Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 104 (2021) 1795–1800

www.elsevier.com/locate/procedia

54th CIRP Conference on Manufacturing Systems

Transfer Learning-enabled Action Recognition for Human-robot Collaborative Assembly

Shufei Li[a], Junming Fan[a], Pai Zheng[a,*], Lihui Wang[b,*]

[a]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[b]Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

* Corresponding author. E-mail address: pai.zheng@polyu.edu.hk

Abstract

Human-robot collaboration (HRC) is critical to today's tendency towards high-flexible assembly in manufacturing. Human action recognition, as one of the core prerequisites for HRC, enables industrial robots to understand human intentions and to execute planning adaptively. However, existing deep learning-based action recognition methods rely heavily on a huge amount of annotation data, which may not be effective or realistic in practice. Therefore, a transfer learning-enabled action recognition approach is proposed in this research to facilitate robot reactive control in HRC assembly. Meanwhile, a decision-making mechanism for robotic planning is introduced as well. Lastly, the proposed approach is evaluated in an aircraft bracket assembly scenario to reveal its significance.

Keywords: Transfer learning; action recognition; human-robot collaboration assembly; domain adaptation

1. Introduction

With modern manufacturing shifting from mass production to mass personalization, industrial robots have been of rising demands for adaptive control and seamless cooperation with human operators in a shared workspace [1]. In the context of flexible automation, human-robot collaboration (HRC) aims to integrate the accuracy and strength of robots with cognitive ability and flexibility of humans in the execution loop [2]. One major pillar to achieve this is that robots dynamically plan safe reactions responded to human activities and intentions [3]. Hence, human action recognition, as the prerequisite, plays a critical role in efficient HRC, which can result in higher overall productivity in customization-oriented manufacturing.

In recent years, cutting-edge deep learning techniques have led to numerous resounding success in the prevailing human action recognition field, where thousands of video samples and millions of frames have been collected from different human subjects in daily activities [4]. For example, in industrial scenarios of surveillance systems, temporal representations of human actions were distilled from sequences of frames via convolutional neural networks (CNN) and long short-term memory (LSTM) [5].

Nevertheless, above prominent capabilities of perception intelligence are restricted by the following two constraints: 1) vast amounts of data should be available and annotated with labels; and 2) training data and testing data are subjected to the same probability distribution, instead of suffering feature variances caused by different working conditions among manufacturing. In real workplace settings, it is difficult or even unrealistic to develop such a typical dataset, which covers potential action representations of all possible subjects. Hence, there is an urgency to transfer knowledge learned from daily human action to intention prediction in those complex assembly scenarios, so that efficient HRC can be achieved for smart manufacturing.

Aiming to fill this research gap, this paper proposes a transfer learning-enabled action recognition approach for HRC

10.1016/j.procir.2021.11.303

assembly, aiming to allow unattained cooperation efficiency between human and robots in a shared workspace. The deep transfer learning approach contains an action recognition module (i.e., spatial temporal GCN (ST-GCN)) and a domain adaptation component, where maximum mean discrepancy (MMD) is utilized to reduce the discrepancy of action features between source and domain datasets. Hence, both stationary and non-stationary human activities of assembly tasks can be predicted, and preprogramming-free robots can adaptively execute assembly instructions in response to the semantic knowledge of operator intentions.

The rest of paper is organized as follows. Section 2 introduces related works of HRC assembly, human action recognition and transfer learning strategies in industrial tasks. Our proposed methods including transfer learning-enabled human activity recognition and the decision-making mechanism for robots, are described in section 3. Comparison experiments of human action recognition and a typical application of HRC bracket assembly in aircraft cabins are depicted in section 4. Conclusions and future works are summarized in section 5.

2. Related work

This section summarizes HRC assembly applications and their adaptive decision-making capabilities. Dominant action recognition and transfer learning methods which can facilitate robot intelligence are introduced as well.

2.1. Human-robot collaboration assembly

Different from traditional industrial robots segregated in a closed space, collaborative robots are free from preprogramming control instructions and work side by side with human operators in close proximity [6]. Therefore, HRC is characterized by the integration of robotic automation and human flexibility, which puts it at the leading edge in terms of increasing flexibility requirements in modern manufacturing, especially for human-centered assembly.

To interact with human operators and act high-level teamwork skills in real manufacturing settings, adaptive decision-making is a persistent objective for collaborative robots, such as active collision avoidance [7] and reactive robotic planning [8]. One of the most notable areas focuses on immersing context-aware intelligence in the entire human-robot organization. With context-aware monitoring of a shared workspace, operators can freely move, and the robot dynamically updates its planning as a response [9]. Another focus is on permeating semantic knowledge in HRC systems. Especially for task oriented HRC, robotic collaborative skills, including trajectory planning and adaptive subtask planning, can be achieved by inferring semantic knowledge of humans' continuous movement and action prediction [10].

2.2. Industrial action recognition and intention analysis

Human action recognition plays a critical role in HRC, enabling robots to understand human behavior and to assist the worker in a proactive manner. Nowadays, the advanced computer vision-based action recognition methods can directly distill human activity representations from videos, such as optical flow trajectories [11] and skeleton motions.

Slightly different from that in daily video surveillance, human action recognition methods in manufacturing pay more attention to the ongoing activity prediction, unlabeled video classification and 3D action trajectory, which ensure robots to react with a safe and suitable response. A progressive filtering approach was introduced to recognize human action expressions captured by Kinect as early as possible [12]. Robots can improve their performance of reaction based on the identified ongoing human action. In particular, 3D human action recognition opens the way to allow collaborative robots to action with accurateness and resilience. Deep learning network was constructed to extract features of 3D human activities in [13].

2.3. Transfer learning in smart manufacturing

Prevailing deep learning methods make it available to learn from industrial data for manufacturing intelligence [14]. Nevertheless, their notable capability of knowledge learning is realized on the essential assumptions: 1) large amounts of data are available 2) training data and testing data are subjected to the same distribution spaces [15]. For human action recognition in HRC, there is few data of operators' assembly motion and these data suffer huge distribution discrepancy caused by different working conditions and human body characteristic. Therefore, traditional deep learning approaches fail to provide efficient implementations.

Transfer learning (e.g., finetune, adaptation layer and generative adversarial nets) builds a bridge for action recognition in relevant but not the same scenarios, as it can learn sharing knowledge and extract invariant features between source and target data. For example, in dynamic production process optimization, finetune-based prediction model was utilized to extract latent sharing features between the historical production records and real on-time delivery of orders [16]. For fault diagnostics across diverse working conditions and devices, MMD was introduced to a domain adaption module to minimize the probability distribution distance between high-level extracted features between source and target datasets [17].

3. Methodology

In this work, a transfer learning-based action prediction approach is proposed for efficient HRC assembly. As shown in Fig. 1, HRC assembly mainly contains two parts, i.e., human action recognition and robotic adaptive control, both of which are introduced accordingly as follows. HRC assembly mainly goes through these steps, 1) data sensing and pre-processing 2) knowledge distilling and action recognition from sampling videos, and 3) robotic decision making and reaction in response to learned semantic knowledge. Among them, human action recognition and robotic adaptive control are two vital parts, both of which are introduced accordingly.

3.1. Transfer learning-enabled human action recognition

Human pose acquisition is the prerequisite for action recognition. As present in the preprocessing part in Fig. 1, Azure Kinect is utilized to capture operators' motion for an ongoing task in assembly workplaces. Then, human pose can be obtained via Openpose toolbox [18], which predicted body joints of the operator in consecutive frames of a video.

With skeleton joints of human bodies, a transfer learning-based ST-GCN [4] architecture is proposed to learn knowledge of human actions from these data. Our proposed deep transfer learning network includes three blocks, i.e., a ST-GCN feature extractor, an action classifier and one domain adaptation module, as shown in the middle part in Fig. 1. The feature extractor distills latent representations of human actions between source and target domains. The following one is the action classifier, which partitions extracted high-level action features into different categories. The feature extractor and action classifier make up the action recognition module. The domain adaptation module is connected to the feature extractor. By reducing the distribution discrepancy of source and target data, the domain adaptation layer enables the extractor to learn sharing and domain-invariant features.

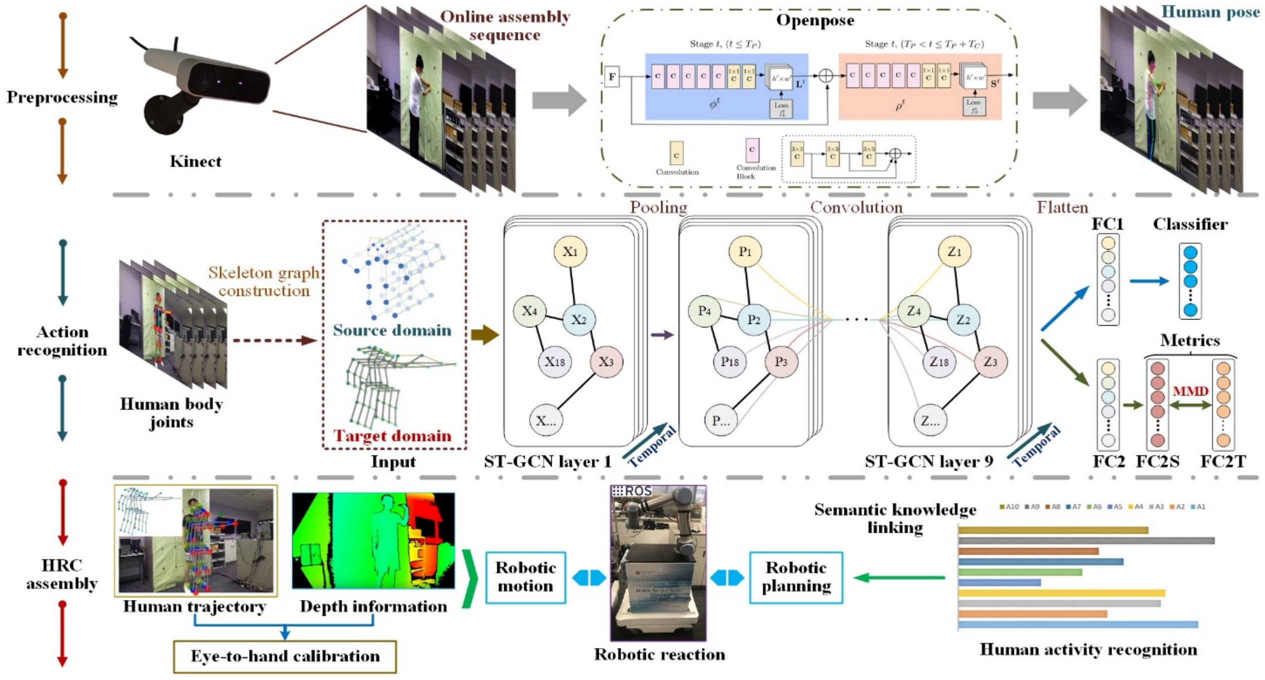

Fig. 1. Pipeline of HRC assembly based on human action recognition

- Action recognition module

The feature extractor of action recognition module consists of nine layers of ST-GCN, which achieved by spatial-temporal graph convolution and pooling operations. Subsequent action classifier is completed by concatenating a fully connected layer (FC1) and one output layer.

Human body joints are firstly connected to construct one spatial-temporal graph $G=(V, E)$ as input. The node set $V=\{v_{ti}|t = 1, ..., T, i = 1, ..., N\}$ denotes there are T frames of a video and N human joints for each frame. Besides, the edge set E contains two subsets. The first one is intra-skeleton connection $E_S=\{v_{ti}v_{tj}|(i, j) \in H\}$ at each frame, where H denotes the natural connection set of human body joints. Another subset is utilized to describe the connection of the same joint across consecutive frames, i.e., inter-frame edges $E_F=\{v_{ti}v_{(t+1)i}\}$. Then, convolution layers of ST-GCN are connected to the input graph to extract latent features of human actions. With an adjacency matrix A and an identity matrix I denoting nodes' connections, the spatial-temporal graph convolution is calculated as follows,

$$f_{out}\left(v_{ti}\right) = \Lambda^{-\frac{1}{2}}\left(A+I\right)\Lambda^{-\frac{1}{2}} f_{in}\left(P\left(v_{ti}\right)\right)\square w\left(l_{st}\left(v_{ti}\right)\right) \tag{1}$$

, where $\Lambda^{ii} = \Sigma_j A^{ij} + I^{ij}$. The operation is composed of sampling function P, feature map f_{in}, partition strategies l_{st} and weight function w. The sampling function P is denoted as,

$$P\left(v_{ti}\right) = \left\{v_{qj} \mid d\left(v_{tj}, v_{ti}\right) \leq D, \left|q-t\right| \leq \left\lfloor \Gamma/2 \right\rfloor\right\} \tag{2}$$

For spatial graph at each frame, $d(v_{tj},v_{ti})$ depicts the minimum length from the node v_{tj} to v_{ti}. While for temporal graph among consecutive frames, Γ controls the temporal range to be included in the neighbor graph. The D and $\Gamma/2$ are set to 1 in our paper. At one frame t, input feature vectors are mapped to dimension c via feature map f_{in}: $V_t \rightarrow R^c$.

Partition strategy l is developed to divide the neighbor set $B(v_{ti})$ of a node v_{ti} into K subsets, i.e., l_{ti}:$B(V_{ti}) \rightarrow \{0,\ldots,K\text{-}1\}$. K is set to 3. For spatial graph, a neighbor node v_{tj} of v_{ti} is partitioned into different subsets as follows,

$$l_{ti}\left(v_{tj}\right) = \begin{cases} 0 & if\ r_j = r_i \\ 1 & if\ r_j < r_i \\ 2 & if\ r_j > r_i \end{cases} \tag{3}$$

, where r_j and r_i is the average distance from the gravity center of body joints to v_{tj} and v_{ti}, respectively. Similarly, the partition strategy l can be extended to spatial temporal graph,

$$l_{st}(v_{qj}) = l_{tj}(v_{tj}) + \left(q - t + \lfloor \Gamma / 2 \rfloor\right) \times K \tag{4}$$

Weight function w is introduced to generate a weight vector in c dimensions. After nine layers of ST-GCN and a layer of FC1, the SoftMax regression is utilized to estimate human action categories in the output layer.

- Domain adaptation module

A fully connected layer FC2 is connected to the last ST-GCN layer to output high-level features of the source domain FC2S and the target domain FC2T, respectively. With n_s data samples from the source domain and n_t examples from the source domain, the distance between the probability distributions of FC2S and FC2T is calculated by MMD,

$$\begin{aligned} D &= \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi\left(X_i^{FC2S}\right) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi\left(X_j^{FC2T}\right) \right\|_H^2 \\ &= \frac{1}{n_s{}^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} \left\langle \phi\left(X_i^{FC2S}\right), \phi\left(X_j^{FC2S}\right) \right\rangle_H + \frac{1}{n_t{}^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \left\langle \phi\left(X_i^{FC2T}\right), \phi\left(X_j^{FC2T}\right) \right\rangle_H \\ &\quad - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \left\langle \phi\left(X_i^{FC2S}\right), \phi\left(X_j^{FC2T}\right) \right\rangle_H \end{aligned} \tag{5}$$

, where H denotes the reproducing kernel Hilbert space (RKHS). If the Gaussian kernel is introduced to calculate the function $\langle f(x), f(y) \rangle$ in RKHS, the unbiased estimation of D_H is defined as,

$$\begin{aligned} \hat{D}\left[X^s, X^t\right] &= \frac{1}{N(N-1)} \sum_{i \neq j}^{N} \exp\left(-\left\|X_i^{FC2S} - X_j^{FC2S}\right\|^2 \big/ 2\sigma^2\right) \\ &\quad + \frac{1}{N(N-1)} \sum_{i \neq j}^{N} \exp\left(-\left\|X_i^{FC2T} - X_j^{FC2T}\right\|^2 \big/ 2\sigma^2\right) \\ &\quad - \frac{2}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left(-\left\|X_i^{FC2S} - X_j^{FC2T}\right\|^2 \big/ 2\sigma^2\right) \end{aligned} \tag{6}$$

- Optimization objective and updated rule

There are two optimization objectives in the proposed transfer learning-based ST-GCN. The former one is to minimize the classification error L_c of action classifier and another one is to minimize the MMD distance between source and target domains. The loss function L is defined as,

$$L = \min_{\theta_f, \theta_c} L_c\left(\theta_f, \theta_c\right) + \lambda \hat{D}\left(\theta_f\right) \tag{7}$$

Where θ_f and θ_c are parameters of the feature extractor and the classifier, respectively. λ balances the influence of two objectives and is set to 1. With learning rate α , parameters θ_f and θ_c of can be updated in the following rules,

$$\begin{aligned} \theta_f &\leftarrow \theta_f - \alpha\left(\frac{\partial L_c}{\partial \theta_f} + \lambda \frac{\partial \hat{D}}{\partial \theta_f}\right) \\ \theta_c &\leftarrow \theta_c - \alpha \frac{\partial L_c}{\partial \theta_c} \end{aligned} \tag{8}$$

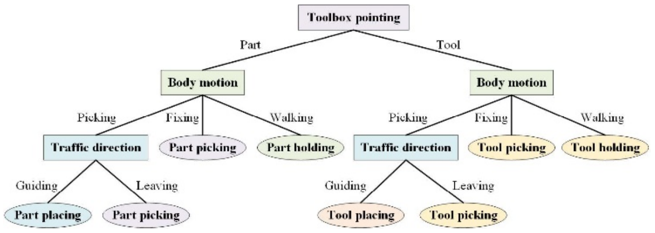

Fig. 2. Decision tree for semantic knowledge linking

Table 1. Samples of human actions in HRC assembly.

Body motion	Gesture action	Traffic direction	Toolbox pointing
Part picking/fixing	Screwing	Robotic guiding	Part selection
Walking	Taping	Robotic leaving	Tool selection

Table 2. Samples of robotic actions in HRC assembly.

Vision detection		Robotic reaction
Scenario perception	Bracket assembly	Obstacle avoidance
	Wire-harness assembly	Vision inspection
Human holding	Part subject	Toolbox picking/holding/placing
	Tool subject	Motion following/pausing/leaving

Table 3. An example of industrial HAR.

Num.	Traffic direction	Body motion	Gesture action	Toolbox pointing	Robotic reaction
1	Guiding	Picking	Taping	Part	Part placing
2	Guiding	Picking	Taping	Tool	Tool placing
3	Guiding	Fixing	Screwing	Part	Part picking
4	Guiding	Fixing	Screwing	Tool	Tool picking
5	Leaving	Fixing	Screwing	Part	Part picking
6	Leaving	Picking	Taping	Part	Part picking
7	Leaving	Picking	Taping	Tool	Tool picking
8	Leaving	Fixing	Screwing	Tool	Tool picking
9	Guiding	Walking	Taping	Part	Part holding
10	Guiding	Walking	Taping	Tool	Tool holding

3.2. *Task-oriented adaptive HRC assembly*

As shown in the bottom part in Fig. 1, predicted human activities provide decision-making for robotic planning in HRC assembly via semantic maps linking. Based on eye-to-hand calibration between human trajectory and depth information, robots can obtain real physical world coordinates, which enable robot to move to a precise location. Therefore, robots can dynamically assist operators and adaptively changing their actions according to human's subtasks, i.e., task-oriented HRC assembly. Some samples of human activities and robotic actions in the human-centered assembly environment are listed in Table 1 and Table 2.

To illustrate the semantic knowledge linking process, a robotic decision-making representation is denoted in Table 3. It is noted that placing means that robots need to deliver the toolbox to human in close distance, while picking means robots leaving the operator and pick a toolbox from storage areas. The robot will hold a toolbox and follow human in close proximity under the 'holding' instruction. Iterative Dichotomiser 3 (ID3) can create a multiway tree based on the largest information gain from decision categories, i.e., classes of robotic reaction in our samples. The information obtained from decision-making process is denoted as entropy,

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \qquad (9)$$

Where, $|y|$ is the number of decision categories and equal to six in our case. P_k is the proportion of the k-the decision. The generated decision tree can link human actions to robotic planning in a generalization manner, as shown in Fig. 2. In this way, flexible and efficient HRC assembly can be achieved due to adaptive robotic reaction.

4. Case study and experiment results

In this section, two comparison experiments are conducted to evaluate the performance of our proposed model for human action recognition. Based on prediction results, efficient HRC is implemented on a bracket assembly task in aircraft cabins.

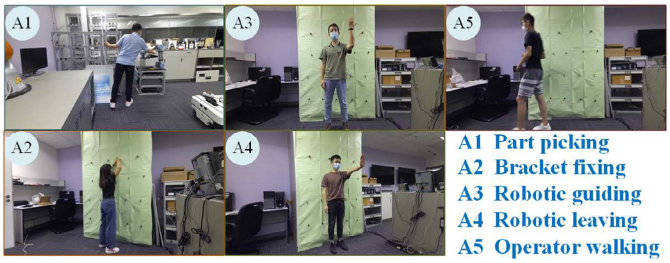

Fig. 3. Samples of human assembly action

4.1. Transfer leaning-enabled action recognition

Dataset. To enable successful assembly action recognition with small dataset, the proposed deep transfer learning network expects to learn knowledge from general daily action representation to manufacturing activities. Hence, one dataset obtained from assembly tasks and one open dataset are used to conduct transfer learning-based action classification.

1) Kinetics. Deepmind Kinetics dataset contains 300,000 video samples which covers up to 400 human action classes [4]. In our experiment, 56 action categories are selected as the source domain data. In preprocessing, each video lasts 10 seconds with 300 frames. Openpose toolbox is used to estimated 18 body joints for each frame (340×256 resolutions). Each joint is recorded with a tuple of (X, Y, C), which means the 2D coordinates and a confidence score.

2) Assembly action. Assembly action dataset (AAD) consists of 222 video clips captured by Azure Kinect. This dataset is developed to simulate human activities of bracket assembly in aircraft cabins and contains five different human actions, including part picking (A1), bracket fixing (A2), robotic guiding (A3), robotic leaving (A4) and operator walking (A5). These videos are captured from three different views and with five volunteers as subjects. Except that each frame of a video has 640×576 resolutions, AAD undergoes the same preprocessing as Kinetics dataset.

Experiment setting. The target domain of AAD is divided into a training dataset and one testing dataset, of which there are 170 video samples and 50 video clips, respectively. All 2D coordinates of human body joints are normalized via dividing by images' resolutions before feeding into our proposed deep transfer learning network. The transfer learning model is optimized using stochastic gradient descent with an initial learning rate of 0.1, which is then multiplied by 0.1 after every 10 epochs. Besides, only 8% of target data are given labels during the training process, while unlabeled data are aligned to extracted domain-invariant features via MMD.

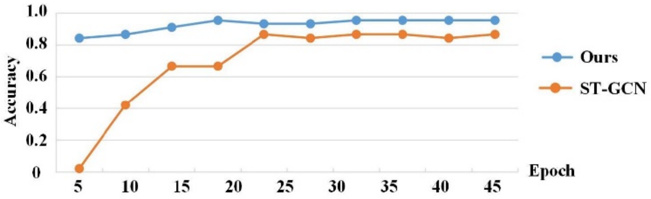

Fig. 4. Accuracy during training process

Comparison results. The human action classification performance of our proposed model can be evaluated by a comparison experiment of original ST-GCN. As shown in Table 4, our proposed network presents an obvious improvement on the mean average precision (mAP). Therefore, the deep transfer learning network has better capabilities to distill latent action expression and classify human assembly activities. Some examples of assembly action are illustrated in Fig. 3. Moreover, our model can achieve a competitive performance much faster during training process (see Fig. 4). In this context, the general daily action representation provides a valuable initialized direction for model training, which in turn accelerates knowledge transferring and learning for assembly action patterns.

Table 4. Accuracy of assembly action classification.

Accuracy	A1	A2	A3	A4	A5	mAP
ST-GCN	100.00	77.78	100.00	88.89	66.67	86.67
Ours	100.00	100.00	100.00	88.89	88.89	95.56

4.2. Efficient HRC for aircraft bracket assembly

Aircraft bracket assembly always suffer from low efficiency due to the narrow and small workspaces in aircraft cabins, where it is not realistic for a worker to hold all tools and parts during the bracket installing operation, and he has to constantly pass between the assembly area and tool storage areas for toolbox change. There is an urgent requirement for mobile robots to collaboratively conduct pick-and-place work and to collaborate with operators as a smart assistant.

With our proposed action recognition model, operators' activity perception and location information can be obtained from the Azure Kinect. As shown in Fig. 5 (a), a mobile robot which consists of Universal Robots UR5 and a mobile base is utilized to implement following robotic plans, 1) picking toolboxes from storage areas (RP1), 2) moving towards an operator with toolboxes (RP2), 3) following operators' motion to enable him to pick tools or parts in a shared workspace (RP3). In this HRC assembly scenario, Azure Kinect is utilized to capture living video streams of a human operator, who installs brackets to the aircraft cabin. The human assembly action can be predicted by the transfer learning-based model. Similarly, based on the semantic knowledge linking process, suitable decision-making can be generated for robotic reaction, as shown in Fig. 5 (b). After eye-to-hand calibration for the world coordinate, the mobile robot can execute the generated path planning and motion to assist the human operator. Therefore, human and robots can efficiently complete bracket assembly tasks with high-level collaboration, rather than simple co-existence.

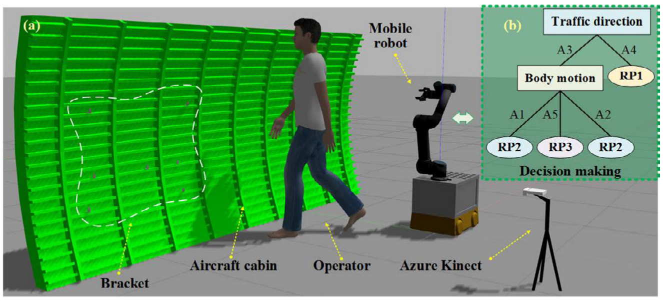

Fig. 5. A simulation environment of HRC for aircraft bracket assembly

5. Conclusion

This paper introduced a novel computer vision based HRC assembly approach based on human action recognition. A deep transfer learning ST-GCN model was proposed to learn domain-invariant action representations between source and target human body joints. Extracted features between source and target domains can be aligned in the domain adaptation module by the application of MMD. Our proposed approach shows great advantages for human activity recognition in manufacturing scenarios without collecting a huge amount of labeled data. Robotic reactions can be generated via semantic knowledge mapping from identified human activities, of which the decision-making mechanism can greatly improve HRC assembly efficiency. Lastly, a case study of the bracket assembly in aircraft cabins was carried out to evaluate its feasibility with better overall performance.

Apart from the abovementioned advantages, potential future research directions are also highlighted here, including: 1) conducting online human action recognition based on 3D images, and 2) developing dynamic robotic control in the adaptive decision-making mechanism. It is hoped this work can provide insightful knowledge to today's industrial HRC research and implementations.

Acknowledgements

This research work was partially supported by the grants from the Innovation and Technology Commission, HKSAR (AiDLab-RP2-1) and Research Committee of the Hong Kong Polytechnic University (U-GAHH).

References

[1] Xiong Q, Zhang J, Wang P, Liu D, Gao RX. Transferable two-stream convolutional neural network for human action recognition. J Manuf Syst 2020;56:605–14.

[2] Wang L, Liu S, Liu H, Wang XV. Overview of Human-Robot Collaboration in Manufacturing. Springer International Publishing; 020. https://doi.org/10.1007/978-3-030-46212-3.

[3] Wang XV, Kemény Z, Váncza J, Wang L. Human–robot collaborative assembly in cyber-physical production: Classification framework and implementation. CIRP Ann - Manuf Technol 2017;66:5–8. https://doi.org/10.1016/j.cirp.2017.04.101.

[4] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. 32nd AAAI Conf Artif Intell 2018:7444–52.

[5] Ullah A, Muhammad K, Del Ser J, Baik SW, de Albuquerque VHC. Activity recognition using temporal optical flow convolutional features and multilayer LSTM. IEEE Trans Ind Electron 2019;66:9692–702.

[6] Liu S, Wang L, Wang XV. Symbiotic human-robot collaboration: multimodal control using function blocks. Procedia CIRP 2020;93:1188–93. https://doi.org/10.1016/j.procir.2020.03.022.

[7] Bi ZM, Luo M, Miao Z, Zhang B, Zhang WJ, Wang L. Safety Assurance Mechanisms of Collaborative Robotic Systems in Manufacturing. Robot Comput Integr Manuf 2021;67:1–10. https://doi.org/10.1016/j.rcim.2020.102022.

[8] Mazhar O, Navarro B, Ramdani S, Passama R, Cherubini A. A real-time human-robot interaction framework with robust background invariant hand gesture detection. Robot Comput Integr Manuf 2019;60:34–48. https://doi.org/10.1016/j.rcim.2019.05.008.

[9] Hietanen A, Pieters R, Lanz M, Latokartano J, Kämäräinen JK. AR-based interaction for human-robot collaborative manufacturing. Robot Comput Integr Manuf 2020;63:101891. https://doi.org/10.1016/j.rcim.2019.101891.

[10] Cheng Y, Sun L, Liu C, Tomizuka M. Towards Efficient Human-Robot Collaboration with Robust Plan Recognition and Trajectory Prediction. IEEE Robot Autom Lett 2020;5:2602–9.

[11] Sun, S., Kuang, Z., Sheng, L., Ouyang, W., Zhang, W., 2018. Optical flow guided feature: A fast and robust motion representation for video action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1390–1399.

[12] Zhu T, Zhou Y, Xia Z, Dong J, Zhao Q. Progressive Filtering Approach for Early Human Action Recognition. Int J Control Autom Syst 2018;16:2393–404.

[13] Dang Y, Yang F, Yin J. DWnet: Deep-wide network for 3D action recognition. Rob Auton Syst 2020;126:103441.

[14] Li S, Zheng P, Zheng L. An AR-Assisted Deep Learning Based Approach for Automatic Inspection of Aviation Connectors. IEEE Trans Ind Informatics 2021; 17(3): 1721 - 1731.

[15] Guo L, Lei Y, Xing S, Yan T, Li N. Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines with Unlabeled Data. IEEE Trans Ind Electron 2019;66:7316–25. https://doi.org/10.1109/TIE.2018.2877090.

[16] Huang S, Guo Y, Liu D, Zha S, Fang W. A Two-Stage Transfer Learning-Based Deep Learning Approach for Production Progress Prediction in IoT-Enabled Manufacturing. IEEE Internet Things J 2019;6:10627–38. https://doi.org/10.1109/JIOT.2019.2940131.

[17] Zhou J, Zheng LY, Wang Y, Gogu C. A Multistage Deep Transfer Learning Method for Machinery Fault Diagnostics across Diverse Working Conditions and Devices. IEEE Access 2020;8:80879–98.

[18] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. Proc - 30th IEEE Conf Comput Vis Pattern Recognition;2017-Janua:1302–10.