УДК 004.932.72’1
DOI 10.17513/snt.39629

АНАЛИЗ ПОДХОДОВ, МЕТОДОВ И РЕШЕНИЙ ДЛЯ ДЕТЕКТИРОВАНИЯ ПОЗЫ ЧЕЛОВЕКА. ВЫБОР ИНСТРУМЕНТА ДЛЯ ЗАДАЧИ ОПРЕДЕЛЕНИЯ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ЧЕЛОВЕКА ПО ЕГО ПОЗЕ

Киселев Ю.В., Богомолов И.А., Розалиев В.Л., Баклан В.А.

ФГБОУ ВО «Волгоградский государственный технический университет», Волгоград,
e-mail: agonmountain@yandex.ru, bogomolov222@gmail.com,
vladimir.rozaliev@gmail.com, baklanv84@gmail.com

Детектирование позы и частей тела человека используется в различных задачах оценки человека, его состояния, поведения и намерений. Определение эмоционального состояния человека по его позе является одной из таких задач. Целью данной работы является анализ и сравнение существующих подходов, методов и решений для детектирования позы и частей тела человека. Авторами были выделены маркерный и безмаркерный подходы к детектированию позы человека. Проанализировав оба подхода, выбрали безмаркерный подход. Были рассмотрены решения безмаркерного подхода AlphaPose, MoveNet, MediaPipe (версии Pose, Holistic), OpenPose (версии CPU, GPU), DeeperCut и YOLOv7. Для рассмотренных решений авторами был проведен эксперимент. На вход каждого решения поступало 10 заранее подготовленных изображений, содержащих человека в различных позах. Изображения отличались как по сложности позы, так и по расположению человека в кадре (человек в полный рост, человек по пояс). Результатом эксперимента стали оценки качества детектирования, времени детектирования и технических требований для каждого решения. На основе анализа результатов эксперимента авторами были выявлены преимущества решений MediaPipe Holistic и OpenPose GPU относительно остальных рассмотренных решений для задачи определения эмоционального состояния человека по его позе.

Ключевые слова: поза, жест, жестикуляция, часть тела, человек, детектировать, безмаркерное, компьютерное зрение, нейронная сеть, метод, сверху вниз, снизу вверх, эмоциональное состояние

ANALYSIS OF APPROACHES, METHODS AND SOLUTIONS FOR DETECTING HUMAN POSTURE. CHOOSING A TOOL FOR THE TASK OF DETERMINING THE EMOTIONAL STATE OF A PERSON BY HIS POSTURE

Kiselev Yu.V., Bogomolov I.A., Rozaliev V.L., Baklan V.A.

Volgograd State Technical University, Volgograd,
e-mail: agonmountain@yandex.ru, bogomolov222@gmail.com,
vladimir.rozaliev@gmail.com, baklanv84@gmail.com

The detection of a person’s posture and body parts is used in various tasks of assessing a person, his condition, behavior and intentions. Determining the emotional state of a person by his posture is one of such tasks. The purpose of this work is to analyze and compare existing approaches, methods and solutions for detecting posture and human body parts. The authors identified marker-based and marker-free approaches to detecting a person’s posture. After analyzing both approaches, a marker-free approach was chosen. Solutions of the marker-free approach AlphaPose, MoveNet, MediaPipe (Pose, Holistic versions), OpenPose (CPU, GPU versions), DeeperCut and YOLOv7 were considered. For the considered solutions, the authors conducted an experiment. At the input of each solution, 10 pre-prepared images containing a person in various poses were received. The images differed both in the complexity of the pose and in the location of the person in the frame (person in full height, person to the waist). The experiment resulted in the evaluation of detection quality, detection time and technical requirements for each solution. Based on the analysis of the experimental results, the authors identified the advantages of MediaPipe Holistic and OpenPose GPU solutions relative to the other solutions considered for the task of determining a person’s emotional state by his posture.

Keywords: pose, gesture, gesticulation, body part, person, detect, marker-free, computer vision, neural network, method, top-down, bottom-up, emotional state

Детектирование позы является актуальной задачей, которая позволяет получать данные о положении частей тела человека, на основе которых возможно проводить анализ по множеству направлений, таких как оценка состояния, качества движений, взаимодействия с объектами реального мира и виртуальной реальности.

Одной из таких задач является оценка эмоционального состояния человека по его позе. Это сложная многоэтапная задача, решение которой позволит формировать оценку состояния человека, прогнозировать его поведение и определять причины совершенных им действий [1, 2]. Предложенную задачу можно разделить на два этапа: на первом этапе происходит детектирование позы человека, на втором – оценка состояния человека по данным о позе.

Целью данной работы является анализ и сравнение существующих подходов, методов и конкретных решений для детек-

тирования позы и частей тела человека. Результаты данной работы будут задействованы при разработке программного решения по определению эмоционального состояния человека по его позе с использованием существующих решений детектирования.

Подходы для детектирования позы и частей тела человека

Можно выделить два подхода для детектирования частей тела и позы человека. Первый подход заключается в использовании специальных инструментов – маркеров, которые располагают на определенных позициях частей тела человека. Координаты маркеров в пространстве определяют положения частей тела и позы человека. Считывание положения маркера в пространстве осуществляется с помощью соответствующего оборудования, которое размещается по периметру комнаты или ее области. Для считывания маркеров детектируемый человек должен находиться внутри данной выделенной области. Детектирование маркеров в пространстве зависит от вида используемого маркера. Существуют разные виды маркеров: отражающие свет, поглощающие свет, использующие системы из гироскопов, акселерометров и магнетометров.

При высокой точности детектирования недостатками данного подхода являются такие особенности, как необходимость наличия дорогостоящего оборудования – маркеры и системы считывания маркеров, пространство для установки системы считывания маркеров, влияние на движения анализируемого человека (появляется скованность движений у человека, на котором физически размещают маркеры и дополнительные устройства для их работы, такие как аккумуляторы, передатчики и др.).

Примером решения с описанным подходом служит работа Ашраф Шарифа и др. [3], которая использует маркерную систему детектирования позы. Решение способно определять шесть действий – ходьбу, прыжок, скакалку, взмах, растяжку, пробежку трусцой.

Второй подход – безмаркерный. Он основывается на детектировании ключевых точек тела на основе анализа частей тела человека на изображении с помощью нейронных сетей. Под ключевыми точками понимаются точки, которые обычно совпадают с местами суставов тела человека или другими важными элементами (уши, нос, глаза и др.). Данный подход является менее точным, чем при использовании маркерной системы, но позволяет при относительно малой потере качества детектирования избавиться от существенных недостатков маркерной системы.

Выбор безмаркерного подхода позволит избежать дополнительных затрат на дорогостоящую систему на базе маркеров и необходимость в создании подходящей среды для ее развертывания. Для задачи определения эмоционального состояния человека по его позе, безмаркерный подход является более успешным, поскольку он не воздействует на анализируемого человека, сковывая движения и влияя на его позу.

Методы безмаркерного детектирования позы и частей тела человека

Методы безмаркерного детектирования позы по изображению с использованием нейронных сетей можно разделить на два типа: методы «сверху вниз» и методы «снизу вверх».

Первый тип методов, «сверху вниз», характеризуется первостепенным детектированием человека на изображении, выделяя ограничительную область, на которой присутствует человек. Далее эта область будет подвергнута анализу для определения отдельных частей тела человека на ней. Частая проблема данного типа методов заключается в неверном определении ограничительной области, в рамках которой будет проводиться дальнейший анализ. Поскольку поиск частей тела выполняется только внутри данной области, возможна потеря тех или иных частей позы, которые оказались за пределами ограничительной области. В некоторых случаях возможна полная потеря позы при ее детектировании.

Второй тип методов, «снизу вверх», изначально детектирует все части тела человека на изображении, которые смог найти детектор, после чего выделяет человека, группируя найденные части тела в единый объект, ориентируясь на их расположение и направленность в кадре. Детектирование частей тела основывается на анализе вероятностной карты, которую строит детектор перед выделением частей тела на изображении. Вероятностная карта – это карта областей на изображении, в которых может находиться та или иная часть тела детектируемого человека. Данный тип методов является более успешным при задачах определения группы людей на изображении, позволяя избежать проблемы пересечения конечностей и соотношения их между разными людьми. Вытекающим недостатком является более долгое выполнение, поскольку методы данного типа детектируют не выделенную область, а большую часть изображения.

Решения для детектирования позы человека по изображению с помощью нейронных сетей

Авторами данной работы были рассмотрены решения AlphaPose, MediaPipe (версии Pose, Holistic), MoveNet, OpenPose (версии CPU, GPU), DeeperCut и YOLOv7.

AlphaPose – модель типа «сверху вниз», обеспечивающая детектирование множества людей в кадре. Решение способно определять до 26 ключевых точек тела, до 42 точек в сумме для обеих кистей рук и до 68 точек лица [4].

MediaPipe Holistic – модель типа «сверху вниз», ориентирующаяся на менее мощные устройства выполнения. В решении используется детектор, вдохновленный моделью BlazeFace (быстрый и легкий детектор ключевых точек лица). Он явно предсказывает две дополнительные виртуальные точки, которые позволяют описать центр тела человека, его вращение и масштаб в виде круга. Данное решение является совокупностью решений MediaPipe Pose, MediaPipe Hands и MediaPipe Face Mesh и способно определять до 33 ключевых точек тела, до 42 точек в сумме для обеих кистей рук и до 468 точек лица [5, 6]. В данной работы в эксперименте были протестированы версии MediaPipe Pose и MediaPipe Holistic.

MoveNet – модель типа «снизу вверх», имеет два исполнения – Thunder и Lightning. Thunder предназначен для приложений, которые требуют высокой точности, Lightning предназначен для приложений, для которых задержка является более критичным фактором. MoveNet доступно определение до 17 ключевых точек тела человека. Данное решение, как и решение MediaPipe, ориентируется на менее мощные устройства выполнения и способно запускаться в том числе на мобильных устройствах [7]. В данной работе была использована версия Thunder.

OpenPose – модель типа «снизу вверх», обеспечивающая детектирование множества людей в кадре. Решение определяет до 21 ключевой точки тела, до 42 точек в сумме для обеих кистей рук и до 70 точек лица. С ростом количества детектируемых людей на изображении – время выполнения остается постоянным, по сравнению с аналогичными решениями [8, 9]. Решение OpenPose имеет две версии, для работы на CPU (центральный процессор) и GPU (графический процессор). В данной работе были протестированы обе версии.

DeeperCut – модель типа «снизу вверх», обеспечивающая детектирование множества людей в кадре и определяющая до 14 ключевых точек тела. Решение имеет ограничение на среду выполнения и выполняется только на системе под управлением Linux [10].

YOLOv7 – модель типа «снизу вверх», способная определять несколько людей в кадре. Она отличается от других моделей типа «снизу вверх» упрощением детектирования частей тела на изображении, не прибегая к построению «тепловой карты» (вероятностной карты), которое стало возможным благодаря улучшению расширения YOLO-Pose [11, 12]. Решение способно определять до 17 ключевых точек тела.

Сравнение решений на основе эксперимента

Характеристики устройства, на котором проводился эксперимент, представлены в табл. 1.

Таблица 1

Характеристики устройства, на котором проводился эксперимент

Тип	Модель / характеристика
OS (операционная система)	Windows 10 / Linux Ubuntu 22.04
CPU (центральный процессор)	Intel Core i5 6700HQ 2.30 GHz
GPU (графический процессор)	Nvidia GeForce 960M
RAM (оперативная память)	8 Gb
SSD (твердотельный накопитель)	Samsung SSD EVO 860 M.2

Авторами был проведен эксперимент для сравнительного анализа рассмотренных решений. Для представленных в работе решений выявлены сравнительные характеристики, исходящие из быстродействия, качества анализа данных исходных изображений, а также требуемые технические мощности и ресурсы для их работы.

Для эксперимента были выбраны десять изображений с одним человеком для детектирования в кадре (источником изображений является личная медиатека одного из авторов данной работы). Изображения отличаются друг от друга сложностью позы и расположением человека в кадре (в кадре человек в полный рост, человек по пояс). Все изображения были уменьшены и приведены к общему разрешению, которое составило до 800 пикселей в ширину и до 800 пикселей в высоту. Оригинальное соотношение сторон изображений сохранено.

Таблица 2

Критерии оценки технических требований

Критерий оценки	Способ вычисления	Оценка критерия и соответствующий ей числовой интервал		
		Низкая	Средняя	Высокая
Время выполнения	Оценка временного интервала выполнения детектирования	< 0,5 с	> 0,5 с	> 1,0 с
Количество потребляемых ресурсов	Оценка процентного потребления ресурсов CPU, GPU и RAM на временном промежутке выполнения детектирования	> 0,06	> 0,18	> 0,30

Тестирование проводилось пять раз для каждого решения. На вход поступало десять изображений, которые последовательно проходили через тестируемое решение. Качество детектирования, как и продолжительность выполнения, определялось средним значением за весь эксперимент. Оценка технических требований определялась на основе времени выполнения и количества потребляемых ресурсов устройства выполнения. Все решения были настроены на детектирование ключевых точек тела одного человека в кадре.

Критерии оценки технических требований приведены в табл. 2.

Оценка технических требований строится на оценках времени выполнения и количества потребляемых ресурсов, т.е. на оценках критериев. Для каждого критерия определяется его оценка, и далее по ним вычисляется итоговая оценка технических требований. В табл. 2 приведены оценки по числовым значениям критериев.

Количество потребляемых ресурсов определяется по формуле

$$y = \left(\frac{p_{CPU}}{100} * \frac{p_{RAM}}{100} * \frac{p_{GPU}}{100} \right), \qquad (1)$$

где параметры p_{CPU}, p_{RAM}, p_{GPU} принимают максимальные значения процентов задействованных ресурсов системы во время всего детектирования для CPU, RAM и GPU соответственно.

Если при выполнении детектирования решением не был задействован GPU, то потребление ресурсов GPU не учитывалось и последний множитель p_{GPU} / 100 из формулы (1) не участвовал в произведении.

Чтобы применить оценки по критериям для вычисления итоговой оценки необходимо перевести их в числовые значения. На основе проведенной работы было установлено, что для оценок по критериям следует использовать числовые значения 60 (низкая оценка), 80 (средняя оценка) и 100 (высокая оценка).

Числовое значение итоговой оценки технических требований определялось по формуле

$$x = \frac{\left(k_a + k_b\right)}{2} + c, \qquad (2)$$

где параметры k_a и k_b являются оценками критериев (время выполнения и количество потребляемых ресурсов) и принимают числовые значения 60, 80 и 100, как это было описано ранее. Значение c выступает в роли компенсирующего параметра и принимает значение 10.

Результатом вычисления по формуле (2) будет являться числовое значение на промежутке от 70 до 110 включительно. Оценка интерпретируется принадлежностью числового значения итоговой оценки одному из интервалов, отображенных в табл. 3.

Таблица 3

Интерпретация числового значения оценки

Оценка	Числовой интервал оценки
Низкая	$x \geq 60$
Средняя	$x \geq 80$
Высокая	$x \geq 100$

Так, итоговая оценка технических требований при оценках критериев – средняя и высокая, по формуле (2) будет составлять значение 100, которое интерпретируется по табл. 3 в оценку – высокая.

Оценка качества детектирования определялась количеством найденных ключевых точек позы человека на изображении и их приближенности к истинным (действительным) позициям ключевых точек тела. Одним из авторов были выделены все ключевые точки тела на всех изображениях для эксперимента. Ключевые точки были выделены в виде областей, описывающих суставы человека.

Таблица 4

Критерии оценки качества детектирования

Критерий оценки	Способ вычисления	Оценка критерия и соответствующий ей числовой интервал		
		Низкая	Средняя	Высокая
Количество найденных ключевых точек	Значение определяется отношением найденных ключевых точек к количеству искомых ключевых точек в процентах	> 40 %	> 70 %	> 90 %
Качество найденных ключевых точек	Значение определяется принадлежностью ключевой точки к одному из колец соответствующей области сустава	Внешнее кольцо	Среднее кольцо	Внутреннее кольцо

Выделенные области были разделены на три кольца одинаковой ширины – внутреннее (сердцевина области), среднее и внешнее кольца.

Алгоритм вычисления оценки качества детектирования идентичен оценке технических требований. Отличие между двумя оценками заключается в критериях оценки, способах их вычислений и числовых значениях.

Критерии оценки качества приведены в табл. 4.

Оценка качества детектирования определялась по формуле (2) и интерпретировалась по табл. 3, аналогично оценке технических требований. Время определялось на промежутке выполнения решением загрузки изображения для детектирования, детектирования по изображению и вывода результатов детектирования. Результаты эксперимента приведены в табл. 5–7.

Таблица 5

Сравнение решений по техническим требованиям (ключевые точки тела)

Решение	Технические требования
AlphaPose	Средние
MoveNet	Низкие
MediaPipe Pose	Низкие
OpenPose (CPU)	Высокие
OpenPose (GPU)	Высокие
DeeperCut	Высокие
YOLOv7	Средние

Поскольку решения MediaPipe (версия Holistic) и OpenPose (версии CPU, GPU) имеют возможность детектирования кистей рук, для них был проведен дополнительный эксперимент. В дополнительном эксперименте решения были настроены на детектирование ключевых точек тела и ключевых точек кистей рук одного человека в кадре. В качестве входных изображений были использованы изображения из ранее проведенного эксперимента для детектирования ключевых точек тела. Результаты эксперимента приведены в табл. 8–10.

Таблица 6

Сравнение решений по времени выполнения (ключевые точки тела)

Решение	Время выполнения для 1 изображения (секунды)	Время выполнения для группы из 10 изображений (секунды)
AlphaPose	2,27	15,11
MoveNet	0,32	1,89
MediaPipe Pose	0,19	2,11
OpenPose (CPU)	1,95	28,21
OpenPose (GPU)	3,12	12,33
DeeperCut	2,13	10,21
YOLOv7	2,49	15,37

Таблица 7

Сравнение решений по качеству детектирования (ключевые точки тела)

Решение	Качество детектирования
AlphaPose	Среднее
MoveNet	Низкое
MediaPipe Pose	Среднее
OpenPose (CPU)	Высокое
OpenPose (GPU)	Высокое
DeeperCut	Высокое
YOLOv7	Среднее

Таблица 8

Сравнение решений
по качеству детектирования
(ключевые точки тела и кистей рук)

Решение	Качество детектирования
MediaPipe Holistic	Среднее
OpenPose (CPU)	Высокое
OpenPose (GPU)	Высокое

Таблица 9

Сравнение решений
по времени выполнения
(ключевые точки тела и кистей рук)

Решение	Время выполнения для 1 изображения (секунды)	Время выполнения для группы из 10 изображений (секунды)
MediaPipe Holistic	0,20	2,23
OpenPose (CPU)	12,72	114,71
OpenPose (GPU)	4,49	12,15

Таблица 10

Сравнение решений
по техническим требованиям
(ключевые точки тела и кистей рук)

Решение	Технические требования
MediaPipe Holistic	Низкие
OpenPose (CPU)	Высокие
OpenPose (GPU)	Высокие

Анализ результатов эксперимента

Легковесные модели (требующие меньше вычислительной мощности устройства выполнения) и модели типа «сверху вниз» уменьшают зависимость от устройства выполнения при детектировании позы человека. Тяжеловесные модели обеспечивают высокое качество и точность детектирования. Модели типа «снизу вверх» лучше способны детектировать отдельные части тела человека.

Решение DeeperCut имеет ограничения на поддерживаемую операционную систему. Данные ограничения будут существенными для пользователей операционных систем семейства Windows. На март 2023 г. количество пользователей операционной системы семейства Windows составляет около 69% от всей массы пользователей, согласно источнику [13].

Решения AlphaPose, OpenPose (версии CPU, GPU), YOLOv7, выполняющие детектирование позы слишком продолжительное время (при сравнении с остальными рассмотренными решениями), не позволят добиться качественного детектирования изображения с видеопотока, к примеру, с веб-камеры, поскольку не смогут обеспечить большое количество обработанных кадров в отведенное время.

Решения MoveNet и MediaPipe (версии Pose, Holistic), при своем преимуществе во времени, затрачиваемом на детектирование, уступают в качестве детектирования тяжеловесным решениям.

При подборе детектора позы человека для конкретной задачи следует ориентироваться на те требования и ограничения, которые за ней стоят. Грамотный подбор детектора обеспечит успешное достижение поставленных целей.

Заключение

Авторами данной работы было проведено исследование существующих решений безмаркерного подхода для детектирования позы и частей тела человека с помощью нейронных сетей. Результатом эксперимента являются выявленные преимущества решений MediaPipe Holistic (легковесное решение) и OpenPose GPU (тяжеловесное решение) относительно остальных рассмотренных решений для задачи определения эмоционального состояния человека по его позе.

Данные решения были выбраны на основе их оценки качества детектирования, времени выполнения и технических требований. Также была учтена возможность дополнительного детектирования кистей рук. Ее использование может позволить оперировать большим количеством данных при оценке эмоционального состояния человека по его позе. Решение MediaPipe Holistic подходит для детектирования позы человека при условиях, когда необходимо обрабатывать поток изображений (видео, видео с веб-камеры) за ограниченное время. Решение OpenPose (версия GPU) подходит для задач, когда задержка по времени не является столь критичной, а качество детектирования выступает первостепенным ориентиром. Данное решение может быть использовано при анализе статичного изображения.

Список литературы

1. Лобанова Е.Н. Анализ невербальной информации как детерминанты управленческого поведения // Вестник Российского университета дружбы народов. Серия: Социология. 2013. № 1. С. 181–191.

2. Пырьев Е.А. Эмоциональные состояния, мотивирующие поведение человека // Известия Российского государственного педагогического университета им. А.И. Герцена. 2012. № 133. С. 288–294.

3. Sharifi A., Harati A., Vahedian A. Marker based human pose estimation using annealed particle swarm optimization with search space partitioning // 4th International Conference on Computer and Knowledge Engineering (ICCKE). 2014. С. 135–140.

4. Fang H.S., Li J., Tang H., Xu C., Zhu H., Xiu Y., Li Y.L., Lu C. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. С. 1–17.

5. Lugaresi C., Tang J., Nash H., McClanahan C., Uboweja E., Hays M., Zhang F., Chang C., Yong M., Lee J., Chang W., Hua W., Georg M., Grundmann M. Mediapipe: A framework for building perception pipelines // arXiv preprint arXiv:1906.08172. 2019.

6. MediaPipe Pose. MediaPipe. [Электронный ресурс]. URL: http://google.github.io/mediapipe/solutions/pose (дата обращения: 30.05.2023).

7. TensorFlow MoveNet: Ultra-fast and accurate pose detection model. TensorFlow. [Электронный ресурс]. URL: http://tensorflow.org/hub/tutorials/movenet (дата обращения: 30.05.2023).

8. Cao Z., Hidalgo G., Simon T., Wei S., Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. С. 7291–7299.

9. OpenPose: The first real-time multi-person system to jointly detect human body, hand, facial, and foot keypoints. OpenPose. [Электронный ресурс]. URL: http://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc (дата обращения: 30.05.2023).

10. Insafutdinov E., Pischulin L., Andres B., Andriluka M., Schiele B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model // Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016. Proceedings, Part VI 14. Springer International Publishing. 2016. С. 34–50.

11. Maji D., Nagori S., Mathew M., Poddar D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. С. 2637–2646.

12. Wang C.Y., Bochkovskiy A., Liao H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors // arXiv preprint arXiv:2207.02696. 2022.

13. Desktop Operating System Market. StatCounter. [Электронный ресурс]. URL: https://gs.statcounter.com/os-market-share/desktop/worldwide (дата обращения: 30.05.2023).