Инженерный вестник Дона, №12 (2025)
ivdon.ru/ru/magazine/archive/n12y2025/10605

Распознавание действий человека с использованием нейросетевых технологий

Д.А. Ризванов

Уфимский университет науки и технологий

Аннотация: В статье представлен гибридный подход к распознаванию действий человека, сочетающий нейросетевое извлечение скелетных признаков с детерминированным геометрическим анализом на основе векторной алгебры и аффинных преобразований. Проведен обзор исследований по данной проблематике. В отличие от традиционных решений, требующих повторного обучения модели при добавлении нового действия, предложенная 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 поз осуществляется через косинусное сходство векторов, а устойчивость к изменениям ракурса обеспечивается за счёт трёхмерных аффинных преобразований. Программное обеспечение реализовано на языке Python с использованием фреймворков MediaPipe и OpenCV, имеет интуитивно понятный графический интерфейс и работает с обычной веб-камерой. Экспериментальная апробация подтвердила корректность распознавания заданных действий с точностью не ниже 85 % в условиях естественного выполнения. Решение ориентировано на применение в системах управления поведением в организационных средах, где важны гибкость настройки, интерпретируемость и низкий порог вхождения.

Ключевые слова: распознавание действий человека, векторная алгебра, аффинные преобразования, гибридная модель, управление поведением, человеко-машинные интерфейсы.

Введение

Современные организационные системы всё чаще сталкиваются с необходимостью автоматизированного мониторинга поведения персонала и посетителей. Это особенно актуально в таких сферах, как безопасность на промышленных объектах, контроль за соблюдением технологических регламентов, медицинский надзор за пациентами и управление взаимодействием в «умных» офисах. Одним из ключевых инструментов для решения этих задач является распознавание действий человека (Human Action Recognition, HAR).

Традиционные методы HAR, основанные на глубоком обучении, требуют полного переобучения модели при каждом изменении списка

распознаваемых действий. Это создаёт высокий порог вхождения для конечных пользователей и делает такие системы негибкими в условиях динамически меняющихся требований. В работе предлагается альтернативный подход, сочетающий нейросетевые технологии для извлечения признаков и классические математические методы.

Целью исследования является разработка математического и программного обеспечения, позволяющего эффективно распознавать произвольный, заранее неопределённый список действий человека в реальном времени без необходимости повторного обучения модели.

Обзор современных подходов к распознаванию действий

Распознавание действий человека остаётся одной из ключевых задач компьютерного зрения, находя применение в системах видеонаблюдения, медицинского мониторинга, человеко-машинного взаимодействия и управления поведением в организационных средах. За последние годы в этой области произошёл значительный сдвиг от ручного проектирования признаков к методам глубокого обучения, что позволило существенно повысить точность и устойчивость систем.

Систематический анализ современных архитектур глубокого обучения для HAR представлен в работе Le et al. [1]. Авторы показывают, что комбинации 3D-сверточных нейронных сетей, двухпоточных сетей и рекуррентных архитектур демонстрируют высокую эффективность на крупных датасетах, однако требуют значительных вычислительных ресурсов и больших объёмов размеченных данных, что ограничивает их применение в условиях ограниченных ресурсов.

В обзоре Zhang et al. подробно рассмотрены как методы, основанные на RGB-видео, так и подходы, использующие данные с датчиков глубины [2]. Особое внимание уделено скелетным методам, которые обеспечивают устойчивость к изменениям освещения и фону, что делает их особенно

перспективными для приложений в «умных» офисах и системах безопасности.

Значительный вклад в развитие скелетных методов внесла работа Yan et al., в которой предложена модель, которая естественным образом представляет скелет человека как динамический граф, где суставы — узлы, а кости — рёбра, что позволяет эффективно моделировать пространственно-временные зависимости между частями тела [3].

Ранним примером гибридного подхода является работа Baccouche et al., где последовательности кадров обрабатываются свёрточной сетью, а временная динамика моделируется с помощью рекуррентных слоёв [4]. Такая архитектура показала высокую эффективность при распознавании коротких действий в реальном времени и легла в основу многих последующих исследований.

С появлением трансформеров в задачах компьютерного зрения возник интерес к их применению в HAR. В работе Mazzia et al. предложена модель, которая обрабатывает последовательности скелетных поз напрямую, без использования рекуррентных слоёв, и демонстрирует конкурентоспособные результаты на датасетах [5].

Особое внимание в последние годы уделяется мультимодальным и гибридным системам. Pham et al. в своём обзоре 2022 года анализируют методы, объединяющие визуальные, аудио- и инерциальные данные для повышения надёжности распознавания в сложных условиях [6]. Авторы подчёркивают, что такие системы особенно актуальны для организационных сред с высокими требованиями к безопасности и устойчивости.

Важным направлением является снижение зависимости от размеченных данных. Работа Vahdani et al. посвящена работе с моделями, когда у вас мало (few-shot) или совсем нет (zero-shot) данных для обучения в HAR, что позволяет распознавать новые действия без полного переобучения

модели — подход, напрямую релевантный задачам управления в динамически меняющихся организационных системах, где требования к поведенческому контролю могут меняться оперативно [7].

Отдельно следует отметить исследования, посвящённые практической реализации HAR-систем на основе датчиков глубины. Guerra et al. предложили рекуррентную архитектуру для распознавания поз, достигнув высокой точности (до 97%) при использовании стандартного оборудования [8]. Однако такие решения требуют наличия специализированного аппаратного обеспечения, что ограничивает их массовое внедрение.

В медицинской сфере HAR применяется для мониторинга критических состояний. Ahmedt-Aristizabal et al. в своём обзоре 2024 года обобщили достижения глубокого обучения в анализе видеоэпилептических припадков, подчеркнув важность интерпретируемости и надёжности моделей в критически важных приложениях [9].

Наконец, в недавнем обзоре Zhou et al. обобщены современные методы, основанные на оценке позы человека с помощью глубоких нейросетей [10]. Авторы отмечают, что фреймворки, такие как MediaPipe и OpenPose, сделали скелетные методы доступными даже для систем с ограниченными ресурсами, что открывает возможности для их внедрения в организационные системы без необходимости в дорогостоящем оборудовании.

Несмотря на значительный прогресс, большинство существующих решений требуют повторного обучения модели при добавлении нового действия, что создаёт высокий порог вхождения для конечных пользователей. Как отмечают авторы работ [1, 6, 7], даже при использовании трансферного обучения или few-shot подходов настройка системы под новые задачи остаётся трудоёмкой и требует экспертизы в области машинного обучения.

Таким образом, в литературе отсутствуют решения, сочетающие гибкость настройки, низкий порог вхождения, работу с обычной веб-камерой и отсутствие необходимости в переобучении. Заполнение этого пробела и является целью настоящего исследования.

Разработка гибридной модели распознавания действий

В основе предложенного подхода лежит идея разделения процесса на два этапа: извлечение признаков и классификация. На первом этапе используется предобученная нейросеть MediaPipe Pose, которая в реальном времени определяет 33 ключевые точки скелета человека на видеокадре. Этот этап не требует дообучения и обеспечивает стабильное качество извлечения признаков.

На втором этапе происходит сравнение текущей позы с эталонной. Пользователь в графическом интерфейсе задаёт эталонную позу, выбирая ключевые точки и фиксируя их взаимное расположение. Для каждой пары точек строится двумерный вектор. Сравнение осуществляется путём вычисления косинусного сходства между векторами эталонной и текущей поз. Усреднённое значение по всем парам даёт меру сходства позы.

Для компенсации искажений, возникающих при изменении угла обзора камеры, применяются аффинные преобразования в трёхмерном пространстве с последующей проекцией на 2D-плоскость. Это значительно повышает устойчивость системы к поворотам тела.

Действие в модели определяется как последовательность поз, каждая из которых должна быть распознана с точностью не ниже заданного порога и в течение временного окна, составляющего от 0,5 до 2,0 от среднего времени выполнения действия. Такой подход позволяет распознавать сложные, многофазные действия, такие как «отдать честь» или «приветствие».

Главное преимущество модели — отсутствие необходимости в обучении. Пользователь может в любой момент добавить, изменить или удалить действие, не обладая знаниями в области машинного обучения.

В ходе выполнения работы была разработана гибридная модель, сочетающая в себе преимущества нейросетевых технологий и классических методов математического анализа. Основная цель такого подхода – обеспечить возможность распознавания произвольного, заранее не фиксированного набора действий без необходимости повторного обучения модели. Это особенно важно в условиях организационных систем, где требования к поведенческому контролю могут меняться динамически и непредсказуемо.

Общая архитектура модели

Модель состоит из трёх последовательно связанных компонентов:

1. Модуль детекции скелета – отвечает за извлечение ключевых точек тела человека из видеопотока.
2. Модуль сопоставления поз – выполняет сравнение текущей позы с эталонной на основе геометрических признаков.
3. Модуль распознавания действий – отслеживает последовательность распознанных поз во времени и принимает решение о завершении действия.

Такая декомпозиция позволяет отделить этап извлечения признаков (реализуемый нейросетью) от этапа классификации (реализуемого детерминированным алгоритмом), что обеспечивает прозрачность принятия решений и гибкость настройки.

Модуль детекции скелета

Для извлечения ключевых точек скелета используется фреймворк MediaPipe Pose, разработанный компанией Google. Этот инструмент был

выбран по результатам сравнительного анализа с такими альтернативами, как OpenPose и PoseNet, на основании следующих критериев:

- высокая скорость обработки кадров (до 30 FPS на центральном процессоре);
- приемлемая точность определения ключевых точек при работе с обычной веб-камерой;
- минимальные требования к вычислительным ресурсам;
- простота интеграции в Python-среду.

MediaPipe Pose возвращает координаты 33 ключевых точек тела человека в двумерном пространстве. Эти точки включают суставы конечностей, ключевые точки туловища и головы. Полученные координаты передаются на следующий этап обработки без дополнительной фильтрации или постобработки.

Модуль сопоставления поз

Этот модуль реализует математический аппарат, позволяющий количественно оценить степень схожести текущей позы с эталонной. Пользователь в графическом интерфейсе задаёт эталонную позу, выбирая подмножество ключевых точек и фиксируя их взаимное расположение. Для каждой пары выбранных точек строится двумерный вектор:

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

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

Для сопоставления эталонной и текущей поз и вычисления меры идентичности (схожести) между ними вычисляется косинус между соответствующими векторами

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

где $\cos\alpha$ – мера идентичности между векторами на промежутке $[-1,1]$: 1 – полная идентичность двух поз, -1 – полная 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.

Усреднённая мера сходства для всей позы определяется как:

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

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

Поза считается распознанной, если $\text{average}_{\text{identity}} \geq \gamma$, где γ – порог точности, задаваемый пользователем для конкретного действия (в диапазоне от 50% до 99%).

Компенсация искажений при изменении ракурса

Заданные пользователем позы имеют 2D формат и задаются с точки зрения, как если бы человек стоял прямо перед камерой и смотрел прямо на нее. Одной из ключевых проблем при распознавании поз является изменение их визуального представления при повороте тела относительно камеры. Чтобы минимизировать влияние этого фактора, в модели применяются аффинные преобразования в трёхмерном пространстве.

При этом используются следующие матрицы преобразования, представленные на формулах.

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

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

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

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

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

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

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

Результирующий вектор проецируется обратно на 2D-плоскость путём отбрасывания координаты z. Это позволяет значительно повысить устойчивость распознавания при наклонах и поворотах тела.

Модуль распознавания действий

Действие в модели определяется как упорядоченная последовательность поз, каждая из которых должна быть распознана с точностью не ниже заданного порога. Кроме того, действие должно быть выполнено в допустимом временном интервале:

$$0{,}5 \cdot T_{ср} \leq T_{факт} \leq 2{,}0 \cdot T_{ср},$$

где $T_{ср}$ – среднее время выполнения действия, задаваемое пользователем, а $T_{факт}$ – фактическое время между началом отслеживания и распознаванием последней позы.

Алгоритм отслеживания действия работает следующим образом:

1. При старте распознавания для каждого действия инициализируется внутренний счётчик текущей позы (начинается с первой).
2. На каждом кадре проверяется, соответствует ли текущая поза ожидаемой (с учётом точности и аффинных преобразований).

3. Если поза распознана, счётчик увеличивается; если достигнута последняя поза – проверяется временное условие.

4. При успешном выполнении всех условий действие фиксируется как распознанное, и выводится в лог с указанием:

- названия действия;
- фактической точности.

Такой подход позволяет распознавать сложные многофазные действия без необходимости в обучении на видеозаписях этих действий.

Преимущества гибридного подхода

Предложенная модель обладает рядом существенных преимуществ по сравнению с традиционными нейросетевыми решениями сквозного тестирования:

- Отсутствие необходимости в обучении: пользователь может добавлять, редактировать или удалять действия в любое время без участия специалиста по машинному обучению.
- Интерпретируемость: каждое решение сопровождается количественными метриками (точность, время), что повышает доверие к системе.
- Низкий порог вхождения: настройка действий осуществляется через интуитивно понятный графический интерфейс с визуальным редактором скелета.
- Аппаратная независимость: система работает с обычной веб-камерой и не требует специализированного оборудования (например, датчиков глубины).

Эти свойства делают разработанную модель особенно пригодной для внедрения в организационные системы управления, где важны гибкость, надёжность и простота эксплуатации.

Реализация программного обеспечения

Программное обеспечение разработано на языке Python с использованием библиотек OpenCV (для работы с видеопотоком), MediaPipe (для детекции скелета) и PyQt5 (для построения графического интерфейса). Все данные о действиях и позах хранятся во внутреннем контейнере ActionData и могут быть экспортированы в/импортированы из файла формата Excel (poses.xlsx).

Интерфейс программы состоит из двух вкладок:

1. Настройка действий: позволяет создавать и редактировать действия и позы с помощью интерактивного холста со скелетом. Пользователь может перетаскивать ключевые точки и включать/выключать их отслеживание двойным кликом.

2. Распознавание: отображает видеопоток с наложенным скелетом и выводит лог распознанных действий с указанием их фактической точности и времени выполнения.

Программа включает в себя систему валидации входных данных и обработки ошибок, что обеспечивает её стабильную работу даже в экстремальных условиях (например, при пустом списке действий или отсутствии файла импорта).

Апробация и оценка качества

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, приближённых 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 мв.

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 $A_{\text{факт}}$ – усреднённая мера косинусного сходства между эталонными и фактическими позами;
- фактическое отношение времени R_T – отношение заданного среднего времени к фактическому времени выполнения.

Действие считалось успешно распознанным, если:

- все позы были распознаны последовательно и без пропусков;
- $A_{\text{факт}} \geq 0{,}80$;
- $0{,}5 \leq R_T \leq 2{,}0$.

Из 357 попыток все 357 (100 %) 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).

Полученные результаты свидетельствуют о высокой надёжности и устойчивости предложенной модели в условиях естественного выполнения действий. Отсутствие ложных срабатываний и пропусков подтверждает корректность алгоритма сопоставления поз с учётом аффинных преобразований и временного окна. Высокие значения фактической точности (в среднем > 0,89) указывают на достаточную чувствительность геометрического аппарата к индивидуальным особенностям исполнения движений.

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

Предложенная гибридная модель распознавания действий человека решает ключевую проблему современных систем — их негибкость. Разработанное программное обеспечение позволяет конечному пользователю без специальных знаний в области ИИ настраивать и модифицировать поведенческие шаблоны в реальном времени.

Это делает решение особенно ценным для применения в системах управления организационными процессами, где требования к контролю поведения могут меняться оперативно. В перспективе планируется расширение функционала за счёт использования 3D-данных, внедрения механизмов адаптации к индивидуальным особенностям пользователя и оптимизации производительности для работы на мобильных платформах.

Литература

1. 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. URL: doi.org/10.1155/2022/8323962.

2. 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. V. 19. №5. p. 1005.

3. Yan, Sijie & Xiong, Yuanjun & Lin, Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition // Proceedings of the AAAI conference on artificial intelligence. 2018. V. 32. №1. URL: doi.org/10.1609/aaai.v32i1.12328.

4. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A. Sequential deep learning for human action recognition // International workshop on human behavior understanding. Berlin, Heidelberg, 2011. pp. 29-39.

5. 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. V. 124. URL: doi.org/10.1016/j.patcog.2021.108487.

6. Либерман А. И. Решение задачи классификации объектов на изображении методами компьютерного зрения в различных сферах человеческой деятельности // Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2024. №. 3. С. 74-91.

7. Yu X., Fang Y., Liu Z., Wu Y., Wen Z., Bo J., Hoi S. C. Few-shot learning on graphs: from meta-learning to pre-training and prompting. 2024. URL: arxiv.org/html/2402.01440v1.

8. Guerra B., Ramat S., Beltrami G., Schmid M. Recurrent network solutions for human posture recognition based on Kinect skeletal data //Sensors. 2023. V. 23. №11. URL: mdpi.com/1424-8220/23/11/5260.

9. Ahmedt-Aristizabal D., Armin 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. V.154. URL: doi.org/10.1016/j.yebeh.2024.109735.

10. Zhou L., Meng X., Liu Zh.., Wu M., Gao Zh., Wang P. Human Pose-based Estimation, Tracking and Action Recognition with Deep Learning: A Survey. 2023. URL: arxiv.org/abs/2310.13039.

References

1. Le V.-T., Tran-Trung K., Hoang V. T. Computational Intelligence and Neuroscience. 2022. URL: doi.org/10.1155/2022/8323962.

2. Zhang, H.-B., Zhang, Y.-X., Zhong, B., Lei, Q., Yang, L., Du, J.-X., & Chen, D.-S. Sensors. 2019. V. 19. №5. pp. 1005.

3. Yan, Sijie & Xiong, Yuanjun & Lin, Dahua. Proceedings of the AAAI conference on artificial intelligence. 2018. V. 32. №1. URL: doi.org/10.1609/aaai.v32i1.12328.

4. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A. International workshop on human behavior understanding. Berlin, Heidelberg, 2011. pp. 29-39.

5. Mazzia V., Angarano S., Salvetti F., Angelini F., Chiaberge M. Pattern Recognition. 2022. V. 124. URL: doi.org/10.1016/j.patcog.2021.108487.

6. Liberman A. I. Vestnik VGU. Seriya: Sistemnyj analiz i informacionnye tekhnologii. 2024. №3. pp. 74-91.

7. Yu X., Fang Y., Liu Z., Wu Y., Wen Z., Bo J., Hoi S. C. Few-shot learning on graphs: from meta-learning to pre-training and prompting. 2024. URL: arxiv.org/html/2402.01440v1.

8. Guerra B., Ramat S., Beltrami G., Schmid M. Sensors. 2023. V. 23. №11. URL: mdpi.com/1424-8220/23/11/5260.

9. Ahmedt-Aristizabal D., Armin A., Hayder Z., Garcia-Cairasco N., Petersson L., Fookes C., Denman S., McGonigal A. Epilepsy & Behavior. 2024. V.154. URL: doi.org/10.1016/j.yebeh.2024.109735.

10. Zhou L., Meng X., Liu Zh.., Wu M., Gao Zh., Wang P. Human Pose-based Estimation, Tracking and Action Recognition with Deep Learning: A Survey. 2023. URL: arxiv.org/abs/2310.13039.

Дата поступления: 4.10.2025

Дата публикации: 27.11.2025