Исследования и разработки в области новых информационных технологий и их приложений

https://doi.org/10.25559/SITITO.021.202502.241-250
УДК 519.872

Распознавание сложно-составных действий человека на основе анализа последовательности скелетных поз

Оригинальная статья

К. М. Максименко[1], Л. Н. Теряев[1*], В. А. Дорохин[1], А. В. Нечаевский[1,2]

[1] ГБОУ ВО Московской области «Университет «Дубна», г. Дубна, Российская Федерация

Адрес: 141982, Российская Федерация, Московская область, г. Дубна, ул. Университетская, д. 19

[2] Международная межправительственная организация Объединенный институт ядерных исследований, г. Дубна, Российская Федерация

Адрес: 141980, Российская Федерация, Московская область, г. Дубна, ул. Жолио-Кюри, д. 6

[*] u1seanon@yandex.ru

Аннотация

Одним из приоритетных направлений развития технологии компьютерного зрения является выделение скелетных данных из изображений людей и последующее использование этих данных для решения целого спектра прикладных задач. В статье дается краткий обзор технологий для решения задачи распознавания действий человека, выделяются основные подходы, описываются ограничения, преимущества и недостатки. Авторами предложен новый подход к распознаванию сложносоставных действий человека на основе анализа динамики скелетных данных и применения машины состояний. Используемый подход является многоступенчатым и сочетает в себе последовательное использование нейросетевой модели определения позы человека *MoveNet*, пользовательского слоя извлечения расширенных признаков (*PoseEnhancementLayer*), а также алгоритм выявления совершаемого действия на основе анализа поз в бифуркационных точках действия. Предложенное авторами решение позволяет определять действия без дополнительного обучения модели, что обеспечивает гибкость и масштабируемость. Тестирование на открытых датасетах показало высокую точность классификации поз человека и устойчивость к неполным или зашумленным последовательностям. Результаты работы актуальны для задач в области спортивной аналитики, интерактивного обучения, реабилитации и медицинского мониторинга.

Ключевые слова: распознавание поз, скелет, human action classification, компьютерное зрение, полносвязные слои, машина состояний, нечеткая логика, конструирование паттернов классификации, XR

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Максименко К. М., Теряев Л. Н., Дорохин В. А., Нечаевский А. В. Распознавание сложно-составных действий человека на основе анализа последовательности скелетных поз // Современные информационные технологии и ИТ-образование. 2025. Т. 21, № 2. С. 241-250. https://doi.org/10.25559/SITITO.021.202502.241-250

Research and Development in the Field of New IT and Their Applications

Recognition of Complex Human Actions Based on Skeletal Pose Sequence Analysis

Original article

K. M. Maksimenko[a], L. N. Teryaev[a*], V. A. Dorokhin[a], A. V. Nechaevskiy[a,b]

[a] Dubna State University, Dubna, Russian Federation

Address: 19 Universitetskaya St., Dubna 141980, Moscow Region, Russian Federation

[b] Joint Institute for Nuclear Research, Dubna, Russian Federation

Address: 6 Joliot-Curie St., Dubna 141980, Moscow region, Russian Federation

[*] u1seanon@yandex.ru

Abstract

One of the priority areas of computer vision technology development is the extraction of skeletal data from human images and the subsequent use of this data to solve a whole range of applied problems. The paper gives a brief overview of technologies for solving the problem of human action recognition, highlights the main approaches, describes limitations, advantages and disadvantages. The authors propose a new approach to the recognition of complex human actions based on the analysis of skeletal data dynamics and application of state machine. The approach used is multi-stage and combines the sequential use of a neural network model of human pose detection MoveNet, a custom enhanced feature extraction layer (PoseEnhancementLayer), as well as an algorithm for detecting a committed action based on the analysis of poses at bifurcation points of the action. The solution proposed by the authors allows action detection without additional model training, which provides flexibility and scalability. Testing on open datasets showed high accuracy of human pose classification and robustness to incomplete or noisy sequences. The results are relevant for applications in sports analytics, interactive learning, rehabilitation and medical monitoring.

Keywords: pose recognition, skeleton, HPE, human action classification, computer vision, fully connected layers, state machine, fuzzy logic, classification pattern construction, XR

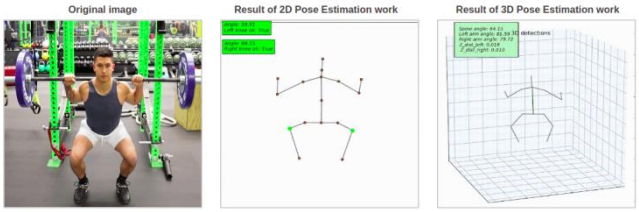
Conflict of interests: The authors declares no conflict of interest.

For citation: Maksimenko K.M., Teryaev L.N., Dorokhin A.V., Nechaevskiy A.V. Recognition of Complex Human Actions Based on Skeletal Pose Sequence Analysis. *Modern Information Technologies and IT-Education.* 2025;21(2):241-250. https://doi.org/10.25559/SITITO.021.202502.241-250

Введение

Развитие современного программного и аппаратного обеспечения позволяет использовать технологии компьютерного зрения, анализа поведения и машинного обучения в режиме реального времени. Это привело к возможности автоматизировать процессы в задачах, где ранее был необходим постоянный контроль человека. Особенно актуальной становится задача распознавания сложносоставных действий человека – ключевое направление, которое может применяться в таких областях, как медицина, спортивная аналитика, интерактивное обучение и других.

Human Pose Estimation (HPE) – это технология, которая по видео определяет и классифицирует определенные реперные точки на теле человека. Данные точки позволяют рассчитывать углы сгиба конечностей и как следствие, получать скелетную модель человека. Ключевые точки используются для создания 2D и 3D-представлений модели человеческого тела [1]. Общий алгоритм оценки позы тела начинается со сбора исходных данных и их загрузки в систему для обработки. Поскольку задача обнаружения движения, требует анализировать последовательность изображений, необходимо понимать, как ключевые точки меняются во время паттерна движения [2].



Р и с. 1. Процесс обработки данных в HPE [3]
F i g. 1. Data processing pipeline in HPE [3]

Процесс работы алгоритма можно разбить на два основных этапа (рис. 1):

- Обнаружение и извлечение ключевых точек из последовательности двумерных изображений. Это приводит к получению схемы расположения скелета относительно текущего угла зрения.
- Конвертация двумерного изображения в трехмерное, добавляя изображению глубину на основе анализа других изображений и известных принципов строения человеческого тела.

На сегодняшний день существует много имплементаций технологии распознавания и трекинга тела человека. Каждая технология имеет преимущество в определенных условиях, как правило, это зависит от параметров задачи, количества людей, расположения камеры и других факторов. Можно выделить технологии, которые показывают лучшие результаты на различных тестовых выборках[1].

OmniPose – однопроходная, сквозная обучаемая структура, которая обеспечивает высокие результаты для оценки позы нескольких человек. Архитектура *OmniPose* использует многомерные представления, которые повышают эффективность экстракторов основных функций без необходимости постобработки [4].

ViTPose – основывается на классическом трансформере. *ViTPose* использует простые и неиерархические преобразователи зрения в качестве основы для извлечения признаков и облегченный декодер для оценки позы. Его можно увеличить со 100 млн до 1 млрд параметров, используя преимущества масштабируемой емкости модели и высокого параллелизма преобразователей, изменяя соотношение между пропускной способностью и производительностью. Кроме того, *ViTPose* очень гибок в отношении точки зрения, разрешения входного изображения, стратегии предварительной подготовки и тонкой настройки, а также решения задач с несколькими позами [5].

MoveNet – ультра быстрая и точная модель, которая определяет 17 ключевых точек тела. Модель предлагается с двумя вариантами, отличающимися в скорости работы и качества распознавания. Обе модели работают быстрее, чем в реальном времени (более 30 кадров в секунду) на большинстве современных настольных компьютеров, ноутбуков и телефонов, что имеет решающее значение для приложений в сфере фитнеса и здоровья [6].

После завершения процесса получения облака трехмерных точек они могут использоваться для решения различных прикладных задач: классификация, оценка соответствия, поиск статистических отклонений от эталона и т.д.

В статье представлен подход к распознаванию сложносоставных действий человека на основе анализа динамики скелетных данных и алгоритма выявления действия, реализованном в виде машины состояний. Такой метод значительно ускоряет и упрощает внедрение модели во многие практические задачи.

Предлагаемая система не требует предварительного обучения на новых данных, устойчива к шуму и допускает гибкое добавление новых действий без необходимости перерабатывать архитектуру. Кроме того, она адаптирована для работы в реальном времени, что особенно важно для работы на мобильных устройствах.

Существующие решения в области отслеживания сложносоставных действий человека

На текущий момент методы распознавания действий человека в большинстве своем представлены тремя основными подходами: по видеоряду, по скелетным данным в динамике, по одному кадру. Данные методы активно развиваются и используются для решения различных прикладных задач.

Подход распознавания по видеоряду подразумевает

[1] Skeleton Based Action Recognition [Электронный ресурс] // Hugging Face, 2025. URL: https://paperswithcode.com/task/skeleton-based-action-recognition (дата обращения: 10.04.2025).

покадровую обработку видео в различных спектрах. Обучение строится исключительно на визуальных данных. Такой подход реализован в следующих моделях:

1. *SlowFast Networks for Video Recognition* – для распознавания действия человека используются 2 параллельные нейросети: быстрая и медленная, быстрая используется для того, чтобы понять контекст действия во времени, а медленная используется для определения того, что человек делает [7].
2. *Inflated 3D Convolutional Networks (I3D)* – это глубокая нейронная сеть, которая расширяет архитектуру *Inception* для обработки трёхмерных данных, включая видео. Использует архитектуру *Inception V1*, где каждая свертка захватывает и временной контекст [8].
3. *Video Transformer Network (VTN)* – этот метод представляет собой трансформерную архитектуру, предназначенную для распознавания действий в видео. В отличие от традиционных 3D-сверточных сетей, *VTN* использует механизм внимания для анализа всей видеопоследовательности, что позволяет эффективно учитывать как пространственные, так и временные зависимости. Это обеспечивает высокую точность при снижении вычислительных затрат [9].

Подход распознавания по **скелетным данным в динамике** состоит из набора этапов, принимает на вход видеоряд видимого спектра. На первом этапе проводится перевод всех кадров в информацию о скелете во времени, после чего уже на базе этих данных обучается классификатор. Такой подход реализован в следующих моделях:

1. *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition* – модель, которая обобщает распознавание действий по скелетным данным как задачу обучения на пространственно-временном графе [10].
2. Исследование движения человека в системах компьютерного зрения на основе скелетной модели – метод анализирует движение человека через выделение фигуры в кадре, построение координатных осей и моделей с использованием сплайн-функций. Отклонения от эталонного поведения (например, абнормальные скачки в движении) считаются девиантными и анализируются для дальнейшего обучения алгоритма [11].
3. *Skeleton-Based Action Recognition via Convolutional Neural Networks (CNN)* – в работе представлен подход к распознаванию действий человека, основанный на использовании сверточных нейронных сетей (CNN) для анализа скелетных данных. Авторы предлагают архитектуру, способную эффективно извлекать пространственные и временные характеристики из последовательностей скелетных точек, что позволяет улучшить точность классификации действий [12].

Также в качестве входных данных обучения и работы может выступать один кадр. Обучение в этом случае проводится по изображению или путем промежуточного преобразования данных в скелетную модель. Такой подход реализован в следующих моделях:

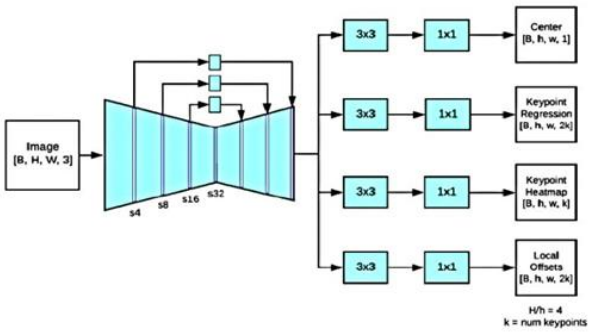
1. *Search-Map-Search: A Frame Selection Paradigm for Action Recognition* – модель, использующая полный набор полученных кадров, из которых выделяет ряд ключевых, которые представляют собой важные моменты для классификации действия [13].
2. *One-Frame Calibration with Siamese Network in Facial Action Unit Recognition* – модель, использующая пару изображений: текущий кадр и нейтральный эталон одного и того же человека, что позволяет учитывать индивидуальные особенности выражения лица. Сеть реализована в виде сиамской архитектуры, которая сравнивает признаки активного и нейтрального состояния, благодаря чему достигается высокая точность распознавания мимики даже при отсутствии обширных данных [14]. Основное назначение модели – эмоции, но может так же использоваться для распознавания действий.

В результате анализа было выявлено, что среди основных подходов к решению задачи распознавания действия человека не используется анализ последовательности кадров в бифуркационных точках действия. Данный подход ограниченно применялся в ряде работ, в основном – для решения четко определенной прикладной задачи, как правило, для подсчета количества выполненных упражнений.

Разработка модели классификатора скелетных поз на основе расширенного набора признаков

В качестве инструмента разработки модели использовался фреймворк *TensorFlow*. На первом этапе метода необходимо получить из кадров видео-источника скелетные данные. Для этого была использована модель *MoveNet*.

TensorFlow – открытая программная библиотека для машинного обучения, разработанная компанией Google для решения задач построения и тренировки нейронной сети с целью автоматического нахождения и классификации образов[2].



Р и с. 2. Архитектура MoveNet [19]
F i g. 2. MoveNet Architecture [19]

[2] Зиганшин И. А., Валиуллина Д. И. Применение библиотеки Tensorflow для обучения нейронных сетей // Лучшие научные исследования 2021: сб. ст. Межд. научно-исследовательского конкурса, Пенза, 27 июля 2021 года. Пенза: Наука и Просвещение, 2021. С. 14-16. EDN: NCYLJQ

Архитектура *MoveNet* состоит из модуля извлечения характеристик изображений *MobileNetV2: Inverted Residuals and Linear Bottlenecks* [16] с сетевым декодером *Feature Pyramid Networks for Object Detection* [17] (с шагом в 4), за которым следуют устройства прогнозирования *Objects as Points* [18] с пользовательской логикой постобработки. В Lightning используется множитель глубины 1,5.

Предлагаемая модель классификации *"PoseClassifier"* принимает на вход массив скелетных данных, полученных в результате обработки исходных изображений моделью *MoveNet*. Основным источником входных данных для классификатора служит массив, состоящий из 17 ключевых точек человеческого тела.

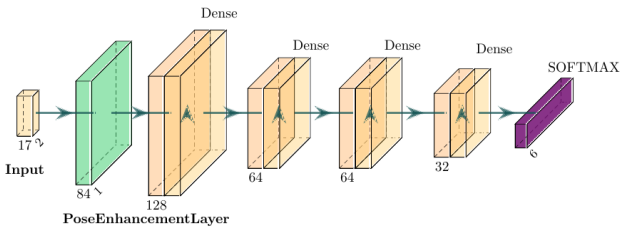
Для повышения качества распознавания и более глубокого извлечения признаков из этих данных был реализован отдельный слой *PoseEnhancementLayer*, расширяющий возможности исходного представления. Данный слой обладает высокой гибкостью и может быть адаптирован для работы с любыми моделями, выполняющими извлечение скелетных данных.

Этот слой выполняет следующие преобразования:

1. Центрирует скелет относительно бедер.
2. Вычисляет нормированные расстояния от всех ключевых точек до центра.
3. Определяет углы между основными сегментами тела: руками, ногами, коленями.
4. Вычисляет углы наклона корпуса относительно вертикали.
5. Измеряет длины отрезков между запястьями, плечами, бёдрами и голеностопами.
6. Оценивает ориентацию тела (стоя/сидя/лёжа) по наклону торса.

В результате все полученные признаки объединяются в единый тензор размерности 84, что позволяет эффективно описать текущую позу человека.

Обработанный вектор признаков подается на вход нейронной сети, состоящей из последовательности полносвязных слоёв с нормализацией. Архитектура нейронной сети изображена на рисунке 3.



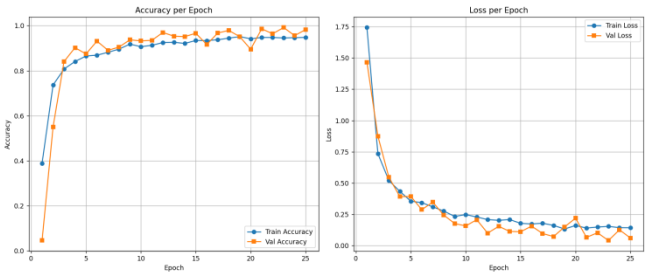
Р и с. 3. Архитектура разработанной модели PoseClassifier [15]
F i g. 3. Architecture of the developed PoseClassifier model [15]

Для обучения нейросети использовался открытый датасет упражнений *Workout/Exercises Video* на площадке *Kaggle*[3]. Пример входных данных для обучения изображен на рисунке 4.

Набор данных для обучения модели включал изображения людей в различных позах, снятые под разными углами и на разных фонах. Для каждого изображения были сгенерированы дополнительно три аугментированных изображения, что расширяет изначальный датасет и улучшает устойчивость модели к колебаниям в пространстве и различным вариациям углов обзора. Следует подчеркнуть, что устойчивость к шумам, теням, плохому освещению и нестандартным ракурсам в первую очередь определяется качеством работы используемой модели извлечения скелетных данных (в настоящей работе – *MoveNet*).

Р и с. 4. Пример обучающей выборки на 2 позы
F i g. 4. Example of a training dataset for two poses

В процессе обучения выборка делится на обучающую и тестовую в соотношении 4:1. Обучение с применением тестовых данных позволяет отслеживать качество обучения для каждой отдельной эпохи и автоматизировать процесс обучения [20, 21]. Общий объем обучающей выборки составил 1300 изображений, охватывающих 3 действия, которые были разложены на 7 отдельных классов поз. Набор возможных комбинаций действий, формируемых из этих классов, может быть значительно шире, чем исходный тренировочный набор. Такой подход позволяет масштабировать систему без переобучения – добавляя новые действия как комбинации уже известных поз. Итоговый график обучения нейронной сети изображен на рисунке 5. По результатам обучения была достигнута точность распознавания поз в районе 95 %.



Р и с. 5. Графики обучения модели
F i g. 5. Model training curves

Источник: здесь и далее в статье все рисунки составлены авторами.
Source: Hereinafter in this article all figures were drawn up by the authors.

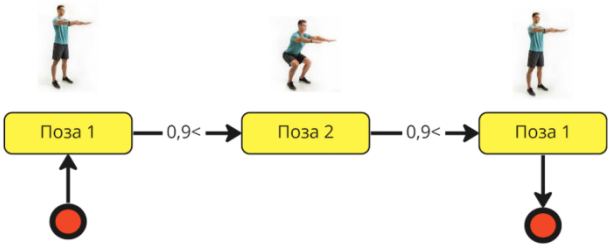
[3] Workout/Exercises Video Dataset [Электронный ресурс] // Kaggle, 2025. URL: https://www.kaggle.com/datasets/hasyimabdillah/workoutfitness-video (дата обращения: 10.04.2025).

Метод отслеживания многосоставных действий человека на основе машины состояний

Существующие подходы динамического анализа используют полный набор кадров для обучения классификатора. Это значительно утяжеляет модель и затрудняет ее использование в режиме реального времени, особенно на слабых устройствах.

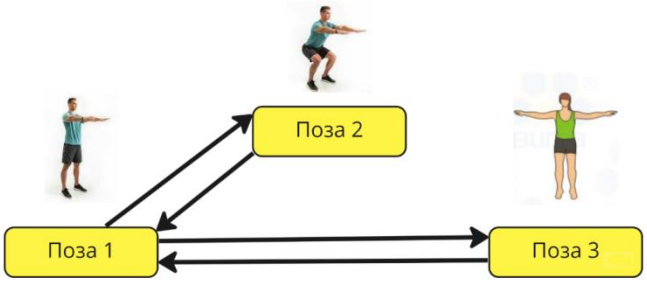
Предлагаемый авторами метод основывается на выделении в действии ключевых кадров и анализе их последовательности на основе классических алгоритмов обработки данных.

Полученный классификатор в качестве выходных данных возвращает массив дискретных значений, описывающих вероятность наличия конкретной позы на входном изображении. Это позволяет на основе выбранного граничного значения определить факт нахождения человека в определенной позе. Сохранение цепочки таких классификаций позволяет проводить их дальнейший анализ для классификации уже много составного действия.



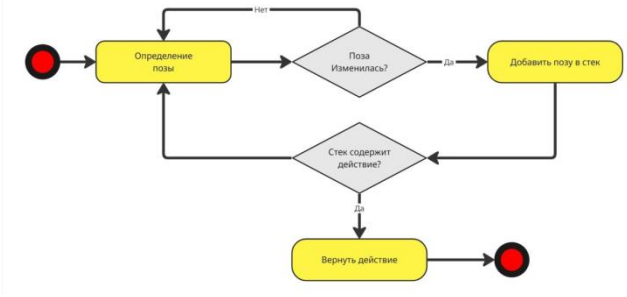
Р и с. 6. Пример схемы упражнения
F i g. 6. Example of an exercise scheme

На рисунке 6 представлен пример работы алгоритма для отслеживания упражнения. В том случае, если известно выполнение какого действия необходимо отслеживать, алгоритм работы линеен. В случае, когда действие состоит из нескольких поз, возможно ветвление переходов (рис. 7).



Р и с. 7. Пример машины состояний
F i g. 7. Example of a state machine

Повышение количества известных системе поз улучшает качество распознавания конкретного действия, поскольку уменьшается время нахождения объекта в промежуточном состоянии. Это в свою очередь дает возможность повышать пороговые значения перехода между состояниями. Помимо этого метод позволяет достичь экспоненциального роста количества потенциально распознаваемых действий относительно количества поз в модели классификатора. Общий алгоритм сопоставления распознавания позы по шаблону представлен на рисунке 8.



Р и с. 8. Алгоритм определения действия
F i g. 8. Action detection algorithm

Для распознавания действий используется структура данных «стек», реализованная на основе ограниченной очереди. Структура хранит историю последних зарегистрированных поз человека. При появлении новой позы, в случае ее несовпадения с предыдущей и наличии приемлемой достоверности, происходит внесение полученной позы в стек.

Такой подход не требует повторного обучения модели для распознавания новых действий в том случае, если искомое действие состоит из поз уже присутствующих в обученной модели. Для добавления нового действия, достаточно внести его шаблон в базу знаний. Это делает систему гибкой и расширяемой, подходящей для применения в интерактивном обучении, спорте, реабилитации и других задачах, где необходимо отслеживать действия человека по позам.

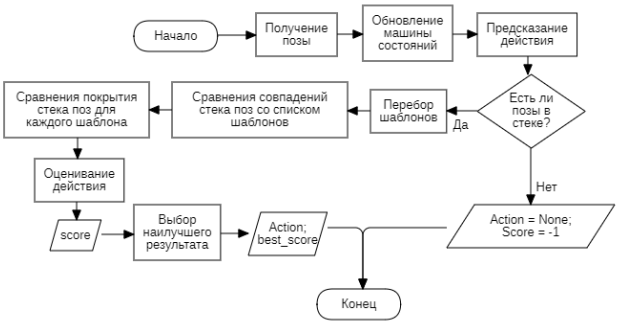
Реализация и тестирование алгоритма определения действий на основе анализа последовательности поз

В алгоритме машины состояний, стек имеет фиксированную длину, равную максимальной длине действия в базе шаблонов. Это значит, что автоматически хранится только последние N поз, где N – максимальная длина цепочки в базе. Такой подход позволяет поддерживать статичный объем памяти, в пределах которого происходит анализ, поскольку максимальная длина шаблонной последовательности заранее известна.

В ходе анализа для каждого шаблона подсчитывается:

1. количество совпавших поз из шаблона и текущего стека (*matches*),
2. доля покрытия шаблона (*coverage*) – насколько полно стек отражает весь шаблон.

На выходе алгоритм выбирает шаблон с наибольшим количеством совпадений. Если таких шаблонов несколько, приоритет отдается тому действию, которое имеет большее покрытие. Таким образом, даже если последовательность поз неполная, содержит пропуски или последовательность действий смещена, система может корректно идентифицировать действие. Пошаговое описание алгоритма изображено на рисунке 9.

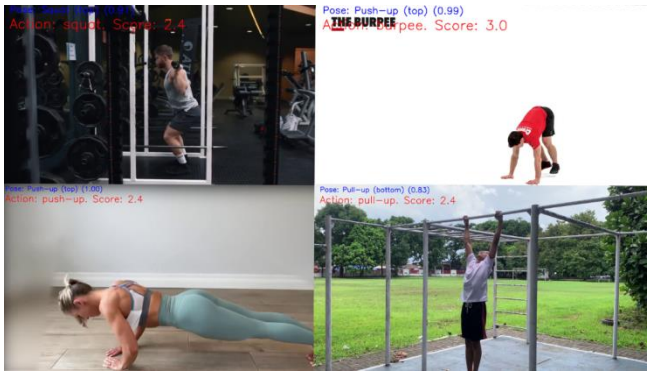


Р и с. 9. Блок-схема алгоритма машины состояний
F i g. 9. Flowchart of the state machine algorithm

База данных паттернов, описывающих действия, состоит из конечного набора элементов *X*:
$T=\{X_1, X_2, ... X_n \vee n>1\}$
При этом для работы алгоритма база данных должна содержать как минимум 2 паттерна.
X представляет собой множество детекций описывающее паттерн действия:
$X=\{x_1, x_2, ... x_n | x \in P \wedge n>1\}$,
где $P=(p_1, p_2, p_3, p_n \vee n>1)$ – конечный словарь поз распознавания.
Стек представляет собой:
$S=\{s_1, s_2, \ldots, s_n \vee s \in P \wedge n>1\}$,
где максимальная длина стека:
$N=max(len(X_1), len(X_2), \ldots, len(X_n))$
Блок сравнения совпадений стека поз со списком шаблонов основан на алгоритме *LCS (Longest Common Subsequence)*. Алгоритм ищет самую длинную подпоследовательность между двумя строками. В общем случае каждая строка имеет 2^n подпоследовательности. Подпоследовательность создается путем удаления 0 или более символов, без изменения относительной последовательности символов строки [22].
Формула LCS (Longest Common Subsequence) между двумя последовательностями определяется следующим рекурсивным соотношением:

$$LCS(i,j) = \begin{cases} 0 & \text{если } i=0 \text{ или } j=0 \\ LCS(i-1, j-1)+1 & \text{если } S_1[i-1]=S_2[j-1] \\ \max(LCS(i-1,j),\ LCS(i,j-1)) & \text{если } S_1[i-1] \neq S_2[j-1] \end{cases}$$

Итоговой длинной последовательности является число
$matches = LCS(N, len(X))$
Следует отметить, что классическая реализация *LCS* имеет вычислительную сложность порядка *O(n × m)*, где *n* и *m* — длины сравниваемых последовательностей. При значительном увеличении длины шаблонов это может приводить к росту времени обработки. В случае применения метода для длинных действий или большого количества шаблонов может потребоваться оптимизация алгоритма. Кроме того, при использовании на мобильных устройствах важно учитывать энергопотребление, которое может быть снижено путём уменьшения частоты кадров. Тем не менее, абсолютное большинство действий состоят из небольшого набора поз, что нивелирует данный недостаток.
Сравнение покрытия стека поз для каждого шаблона выполняет данное действие:
$coverage = \frac{matches}{¿X \vee ¿¿}$
Выбор наилучшего результата осуществляется на основе взвешенной оценки качества совпадения. Итоговый балл рассчитывается по следующей формуле:
$score = match_{weight} \cdot match + coverage_{weight} \cdot coverage$,
Где *match_weight* и *coverage_weight* - пользовательские настройки весов каждого качества. По умолчанию *match_weight* равняется 0.7, а *coverage_weight* равняется 0.3. Параметры весов были выбраны эмпирически на основе валидационной выборки, обеспечивая оптимальный баланс между точностью определения позы (*match*) и полнотой покрытия действия (*coverage*). Такой подход позволил минимизировать количество ложных срабатываний при сохранении устойчивости к пропускам поз.
После создания машины состояний система готова к эксплуатации в реальном времени. Изображения с результатами работы алгоритма изображены на рисунке 10.



Р и с. 10. Результат тестирования алгоритма
F i g. 10. Algorithm testing result

Тестирование проводилось на персональном компьютере под управлением Windows 11 с процессором Intel Core i5-i5-9400f, графическим адаптером NVIDIA GeForce GTX 1650 и 16 ГБ оперативной памяти.
Результаты тестирования показали, что предложенный метод обрабатывает ~79.08 кадров в секунду, при этом точность классификации варьируется в диапазоне 71-75%.
В текущей реализации алгоритм ориентирован на обработку данных с одним человеком в кадре. Это обусловлено тем, что входной поток поз формируется моделью *MoveNet* в режиме *single-person detection*. Для расширения применения в многопользовательских сценах потребуется адаптация входного блока под многопотоковую обработку *(multi-person pose estimation)* и доработка механизма сопоставления последовательностей поз для нескольких объектов одновременно.
Данное решение позволяет применять модель в таких задачах, как медицина, спортивная аналитика,

интерактивное обучение, иммерсивные технологии или для реализации методов интеллектуального управления в различных прикладных сферах.

Предложенный подход применяется авторами для разработки системы управления сценическим пространством. Распознавание поз и действий актеров служит триггерным действием для переключения световых партитур.

Для задач глобального внедрения метапространства и георасределенных иммерсивных технологий распознавание поз и действий может быть использовано для введения в виртуальное пространство не только статических объектов реального мира, но и самих пользователей. Современный уровень научно-технического развития уже в ближайшее время будет позволять глобальное внедрение подобных решений [23]. Использование результатов работы технологии позволит реализовывать новые способы взаимодействия пользователей в синхронизированном виртуальном пространстве [24].

Текущая реализация технологии позволяет создавать виртуальных тренеров ведущих для пользователя тренировку с автоматическим подсчетом количества выполненных упражнения [25]. Дальнейшая доработка может позволить проводить оценку качества выполнения упражнений и давать рекомендации по поводу их более правильного выполнения.

Заключение

На сегодняшний день для распознавания действий человека существует несколько распространенных подходов, каждый из которых имеет свои достоинства и недостатки:

1. Обучение по видеоряду – хорошо распознает временный контекст, но требует больших вычислительных ресурсов, плохо масштабируется и не подходит для *real-time* приложений.
2. Классификация по скелетным данным в динамике – метод снижает объем данных, не зависит от фона, освещения, что позволяет работать даже при сильных помехах качества, но по-прежнему требует предварительной сегментации и накопления всего видеофрагмента;
3. Классификация по одному изображению - высокая скорость, не требует больших вычислительных ресурсов, но теряет временной контекст, из-за чего возможны ложные срабатывания и путаница между схожими позами.

Все эти методы в той или иной степени зависят от обучения нейросети на конкретных данных, что делает их чувствительными к изменению условий (углы съёмки, одежда, освещение), а также требует повторного обучения при добавлении новых действий.

Авторами предложен подход, на основе машины состояний и шаблонов поз, который имеет ряд преимуществ:

1. Не требует переобучения модели для добавления новых действий – достаточно ввести шаблон последовательности;
2. Имеет низкое время отклика – возможность эффективно применять в реальном времени;
3. Устойчив к пропускам ключевых поз и неполным последовательностям, что особенно важно для реальных условий эксплуатации.
4. Метод может быть модифицирован для задачи анализа соответствия действия заданному эталону.

Таким образом, предложенный метод, модель и алгоритмы для определения действий человека представляют собой гибкую, масштабируемую альтернативу традиционным методам классификации действий, открывая возможности для создания умных, адаптивных систем сопровождения и анализа активности человека.

References

1. Song X., et al. Quater-GCN: enhancing 3D human pose estimation with orientation and semi-supervised training. *Frontiers in Artificial Intelligence and Applications*. Vol. 392:(ECAI). IOS Press; 2024. p. 121-128. https://doi.org/10.3233/FAIA240479
2. Zhou L., et al. Human pose-based estimation, tracking and action recognition with deep learning: a survey. *arXiv*:2310.13039. 2023. https://doi.org/10.48550/arXiv.2310.13039
3. Jan M.T., Kumar A., Sonar V.G., et al. Comprehensive survey of body weight estimation: techniques, datasets, and applications. *Multimedia Tools and Applications*. 2025;84:28807-28837. https://doi.org/10.1007/s11042-024-20318-4
4. Artacho B., Savakis A. Omnipose: A multi-scale framework for multi-person pose estimation. *arXiv:*2103.10180. 2021. https://doi.org/10.48550/arXiv.2103.10180
5. Xu Y., et al. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;46(2):1212-1230. https://doi.org/10.1109/TPAMI.2023.3330016
6. Bajpai R., Joshi D. MoveNet: A Deep Neural Network for Joint Profile Prediction Across Variable Walking Speeds and Slopes. *IEEE Transactions on Instrumentation and Measurement*. 2021;70:2508511. https://doi.org/10.1109/TIM.2021.3073720
7. Feichtenhofer C., Fan H., Malik J., He K. SlowFast Networks for Video Recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE Press; 2019. p. 6201-6210. https://doi.org/10.1109/ICCV.2019.00630

8. Freire-Obregón D., Barra P., Castrillón-Santana M., et al. Inflated 3D ConvNet context analysis for violence detection. *Machine Vision and Applications*. 2022;33:15. https://doi.org/10.1007/s00138-021-01264-9
9. Neimark D., Bar O., Zohar M., Asselmann D. Video Transformer Network. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, BC, Canada: IEEE Press; 2021. p. 3156-3165. https://doi.org/10.1109/ICCVW54120.2021.00355
10. Yan S., Xiong Y., Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). Article number: 912. AAAI Press; 2018. p. 7444-7452.
11. Kazakova S.A., Leonteva P.A., Frolova M.I., Donetskaya Ju.V., Popov I.Yu., Kuznetsov A.Yu. A study of human motion in computer vision systems based on a skeletal model. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics.* 2021;21(4):571-577. (In Russ., abstract in Eng.) https://doi.org/10.17586/2226-1494-2021-21-4-571-577
12. Ali A., et al. Skeleton-based human action recognition via convolutional neural networks (CNN). *arXiv:*2301.13360. 2023. https://doi.org/10.48550/arXiv.2301.13360
13. Zhao M., Yu Y., Wang X., Yang L., Niu D. Search-Map-Search: A Frame Selection Paradigm for Action Recognition. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE Press; 2023. p. 10627-10636. https://doi.org/10.1109/CVPR52729.2023.01024
14. Feng S., de Sa V.R. One-Frame Calibration with Siamese Network in Facial Action Unit Recognition. *arXiv:*2409.00240. 2024. https://doi.org/10.48550/arXiv.2409.00240
15. Pham D.T., Diem C.H. Deep Learning for Hand Gesture Recognition Using Channel-Wise Topology Refinement. In: Le Thi H.A., Pham Dinh T., Le H.M. (eds.) Modelling, Computation and Optimization in Information Systems and Management Sciences. MCO 2025. *Lecture Notes in Networks and Systems*. Vol. 1689. Cham: Springer; 2026.p. 250-259. https://doi.org/10.1007/978-3-032-08384-5_21
16. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L. -C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE Press; 2018. p. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474
17. Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature Pyramid Networks for Object Detection. In; 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE Press; 2017. p. 936-944. https://doi.org/10.1109/CVPR.2017.106
18. Zhou X., Wang D., Krähenbühl P. Objects as points. *arXiv:*1904.07850. 2019. https://doi.org/10.48550/arXiv.1904.07850
19. Mishra A.K., Sahoo D., Subhankar I., Samal I. YogaSiddhi: AI-powered pose analysis using MoveNet for yoga refinement. *International Journal of Computer Applications*. 2024;186(4):33-39. https://doi.org/10.5120/ijca2024923427
20. Al-Kababji A., Bensaali F., Dakua S.P. Scheduling Techniques for Liver Segmentation: ReduceLRonPlateau vs OneCycleLR. In: Bennour A., Ensari T., Kessentini Y., Eom S. (eds.) Intelligent Systems and Pattern Recognition. ISPR 2022. *Communications in Computer and Information Science*. Vol. 1589. Cham: Springer; 2022. p. 204-212. https://doi.org/10.1007/978-3-031-08277-1_17
21. Ziebell E., et al. EarlyStopping: Implicit Regularization for Iterative Learning Procedures in Python. *arXiv:*2503.16753. 2025. https://doi.org/10.48550/arXiv.2503.16753
22. Bergroth L., Hakonen H., Raita T. A survey of longest common subsequence algorithms. In: Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000. A Curuna, Spain: IEEE Press; 2000. p. 39-48. https://doi.org/10.1109/SPIRE.2000.878178
23. Dorokhin V.A., Teryaev L.N., Zorin R.A. Augmented Reality Technologies and Prospects for Their Global Application within the Framework of the Geo-Oriented XR-Internet Concept. *Modern Information Technologies and IT-Education.* 2023;19(2):403-411. (In Russ., abstract in Eng.) https://doi.org/10.25559/SITI-TO.019.202302.403-411
24. Dorokhin V.A. Augmented Reality and Synchronisation of its Events. *System analysis in science and education.* 2017;(2):1-5. (In Russ., abstract in Eng.) EDN: ZTQADL
25. Gupta L., Gurbuxani S., Madan K. Virtual Fitness Trainer using Artificial Intelligence. In: Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing (IC3-2024). New York, NY, USA: Association for Computing Machinery; 2024. p. 226-233. https://doi.org/10.1145/3675888.3676056

Поступила 10.04.2025; одобрена после рецензирования 21.05.2025; принята к публикации 19.06.2025.

Submitted 10.04.2025; approved after reviewing 21.05.2025; accepted for publication 19.06.2025.

Об авторах:

Максименко Кирилл Михайлович, студент Института системного анализа и управления, ГБОУ ВО Московской области «Университет «Дубна» (141982, Российская Федерация, Московская область, г. Дубна, ул. Университетская, д. 19), **ORCID: https://orcid.org/0009-0000-9251-3812,** pu1seanon@yandex.ru

Теряев Лев Николаевич, аспирант кафедры распределенных и вычислительных систем Института системного анализа и управления, ГБОУ ВО Московской области «Университет «Дубна» (141982, Российская Федерация, Московская область, г. Дубна, ул. Университетская, д. 19), **ORCID: https://orcid.org/0009-0008-3188-7616,** ttrchik228@gmail.com

Дорохин Виктор Александрович, старший преподаватель кафедры распределенных и вычислительных систем Института системного анализа и управления, ГБОУ ВО Московской области «Университет «Дубна» (141982, Российская Федерация, Московская область, г. Дубна, ул. Университетская, д. 19), **ORCID: https://orcid.org/0000-0001-5283-614X,** victor.doroh@gmail.com

Нечаевский Андрей Васильевич, и.о. проректора по цифровому развитию, ГБОУ ВО Московской области «Университет «Дубна» (141982, Российская Федерация, Московская область, г. Дубна, ул. Университетская, д. 19); старший научный сотрудник Лаборатории информационных технологий имени М.Г. Мещерякова, Международная межправительственная организация Объединенный институт ядерных исследований (141980, Российская Федерация, Московская область, г. Дубна, ул. Жолио-Кюри, д. 6), **ORCID: https://orcid.org/0000-0001-6751-8195,** nechav@jinr.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Kirill M. Maximenko, Student at the Institute of System Analysis and Management, Dubna State University (19 Universitetskaya St., Dubna 141980, Moscow Region, Russian Federation), **ORCID: https://orcid.org/0009-0000-9251-3812,** pu1seanon@yandex.ru

Lev N. Teryaev, Postgraduate student at the Department of Distributed and Computing Systems, Institute of System Analysis and Management, Dubna State University (19 Universitetskaya St., Dubna 141980, Moscow Region, Russian Federation), **ORCID: https://orcid.org/0009-0008-3188-7616,** ttrchik228@gmail.com

Viktor A. Dorokhin, Senior Lecturer of the Department of Distributed and Computing Systems, Institute of System Analysis and Management, Dubna State University (19 Universitetskaya St., Dubna 141980, Moscow Region, Russian Federation), **ORCID: https://orcid.org/0000-0001-5283-614X,** victor.doroh@gmail.com

Andrey V. Nechaevsky, Acting Vice-Rector for Digital Development, Dubna State University (19 Universitetskaya St., Dubna 141980, Moscow Region, Russian Federation); Senior Researcher of the Mescheryakov Laboratory of Information Technologies, Joint Institute for Nuclear Research (6 Joliot-Curie St., Dubna 141980, Moscow region, Russian Federation), **ORCID: https://orcid.org/0000-0001-6751-8195,** nechav@jinr.ru

All authors have read and approved the final manuscript.