данных . . . . .	30
4.3 Результаты эксперимента . . . . .	33
<b>5 Заключение</b>	<b>38</b>
<b>Список литературы</b>	<b>39</b>
<b>А Приложение</b>	<b>44</b>

# 1 Введение

Понимание движений человека является необходимой частью нашей жизни. При общении людьми часто используется жестикуляция, так как это помогает выражать чувства, эмоции и доносить свои мысли до окружающих . Из анализа позы человека можно сделать вывод о его состоянии. К примеру, хромота или нахождение в неестественном положении говорят о необходимости не только медицинской, но, возможно, и вашей помощи. Ещё можно обратиться к психоанализу, а точнее к разделу о языке телодвижений. В нем по позе можно сделать вывод о характере человека или о текущем состоянии, его заинтересованности в беседе. Также работает распознавание движений. Если мы видим бегущих в панике людей, то наш мозг получает сигнал об опасности и спасает нас. Из приведенных ситуаций становится понятно, почему определение позы и классификация движений являются важными аспектами нашей жизни. В связи с развитием информационных технологий, человечество задумалось над выполнением данной задачи с помощью компьютера. Тогда можно будет добавить дополнительный источник информации для взаимодействия искусственного интеллекта с человеком.

При рассмотрении данной задачи через призму машинного обучения, получим, что нам нужно классифицировать положение человека, данные о котором необходимо каким-то образом получать. Первый способ - надеть на добровольца датчики и, считывая координаты каждого из них, построить на компьютере его позу и, таким образом, восстановить скелет для последующего анализа. Второй способ - искать особые точки на фотографии с помощью компьютерного зрения. Установим камеру и начнем анализировать положение и скелет человека, исходя из картинки. Тогда не придется закупать большое количество датчиков для снятия данных, а нужна будет только камера и вычислительные мощно-

сти для работы алгоритмов глубокого обучения. Несмотря на сложность реализации первого варианта, его удобно использовать для подготовки тренировочных датасетов [1].

Движение - это растянутый во времени процесс. Он анализируется по видеозаписям, каждая из которых представляет собой последовательность кадров. Поэтому первостепенно научиться работать с изображением. Как же собирать данные для модели классификации?

Восстановление скелета (Skeletal Representation), детекция (Pose Detection) и оценка позы (Pose Estimation), распознавание движения (Action Recognition) являются расширением одной задачи: распознавание ключевых точек на теле человека (Key-points Detection). Она имеет прикладной смысл не только в связке с классификацией движений. В работе будет рассматриваться распознавание на картинке, то есть в 2-х мерном пространстве. Но ведь можно восстанавливать положение человека (скелет человека) в 3-х мерном пространстве [2, 3]. Используя генеративные нейронные сети возможно воссоздавать не только скелет человека, но и тело человека [4]. Объединяя две предыдущие задачи можно получить набор данных из 3-х мерных людей в различных позах. Некоторые исследователи уже пробуют реализовать этот симбиоз на практике [5].

Если перейти в тематику биологических и медицинских наук, то есть возможность развить данную тему на примере восстановления структуры тканей человека. Получается, можно по фотографии моделировать распределение мышечных, жировых и других тканей в теле человека. Это поможет более детально изучать проблемы каждого человека персонально и подбирать индивидуальные курсы лечения или диеты.

Восстановление скелета человека упростит спасателям анализ положения человека под завалами и облегчит им построение планов по его спасению. Однако в данном случае необходимо быстродействие алгорит-

ма и очень важно получить изображение человека.

В современном мире, где повсюду слышны разговоры о технологиях дополненной реальности и мета вселенной, найдем ещё одно применение для алгоритмов детектирования позы. Для нахождения в виртуальной вселенной необходимо транслировать человека туда, а значит можно с помощью видеокамер определять положение, восстанавливать скелет и получать итоговое изображение или 3-х мерную модель. Чем-то напоминает фильм "Первому Игроку Приготовиться" Стивена Спилберга. Добавим алгоритм генерации аватара вместо реального человека и получим рабочий алгоритм трансляции живого человека в мета вселенную.

Второй частью работы является задача классификации (Pose Classification), которая использует данные, полученные в первой части. Таким образом, построен алгоритм анализа движений человека на статическом изображении. По нему трудно давать оценку поведению человека, но своеобразный "помощник" из полученного алгоритма будет хороший. Рассмотрим некоторые идеи применения.

Начать можно с медицины. Восстановление больных после операций, травм и несчастных случаев - это длительный и трудоемкий процесс, требующий постоянного присмотра врача. Если человек учится двигаться, то нужен тренер, который укажет на ошибки и исправит вас. Решение нашей задачи помогает таким пациентам. Анализ движений может сравнивать человека с эталоном и указывать на ошибки. Также при наблюдении за больным алгоритм может идентифицировать отклонения от нормального поведения и вызвать врача (к примеру увидеть приступы эпилепсии у человека). Это может не только спасти множество жизней по всему миру, просто вызывая врача в необходимый момент, но и помочь в восстановлении.

Также можно выявлять у человека заболевания или дефекты скелета

та. Тогда появляется возможность выявлять сколиоз или сутулость и подсказывать людям, что надо стараться держать спину прямо. Хотя лучше направлять к врачу на консультацию и лечить дефекты позвоночника сразу. Проведя исследование населения, получится собрать статистику тех или иных отклонений. Так уже сделали производители кроссовок и с помощью gait-анализа [6] помогают выбрать подходящую обувь.

Посмотрим теперь на спорт. Из классификатора можно сделать хорошего судью соревнований в тех видах, где надо различать, отслеживать положения тела. К примеру, GOOGLE придумали использовать классификатор как счетчик подтягиваний, приседаний или отжиманий [7] и это можно поместить в современный смартфон. Если углубиться дальше, то решение можно обернуть интерфейсом и создать хорошего робота-фитнес-тренера. Ведь настроив камеру смартфона на наблюдение за вами во время тренировки, приложение будет подсказывать правильную позу для упражнения и укажет на ошибки, если таковые имеются.

При развитии моделей в будущем, можно будет найти другие варианты применения технологии классификации движений человека. Можно анализировать поведение группы людей, но для этого надо хорошо восстанавливать скелет нескольких человек на одном изображении [8, 9, 10]. Также есть возможность предсказывать будущие действия человека при изучении уже имеющихся [11]. Если опять затронуть идею генеративных нейронных сетей, то можно генерировать движения человека по заданному начальному условию. Следовательно, можно создавать искусственные видеозаписи или добавлять неигровых персонажей (пр - non-player character) в виртуальную реальность. В медицине можно моделировать восстановление двигательной активности человека или моделирование протезов индивидуально под каждого пациента.

Применений для решений задачи оценки движения множество и в

мире уже существует большое количество моделей анализа движения и восстановления скелета. В данной работе будут рассмотрены некоторые из них, а также будет приведен качественный и количественный анализ работы этих моделей на основе нескольких метрик.

## 2 Постановка задачи

Как уже было сказано в разд. 1, в работе будет произведена оценка систем классификации движений человека на изображении. Из рисунка надо получить данные о принимаемой субъектом позе и классифицировать её на род деятельности. Получается решается две задачи: предобработка данных, то есть извлечение расположения ключевых точек на теле человека, и их последующая категоризация. Рассмотрим их по отдельности.

### 2.1 Задача распознавания ключевых точек на теле человека

Первоначально необходимо понять каким образом можно распознать позу человека, чтобы в дальнейшем взять оттуда информацию для классификации. Человек смотрит на другого человека и анализирует его позицию исходя из данных о его расположении частей тела анализируемого. Получается нам необходимо найти части тела человека, каждая из которых ограничена какими-либо суставами. Последние можно и взять за ключевые точки, которые будут распознаваться моделью. Если соединить выходные данные, то получим своеобразный "скелет" человека.

Необходимо определиться сколько точек на теле человека необходимо различать. На текущий момент стандартом является топология CO-CO (см. рис. 1а), которая включает в себя 17 ориентиров на теле человека [12, 13]. Данная топология не учитывает расположение ступней и кистей рук, а также рассматривает всего 5 точек на лице человека: нос, два глаза и два уха. Но стандартом многие исследователи не ограничиваются и добавляют дополнительные точки. Приведу два примера:

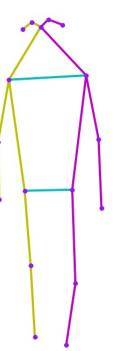
1. Топология от BlazePose (см. рис. 1б)

Включает в себя 33 точки расположенные на теле человека. Данная

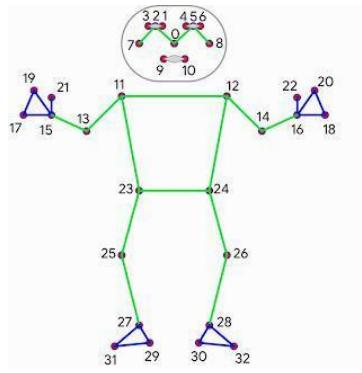
топология представляет собой объединение COCO, BlazeFace [14] и BlazePalm [15]. В итоге, извлекается дополнительная информацию о направлении стоп и кистей, а также больше понимаем насчет точек на лице. Данная модель расположения точек используется в одноименной модели (BlazePose [16]) и ориентирована на использование в фитнес приложениях. Также у данной компании есть более развитая модель, которая определяет положение всех пальцев кисти и распознает мимику на лице [17].

## 2. Halpe (см. рис. 1c)

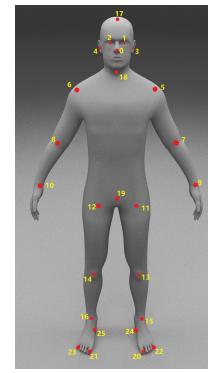
Данная топология - это совместный проект AlphaPose [10] и НАКЕ [18]. Представлено две модели: на 26 и на 136 точек. Здесь добавлено рассмотрение ориентации стоп, распознавание шеи, паха и макушки головы. В расширенной модели присутствует ещё 68 точек на лице, а также по 21 на ладонях.



(a) Топология COCO



(b) Топология BlazePose



(c) Топология Halpe

Рис. 1: Примеры расположения точек на теле человека.

В итоге получилось разобраться с тем, что надо искать на изображении. Теперь надо выяснить каким образом искать ключевые точки на фотографии. На сегодняшний день есть два подхода:

- Bottom-up

Изначально строится карта предсказания для каждой ключевой точ-

ки из топологии и потом эти точки определенным образом собираются в скелет человека. Важно отметить, что при распознавании нескольких человек необходимо правильно сопоставить точки каждому человеку.

Bottom-up удобно использовать при покадровой обработке видеофрагментов.

- Top-down

В данном подходе требуется локализация человека, получение прямоугольника с ним и перенаправление этого прямоугольника в модель распознавания ключевых точек. Модели, использующие данный подход выдают точные результаты, но чувствительны к ошибкам детекции и к ориентации человека в кадре. Поэтому им требуется предобработка прямоугольника с человеком.

В обоих подходах результатом необходимо получить набор координат точек по заранее заданной топологии для дальнейшего преобразования в скелет человека.

## 2.2 Задача классификации движений человека

Задача классификации подразумевает обучения алгоритма на размеченных данных предсказывать класс объекта исследования. К примеру можно предсказывать класс объекта на фото, тему текста, жанр музыки или позу человека.

Получается имеется множество объектов  $X$ , множество классов  $Y$  и некоторая зависимость  $cls : X \rightarrow Y$ , то есть  $\forall x \in X \exists y \in Y : y = cls(x)$ .

Задача классификатора состоит в том, чтобы построить алгоритм  $a_{cls}$ , который мог бы предсказывать метку из  $y \in Y$  для любого объекта  $x \in X$ .

Тогда обучающая выборка будет представлять собой пространство размерности 2, где каждому вектору будет ставиться в соответствие другой вектор:

$$TrainData \triangleq X^2 : (x_i, y_i), \quad x_i \in X, y_i \in Y$$

В текущей задаче используется предобработка данных, получается, что классификация происходит не над объектами из  $X$ , которые являются изображениями, а над пространством признаков, извлеченных из изображения. То есть имеется функция, которая строит биекцию между объектами исходного пространства  $X$  и пространства признаков  $D_f$ :

$$\vec{g}(x) : X \rightarrow D_{f_1} \times D_{f_2} \times \dots \times D_{f_m}$$

$$\forall x \in X \exists \vec{f} \in D_f^m : \vec{f} = \vec{g}(x) = (f_1(x), f_2(x), \dots, f_m(x))$$

Функции  $f_1, f_2, \dots, f_m$  называются признаками.

Тогда получается предсказание будет строится не над объектами из  $X$ , а над векторами признаков из  $D_f^m$  и задача ставится научиться классифицировать метки классов из  $Y$  для  $\vec{f}$  из  $D_f^m$ :

$$\forall \vec{f} \in D_f^m \exists y \in Y : y = cls(\vec{f})$$

Если суммировать все вышесказанное, то задача классификации движений человека по данным оценки его позы требует построить алгоритмы  $a_g$ , способный переводить изображение в пространство признаков, и  $a_{cls}$ , способный предсказывать метку класса по полученному вектору:

$$\left[ \forall x \in X \exists \vec{f} \in D_f^m : \vec{f} = \vec{g}(x) = (f_1(x), f_2(x), \dots, f_m(x)) \right]$$

&

$$[\exists y \in Y : y = cls(\vec{g}(x))]$$

### 3 Обзор существующих моделей

#### 3.1 Модели для распознавания ключевых точек на теле человека

В данном разделе будет рассмотрено 6 различных моделей. В разд. 4 будут выбраны 4 наиболее удобные в использовании и в обучении и будет проведен эксперимент по оценке данных моделей.

Так же хочется сказать, что, помимо приведенных, есть множество моделей от одиночных авторов, не объединенных в лаборатории [19, 20]. Они в основном брали какую-то из представленных ниже моделей и проводили небольшое улучшение.

А теперь перейдем к моделям.

##### 3.1.1 DeepPose

DeepPose является одним из самых первых решений задачи распознавания ключевых точек на теле человека с помощью глубокого обучения. Статья "DeepPose: Human Pose Estimation via Deep Neural Networks" [21] была представлена исследователями из GOOGLE на конференции CVPR в 2014 году.

В своей работе они представили каскад из DNN-регрессоров для локализации суставов тела. На тот момент принятых топологий ещё не было и поза кодировалась координатами суставов, нормализованными на размер изображения. Первым этапом применялась CNN для локализации точки, а вторым каскадом применялись DNN для уточнения результата (см. рис. 2). Таким образом получалось довольно точно распознавать ключевые точки на фотографии.

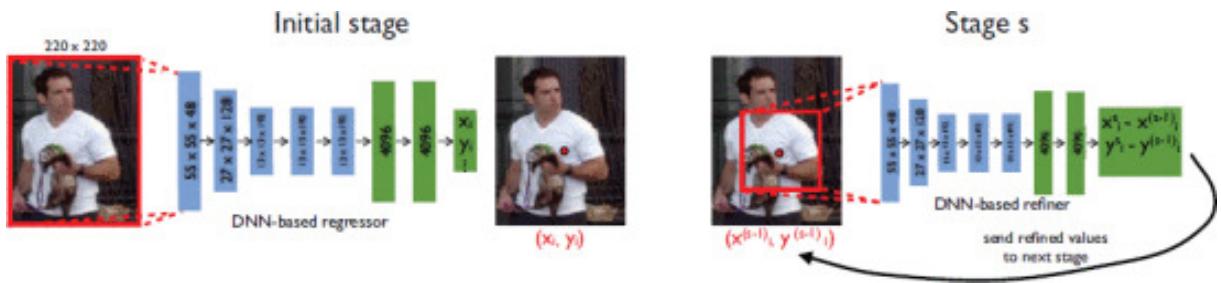


Рис. 2: Архитектура сети DeepPose. [21]

### 3.1.2 AlphaPose

AlphaPose является основной наработкой проекта Machine Vision and Intelligence Group из Шанхайского университета транспорта (Shanghai Jiao Tong University или SJTU). Это первая модель, которая получила значение метрики mAP на датасете COCO выше 70 (0.7) и выше 80 на MPII. Поддерживается как на Linux, так и на Windows. Обрабатывает видео и поддерживает слежение за человеком в реальном времени через восстановление его скелета.

Исследователи не против объединяться с другими проектами для улучшения качества модели и проработки новый фичей. Одной из таких коопераций стала топология Halpe (см. разд. 2.1), на основании которой и производится оценка позы человека. Доступными являются варианты также с 17 точками от COCO и 25 и 135 точек от Halpe.

В работе Alpha использует top-down подход. Первым этапом идет Faster R-CNN для детектирования человека и выдачи прямоугольников. После используется модель RMPE для предсказания различных поз, которые может принимать человек. Последним этапом идет работа p-Pose NMS для устранению избыточных предсказаний. На выходе получается изображение с восстановленными позами людей. Все шаги представлены на рис. 3. [10]

Рассмотрим поближе модель предсказания. Она использует симметричное преобразование spatial transformer network (STN) и обратное ему

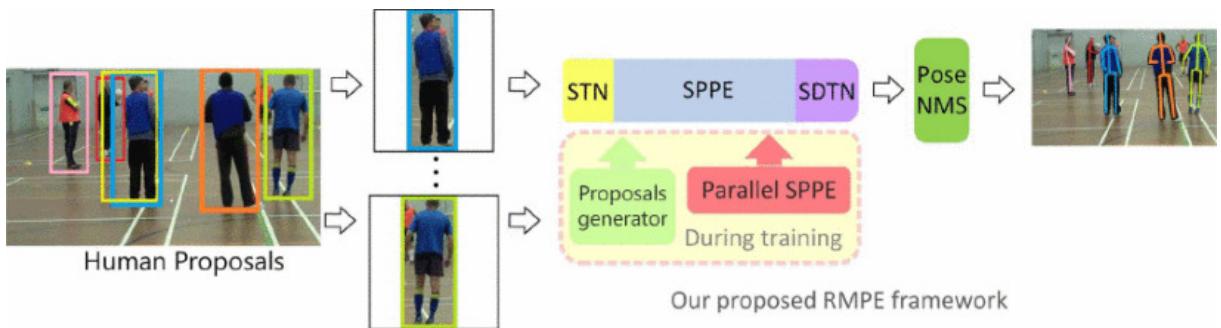


Рис. 3: Пример работы сети AlphaPose. [10]

spatial de-transformer network (SDTN). Они введены для исправления ошибок локализации, так как SPPE, представленная между ними (см. рис. 4), является очень чувствительной к неточностям прямоугольника. При обучении параллельно основной модели добавляется ещё одна SPPE (см. рис. 4) для корректировки преобразования STN. Таким образом получится приблизить данные, получаемые предсказателем, к идеальным и получить наиболее точный результат.

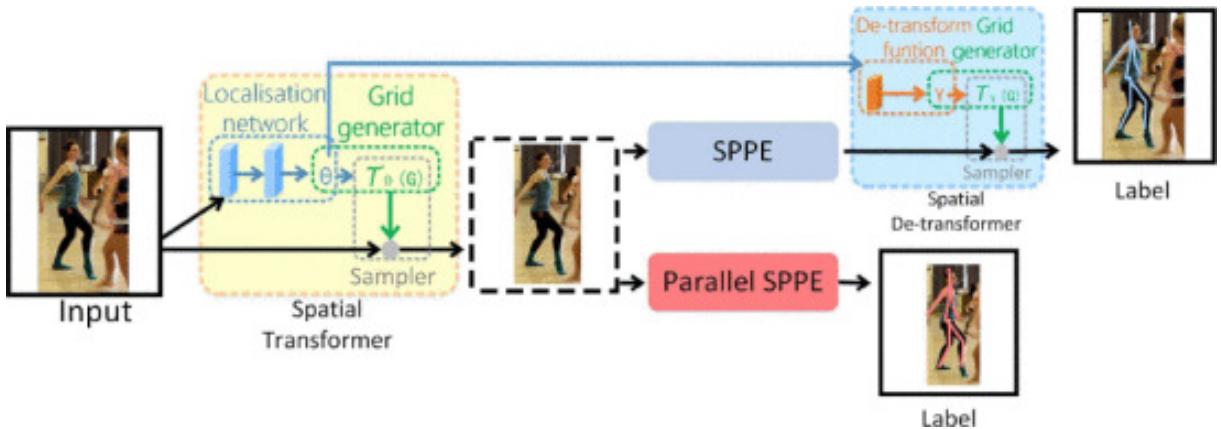


Рис. 4: Архитектура сети RMPE. [10]

### 3.1.3 BlazePose

MediaPipe является одним из проектов компании GOOGLE и в своей работе решает задачи компьютерного зрения. В нем уже были представлены модели для распознавания лица (Face Detection) и его поверхности (Face Mesh), ладоней (Hands), объектов (Object Detection и Objectron)

и другие [22]. Для нас же интересна задача поиска ключевых точек, которую и решает модель BlazePose [16]. На момент исследования модель умеет отслеживать движения человека на видеофрагменте и строить покадровую маску человека.

Для предложенной модели была создана топология, которая представляет собой суперпозицию топологии COCO и двух других топологий, уже использовавшихся в других подпроектах MediaPipe. Об этом более подробно написано в разд. 2.1.

В BlazePose используется top-down подход оценки позы человека. Сначала запускается Pose Detector (см рис. 5), который возвращает координаты интересующей нас области (region-of-interest или ROI). Алгоритм использует расширение модели BlazeFace для определения наличия человека в кадре. Поэтому данная модель чувствительна к видимости головы, лица в частности, на фотографии. Взяв идею витрувианского человека Леонардо Да Винчи, исследователям понадобилось ещё две точки для точной локализации человека на изображении.

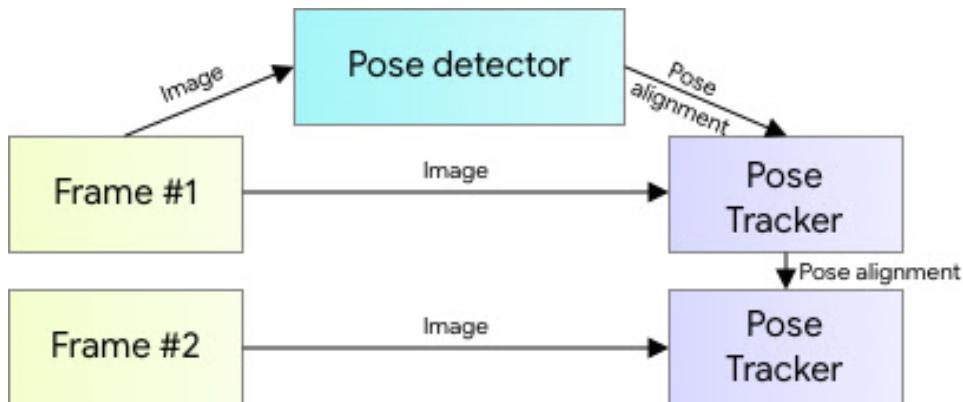


Рис. 5: Структура модели BlazePose для работы в реальном времени.

[16]

Следующим шагом Pose Tracker производит локализацию каждой точки в заданной ROI. Данное действие производится путем комбинированной обработки тепловой карты и данных о смещении с использованием регрессионной модели (см рис. 6).

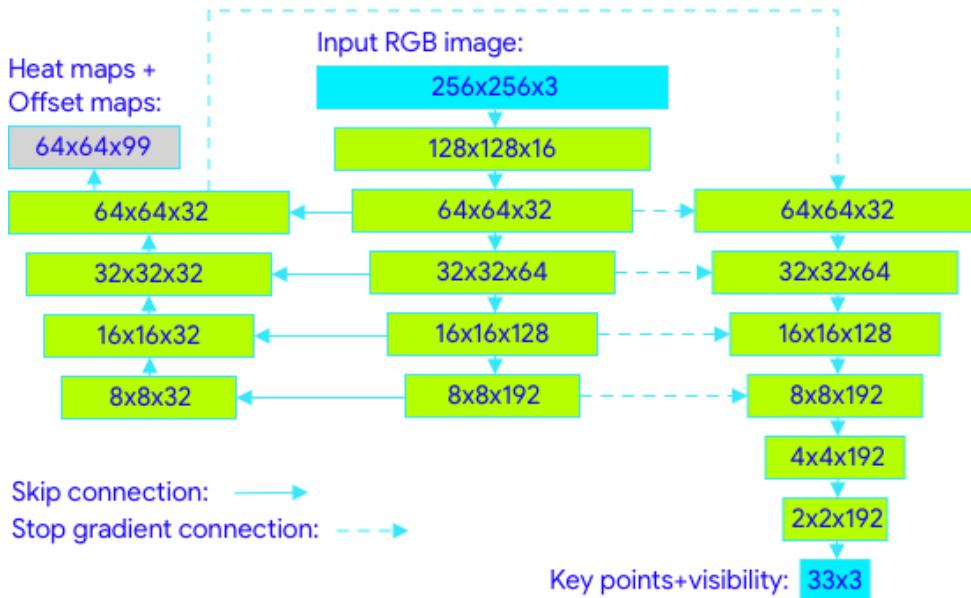


Рис. 6: Архитектура модели Pose Tracker. [16]

Как можно заметить из рис. 5, при аналитике видеофрагмента Pose Detector используется только на первом кадре, ведь позже данные об интересующей нас области передаются от кадра к кадру. Это упрощает вычисления и позволяет ускорить работу модели в реальном времени.

Развитием данной модели есть ее полное объединение с моделями BlazeFace и BlazeHand в модель Holistic [17]. Она рассматривает намного большее количество точек на лице и ладонях.

### 3.1.4 MoveNet.SinglePose

SinglePose создана для работы в веб-интерфейсах или на мобильных устройствах в режиме реального времени. [23] Модель представлена в двух спецификациях: lightning и thunder. Первая является менее требовательной в плане мощностей и вычислений и способна обрабатывать до 50 кадров в секунду. В то же время, по заверениям создателей, вторая модель имеет большие запросы по ресурсам, но дает лучшую точность распознавания, правда со скоростью до 30 кадров в секунду.

За расположение ключевых точек выбрана классическая топология COCO. Поэтому модель возвращает координаты 17 точек, которые нор-

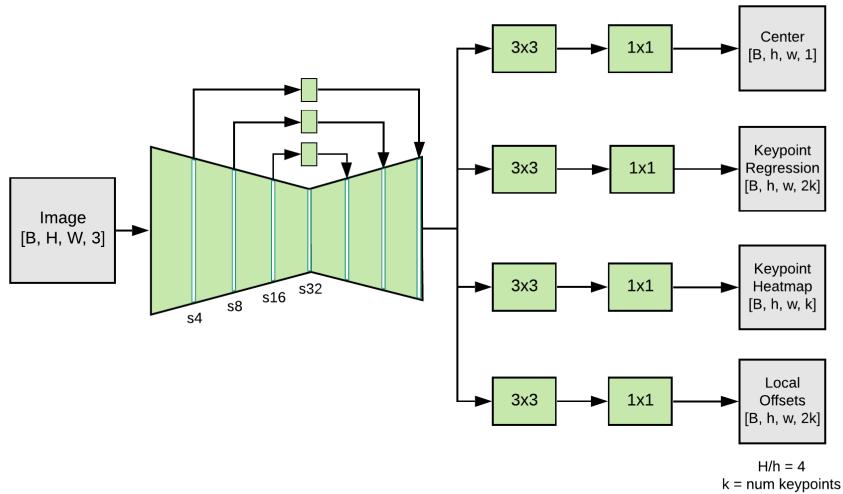


Рис. 7: Архитектура модели MoveNet. [23]

мированы на размер изображения (лежат в отрезке  $[0, 1]$ ).

Представленная модель является восходящей, то есть использует bottom-up подход к решению задачи. Она реализована на архитектуре MobileNetV2 [24] с Feature Pyramid Networks [25], которая используется для извлечения признаков. Обработка результатов магистральной части сети происходит с помощью прогнозирующих головок (см. рис. 7), которые используют CenterNet [26] с изменениями, которые повышают быстродействие модели.

Как можно заметить на рис. 7, 4 прогнозирующие головки отвечают за прогноз карты центральной точки человека, поле регрессионных векторов ключевых точек, карту предсказанных положений ключевых точек и поле смещения ключевых точек. Результаты каждого предсказаниярабатываются параллельно и далее путем последовательной обработки (см. рис. 8) уточняются координаты каждой ключевой точки.

Такая система позволяет получать точные результаты оценки позы в реальном времени и настроена на обработку в реальном времени изображений, поступающих с камер устройств пользователя.

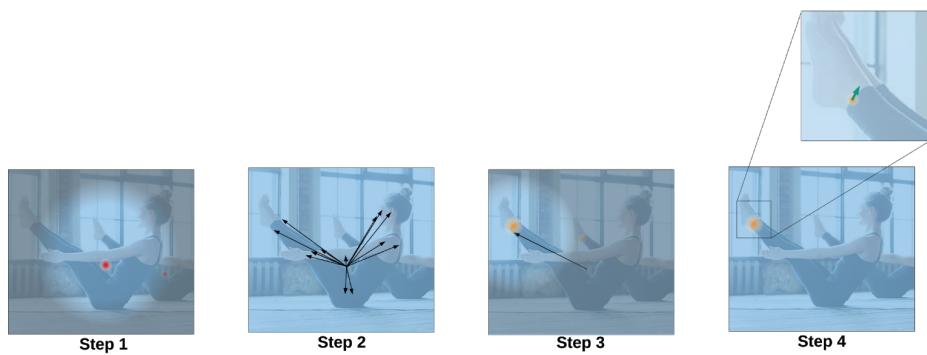


Рис. 8: Шаги работы prediction head в MoveNet. [23]

Следующим этапом развития проекта стала модель MoveNet.MutiPose для распознавания сразу нескольких людей на изображении. Разработчики не стали изменять традициям и представили ее также в двух вариантах: lightning и thunder. А также она использует ту же MobileNetV2 в основе своей работы.

### 3.1.5 OpenPose

OpenPose (OP) - это подпроект CMU-Perceptual-Computing-Lab из университета Карнеги-Меллона. В нем представлены различные модели: от локализации точного положения кистей и лица, до определения позы, исследуя 135 точек на теле человека. Также OP может распознавать нескольких человек одновременно и поддерживает отслеживание скелета человека на видеозаписи в реальном времени через веб-камеру.

В текущей работе рассмотрим модель,工作的 по топологии из 25 точек - чем-то похожую на топологию Halpe (см. разд. 2.1). Особенностью является определение положения стоп за счет детекции 3 дополнительных точек на каждой из них.

OpenPose тоже использует сверточные нейронные сети для решения задачи распознавания ключевых точек с помощью bottom-up подхода. В своей работе исследователи используют понятия карты достоверности обнаружения точки, карты двумерных векторных полей ориентации

конечностей (Part Affinity Fields или PAFs) и графы соответствия обнаруженных точек определенным людям на изображении.

Первым шагом происходит обработка изображения и построение карт достоверности и PAFs, про которое поговорим позже. Вторым шагом происходит сопоставление точек и конечностей отдельным людям с использованием графов соответствия. Они помогают решить данную задачу и быстрее решать задачу построения скелетов нескольких человек. В итоге на выходе имеются скелеты нескольких человек. Все шаги представлены на рис. 9.

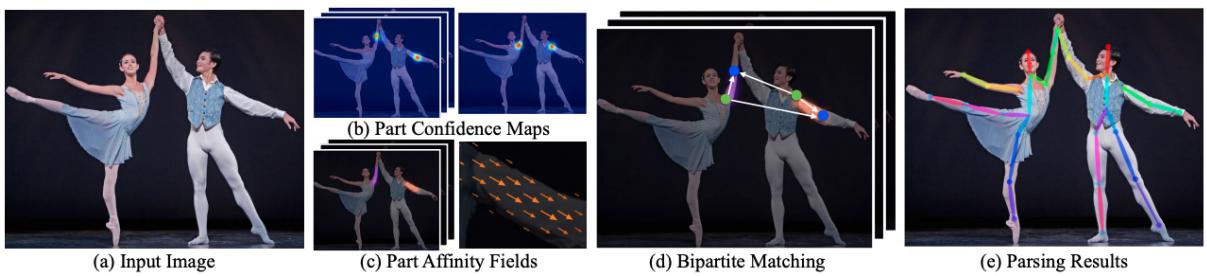


Рис. 9: Последовательность распознавания ключевых точек моделью OpenPose. [8]

В первоначальной версии модели [27] вычисление карт достоверности и PAFs происходило параллельно, в два этапа. Но в последней работе [8] предсказания поставили последовательно, так как из PAFs интуитивно можно предсказать и уточнить карты достоверности обнаружения точки (см. рис. 10). Ещё были заменены ядра свертки размером 7x7 на блоки из 3-х последовательных ядер 3x3 (см. рис. 10). Эти преобразования помогло почти в 200 раз увеличить скорость распознавания и в 7 раз улучшить точность.

### 3.1.6 MMPose

MMPose - является подпроектом лаборатории Open-MMLab [28]. Первым проектом лаборатории было решение задачи детектирования объектов,

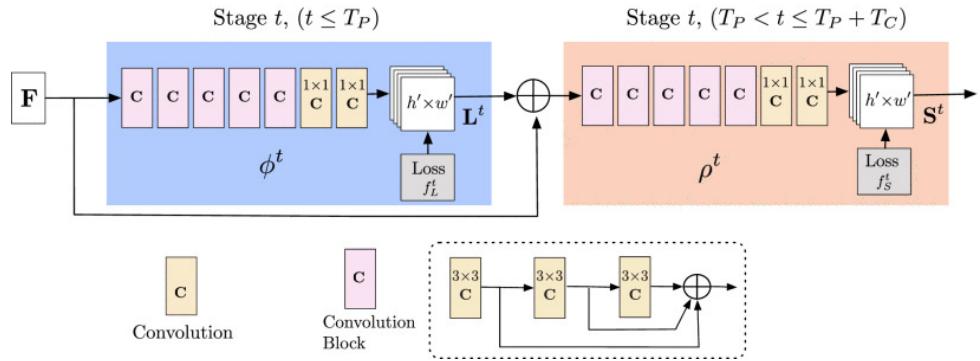


Рис. 10: Архитектура модели OpenPose 2D Pose Estimation. [8]

но позже развивались и другие связанные с компьютерным зрением. Изначально было исследование классификации движений с помощью сетей ST-GCN [29], но оценку позы решили вынести в отдельную работу. Так появился проект MMpose, включающий в себя распознавание ключевых точек и восстановление скелета человека, модель Animal для тех же задач, но на теле животных. Причем задачи, связанные с человеком, имеют решения с получением предсказаний в 2-х мерном и в 3-х мерном вариантах.

Данная модель постоянно улучшается и подключает новые мировые достижения в свои работы. MMpose использует top-down подход в своей работе и другие наработки лаборатории Open-MMLab. Для работы необходимо сначала использовать detection model, а потом уже передать данные о локализованных людях в pose model.

Детектор натренирован на датасете COCO и выдает предсказания в соответствии с его 17 точечной топологией. Также есть скрипты для обучения на других наборах данных, за что спасибо разработчикам. Правда эти наборы данных надо сначала скачать, но об этом в разд. 4.

## 3.2 Модели для классификации позы человека

Далее будет представлено рассмотрение 3 различных модели, которые можно использовать для классификации движений человека.

### 3.2.1 Классификатор от MediaPipe

Проект представил модель для оценки позы человека (см. разд. 3.1.3) и, добавив к нему k-NN, получили алгоритм классификации позы человека. Таким образом MediaPipe представило программу для счета приседаний или отжиманий с камеры смартфона в реальном времени. [30]

Для работы использовалась топология BlazePose, а точнее вектор расстояний между некоторыми точками этой топологии (см. рис. 11). Так как в реальных изображениях бывают разные масштабы и размеры, то все позы нормализуются и переориентируются в вертикальное положение.

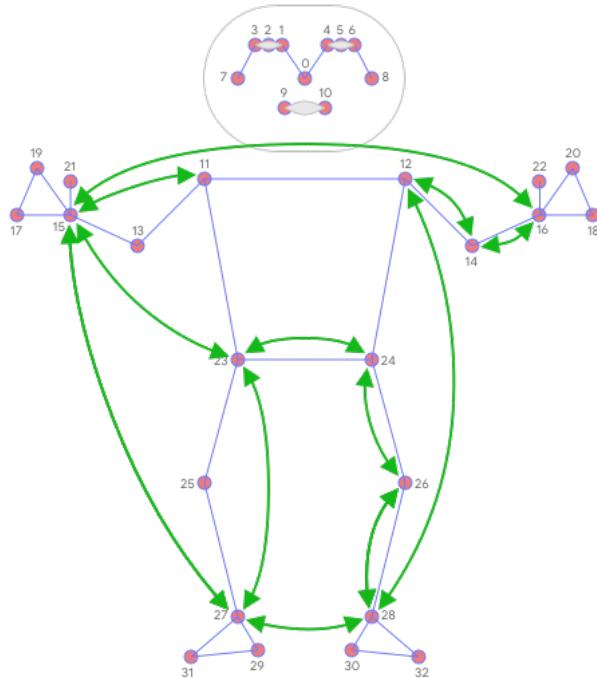


Рис. 11: Расстояния между точками топологии BlazePose, которые используются в классификации. [30]

Для обеспечения точности классификации разработчики решили запускать k-NN дважды: первый раз для отбора векторов признаков для похожих с целевой позой, а во второй раз для отбора точной позы среди выбранных ранее. Различия в том, что используются различные метрики минимальное покоординатное расстояние и среднее покоординатное расстояния для первого и второго прогона соответственно.

### 3.2.2 mmakos

Модель является бакалаврской работой студента Варшавского университета Михала Макоса [31]. Классификатор работает с видеоданными и разделяет модели на два супер класса: статические и динамические.

Разработчик взял за основу модель распознавания позы человека OpenPose (см. разд. 3.1.5) и преобразовал входные данные из 4-х мерного формата в 2-мерное изображение (см. рис. 12). По сути три координаты каждой ключевой точки преобразовывались в BGR формат и создавался столбец пикселей для текущего момента времени. При рассмотрении 15 точек и 32 кадров получаются переходные изображения размером (15, 32, 3).

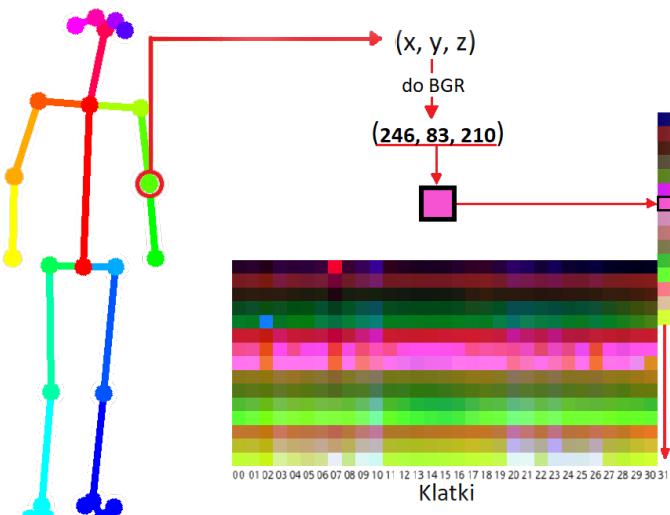


Рис. 12: Преобразование данных в модели от M. Makos. [31]

Для классификации используется версия сети VGG. По решению автора все позы делятся на два супер-класса: динамические и статические, а потом уже классифицируются внутри супер-классов более точно.

### 3.2.3 MMaction2 by OpenMMLab

MMAction2 является подпроектом лаборатории Open-MMLab [32]. Идея классификации движения начиналась с проекта MMSkeleton построенным на исследовании про ST-GCN [29] и позже переросло в первую версию MMAction для классификации движений одного человека. Текущая версия может обрабатывать взаимоотношения между людьми, такие как обнимания или рукопожатия.

Для классификации необходимо сначала обработать видеофрагмент. Для этого производится поиск и локализация действия во времени, а также восстановление и анализ скелета.

В первом случае используется TimeSformer [33] - создан на основе идеи Transformer для обработки видеопотоков. Он использует несколько видов блоков Attention (см. рис. 13) для объединения информации на нескольких кадрах до и после текущего. Это дает возможность точно распознавать действия в кадре и отслеживать их.

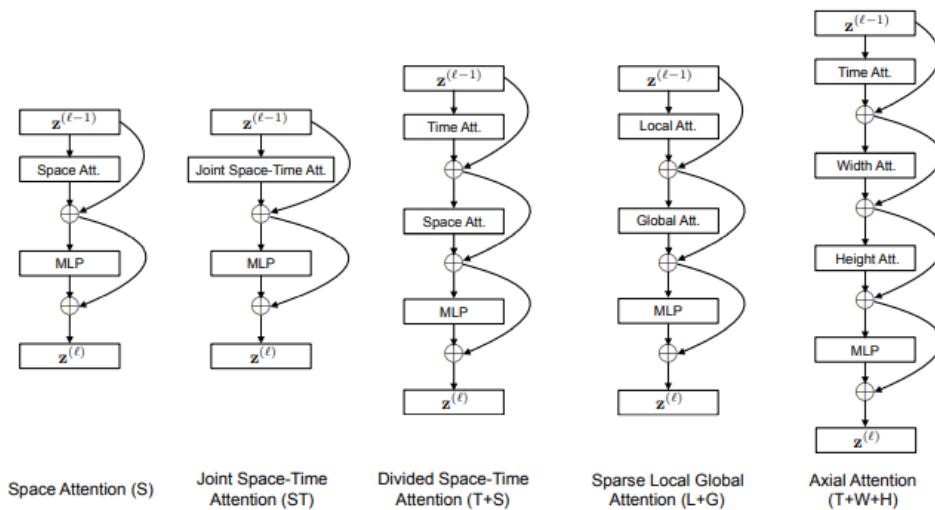


Рис. 13: Примеры блоков внимания в TimeSformer. [33]

Для обработки скелетов используется модель PoseC3D [34]. Она использует несколько тепловых 3D-карт в качестве представления скелета человека и за счет этого обходит GCN и более устойчив к шумам на изображении. Также без особенно лишних затрат можно использовать PoseC3D для классификации взаимодействия между людьми.

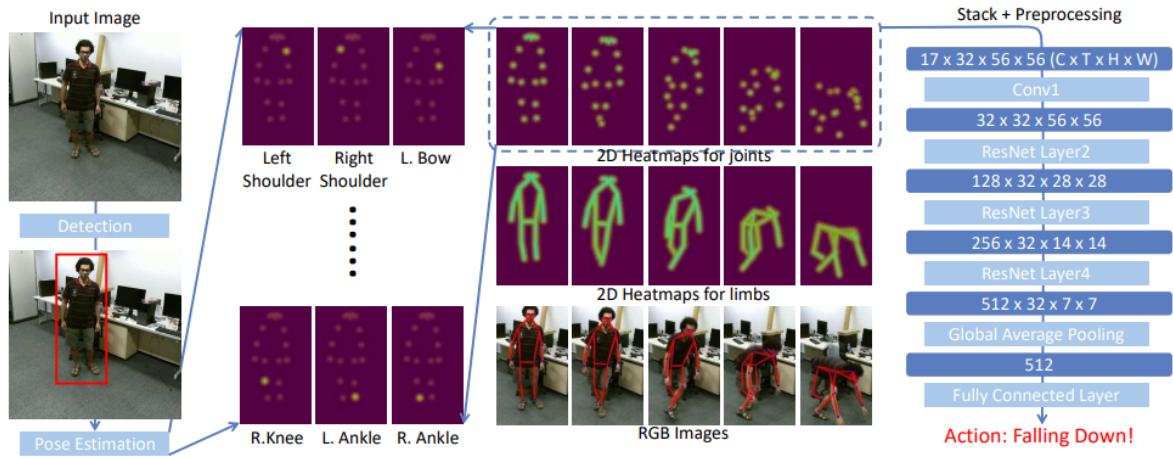


Рис. 14: Структура сети PoseC3D. [34]

Все исследования проекта основаны на классификации и анализе видео, что полезно для будущих исследований и развитии темы работы.

## 4 Исследование моделей

В данной главе будет представлено описание и результаты исследования моделей: BlazePose, MoveNet.SinglePose, OpenPose и MPMpose.

### 4.1 Описание эксперимента

Эксперимент включает рассмотрение только моделей распознавания ключевых точек на теле человека, так как это является основной частью решения задачи классификации движений и многие классификаторы движений строятся на выходных данных о позе.

Из представленных в разд. 3.1 моделей были выбраны 4 представителя по некоторым критериям:

- Доступность модели для исследований

Необходимо оценить длительность установки и возможности работы с различными операционными системами. Эксперимент проводился на платформе Google Colab, поэтому необходимо было рассмотреть возможность использования модели на в Colaboratory.

- Новизна модели

Представленная выборка была создана в основном в 2010-х, но модель DeepPose является самой старой. Новые разработки опирались на результаты, полученные в ней, и таким образом получали более хорошие результаты.

- Наличие документации

Все модели производят классификацию по двум осям изображения: высота и ширина, а также по параметру видимость ключевой точки. Некоторые модели выдают данные нормированные на размер изображения (число из отрезка  $[0,1]$ ), а некоторые точное значение

в пикселях. Поэтому для работы необходимо было понимать как работает API модели, какие у нее входные - выходные данные.

- Тренировка модели на датасете COCO

Все используемые претренированные модели были обучены на наборе данных COCO [12] в совместительстве с каким-либо другим датасетом. В некоторых примерах не было возможности использовать претренированную модель и из-за этого они были отсечены.

В итоге было выбрано 4 модели наиболее подходящие под критерии:

1. BlazePose
2. MoveNet.SinglePose
3. OpenPose
4. Mmpose

Для проведения качественного анализа и выявления лучшей модели необходимо их сравнить. Поэтому рассмотрим метрики, которые подходят для задач в 2-х мерном пространстве:

- Percentage of Detection Joints

PDJ оценивает точность распознавания ключевой точки в зависимости от диагональных размеров человека. При рассмотрении задачи распознавания человека на выходе имеются координаты точек, которые характеризуют прямоугольник, внутри которого вписан человек. Диагональ этого прямоугольника используется при вычислении метрики PDJ (см. рис. 15). Формулу можно представить в следующем виде:

$$PDJ = \frac{\sum_{i=1}^n \text{bool}(d_i < \text{threshold} * \text{diag})}{n}, \quad (1)$$

где

$d_i$  - расстояние между предсказанной и правильной точкой,

$threshold$  - порог, задаваемый исследователем,

$diag$  - размер диагонали прямоугольника, внутри которого находится человек,

$bool()$  - логическое условие, возвращает 1, если оно верно и 0 в ином случае,

$n$  - размер выборки.

С помощью значения порога можно варьировать допустимую погрешность расстояния между истинной и предсказанной точками.

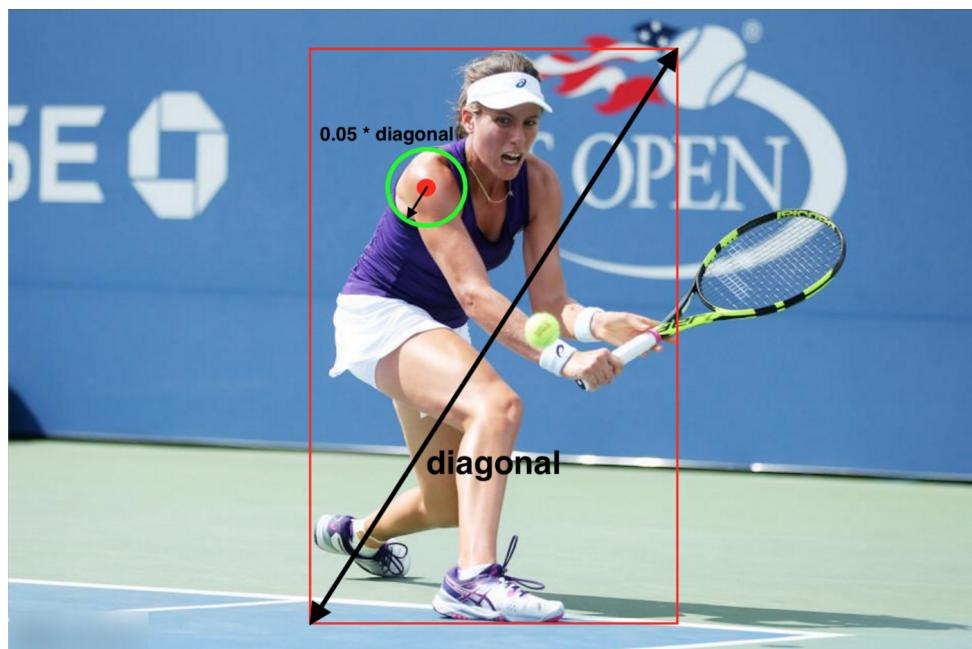


Рис. 15: Визуальное представление метрики PDJ.

Зеленый круг ограничивает область допустимого расположения распознанной ключевой точки.

- Percentage of Correct Key-points

ПСК очень похожа на предыдущую метрику, только погрешность рассматривается относительно высоты человека. Формулу можно

представить в следующем виде:

$$PDJ = \frac{\sum_{i=1}^n \text{bool}(d_i < \text{threshold} * \text{bodyheight})}{n}, \quad (2)$$

где

$d_i$  - расстояние между предсказанной и правильной точкой,

$\text{threshold}$  - порог, задаваемый исследователем,

$\text{bodyheight}$  - высота прямоугольника, внутри которого находится человек,

$\text{bool}()$  - логическое условие, возвращает 1, если оно верно и 0 в ином случае,

$n$  - размер выборки.

- Object Key-point Similarity

OKS является основной при оценке задачи Keypoint Detection COCO [12]. Она использует третью координату выходного предсказания и расстояние между реальной и предсказанной точками. Формулу можно представить в следующем виде:

$$OKS = \frac{\sum_i \exp\left(-d_i^2/2s^2k_i^2\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \quad (3)$$

где

$d_i$  - расстояние между предсказанной и правильной точкой,

$s$  - площадь объекта,

$k_i$  - константа ключевой точки, контролирующая спад,

$v_i$  - видимость.

При оценке задачи детекции ключевых точек COCO вводится метрика Average Precision (AP) через OKS. Изменяя границу допустимого значения OKS можно получать различные значения precision и AP.

В метриках РСК и РДЖ необходимо знать размеры прямоугольника, ограничивающего человека. Для этого необходимо использовать модель распознавания объектов, которая будет давать нам эти данные. При работе с метрикой РСК можно обойтись без такой модели, потому что во всех топологиях есть точки, которые обозначают верхнюю и нижнюю границы человека. Отсюда погрешность вычисления через модель и вычисления разности ординат верхней и нижней точек будет мала. Что требует меньше затрат для оценки.

## 4.2 Поиск данных

Первым делом необходимо было проверить модели на неразмеченных данных. Для этого были выбраны фотографии высокого разрешения, где человек изображен во весь рост (см. рис. 16).

Для качественной оценки работы моделей с помощью метрик необходимо было найти размеченные данные. Приведу описание датасетов, которые были мной рассмотрены:

- COCO Dataset и MPII

Данные наборы являются основными при работе с компьютерным зрением и большинство исследователей используют их как тренировочные для своих моделей. Поэтому использовать их в качестве тестовых не целесообразно. [13, 35]

- HUMAN 3.6M

Данные собирались специально для задачи классификации движений человека в студии. С помощью датчиков фиксировались положения всех суставов и ключевых точек. Это идеальный датасет, но доступ к нему ограничен и создатель не выходит на связь. [1]



(a)



(b)



(c)



(d)

Рис. 16: Изображения для визуальной оценки моделей.

- LSP

Данные собраны со спортивных соревнований и пред обработаны до обозначения одного человека на изображении размером не менее 150 пикселей в высоту. Единственная проблема - не описаны точки на лице, поэтому оценивать можно только распознавание 12 суставов или, другими словами, точек на теле человека. [36]

- LSP Extended

Является расширенной версией предыдущего набора данных. По объёму превосходит его в пять раз. Остальные характеристики не

поменялись. [37]

- Halpe

Датасет для распознавания точек на всем теле человека. К каждой фотографии идет аннотация из 136 точек, краев ограничивающего прямоугольника и категории детектируемого объекта (во всех фотографиях стоит категория человек).

Набор состоит из двух частей: HICO-DET и COCO. Все они аннотированы описанные выше способом. Для наших моделей необходимо будет отобрать только 17 точек, совпадающих с топологией COCO. [38]

Выше были представлены наборы данных для задачи распознавания точек на теле человека. Но основной темой является классификация движений человека, поэтому необходимы фотографии с меткой класса для позы, представленной на данных. Некоторые из уже представленных (COCO, MPII) тоже могут использоваться для классификации позы, но по тем же причинам, что и описаны выше, они не будут рассмотрены в эксперименте. Приведу описание датасетов для классификации движения человека по позе на изображении:

- Stanford-40

Набор данных состоит примерно из 10 тыс. изображений, на каждом из которых человек делает одно из 40 действий. На один класс приходится от 180 до 300 фотографий. [39]

- Yoga-82

Сложность распознавания поз йоги в том, что многие из них не могут быть точно аннотированы. Для решения этой проблемы и был собран этот набор данных вмещающий примерно 28 тыс. изображений. 82 различные позы разделены по 6 подклассам. При желании

можно использовать метки и классов, и подклассов для классификации. [40]

Итого были выбраны наборы данных LSP, LSPE и Halpe.



Рис. 17: Примеры изображений из датасета LSP.

### 4.3 Результаты эксперимента

Для каждой модели была рассмотрена локализация ключевых точек на фотографиях высокого качества (см. рис. 16), а также набор данных низкого качества с обработанными изображениями.

Метрики были рассчитаны с порогами 0.05, 0.2, 0.5. В дополнение к этому был проведен временной анализ классификации одного изображения в среднем по целому датасету.

Перейдем к результатам по каждой модели.

#### BlazePose

Результаты работы модели можно посмотреть на рис. 21. Как можно заметить, некоторые точки не распознаются и не отображаются на итоговом результате. А также есть небольшая погрешность при распознавании глаз. Но модель хорошо сработала на человека в маске.

Исходя из значений метрик (см. в табл. 1) можно сделать вывод, что модель имеет высокую погрешность при среднем времени обработки одной фотографии 0.063 секунды.

Metric	PCK@0.05	PCK@0.2	PCK@0.5	PDJ@0.05	PDJ@0.2	PDJ@0.5
LSP	0.035	0.304	0.741	0.042	0.359	0.815
LSPE	0.005	0.059	0.277	0.009	0.108	0.457
Halpe	0.004	0.051	0.222	0.005	0.077	0.313

Metric	AP@0.5	AP@0.75	AP
LSP	0.012	0.004	0.006
LSPE	0.03	0.001	0.001

Таблица 1: Результаты вычисления метрик для BlazePose.

## MoveNet.SinglePose

Результаты работы модели можно посмотреть на рис. 22. Заметны ограхи в распознавании локтей и ладоней. Как и в прошлой модели есть неточности при распознавании частей лица.

Исходя из значений метрик (см. в табл. 2) можно сделать вывод, что модель показала себя плохо при оценивании с порогом 0.05, но лучший результат среди всех для порогах 0.2 и 0.5. Среднее время обработки одной фотографии составило 0.009 секунды, что также является наилучшим показателем.

Metric	PCK@0.05	PCK@0.2	PCK@0.5	PDJ@0.05	PDJ@0.2	PDJ@0.5
LSP	0.356	0.928	0.983	0.43	0.95	0.993
LSPE	0.144	0.539	0.745	0.227	0.635	0.806
Halpe	0.04	0.052	0.223	0.006	0.078	0.32

Metric	AP@0.5	AP@0.75	AP
LSP	0.168	0.064	0.08
LSPE	0.312	0.059	0.111

Таблица 2: Результаты вычисления метрик для MoveNet.SinglePose.

## OpenPose

Результаты работы модели можно посмотреть на рис. 23.

Значения метрик см. в табл. 3.

Среднее время обработки одной фотографии составило 0.037 секунды.

Визуально лучший результат обработки изображений, хотя по метрикам немного уступает следующей модели. Но полностью обыгрывает ее во времени обработки.

Metric	PCK@0.05	PCK@0.2	PCK@0.5	PDJ@0.05	PDJ@0.2	PDJ@0.5
LSP	0.708	0.837	0.882	0.746	0.845	0.899
LSPE	0.362	0.524	0.633	0.42	0.569	0.69
Halpe	0.575	0.638	0.686	0.613	0.658	0.72

Metric	AP@0.5	AP@0.75	AP
LSP	0.166	0.085	0.092
LSPE	0.538	0.362	0.362
Halpe	0.562	0.398	0.384

Таблица 3: Результаты вычисления метрик для OpenPose.

## MMPose

Результаты работы модели можно посмотреть на рис. 24.

Значения метрик см. в табл. 4.

Среднее время обработки одной фотографии составило 0.409 секунды.

Для своей работы требует детекции человека, из-за чего показывает лучший результат в точности, но худший результат по времени обработки одного изображения.

Metric	PCK@0.05	PCK@0.2	PCK@0.5	PDJ@0.05	PDJ@0.2	PDJ@0.5
LSP	0.728	0.85	0.914	0.76	0.858	0.933
LSPE	0.403	0.56	0.651	0.463	0.587	0.72
Halpe	0.434	0.49	0.545	0.462	0.507	0.585
	AP@0.5	AP@0.75	AP			
LSP	0.225	0.223	0.123			
LSPE	0.598	0.462	0.443			
Halpe	0.692	0.579	0.544			

Таблица 4: Результаты вычисления метрик для MMPose.

## Визуализация результатов

На рис. 18 и рис. 19 представлены средние для трех датасетов (LSP, LSPE и Halpe) метрик PCK и PDJ соответственно. Модель BlazePose является явным аутсайдером. OpenPose и MMPose показывают качественные результаты при любых порогах допустимой погрешности. А MoveNet имеет лучшие результаты при большей ошибке.

Результаты растут с увеличением порога, так как увеличивается и допустимая погрешность локализации. Следовательно и количество верных точек тоже становится больше.

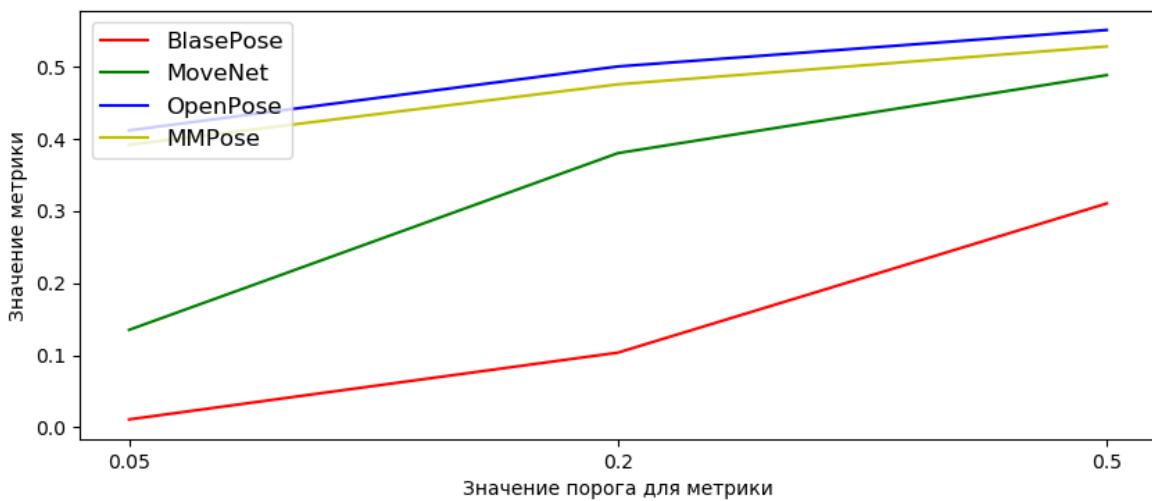


Рис. 18: Средние значения метрики PCK.

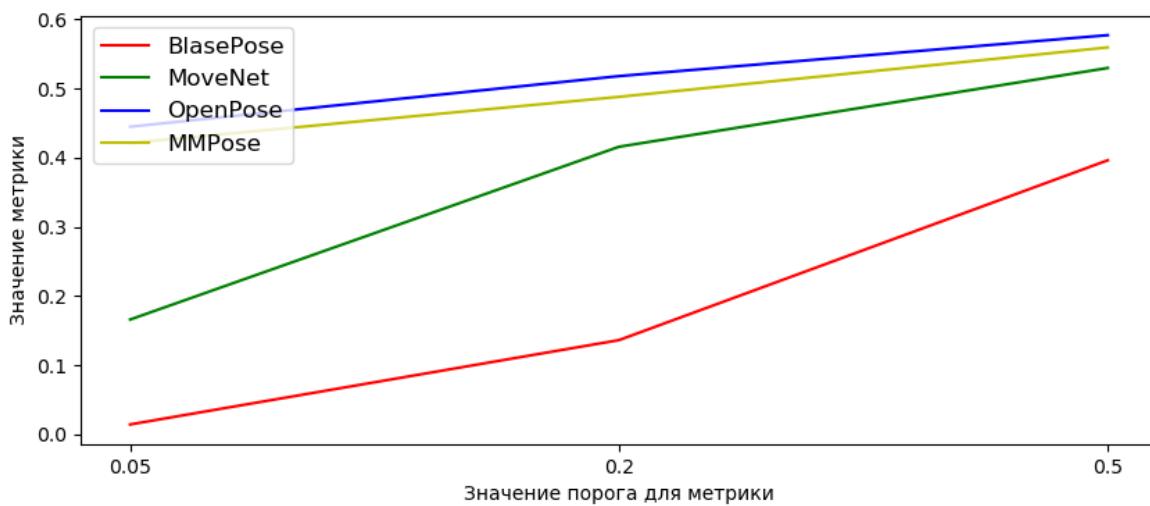


Рис. 19: Средние значения метрики PDJ.

На рис. 20 представлены значения метрики AP также по всем трем наборам данных. Здесь первые две модели (BlazePose и MoveNet) ушли в аутсайдеры из-за плохого результата на изображениях из Halpe, в то время как две другие (OpenPose и MMPose) наоборот выросли за счет этих фотографий.

Результаты данной метрики убывают, так как с увеличением порога метрики растет минимально допустимое значение OKS, которое необходимо для определения верного распознавания.

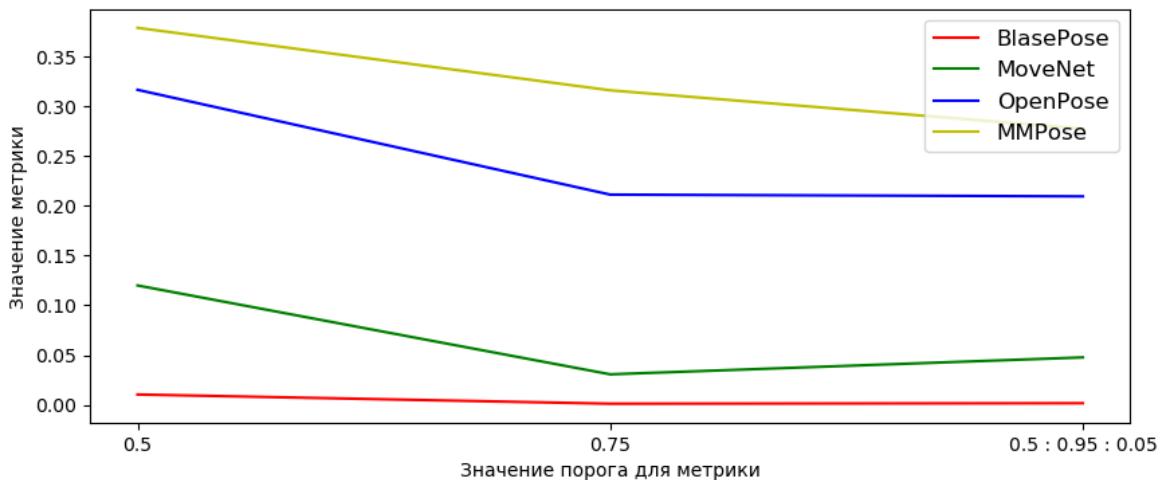


Рис. 20: Средние значения метрики AP.

## 5 Заключение

В ходе работы было рассмотрено несколько решений задачи классификации движений человека, в основе которых лежат модели распознавания ключевых точек. Так как данная задача выглядит более простой в сравнении с задачей получения векторов признаков - были рассмотрены современные проекты для оценки позы.

Результаты эксперимента показали:

- При визуальной оценке работы различных моделей (см. прил. А) выделяются OpenPose и MoveNet как адекватностью локализации, так и качеством отрисовки изображения;
- Модель MoveNet показывает самое быстрое распознавание точек, правда точность модели уступает предыдущим двум;
- Модели OpenPose и MMPose являются лучшими решениями из представленных по точности детектирования ключевых точек.

Подводя итог, если не обращать внимание на высокую точность (порог метрик PDJ и РСК 0.2 и больше) можно выбрать модель MoveNet - так как она показала себя хорошо в данном вопросе. Если необходима достоверность рассматриваемых результатов, то стоит обратиться к OpenPose или MMPose. Но стоит отметить, что первая работает значительно быстрее, чем вторая.

Дальнейшая цель исследования заключается в том, чтобы рассмотреть задачу анализа движений человека по видеофрагментам и попробовать улучшить некоторые модели посредством архитектурных изменений или используя обучение на узко ориентированном наборе данных.

## Список литературы

- [1] Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments / Catalin Ionescu, Dragos Papava, Vlad Olaru, Cristian Sminchisescu // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. — 2014. — Pp. 1325–1339.
- [2] Deep 3D human pose estimation: A review / Jinbao Wang, Shujie Tan, Xiantong Zhen et al. // *Computer Vision and Image Understanding*. — 2021. — P. 103225.
- [3] Tome, Denis. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image / Denis Tome, Chris Russell, Lourdes Agapito // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2017. — Pp. 5689–5698.
- [4] Detailed, Accurate, Human Shape Estimation From Clothed 3D Scan Sequences / Chao Zhang, Sergi Pujades, Michael J. Black, Gerard Pons-Moll // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2017.
- [5] Learning from Synthetic Humans / Gül Varol, Javier Romero, Xavier Martin et al. // CVPR. — 2017.
- [6] Whittle, Michael W. Clinical gait analysis: A review / Michael W. Whittle // *Human Movement Science*. — 1996. — Pp. 369–387.
- [7] google.github.io. Pose Classification. — [https://google.github.io/mediapipe/solutions/pose\\_classification.html](https://google.github.io/mediapipe/solutions/pose_classification.html).

- [8] OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / Z. Cao, G. Hidalgo Martinez, T. Simon et al. // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. — 2019.
- [9] Kocabas, Muhammed. MultiPoseNet: Fast Multi-Person Pose Estimation using Pose Residual Network. — 2018.
- [10] RMPE: Regional Multi-person Pose Estimation / Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu // ICCV. — 2017.
- [11] Prediction of Human Activities Based on a New Structure of Skeleton Features and Deep Learning Model / Neziha Jaouedi, Francisco J. Perales, José María Buades et al. // *Sensors*. — 2020. — no. 17.
- [12] Tsung-Yi Lin Matteo Ruggero Ronchi, Alexander Kirillov. COCO 2020 Keypoint Detection Task. — <https://cocodataset.org/#keypoints-2020>.
- [13] Lin, Tsung-Yi. Microsoft COCO: Common Objects in Context. — 2014.
- [14] Bazarevsky, Valentin. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. — 2019.
- [15] Zhang, Fan. MediaPipe Hands: On-device Real-time Hand Tracking. — 2020.
- [16] Bazarevsky, Valentin. BlazePose: On-device Real-time Body Pose tracking. — 2020.
- [17] Ivan Grishchenko, Valentin Bazarevsky. MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device. — <https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html>.

- [18] PaStaNet: Toward Human Activity Knowledge Engine / Yong-Lu Li, Liang Xu, Xinpeng Liu et al. // CVPR. — 2020.
- [19] Osokin, Daniil. Real-time 3D Multi-person Pose Estimation Demo. — <https://github.com/Daniil-Osokin/lightweight-human-pose-estimation-3d-demo.pytorch>.
- [20] Saksena, Saumya Kumaar. Human Action Classification. — <https://github.com/dronefreak/human-action-classification>.
- [21] Toshev, Alexander. DeepPose: Human Pose Estimation via Deep Neural Networks / Alexander Toshev, Christian Szegedy // 2014 IEEE Conference on Computer Vision and Pattern Recognition. — 2014. — Pp. 1653–1660.
- [22] google.github.io. MediaPipe.Home. — <https://google.github.io/mediapipe/>.
- [23] Votel, Ronny. Next-Generation Pose Detection with MoveNet and TensorFlow.js. — <https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html>.
- [24] Sandler, Mark. MobileNetV2: Inverted Residuals and Linear Bottlenecks. — 2019.
- [25] Lin, Tsung-Yi. Feature Pyramid Networks for Object Detection. — 2016.
- [26] Zhou, Xingyi. Objects as Points. — 2019.
- [27] Cao, Zhe. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. — 2016.
- [28] Contributors, MMPose. OpenMMLab Pose Estimation Toolbox and Benchmark. — <https://github.com/open-mmlab/mmpose>. — 2020.

- [29] *Yan, Sijie.* Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. — 2018.
- [30] *google.github.io.* Pose Classification. — [https://google.github.io/mediapipe/solutions/pose\\_classification.html](https://google.github.io/mediapipe/solutions/pose_classification.html).
- [31] *MAKOŚ, MICHAŁ.* Human Pose Classification - BEng Thesis. — <https://github.com/mmakos/HPC>.
- [32] *Contributors, MMACTION2.* OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. — <https://github.com/open-mmlab/mmaction2>. — 2020.
- [33] *Bertasius, Gedas.* Is Space-Time Attention All You Need for Video Understanding? / Gedas Bertasius, Heng Wang, Lorenzo Torresani // Proceedings of the International Conference on Machine Learning (ICML). — 2021. — July.
- [34] *Duan, Haodong.* Revisiting Skeleton-based Action Recognition. — 2021.
- [35] 2D Human Pose Estimation: New Benchmark and State of the Art Analysis / Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Bernt Schiele // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2014.
- [36] *Johnson, Sam.* Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation / Sam Johnson, Mark Everingham // Proceedings of the British Machine Vision Conference. — 2010.
- [37] *Johnson, Sam.* Learning Effective Human Pose Estimation from Inaccurate Annotation / Sam Johnson, Mark Everingham // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. — 2011.

- [38] *Haoshu, Fang.* Halpe Full-Body Human Keypoints and HOI-Det dataset.  
— <https://github.com/Fang-Haoshu/Halpe-FullBody>.
- [39] Human action recognition by learning bases of action attributes and parts / Bangpeng Yao, Xiaoye Jiang, Aditya Khosla et al. — 2011. — Pp. 1331–1338.
- [40] Yoga-82: A New Dataset for Fine-grained Classification of Human Poses / Manisha Verma, Sudhakar Kumawat, Yuta Nakashima, Shanmuganathan Raman // IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). — 2020. — Pp. 4472–4479.

## A Приложение

### Визуальные результаты работы моделей



(a)



(b)



(c)



(d)

Рис. 21: Пример результатов работы модели BlazePose.



Рис. 22: Пример результатов работы модели MoveNet.SinglePose.



Рис. 23: Пример результатов работы модели OpenPose.

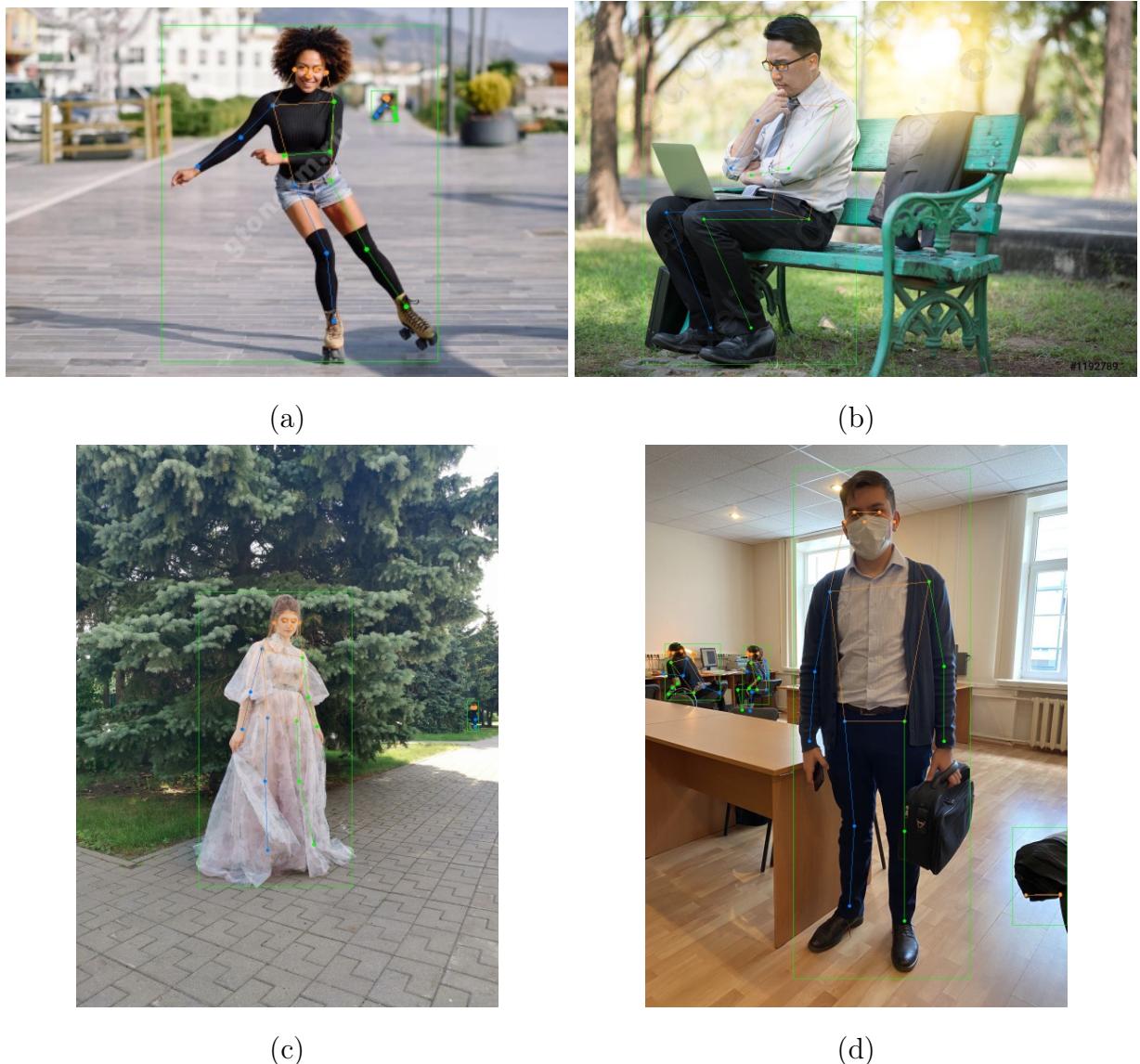


Рис. 24: Пример результатов работы модели MMPose.