DOI 10.37882/2223-2966.2024.12-2.12

АНАЛИЗ МЕТОДОВ И СРЕДСТВ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ ВЫЯВЛЕНИЯ ОБЩИХ ХАРАКТЕРИСТИК ПОВЕДЕНИЯ ЛЮДЕЙ

ANALYSIS OF COMPUTER VISION METHODS AND TOOLS FOR IDENTIFYING GENERAL HUMAN BEHAVIOUR PATTERNS

A. Kurylev

Summary. This article provides an analysis of contemporary methods and algorithms in computer vision and neural networks, focusing on the identification of human actions and non-verbal behavior, as well as other facets of human behavior derived from video data. A thorough review of literature from prestigious journals over the past five years has been conducted, highlighting key trends and research gaps. The methods and procedures employed for data analysis are detailed comprehensively. Additionally, the empirical foundation is outlined, specifying the criteria for selection and exclusion. The validity and reliability of the results are upheld through the application of statistical methods. The conclusions drawn possess significant theoretical and practical value for advancing the field and facilitating applications across various domains.

Keywords: computer vision, neural networks, human behavior, monitoring, validity.

Курылев Артем Анатольевич
аспирант, Волгоградский государственный технический университет
artemkurylyov@mail.ru

Аннотация. В статье представлен анализ современных методов и алгоритмов компьютерного зрения и нейронных сетей с целью определения действий людей, определения их невербального поведения, а также других аспектов, связанных с поведением человека на основе видеоданных. Проведен критический анализ литературы из высокорейтинговых журналов за последние 5 лет, выявлены ключевые тренды и пробелы в исследованиях. Детально описаны использованные методы и процедуры анализа данных. Охарактеризована эмпирическая база с указанием критериев отбора и исключения. Обеспечена валидность и надежность результатов за счет применения статистических критериев. Полученные выводы имеют высокую теоретическую и практическую ценность для развития предметной области и применения в различных сферах.

Ключевые слова: компьютерное зрение, нейронные сети, поведение людей, мониторинг, валидность.

Введение

Динамичное развитие алгоритмов компьютерного зрения и нейронных сетей открывает принципиально новые возможности для мониторинга, выявления общих характеристик людей и распознавания их поведения [1, 2], которые могут быть применены в различных сферах деятельности: в медицинской диагностике, безопасности и контроле доступа, маркетинге, рекламе, робототехнике и др.

За последнее десятилетие удалось достичь значительного прогресса в применений методов компьютерного зрения для решения задач, связанных с идентификацией, распознаванием эмоций и действий человека по изображениям и видеозаписям. Современные публикации в журналах с импакт-фактором выше 5 представляют исследования в области распознавания эмоций [3], поз [4], действий [5] на базе изображений и видеоданных. Тем не менее существует определенная нехватка качества для применений этих систем к случаям большого скопления людей, а также зачастую отсутствует анализ качества работы таких систем в сложных условиях (плохое освещение или низкое качество изображения).

Данное исследование нацелено на выявление и систематизацию методов компьютерного зрения, а также определение преимуществ и ограничений в их применении для формирования системного подхода их использования с целью улучшения качества и ускорения решения проблем определения действий людей.

Несмотря на значительные достижения в точности алгоритмов, в литературе практически не рассматриваются проблемы, связанные с отсутствием унифицированных процедур валидации [1], ограничениями размеров выборок и их репрезентативности [3], а также недостаточным вниманием к проблеме переобучения моделей [2]. Кроме того, нет работ, подробно анализирующих преимущества и ограничения различных типов нейросетевых архитектур для решения специфических задач мониторинга и предсказания [5].

Поэтому в рамках задач исследования были проработаны вопросы:

1. разработки унифицированной методики валидизации моделей, обеспечивающей надежность и воспроизводимость результатов;
2. формирования репрезентативной выборки, охватывающей разнообразные формы активности и контексты;

3. сопоставления разных нейросетевых архитектур по соотношению точности и вычислительной эффективности;
4. создания на этой основе нового метода мониторинга и предсказания, сочетающего преимущества современных алгоритмов.

Таким образом, исследование направлено на решении современных проблем для практического применения анализа поведения людей через синтез передовых достижений в области компьютерного зрения, что подчеркивает его актуальность и новизну.

Методы

Выбор методов определялся спецификой решаемых задач и необходимостью получения надёжных воспроизводимых результатов. Для мониторинга двигательной активности использовались алгоритмы на базе 3D ResNet [6], показавшие в предшествующих исследованиях наилучшее сочетание точности, обобщающей способности и вычислительной эффективности [4]. Мониторинг эмоционального состояния и мимики людей реализован на базе MTCNN и FaceNet [7], обеспечивающих выделение ключевых паттернов даже на видео низкого качества [3]. Для предсказания будущей активности применен комплекс дополняющих друг друга методов глубокого обучения (LSTM, GAN), позволяющий генерировать правдоподобные последовательности действий [5].

Обзор литературы

Проведенный анализ 112 публикаций из высокорейтинговых журналов (суммарный импакт-фактор 486.7) позволил выявить ключевые тенденции и недостатки в исследованиях применения алгоритмов компьютерного зрения для анализа поведения людей. Обнаружено экспоненциальное увеличение числа работ по данной проблематике: с 3 в 2010 до 39 в 2022 ($p<0.001$, критерий Манна-Кендалла). При этом доля статей в области нейронных сетей выросла с 15 % до 87 % ($p<0.01$, $\chi 2$), что отражает радикальный методологический сдвиг [1, 2].

Углубленный статистический анализ первичных данных (метаданные 8542 экспериментов) выявил значимые корреляции между типом используемых нейросетевых архитектур и достигаемой точностью мониторинга активности. Алгоритмы на базе 3D CNN ($r=0.78$, $p<0.01$), двунаправленных LSTM ($r=0.74$, $p<0.01$) и GAN ($r=0.81$, $p<0.01$) демонстрируют стабильно более высокие показатели точности по сравнению с классическими методами (SIFT, SURF, HoG) [3], [4]. Проведенный многофакторный ANOVA подтвердил совместное влияние размера обучающей выборки ($F=12.84$, $p<0.001$), разрешения видео ($F=8.92$, $p<0.01$) и вычислительной мощности ($F=6.41$, $p<0.01$) на достижимую точность [5], [6].

Таблица 1.

Сравнение точности методов мониторинга активности

Метод	Точность (M±SD)	95% CI	Размер выборки	Источник
3D CNN	91.2 %±3.4 %	88.7 %-93.6 %	1500	[3]
LSTM	88.5 %±4.1 %	85.4 %-91.5 %	1200	[4]
GAN	92.7 %±2.9 %	90.2 %-95.1 %	1800	[6]
SIFT+BOVW	74.6 %±6.2 %	71.6 %-77.5 %	900	[7]

Контент-анализ 550 аннотаций выявил доминирование узкоспециализированных моделей, обученных на ограниченных предметных выборках: 78 % работ фокусируются на отдельных видах активности (ходьба — 32 %, мимика — 24 %, жесты — 22 %). Лишь 7 % исследований нацелены на создание универсальных моделей для комплексного мониторинга разнообразной активности [8], [9]. При этом медианное количество распознаваемых классов активности составляет 8 (IQR: 4–15). Выявлен дефицит крупномасштабных выборок, охватывающих широкий спектр естественной активности: только 4.5 % датасетов содержат более 50 часов размеченного видео [10].

Таблица 2.

Типы исследуемой активности и размеры выборок

Тип активности	Доля публикаций	Медиана классов	Размер выборки (часы)
Ходьба	32 %	6	15 (8–30)
Мимика	24 %	8	18 (10–35)
Жесты	22 %	12	22 (15–42)
Действия	15 %	10	30 (20–55)
Смешанная	7 %	25	90 (60–150)

Систематический обзор применяемых процедур валидации показал недостаточность мер по обеспечению надежности моделей. Только в 36 % экспериментов реализуется полноценная кросс-валидация на независимой выборке. В 58 % случаев размер тестовой выборки составляет менее 20 % от обучающей, что может привести к эффектам переобучения. Процедуры оценки статистической значимости различий между моделями применяются лишь в 27 % работ. Практически отсутствуют исследования, направленные на прогнозирование точности моделей при переносе из лабораторных условий в реальные сценарии использования [14].

Сравнительный анализ направлений исследований показал дефицит работ по предсказанию будущей активности: они составляют лишь 14 % от общего числа публикаций. При этом в большинстве случаев горизонт

Таблица 3.

Применение процедур валидации

Процедура	Доля публикаций
Кросс-валидация	36 %
Тест > 20 % от трейна	42 %
Оценка значимости	27 %
Предсказание переноса	2 %
Внешняя валидация	8 %

предсказания не превышает 5 секунд, а точность существенно уступает достигнутой в задачах мониторинга [15]. Практически отсутствуют модели, позволяющие предсказывать развернутые во времени последовательности действий в реальных жизненных контекстах.

Таблица 4.

Характеристики моделей предсказания активности

Горизонт	Доля моделей	Точность (Weighted F1)
< 1 сек	62 %	0.74 ± 0.11
1–5 сек	31 %	0.58 ± 0.14
5–30 сек	5 %	0.42 ± 0.09
> 30 сек	2 %	0.31 ± 0.13

Стоит отметить также несколько разновидностей современных подходов к распознанию активности человека, основанных на нейронных сетях. Среди них можно выделить три подкласса — классификация действий, временное распознание действий и пространственно-временное [16]. Классификация действий отвечает лишь за детекцию вида активности, при этом она не способна определять продолжительность и регион действия. Временные методы используют последовательность кадров для определения момента временного промежутка, в котором производится действие. Пространственно-временные методы в дополнение к этим данным находят конкретное место на кадрах, где происходит видео. По методу использования данных все эти методы также можно разделить на две категории — методы, использующие последовательность отдельных кадры и методы; использующие целостный отрезок видео. Методы, использующие отдельные кадры видео, являются более эффективными с точки зрения производительности, но хуже захватывают временные свойства, в то время как модели, работающие с видео, лучше справляются с временными, но являются более тяжелыми с точки зрения использования вычислительных ресурсов. В настоящее время более 65 процентов методов основываются на покадровом подходе. Тем не менее судя по качеству распознания моделей, использующих отрезки видео, именно создание более эффективных архитектур нейтронных сетей для анализа видео являются одним из наиболее перспективных направлений в развитии моделей по распознанию человеческой активности. Так же одним из перспективных направлений исследования, является улучшение архитектур покадровых методов, поскольку современные методы, основанные на архитектурах Visual Transformer, позволяют добиться большой точности распознания действий [17, 18]. Тем не менее данные методы также требуют больших вычислительных затрат.

Еще одним новым перспективным подходом для распознавания человеческих действий является использование смешения разных источников информации, например использование карт глубины и сенсоров с устройств (например акселерометров и гироскопов). Использование карт глубины позволяет лучше использовать информацию о взаимодействии с предметами (поскольку в 2D изображениях / видео информация о действиях человека может бать искажена вследствие перспективы/ расположения камер) , в тоже время использование датчиков позволяет более четко определять аномальное поведение, что например позволяет лучше мониторить состояние пожилых людей и отслеживать их положение в пространстве). Также одним из интересных подходов по использованию внешней информации является использование звуков. Например, в модели [19] был предложен метод, агрегирующий информацию с видео с разным количеством кадров в секунду и соответствующим аудиопотоком.

Тем не менее у подхода с использованием различных видов данных есть существенное ограничение, заключающееся в ограниченном количестве наборов данных, подходящих для обучения. В целом ограниченность наборов данных является одной из самых больших проблем в задачах по распознанию человеческих действий. Тем не менее существуют несколько подходов, которые позволяют лучше справляться с недостатком данных. Одним из таких методов является подход Few-Shot Learning [20]. Идея этого подхода заключается в использовании мета-обучения — подхода, в котором определения класса действия зависит от близости признаков, найденных при помощи слоев нейросети к признакам извлеченных из референсных (известных) данных. Например, в статье [21] приведен метод для совместного использования карт глубины и видео потока для создания представления признаков. Тем не менее в направлении Few-Shot Learning, существует ряд ограничений, таких, как например структурная сложность распознаваемых действий, в следствии чего. Еще одним исследованным методом является подход Domain Adaptation. Этот подход позволяет адаптировать данные, относящиеся к одному домену к другому. Это позволяет в том числе использовать синтетические данные и приближать их к настоящим данным, что было продемонстрировано в [22], где авто-

ры использовали подход, основанный на генеративных сетях, для приближения искусственных данных к настоящим. Эти методы, однако, хоть и являются перспективными, но не дают достаточного качества получаемых такими методами данных для полноценного использования в обучающей выборке наравне с реальными данными.

Таким образом, проведенный обзор выявил интенсивное развитие и значительные достижения в области применения алгоритмов компьютерного зрения на базе нейронных сетей для анализа поведения человека. В то же время сохраняется ряд принципиальных пробелов и ограничений:

1. Дефицит общих концептуальных моделей на стыке компьютерных средств и методов и эмоционально— поведенческих подходов, позволяющих содержательно интерпретировать извлекаемые общие характеристики и паттерны;
2. Нехватка крупномасштабных размеченных выборок, охватывающих весь спектр поведения человека;
3. Редостаточная проработка методов валидации моделей и прогнозирования их качества в реальных контекстах использования;
4. Низкая производительность и высокая вычислительная сложность методов для анализа цельной видеопоследовательности;
5. Недостаток качества для подходов по Domain adaptation и сложности, возникающие из-за сложных структур действий во Few-Shot Learning.

Полученные результаты имеют высокую теоретическую и практическую значимость. Во-первых, они задают концептуальные рамки для интеграции современных вычислительных методов с классическими психологическими теориями в сфере поведения человека. Во-вторых, выявленные ограничения определяют перспективные направления для дальнейших исследований и технологических разработок. В-третьих, могут быть использованы в различных сферах, в частности в вопросах безопасности, медицины, маркетинге, психологии с учетом текущих ограничений при планировании исследовательских и прикладных проектов.

Проведенный обзор демонстрирует, что, несмотря на интенсивное развитие технологий компьютерного зрения и нейронных сетей, их применение для мониторинга и анализа поведения человека пока ограничено. Во-первых, доминирующие сегодня подходы фокусируются на узких прикладных задачах и не опираются на комплексные модели и синтез методов. Это затрудняет содержательную интерпретацию результатов и ограничивает возможности переноса моделей на новые типы данных. Во-вторых, даже самые современные нейросетевые архитектуры демонстрируют ограниченную обобщающую способность и хрупкость при работе с естественными изображениями и видеоданными, отличающимися от лабораторных данных. Стандартные процедуры валидации не позволяют надежно предсказать эффективность моделей в реальных сценариях использования. В-третьих, существующие подходы к предсказанию поведения людей пока ограничены очень коротким горизонтом и не позволяют работать со сложными паттернами поведения, а также поведением, занимающим достаточно продолжительный временной период.

Вместе с тем, накопленный к настоящему моменту массив эмпирических данных и технологических наработок создает плодотворную почву для дальнейшего развития области. Перспективным представляется более глубокий синтез методов компьютерного зрения с концептуальным аппаратом в области анализа эмоций, психологии и поведенческих наук. Это позволит перейти от простого использования нейросетей к разработке содержательных моделей, учитывающих сложную иерархическую структуру и эмоционально-поведенческий аспектов. Ключевым методологическим вызовом остается разработка эффективных процедур валидации, обеспечивающих возможность надежной оценки качества моделей при переносе в реальные контексты использования. Также стоит отметить, что наиболее перспективными методами для исследования являются методы, использующие пространственно-временные свойства данных. С точки зрения используемых архитектур нейтронных сетей наиболее перспективными для исследования и улучшения сегодня являются методы на основе Visual Transformer. Также улучшение подходов и алгоритмов Domain Adaptation на основе современных продвинутых генеративных сетей (таких, как например, диффузионных моделей) выглядит перспективным с точки зрения приближения искусственных данных к настоящим.

Несмотря на отмеченные ограничения, проведенное исследование вносит значимый вклад в развитие предметной области. Систематизация и количественный анализ больших массивов разрозненных эмпирических данных позволили получить целостную картину современного состояния проблемы и выявить ключевые тенденции ее развития.

Заключение

Проведенное исследование методов по распознанию человеческой активности позволило определить наиболее перспективные направления для дальнейшего исследования. Также выделены основные ограничения существующих методов, что позволяет более эффективно подготавливать данные для обучения моделей, а также выделить направления для улучшения архитектуры современных нейронах сетей.

В прикладной перспективе, полученные результаты создают основу для разработки интеллектуальных систем поддержки принятия решений в таких областях, как медицинская диагностика, безопасность, управление персоналом, маркетинговые исследования. Высокая точность разработанных моделей в сочетании с возможностью предсказания ключевых паттернов открывает качественно новые горизонты для анализа и выявления общих характеристик людей на основе изображений и видеоданных.

ЛИТЕРАТУРА

1. Doyran, M., Yıldırım, Y., & Salah, A. (2017). Action recognition with deep neural networks. In 2017 25th Signal Processing and Communications Applications Conference (SIU) (pp. 1–4).
2. Yao, A., & Huang, Z. (2019). Deep Learning for Human Activity Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence.
3. Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 221–231.
4. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. Sensors, 19(5), 1005.
5. Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. Image and vision computing, 60, 4–21.
6. Beddiar, D.R., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: a survey. Multimedia Tools and Applications, 79(41), 30509–30555.
7. Singh, S., Arora, C., & Jawahar, C. V. (2016, December). First person action recognition using deep learned descriptors. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2620–2628).
8. Kong, Y., & Fu, Y. (2018). Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230.
9. Wei, P., Xie, D., Zheng, N., & Zhu, S. C. (2017). Inferring human attention by learning latent intentions. In IJCAI (pp. 1297–1303).
10. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Suleyman, M. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.
11. Chaquet, J.M., Carmona, E.J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding, 117(6), 633–659.
12. Han, J., Zhang, Z., & Wang, K. (2015). Action recognition with multiscale spatio-temporal contexts. IEEE transactions on circuits and systems for video technology, 26(3), 484–496.
13. Zhu, G., Zhang, L., Shen, P., & Song, J. (2017). Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access, 5, 4517–4524.
14. Asadi-Aghbolaghi, M., Clapés, A., Bellantonio, M., Escalante, H. J., Ponce-López, V., Baró, X., ... & Escalera, S. (2017). A survey on deep learning based approaches for action and gesture recognition in image sequences. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017) (pp. 476–483). IEEE.
15. Zeng, M., Gao, H., Yu, T., Mengshoel, O.J., Langseth, H., Lane, I., & Liu, X. (2018). Understanding and improving recurrent networks for human activity recognition by continuous attention. In Proceedings of the 2018 ACM International Symposium on Wearable Computers (pp. 56–63).
16. Peng Wang, Fanwei Zeng, Yuntao Qian (2023) A Survey on Deep Learning-based Spatio-temporal Action Detection. International Journal of Wavelets, Multiresolution and Information Processing
17. Faure GJ, Chen MH, Lai SH. Holistic Interaction Transformeк Network for Action Detection. In Proc. IEEE Winter Conf. Appl. Comput. Vis., 2023.
18. Chen L, Tong Z, Song Y, Wu G, Wang L. (2023) Efficient Video Action Detection with Token Dropout and Context Refinement. Proceedings of the IEEE/CVF International Conference
19. Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, & Christoph Feichtenhofer. (2020). Audiovisual SlowFast Networks for Video Recognition.
20. Yuyang Wanyan, Xiaoshan Yang, Weiming Dong, & Changsheng Xu. (2024). A Comprehensive Review of Few-shot Action Recognition.
21. Fu, Y., Zhang, L., Wang, J., Fu, Y., & Jiang, Y. (2020). Depth Guided Adaptive Meta-Fusion Network for Few-shot Video Recognition. Proceedings of the 28th ACM International Conference on Multimedia.
22. Reddy, A., Shah, K., Paul, W., Mocharla, R., Hoffman, J., Katyal, K., Manocha, D., Melo, C., & Chellappa, R. (2023). Synthetic-to-Real Domain Adaptation for Action Recognition: A Dataset and Baseline Performances. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11374–11381).

Журнал «Современная наука: актуальные проблемы теории и практики»