УДК 004.896 https://doi.org/10.26160/2541-8637-2025-15-62-64

ГИБРИДНЫЙ ПОДХОД К РАСПОЗНАВАНИЮ ДЕЙСТВИЙ ЧЕЛОВЕКА-ОПЕРАТОРА В КОЛЛАБОРАТИВНЫХ РОБОТИЗИРОВАННЫХ СРЕДАХ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ И КОМПЬЮТЕРНОГО ЗРЕНИЯ

Грабарь Д.М., Иванов Ю.С.
Комсомольский-на-Амуре государственный университет, Комсомольск-на-Амуре

Ключевые слова: определение действий, LLM, ключевые точки, роботизированные системы, коботы, искусственный интеллект.
Аннотация. Предложен гибридный подход по распознаванию действий человека-оператора в коллаборативных роботизированных средах, сочетающий методы компьютерного зрения и большие языковые модели. Разработана модифицированная метрика WSAA, позволяющая оценивать полученные результаты с учетом точности и адаптивности модели. Наилучшие результаты показала модель Llama3.2-Vision, продемонстрировав высокую точность распознавания и устойчивость к изменениям в условиях окружающей среды. Предложенный подход может быть применен для повышения эффективности взаимодействия человека и робота в промышленных условиях.

A HYBRID APPROACH TO RECOGNIZING HUMAN OPERATOR ACTIONS IN COLLABORATIVE ROBOTIC ENVIRONMENTS USING LARGE LANGUAGE MODELS AND COMPUTER VISION

Grabar D.M., Ivanov Y.S.
Komsomolsk-on-Amur State University, Komsomolsk-on-Amur

Keywords: action detection, LLM, keypoints, robotics system, cobot, artificial intelligence.
Abstract. A hybrid approach for recognizing human operator actions in collaborative robotic environments is proposed, combining computer vision methods and large language models. A modified WSAA metric has been developed, which makes it possible to evaluate the results obtained taking into account the accuracy and adaptability of the model. The Llama3.2-Vision model showed the best results, demonstrating high recognition accuracy and resistance to changes in environmental conditions. The proposed approach can be applied to improve the efficiency of human-robot interaction in industrial environments.

Введение

Современные методы распознавания действий человека в промышленных коллаборативных средах основаны на алгоритмах компьютерного зрения и машинного обучения, таких как SlowFast Networks [1] и TimeSformer [2]. Однако эти методы требуют больших объемов размеченных данных и имеют ограниченную способность к обобщению в изменяющихся условиях. Использование больших языковых моделей (LLM) для автоматической аннотации видеоданных представляется перспективным направлением, но существующие подходы, такие как CLIP [3], VideoBERT [4] и ActBERT [5], имеют недостатки, включая высокую вычислительную сложность и неточности в генерации аннотаций. В работе предлагается гибридный подход, сочетающий методы компьютерного зрения и семантические возможности LLM, для

повышения точности и интерпретируемости распознавания действий человека-оператора.

Основная часть

Предлагается гибридный подход по распознаванию действий человека-оператора в рабочей зоне коллаборативного робота, который сочетает методы компьютерного зрения и семантические возможности LLM для распознавания действий человека-оператора в коллаборативных роботизированных средах. Основой метода является детекция ключевых точек с использованием алгоритмов компьютерного зрения, таких как YOLO11-POSE [6], и последующий семантический анализ с помощью LLM. Это позволяет генерировать текстовые описания атомарных действий, таких как «наклон», «поворот» и «подъем», на основе динамики изменений координат и углов суставов, даже в условиях частично перекрытых точек. Для адаптации LLM к предметной области разработан структурированный промпт, который формализует правила классификации действий на основе ключевых точек. Для оценки точности предсказаний предложена модифицированная метрика Weighted Semantic Action Accuracy (WSAA), которая объединяет семантическое сходство на основе BERT-эмбеддингов и экспертные правила из предопределенной матрицы соответствий:

$$\mathrm{WSAA} = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{\mathrm{attr}} w_{\mathrm{attr}} \cdot s(\mathrm{gt}_t(\mathrm{attr}), \mathrm{pred}_t(\mathrm{attr}))}{\sum_{\mathrm{attr}} w_{\mathrm{attr}}}, \quad (1)$$

где N – общее количество данных, $\mathrm{gt}_t(\mathrm{attr})$ и $\mathrm{pred}_t(\mathrm{attr})$ – истинное и предсказанное действия для атрибута attr в момент времени t, а $s(\mathrm{gt}_t(\mathrm{attr}), \mathrm{pred}_t(\mathrm{attr}))$ – сходство между этими действиями.

Метрика корректирует оценки для специфичных промышленных действий, таких как «смещение торса» или «сгибание локтя», что повышает точность валидации предсказаний. Эксперименты проводились (рис. 1) с использованием мультикамерной системы, состоящей из трех камер Intel RealSense, и показали, что наилучшие результаты достигаются с моделью Llama3.2-Vision ($\mathrm{WSAA} = 0.7438$), которая эффективно анализирует движения рук и головы. Однако выявлены ограничения при анализе резких движений и частичных окклюзий, что требует дальнейшей интеграции алгоритмов интерполяции ключевых точек.

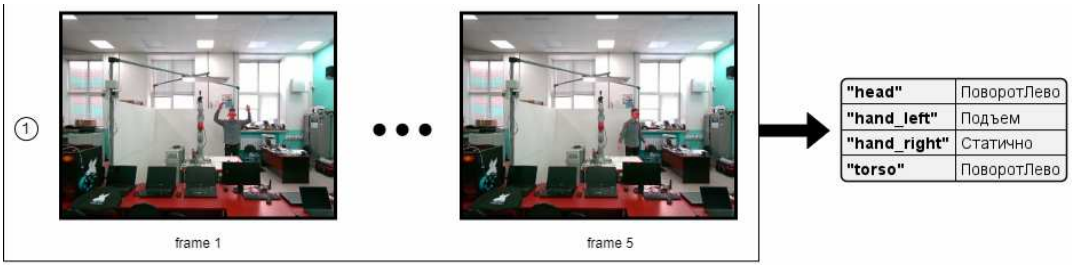

Рис. 1. Результат эксперимента

Предложенный подход позволяет интерпретировать действия человека-оператора в реальном времени, обеспечивая безопасное и эффективное взаимодействие между человеком и коллаборативным роботом. Разработанные

методы служат основой для создания адаптивных коллаборативных систем, способных анализировать как атомарные действия, так и их технологические последовательности, что открывает новые перспективы для автоматизации промышленных процессов.

Заключение

Проведенное исследование демонстрирует эффективность гибридного подхода, сочетающего методы компьютерного зрения и семантические возможности LLM, для детектирования действий человека-оператора в промышленных коллаборативных средах. Научный вклад работы заключается в разработке уникального структурированного промпта и модифицированной метрики WSAA, которые устраняют зависимость от объемных размеченных датасетов и повышают точность валидации предсказаний. Результаты подтверждают потенциал LLM в задачах промышленного мониторинга, особенно при использовании структурированных дескрипторов позы и модифицированных метрик оценки.

Финансирование. Работа выполнена при поддержке Российского научного фонда (проект № 22-71-10093) [7].

Список литературы

1. Feichtenhofer C., Fan H., Malik J., He K. Slowfast networks for video recognition // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 6202-6211.
2. Bertasius G., Wang H., Torresani L. Is space-time attention all you need for video understanding? // ICML. 2021, vol. 2, no. 3, p. 4.
3. Metzger C. Training CLIP models on Data from Scientific Papers // arXiv preprint arXiv:2311.04711. – 2023.
4. Sun C., Myers A., Vondrick C., Murphy K., Schmid C. Videobert: A joint model for video and language representation learning // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 7464-7473.
5. Zhu L., Yang Y. Actbert: Learning global-local video-text representations // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 8746-8755.
6. Khanam R., Hussain M. Yolov11: An overview of the key architectural enhancements // arXiv preprint arXiv:2410.17725. – 2024.
7. Разработка и синтез перспективных мультимодальных адаптивных алгоритмов и методов управления поведением коллаборативных робототехнических систем с учетом нештатных ситуаций и экстремальных условий в недетерминированной среде [Электронный ресурс]. – Режим доступа: https://rscf.ru/project/22-71-10093/.

Сведения об авторах:
Грабарь Даниил Михайлович – аспирант;
Иванов Юрий Сергеевич – к.т.н., доцент.