Gesture Recognition in Augmented Reality Assisted Assembly Training

Jiaqi Dong¹, Zisheng Tang², Qunfei Zhao¹,*

¹ School of Electronic Information and Electric Engineering, Shanghai JiaoTong University, Shanghai, China
² Shanghai Ninth People’s Hospital, Shanghai JiaoTong University School of Medicine, Shanghai, China

*Corresponding author e-mail: zhaoqf@sjtu.edu.cn

Abstract. The Augmented Reality assisted assembly training (ARAAT) is an effective and low-cost way in motor and electronic industry. In ARAAT, assembly operations are the processes which mainly use the AR device to recognize the gestures and match the virtual workpieces to the hand based on the consistency of time and space. Operations are recorded to the video by an AR device. By dealing with frames of the input video, operations can be distinguished. In this paper, a gesture recognition algorithm in ARAAT are proposed. The assembly training operations consist of several actions and actions we concerned are all conducted by gestures. Operations can be classified from the sequences of actions, that is, gestures. Every 20 frames of the input video are denoted as an action unit and the action unit slides with a time window. According to the 2D and 3D features of the action unit, a scorer trained by the samples of specific actions gives the optimal label of each frame to recognize the action. To avoid the disturbance of the transition and invalid action to the action boundary during the recognition, the boundary is iteratively optimized by the probability density distribution. The proposed algorithm implemented on HoloLens is compared with other algorithms. The experimental results indicate that the proposed algorithm achieves a high recognition rate and can reduce the computational complexity. The results prove the efficiency of recognition algorithm in ARAAT and ensure a friendly experience for the human-machine interaction.

1. Introduction

The Augmented Reality assisted assembly (ARAA) is an effective and low-cost way in motor and electronic industrial assembly. The one of main ARAA applications is ARAA training (ARAAT). In the ARAAT, trainees need to interact with virtual objects with real hands or equipment. Therefore, it is necessary to correctly match the virtual workpieces to the hand or the equipment based on the consistency of time and space. In the ARAAT, operations are recorded to the video by an AR device. From dealing with each frame of the input video, operations can be distinguished. Hand tracking, hand localizing and gesture recognition are the three essential steps for the ARAAT operation recognition. Because most current AR devices are with the function of hand tracking, more and more researchers focus on the hand localizing and the gesture recognition.

In ARAAT, doing the assembly tasks more realistically is an important topic. ARAAT is helpful and suitable for assembling small devices. Trainee can obtain experiences from doing assembly tasks
in real and virtual environment by AR devices. Moreover, ARAAT can rehearse at any time and any place with a minimum cost. To perform ARAAT more realistically, many methods are proposed. Boonbrahm [1] made assembly tasks more natural by adding physical attributes to the virtual objects. The experiment results proved it is easy for trainees to perform assembly tasks, however it lacks of the feeling of touch. To make a natural interaction with virtual objects by real hands, a natural hand interaction method is proposed by Lee [2] which is based on hand direction calculation and collision detection. However, there are some errors from calculation in one-finger interaction. Therefore, the application is limited. Wang [3] and Li [4] developed the bare hand interaction in the AR assembly environment not only by a single finger. However, the result is with a low recognition accuracy due to the fingertips tracking algorithms without depth information. Choi [5] and Figueiredo [6] proposed an effective and natural hand-based interaction in AR which used the gesture of grasp and release.

The researches mentioned above seem to be suitable for the assembly of small devices by hands, however, the interaction gestures in the given applications are limited only to grasp and release. Therefore, some training operations which need special equipment or special setups are too complicated to be done by such simple interaction gestures. HoloLens [7] is an advanced AR equipment which is widely used in many fields. Limited to the gestures of pointing and blooming, HoloLens cannot play a full role in ARAAT. In order to make ARAAT more efficient for trainees, it is still needed more works done to improve the types of interaction gestures and accuracy of the recognition results. In this paper, a gesture recognition algorithm in Augmented Reality assisted assembly training (ARAAT) operations are proposed. By dealing with each frame of the input video recorded by an AR device, each operation is divided into a sequence of actions and actions we concerned are all conducted by gestures. On the basis of the practical industrial assembly tasks, several common assembly operations and actions are concluded to be recognized. Every 20 frames of the input video are denoted as an action unit and the action unit slides with a time window. According to the 2D and 3D features of the action unit, a scorer trained by the samples of specific actions gives the optimal label of each frame to recognize the gestures of each action. By the recognition of action sequences, the operations can be easily distinguished. To improve the accuracy of recognition, the action boundary is iteratively optimized by the probability density distribution to avoid the disturbance of the transition and invalid action during the recognition.

The rest of the paper is organized as follows. The description and modeling for ARAAT are presented in Section 2. The action and operation recognition is detailed in section 3. Experiments are conducted on a homemade dataset and compared with other algorithms. The experimental results are analyzed in the section 4. Finally, a brief conclusion of this paper and some future works are given in section 5.

2. Description and modeling for ARAAT

There are many assembly tasks in ARAAT. Each task consists of different operations. Each operation is concerned to be partitioned to several actions. Most of actions in ARAAT can be mainly conducted by gestures. To recognize operations, the gestures in each action should be recognized first. Because the operations are recorded to the video by an AR device, operations can be distinguished by dealing with each frame of the input video.

The problem is formally stated as follow.

Given an input video of a task obtained from an AR device denoted by

\[ V = \{ f_t \mid f_t \text{ is the } t\text{th frame of the video, } t=1,2, \ldots \}. \]  

(1)

The input video \( V \) contains a series of operations which denotes as \( O_i \), namely,

\[ V = \{ O_i \mid O_i \text{ is the } i\text{th operation in the video } V, \ i=1,2, \ldots \}. \]  

(2)

The actions denoted by \( A^j_i \) are continuous segments in each operation \( O_i \). The frame of the beginning and end of the action \( A^j_i \) are denoted as \( f^j_{b,i} \) and \( f^j_{e,i} \). Then,

\[ O_i = \{ A^j_i \mid A^j_i \text{ is the } j\text{th action in the operation } O_i, j=1,2, \ldots, N_i \}, \]  

(3)

\[ A^j_i = \{ f^l_{b,i}, \ldots, f^l_{e,i} \mid f^l_{e,i} = f^l_{b,i+1} \}, \]  

(4)
One of the solutions of the operation recognition is to label each action and sequences of $O_i$ with the corresponding $\{A^i\}$ in the given video. The model of ARAAT are illustrated in Fig.1. The difficulty of the operation recognition is to segment and label actions in an operation at the same time.

3. Action recognition

3.1 Category of the action

In the ARAAT, the trainee directly participates in the assembly work by hands or various virtual equipment. The assembly work generally falls into five types as shown in Fig.2: match, conjugate, join, fasten and mesh. According to the corresponding work, the actions can be concluded as shown in Fig.3: insert, equip and screw.

As mentioned above, the actions of the assembly work in ARAAT can be summarized as shown in Fig.4: Point, Move, Grasp, Release, Scale and Rotate. “Point” is to select a virtual object or an option in AR environment. “Scale” is the gesture that the trainee tries to change the size of some objects and “Rotate” is to change the orientation. “Move” can be the shift of a workpiece or an equipment in the hand after a “Grasp” action, or just a movement of a virtual object after a “Point” action. The first is to recognize the action $A \in \{\text{Point, Move, Grasp, Release, Scale, Rotate}\}$ by a hierarchical classifier from the video which only contains one action.
3.2 Action recognition

The action recognition contains two stages: the learning stage and the recognition stage.

The learning stage is a two-layer classifier. In the lower layer, the gesture is divided into three types \{I, TI, AF\} which means the gesture applies index finger only, thumb and index finger, or all fingers. When the frame \(f_t \in V\) inputs, where \(t > 20\), the continuous 20 frames before \(f_t\) in the video and frame \(f_t\) constitute an action unit denoted by \(U\). According to the experimental validation, 20 is best choice for the recognition. The action unit \(U\) is inputted into a feature extractor to gain the gesture type, gesture trajectory, and hand-object effects. The multiple features yielded from the lower layer are sent to the top layer. In the top layer, the score of each action and a probability distribution \(p(A)\) for the action unit are produced, which are shown in Fig.5. The distribution indicates the probability of the action being sorted to each type. At last, \(p(A)\) is transferred to a log density distribution \(\log p(A)\) for the convince of the action segmentation.

The recognition stage utilizes PCNN to search and label the optimal boundary. The recognition stage finds the time step first for the initial boundaries and then optimized iteratively to the final boundaries.

3.3 Action segmentation

According to the action score, the given video is segmented to each action by labelling each frame. In the practical assembly operation, it is inevitably that some transition actions between two meaningful actions will happen, which means nothing. The transition action can affect the classifier to generate an incorrect result and decrease the accuracy of the recognition. To cope with the transition action, the actions are augmented with a “Null” action, that is \(A \in \{\text{Point}, \text{Move}, \text{Grasp}, \text{Release}, \text{Scale}, \text{Rotate}, \text{Null}\}\). Because each assembly action or the transition action will take 20–30 frames, it is possible for the action unit to contain only one action. The action unit is inputted into the action recognizer and each frame is labeled. Denote \(p(A|f_k)\) is the distribution of each action on frame \(f_k\), so

\[
A_m = \arg \max p(A|f_k),
\]

\(m\) is the class of the action \(A\). when a frame \(f_k\) is inputted, if

\[
\arg \max p(A|f_{k-1}) \neq \arg \max p(A|f_k),
\]

the \(f_k\) can be seen as the switch point of the initial boundaries, as shown in Fig.6.

With the initial boundaries \(\{A_j\}\), the segmentation is refined by tuning \(\{f_b^{l,j}, f_e^{l,j}\}\), so that

\[
p = \prod_j p(A_j), s.t = A_j = \{f_b^{l,j}, ..., f_e^{l,j}\}.
\]

Thus, the optimal segmentation can be obtained by

\[
\arg \max \Sigma_j \log p(A_j).
\]

The initial number of actions in the segmentation is denoted by \(N\). \(N\) is reduced as each iteration. (9) can be solved by PCNN and obtain the optimal boundaries \(\{f_b^{l,j}, ..., f_e^{l,j}\}\). Because the time for each action in practical assembly operation generally longer than 10s, the short actions in an operation can be seen as the noise. When the neighboring \(f_b^{l,j}\) and \(f_e^{l,j}\) are shorter than 10 frames, the segments are...
merged and $N$ is reduced, then (9) is solved again. The process is iterated until neighboring $f_{b}^{i,j}$ and $f_{e}^{i,j}$ are far and stable.

![Figure 6. the initial switch points in the distribution of actions](image)

### 3.4 Operation recognition

Because the operation consists of different action in a given order, according to the sequences of the actions, the operation can be easily recognized.

In ARAAT, the trainee needs to participate in the assembly works by hands or various virtual equipment. The common assembly operations mentioned in section 3.1 are matching, conjugating, joining, fastening and meshing which are conducted by the equipment of the wrench and the screwdriver, also with hands. The operations of the assembly training are mainly the rotation and translation of workpieces. Thus, in the virtual environment, the trainee needs to move or rotate the virtual workpiece by hands and equipment to achieve the requirement of the assembly. According to the ways of the used equipment, the operations can be concluded in three situations: insert, equip and screw. The action sequences for each operation in the ARAAT are given below.

**Insert**: grasp $\rightarrow$ move $\rightarrow$ rotate/scale $\rightarrow$ release

**Equip**: grasp the wrench $\rightarrow$ move $\rightarrow$ move around $\rightarrow$ release the wrench

**Fasten**: grasp the screwdriver $\rightarrow$ move $\rightarrow$ rotate $\rightarrow$ release the screwdriver

### 4. Experiments and discussions

The experiments for action recognition are made on a self-made action dataset to evaluate the propose algorithm and compare with other approaches. The dataset is the action videos captured by the HoloLens RGB camera shown in Fig.7. The resolution of the camera is $1024 \times 720$.

The dataset contains 6 types of actions which are Point, Move, Grasp, Release, Scale, Rotate and Null, respectively. Actions in the dataset are performed by 10 people and 10 times for each action. So, the dataset contains 60 samples for each person and 600 videos in total. Each sample video has 100-150 frames. Each sample video has 100-150 frames. In the experiment, the results are tested by 5-fold cross-validation. The dataset is randomly divided into five equal subsets. The training set consists of four and the testing set consists of the remaining one. This procedure is repeated five times and every subset is used once for testing. The final recognition results are the average of the five testing results which is shown in Table 1. High action recognition accuracy can be observed and the average of all actions is 93.3% with several actions achieving outstanding recognition.

![Figure 7. the experiment device HoloLens](image)
The proposed algorithm is also compared with SSBoW[8], DSBoW[9], DTBoW[10,11], and DFW[12]. The recognition results are shown in Table 2. The performances of the proposed algorithm on the action dataset considerably outperform all the other ones. In the action recognition experiments, high accuracies which are all over 90% of each action are achieved. The SSBoW gets the lowest results in all actions because the type of each input frame cannot be explicitly classified which may result in the total low recognition rates. The recognition results of DFW are highest relative to the others, though lower than the one of proposed algorithm. It is due to the reason that the DFW cannot clearly segment each action. The wrong segmented frame may affect the precision of recognition.

Table 1: The recognition accuracy of ARAA actions with other

| Action | Point | Move | Grasp | Release | Scale | Rotate | null |
|--------|-------|------|-------|---------|-------|--------|------|
| Result | 93.4  | 96.2 | 91.6  | 91.2    | 91.7  | 90.6   | 95.2 |
|        | 94.2  | 95.1 | 92.3  | 91.5    | 90.9  | 91.7   | 96.8 |
| Average| 93.1  | 97.2 | 91.7  | 92.1    | 92.5  | 90.4   | 95.9 |
|        | 93.8  | 95.4 | 91.5  | 91.6    | 93.0  | 91.7   | 94.3 |
|        | 94.0  | 96.5 | 93.1  | 92.9    | 92.2  | 92.1   | 97.6 |

The experiment results are shown in Table 3. In the ARAAT experiments, high recognition rates which are all over 90% of each operation are achieved. The results confirm the efficiency and reliability of the proposed algorithm in ARAAT.

Table 2: The recognition accuracy of ARAA actions with other

| Action | Point | Move | Grasp | Release | Scale | Rotate | null |
|--------|-------|------|-------|---------|-------|--------|------|
| SSBoW  | 66.8  | 83.3 | 87.3  | 89.6    | 64.0  | 80.1   | 78.0 |
| DSBoW  | 72.0  | 89.7 | 90.1  | 90.3    | 70.0  | 82.4   | 83.3 |
| DTBoW  | 74.0  | 94.2 | 91.7  | 90.9    | 91.1  | 88.6   | 91.9 |
| DFW    | 89.6  | 93.7 | 88.4  | 86.1    | 85.3  | 87.7   | 90.5 |
| Proposed| 93.7  | 96.1 | 92.0  | 93.2    | 92.1  | 91.3   | 96.0 |

5. Conclusion

In this paper, a gesture recognition algorithm in ARAAT are proposed. According to the 2D and 3D features of the action, a scorer trained by the samples of specific actions gives the optimal label of each frame to recognize the action. To avoid the disturbance of the transition and invalid action to the action boundary during the recognition, the boundary is iteratively optimized by the probability density distribution. The proposed algorithm implemented on HoloLens is compared with other algorithms. The experimental results indicate that the proposed algorithm achieves a high recognition
rate and can reduce the computational complexity. The results prove the efficiency of recognition algorithm in ARAAT and ensure a friendly experience for the human-machine interaction.

References
[1] Imbert, Nicolas, et al. "Adding Physical properties to 3D models in augmented reality for realistic interactions experiments." Procedia Computer Science 25 (2013): 364-369.
[2] Lee, Minkyung, Richard Green, and Mark Billinghurst. "3D natural hand interaction for AR applications." Image and Vision Computing New Zealand, 2008. IVCNZ 2008. 23rd International Conference. IEEE, 2008.
[3] Wang, Zhenbiao, et al. "Assembly design and evaluation based on bare-hand interaction in an augmented reality environment." 2009 International Conference on CyberWorlds. IEEE, 2009.
[4] Li, Yan, and Xukun Shen. "A real-time collision detection between virtual and real objects based on three-dimensional tracking of hand." Audio Language and Image Processing (ICALIP), 2010 International Conference on. IEEE, 2010.
[5] Choi, Junyeong, et al. "Bare-hand-based augmented reality interface on mobile phone." Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on. IEEE, 2011.
[6] Hilliges, Otmar, et al. "Grasping virtual objects in augmented reality." U.S. Patent No. 9,552,673. 24 Jan. 2017.
[7] Laptev, Ivan. "On space-time interest points." International journal of computer vision 64.2-3 (2005): 107-123.
[8] Laptev, Ivan, et al. "Learning realistic human actions from movies." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.
[9] Hao, Zongbo, et al. "Human action recognition by fast dense trajectories." Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013.
[10] Wang, Heng, et al. "Dense trajectories and motion boundary descriptors for action recognition." International journal of computer vision 103.1 (2013): 60-79.
[11] Kulkarni, Kaustubh, et al. "Continuous action recognition based on sequence alignment." International Journal of Computer Vision 112.1 (2015): 90-114.