Exploring Vision Language Models for Egocentric Action Localization

Valentin Knoben[1,2], Julia Kramme[1], Björn Hein[1], Christian Wurll[1]

[1]Institute for Robotics and Autonomous Systems, Karlsruhe University of Applied Sciences
[2]Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology
{valentin.knoben, julia.kramme, bjoern.hein, christian.wurll}@h-ka.de

Abstract. Context-aware systems can support humans at work by automatically performing quality control, providing assistance, or generating instructions and documentation for latter use. However, the adaptation of such intelligent systems to custom use cases demands training data, expertise, and effort. With the dissem-ination of Vision Language Models (VLMs), recognition capabilities are becoming more accessible. We explore the use of readily available VLMs for understanding egocentric video footage of common manual tasks in production environments. Results demonstrate the feasibility of using VLMs in such contexts.

Keywords: Vision language model, action recognition, temporal localization

1 Introduction

In production-related environments, recognizing a worker's actions enables online quality checks [1] and real-time assistance [2]. Apart from simply understanding a performed task, extracting temporal information further supports the documentation of a workflow or the creation of digital instructions, e.g., for augmented reality-based manuals [3]. Research has particularly explored exocentric scene understanding, for instance, to enhance human robot interaction [4]. With the ongoing advancement of head-worn devices and their application in work environments [5], egocentric input data is becoming a potential source of information, too. Yet, reliably integrating such contextual understanding into a system requires expertise and effort in training a model on custom data or hiring a dedicated system integrator which are key barriers faced by Small and Medium-sized Enterprises (SMEs) in particular [6]. With the dissemination of Vision Language Models (VLMs) such as ChatGPT, Gemini, or Claude which can handle multimodal input, the barrier to entry has been reduced significantly in that regard. Therefore, we investigate the feasibility of using an off-the-shelf VLM to recognize and temporally locate common, production-related actions from videos. We focus on short clips showing a single action which might be performed as part of a more comprehensive workflow.

2 Methods

We recorded egocentric videos at different resolutions using the Microsoft HoloLens 2. Some VLMs can handle video input directly. However, internally the video input is usually sampled at 1fps which is insufficient for understanding short action clips. Therefore, we streamed footage to a near-edge unit at 30fps and extracted frames at 3fps. Frames were numbered to provide temporal anchors [7] (Fig. 1) and sent with a prompt asking the

VLM to return the action type and key frame(s) as structured JSON. The candidate actions were: pick-and-place, use wrench/Allen/screwdriver, press button, rotate switch, open/close element, fasten/loosen by hand. For each action we recorded three scenarios, provided ground truths for the correct action type and eligible key frames, and ran five trials each, resulting in 15 data points per action type. After initially testing different VLMs, we proceeded with gpt-4.1.

Fig. 1. Exemplary frames extracted from an egocentric image stream, fed to VLM. Returned json output: *{"action type"*: "rotate switch", *"key frames"*: "5, 6, 7"*}*

3 Results and Discussion

Fig. 2 illustrates the number of correct action recognitions and temporal localizations per action type, revealing high correct recognition counts alongside much more variable frame-level localizations. We identified two common failure modes. In case of incorrectly recognized action types, actions were simply confused, e.g., rotating a switch with pressing a button, potentially due to bad lighting conditions and both elements being visible in the scene. Causes of incorrect temporal localization are the strict classification as incorrect even when off by only 1 frame in ambiguous scenarios and the confusion between object visibility and its actual engagement. Apart from that, a temperature of around 0.3 and top p of 1 produced concise, low-hallucination outputs. Action recognition degraded noticeably below a resolution of 1504×846, while key-frame localization remained robust at lower resolutions, consistent with recognition needing fine spatial detail such as object geometry whereas localization relies on coarser motion cues. Increasing resolution beyond 1500 pixels in width gave diminishing returns, likely because VLMs tile large images to fixed sizes, fragmenting objects and limiting cross-tile context [8]. Possible leverage points for improvement are to perform hand- or object-based cropping of frames or concatenate multiple frames into a single image to reduce the number of pixels or frames to evaluate. However, this must carefully be balanced with recognition quality which could be aided by dynamic instead of fixed sampling rates.

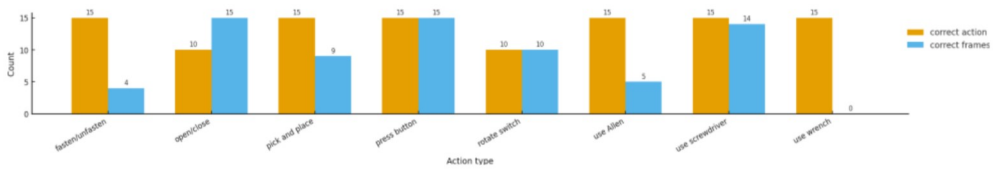

Fig. 2. Correct action recognitions (orange) and temporal localizations (blue) per action type

Acknowledgments. This research was conducted within The Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) which is funded by the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) under Grant BW6 03.

References

1. Islam, M.R., Zamil, M.Z.H., Rayed, M.E., Kabir, M.M., Mridha, M., Nishimura, S., Shin, J.: Deep learning and computer vision techniques for enhanced quality control in manufacturing processes. IEEE Access (2024)
2. Li, W., Xu, A., Wei, M., Zuo, W., Li, R.: Deep learning-based augmented reality work in-struction assistance system for complex manual assembly. Journal of Manufacturing Systems **73** (April 2024) 307–319
3. Klimant, P., Kollatsch, C.: Concepts for creating augmented reality based technical docu-mentations for the maintenance of machine tools. International Journal on Interactive Design and Manufacturing (IJIDeM) **16**(2) (June 2022) 765–773
4. Ji, Y., Yang, Y., Shen, F., Shen, H.T., Li, X.: A survey of human action analysis in hri applications. IEEE Transactions on Circuits and Systems for Video Technology **30**(7) (2020) 2114–2128
5. Fang, W., Chen, L., Zhang, T., Chen, C., Teng, Z., Wang, L.: Head-mounted display aug-mented reality in manufacturing: A systematic review. Robotics and Computer-Integrated Manufacturing **83** (October 2023) 102567
6. Oldemeyer, L., Jede, A., Teuteberg, F.: Investigation of artificial intelligence in SMEs: a sys-tematic review of the state of the art and the main implementation challenges. Management Review Quarterly **75**(2) (June 2025) 1185–1227
7. Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., Ikeuchi, K.: Open-vocabulary action localization with iterative visual prompting (August 2024)
8. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Zhang, S., Duan, H., Zhang, W., Li, Y., Yan, H., Gao, Y., Chen, Z., Zhang, X., Li, W., Li, J., Wang, W., Chen, K., He, C., Zhang, X., Dai, J., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2-4khd: a pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Volume 37 of NIPS '24., Red Hook, NY, USA, Curran Associates Inc. (June 2025) 42566–42592