

# Egocentric Industrial Human–Object Interaction Recognition Based on Hand Interaction Point Modeling

Yen-Chun Chen,<sup>1\*</sup> Tsai-Yu Yao,<sup>2</sup> and Po-Yu Chien<sup>3</sup>

<sup>1,3</sup>Department of Intelligent Commerce, National Kaohsiung University of Science and Technology,  
Kaohsiung 82444, Taiwan

<sup>2</sup>Kang Chiao International School: Taipei Campus, Taipei 231049 Taiwan

(Received June 4, 2026; accepted September 1, 2026)

**Keywords:** egocentric vision, industrial human–object interaction recognition, hand interaction point modeling, hand segmentation, human–object interaction analysis

In intelligent manufacturing environments, first-person view imagery has become increasingly important for equipment maintenance, operation monitoring, and intelligent assistance systems. However, frequent hand occlusions, small interaction targets, and complex operational procedures make fine-grained human–object interaction (HOI) recognition challenging. In this paper, we propose an egocentric industrial HOI recognition method based on hand interaction point modeling, which integrates hand segmentation, object detection, a minimum-distance interaction search mechanism, and temporal analysis to establish hand–object associations and recognize interaction events. Experiments were conducted on the ENIGMA-51 industrial egocentric dataset and evaluated using the Egocentric HOI Detection task. Experimental results demonstrate that the proposed contour-based interaction point modeling significantly improves the HOI recognition performance, achieving 59.60% average precision for the hand and contact state (AP H.+State), 78.60% mean average precision for the hand and object state (mAP H.+Obj), and 45.63% mean average precision for the hand, object, and contact state (mAP H.+All), outperforming the conventional center-based interaction modeling approach. The results demonstrate that contour-based interaction modeling more accurately captures real hand–object contact relationships and provides more reliable interaction cues for industrial HOI recognition, offering an effective solution for intelligent factory operation comprehension and standard operating procedure analysis.

## 1. Introduction

In modern manufacturing environments, equipment operation and maintenance activities are generally performed in accordance with strict standard operating procedures (SOPs) to ensure product quality, equipment reliability, and workplace safety. However, even when comprehensive operating guidelines and training programs are provided, human errors may still occur owing to differences in experience, accumulated fatigue, distraction, or environmental disturbances. Such

---

\*Corresponding author: e-mail: [alexchen@nkust.edu.tw](mailto:alexchen@nkust.edu.tw)  
<https://doi.org/10.18494/SAM6464>

errors can not only reduce operational efficiency and product quality but may also result in equipment damage, personal injury, and even serious industrial accidents. Therefore, the development of intelligent and proactive safety management mechanisms has become an important research topic in the fields of smart manufacturing and industrial safety.<sup>(1)</sup> With the rapid advancement of wearable sensing technologies, various sensing devices have been introduced into occupational monitoring and safety management applications. For example, smartphones equipped with accelerometers and gyroscopes can be utilized to track human motion patterns and posture changes.<sup>(1)</sup> Smartwatches can monitor workers' physiological conditions through heart-rate and motion sensors. Furthermore, smart glasses, which provide real-time image acquisition and information visualization capabilities, have emerged as a promising platform for both operational assistance and workplace monitoring.<sup>(1)</sup> Among these wearable devices, smart glasses are capable of directly capturing first-person-view (FPV) visual information from the user's perspective.<sup>(2)</sup> Such visual data not only provide a complete record of the operator's observations and actions but also accurately reflect interactions among workers, equipment, tools, and the surrounding environment. Consequently, FPV-based monitoring systems offer significant potential for real-time observation, behavior analysis, and the safety supervision of industrial operations.

However, in FPV imagery, accurately characterizing the contact relationship between hands and manipulated objects remains a challenging task owing to frequent hand occlusions, rapid movements, and continuous viewpoint changes. This research area in computer vision is commonly referred to as egocentric or first-person vision. Although early studies have explored FPV-based visual understanding,<sup>(3)</sup> existing approaches primarily rely on hand keypoint,<sup>(4)</sup> bounding boxes,<sup>(5)</sup> or segmentation masks<sup>(6)</sup> to determine interaction relationships. In industrial operating environments, however, these methods often fail to precisely represent actual contact locations and operator intentions. To address this limitation, we propose an egocentric industrial human–object interaction (HOI) recognition method based on hand interaction point modeling (HIPM), aiming to improve the accuracy of hand–object interaction analysis and our understanding of operation behavior. To facilitate research in industrial FPV applications, several public datasets have been introduced in recent years, including EPIC-KITCHENS,<sup>(7)</sup> Ego4D,<sup>(7)</sup> and Assembly101.<sup>(7)</sup> However, most of these datasets focus on kitchen activities, daily-life scenarios, or assembly tasks, which differ significantly from real-world industrial maintenance environments. Because of safety regulations, privacy concerns, and proprietary information restrictions, acquiring data from actual industrial settings remains difficult. Consequently, many existing studies rely on simulated equipment or simplified environments, limiting their ability to fully capture the characteristics of industrial operations. The ENIGMA-51 dataset<sup>(4)</sup> was developed to bridge this gap by simulating realistic industrial maintenance scenarios. It contains recordings of participants performing electrical control board maintenance tasks while interacting with equipment such as electric screwdrivers, power supplies, oscilloscopes, and control buttons. In addition, the dataset provides detailed annotations for hands, objects, and interaction relationships, making it an important benchmark for egocentric HOI research in industrial environments. On the basis of these fine-grained annotations, the proposed method is positioned as the perception layer within the overall SOPs

analysis framework. By leveraging HIPM, the proposed framework recognizes hand–object contact relationships, manipulated objects, and interaction events such as contact, take, and release, thereby providing the low-level semantic information required for subsequent procedure understanding, intention recognition, and intelligent operation assistance. Although the proposed method does not directly perform complete SOPs recognition, it establishes a fundamental perception module that provides critical low-level semantic cues for downstream procedure analysis, procedure understanding, and intelligent operation assistance.

To address the limitations of existing HOI recognition methods in industrial environments, we propose a procedure-aware egocentric industrial HOI recognition framework based on HIPM and a temporal voting mechanism. HIPM samples hand contour points extracted from hand segmentation and estimates hand–object interactions by calculating the minimum distance to detected objects, enabling a more accurate recognition of contact and release events than conventional center-based approaches. A temporal voting mechanism further exploits interaction consistency across consecutive frames to reduce recognition errors caused by occlusion and rapid hand movements. Experimental results on the ENIGMA-51 industrial FPV dataset demonstrate that the proposed framework effectively improves the robustness and accuracy of industrial HOI recognition, providing a practical solution for intelligent factory operation understanding and standard operating procedure analysis.

## 2. Related Work

Research on human behavior understanding in FPV scenarios has primarily focused on tasks such as action recognition and action localization, where data-driven approaches are employed to learn spatiotemporal features from visual data for identifying user activities. However, compared with the rapid progress in action recognition, HOI detection remains a relatively underexplored research area. In egocentric vision, hands serve not only as the primary medium through which operators perform tasks but also as a critical bridge connecting human actions with surrounding objects. Therefore, accurately recognizing hand–object interaction relationships is essential for operation understanding, procedure analysis, and future action anticipation. Existing studies can generally be categorized into two groups depending on the visual features employed: global-feature-based and local-feature-based methods. Global-feature-based methods extract information from the entire image to capture the overall scene context and activity patterns, whereas local-feature-based methods focus on specific regions of interest, such as hands, manipulated objects, and interaction areas, to model fine-grained interaction details. Both methods have demonstrated effectiveness in FPV-based behavior understanding; however, accurately capturing hand–object interaction relationships in complex industrial environments remains a challenging problem.

### 2.1 Global features

Global features primarily utilize holistic scene information for behavior analysis, effectively reducing preprocessing complexity while improving recognition efficiency. However, in

scenarios involving fine-grained manipulation tasks, relying solely on global information is often insufficient to fully capture the interaction details between operators and objects. In contrast, local features focus on specific human-related regions, such as hands, head movements, gaze information, and manipulated objects, enabling a more accurate modeling of critical actions and interaction behaviors during task execution. Consequently, local-feature-based methods have gradually become an important research direction in FPV behavior understanding. Regarding global feature analysis, in early studies, entire video frames were directly employed as inputs for action recognition. Among such methods, the Two-Stream Network simultaneously learned spatial and temporal representations from RGB images and optical flow, significantly improving video action recognition performance.<sup>(8)</sup> Subsequently, the SlowFast Network adopted a dual-pathway architecture with different temporal resolutions to capture both semantic and motion information, reducing the computational cost of optical flow while maintaining strong recognition performance.<sup>(8)</sup> More recently, research has evolved from convolutional neural networks (CNNs) toward architectures such as the vision and multiscale vision transformers, which leverage self-attention mechanisms to model long-range temporal dependences and have demonstrated promising performance in high-resolution video analysis and long-term activity recognition tasks.<sup>(8)</sup> Beyond action recognition, FPV research has increasingly expanded to the field of action anticipation, which aims to predict a user's future actions and interaction targets before they occur. One representative work is the rolling-unrolling long short-term memory framework proposed by Furnari and Farinella.<sup>(9)</sup> The authors argued that conventional methods often combine historical information encoding and future event inference within a single model, making it difficult to effectively summarize past behaviors while accurately predicting future actions. To address this issue, rolling-unrolling long short-term memory (RU-LSTM) separates these two processes by employing a rolling LSTM for historical information encoding and an unrolling LSTM for future action prediction, thereby improving anticipation performance. Despite the success of global-feature-based methods in action recognition and action anticipation tasks, several limitations remain. First, RU-LSTM primarily relies on RGB images, optical flow, and object detection<sup>(10)</sup> features, without explicitly modeling hand information. In FPV scenarios, hands serve as the most direct medium through which users interact with their environment; therefore, the absence of hand-contact information may limit the model's ability to understand fine-grained manipulation behaviors. Furthermore, Verma *et al.* proposed a unified CNN-LSTM framework<sup>(11)</sup> that utilizes a twostream CNN to extract RGB and optical flow features and an LSTM to model temporal dependences. Although the method achieved strong performance on several public datasets, it remains fundamentally a global-feature-based approach. The model infers user actions primarily from the overall scene content without explicitly modeling hand–object interactions.

As a result, the model cannot determine whether a hand is in contact with an object, identify the contact location, or describe detailed interaction states during task execution. The above analysis revealed that global-feature-based methods offer strong generalization capabilities and computational efficiency for action recognition and anticipation tasks. However, for industrial scenarios requiring detailed descriptions of hand contact, grasping, releasing, and fine-grained manipulation processes, these methods remain insufficient for accurately capturing the actual interactions between operators and equipment.

## 2.2 Local features

In local-feature-based methods, hand trajectory analysis and gesture recognition<sup>(12)</sup> have long been important research topics in egocentric behavior understanding. Previous studies have shown that static gesture recognition primarily relies on hand appearance features, whereas dynamic gesture recognition requires temporal information to model motion patterns over time. Since dynamic gestures can more comprehensively describe operational procedures, user intentions, and behavioral transitions, they have demonstrated significant value in industrial operation analysis, procedure understanding, and human–computer interaction applications. In recent years, FPV research has gradually evolved from action recognition toward attention-aware and HOI-based frameworks. The gaze-based action recognition method<sup>(13)</sup> modeled gaze information as a stochastic variable within a deep probabilistic framework, wherein uncertainty-aware attention maps were generated from gaze distributions to guide the model toward task-relevant regions, thereby improving action recognition performance. Although this approach effectively integrates gaze cues and action features, its performance remains sensitive to gaze-tracking errors and user attention shifts, while lacking an explicit representation of hand–object contact relationships. For HOI analysis, Mur-Labadia *et al.*<sup>(14)</sup> proposed a mutual learning framework that jointly models hand–object contact and object state localization. By exploiting the correlation between contact states and object state changes, the method improves robustness in few-shot learning and unseen environments. However, it requires large amounts of detailed manual annotations, significantly increasing data collection and labeling costs. In the area of short-term anticipation, the StillFast framework was introduced,<sup>(15)</sup> which employs a dual-branch architecture consisting of high-resolution still images and low-resolution video streams<sup>(16)</sup> to balance recognition accuracy and computational efficiency. Compared with conventional two-stage approaches, StillFast achieves superior inference speed and prediction accuracy. Nevertheless, the multitask learning paradigm may introduce task competition, potentially leading to unstable training behavior. This method has consistently ranked among the top-performing approaches on the Ego4D STA benchmark. However, its reliance on scene-specific operational patterns may limit generalization across different environments. Furthermore, the state–object interaction anticipation via a decoupled object detection framework was proposed.<sup>(17)</sup> By first detecting potential interaction objects using YOLOv9 and subsequently performing state classification and anticipation, the method significantly reduces the complexity associated with multitask learning. However, because inference is primarily based on the final video frame, its time-to-contact prediction accuracy remains limited. For long-term action forecasting, a method that utilizes pairwise HOI features together with cross-correlation mechanisms was used to infer future behaviors.<sup>(18)</sup> While this method effectively handles varying numbers of interacting objects and demonstrates strong long-term forecasting capabilities, its performance may degrade in unstructured environments owing to object occlusions and camera motion. To introduce higher-level semantic reasoning, phrase-guided egocentric action reasoning,<sup>(19)</sup> which employs phrase prompts to establish bidirectional constraints between manipulation intentions and hand actions, was proposed. This design enables simultaneous reasoning over manipulated objects, hand poses, and behavioral intentions.

However, ambiguous phrase descriptions may introduce uncertainty and instability into the reasoning process. Regarding cross-view research, the temporal prior driven flow (TPDF) network<sup>(20)</sup> was proposed to generate egocentric videos from third-person perspectives. By leveraging motion-guided prioritization strategies, TPDF achieves greater flexibility than conventional methods that require explicit target-view specifications. Nevertheless, its training process demands substantial computational resources and training time, limiting practical deployment.

With the increasing adoption of smart glasses and wearable sensing devices, FPV has become an important research direction for industrial operation assistance systems. In addition to high-level behavior understanding models, image preprocessing techniques have also continued to evolve. Various region-based preprocessing strategies have been proposed<sup>(21)</sup> to improve hand localization and interaction analysis. For instance, sample space dimensionality reduction and directional lifting discrete wavelet transform have been employed to enhance target localization in low-resolution images while removing redundant information. These statistical approaches can achieve high recognition accuracy with response times as fast as 300  $\mu$ s. However, they remain sensitive to illumination variations and environmental changes. To improve the automation of hand region extraction, DeepLabV3+ combined with radial distance function analysis was subsequently introduced for hand contour localization and segmentation.<sup>(22)</sup> These methods effectively address challenges related to hand rotation, translation, and scale variation, thereby improving robustness. Nevertheless, the multistage preprocessing pipeline introduces considerable computational complexity, making real-time deployment difficult. The above studies have revealed that hand segmentation serves as an important preprocessing step for isolating hand regions from complex backgrounds and providing accurate hand contour information for subsequent interaction analysis. In this work, hand segmentation is adopted not as the final objective but as the foundation for the proposed HIPM. Specifically, the segmented hand contour enables HIPM to accurately identify fine-grained interaction points between the hand and manipulated objects. Accordingly, hand segmentation establishes the essential link between object detection and the proposed HIPM framework, enabling coarse object localization and fine-grained hand–object interaction analysis within a unified pipeline.

As seen from the above literature review, global-feature-based and local-feature-based methods play different roles in FPV behavior understanding. Global-feature-based methods mainly exploit holistic scene information to model spatiotemporal relationships between human actions and the surrounding environment. They generally exhibit strong scene understanding, better generalization, and lower preprocessing requirements, making them suitable for action recognition, action localization, and action anticipation. However, because these methods rely primarily on holistic visual representations, they often fail to accurately capture hand–object contact locations, contact states, and fine-grained manipulation details, limiting their applicability in industrial scenarios involving tool operation, precision assembly, and procedure analysis. In contrast, local-feature-based methods focus on hands, manipulated objects, gaze, and their interaction relationships, enabling a more accurate modeling of HOIs, grasp states, and manipulation details. Consequently, they are more suitable for industrial FPV applications such as procedure understanding, operational error detection, and skill assessment, although they

generally require additional preprocessing, including hand/object detection, segmentation, and tracking. Despite these advances, most existing studies still focus on action recognition, future behavior prediction, or gaze analysis, while providing limited descriptions of actual hand–object contact locations and interaction processes. Therefore, in this study, we adopt a local-feature-based approach and propose the HIPM framework to achieve more accurate fine-grained hand–object contact recognition for industrial FPV applications.

### 3. Proposed Method

In this study, we address the challenge of accurately modeling hand–object interactions in egocentric industrial environments. Compared with conventional FPV-based behavior recognition approaches that primarily rely on global visual features, hand bounding boxes, or hand keypoint, existing methods can effectively recognize action categories but often fail to precisely describe the actual contact locations and manipulation states involved in industrial maintenance and equipment operation tasks. To overcome this limitation, the proposed framework employs a hand segmentation model to obtain accurate hand regions and further integrates hand keypoint information to construct extreme hand anchors, enabling a more precise representation of the interaction relationship between operators and industrial equipment. Subsequently, a YOLOv8-based object detection model is utilized to localize industrial devices and operational components. The minimum distance between extreme hand anchors and object bounding boxes is then calculated to establish a HOI model. By using the interaction strength derived from spatial proximity, the proposed framework is capable of inferring interaction events such as contact, grasping, and release, thereby providing a more detailed understanding of industrial operation behaviors and procedural execution processes.

#### 3.1 System architecture of the proposed egocentric industrial HOI recognition method

HOI recognition in FPV scenarios primarily focuses on analyzing the interaction relationships between the operator's hands and surrounding objects. Therefore, we propose an industrial HOI recognition framework based on HIPM, as illustrated in Fig. 1. The input to the proposed system consists of continuous egocentric video streams. Because of the strong temporal continuity of video data, the input video is first divided into a sequence of frames in accordance with a predefined frame rate, providing the fundamental data for subsequent analysis. To accurately localize the operator's hands, a hand segmentation module is first employed to extract hand contours. Since hands serve as the primary interaction medium in FPV environments, the accuracy of hand localization directly affects the performance of subsequent HOI recognition tasks. Inspired by the hand segmentation frameworks proposed,<sup>(22)</sup> we adopt a DeepLabV3-based architecture for pixel-level hand segmentation. The model utilizes convolutional neural networks to identify hand regions and combines contour analysis techniques to obtain complete hand mask information. To improve generalization across diverse environments, the model is further pretrained on several publicly available egocentric hand datasets, including EgoHands,<sup>(7)</sup> EgoYouTubeHands,<sup>(7)</sup> and GTEA.<sup>(7)</sup> This enables robust hand segmentation under varying illumination conditions, background clutter, and hand pose variations.

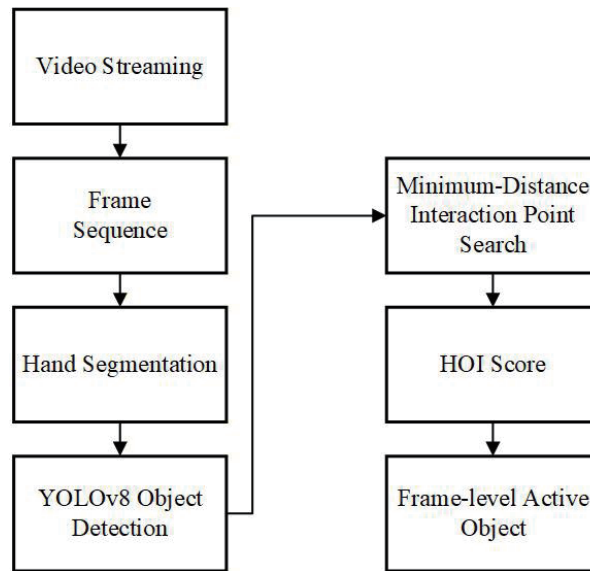


Fig. 1. System architecture of the proposed egocentric industrial HOI recognition framework.

After hand segmentation, YOLOv8<sup>(17)</sup> is employed as the industrial object detection model. YOLOv8 inherits the real-time object detection capabilities of the YOLO family while incorporating a CSPNet-based backbone and a neck architecture composed of feature pyramid and path aggregation networks for multiscale feature fusion. Furthermore, an anchor-free detection head is adopted to directly predict object centers and bounding boxes, significantly reducing the hyperparameter tuning requirements associated with conventional anchor-based approaches. The YOLOv8 model outputs object class labels and corresponding bounding boxes for industrial equipment and operational components, which serve as the basis for subsequent hand–object interaction analysis.

### 3.2 Hand interaction target point modeling

During equipment operation and object manipulation, the actual contact regions are typically concentrated around fingertips and hand boundaries rather than the geometric center of the hand. Therefore, conventional approaches that use the hand center as the primary reference for interaction analysis often fail to accurately represent real hand–object contact relationships. To address this limitation, we propose a HIPM mechanism to establish a more realistic representation of hand–object interactions during industrial operations. As illustrated in Fig. 2, a hand segmentation model is first employed to obtain the hand mask, followed by the extraction of the complete hand contour. Since the hand contour preserves the geometric shape and boundary information of the hand, it provides a more accurate description of the actual contact regions between the operator and the manipulated objects. In this study, the inner contour is adopted as the candidate set of interaction points. Subsequently,  $n$  feature points are uniformly sampled along the contour to construct a hand contour point set, defined as

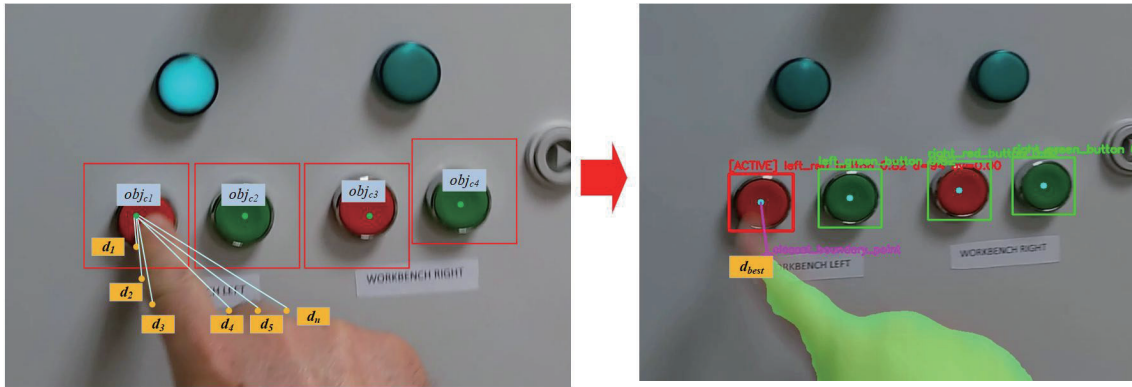


Fig. 2. (Color online) Shortest distance interactive search for hand objects.

$$P_H = \{p_1, p_2 \dots p_n\}, \quad (1)$$

where

$$p_i = (x_i, y_i) \quad (2)$$

represents the coordinates of the  $i$ -th hand contour point. On the other hand, the bounding boxes of operational objects in the scene are obtained using the YOLOv8 object detection model, and the center coordinates of each detected object are computed as follows.

$$c_j = \{c_{xj}, c_{yj}\} \quad (3)$$

$$d_{ij} = \sqrt{(x_i - c_{xj})^2 + (y_i - c_{yj})^2} \quad (4)$$

For each detected object  $obj_j$ , the Euclidean distances between its center  $c_j$  and all sampled hand contour points  $p_i$  are calculated using Eq. (4). As illustrated in the left panel of Fig. 2,  $d_{1j}, d_{2j}, \dots, d_{nj}$  represent the candidate distances from the sampled hand contour points to object  $obj_j$ . The shortest distance  $d_j^{min}$  is then selected as the object-specific interaction distance.

$$d_j^{min} = \min_{p_i \in P_H} d_{ij} \quad (5)$$

The contour point corresponding to  $d_j^{min}$  is regarded as the hand interaction target point for  $obj_j$ . In Fig. 2,  $d_{best}$  denotes the smallest object-specific interaction distance among all detected objects.

$$d_{best} = \min_{obj_j \in O} d_j^{min} \quad (6)$$

To convert the interaction distance into an interpretable interaction likelihood, the HOI score for object  $obj_j$  is defined as

$$S_j^{HOI} = \exp\left(-\frac{\hat{d}_j^{min}}{\tau}\right) \quad (7)$$

where  $\tau$  is a distance-sensitive parameter and  $\hat{d}_j^{min}$  denotes the normalized minimum distance. The distance can be normalized by the image diagonal to reduce the effect of image resolution.

$$\hat{d}_j^{min} = \frac{d_j^{min}}{\sqrt{W^2 + H^2}} \quad (8)$$

Equation (8) represents the minimum interaction distance between the hand and the corresponding object. Since actual contact behaviors typically occur at the nearest contour location, the minimum distance provides a more realistic representation of the operator's current interaction target. Subsequently, a HOI score is established on the basis of the minimum distance associated with each object. A smaller hand–object distance indicates a higher likelihood of interaction. The system then selects the object with the minimum interaction distance as the frame-level active object for the current frame, which can be expressed as follows.

$$obj^* = \arg \min_{obj_j \in O} d_j^{min} \quad (9)$$

In Eq. (9),  $O$  denotes the set of all detected objects in the scene. Note that the proposed HIPM is not a standalone HOI recognition module but rather a fine-grained interaction modeling mechanism that provides discriminative interaction cues for subsequent HOI recognition. By explicitly modeling the nearest hand–object interaction point using the hand contour geometry, HIPM establishes a more realistic spatial relationship between the operator's hand and the manipulated objects than do conventional hand-center-based methods. The resulting interaction features enable a more accurate identification of the active object and contact state, thereby improving the robustness and accuracy of frame-level HOI recognition. However, predictions based solely on a single frame are susceptible to errors caused by object occlusion, hand motion blur, object detection inaccuracies, and transient movements. Consequently, frame-level interaction analysis may lead to unstable or incorrect recognition results. Therefore, temporal information is further incorporated to analyze the distance variation trends between hands and objects across consecutive frames. On the basis of these temporal dynamics, a set of interaction

states is defined, including first contact, contact, take, release, and de-contact. Through temporal voting and distance variation analysis, the proposed framework effectively suppresses unstable single-frame predictions and enhances the accuracy and stability of egocentric industrial HOI recognition.

## 4. Experiment

In this study, the ENIGMA-51 dataset is employed as the experimental benchmark for performance evaluation. ENIGMA-51<sup>(4)</sup> is a publicly available dataset specifically developed for industrial FPV research, aiming to support studies on human behavior understanding, HOI, object interaction anticipation, and intent and entity understanding for industrial assistant systems. The dataset was collected from 19 participants wearing Microsoft HoloLens 2 head-mounted devices and contains 51 industrial maintenance videos with a total duration of approximately 22 h. Participants performed maintenance procedures on high-voltage and low-voltage control boards by following AR instructions and voice guidance. During maintenance, various industrial tools and equipment, including oscilloscopes, power supplies, electric screwdrivers, soldering irons, and control buttons, were operated, providing a realistic simulation of industrial maintenance scenarios. To establish the industrial object detection model, the object annotations provided by ENIGMA-51 were first utilized for training. YOLOv8 was adopted as the baseline object detection framework, and pretrained weights from the Microsoft Common Objects in Context (COCO) dataset were used for model initialization to improve convergence speed and detection performance in industrial environments. Subsequently, transfer learning and fine-tuning were performed using the ENIGMA-51 training set, enabling the model to accurately recognize industrial equipment and operational components contained in the dataset. For the experimental hardware configuration, an NVIDIA GeForce RTX 2080 Ti graphics processing unit (GPU) was used as the training platform. Model development and training were implemented using the CUDA toolkit and the PyTorch deep learning framework. The YOLOv8 model was trained for 35 epochs, while the validation set was used to monitor the convergence process and determine the optimal model configuration for industrial object detection.

### 4.1 Impact of different hand segmentation methods on HOI recognition

Hand segmentation plays a critical role in egocentric HOI recognition, as its quality directly affects subsequent hand–object interaction analysis. In industrial environments, factors such as operator skill levels, motion speed, hand occlusions, and image blur often make accurate hand localization challenging. Therefore, the development of a robust and precise hand segmentation model is a fundamental prerequisite for reliable HOI recognition. This study is focused on HOI recognition in FPV scenarios, where the primary objective is to identify the object being manipulated by the operator and determine the timing of interaction events. In industrial applications, common interaction targets include buttons, knobs, connectors, probes, and various electronic components. Actual contact locations are typically concentrated around fingertips,

hand boundaries, or tool-contact regions. Compared with hand detection approaches that only provide coarse spatial information through bounding boxes, hand segmentation models generate pixel-level hand contour information, which more accurately reflects actual contact regions and facilitates precise hand–object relationship modeling.

Although multiclass segmentation models are capable of simultaneously segmenting hands and industrial objects, their learning objectives are distributed across multiple categories, potentially compromising the quality of hand contour representations. Since the proposed HIPM framework relies heavily on accurate hand contour information, improving hand segmentation quality is considered more beneficial than improving object segmentation performance for HOI recognition. To investigate the effects of different hand segmentation methods on HOI recognition, two representative hand segmentation approaches were selected for comparison. The first approach is the lightweight real-time hand segmentation model proposed by Sánchez-Brizuela *et al.*,<sup>(23)</sup> which combines MediaPipe hand landmarks with color-based features to generate hand masks, offering high inference speed and low deployment cost. The second approach is the enhanced mmEgoHand framework,<sup>(24)</sup> which employs deep-learning-based pixel-level hand segmentation to obtain more complete hand contour information. Building upon these methods, the proposed approach further optimizes hand contour extraction and interaction point modeling, forming the core component of the proposed framework. Figure 3 presents the results of different hand segmentation methods. Experimental results demonstrate that the proposed method achieves superior performance in terms of hand contour completeness, boundary stability, and robustness under occlusion conditions. It effectively reduces recognition errors caused by motion blur and partial hand occlusions. Furthermore, the more stable hand contour information facilitates subsequent minimum-distance interaction point search and hand–object association analysis, thereby improving the overall accuracy and robustness of industrial HOI recognition. In contrast, the lightweight and CSW segmentation methods show inferior performance in industrial FPV scenarios. The lightweight model is more sensitive to challenging imaging conditions, whereas the CSW segmentation module often produces incomplete hand contours under occlusion and rapid hand movements. These segmentation errors degrade hand contour extraction and subsequently affect interaction point estimation, hand–object distance calculation, and active object selection in HIPM. In comparison, the proposed method preserves more complete hand contours, resulting in more robust and accurate HOI recognition.

#### 4.2 Comparison of hand interaction point modeling methods

In this section, we investigate the impact of different HIPM strategies on HOI recognition performance. First, a hand segmentation model is employed to obtain the complete hand mask, followed by the extraction of the hand contour. On the basis of different hand representations, five interaction point modeling methods are constructed, including center, moments, heatmap center, outer contour, and inner contour. Since each interaction point definition captures different spatial characteristics of the hand, the resulting relationships with target objects may vary accordingly. To evaluate the effectiveness of these interaction point modeling strategies, a

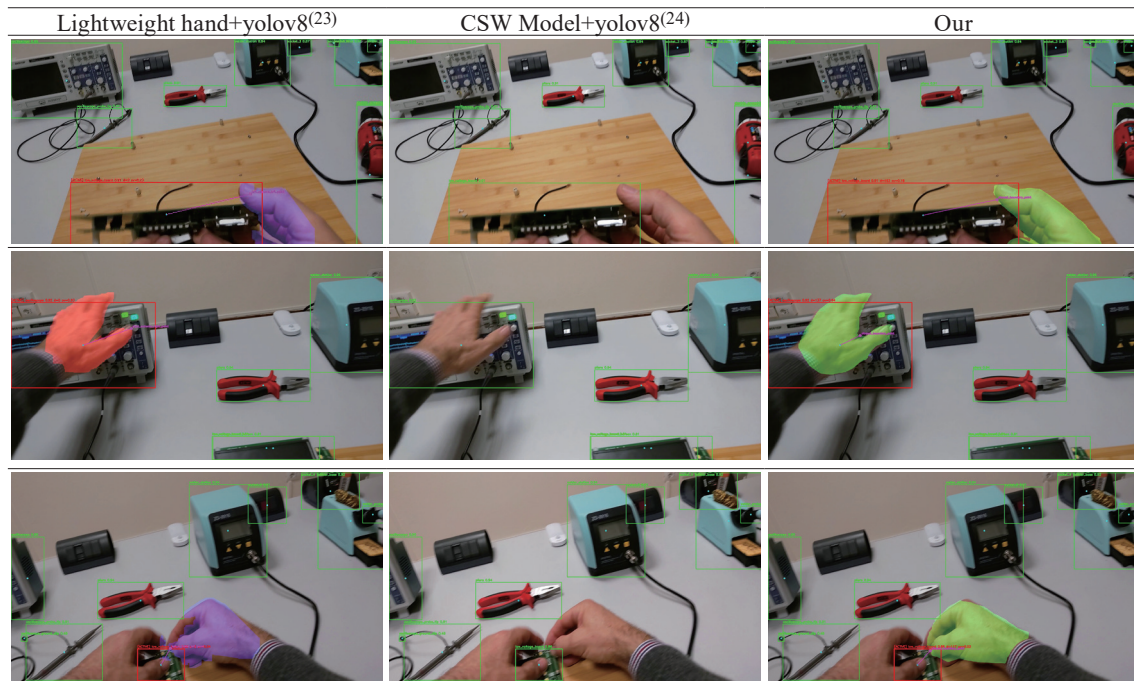


Fig. 3. (Color online) Hand segmentation results with different objects and viewpoints.

minimum-distance interaction point search mechanism is proposed. The minimum distance between each hand interaction point and object bounding box is calculated to estimate the likelihood of interaction. A smaller distance indicates a higher probability that the hand is interacting with the corresponding object. Subsequently, a HOI score is computed on the basis of the minimum distance, and the object with the highest interaction relevance is selected as the frame-level active object for the current frame. Figure 4 illustrates the different hand interaction point modeling methods. Among them, the center, moments, and heatmap center approaches utilize global hand region information to represent hand location, whereas the outer contour and inner contour approaches explicitly incorporate hand boundary characteristics. In practical industrial operations, contact between the hand and an object typically occurs first at the fingertips or along the hand boundary. Therefore, contour-based interaction points provide a more realistic representation of actual contact locations. In particular, the outer contour approach directly captures the spatial relationship between the outermost hand boundary and the target object, enabling an earlier detection of contact events and improving the accuracy of HOI analysis. Experimental results further demonstrate that, compared with conventional center-based methods, contour-based interaction point modeling achieves more accurate and reliable HOI recognition performance.

#### 4.3 Analysis of egocentric human–object interaction (EHOI) detection task results

To evaluate the effectiveness of the proposed HIPM-based egocentric industrial HOI recognition method, the EHOI detection task defined in the ENIGMA-51 dataset was adopted as

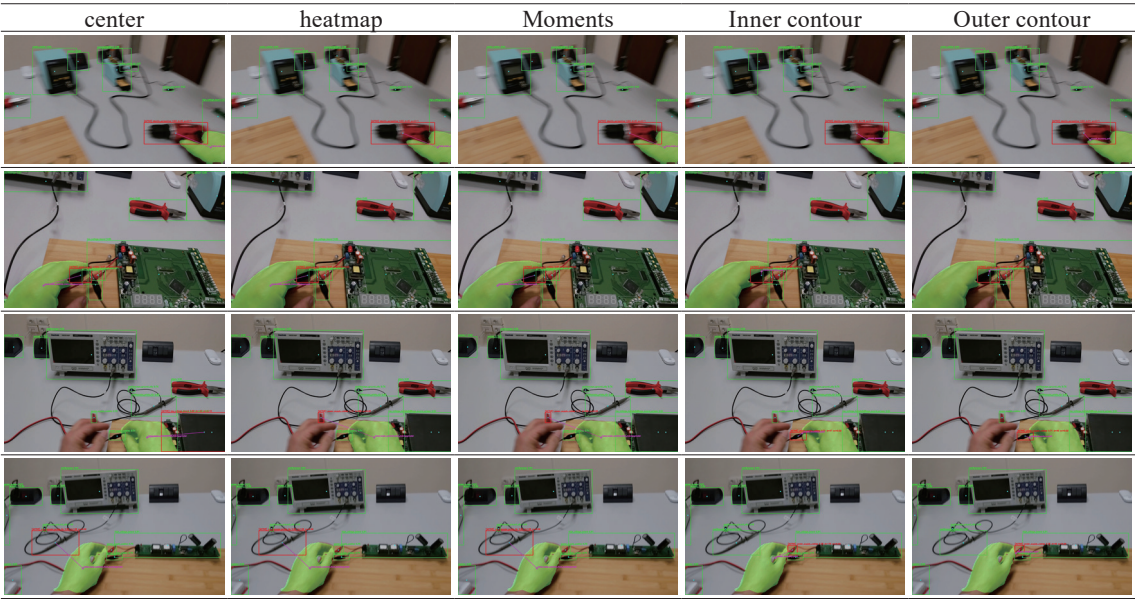


Fig. 4. (Color online) Hand–object interaction analysis with different objects and viewing conditions using different interaction point modeling methods.

Table 1  
EHOI detection performances characteristics of various hand interaction point modeling methods.

| Method                           | AP H.+Side | AP H.+State | mAP H.+Obj | mAP H.+All |
|----------------------------------|------------|-------------|------------|------------|
| StillFast                        | 90.81      | 73.31       | 46.51      | 46.24      |
| Lightweight hand+yolov8          | 59.50      | 16.83       | 37.59      | 33.93      |
| Csw+yolov8                       | 3.14       | 0.85        | 2.3        | 0.32       |
| Proposed<br>(center-based HIPM)  | 99.26      | 36.39       | 77.19      | 29.4       |
| Proposed<br>(outer-contour HIPM) | 99.26      | 56.42       | 78.03      | 45.63      |
| Proposed<br>(inner-contour HIPM) | 99.26      | 59.6        | 78.60      | 43.23      |

the evaluation benchmark. This task assesses the capability of a model to recognize hands, interaction states, and manipulated objects in industrial FPV scenarios. Performance is quantified using four evaluation metrics: AP Hand, AP H.+State, mAP H.+Obj, and mAP H.+All. Specifically, AP Hand is a measure of hand detection performance; AP H.+State is the joint recognition capability of hands and interaction states (contact/no-contact); mAP H.+Obj is the accuracy of hand–object association; and mAP H.+All represents the most stringent evaluation criterion, where the simultaneous correct identification of the hand, interaction state, and manipulated object is required for a prediction to be considered successful. To address the limitations of conventional center-based interaction analysis, we introduce a hand interaction point modeling mechanism that utilizes contour information extracted from hand segmentation results to establish hand–object associations. Furthermore, a minimum-distance interaction point search strategy is employed to analyze the spatial relationship between the hand and candidate objects, enabling a more accurate representation of actual contact locations. The experimental results are summarized in Table 1. Compared with the ENIGMA-51 baseline

model and other comparative approaches, the proposed method achieves superior performance in AP H.+State, mAP H.+Obj, and mAP H.+All. The contour-based interaction point modeling strategy more effectively captures real hand–object contact locations, allowing the system to accurately recognize interaction events such as contact, grasping, and release during industrial operations. Consequently, the proposed framework significantly improves the overall HOI recognition performance in industrial FPV environments.

## 5. Conclusions

Industrial operations are characterized by strong temporal dependences, frequent hand occlusions, and small interaction targets, making it challenging for conventional action recognition and object detection approaches to accurately model real hand–object interaction relationships. To address this issue, we proposed an egocentric industrial HOI recognition method based on HIPM. The proposed framework first utilizes a hand segmentation model to obtain complete hand contour information and then constructs multiple hand interaction point representations. Subsequently, a minimum-distance interaction point search mechanism is employed to establish spatial relationships between hands and objects. By further incorporating temporal information, interaction events such as contact, grasping, and release can be effectively recognized in industrial scenarios. To evaluate the effectiveness of the proposed method, experiments were conducted using the EHOI detection task defined in the ENIGMA-51 dataset. Experimental results indicate that contour-based interaction point modeling provides a more accurate representation of actual hand–object contact locations than do conventional center-based approaches. In particular, the inner contour and outer contour methods achieved significant improvements in AP H.+State metric, indicating that interaction behaviors in industrial operations are primarily concentrated along hand boundaries rather than at the geometric center of the hand. Furthermore, the proposed contour-based approaches achieved superior performance in hand–object association analysis and interaction event recognition, validating the importance of hand contour information for industrial HOI understanding.

Although the current framework mainly relies on spatial distance relationships between hands and objects for interaction inference, future work will focus on integrating vision-language models (VLMs) and large language models (LLMs) to further incorporate semantic information, procedural knowledge, and task intentions. Such integration is expected to enable higher-level procedure understanding, operation guidance, anomaly detection, and intelligent assistance capabilities for next-generation smart manufacturing systems.

## References

- 1 A. A. Fasoyinu, S. Azhar, A. Sattineni, and J. O. Toyin: *Autom. Constr.* **179** (2025) 106424. <https://doi.org/10.1016/j.autcon.2025.106424>
- 2 A. Bandini and J. Zariffa: *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (2023) 6846. <https://doi.org/10.1109/TPAMI.2020.2986648>
- 3 S. Mann: *Proc. 2nd Int. Symp. Wearable Comput.* (1998) 124–131. <https://doi.org/10.1109/ISWC.1998.729538>
- 4 F. Ragusa, R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, and A. Furnari: *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)* (2024) 4345–4354. <https://doi.org/10.1109/WACV57701.2024.00449>

- 5 H. Coskun, M. Z. Zia, B. Tekin, F. Bogo, N. Navab, and F. Tombari: IEEE Trans. Pattern Anal. Mach. Intell. **45** (2023) 6659. <https://doi.org/10.1109/TPAMI.2021.3058606>
- 6 Y. Lu and Y. Liu: Proc. IEEE SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta (2022) 176–183. <https://doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00031>
- 7 K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, *et al.*: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2022) 18973–18990. <https://doi.org/10.1109/CVPR52688.2022.01842>
- 8 H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) (2021) 6804–6815. <https://doi.org/10.1109/ICCV48922.2021.00675>
- 9 A. Furnari and G. M. Farinella: IEEE Trans. Pattern Anal. Mach. Intell. **43** (2021) 4021. <https://doi.org/10.1109/TPAMI.2020.2992889>
- 10 T. Kurniati, C.-K. Yang, T.-S. Chen, Y.-F. Chung, Y.-M. Huang, and C.-C. Chen: Sens. Mater. **33** (2021) 4057. <https://doi.org/10.18494/SAM.2021.3528>
- 11 S. Verma, P. Nagar, D. Gupta, and C. Arora: Proc. IEEE Int. Conf. Image Process. (ICIP) (2018) 2456–2460. <https://doi.org/10.48550/arXiv.1910.07766>
- 12 H.-Y. Lai, H.-Y. Ke, and Y.-C. Hsu: Sens. Mater. **30** (2018) 869. <https://doi.org/10.18494/SAM.2018.1790>
- 13 Y. Li, M. Liu, and J. M. Rehg: IEEE Trans. Pattern Anal. Mach. Intell. **45** (2023) 6731. <https://doi.org/10.1109/TPAMI.2021.3051319>
- 14 L. Mur-Labadia, R. Martinez-Cantin, J. J. Guerrero, G. M. Farinella, and A. Furnari: Lect. Notes Comput. Sci. **15081** (2024) 167.
- 15 T. Shiota, M. Takagi, K. Kumagai, H. Seshimo, and Y. Aono: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV) (2024) 6331–6340. <https://doi.org/10.1109/WACV57701.2024.00641>
- 16 Y. Kanaya, S. Kawanaka, H. Suwa, Y. Arakawa, and K. Yasumoto: Sens. Mater. **32** (2020) 599. <https://doi.org/10.18494/SAM.2020.2616>
- 17 H. Cho, D. U. Kang, and S. Y. Chun: arXiv:2407.05713 (2024). <https://doi.org/10.48550/arXiv.2407.05713>
- 18 D. Roy and B. Fernando: IEEE Trans. Image Process. **30** (2021) 8116. <https://doi.org/10.1109/TIP.2021.3113114>
- 19 Z. Zhang, H. Luo, W. Zhai, Y. Cao, and Y. Kang: Sci. China Inf. Sci. **68** (2025) 150209. <https://doi.org/10.1007/s11432-024-4405-4>
- 20 W. Hu, J. T. Hoe, R. Zhang, Y. Yang, H. Hu, and Y.-P. Tan: Inf. Fusion **123** (2025) 103273. <https://doi.org/10.1016/j.inffus.2025.103273>
- 21 Y. Y. Chen, C.-H. Hsia, and P.-H. Chen: IEEE Access **9** (2021) 149796. <https://doi.org/10.1109/ACCESS.2021.3124631>
- 22 K. Y. Chou, C.-S. Lin, C.-H. Chien, J.-S. Chiang, and C.-H. Hsia: Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS) (2016) 1–4. <https://doi.org/10.1109/ISPACS.2016.7824769>
- 23 G. Sánchez-Brizuela, A. Císal, E. de la Fuente-López, J.-C. Fraile, and J. Pérez-Turiel: Virtual Reality **27** (2023) 3125. <https://doi.org/10.1007/s10055-023-00856-2>
- 24 Y. Lv, T. Zhang, Z. Wang, Y. Song, H. Ding, J. Han, and F. Wang: arXiv:2501.13805 (2025). <https://doi.org/10.48550/arXiv.2501.13805>