Human Factors in Robots, Drones and Unmanned Systems, Vol. 138, 2024, 66–74
https://doi.org/10.54941/ahfe1005009

A Method for Human-Robot Collaborative Assembly Action Recognition Based on Skeleton Data and Transfer Learning

Shangsi Wu, Haonan Fang, Peng Wang, Xiaonan Yang, Yaoguang Hu, Jingfei Wang, and Chengshun Li

Industrial and Systems Engineering Laboratory, Beijing Institute of Technology, Beijing, China

ABSTRACT

Human-robot collaborative assembly (HRCA) has become a vital technology in the current context of intelligent manufacturing. To ensure the efficiency and safety of the HRCA process, robots must rapidly and accurately recognize human assembly actions. However, due to the complexity and variability of the human state, it is challenging to accurately recognize such actions. Furthermore, with the lack of a large-scale assembly action dataset, the model only constructed from the data obtained in a single assembly scenario demonstrates limited robustness when applied to other situations. To achieve rapid and cost-effective action recognition, this paper proposes a method for human action recognition based on skeleton data and transfer learning. First, we screen the action samples which are similar to assembly actions from the NTU-RGB+D dataset to build the source dataset and reduce the dimension of its skeleton data. Afterwards, the Long Short-Term Memory (LSTM) network is used for learning universal features from the source dataset. Second, we use Microsoft Kinect to collect skeleton data of human assembly actions as the initial target dataset and use the sliding time window method to expand its size. After aligning the data of two datasets, the gradient freezing strategy is adopted during the transfer learning process to transfer the features learned from the source dataset into the recognition of HRCA actions. Third, the transfer model is validated through a small-scale reducer assembly task. The experimental results demonstrate that the method proposed can achieve assembly actions recognition rapidly and cost-effectively while ensuring a certain level of accuracy.

Keywords: Action recognition, Human-robot collaboration, Transfer learning, Skeleton data

INTRODUCTION AND BACKGROUND

In today's smart factories, human-robot collaboration is increasingly applied in product assembly processes due to its significant role in enhancing production efficiency and assembly line flexibility. In industrial scenarios, typical human-robot interaction methods include interactive interface (Dos Santos et al., 2020), gesture recognition (Coupeté et al., 2015), speech recognition (Bingol and Aydogmus, 2020) and etc. However, these interaction modes require additional human actions, which are inflexible and hard to adjust. This deficiency prevents operators to fully and effectively engage in the

assembly tasks at hand. Considering the issues above, people began to allow robots to proactively recognize and respond to the natural human actions during assembly, known as human action recognition (HAR). A significant challenge in this technology is enabling robots to accurately and quickly recognize human actions.

In recent years, the rapidly developing deep learning technology has been increasingly employed in HAR processes for its superior feature extraction abilities. The action dataset used for feature extraction significantly influence the final result of HAR. Aiming at diagnosing Autism Spectrum Disorder (ASD) children, Zhang et al. (Zhang et al., 2021) developed an ASD dataset and classified stereotyped actions of ASD children in their daily life. To design a service robot's action recognition system, Wang et al. (Wang et al., 2023) proposed a multi-modal visual dataset named THU-HRIA dataset including eight prevalent human actions in a restaurant environment. Nevertheless, in the domain of HRCA, there isn't an extensive dataset that comprehensively considers diverse environment factors and camera settings, and includes various common assembly actions. Only using the data obtained in a single assembly scenario to train models for action recognition poses issues as follow: If the size of the dataset needed is too large, it entails significant time and labor costs. Moreover, it is impractical to collect sufficient annotated data from complex industrial settings (Li et al., 2021). If the size is too small, it leads to poor robustness and is prone to over-fitting issues. The recognition accuracy is profoundly impacted by the complexity and variability of human states and alterations in camera settings and environmental factors.

Transfer learning is an effective approach to address the issue of sample scarcity in target dataset. The transfer recognition model learns universal features from daily actions similar to assembly actions in the source domain and then further learns more accurate assembly action features from samples in the target domain. Current researches using transfer learning for action recognition show high levels of the efficiency and accuracy, demonstrating the feasibility of constructing the HAR model through transfer learning during the HRCA process (Wei et al., 2023, Xiong et al., 2020).

Therefore, this paper proposes an action recognition method for HRCA based on skeleton data and transfer learning. Skeleton data has a lower dimension, leading to faster model training and recognition speed. Moreover, compared to RGB, which is also commonly used as HAR input, skeleton data is less affected by environment factors and variations in human body size. Finally, we present a validation of the method through a small-scale reducer assembly task. The results of the task show that the HAR model trained by this method can rapidly and cost-effectively recognize the assembly actions.

METHOD

As the framework of the method shown in Figure 1, the whole method can be divided into three sections: A) Classification of assembly actions. In this section, we classify the assembly actions into three categories according to the assembly scenario: selecting parts, assembling parts, and checking; B) Pre-training process. In this section, we collect the source dataset from the NTU-RGB+D dataset, reducing the dimension and extracting the frame of the data. Then the Long Short-Term Memory (LSTM) network is used to pre-train the

HAR model. C) Transfer training process. In this section, we collect the target dataset through an experiment, reducing the dimension and expanding the size of the data by the sliding time window. After that, the transfer learning model is trained by using the gradient freezing strategy.

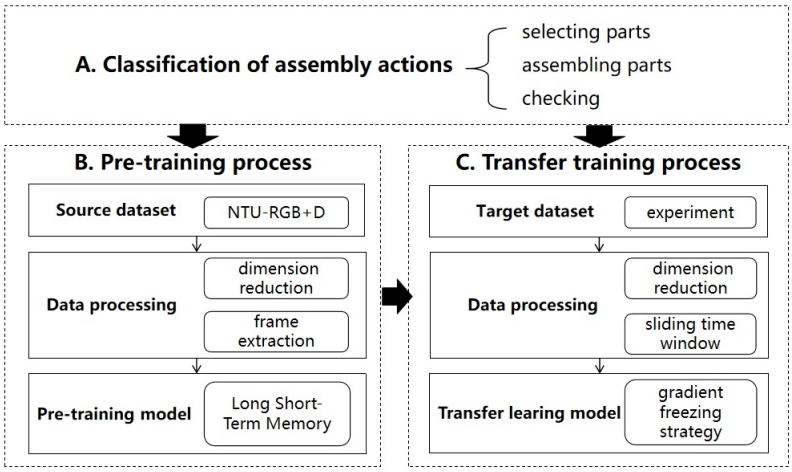

Figure 1: Method framework.

Classification of Assembly Actions

In this paper, we design a small-scale reducer assembly task as shown in Figure 2. As the operator sits at a fixed workbench to assemble the parts, it is only necessary to focus on the upper body of the operator. The classified actions should cover all assembly actions and have significant differences of characteristics between them. Under the premise of excluding actions irrelevant to the assembly process, the human actions during the small-scale reducer assembly task are classified into the following three categories:

A) Selecting parts: The operator picks up parts from the parts area using one or both hands. This action's characteristics are obvious movements of the operator's hands and arms and minimal movement of the head and torso.
B) Assembling parts: The operator installs the parts onto components in the assembly area. This action's characteristics are obvious movements of the hands, limited movements of the arms, and almost no movement of the head and torso.
C) Checking: The operator bends over to check whether the parts are assembled correctly. This action's characteristics are obvious movements of the head and torso and almost no movement of the arms and hands.

Pre-Training Process

Obtaining Source Dataset

We choose the NTU-RGB+D dataset as the source dataset. The NTU-RGB+D (Liu et al., 2020) dataset comprises 114,480 video samples involving

106 human subjects. For each video sample, the dataset stores the skeleton data of human actions. Most importantly, the ages of human subjects in this dataset are between 10 to 57 and the heights are between 1.3m to 1.9m, providing diverse human states and different human body size. Additionally, the Microsoft Kinect used to collect the data is set at different angles, heights, and distances, which is also beneficial for enhancing the generalization ability of the HAR model. Based on the action classification above, we screen daily actions similar to each of the three categories of assembly actions from the NTU-RGB+D dataset, according to their respective action characteristics.

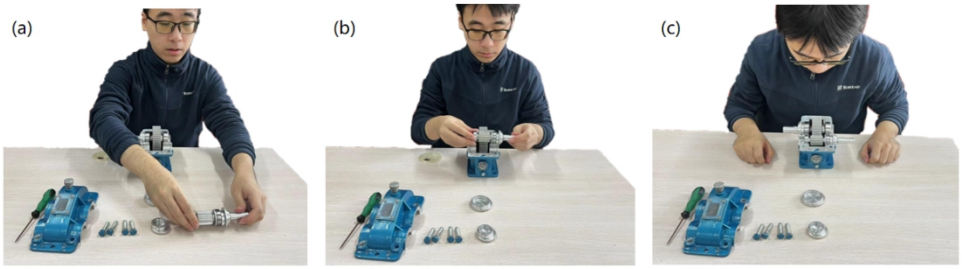

Figure 2: The classification of assembly actions. (a) Selecting parts; (b) assembling parts; (c) checking.

Data Processing

In the skeleton data obtained from the NTU-RGB+D dataset, each frame contains the data for 12 items of 25 human skeletal joints. The spatial coordinate-related joint data is valued in meters. Given that the task focuses only on the upper body assembly actions, the data for the 8 skeletal joints of the lower body can be omitted. Only the 17 skeletal joints located on the head, shoulder, elbow, arm, hand, and torso are retained.

Each row of data corresponds to the relevant data for a skeletal joint in that frame. Only 7 items of data describing the spatial position of each skeleton joint are retained and unnecessary skeletal joint data are removed as shown in Figure 3. Finally, the skeleton data for each frame is reduced from a total of 300 data (25 joints × 12 items) to 119 data (17 joints × 7 items). This process reduces the dimension of each frame's data to approximately 60%.

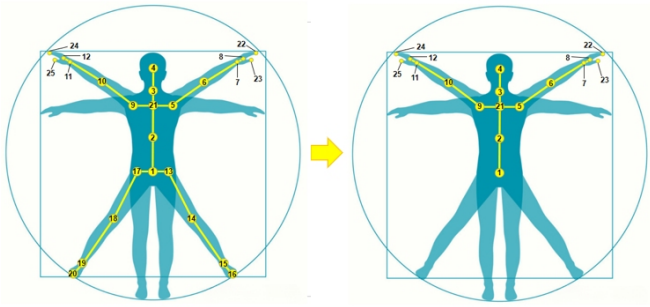

Figure 3: Retaining necessary skeletal joints.

In the NTU-RGB+D dataset, the action frame rate is 30 fps, while the camera used to capture assembly action skeletal data outputs at approximately 7 to 8 fps. Therefore, for transfer learning process, we extract the frame of the data from the source dataset. This step also quadruples the number of samples in the source dataset.

Additionally, to accelerate the training speed of the model, the precision of the skeletal data is rounded to four decimal places, which also means that the data precision is 0.1 mm. Each action sample, stored as a sequence of 24 frames in 3 seconds, is annotated according to the three categories of actions. The total number of samples in the final constructed source dataset is 5,224. We set the ratio of the training samples to the validation samples to 4:1, so that the number of the training samples is 4,179, and the number of the validation samples is 1,045.

Pre-Training Model

Due to the temporal correlation of the skeleton data, we use the LSTM network to extract features of the skeleton data. We choose Adam as the network's optimizer and Negative Log Likelihood Loss as the loss function.

We evaluate the number of hidden states and the number of training epochs by comparing the validation accuracy and loss values of different models, which are shown in Figure 4. Hidden state is an important parameter related to the network structure. Too many hidden states will cause over-fitting of neural networks, while too few will reduce the adaptive ability. Training epoch straightly influences the training speed and quality. Too many training epochs will increase the training duration linearly, while too few will obtain a bad convergence effect.

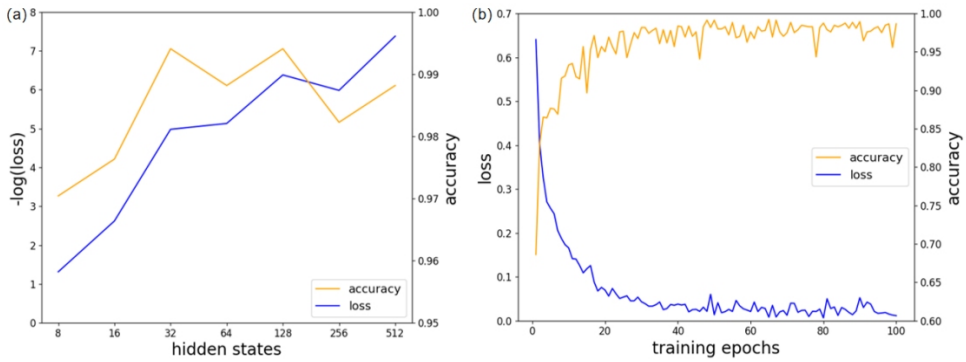

Figure 4: Values of loss and accuracy in different (a) hidden states; (b) training epochs.

According to Figure 4(a), regardless of the hidden state we choose, the final accuracy values of the validation consistently remain at a high level. For a easy observation, values of the final loss are taken the negative logarithm. As the number of hidden states increases, values of the minimum loss decrease continuously. Considering both over-fitting problem and adaptive ability, we set the number of hidden states to 64.

According to Figure 4(b), in the early stages of training, there are significant changes in values of accuracy and loss. As training around 20 to 30 epochs, values of accuracy and loss exhibit converging trends. From approximately the 40th epoch onward, the values of accuracy and loss exhibit oscillation within a certain range. To ensure training efficiency and HAR performance, we set the number of training epochs to 50.

The rest of the important neutral network parameters are chosen as follows: the number of hidden layers is 2 and the initial learning rate is 0.001. The pre-trained model achieves an accuracy of 98.4% on the validation samples in the source dataset.

Transfer Training Process

Obtaining Target Dataset

We design an experiment to obtain skeleton data of assembly actions and used a Microsoft Kinect sensor to record these actions. Two experimenters participate in this experiment. Each person records three categories of assembly actions and each category is recorded for 5 minutes.

Data Processing

To transfer the action features from the source dataset to the new model, it is essential to ensure that the data formats of the source and target datasets are consistent. First, we use the same method to reduce the dimension of the skeleton data collected by removing unnecessary skeletal joints and other parameters. The processed data for each frame also includes 119 data, which is consistent with the source dataset.

To address the issue of insufficient sample quantity in the target dataset collected from the experiment, we use the sliding time window strategy. The window length is set to 3 seconds, consistent with the sample length of the source dataset, and the sliding step is set to 2 seconds. This approach increased the number of the target dataset samples by 1.5 times.

In the same way, the precision of the skeletal data is rounded to four decimal places and each action sample is annotated according to the three categories of actions. The number of samples processed in the two datasets is shown in the Table 1. The final ratio of samples between the source dataset and the target dataset is approximately 6:1.The ratio of the training samples to the validation samples is set to 4:1.

Table 1. The numbers of samples.

	Total Samples	Training Samples	Validation Samples
Source dataset	5224	4179	1045
Target dataset	842	674	168

Transfer Learning Model

To transfer the features learned from the source dataset into the model, we use a gradient freezing strategy. Under the premise of setting two hidden layers,

we choose to freeze only the weights of the first layer. The weights of the second layer and the fully connected layer are retrained. All neutral network parameters are the same as the pre-training's, except for the initial learning rate which is fine-tuned to 0.0001. The transfer-trained model achieves an accuracy of 96.3% on the validation samples in the target dataset.

CASE STUDY

To validate whether the model trained through the method above can accurately identify the operator's assembly actions, we construct a working platform as shown in Figure 5. Parts and tools are placed on the left side of the platform, while the collaborative robotic arm is positioned on the right side. The operator sits in front of the working platform and completes the small-scale reducer assembly task. Kinect is placed above the platform to recognize the operator's actions, and the recognition results are transmitted to the robotic arm as input.

Figure 5: Action samples of the operator and robotic arm. (a) Selecting parts; (b) assembling parts; (c) checking.

We define the collaborative actions performed by the robotic arm based on the recognition results for different human actions:

A) Selecting parts: The robotic arm picks up the part which this assembly step needs and give them to the operator.
B) Assembling parts: The robotic arm picks up the tool which this assembly step needs and give it to the operator or just returns to the beginning position to avoid obstruction.
C) Checking: The robotic arm arrives at the beginning position.

The experiment results indicate that with the assistance of the robotic arm, the operator can successfully complete the assembly task. In all three action categories of selecting parts, assembling parts, and checking, the robotic arm can accurately and swiftly perform the pre-designed collaborative actions.

Furthermore, we test another recognition model which is trained only using samples from the target dataset. Although this model still achieves a recognition accuracy of over 95% on the validation samples of the target dataset, it exhibits high recognition error rate during the assembly task. Therefore, the comparison between the two recognition results indicates that

using the method proposed in this paper can enhance the robustness of the HAR model.

CONCLUSION

This paper proposes an HRCA action recognition method based on skeleton data and transfer learning. By categorizing assembly actions and creating source and target datasets, an HAR model is constructed using a LSTM network. Finally, the recognition speed and accuracy of the recognition model are validated on a small-scale reducer assembly task. The experiment demonstrates that this method can not only achieve assembly actions recognition rapidly and cost-effectively, but effectively enhance the robustness of the recognition model in the absence of a large-scale HRCA action dataset.

However, the assembly action classification mentioned in this method is idealized, as it does not considerate transitional actions which are unrelated to the assembly actions during the assembly process (Carrara et al., 2019). Our model lacks the ability to handle such spontaneous actions. Furthermore, skeleton data exhibits spatial-temporal correlation, but the LSTM network only utilizes the temporal aspect. In future work, we plan to use graph convolution network, which can effectively extract both temporal and spatial features of skeleton data, to improve our method.

To increase the efficiency of human-robot collaboration, many researchers are exploring how to allow robots to proactively predict human intentions and avoid collision. These abilities ensure that humans and robots safely and efficiently accomplish collaborative tasks even when coexisting in narrow spaces (Lyu et al., 2023). This will also be an important direction for our future work.

ACKNOWLEDGMENT

The authors would like to thank the National Natural Science Foundation (52175451 and 52205513).

REFERENCES

Bingol, M. C. & Aydogmus, O. 2020. Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot. Engineering Applications of Artificial Intelligence, 95, 103903.

Carrara, F., Elias, P., Sedmidubsky, J. & Zezula, P. 2019. LSTM-based real-time action detection and prediction in human motion streams. Multimedia Tools and Applications, 78, 27309–27331.

Coupeté, E., Moutarde, F. & Manitsaris, S. 2015. Gesture Recognition Using a Depth Camera for Human Robot Collaboration on Assembly Line. Procedia Manufacturing, 3, 518–525.

Dos Santos, C. W., Filho, N. L. D., Espíndola, D. B. & Botelho, S. S. C. 2020. Situational Awareness Oriented Interfaces on Human-Robot Interaction for Industrial Welding Processes. IFAC-PapersOnLine, 53, 10168–10173.

Li, S., Fan, J., Zheng, P. & Wang, L. 2021. Transfer Learning-enabled Action Recognition for Human-robot Collaborative Assembly. Procedia CIRP, 104, 1795–1800.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y. & Kot, A. C. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42, 2684–2701.

Lyu, J., Ruppel, P., Hendrich, N., Li, S., Görner, M. & Zhang, J. 2023. Efficient and Collision-Free Human–Robot Collaboration Based on Intention and Trajectory Prediction. IEEE Transactions on Cognitive and Developmental Systems, 15, 1853–1863.

Wang, J., Zhang, T., Wu, X. & Zeng, L. 2023. A Dataset and System for Service Robot Action Interaction Based on Skeleton Action Recognition. 2023 8th International Conference on Signal and Image Processing (ICSIP).

Wei, C., Wang, C., Bai, S., Li, Y., Tian, X. & Zhou, L. 2023. Transfer Learning Based Multi-Perception Safety Strategy for Human-Robot Collaboration. 2023 IEEE International Conference on Real-time Computing and Robotics (RCAR).

Xiong, Q., Zhang, J., Wang, P., Liu, D. & Gao, R. X. 2020. Transferable two-stream convolutional neural network for human action recognition. Journal of Manufacturing Systems, 56, 605–614.

Zhang, Y., Tian, Y., Wu, P. & Chen, D. 2021. Application of Skeleton Data and Long Short-Term Memory in Action Recognition of Children with Autism Spectrum Disorder. Sensors (Basel), 21.