https://doi.org/10.1016/j.aei.2022.101792

Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly

Yaqian Zhang[1,2], Kai Ding[1,2*], Jizhuang Hui[1,2*], Jingxiang Lv[1,2], Xueliang Zhou[3], Pai Zheng[4]

[1] Institute of Smart Manufacturing Systems, Chang'an University, Xi'an, China;
[2] Key Laboratory of Road Construction Technology and Equipment, Chang'an University, Xi'an, China.
[3] School of Mechanical Engineering, Hubei University of Automotive Technology, Shiyan, China.
[4] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China.
***** Correspondence: kding@chd.edu.cn; huijz@chd.edu.cn

Abstract: Human-robot collaborative (HRC) assembly combines the advantages of robot's operation consistency with human's cognitive ability and adaptivity, which provides an efficient and flexible way for complex assembly tasks. In the process of HRC assembly, the robot needs to understand the operator's intention accurately to assist the collaborative assembly tasks. At present, operator intention recognition considering context information such as assembly objects in a complex environment remains challenging. In this paper, we propose a human-object integrated approach for context-aware assembly intention recognition in the HRC, which integrates the recognition of assembly actions and assembly parts to improve the accuracy of the operator's intention recognition. Specifically, considering the real-time requirements of HRC assembly, Spatial-Temporal Graph Convolutional Networks (ST-GCN) model based on skeleton features is utilized to recognize the assembly action to reduce unnecessary redundant information. Considering the disorder and occlusion of assembly parts, an improved YOLOX model is proposed to improve the focusing capability of network structure on the assembly parts that are difficult to recognize. Afterwards, taking decelerator assembly tasks as an example, a rule-based reasoning method that contains the recognition information of assembly actions and assembly parts is designed to recognize the current assembly intention. Finally, the feasibility and effectiveness of the proposed approach for recognizing human intentions are verified. The integration of assembly action recognition and assembly part recognition can facilitate the accurate operator's intention recognition in the complex and flexible HRC assembly environment.

Keywords: Human-robot collaborative assembly; human intention recognition; ST-GCN; part recognition; improved YOLOX

1. Introduction

With the development of advanced machining technologies, the machining accuracy and consistency of parts have improved much, which highlights the importance of assembly to ensure product quality [1]. Since complex product assembly work occupies large labor intensity and cost, it is of vital importance to improve the efficiency and flexibility of complex product assembly tasks [2]. In automated production workshops, robots have been widely used to execute repeatable and heavy work to reduce labor costs and improve operation accuracy, especially in assembly processes. However, in complex product assembly tasks, human operations are still essential because robots have little cognitive ability and flexibility. Therefore, Human-robot collaborative (HRC) assembly [3, 4], as a new model combining the advantages of humans and robots, has gradually become a hot research topic. Compared with traditional manufacturing systems, collaborative robots manage their behaviors not based on the traditional pre-programmed instructions but the visual [5], tactile [6], and other ways [7, 8] to perceive the operator's intention, to better accomplish the HRC assembly work. Therefore, it is crucial for robots to accurately recognize the operator's intention in the HRC assembly process.

The operator's intention recognition can be inferred by recognizing the assembly action. Assembly action recognition can be realized in different modalities of data, such as RGB images [9], optical flow [10], body skeletons [11], etc. However, RGB image-based methods are usually susceptible to complex backgrounds, illumination changes, and other external factors. Optical flow only represents the pixel-level differences between adjacent frames. Traditional human action recognition methods based on optical flow are slow in computation. In contrast, skeleton-based action recognition methods are robust to the above factors and have less computational consumption because they only need to process the skeleton data. In addition, Microsoft® Kinect visual camera and human pose estimation algorithm provide the basis for skeleton-based action recognition. However, it should be noted that in the process of HRC assembly, due to the influence of complex background changes, the similarity of different actions, the occlusion of the human body, and other factors, it is still of low confidence to recognize the operator's intention only through action recognition. Moreover, the operator's intention will change with different assembly parts.

To promote better collaboration between human and robot in the HRC assembly, we propose a framework combining skeleton-based assembly action recognition and assembly part recognition to recognize the operator's intentions. On the one hand, we recognize the operator's assembly action based on the Spatial-Temporal Graph Convolutional Networks (ST-GCN) model. On the other hand, an improved YOLOX model integrating the Convolutional Block Attention Module (CBAM) and Focal Loss function is proposed to recognize the assembly part. On that basis, we design a rule-based reasoning method to accurately recognize the operator's intentions in the complex and flexible HRC assembly environment.

The main contributions of our paper are as follows:

- A framework for operator intention recognition in the HRC assembly is built based on the integration of assembly action recognition and assembly part recognition.
- An assembly action dataset (AAD) of the decelerator assembly tasks is built by using the Azure Kinect DK camera to capture a series of 5~6s short videos, and the ST-GCN model is adopted to recognize the operator's assembly action in the HRC assembly.
- The corresponding assembly part dataset (APD) of the decelerator assembly tasks is built by snapshotting images from the short videos, and an improved YOLOX model integrating CBAM and Focal Loss function is developed to recognize the assembly part in the HRC assembly.
- A rule-based reasoning method is designed to infer the operator's assembly intention and the responsive operation of the robot.

The rest of this paper is organized as follows. A literature review related to human action recognition for HRC assembly and object detection for HRC assembly is provided in Section 2. Section 3 describes the overall framework of the operator's intention recognition in the HRC assembly. In section 4, the ST-GCN model for recognizing the operator's assembly action is constructed. In Section 5, we establish an improved YOLOX model to recognize assembly parts. In Section 6, the feasibility and effectiveness of the proposed approach are verified based on the AAD and APD. Finally, some concluding remarks are presented in Section 7.

2. Related work

In the HRC assembly, the operator's intention is highly related to the assembly actions and the assembly parts. In section 2, we review human action recognition for HRC assembly and object detection for HRC assembly, respectively.

2.1. Human action recognition for HRC assembly

In the process of HRC assembly, the operator's assembly action information is an essential part of obtaining assembly intention. The action recognition can be realized by optical flow, RGB, skeleton, and other modalities of data. Zhu et al. [12] used an optical flow model to extract local optical flow features and combined the global silhouette features to recognize human action. Sidor et al. [13] converted the depth maps into a 3D point cloud, and then realized the classification of human activities through the classifier. These image sequence-based methods need to deal with a large amount of data information, and still have shortcomings when applied in scenarios with real-time requirements.

In contrast, skeleton sequence-based action recognition methods are robust to the background changes and do not have excessive redundant information, which can better realize HRC assembly tasks with real-time requirements. The human skeleton is like a topology, naturally constructed as a graph in a non-Euclidean space. There are two ways to process skeleton sequences. One way is to encode skeleton sequences into images, and then typically use the recurrent neural network or convolutional neural network (CNN) to extract features. Urgo et al. [14] recognized operator's locations based on OpenPose, and then built monitoring methods based on a hidden Markov model to recognize missing operations or unsafe behavior. Hu et al. [15] proposed a framework for skeleton-based action recognition that can select temporal scales automatically with a single layer Long Short Memory Networks (LSTM). The action recognition methods based on RNN can effectively process sequence data but has limitations in extracting spatial features of the human skeleton. However, the skeleton sequence has abundant spatial and temporal information, and CNN has excellent advanced information extraction ability, which has been widely used. Naveenkumar et al. [16] presented a deep learning approach for skeleton-based action recognition using CNN and LSTM, which achieved competitive results on open datasets. Al-Amin et al. [17] proposed a personalized system of the skeleton data-based CNN classifier to recognize the operator's assembly actions, which improves the action recognition accuracy of heterogeneous workers. The system comprised six 1-channel CNN classifiers, which can be adapted to new workers by transfer learning.

Another way is to construct a Graph convolutional network (GCN). The application of GCN to skeleton-based action recognition has been proved to achieve excellent results, extending traditional CNN from images to graphs with arbitrary structure. Yan et al. [18] first proposed the ST-GCN model for skeleton-based action recognition, which can automatically learn spatial-temporal patterns from skeleton data. This work has drawn more attention to the advantages of GCN for skeleton-based behavior recognition. Some researchers have also made improvements on the basis of the ST-GCN model [19, 20]. In this paper, we apply the ST-GCN model to the field of HRC assembly and recognize the operator's actions by exploring the spatial-temporal features of the human skeleton, providing a decision-making basis for HRC.

2.2 Object detection for HRC assembly

Object detection is a hot research topic in the machine vision field and it is widely used in real-life scenarios, such as assembly elements recognition [21], ship detection [22], etc. Especially, with the rapid development of deep learning, the performance of object detection algorithms has been greatly improved. According to the existence of candidate regions, object detection algorithms can be divided into two types, i.e. one-stage detection and two-stage detection [23].

Two-stage object detection algorithm includes two stages, i.e. candidate region extraction and classification regression. Typical two-stage algorithms, especially the R-CNN series, show high accuracy in the recognition of assembly parts. Wang et al. [24] adopted the Faster R-CNN algorithm to recognize assembly parts related to specific tasks, achieving 99% accuracy. Back et al. [25] proposed a Mask R-CNN with a confidence map estimator for the accurate detection of texture-less and metallic industrial components. The two-stage object detection algorithm achieves good results in precision, but the speed is limited, and it is often difficult to meet the real-time detection requirements in the HRC assembly scene [26].

The one-stage object detection algorithm has a smaller network model and faster operation speed, which has great advantages in application scenarios requiring real-time recognition and fast decision-making [26]. Typical algorithms for one-stage object detection include the YOLO series [27], single shot detector (SSD) series [28], etc. Andrianakos et al. [29] applied the SSD algorithm to recognize assembly parts and the operator's hand for automatic monitoring of assembly operation execution. With the advantages of simple structure, fast, and higher accuracy, some researchers applied the YOLO algorithm to assembly part recognition. Chen et al. [30] applied the YOLOv3 algorithm to the location and judgment of assembly tools, so as to recognize the operator's assembly actions. Wang et al. [21] utilized YOLOv3 to predict the positions of elements (operator, robot, assembly parts and tools, etc.) in the assembly line, calculated the corresponding target movement speed based on the position information, and finally carried out motion recognition based on the above information. However, these YOLO series algorithms, which adopt the structure of coupled head and anchor-based, still have some disadvantages in balancing speed and accuracy.

As a new YOLO series algorithm for object detection, YOLOX improves the accuracy and optimizes the inference speed [27]. We apply YOLOX to the recognition of assembly parts in HRC assembly. Assembly parts often have the problem of disordered parts placement and occlusion, which will affect the recognition of assembly parts. To solve these problems in the HRC assembly, and recognize assembly parts accurately, an improved YOLOX algorithm is designed based on the YOLOX-S network. The improved YOLOX algorithm makes the network focus on assembly parts by adding CBAM [31] at the end of the backbone network and replaces the confidence loss function in the original algorithm with the Focal Loss function [32] to improve the recognition performance of assembly parts that are difficult to recognize.

2.3 Research gap

The operator's intention inference is closely related to assembly actions and assembly parts, and directly affects the robot's responsive operation. The assembly parts corresponding to the same assembly action may be different, so the operator's intention will also change, and the robot's assistance will also be different. In a complex and flexible assembly environment, there are many kinds of assembly parts and different assembly sequences, so it is challenging to recognize the operator's intention. Chen et al. applied YOLOv3 to locating and judging assembly tools to directly recognize assembly actions, and the convolutional pose machine was used to estimate the operating times of the repetitive assembly action. Wang et al. [33] investigated the transfer learning-based AlexNet network for synchronous recognition of human actions and corresponding assembly parts, providing a basis for high-performance HRC. Zhang et al. [34] developed Bi-stream CNN for human action recognition, which combined action and object recognition by simultaneously parsing and fusing video frames from two perspectives of workspace and nearby objects to avoid confusion caused by similar actions. The aforementioned studies can capture detailed information well, but there are some limitations. Although researchers adopted some processing methods of extracting frames or down-sampling, the computational cost is still relatively high when processing video streams.

In this paper, we propose the human intention recognition method from three aspects. 1) To improve the real-time performance of HRC assembly, we adopt the ST-GCN lightweight model to recognize assembly actions. 2) Considering the problem of disordered parts placement and occlusion, an improved YOLOX algorithm is designed to recognize assembly parts. 3) Considering the flexibility of the assembly process, we study different assembly sequences. By combining the information of assembly actions and assembly parts, we can more accurately recognize the operator's current assembly intention in the complex assembly environment and infer the robot's responsive operation.

3. Framework for operator's intention recognition in the HRC assembly

In the HRC assembly, it is a premise for the robot to accurately recognize the operator's assembly action and understand the operator's intention. As shown in Figure 1, to better realize the collaboration between operator and robot to complete the assembly tasks, this paper proposes the framework for operator's intention recognition in the HRC assembly. The framework consists of two modules. The first module is based on the ST-GCN model to extract skeleton features and recognize the operator's assembly actions. The second module is based on the improved YOLOX algorithm integrating CBAM and the Focal Loss function to recognize assembly parts. Finally, the operator's intention can be accurately recognized by combining assembly action information and assembly part information.

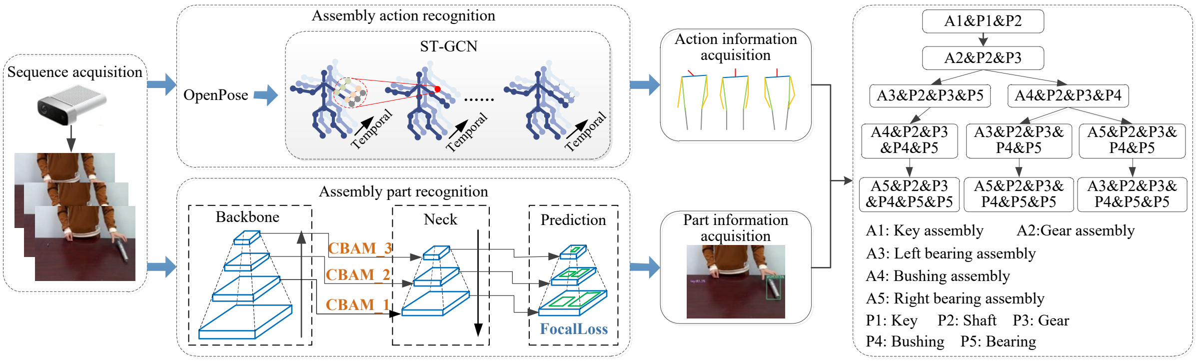


Fig. 1 Framework for operator's intention recognition in the HRC assembly

In the framework, we take the assembly video sequence as input, extract skeleton data based on OpenPose [35], and adopted the ST-GCN model to recognize the operator's assembly actions. Meanwhile, we recognize assembly parts based on the improved YOLOX algorithm. Finally, for the flexibility of assembly sequence, we design a rule-based reasoning method to recognize the operator’s intention by combining the action information with the part information.

In this paper, decelerator assembly tasks are taken as an example to study the operator's assembly action recognition method and assembly part recognition method. Table 1 lists five assembly actions in the decelerator assembly tasks. The decelerator assembly tasks include five types of parts: key, shaft, gear, bushing, and bearing.

Table 1 Assembly actions in the decelerator assembly tasks

Assembly tasks	Assembly actions
1	Key assembly
2	Gear assembly
3	Left bearing assembly
4	Bushing assembly
5	Right bearing assembly

4. ST-GCN model for assembly action recognition

The operator's assembly action recognition in the HRC assembly should be fast and accurate. Compared with the RGB image and optical flows-based action recognition methods, the skeleton-based action recognition method is more lightweight and has a faster inference speed. Therefore, we introduce the ST-GCN model based on the skeleton to recognize the operator’s assembly action.

GCN can process non-Euclidean distance data and extract topological graph features. The spatial-temporal graph $G=(V,E)$ can be constructed on a skeleton sequence. The node set $V=\{v_{ti} \mid t=1,\cdots,T, i=1,2,\cdots,N\}$ represents that a skeleton sequence includes T frames, and each frame contains N joints of the operator. E represents the edge set. The graph covers the joints change information of the assembly action sequence.

The structure of intra-skeleton and inter-frame connection is similar to the convolution operation on images. The CNN model can be extended to space graph to realize space graph convolution operation, which can be written as:

$$f_{out}(v_{ti}) = \sum_{v_{tj}\in B(v_{ti})} f_{in}(P(v_{ti},v_{tj}))\cdot\omega(v_{ti},v_{tj})\big/Z_{ti}(v_{tj}) \qquad (1)$$

This operation consists of the normalizing term $Z_{ti}(v_{tj})$, sampling function $P(v_{ti},v_{tj})$, and weight function $w(v_{ti},v_{tj})$. The sampling function $P(v_{ti},v_{tj})$ is defined on the neighbor set $B(v_{ti})=\{v_{tj}\mid d(v_{tj},v_{ti})\le D\}$ of a node v_{ti}. $d(v_{ti},v_{tj})$ depicts the minimum distance from v_{tj} to v_{ti}.

Then, the concept of the neighborhood is extended to also include temporally connected joints as:

$$B(v_{ti})=\{v_{qj}\mid d(v_{tj},v_{ti})\le K, |q-t|\le \gamma/2\} \qquad (2)$$

The parameter γ controls the temporal range to be included in the neighbor graph.

As shown in Figure 2, the ST-GCN model has 9 layers, each of which contains a spatial GCN and a temporal GCN. Firstly, we can obtain the assembly sequences from the assembly action video streams. Then, skeleton features are extracted from the corresponding frames. Finally, the assembly actions are classified through average pooling and the full connection layer.

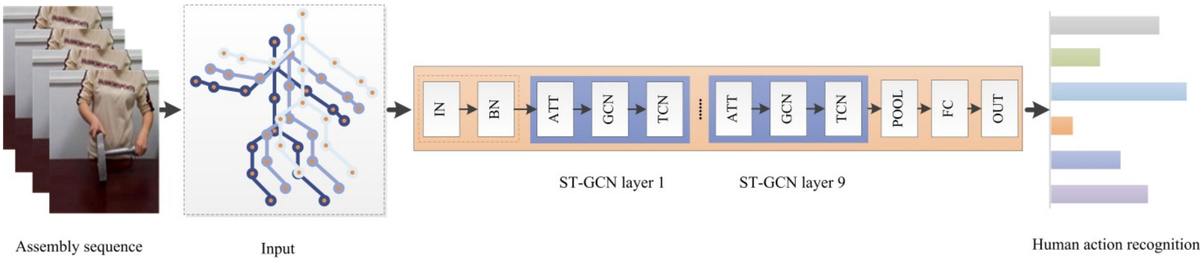


Fig. 2 ST-GCN model

ST-GCN model has three partitioning strategies, i.e. uni-labeling, distance partitioning, and spatial configuration partitioning [18]. In this paper, spatial configuration partitioning is adopted to recognize assembly actions.

5. Improved YOLOX model for assembly part recognition

In HRC assembly, the operator frequently interacts with different assembly parts to accomplish complex product assembly tasks. Since the actions of the operator have high similarity and the same action may relate to different assembly parts corresponding to different assembly sequences, the operator intention recognition in HRC assembly is difficult and has low accuracy, if we only use the results of the skeleton-based operator's assembly action recognition.

To improve the accuracy and effectiveness of operator intention recognition in HRC assembly, the operator's assembly action recognition should be combined with the assembly part recognition. In this section, we use the YOLOX-S network to design an improved YOLOX model that embeds CBAM and the Focal Loss function to recognize the assembly parts.

On the one hand, the attention mechanism refers to the selective attention of human vision to local information, which can focus on the key information and improve the computing performance of the YOLOX-S network. Since CBAM is a lightweight attention module with strong generality, the CBAM module is embedded into the YOLOX algorithm to reduce the background interference, so that the network can better focus on the assembly parts.

On the other hand, YOLOX belongs to the one-stage object detection algorithm, which has the common sample imbalance problem. To solve this problem, we use the Focal Loss function to replace the confidence loss function in the original YOLOX algorithm. The Focal Loss function focuses on increasing the weight of assembly parts that are difficult to classify and improving the recognition performance.

5.1 CBAM

The attention mechanism initially achieved ideal results in machine translation [36] and is gradually applied in the field of computer vision [37]. The CBAM combines the channel attention module (CAM) and spatial attention module (SAM), which is illustrated in Figure 3.

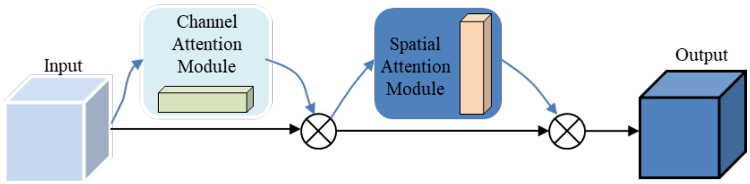


Fig. 3 CBAM schematic

The calculation formula is as follows:

$$F' = M_c(F) \otimes F$$
$$F'' = M_s(F) \otimes F' \quad (3)$$

In CAM, spatial information of the input F is aggregated by using average pooling and max pooling. The generated descriptors are forwarded to multilayer perception and then added. After activation by sigmoid function, channel attention vector $M_C(F)$ is generated, and channel attention output F' is obtained by multiplying $M_C(F)$ and F.

In SAM, the pooling operation is applied along the channel axis, and then the generated feature descriptor is concatenated. The spatial attention vector $M_s(F)$ is obtained after convolution reduction and sigmoid function activation, and the final feature F'' is obtained by multiplying F' and $M_s(F)$. The optimized YOLOX-S network structure embedded with CBAM is shown in Figure 4.

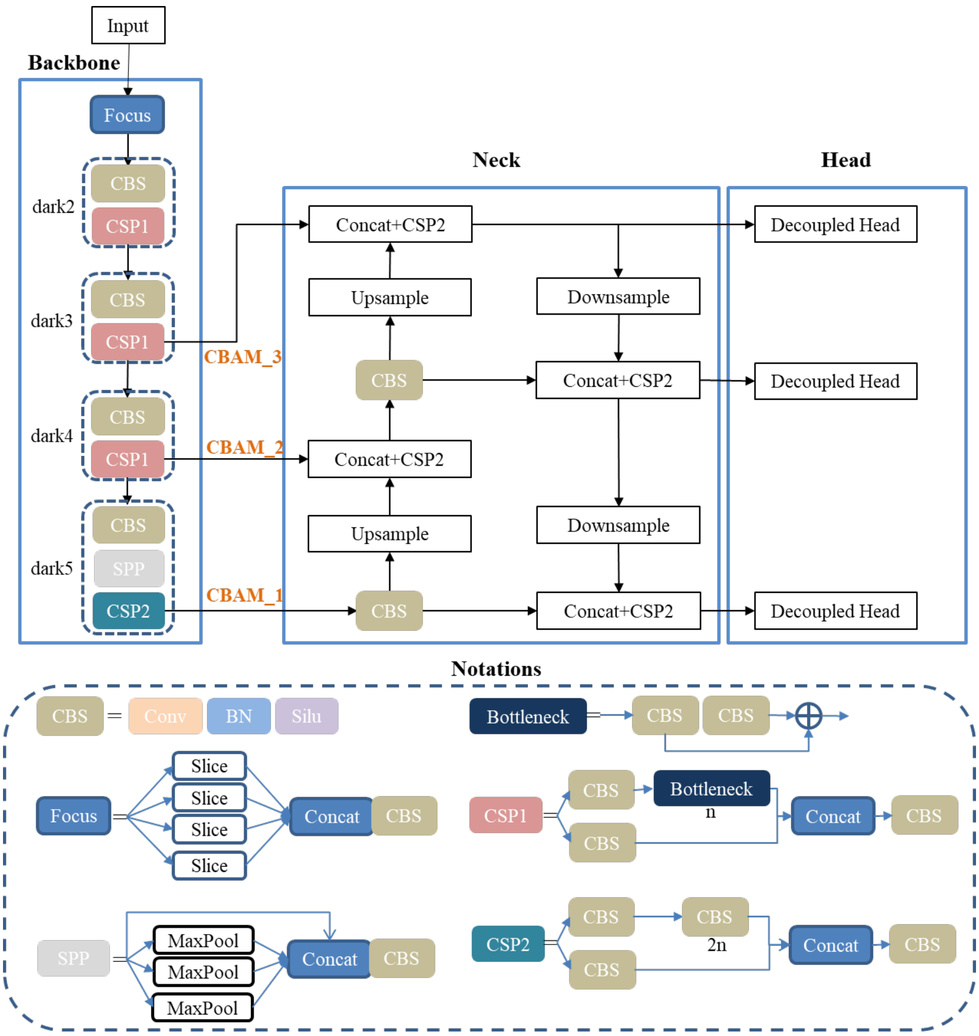


Fig. 4 YOLOX-S structure embedded with CBAM

As shown in Figure 4, we introduce CBAM behind the three valid feature layers of the backbone output, namely dark3, dark4, and dark5 branches. CBAM_1 corresponds to the 1024-dimension channel of the dark_5 branch, CBAM_2 corresponds to the 512-dimension channel of the dark_4 branch, and CBAM_3 corresponds to the 256-dimension channel of the dark_3 branch. On the one hand, the YOLOX model embedded with CBAM can improve its focusing ability on assembly parts. On the other hand, the introduction of CBAM in this paper does not change the number of channels, so it has little effect on the inference speed.

5.2 Focal Loss

Focal Loss is proposed to solve the problem of sample imbalance, which can make the model focus more on the samples that are difficult to classify during training. Chen et al. [38] proposed an extended Focal Loss and generated the class-discriminative Focal Loss for extremely imbalanced object detection toward autonomous driving, which improved the accuracy without requiring more training and inference time. Lee et al. [39] proposed a new deconvolution deep neural network with focal regression loss to detect small traffic lights, and the results show that the introduction of focal regression loss improves detection accuracy.

Assembly parts often have problems of disorder and occlusion, which will affect the recognition of assembly parts. We introduce the Focal Loss function to solve the sample imbalance problem, which makes the network structure more focused on the recognition of disordered and occluded assembly parts. The Focal Loss function is shown in Formula (4):

$$FL(\theta)=\begin{cases}-\alpha(1-\theta)^{\eta}\log(\theta) & if \quad y=1\\ -(1-\alpha)\theta^{\eta}\log(1-\theta) & otherwise\end{cases} \tag{4}$$

where $y\in\{1,-1\}$ specifies the ground-truth class and $\theta\in[0,1]$ is the model's estimated probability for the class with the label $y=1$. α is used to balance the ratio of positive and negative samples. The modulating factor $(1-\theta)^{\eta}$ can reduce the loss contribution of easily classified parts, where $\eta\in[0,5]$.

We used the Focal Loss function to replace the binary cross entropy loss function of the original confidence loss function. As shown in Equation (5), the optimized loss function consists of Intersection over Union (IoU) loss value $Loss_{IoU}$, confidence loss value $Loss_{Focal}$, and classification loss value $Loss_{Class}$.

$$Loss = Loss_{IoU} + Loss_{Focal} + Loss_{Class} \tag{5}$$

6 Case study

This section takes the HRC-based decelerator assembly tasks as an example and establishes datasets for assembly action recognition and assembly part recognition based on the Azure Kinect DK camera to verify the proposed method.

6.1 Assembly action recognition

(1) Creation of assembly action dataset (AAD): AAD is created based on the operator's assembly actions on the HRC-based decelerator assembly tasks. The assembly operations are collected in six directions, i.e. front-left, upper-left, dead ahead, upper-front, front-right, and upper-right. The assembly actions of five operators are recorded. The RGB-D video comprises depth mode (640×576 resolutions) and color mode (1280×720 resolutions). A total of 450 video clips are collected, and each video is 5~6s, to generate the AAD. The dataset prepared in this paper follows the format of the Kinetics dataset [40].

(2) Computing platform: The experiment is carried out with Windows 10 (64bit) system. The CPU card and graphics card are Intel i7-10875H and NVIDIA RTX 2060 (6G), respectively. ST-GCN model is built based on python3.7 language and PyTorch deep learning framework. The training parameters are shown in Table 2, in which the initial learning rate is 0.1 and the learning rate attenuates 0.1 times when the iterative times reach 20, 30, 40, and 50.

Table 2 Training parameters of the ST-GCN model

Parameters	Values
Batch Size	64
Initial learning rate	0.1
Weight decay coefficient	0.0001
Epochs	70

(3) Evaluation index:

1) Top-1 refers to taking the largest probability vector as the assembly action predicted result. If the classification result is correct, then the prediction is correct. This paper uses the Top-1 index to evaluate the assembly action recognition performance of the ST-GCN model on AAD. The calculation of the Top-1 index is shown in Equation (6):

$$Top-1=\sum_{k}^{L}\varphi(class_k^{true}=rank_1(class_k^{pred}))\Big/L \tag{6}$$

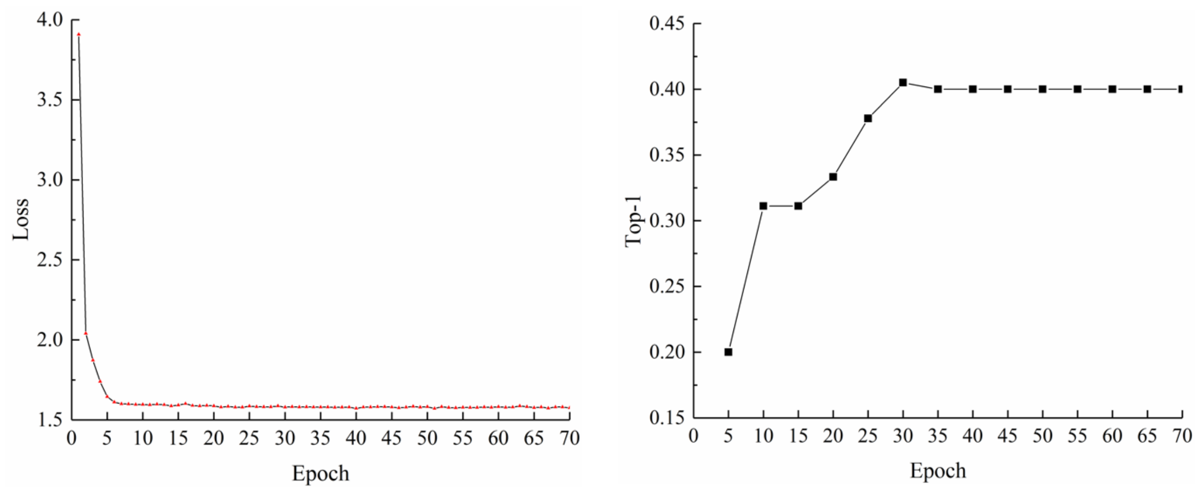
where φ is the judgement function. If the condition is true, the value is 1; otherwise, it is 0. $class_k^{true}$ represents the real classification of the k-th assembly action, and $rank_1(class_k^{pred})$ represents the highest probability in the prediction classification of the k-th assembly action. L is the number of assembly actions.

In this paper, $L = 5$.

2) The parameter is an important index to evaluate the model. The parameters directly affect the memory of the model operation. In this paper, we use the parameters to measure the processing efficiency of the ST-GCN model.

(4) Experimental results and discussions

1) We randomly select 405 video clips as the training dataset and 45 video clips as the test dataset. Figure 5 (a) and Figure 5 (b) demonstrate the loss curve and Top-1 curve of the ST-GCN model training on our AAD.



(a) Loss curve (b) Top-1 curve

Fig. 5 Loss curve and Top-1 curve of ST-GCN model training on AAD

As shown in Figure 5, with the increase of iterative times, the training loss value keeps decreasing and the Top-1 index keeps rising. From the 35th iteration, the Top-1 index levels off and remains at 0.40. Table 3 shows the Top-1 index comparison of the ST-GCN model on the Kinetics dataset [40] and our AAD.

Table 3 Top-1 index comparison of ST-GCN model on Kinetics dataset and our AAD

Datasets	Kinetics	AAD
Top-1	30.7%	40.0%

The Top-1 value obtained by ST-GCN training on AAD is 40.0%, which is better than the recognition effect of the ST-GCN model on the Kinetics dataset, but the accuracy of assembly action recognition is still not very high in general.

Table 4 shows the Top-1 index and sample quantity of five assembly action recognition in the test dataset. In this paper, the test samples of each assembly action in the test dataset are 9.

Table 4 Top-1 index and sample quantity of five assembly action recognition in the test dataset

Assembly actions	Top-1	Number of samples
Key assembly	88.89%	9
Gear assembly	55.56%	9
Left bearing assembly	22.22%	9
Bushing assembly	22.22%	9
Right bearing assembly	11.11%	9
Average/Summation	40%	45

Among the five assembly actions, the recognition accuracy of key assembly action is relatively high, and that of gear assembly action is 55.56%. The right bearing assembly and bushing assembly actions are only different in the distance moved on the shaft, and the overall similarity is high, resulting in low recognition accuracy. In the process of the left bearing assembly and the key assembly, the height of the right hand is different, and the left hand reaches a different height and horizontal position at the end. That is, except for the wrist joint, the positions of the other joint nodes are basically unchanged, resulting in the left bearing assembly action is recognized as the key assembly action. We can know that the accuracy of assembly actions is not high, it is necessary to combine the recognition of assembly parts.

2) As shown in Table 5, we compare the parameters of the ST-GCN model with some classical CNN models, including AlexNet, ResNet18, and VGG16 [41].

Table 5 Comparison of parameters index between ST-GCN model and classical CNN model

Models	ST-GCN	AlexNet	ResNet18	VGG16
Parameters	3.1M	61.1M	11.69M	138.36M

According to the results, the ST-GCN model is naturally more lightweight than CNN-based models. We use the trained model to test the video on the existing workstation. The ST-GCN model runs at approximately 15 frames/s, which basically meets the requirement of online assembly action recognition.

6.2 Assembly part recognition

(1) Creation of assembly part dataset (APD): Based on the above video clips, 4500 images containing five operators, six positions, and five assembly actions are extracted to generate the APD. 1500 images are selected from each assembly action. The number of images containing different assembly parts is shown in Table 6.

Table 6 Number of assembly parts

Classes	Number of samples
Key	1500
Shaft	4500
Gear	3000
Bushing	1500
Bearing	1500

(2) Image annotation: The assembly parts are labeled by the labelImg software [42].

(3) The hardware configuration is consistent with that of the ST-GCN model. The optimized YOLOX network is built based on the python3.7 language and PyTorch deep learning framework. The training parameters of YOLOX are shown in Table 7.

Table 7 Training parameters of the YOLOX algorithm

Parameters	Values
Batch Size	4
Initial learning rate	0.001
Weight decay coefficient	0.0005
Epochs	300

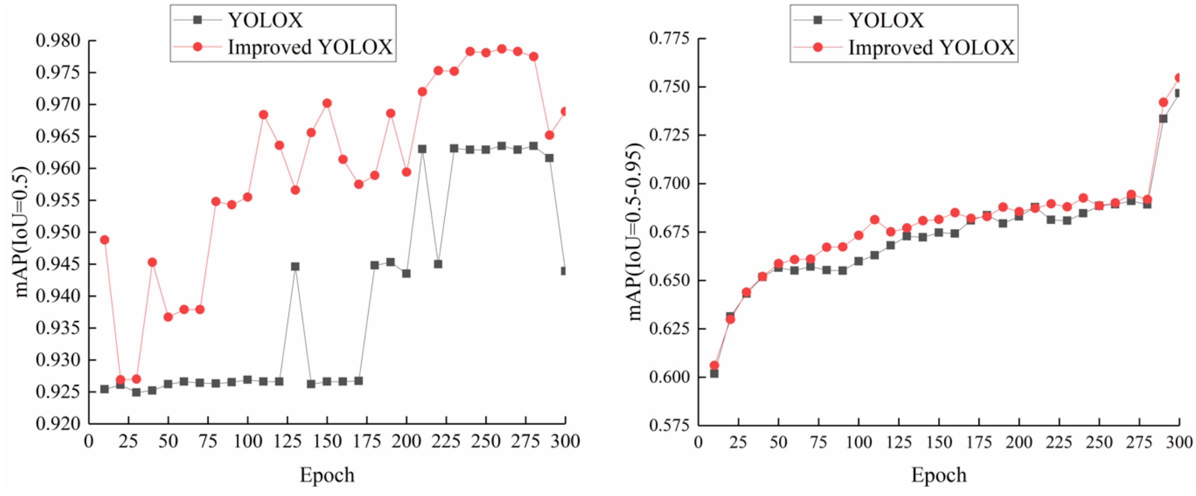
(4) Evaluation index: We use the average precision (AP) and mean average precision (mAP) as the performance evaluation indexes. In this paper, AP is used to evaluate the recognition effect of a certain type of assembly part, as shown in Equation (7):

$$AP = \int_0^1 p(r)dr \tag{7}$$

where p represents the precision and r represents the recall. The mAP is the mean AP value of five categories of assembly parts.

(5) Experimental results and discussions

90% (4050) of 4500 images are randomly selected as the training dataset and 10% (450) as the test dataset. Figure 6(a) shows the mAP (IoU=0.5) comparison curve of original YOLOX and improved YOLOX for assembly part recognition, and Figure 6(b) shows the mAP (IoU=0.5~0.95) comparison curve.



(a) IoU=0.5 (b) IoU=0.5~0.95

Fig. 6 The mAP comparison curve of original YOLOX and improved YOLOX for assembly part

According to the changing trend of mAP in Figure 6, the mAP of the improved YOLOX algorithm for assembly part recognition is generally improved. Table 8 shows the specific comparison results.

Table 8 Comparison results

Model	Key	Shaft	Gear	Bushing	Bearing	mAP (IoU=0.5)	mAP (IoU=0.5~0.95)
Original YOLOX	90.21%	99.78%	100.00%	90.95%	91.01%	94.39%	74.67%
Improved YOLOX	91.29%	99.83%	100.00%	93.32%	100.00%	**96.89%**	75.47%

Table 8 shows that the mAP of the improved YOLOX algorithm for assembly part recognition reaches 96.89% when IoU=0.5, which is 2.50 percentage points higher than that of the original YOLOX algorithm for assembly part recognition. When IoU=0.5~0.95, the mAP of the improved YOLOX algorithm reaches 75.47%, which is 0.80 percentage points higher than the original YOLOX algorithm. According to the AP of key, shaft, gear, bushing, and bearing parts in Table 8, the AP of the gear part is 100.00%, and the AP values of other assembly parts have been improved to different degrees. In particular, the improved YOLOX model shows good performance in the recognition of bushing and bearing, which indicates that the improvement can enhance the recognition effect of occluded or disordered parts. The improved algorithm also enhances the recognition performance of the key to a certain extent, although it is not obvious, which indicates that the improvement is effective for small-size parts recognition. Based on the above analysis, the introduction of CBAM and Focal Loss function can make the YOLOX network focus more on shielded or small-sized assembly parts, and improve the system performance.

The two-stage object detection algorithms can not be well guaranteed in real time. Here, we conduct a comparative experiment with two one-stage object detection algorithms (SSD, YOLOv3) to verify the performance of improved YOLOX for assembly part recognition. We selected mAP (IoU=0.5), mAP (IoU=0.5~0.95), and FPS as the evaluation indexes of algorithm accuracy. The comparison results of different models are shown in Table 9.

Table 9 Performance comparison between SSD, YOLOv3, original YOLOX, and improved YOLOX

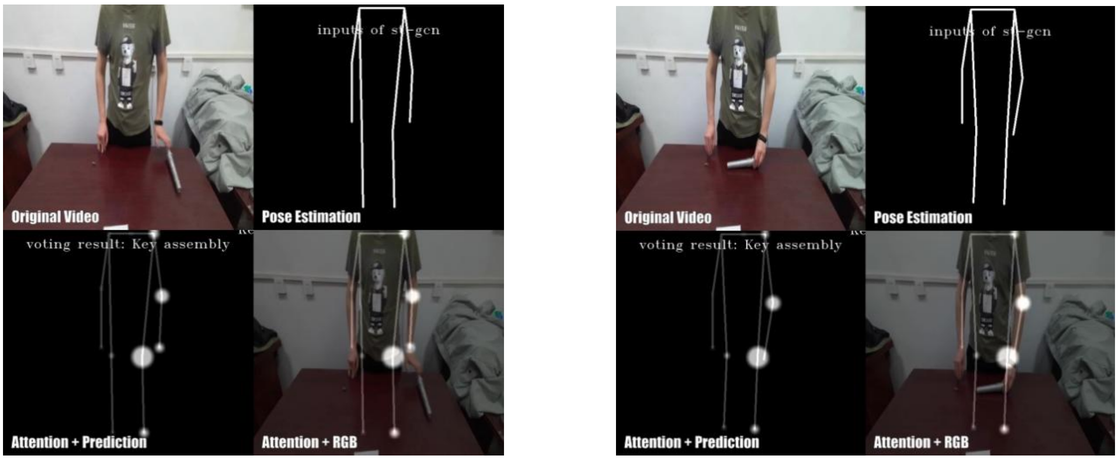
Model	Backbone	mAP (IoU=0.5)	mAP (IoU=0.5~0.95)	FPS
SSD	VGG16	84.75%	45.72%	41.67
YOLOv3	Darknet-53	97.48%	73.89%	21.79
Original YOLOX	Modified CSP v5	94.39%	74.67%	55.49
Improved YOLOX	Modified CSP v5	96.89%	75.47%	54.37

Table 9 shows that the improved YOLOX model proposed in this paper is superior to SSD and original YOLOX models in the recognition accuracy of assembly parts. The inference speed of the YOLOX algorithm is faster than that of the SSD algorithm, and the FPS value of the improved YOLOX algorithm has little change, only decreasing by 1.12. YOLOv3 has the higher recognition accuracy for assembly parts than improved YOLOX but has a low FPS value. In conclusion, considering the recognition accuracy and speed, the improved YOLOX model can accurately capture the features of assembly parts and improve the recognition performance.

6.3 Operator's intention recognition based on assembly action and assembly part information

6.3.1 Assembly action information

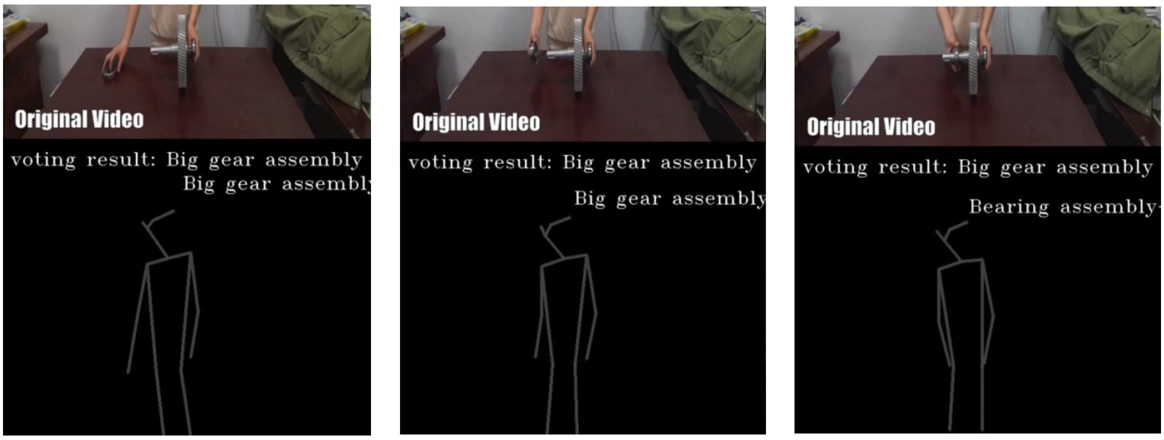
As shown in Figure 7, to observe the performance of the trained ST-GCN model for assembly action recognition, frame 1 and frame 40 of "key" assembly action recognition are selected for analysis. It can be seen from Figure 7 that the key assembly action is correctly classified by the ST-GCN model.



(a) Frame 1 (b) Frame 40

Fig. 7 Recognition results of key assembly action

Figure 8 shows partially captured pictures from frame 1, frame 30, and frame 70 of right bearing assembly action recognition. When recognizing the right bearing assembly action, the assembly actions in frame 1 and frame 30 are wrongly recognized as the gear assembly, and the assembly action in frame 70 is correctly recognized as the right bearing assembly. On that basis, this assembly action is wrongly classified as the gear assembly.



(a) Frame 1 (b) Frame 30 (c) Frame 70

Fig. 8 Recognition results of right bearing assembly action

According to our test results and analysis, it can be concluded that the trained ST-GCN model can recognize some of the decelerator assembly actions accurately. However, for some other assembly actions, there are problems of occlusion, the similarity of different actions, and other factors, besides, large errors exist in inferring the location of occluded joints, which leads to errors in recognizing assembly actions.

The accuracy of the operator's assembly action recognition is not very high due to high action similarity and limited body movement range reasons in HRC decelerator assembly, and sometimes the assembly actions are wrongly recognized, so we need to combine it with assembly part recognition results to better recognize the operator's intention for processing the assembly tasks.

6.3.2 Assembly part information

The assembly part recognition information can assist recognize the operator's intentions. In the situation that the operator's assembly action is accurately recognized, the assembly parts recognition strengthens the correct result. In the situation that the operator's assembly action is not correctly recognized, the assembly parts recognition will help to rectify the wrong result.

Figure 9 shows the assembly part recognition results of frame 1 and frame 40 in the key assembly process. It can be seen from Figure 9 that two parts have been recognized in frame 1 and frame 40, including the shaft and key. The AP values of key part recognition in frame 1 and frame 40 are 86.8% and 79.5% respectively. While the AP values of shaft part recognition in frame 1 and frame 40 are 93.1% and 92.6% respectively, a bit higher than the key part recognition results since the size of the key is relatively small and the YOLOX model is not very good at small-size object detection.

(a) Frame 1 (b) Frame 40

Fig. 9 Part recognition results in the key assembly process

(a) Frame 1 (b) Frame 30 (c) Frame 70

Fig. 10 Part recognition results in the right bearing assembly process

Figure 10 shows the assembly part recognition results of frames 1, 30, and 70 in the right bearing assembly process. Five parts have been recognized in frames 1, 30, and 70, including the left bearing, gear, shaft, bushing, and right bearing. To clearly show the part recognition results in the right bearing assembly process, Table 10 lists the AP corresponding to frames 1, 30, and 70 in Figure 10. It can be seen that the AP of shaft recognition in Figure 10 is relatively lower than that in Figure 9 because there are some shelters (i.e. gear, bearing, bushing) that affect the recognition accuracy. The AP of gear recognition is relatively high due to its large size and distinct features, and the AP of the key recognition is relatively low due to its small size. The AP of bearing recognition also varies with the change of position.

Table 10 AP of part recognition in the right bearing assembly process

Frame	Frame 1		Frame 30		Frame 70	
	Part	AP	Part	AP	Part	AP
Part recognition	Left bearing	65.4%	Left bearing	66.2%	Left bearing	74.4%
	Gear	80.9%	Gear	81.9%	Gear	83.7%
	Shaft	66.1%	Shaft	65.7%	Shaft	73.5%
	Bushing	65.0%	Bushing	63.9%	Bushing	78.1%
	Right bearing	81.9%	Right bearing	72.9%	Right bearing	74.0%

6.3.3 Operator's intention recognition

The case study includes five types of assembly actions, considering the flexibility in the assembly process, the sequence of assembly actions is optional. As shown in Table 11, in order to express the assembly sequences more clearly, we number the assembly actions and assembly parts.

Table 11 Numbers of assembly actions and parts

Assembly actions and parts	Numbers
Key assembly	A1
Gear assembly	A2
Left bearing assembly	A3
Bushing assembly	A4
Right bearing assembly	A5
Key	P1
Shaft	P2
Gear	P3
Bushing	P4
Bearing	P5

The assembly sequences in this case study are shown in Figure 11(a). We can know that it is optional to perform the bushing assembly or left bearing assembly after the gear assembly, and the subsequent assembly actions will also change accordingly. As shown in Figure 11(b), the sequences of corresponding assembly parts will also change.

(a) The sequences of assembly actions (b) The sequences of assembly parts

Fig. 11 The sequences of actions and parts in the assembly process

In this case study, there are three alternatives. In different alternatives, the recognizable assembly parts corresponding to the same assembly action may be different, which will change the recognition result of human behavior intention and the response of the robot. In this paper, we combine the recognition of assembly actions and assembly parts to accurately recognize the operator's assembly behavior intention, and infer the assembly tasks that have been completed and the next assembly task to be carried out.

As shown in Figure 7, it is recognized that the operator is performing the key assembly action. Combined with the shaft and key parts recognized in Figure 9, we can accurately recognize that the operator is assembling the key part. Clearly, the next assembly task is gear assembly.

As shown in Figure 8, there is uncertainty about assembly action recognition, sometimes the assembly action is recognized as right bearing assembly and sometimes as gear assembly. Combined with the recognized parts information in Section 6.3.2, if left bearing, gear, shaft, bushing, and right bearing parts are recognized, and the total number of parts is five, the assembly behavior intention will be recognized as performing the fifth assembly task. If gear and shaft parts are recognized and the number of parts is two, the assembly behavior intention will be recognized as performing the first assembly task. Combining the recognition information of assembly action in Figure 8 with the assembly parts in Figure 10 and Table 10, it can be determined with high confidence that the operator is performing the right bearing assembly task because the assembly parts include left bearing, gear, shaft, bushing, and right bearing and the number of parts is five. At the same time, according to the recognition information of assembly parts and assembly actions, we can infer that all the five assembly tasks have been completed and the robot will leave.

We comprehensively analyze the integration of assembly actions and assembly parts. Figure 12 shows the human assembly behavior intention recognition based on logical rules. The assembly action and assembly parts corresponding to each task are clearly defined. When the key and shaft are recognized simultaneously, we can infer that the operator is currently performing the first assembly task. According to the recognition results of assembly actions, we can also judge the ongoing assembly task. In this paper, we consider that the operator is currently performing the first assembly task when the conditions for both the assembly action and the assembly parts are met simultaneously. Based on logical rules, we can know that the operator needs to obtain gear before performing the second assembly task and then assemble gear from the right side. We can infer that the "A1&P1&P2" condition triggers the robot to perform the response of gear grabbing. By analogy, the operator can be recognized as performing the second assembly task based on the "A2&P2&P3" condition.

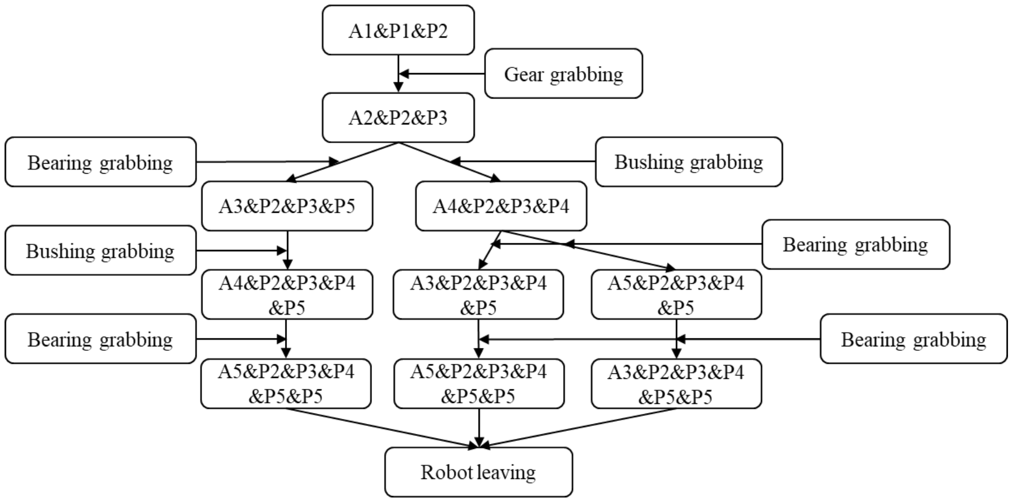


Fig. 12 Human assembly behavior intention recognition based on logical rules

However, in the actual assembly process, the assembly sequence selected by different operators is inconsistent. As shown in Figure 12, after completing the gear assembly task, the operator has the option to perform either bushing assembly or left bearing assembly. If the bushing assembly task is completed first, after the robot grabs the bearing, the operator can also choose to finish the left bearing assembly or the right bearing assembly first. We can know that the same assembly action corresponds to different assembly parts, and the operator's behavioral intention will also change accordingly. When we recognize the operator performing the right bearing assembly action through the ST-GCN model, we recognize the operator's assembly intention according to the detected assembly parts. If five assembly parts are recognized, the operator is inferred to be currently completing the last task and the robot can leave. If four assembly parts are recognized, it is inferred that the operator also needs to complete the left bearing assembly task next, and the robot also needs to grasp the bearing. We clearly understand that the operator's behavioral intention can be accurately recognized from the combined information of the assembly action and the assembly parts.

6.3.4 Discussion

(1) The human behavior intention recognition method proposed in this paper mainly solves the following problems. First, the skeleton-based lightweight model is adopted to recognize the operator’s assembly actions. However, the accuracy of assembly action recognition is limited, which cannot always ensure the correct recognition of assembly action. Moreover, for different assembly sequences, the assembly parts corresponding to the same assembly action may be different. Therefore, based on logic rules, we combine skeleton-based assembly action and image-based assembly part recognition information to accurately recognize the operator's assembly intention in a complex assembly environment. As for the human behavioral intention recognition method, an improved YOLOX model integrating CBAM and Focal Loss function is developed to recognize the assembly part in the HRC assembly. We can understand what kind of assembly task the operator is currently performing and what kind of assembly parts are being assembled more accurately. In the future, we will consider the weight distribution and dependence between skeleton features and part features to optimize the decision model for operator intention recognition.

(2) We use the ST-GCN model to recognize the operator’s assembly actions based on the body skeleton. However, the accuracy is relatively low due to many reasons, such as occlusion, the similarity of different actions, and the inconsistency of different operators' actions. In the future, we will on the one hand optimize the ST-GCN model by integrating the time series model and improving the pose estimation algorithm from the perspective of the joints’ number and location. On the other hand, we will standardize the body movement of

different assembly actions and perfect AAD by adding assembly action samples.

(3) In this paper, five assembly tasks are taken as research cases to describe the trigger rules of the robot. In the future, we expect to improve the robot's proactive decision-making ability through the deep reinforcement learning model, and then combine the operator's intention recognition with the adaptive control of the robot. In this process, we need to think about how the robot can actively make decisions to assist the operator when performing the assembly task. For example, the operator's assembly action is biased because of the different amplitude, and the robot needs to adaptively adjust the running state. If the robot makes a mistake, the robot needs to make a proactive decision to update the assembly operation. At the same time, we will take the whole decelerator assembly process as the research object, improve the experiments related to human behavior intention recognition and robot adaptive control, and verify the feasibility of the theoretical method. By improving the robot's proactive decision-making ability, HRC assembly can be better promoted.

7. Conclusion

Assembly is important to ensure product quality. Human-robot collaborative (HRC) assembly has become prevailing due to its advantages of repeatability, high accuracy, hard work bearing, and flexibility. In HRC assembly, how to recognize the operator’s assembly intentions accurately is a vital problem for the robot during the assembly process. In this paper, we propose a human intention recognition method by combining assembly action information and assembly part information, which improves the accuracy of intention recognition by combining assembly context information. On the one hand, ST-GCN is adopted to dynamically recognize the operator's actions in HRC assembly based on a video dataset. On the other hand, an improved YOLOX algorithm is designed to recognize assembly parts based on the image dataset derived from the video dataset. In the improved YOLOX algorithm, CBAM is introduced to improve the focusing capability of the YOLOX network on the assembly parts, and the Focal Loss function is introduced to focus on disordered, occluded, and small-sized assembly parts.

In the case study, taking the HRC decelerator assembly task as an example, ST-GCN is used to recognize the operator’s assembly actions. The results show that the assembly actions with high similarity have low recognition accuracy, and the operator's intention could not be accurately inferred only through the recognition of assembly actions. The improved YOLOX algorithm is used to recognize assembly parts on APD, and the results show that the improved YOLOX model can improve the recognition accuracy of the parts with obscure features, occlusion, and small size. On this basis, combined with the recognition results of assembly actions and assembly parts, the current process can be further inferred based on the rule reasoning, which effectively recognizes the operator's intention and infers the robot’s responsive operation in the HRC assembly. This is beneficial to promote the robot’s cognitive intelligence and accelerate HRC assembly.

Future research can be also done to: 1) improve the recognition accuracy of the operator’s assembly actions by designing an improved ST-GCN model and standardizing the operator’s body movements; 2) optimize the decision model of operator intention recognition and study the end-to-end intention recognition method; 3) study the adaptive decision-making of the robot based on deep reinforcement learning.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No.51705030), China

Postdoctoral Science Foundation (No.2021M700528), Fundamental Research Funds for the Central Universities, CHD (No.300102250201), and the General Research Fund (GRF) from the Research Grants Council of the Hong Kong Special Administrative Region, China (No. PolyU 15210222).

References

[1] J. Liu, Q. Sun, H. Cheng, X. Liu, X. Ding, S. Liu, H. Xiong, The State-of-the-art, Connotation and Developing Trends of the Products Assembly Technology, J. Mech. Eng. 54 (11) (2018) 2-28, https://doi.org/10.3901/jme.2018.11.002.

[2] J.G. Huo, F.T.S. Chan, C.K.M. Lee, J.O. Strandhagen, B. Niu, Smart control of the assembly process with a fuzzy control system in the context of Industry 4.0, Adv. Eng. Inform. 43 (2020) 101031, https://doi.org/10.1016/j.aei.2019.101031.

[3] L. Wang, R. Gao, J. Vancza, J. Kruger, X.V. Wang, S. Makris, G. Chryssolouris, Symbiotic human-robot collaborative assembly, CIRP Ann-Manuf. Technol. 68 (2) (2019) 701-726, https://doi.org/10.1016/j.cirp.2019.05.002.

[4] S.F. Li, P. Zheng, J.M. Fan, L.H. Wang, Toward Proactive Human-Robot Collaborative Assembly: A Multimodal Transfer-Learning-Enabled Action Prediction Approach, IEEE Trans. Ind. Electron. 69 (8) (2022) 8579-8588, https://doi.org/10.1109/tie.2021.3105977.

[5] J.M. Fan, P. Zheng, S.F. Li, Vision-based holistic scene understanding towards proactive human-robot collaboration, Robot. Comput.-Integr. Manuf. 75 (2022) 102304, https://doi.org/10.1016/j.rcim.2021.102304.

[6] A. Casalino, C. Messeri, M. Pozzi, A.M. Zanchettin, P. Rocco, D. Prattichizzo, Operator Awareness in Human-Robot Collaboration Through Wearable Vibrotactile Feedback, IEEE Robot. Autom. Lett. 3 (4) (2018) 4289-4296, https://doi.org/10.1109/lra.2018.2865034.

[7] F. Wallhoff, J. Blume, A. Bannat, W. Rosel, C. Lenz, A. Knoll, A skill-based approach towards hybrid assembly, Adv. Eng. Inform. 24 (3) (2010) 329-339, https://doi.org/10.1016/j.aei.2010.05.013.

[8] A. Mohammed, L.H. Wang, Brainwaves driven human-robot collaborative assembly, CIRP Ann-Manuf. Technol. 67 (1) (2018) 13-16, https://doi.org/10.1016/j.cirp.2018.04.048.

[9] Z. Gao, J.M. Song, H. Zhang, A.A. Liu, Y.B. Xue, G.P. Xu, Human Action Recognition Via Multi-modality Information, J. Electr. Eng. Technol. 9 (2) (2014) 739-748, https://doi.org/10.5370/jeet.2014.9.2.739.

[10] L.M. Wang, Y.J. Xiong, Z. Wang, Y. Qiao, D.H. Lin, X.O. Tang, L. Van Gool, Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, in: B. Leibe, J. Matas, N. Sebe, M. Welling (eds), Computer Vision - ECCV 2016, 2016, pp. 20-36, https://doi.org/10.1007/978-3-319-46484-8_2.

[11] N. Sun, L. Leng, J.X. Liu, G. Han, Multi-stream slowFast graph convolutional networks for skeleton-based action recognition, Image Vis. Comput. 109 (2021) 104141, https://doi.org/10.1016/j.imavis.2021.104141.

[12] S.P. Zhu, L.M. Xia, Human Action Recognition Based on Fusion Features Extraction of Adaptive Background Subtraction and Optical Flow Model, Math. Probl. Eng. 2015 (2015) 387464, https://doi.org/10.1155/2015/387464.

[13] K. Sidor, M. Wysocki, Recognition of Human Activities Using Depth Maps and the Viewpoint Feature Histogram Descriptor, Sensors 20 (10) (2020) 2940, https://doi.org/10.3390/s20102940.

[14] M. Urgo, M. Tarabini, T. Tolio, A human modelling and monitoring approach to support the execution of manufacturing operations, CIRP Ann-Manuf. Technol. 68 (1) (2019) 5-8, https://doi.org/10.1016/j.cirp.2019.04.052.

[15] L.Z. Hu, J.H. Xu, Learning Discriminative Representation for Skeletal Action Recognition Using LSTM

Networks, in: M. Felsberg, A. Heyden, N. Krüger (eds), Computer Analysis of Images and Patterns, 2017, pp. 94-104, https://doi.org/10.1007/978-3-319-64698-5_9.

[16] M. Naveenkumar, S. Domnic, Deep ensemble network using distance maps and body part features for skeleton based action recognition, Pattern Recognit. 100 (2020) 107125, https://doi.org/10.1016/j.patcog.2019.107125.

[17] M. Al-Amin, R.W. Qin, M. Moniruzzaman, Z.Z. Yin, W.J. Tao, M.C. Leu, An individualized system of skeletal data-based CNN classifiers for action recognition in manufacturing assembly, J. Intell. Manuf. (2021) 1-17, https://doi.org/10.1007/s10845-021-01815-x.

[18] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: 32nd AAAI Conference on Artificial Intelligence, 2018, pp. 7444-7452, https://doi.org/10.48550/arXiv.1801.07455.

[19] L. Shi, Y.F. Zhang, J. Cheng, H.Q. Lu, Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12018-12027, https://doi.org/10.1109/cvpr.2019.01230.

[20] L. Shi, Y.F. Zhang, J. Cheng, H.Q. Lu, Skeleton-Based Action Recognition with Directed Graph Neural Networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7904-7913, https://doi.org/10.1109/cvpr.2019.00810.

[21] K.J. Wang, D. Santoso, A smart operator advice model by deep learning for motion recognition in human-robot coexisting assembly line, Int. J. Adv. Manuf. Technol. 119 (1-2) (2022) 865-884, https://doi.org/10.1007/s00170-021-08319-1.

[22] J.M. Fu, X. Sun, Z.R. Wang, K. Fu, An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images, IEEE Trans. Geosci. Remote Sensing 59 (2) (2021) 1331-1344, https://doi.org/10.1109/TGRS.2020.3005151.

[23] J. Lu, J. He, Z. Li, Y. Zhou, A Survey of Target Detection Based on Deep Learning, Electron. Opt. Control 27 (5) (2020) 56-63, https://doi.org/10.3969/j.issn.1671-637X.2020.05.012.

[24] K.J. Wang, D.A. Rizqi, H.P. Nguyen, Skill transfer support model based on deep learning, J. Intell. Manuf. 32 (4) (2021) 1129-1146, http://doi.org/10.1007/s10845-020-01606-w.

[25] S. Back, J. Kim, R. Kang, S. Choi, K. Lee, Segmenting unseen industrial components in a heavy clutter using rgb-d fusion and synthetic data, in: 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 828-832, http://doi.org/10.1109/ICIP40778.2020.9190804.

[26] Z.J. Duan, S.B. Li, J.J. Hu, J. Yang, Z. Wang, Review of Deep Learning Based Object Detection Methods and Their Mainstream Frameworks, Laser Optoelectron. Prog. 57 (12) (2020) 120005, https://doi.org/10.3788/lop57.120005.

[27] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: Exceeding yolo series in 2021, arXiv preprint arXiv:2107.08430 (2021), https://doi.org/10.48550/arXiv.2107.08430.

[28] Z. Li, F. Zhou, FSSD: feature fusion single shot multibox detector, arXiv preprint arXiv:2107.08430 (2017), https://doi.org/10.48550/arXiv.1712.00960.

[29] G. Andrianakos, N. Dimitropoulos, G. Michalos, S. Makris, An approach for monitoring the execution of human based assembly operations using machine learning, in: F. Dietrich, N. Krenkel (eds), 7th CIRP Global Web Conference on Towards Shifted Production Value Stream Patterns through Inference of Data, Models, and Technology (CIRPE 2019), 2019, pp. 198-203, https://doi.org/10.1016/j.procir.2020.01.040.

[30] K. Zidek, P. Lazorik, J. Pitel, A. Hosovsky, An Automated Training of Deep Learning Networks by 3D Virtual Models for Object Recognition, Symmetry-Basel 11 (4) (2019) 496, https://doi.org/10.3390/sym11040496.

[31] S.H. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional Block Attention Module, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss, Computer Vision - ECCV 2018, 2018, pp. 3-19, https://doi.org/10.1007/978-3-030-01234-2_1.

[32] T.Y. Lin, P. Goyal, R. Girshick, K.M. He, P. Dollar, Focal Loss for Dense Object Detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 318-327, https://doi.org/10.1109/tpami.2018.2858826.

[33] P. Wang, H.Y. Liu, L.H. Wang, R.X. Gao, Deep learning-based human motion recognition for predictive context-aware human-robot collaboration, CIRP Ann-Manuf. Technol. 67 (1) (2018) 17-20, https://doi.org/10.1016/j.cirp.2018.04.066.

[34] J.J. Zhang, P. Wang, R.X. Gao, Hybrid machine learning for human action recognition and prediction in assembly, Robot. Comput.-Integr. Manuf. 72 (2021) 102184, https://doi.org/10.1016/j.rcim.2021.102184.

[35] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, Y. Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2021) 172-186, https://doi.org/10.1109/tpami.2019.2929257.

[36] L. Huang, W.Y. Chen, Y.G. Liu, H. Zhang, H. Qu, Improving neural machine translation using gated state network and focal adaptive attention networtk, Neural Comput. Appl. 33 (23) (2021) 15955-15967, https://doi.org/10.1007/s00521-021-06444-2.

[37] M.H. Guo, T.X. Xu, J.J. Liu, Z.N. Liu, P.T. Jiang, T.J. Mu, S.H. Zhang, R.R. Martin, M.M. Cheng, S.M. Hu, Attention mechanisms in computer vision: A survey, Comput. Vis. Media 8 (3) (2022) 331-368, https://doi.org/10.1007/s41095-022-0271-y.

[38] G.C. Chen, H.B. Qin, Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving, Visual Comput. 38 (3) (2022) 1051-1063, http://doi.org/10.1007/s00371-021-02067-9.

[39] E. Lee, D. Kim, Accurate traffic light detection using deep neural network with focal regression loss, Image Vis. Comput. 87 (2019) 24-36, http://doi.org/10.1016/j.imavis.2019.04.003.

[40] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, The kinetics human action video dataset, arXiv preprint arXiv:1705.06950, 2017, https://doi.org/10.48550/arXiv.1705.06950.

[41] A. Ben Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangere, F. Ghiringhelli, G. Forestier, C. Wemmert, Deep learning for colon cancer histopathological images analysis, 136 (2021) 104730, https://doi.org/10.1016/j.compbiomed.2021.104730.

[42] D. Tzutalin, LabelImg, 2015, https://github.com/tzutalin/labelImg.