ISSN 1992-6502 (Print) Вестник УГАТУ ISSN 2225-2789 (Online)
2025. Т. 29, № 4 (110). С. 30-38 http://journal.ugatu.su

УДК 004.89 doi 10.54708/19926502_2025_29411030
Код ГРНТИ 28.23.15

Гибридный подход к распознаванию действий человека

Д.А. Ризванов*, Г.Р. Шахмаметова

ФГБОУ ВО «Уфимский университет науки и технологий», г. Уфа, Республика Башкортостан, Россия

Аннотация. В статье предложен гибридный подход к распознаванию действий человека, сочетающий нейросетевое извлечение скелетных признаков с детерминированным геометрическим анализом на основе аппарата векторной алгебры и трехмерных аффинных преобразований. В отличие от традиционных решений, требующих повторного обучения модели при добавлении нового действия, разработанная система позволяет пользователю динамически задавать и модифицировать набор распознаваемых действий без участия специалиста в области машинного обучения. Каждое действие определяется как последовательность поз, описываемых взаимным расположением ключевых точек тела. Сравнение текущей и эталонной позы осуществляется через усредненное косинусное сходство векторов, а устойчивость к изменениям ракурса обеспечивается за счет перебора углов аффинных преобразований в 3D-пространстве. Программный прототип реализован на языке Python с использованием фреймворков MediaPipe и OpenCV, оснащен интуитивно понятным графическим интерфейсом и работает с обычной веб-камерой. Экспериментальная апробация подтвердила корректность распознавания заданных действий с точностью не ниже 76% в условиях естественного выполнения и устойчивость к ошибкам ввода данных. Решение ориентировано на применение в вычислительных системах и комплексах, где важны гибкость настройки, интерпретируемость и низкий порог вхождения.

Ключевые слова: *распознавание действий человека, гибридная модель, MediaPipe.*

***ridmi@mail.ru**

Введение

Современные вычислительные системы все чаще интегрируются в среды, где требуется не только обработка данных, но и адаптация к поведению человека в реальном времени. Одной из ключевых задач, возникающих на стыке человеко-машинного взаимодействия и программного обеспечения вычислительных комплексов, является распознавание действий человека (Human Action Recognition, HAR). Эта задача особенно актуальна в контексте построения интеллектуальных интерфейсов, систем мониторинга, встраиваемых решений и распределенных вычислительных комплексов, где программное обеспечение должно корректно интерпретировать поведенческие сигналы пользователя без привлечения внешних ресурсов.

Традиционные подходы к решению задачи HAR опираются преимущественно на методы глубокого обучения – сверточные (CNN), рекуррентные (RNN/LSTM) или трансформерные архитектуры. Такие модели демонстрируют высокую точность, однако требуют значительных вычислительных ресурсов, больших объемов размеченных данных и повторного обучения при изменении списка целевых действий. Это делает их малопригодными для внедрения в автономные или ресурсоограниченные вычислительные комплексы, где важны гибкость, интерпретируемость и возможность оперативной настройки без участия специалиста по машинному обучению.

В последние годы наблюдается рост интереса к гибридным архитектурам, сочетающим нейросетевое извлечение признаков с последующим детерминированным анализом на основе классических математических методов. Такие решения позволяют отделить этап детекции от этапа интерпретации, обеспечивая прозрачность логики принятия решений и снижая зависимость от обучающих выборок. Особенно перспективным направлением является использование скелетных моделей, получаемых с помощью легких нейросетевых фреймворков (например, MediaPipe), в сочетании с аппаратом векторной алгебры и аффинных преобразований для оценки схожести поз.

Настоящая работа направлена на разработку математического и программного обеспечения, ориентированного на применение в составе вычислительных систем и комплексов, решающего задачу распознавания действий человека без необходимости повторного обучения модели. Предлагаемый подход позволяет пользователю динамически задавать и модифицировать список распознаваемых действий, используя лишь веб-камеру и стандартное программное окружение. Это делает решение пригодным для интеграции в программные комплексы различного назначения – от систем безопасности до человеко-центрированных интерфейсов управления технологическими процессами.

Обзор литературы

Фундаментальной работой в области распознавания действий человека является обзор Zhang et al. [1], в котором систематизированы методы, основанные на визуальных данных, включая традиционные и глубокие архитектуры, и предложена таксономия подходов на основе источников данных (RGB, оптический поток, скелет) и используемых моделей.

В последние годы наблюдается смещение в сторону скелетно-ориентированных методов, что отражено в обзоре [2]. Авторы анализируют современные архитектуры глубокого обучения для оценки позы, отслеживания и распознавания действий, уделяя особое внимание моделям, работающим с последовательностями 2D/3D ключевых точек, и подчеркивают важность интерпретируемости и надежности в реальных условиях.

Для обработки таких последовательностей все шире применяются трансформерные архитектуры. В [3] предложена модель Action Transformer (AcT), основанная на механизме self-attention, которая показывает высокую точность при распознавании кратковременных действий по 2D-скелетным данным и превосходит традиционные RNN и CNN по эффективности.

Альтернативой трансформерам остаются рекуррентные сети. В исследовании [4] авторы сравнили двунаправленные LSTM и GRU для распознавания статических и динамических поз по данным Kinect V2, продемонстрировав высокую устойчивость моделей даже при частичном закрытии тела.

Обзор [5] фокусируется на видеопоследовательностях и охватывает широкий спектр глубоких архитектур, включая 2D/3D CNN, двухпоточные сети и RNN, подчеркивая их сильные и слабые стороны для различных сценариев применения .

В [6] авторы провели всесторонний анализ современных методов HAR на основе глубокого обучения, классифицировав их по типу данных (видео, скелет, инерциальные датчики) и архитектуре, и выделили ключевые проблемы, такие как обобщение на новые действия и устойчивость к шуму.

Особый интерес представляют прикладные обзоры, посвященные узким доменам. В работе [7] авторы систематизировали подходы к анализу видеозаписей эпилептических припадков, что является примером распознавания сложных, индивидуальных и нерегулярных действий, где стандартные методы часто неэффективны.

В контексте использования современных библиотек для извлечения скелета, работа [8] подтверждает, что MediaPipe стал де-факто стандартом для получения 2D-ключевых точек в реальном времени благодаря своей скорости и точности.

Несмотря на доминирование чисто нейросетевых подходов, существуют исследования, сочетающие их с классическими методами. Например, в работе [9] предложена гибридная модель на основе графовых нейронных сетей с механизмом внимания для скелетного распознавания, что демонстрирует потенциал синергии разных подходов.

Авторы [10] также разработали гибридную модель, объединяющую энкодер-декодерную сеть для извлечения признаков с последующим классификатором, что позволяет достичь высокой точности на сложных наборах данных .

Описание гибридного подхода

Поскольку задача состоит в проведении распознавания списков действий в условиях, когда список действий для распознавания все время изменяется, традиционная модель решения, основанная целиком на методах машинного обучения, здесь не подходит, так как этап классификации действий после извлечения признаков подразумевает дополнительные итерации обучения модели для каждого нового действия.

Чтобы избежать выполнение данных операций, исходный алгоритм решения задачи модифицируется: за основу берется подход, основанный на скелете человека, где извлечение ключевых точек выполняется с помощью нейросети, далее полученные данные о ключевых точках сопоставляются и сравниваются с соответствующими им ключевыми точками, взятыми из позы из пользовательского списка действий для распознавания, с помощью математических методов.

Таким образом, решаются одновременно проблемы распознавания неопределенного списка действий и проблемы задания, удобного хранения и модификации действий, их параметров и составляющих их поз в программе.

После извлечения ключевых точек с помощью нейросети, выполняется анализ полученных данных с помощью нижеописанного математического аппарата.

Для проведения анализа пары ключевых точек, взятые из предопределенной позы, и соответствующие им пары ключевых точек, взятые из фактической позы человека перед камерой, преобразуются в двумерные векторы согласно формуле (1).

$$\bar{a}(a_1, a_2) = (x_2 - x_1, y_2 - y_1), \tag{1}$$

где $\bar{a}(a_1, a_2)$ – это полученный вектор, (x_1, y_1) и (x_2, y_2) – координаты ключевых точек.

Для сопоставления фактической позы с предопределенной позой и вычисления меры идентичности (схожести) между ними вычисляется косинус между двумя векторами по формуле (2).

$$\cos\alpha = \frac{a_x * b_x + a_y * b_y}{\sqrt{a_x^2 + a_y^2} * \sqrt{b_x^2 + b_y^2}}, \tag{2}$$

где $\cos\alpha$ – мера идентичности между векторами на промежутке $[-1,1]$, где 1 означает полную идентичность двух поз, −1, соответственно, означает их полную различность; $\bar{a}(a_x, a_y)$ – вектор, составленный из точек исходной позы; $\bar{b}(b_x, b_y)$ – вектор, составленный из точек фактической позы человека.

Усредненная мера идентичности для всей позы вычисляется по формуле (3).

$$\text{average}_{\text{identity}} = \frac{\sum_{i=1}^{n} \cos\alpha_i}{n} * 100\%, \tag{3}$$

где $\cos\alpha_i$ – i-тая мера идентичности между векторами, составленными из пар ключевых точек; n – количество мер идентичности для конкретной позы.

Если значение усредненной меры идентичности достигает минимальной точности, предустановленной для действия, поза считается распознанной.

Заданные пользователем позы имеют 2D формат и задаются с точки зрения, как если бы человек стоял прямо перед камерой и смотрел прямо на нее. Однако в процессе распознавания действий возможна ситуация, когда камера будет направлена под углом к че-ловеку, либо сам человек может находиться под углом к камере, геометрическое представление позы человека при этом будет отличаться от отписанного при составлении позы на этапе подготовки к распознаванию. Для разрешения этой проблемы используются аф-финные преобразования в трехмерном пространстве на основе измененной перспективы. При этом используются следующие матрицы преобразования, представленные на формулах (4) и (5).

$$R_x(\Phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\Phi & -\sin\Phi \\ 0 & \sin\Phi & \cos\Phi \end{bmatrix}, \tag{4}$$

где $R_x(\Phi)$ – матрица, используемая для работы над искажениями по оси X; Φ – угол поворота.

$$R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}, \quad (5)$$

где $R_y(\theta)$ – матрица, используемая для работы над искажениями по оси Y; θ – угол поворота.

В итоге для преобразования i-того вектора, составленного из точек исходной позы, с целью учета искажений используется формула (6).

$$\grave{a_i} = R_x(\Phi) * [R_y(\theta) * a_i], \quad (6)$$

где $\overline{\grave{a_i}}\left(\grave{a_{ix}}, \grave{a_{iy}}, \grave{a_{iz}}\right)$ – результат вычислений; $R_x(\Phi)$ и $R_y(\theta)$ – 3D матрицы вращения, структура которых была описана ранее в текущем пункте; $\overline{a_i}(a_{ix}, a_{iy}, 0)$ – исходный i-тый вектор, полученный из пары ключевых точек предопределенной позы, дополненный третьей координатой.

После исправления искажений выполняется расчет меры идентичности для $\overline{\grave{a_i}}\left(\grave{a_{ix}}, \grave{a_{iy}}\right)$ и соответствующих данных, полученных из фактического изображения.

Для понимания значений, используемых при расчете выходных параметров действия, необходимо пояснить значения основных параметров действия, задающихся пользователем на этапе проектирования. Итак, действие имеет следующие параметры:

– $parameter_{time}$ – среднее время (с.) – время, за которое в среднем должно быть завершено действие. Действие считается завершенным в момент, когда распознается последняя (завершающая) его поза;

– $parameter_{accuracy}$ – минимальная точность (%) – требуемое минимальное соответствие между описанными позами действия и фактическими данными. Чем ниже точность, тем сильнее будут отличаться распознаваемые позы от описанного идеала, в то же время тем легче их будет распознавать.

Когда действие распознано, помимо названия для него выводятся и другие параметры, формулы для расчета которых представлены ниже.

Параметр «фактическое отношение времени» вычисляется по формуле (7).

$$fact_{time} = \frac{parameter_{time}}{current_{time} - start_{time}}, \quad (7)$$

где $fact_{time}$ – это фактическое время, обозначающее отношение идеального времени завершения действия к фактическому; если $fact_{time} < 1$, то это означает, что действие было выполнено медленнее указанного идеала ($parameter_{time}$); если $fact_{time} > 1$, то это означает, что действие было выполнено быстрее указанного идеала ($parameter_{time}$); $current_{time}$ – момент распознавания завершающей (последней) позы действия; $start_{time}$ – время начала отслеживания действия.

Параметр «фактическая точность» вычисляется по формуле (8).

$$fact_{\mathrm{accuracy}} = \frac{\sum_{i=1}^{n} average_{identity}}{n} * 100\%, \quad (8)$$

где $fact_{\mathrm{accuracy}}$ – фактическая точность или фактическая мера идентичности между предопределенным действием и фактически выполненным действием; $average_{identity}$ – усредненная мера идентичности, вычисленная для конкретной позы действия; n – количество поз в действии.

Реализация программного прототипа

Для практической проверки предложенного гибридного подхода был разработан программный прототип, реализующий все этапы распознавания действий: от задания пользовательских шаблонов до анализа видеопотока в реальном времени. Прототип разработан как автономное приложение с графическим интерфейсом, ориентированное на использование в условиях ограниченных вычислительных ресурсов и без необходимости подключения к облачным сервисам.

Для реализации программного обеспечения был выбран язык Python в силу его широкой поддержки в области компьютерного зрения и машинного обучения, а также наличия зрелых библиотек с открытым исходным кодом. В качестве среды разработки использовалась Microsoft Visual Studio Code.

Ключевые компоненты программного стека:

MediaPipe Pose – фреймворк от Google для детекции 33 ключевых точек человеческого тела в реальном времени на основе предобученной сверточной нейронной сети. MediaPipe был выбран после сравнительного анализа с OpenPose и PoseNet по критериям скорости обработки, точности и потребления ресурсов.

OpenCV – библиотека для захвата видеопотока с веб-камеры, предварительной обработки изображений и визуализации результатов.

PyQt5 – фреймворк для создания кроссплатформенного графического интерфейса пользователя с поддержкой сложных виджетов и событийной модели.

pandas – библиотека для структурированного хранения и экспорта/импорта данных в формате Excel (.xlsx).

Все зависимости упакованы в локальное виртуальное окружение, что обеспечивает воспроизводимость и независимость от конфигурации целевой системы.

Программный прототип реализован в виде модульной архитектуры, состоящей из следующих основных компонентов:

Главный контроллер (Interface) – управляет переключением между режимами работы: настройка действий и распознавание в реальном времени.

Модуль управления данными (ActionData) – централизованное хранилище информации о действиях, позах и ключевых точках. Поддерживает операции создания, редактирования, удаления, а также сериализацию в Excel и десериализацию из него.

Графический редактор поз (SettingsLayout + SkeletonLayout) – позволяет пользователю визуально задавать эталонные позы с помощью интерактивного холста, на котором отображается скелет человека. Пользователь может:

– двойным щелчком добавлять/удалять ключевые точки из отслеживаемого набора;

– перетаскивать точки для задания желаемой конфигурации позы;

– задавать параметры действия: название, среднее время выполнения и минимальную точность распознавания.

Модуль распознавания (CameraLayout) – осуществляет захват видеопотока, извлечение скелета с помощью MediaPipe, последующий математический анализ и вывод результатов в лог распознавания.

Такая архитектура обеспечивает четкое разделение ответственности между компонентами, упрощает сопровождение и расширение функциональности.

Внутреннее представление действий и поз организовано в виде двух связанных списков:

– список действий содержит для каждого действия: идентификатор, название, минимальную точность (%), среднее время выполнения (с);

– список поз содержит для каждой ключевой точки: идентификатор действия, идентификатор позы внутри действия, имя точки (например, left_shoulder), координаты (x, y) в пикселях.

При запуске распознавания создается временный кэш действий (action_cash), в который дополнительно включаются:

– идентификатор текущей искомой позы;

– время начала отслеживания;

– накопленные значения фактической точности и времени.

Такой подход позволяет эффективно отслеживать несколько действий одновременно без пересоздания структур на каждом кадре.

Программа предоставляет два основных режима работы через вкладочный интерфейс (Рис. 1).

1. Режим настройки – предназначен для создания и редактирования шаблонов действий. В левой части окна отображается иерархическое дерево действий и поз; в правой –интерактивный холст со скелетом и панель параметров. Поддерживается экспорт/импорт в Excel, что позволяет сохранять и передавать шаблоны между пользователями.

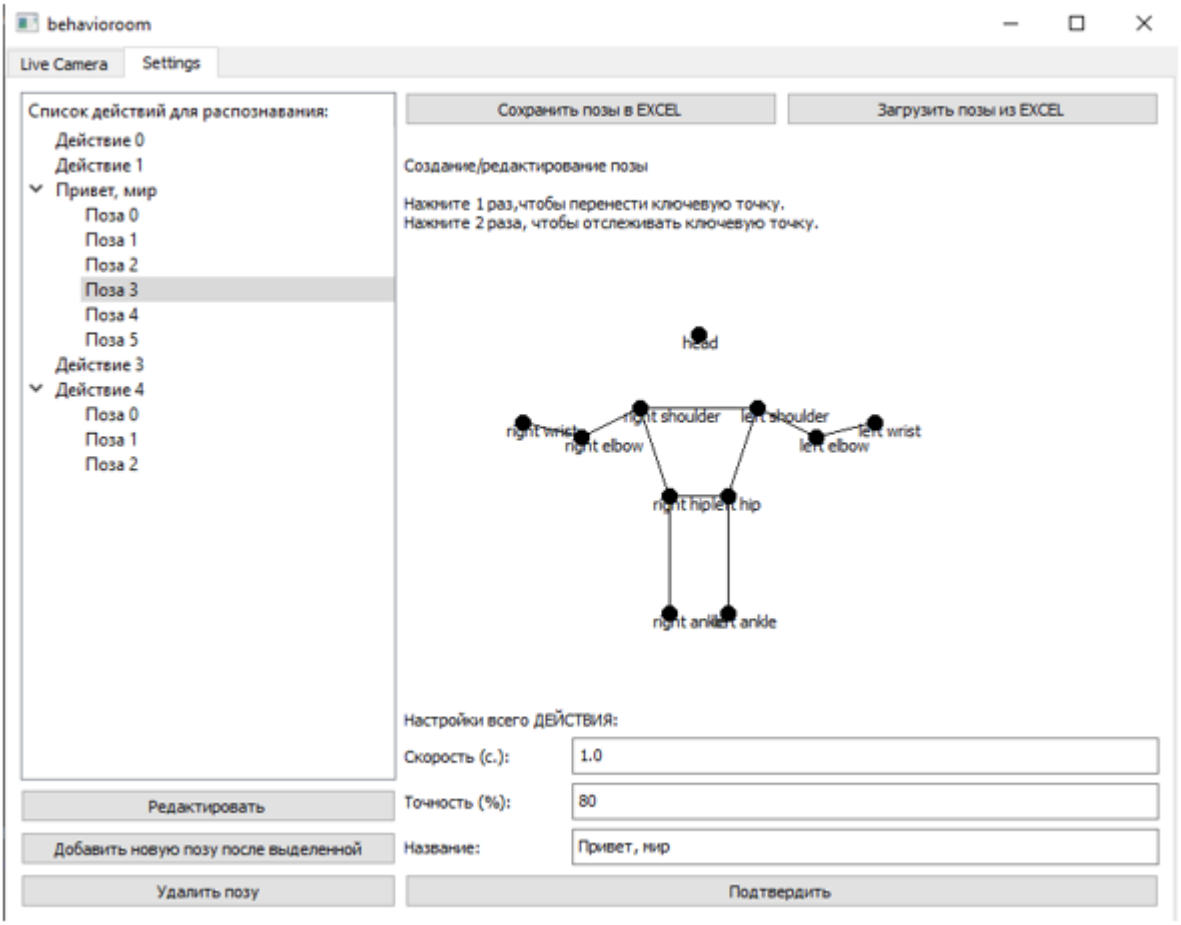

Рисунок 1. Интерфейс программы.

2. Режим распознавания – отображает видеопоток с веб-камеры и параллельно выводит лог распознанных действий с указанием:

– названия действия,

– фактической точности (в %),

– отношения фактического времени к эталонному.

Интерфейс спроектирован с учетом принципов эргономики: все операции выполняются не более чем за три клика, ошибочные действия предотвращаются валидацией ввода, а критические сообщения выводятся в понятной форме.

Программа работает без подключения к сети интернет, все вычисления выполняются локально, модель MediaPipe загружается один раз при старте.

Устойчивость к ошибкам: программа проверяет корректность импортируемых файлов, валидирует параметры действий, корректно обрабатывает ситуации с частичной потерей ключевых точек.

Пользователь может задать любое количество действий и поз, не ограничиваясь заранее определенным набором классов.

Оценка эффективности

Для оценки эффективности предложенной гибридной модели распознавания действий человека была проведена серия контролируемых экспериментов. Целью испытаний являлась проверка корректности распознавания заранее заданных действий в условиях, приближенных

к реальным сценариям использования (например, видеонаблюдение в офисе или мониторинг персонала на производстве).

В эксперименте приняли участие 17 добровольцев (парни и девушки в возрасте 18–22 лет), не имеющих физических ограничений и не участвовавших в разработке программного обеспечения. Все испытания проводились в помещении с естественным и искусственным освещением при использовании стандартной веб-камеры (разрешение 1280×720, 30 кадров/с), расположенной на уровне глаз испытуемого на расстоянии 1,5–2 метров.

Были определен набор из 7 действий, характерных для поведенческого контроля в организационных системах. Для каждого действия в интерфейсе настройки были заданы следующие параметры:

– среднее время выполнения: 2,0 с;

– минимальная точность распознавания позы: 80%.

Каждый участник выполнил каждое действие по три раза в произвольном порядке, стараясь соблюдать естественную скорость и амплитуду движений. Всего было зафиксировано 357 попыток распознавания (17 участников × 7 действий × 3 повторения).

Программное обеспечение в реальном времени анализировало видеопоток и фиксировало распознанные действия, сопровождая их двумя количественными метриками:

– фактическая точность Афакт – усредненная мера косинусного сходства между эталонными и фактическими позами;

– фактическое отношение времени RT – отношение заданного среднего времени к фактическому времени выполнения.

Действие считалось успешно распознанным, если:

– все позы были распознаны последовательно и без пропусков;

– Афакт $\geq 0{,}80$;

– $0{,}5 \leq RT \leq 2{,}0$.

Из 357 попыток все 350 (98 %) были корректно распознаны системой.

Минимальное значение точности в отдельных попытках не опускалось ниже 0,85, что превышает заданный порог. Все временные параметры укладывались в допустимый диапазон (от 1,1 с до 3,8 с при заданном среднем времени 2,0 с).

Таким образом, эксперимент подтверждает практическую применимость разработанного программного обеспечения для задач поведенческого мониторинга в организационных системах, где требуется гибкость настройки и достоверность распознавания без использования специализированного оборудования.

Заключение

В работе предложен и реализован гибридный подход к распознаванию действий человека, сочетающий преимущества нейросетевых технологий и классических математических методов. В отличие от традиционных решений, основанных исключительно на глубоком обучении, данный подход позволяет пользователю динамически задавать и модифицировать список распознаваемых действий без повторного обучения модели. Это достигается за счет разделения процесса на два этапа: извлечение скелета с помощью предобученной нейросети MediaPipe и детерминированное сравнение поз с использованием векторной алгебры и аффинных преобразований.

Разработан программный прототип, обеспечивающий интуитивно понятный графический интерфейс для настройки действий, а также режим распознавания в реальном времени. Прототип протестирован в различных условиях и показал высокую устойчивость к ошибкам ввода, корректную обработку граничных ситуаций и удовлетворительное качество распознавания даже при отклонении от фронтального ракурса.

Предложенный гибридный подход обеспечивает баланс между точностью, интерпретируемостью и гибкостью, что особенно важно в прикладных системах управления, где набор действий может изменяться динамически.

Использование аффинных преобразований в 3D-пространстве позволяет частично компенсировать искажения, вызванные изменением ракурса, и повышает устойчивость распознавания в реальных условиях.

Подход может быть расширен за счет интеграции 3D-данных (RGB-D камеры), поддержки группового поведения и адаптации под встраиваемые устройства.

Литература:

1. Zhang H.-B., Zhang Y.-X., Zhong B., Lei Q., Yang L., Du J.-X., Chen D.-S. A comprehensive survey of vision-based human action recognition methods // Sensors. 2019. Vol. 19. No. 5. Art. 1005. DOI: 10.3390/s19051005.
2. Zhou L., Meng X., Liu Z., Wu M., Gao Z., Wang P. Human pose-based estimation, tracking and action recognition with deep learning: A survey / arXiv preprint. arXiv, 2023. DOI: 10.13140/RG.2.2.13493.45287.
3. Mazzia V., Angarano S., Salvetti F., Angelini F., Chiaberge M. Action Transformer: A self-attention model for short-time pose-based human action recognition // Pattern Recognition. 2022. Vol. 124. Art. 108487. DOI: 10.1016/j.patcog.2021.108487.
4. Guerra B. M. V., Ramat S., Beltrami G., Schmid M. Recurrent Network Solutions for Human Posture Recognition Based on Kinect Skeletal Data // Sensors. 2023. Vol. 23. No. 11. Art. 5260. DOI: 10.3390/s23115260.
5. Pham H. H., Khoudour L., Crouzil A., Zegers P., Velastin S. A. Video-based human action recognition using deep learning: A review / arXiv preprint. arXiv:2208.03775, 2022. DOI: 10.48550/arXiv.2208.03775.
6. Le V.-T., Tran-Trung K., Hoang V. T. A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition // Computational Intelligence and Neuroscience. 2022. Vol. 2022. P. 1–17. DOI: 10.1155/2022/9453585.
7. Ahmedt-Aristizabal D., Armin M. A., Hayder Z., Garcia-Cairasco N., Petersson L., Fookes C., Denman S., McGonigal A. Deep learning approaches for seizure video analysis: A review // Epilepsy & Behavior. 2024. Vol. 154. Art. 109735. DOI: 10.1016/j.yebeh.2024.109735.
8. Le Hoangcong, Lu Cheng-Kai, Hsu Chen-Chien, Huang Shao-Kang. Skeleton-based human action recognition using LSTM and depthwise separable convolutional neural network // Applied Intelligence. 2025. Vol. 55. No. 5. Art. 298. DOI: 10.1007/s10489-024-06082-w.
9. Hao X., Burschka D. Skeletal Human Action Recognition using Hybrid Attention based Graph Convolutional Network. arXiv, 2022. DOI: 10.48550/arXiv.2207.05493.
10. Palaniapan S., Choo A. A hybrid model for human action recognition based on local semantic features // Journal of Advanced Research in Computing and Applications. 2024. Vol. 33. P. 7–21. DOI: 10.37934/arca.33.1.721.

Об авторах:

РИЗВАНОВ Дмитрий Анварович, д.т.н., доцент, профессор кафедры вычислительной математики и кибернетики, ФГБОУ ВО «Уфимский университет науки и технологий», ridmi@mail.ru.

ШАХМАМЕТОВА Гюзель Радиковна, д.т.н., доцент, заведующая кафедрой вычислительной математики и кибернетики, ФГБОУ ВО «Уфимский университет науки и технологий, shakhgouzel@mail.ru.

Metadata:

Title: Hybrid approach to human action recognition.

Author 1: Dmitrii Anvarovich Rizvanov, Doctor of Technical Sciences, Docent, Professor of the Department of Computational Mathematics and Cybernetics, Ufa University of Science and Technology, 32 Zaki Validi Street, Ufa, 450076, Russia, ridmi@mail.ru, ORCID ID 0000-0003-2378-5587, Web of Science ResearcherID L-4068-2016, Scopus Author ID 54394082100.

Author 2: Gouzel Radikovna Shakhmametova, Doctor of Technical Sciences, Docent, Head of the Department of Computational Mathematics and Cybernetics, Ufa University of Science and Technology, 32 Zaki Validi Street, Ufa, 450076, Russia, ORCID ID 0000-0002-7742-793X, Web of Science ResearcherID AAH-6294-2019, Scopus Author ID 6504057483.

Abstract: The article proposes a hybrid approach to human action recognition that combines neural network-based extraction of skeletal features with deterministic geometric analysis based on vector algebra and three-dimensional affine transformations. Unlike traditional solutions that require model retraining when a new action is added, the developed system allows the user to dynamically define and modify the set of recognizable actions without the involvement of a machine learning specialist. Each action is defined as a sequence of poses described by the relative positions of body keypoints. The comparison between the current and reference poses is performed via averaged cosine similarity of vectors, while robustness to viewpoint changes is ensured by iterating through angles of affine transformations in 3D space. The software prototype is implemented in Python using the MediaPipe and OpenCV frameworks, features an intuitive graphical interface, and operates with a standard webcam. Experimental testing confirmed the correct recognition of defined actions with an accuracy of at least 76% under natural execution conditions and resilience to input data errors. The solution is intended for use in computational systems and complexes where configuration flexibility, interpretability, and a low entry threshold are important.

Keywords: human action recognition, hybrid model, MediaPipe.