Egocentric Human-Object Interaction in Industrial Scenarios: a Case Study

Marco Moltisanti[1,a], Ketty Cantone[1], Antonino Lopes[1], Emanuele Ragusa[1], Rosario Leonardi[2], Francesco Ragusa[2], Antonino Furnari[2], Giovanni Maria Farinella[2]

[1] *Xenia Progetti srl, Acicastello, Italy*
[2] *University of Catania, Catania, Italy*

Abstract

This study presents a system for real-time monitoring of personal protective equipment (PPE) compliance using egocentric vision. By analyzing hand-object interactions captured from wearable camera footage, the system detects glove usage during high-risk tasks. A combination of real and synthetic data improves model robustness, while task-specific business logic enables context-aware safety alerts.

Keywords
Egocentric vision, PPE compliance, Human-Object Interaction

1. Introduction

Ensuring compliance with personal protective equipment (PPE) protocols is a fundamental priority in industrial and laboratory environments to safeguard workers against potential hazards. PPE such as gloves, helmets, and safety goggles are essential in mitigating the risk of injuries caused by physical, chemical, or electrical hazards. However, enforcing strict adherence to PPE regulations remains a challenge, often relying on manual supervision, which is costly and subject to human error. Near misses — i.e. incidents where safety protocols are violated but no injury occurs — serve as critical indicators of potential safety failures and warrant timely detection and intervention.

Recent advancements in wearable technology and computer vision have opened new avenues for automated monitoring of operator behavior. Specifically, wearable cameras provide a first-person viewpoint that can capture the operator's actions and interactions with tools and equipment. Egocentric Human-Object Interaction (EHOI) analysis focuses on understanding these interactions from the wearer's perspective, providing valuable contextual information that can be used to assess compliance with safety protocols.

In this work, we present a case study focused on monitoring glove usage during high-risk activities in an indoor laboratory environment. Operators wear smart glasses (Microsoft HoloLens 2), and a deep learning-based system analyzes the video stream to detect whether gloves are worn during interactions with hazardous tools such as welders and electric drills. Our approach integrates visual detection with temporal reasoning to identify near misses, where gloves are absent during critical interactions.

[a] Corresponding author: mmoltisanti@xeniaprogetti.it (Marco Moltisanti)
Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

Our approach builds upon a multimodal dataset named ENIGMA-51 [6], designed for egocentric human-object interaction analysis in industrial environments. To overcome data limitations and improve model robustness, we extend a synthetic image generation pipeline [1] that simulates diverse PPE compliance conditions.

This paper is structured as follows: Section 2 reviews related work in egocentric vision and PPE detection. Section 3 describes the multimodal dataset used for training and evaluation. Section 4 discusses the adaptation of the synthetic image generator for data augmentation. Section 5 presents the business logic implemented for alert generation and system integration. Conclusions and possible future works are outlined in Section 6.

2. Related Work

Egocentric vision has emerged as a dynamic research area, thanks to the growth of wearable cameras and the increasing need for context-aware assistive and monitoring systems. Human-object interaction recognition from a first-person viewpoint offers unique challenges, including occlusions, varying viewpoints, and dynamic backgrounds, but also enables fine-grained analysis of user behavior.

One of the main contributions to the field is the EPIC-KITCHENS dataset [2], which provides an extensive collection of annotated egocentric videos capturing everyday activities in kitchen environments. The dataset has significantly advanced research in object detection, hand segmentation, and action recognition within egocentric vision.

In the industrial domain, the MECCANO dataset introduced by Ragusa et al. [3] represents a significant advancement for egocentric vision applications focused on human-object interaction. MECCANO comprises annotated first-person videos capturing assembly tasks performed by operators in controlled industrial-like settings. The dataset includes detailed annotations for hands, tools, and objects, making it particularly relevant for understanding hand-tool interactions.

In industrial safety, the detection of PPE has predominantly relied on third-person camera perspectives. For instance, Quattrocchi et al. [4] employed convolutional neural networks to classify the presence of helmets and gloves on workers in factory settings, demonstrating promising results in PPE compliance detection.

Ego4D [**5**] represents one of the most ambitious efforts in egocentric data collection, with over 3,500 hours of video captured from 931 subjects in daily activities. It offers 3,025 hours of daily activity video covering hundreds of scenarios (home, outdoor, workplace, leisure, etc.) captured by 855 unique camera wearers from 74 locations worldwide and 9 different countries. Part of the video is accompanied by audio, 3D meshing of the environment, gaze tracking, stereo and/or synchronised video from multiple egocentric cameras in the same event. Furthermore, a series of novel benchmark challenges are presented, focusing on understanding first-person visual experience in the past (interrogating episodic memory), present (analysing hand-object manipulation, audio-visual conversations and social interactions) and future (predicting activities).

3. Multimodal Dataset

The dataset employed in this study is the ENIGMA-51 dataset [6], a comprehensive multimodal egocentric dataset specifically collected in an industrial environment to facilitate the study of human behavior during complex manipulation tasks involving industrial tools.

The dataset comprises 51 egocentric video sequences recorded from smart glasses (Microsoft Hololens 2) worn by 19 different operators performing repair and assembly tasks on electrical boards in a controlled indoor laboratory setting. The total recorded footage amounts to approximately 22 hours, providing a rich source of data for training and evaluating vision-based models. ENIGMA-51 offers extensive annotations that include:

- RGB video streams at 30 frames per second;
- Segmentation masks;
- Bounding boxes and class labels for manipulated objects and tools;
- Hand-Object Interaction annotations;
- Hand keypoints;

These rich multimodal annotations provide an ideal foundation for supervised training of deep learning models aimed at detecting glove usage during interaction with high-risk tools such as welders and electric drills.

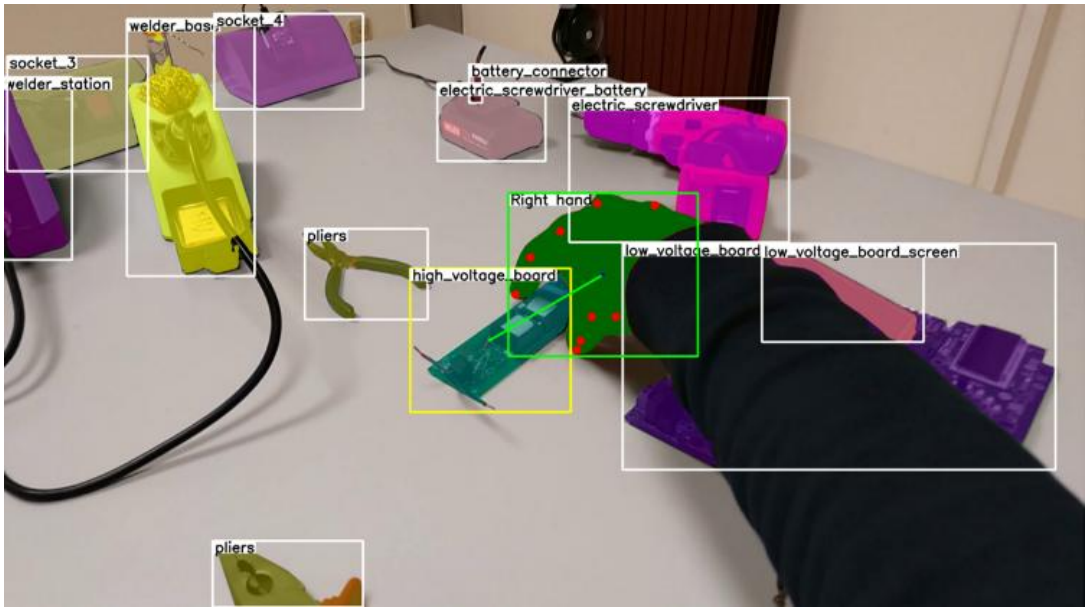

Figure 1. Sample of multimodal annotated images in ENIGMA-51.

By leveraging ENIGMA-51, our study benefits from real-world industrial scenarios and highly detailed labeling, which are crucial for developing robust and generalizable safety monitoring systems in egocentric vision applications.

4. Synthetic Data Generator

To improve the generalization capabilities of our models, we utilized a synthetic image generation framework. This framework, initially proposed by Leonardi et al. [1], is designed to produce photorealistic synthetic egocentric images simulating human-object interactions in controlled virtual environments.

In this work, we extended the simulator to generate hand keypoints annotations and simulate various PPE configurations, including correct and incorrect usage of gloves. This enhancement enables the creation of rich, labeled datasets that support training and evaluation of models for hand pose estimation and PPE recognition.

The synthetic data generated (Figure 2) was leveraged for pretraining deep convolutional neural networks focused on hand segmentation, object detection, and PPE recognition. Subsequently, a domain adaptation phase was performed through fine-tuning on the real ENIGMA-51 dataset to reduce the sim-to-real gap, as demonstrated by Leonardi et al. [1].

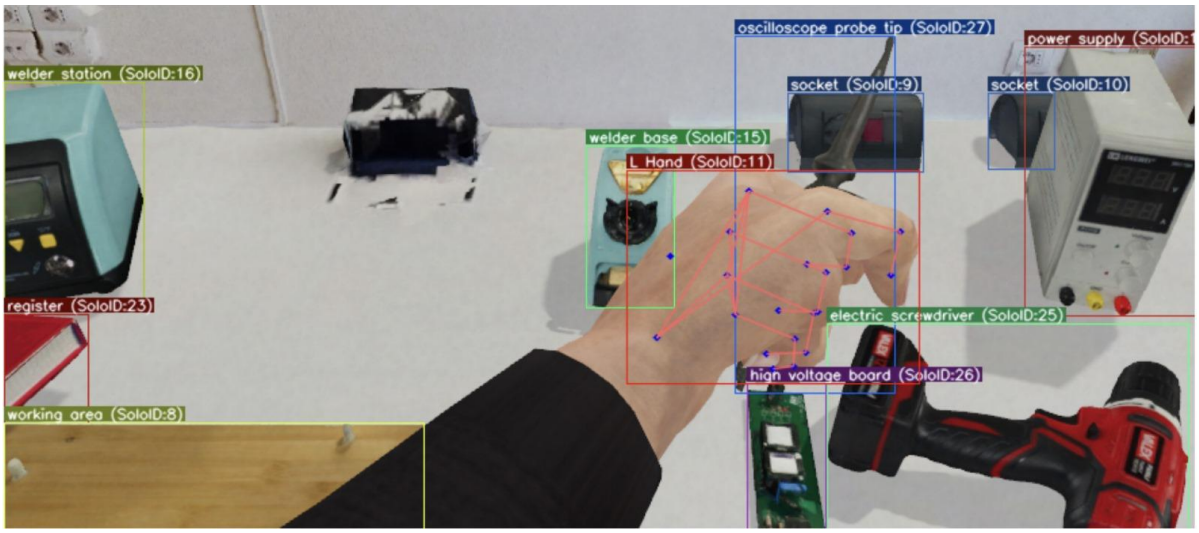

Figure 2. Sample of a synthetically generated frame with interaction.

5. Business Logic

The deployment of the PPE compliance detection system incorporated a robust business logic layer implemented in Python. The system architecture employs a publish/subscribe (pub/sub) messaging framework based on the MQTT protocol, which facilitates scalable, low-latency communication between inference components and alerting modules.

Model inferences—comprising predictions on glove presence and object identity during each frame—are published to designated MQTT topics. The business logic module subscribes to these topics and processes incoming streams using rule-based logic designed to ensure reliable detection and alert generation.

The core elements of the business logic include:

- PPE Presence Monitoring: Continuous evaluation of glove detection results during frames where tool interaction occurs.

- Temporal Smoothing: Application of a sliding window filter to aggregate recent predictions and filter out transient misclassifications caused by motion blur, occlusions, or sensor noise. This temporal consistency check reduces false alarms by requiring sustained evidence of PPE violations before triggering alerts.

- Contextual Filtering: Differentiation of tool types, such as welders and electric drills, to apply specific PPE requirements and tailor alert thresholds accordingly.

When a near miss event—defined as a confirmed interaction with a hazardous tool absent the required gloves—is detected, an alert is generated and dispatched. The alert delivery mechanism supports multiple channels simultaneously, such as visual, textual or auditory.

The MQTT-based pub/sub infrastructure enables seamless integration of additional alert recipients or modalities as required by the industrial setting. This modular and extensible design ensures that the system can adapt to evolving safety policies and operational needs.

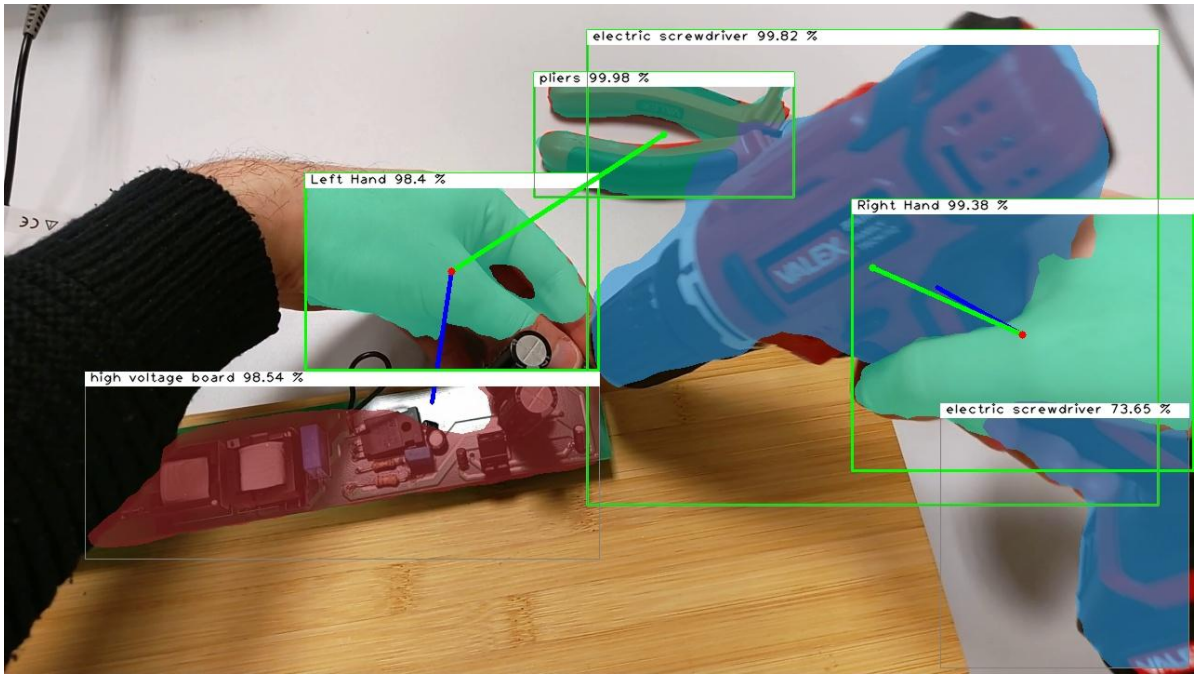

Figure 3. Qualitative results of interaction detection using an electric screwdriver.

6. Conclusions

This study demonstrates the feasibility and effectiveness of using egocentric human-object interaction detection for real-time PPE compliance monitoring in industrial environments. Leveraging a multimodal dataset (ENIGMA-51), augmented with synthetic data generation, we developed robust deep learning models capable of detecting glove usage during tool interactions from a first-person viewpoint.

Future work will explore extending the system to additional PPE types and more complex industrial scenarios. Further improvements in domain adaptation and real-time inference efficiency will also be pursued to enhance practical applicability.

Acknowledgements

Piano Nazionale Ripresa e Resilienza (PNRR) Missione 4, “Istruzione e Ricerca” - Componente 2, “Dalla Ricerca all’impresa” - Linea di investimento 1.3, finanziato dall’Unione Europea – NEXTGENERATIONEU”, Progetto “Future Artificial Intelligence – Fair” PE0000013, Cup J53C22003010006

Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-4 in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella (2024). Exploiting Multimodal Synthetic Data for Egocentric Human-Object Interaction Detection in an Industrial Scenario. Computer Vision and Image Understanding (CVIU)

[2] D. Damen, H. Doughty, G. M. Farinella, et al., “Scaling Egocentric Vision: The EPIC-KITCHENS Dataset,” in Proc. European Conference on Computer Vision (ECCV), 2018.

[3] F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding 2023.

[4] C. Quattrocchi, D. Di Mauro, A. Furnari, A. Lopes, M. Moltisanti, G. M. Farinella (2023). Put Your PPE On: A Tool for Synthetic Data Generation and Related Benchmark in Construction Site Scenarios . In International Conference on Computer Vision Theory and Applications (VISAPP).

[5] Grauman, K., Westbury, A., Byrne, E., Chavis, Z.Q., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Fuegen, C., Gebreselasie, A., Gonz´ alez, C., Hillis, J.M., Huang, X., Huang, Y., Jia, W., Khoo, W.Y.H., Kol´ ar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K.K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhu, Y., Arbel´ aez, P., Crandall, D.J., Damen, D., Farinella, G.M., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R.A., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J., (2021). Ego4d: Around the world in 3,000 hours of egocentric video, in: Conference on Computer Vision and Pattern Recognition, pp. 18995–19012.

[6] F. Ragusa, R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, A. Furnari, G. M. Farinella (2024). ENIGMA-51: Towards a Fine-Grained Understanding of Human-Object Interactions in Industrial Scenarios . In IEEE Winter Conference on Application of Computer Vision (WACV)