https://doi.org/10.1007/s10055-022-00675-x

ORIGINAL ARTICLE

Human–machine integration based augmented reality assisted wire-bending training system for orthodontics

Jiaqi Dong[1] · Zeyang Xia[2] · Qunfei Zhao[1] · Ning Zhao[3]

Received: 15 October 2021 / Accepted: 27 June 2022 / Published online: 6 August 2022

Abstract
With the increasing demand for orthodontic treatment, the skill of wire bending is more and more important for orthodontists. Traditional wire bending training needs a high cost of time and resources. In this paper, an augmented reality assisted wire-bending training system (ARAWTS) is proposed. ARAWTS provides 4 typical wire bending training tasks for the trainee and can give training feedback and improvement advice to the trainee by gesture recognition during the training. For the elaborate and vague wire bending gesture recognition, we develop a temporal logical relation (TLR) module to sparsely sample dense frames and learn the TLRs between frames of gestures. To reduce the computational cost and time, we introduce a new type of sparse optical flow called Focus Grid Optical Flow (FGOF). From the results of experiments, the proposed algorithm implemented on an AR device (HoloLens) achieves a high recognition rate with low computational complexity and ARAWTS is proved reliable.

Keywords Augmented reality · Wire bending · Gesture recognition · Temporal logical relation · Focus Grid Optical Flow · Orthodontic treatment

1 Introduction

An increasing number of patients are demanding orthodontic treatment for improved esthetics and a better mastication system. The orthodontic treatment needs the correct diagnosis, analysis, and rehabilitation design. Also, the medical skills of clinical operations are important and wire bending is a significant part (Lau et al. 2021; Sivarajan et al. 2021; Kono and Kikuchi 2020). In the training for orthodontists, wire bending training is indispensable. The traditional teaching and training pattern for wire bending is that the skilled teacher demonstrates the main wire bending operations and the trainee imitates the gestures. Only by repeatedly inquiring about the key points and techniques of teachers' operations and practice over and over again, the trainee can improve the wire bending skills. The wire bending training is a process of visual imitations and repeating practices under the important hints, which is different from the theoretical study of diagnosis and rehabilitation design. Thus, the wire bending training is time and resource cost.

Virtual reality (VR) technology is used for training since the 1990s. VR technology can reproduce physical objects of the real world in the virtual computer environment (Tang et al. 2021; Lee et al. 2021a). With sensor devices, people can achieve direct natural interaction with the virtual world. In the field of training, the effectiveness and practicability of VR technology have been proved in many researches (Zhou 2021; Osti et al. 2021; Lee et al. 2021b). However, the virtual 3D world created by VR technology is completely separated from reality. People are seeking a new way to better combining the virtual and real world.

Based on VR technology, a new technology named augmented reality (AR) is developed (Hughes et al. 2005; Rios et al. 2011; Vakaliuk and Pochtoviuk 2021). AR technology merges augmented information or virtual objects into the real world and improves the perception of reality. By applying virtual information to reality, AR technology can superimpose virtual objects and scenarios onto the real scenario to achieve a reality augmentation. Under this merged

✉ Zeyang Xia
zy.xia@siat.ac.cn

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China

[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[3] Shanghai Ninth People's Hospital, Shanghai JiaoTong University School of Medicine, Shanghai, China

environment, users can real-time interact with real and virtual objects in a more natural way. Different from the traditional interaction mode in VR environment which is human-dominating and machine assisting, AR technology can extend to a new interaction mode, that is human–machine integration (Kethman 2021; Ballesté and Torras 2013; Nyre-Yu 2019). Human–machine integration combines the subjective information from humans and the objective data from machine, constructs a new understanding method and gives optimization judgments by inter coordination between humans and machines. Human–machine integration converts the understanding between human and machine from one-way to two-way.

AR technology not only possesses the advantages of VR techniques in the field of training, but also has more merits to training in perception. By applying AR technology in the wire bending training, the trainees will not be isolated from reality and can be more natural to interact with the real scenario and virtual objects. Thus, authenticity, interactivity, and practicability are enhanced. AR assisted wire bending training is presently a leading-edge technology and a hot topic in the researches (Lo et al. 2021). The AR technology provides the repeating active goal practices which in the training process can effectively improve the skills of trainees and have a lower resource cost. In this way, the trainees can train their wire bending operations in the AR assisted environment and apply the learned skills to real life.

Because the wire bending is conducted by hand operations, gestures of the bending operations are important to the wire bending training in the AR environment. According to the recognition of wire bending gestures, we can give the trainee evaluations and suggestions to improve their skills. Gesture recognition is a core problem in computer vision. Many researches focus on convolutional neural networks (CNNs) to recognize gestures by frames. Simonyan and Zisserman (2014) proposed a two-stream CNN to capture the complementary information on appearance from frames for gesture recognition. Although CNNs can achieve state-of-the-art performance on many tasks of gesture recognition (Cheng et al. 2019), the high computational complexity and expensive training cost on dense frames made CNNs not meet the demands of real-time applications. Therefore, more and more works tend to design effective CNNs for gesture recognition (Wu et al. 2018). For example, Karpathy et al. (2014) fused the RGB frames on the temporal dimension and was evaluated on the Sport1M dataset. Tran et al. (2015) extracted the frame features by 3D convolution kernels on dense RGB input frames. Wang et al. (2016) proposed a Temporal Segment Network to sample frames and extract characteristics on different time segments for gesture recognition. Carreira and Zisserman (2017) studied an I3D network which used two steam CNNs to combine the RGB data and optical flows of the Kinetics dataset. These methods are mostly validated on many video datasets, such as Sport1M (Karpathy et al. 2014), Kinetics (Kay et al. 2017) and UCF101 (Soomro et al. 2012). These datasets include the gestures without the long-term temporal logical relation which can be identified enough by the frames of the labeled gestures. Thus, CNNs can perform well on these datasets for gesture recognition. However, CNNs are still struggling to deal with gestures which contain long-term temporal logical relations rather than the appearance of objects. Wire bending gestures are elaborate and vague which are hard to divide into several single gestures for recognition. Also, wire bending gestures contain strong temporal logical relations between frames. Because of the inherent ambiguity in the temporal extent, it is challenging for CNNs to accurately recognize the wire bending gestures (Wu 2020).

To address these issues, this paper designs an augmented reality assisted wire-bending training system (ARAWTS) which provides a natural environment of wire bending training for trainees. We define 20 key points on the hand and simultaneously extract the optical flows and positions of key points as the dynamic and static features of gestures. To reduce the computational cost and time, we introduce a modified sparse optical flow called Focus Grid Optical Flow (FGOF). We also develop a temporal logical relation (TLR) module for elaborate and vague gesture recognition. TLR module sparsely samples dense frames and learns the temporal logical relations between frames of gestures. In ARAWTS, the trainee can receive real-time training advice and evaluations to improve their wire bending skills.

This paper is organized in the following way. In the Second Chapter, ARAWTS for orthodontics and the definition of wire bending gestures are shown. The algorithm of gesture recognition in ARAWTS is introduced in the Third Chapter. The Fourth Chapter explains the human–machine integration in ARAWTS. The experiment results of gesture recognition and wire bending training are presented in the Fifth Chapter. Finally, the conclusion is provided.

2 Augmented reality assisted wire-bending training system (ARAWTS) for orthodontics

In this chapter, we will introduce the overall framework of ARAWTS and the definitions of wire bending gestures in ARAWTS.

2.1 Overall framework

ARAWTS is built on the AR device HoloLens (Hilliges et al. 2017) and can give real-time training advice and evaluations according to the trainee's operation videos. The framework of ARAWTS is shown in Fig. 1. The trainee carries on the

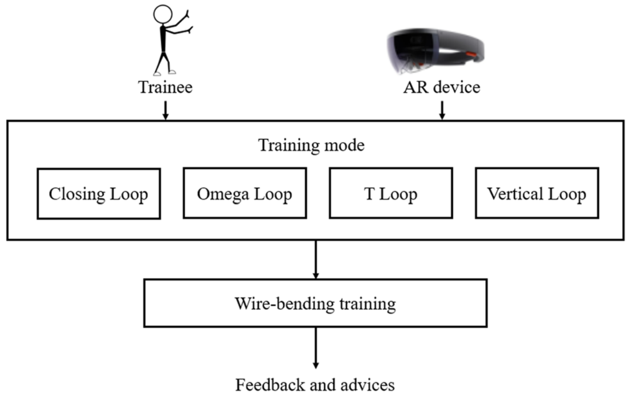

Fig. 1 The framework of ARAWTS

wire bending training in an AR training program. In the training program, there are multiple typical wire bending training tasks. The trainee can choose any training task at will. During the wire bending training, the relevant data are recorded. According to real-time gesture recognition, the system can give training advice for each step of the trainee's operations. At the end of the training task, a total evaluation is given and the system can tell the trainee which step maybe exist the problems and where needs to be improved.

2.2 Definitions of wire bending gestures

The tasks of ARAWTS are mainly conducted by hand operations. By the recognition of gestures, ARAWTS can analyze the standard and achievement of the hand operations and evaluate the performance of the training.

In this paper, we design four typical wire bending training tasks: Closing loop, Omega loop, T loop, and Vertical loop (Waters et al. 1975), which are illustrated in Fig. 2.

The training tasks are recorded to videos by cameras on an AR device. Let T be a given task of wire bending training and V be the recorded video corresponding to T. V can be expressed as a series of frames in a digital image, that is,

$$V = \{f_t, \quad t = 1, 2, \dots\} \tag{1}$$

where f_t is the tth frame.

To conduct the training task, the trainee needs to hold the wire with one hand and move the orthodontic plier to the appropriate position and rotate with both two hands. During the training operations, there are mainly two gestures, that is moving and rotating. According to the different distances of moving and degrees of rotating, various loops can be bent. However, the gestures in the wire bending operations are elaborate and vague which are hard to recognize by the common CNN methods. According to the strong temporal logical relations between frames, this paper proposes a gesture recognition algorithm below for elaborate and vague gesture operations like orthodontics wire bending which is detailed in chapter 3.

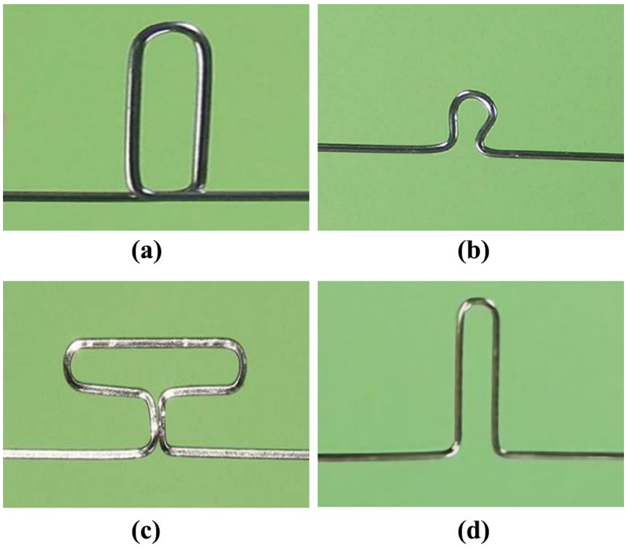

Fig. 2 Four types of wire bending training: **a** closing loop, **b** Omega loop, **c** T loop, and **d** vertical loop

3 Gesture recognition in augmented reality assisted wire-bending training system (ARAWTS)

In this chapter, the algorithms for gesture recognition in ARAWTS are introduced. A modified sparse optical flow called Focus Grid Optical Flow (FGOF) is proposed to reduce the computational cost and provide a smooth interaction in ARAWTS. Also, we develop a temporal logical relation (TLR) module for elaborate and vague gesture recognition which can obtain the temporal logical relations between each frame.

This chapter mainly involves the following parts: feature extraction, temporal logical relation (TLR) module and gesture recognition. The specific processing steps are as follows.

3.1 Feature extraction

For the tth frame f_t from the input video V of wire bending training tasks, we first use the bounding box (Mehta et al. 2017) to crop the target hand. Then, static and dynamic features are simultaneously extracted from cropped frames.

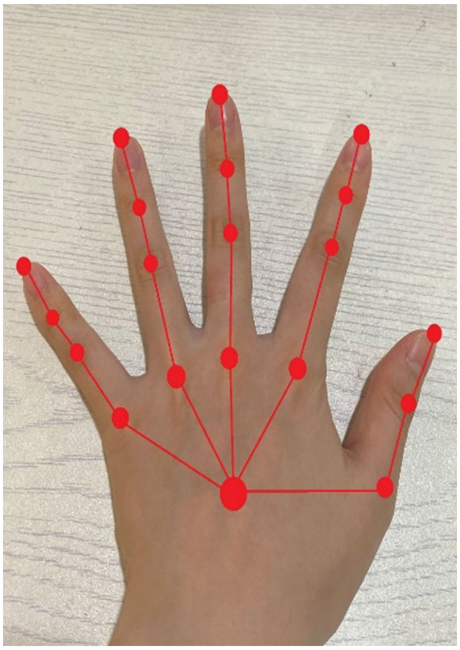

Fig. 3 20 key points of a hand

3.1.1 Static feature

We define 20 key points of the hand which are illustrated in Fig. 3. The two-dimensional key point positions of hands from the input cropped image denoted by P_t are extracted as the static features. P_t in frame f_t can be presented by

$$P_t = \left(p_t^1(x,y), p_t^2(x,y), \dots, p_t^m(x,y), \dots, p_t^{20}(x,y)\right) \quad m = 1, 2, \dots, 20 \tag{2}$$

where $p_t^m(x,y)$ is the two-dimensional position of the mth key point.

3.1.2 Dynamic feature

The optical flow from each key point is calculated as the dynamic feature for gesture recognition. However, the typical optical flow is dense and it will increase the computational complexity during gesture recognition. Also, the typical dense optical flow of the useless pixels may bring negative effects to the final recognition results. To reduce the negative effects of the dense optical flow, we propose a modified sparse optical flow named Focus Grid Optical Flow (FGOF).

First, some grids are built on the input image. d is denoted as the grid distance, H and W are the height and width of the input image. The number of grid points denoted by N_{lp} can be expressed as

$$N_{lp} = H/d \times W/d \tag{3}$$

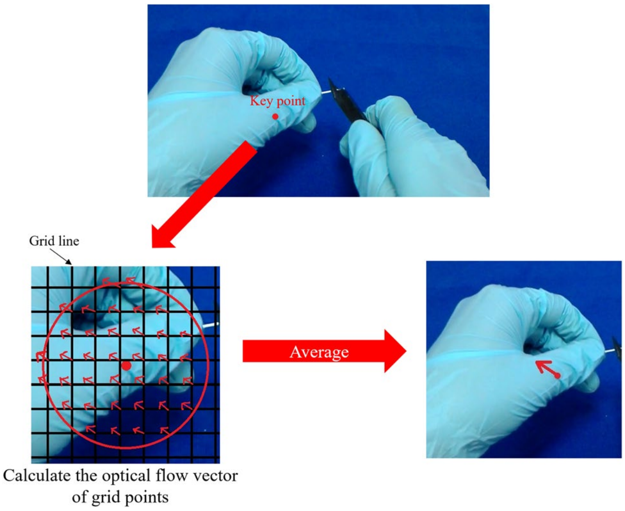

Fig. 4 The illustration of Focus Grid Optical Flow

Then, we don't calculate all the optical flow but only focus on key points. According to static features, that is the two-dimensional positions of key points, we obtain the optical flow vectors of the grid points which are in the circle area with a radius D around each key point. GOF(x, y) is defined as the optical flow vector of the grid point in each corresponding key point area and calculated by the Lucas-Kanade algorithm (Lucas and Kanade 1981).

Last, the average optical flow of these grid points denoted by AF^m represents the movement of each key points, as shown in Fig. 4. The average optical flow AF^m is defined as follows:

$$\mathrm{AF}^m(x,y) = \left\{ \frac{\sum \mathrm{GOF}_k^m(x,y)}{n^m} | D^m \right\} \tag{4}$$

where $\mathrm{GOF}_k^m(x,y)$ and n^m stand for the optical flow vector and the number of the k^{th} grid point for the mth key point, respectively, and D^m is the area radius of the mth key point. The setting for parameters d and D^m needs to be a good trade-off. A larger d may lead to low accuracy but a smaller d may result in heavier computations. Also, a larger D^m may increase the computational cost and a small D^m may reduce the recognition rate. Because only the optical flow of gird points needs to be calculated, the computation cost is reduced by at least d^2 times compared with typical dense optical flow. Also, due to the reason that the region of optical flow vectors is restricted, the influence of useless information can be reduced and the precision will increase.

Therefore, the optical flow in frame f_t can be presented by

$$\mathrm{AF}_t = \left(\mathrm{AF}_t^1(x,y), \mathrm{AF}_t^2(x,y), \dots, \mathrm{AF}_t^m(x,y), \dots, \mathrm{AF}_t^{20}(x,y)\right) \quad m = 1, 2, \dots, 20 \tag{5}$$

3.1.3 Feature matrix

In summary, the final feature matrix Ft of frame ft can be concluded as

$$F_t = \begin{bmatrix} P_t \\ \mathrm{AF}_t \end{bmatrix} \tag{6}$$

3.2 Temporal logical relation (TLR) module

Some elaborate and vague gestures, such as wire bending gestures, are hard to divide into several single gestures for recognition. Thus, in this part, we propose a temporal logical relation (TLR) module which can learn the strong temporal logical relations between frames of elaborate and vague gestures.

For a given input video V, we uniformly sample n frames and build a new TLR video set in a chronological order which is denoted by V', that is

$$V' = \{f_1', f_2', \dots, f_n'\} \tag{7}$$

F_i' and F_j' are the feature matrices of f_i' and f_j' which are the ith frame and the jth frame of V', respectively, $i,j \in \{1,2,\dots,n\}$. We define the temporal logical relations between two frames as below:

$$\mathrm{TLR}_2(V') = A_\alpha\left(\sum_{i<j} B_\beta\left(F_i', F_j'\right)\right) \quad i,j \in \{1,2,\dots,n\} \tag{8}$$

where A_α and B_β are the feature fusion functions and we use multilayer perceptrons (MLP) with parameters α and β, respectively, as the feature fusion functions. For efficient computation, we don't add all the combination pairs but uniformly sample pairs and sort them.

The two-frame temporal logical relations can be further extended to a higher level, for example, the three-frame temporal logical relations can be expressed by

$$\mathrm{TLR}_3(V') = A_\alpha'\left(\sum_{i<j<l} B_\beta'\left(F_i', F_j', F_l'\right)\right) \quad i,j,l \in \{1,2,\dots,n\} \tag{9}$$

The same as the two-frame TLR, we sample and sort the three-frame combinations for the computation rather than adding all the combinations.

To obtain the multiple time-scale temporal logical relations, we extend the TLR to N frames and define the multiple time-scale TLR as follows:

$$\mathrm{MTLR}_N(V') = \mathrm{TLR}_2(V') + \mathrm{TLR}_3(V') + \cdots + \mathrm{TLR}_N(V') \tag{10}$$

where each TLR_γ represents the temporal logical relation between γ sorted frames and each TLR_γ has separate feature fusion functions A_α^γ and B_β^γ. The illustration of MLP module is shown in Fig. 5.

3.3 Gesture recognition

The multiple time-scale temporal logical relations captured by the TLR module then are inputted into the gesture

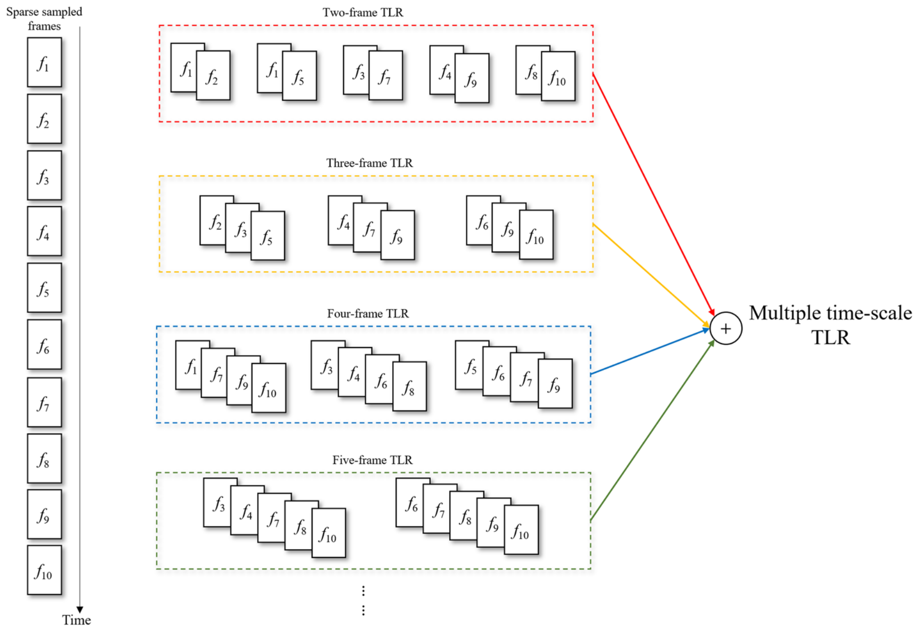

Fig. 5 The illustration of multiple time-scale TLR

recognition network. We choose the directional pulse coupled neuron network (Dong et al. 2019) as the gesture recognition network because DPCNN can classify and recognize dynamic gestures by template matching and is often applied to real-time applications.

For the sake of smooth interaction, it is essential to early recognize gestures of trainees. Through precise gesture prediction, the system can give reasonable training advice to trainees. Because obtaining full temporal logical relation features of hand movement is the most important part of gesture prediction, we use the static key point positions, dynamic FGOF information and multiple time-scale TLRs to do a regression on LSTMs (Zhu et al. 2017) and predict gestures in real time.

4 Human–machine integration system

During the wire bending, the trainee cannot complete wire bending operations in each step at once time. Because the materials of wire have different elastic coefficients, the wire may have a rebound during the bending operations. Even the gesture is correct and standard, the wire may not achieve the ideal shape. Thus, the trainee needs to continuously adjust to make the wire reach an ideal shape.

For this reason, only gesture recognition is not enough in ARAWTS. After the gesture recognition, we need to recognize the deformation of the wire as the assist in the system.

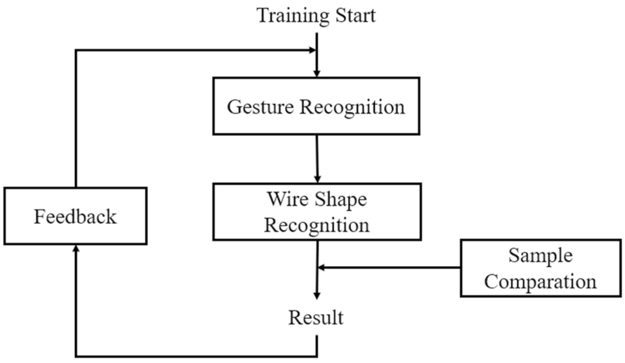

Fig. 6 Flowchart of augmented reality assisted wire-bending training system (ARAWTS)

After the evaluation of the trainee's gestures, the system will compare the practical wire shape with the sample template shape, and then, give feedback on improvement to the trainee. The trainee receives the feedback and can do the adjustment intentionally to achieve a standard loop shape. Then, such mutual feedback procedure is repeated until the whole training is completed. The flow of ARAWTS is illustrated in Fig. 6. ARAWTS is not a single machine system, it needs the trainee to take part in the whole training process. Through continuous feedback from humans and machines, the trainee can improve to the ideal state. Thus, ARAWTS forms a human–machine integration system.

5 Experiments

The experiments are designed into two parts. The first part is the gesture recognition experiments. We compared with several methods on three public datasets to evaluate the accuracy and efficiency of the proposed gesture recognition algorithm with the TLR module and FGOF. The second part is the wire bending training experiments to validate the reliability of the ARAWTS.

5.1 Gesture recognition

To evaluate the effectiveness of the proposed gesture recognition method, we conduct experiments on three public datasets: Cambridge Hand Gesture dataset (Kim et al. 2007), UCF sport dataset (Rodriguez et al. 2008) and MSR Daily Activity 3D dataset (Wang et al. 2012).

Cambridge Hand Gesture dataset is a commonly used benchmark for gesture recognition which consists of 900 image sequences of 9 gesture classes defined by 3 primitive hand shapes and 3 primitive motions. For each gesture class, there are 2 subjects' 10 arbitrary motions sequences under 5 different illuminations which are illustrated in Fig. 7.

UCF sport dataset consists of a set of actions collected from various sports which are typically featured on broadcast television channels such as the BBC and ESPN. The dataset contains 10 different sports and a total of 150 sequences. The sports are Diving, Golf Swing, Kicking, Lifting, Riding

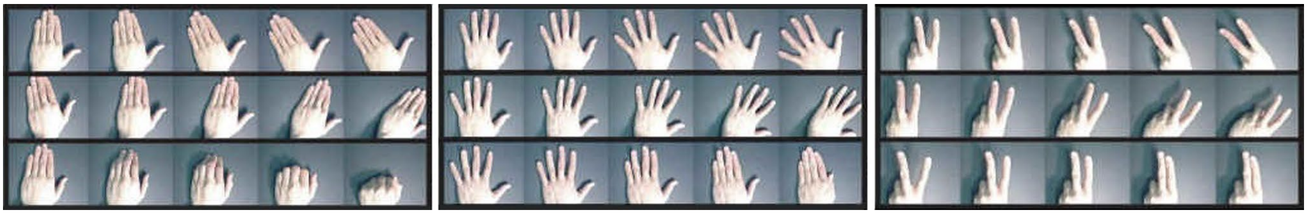

Fig. 7 Image frames from Cambridge Hand Gesture Dataset

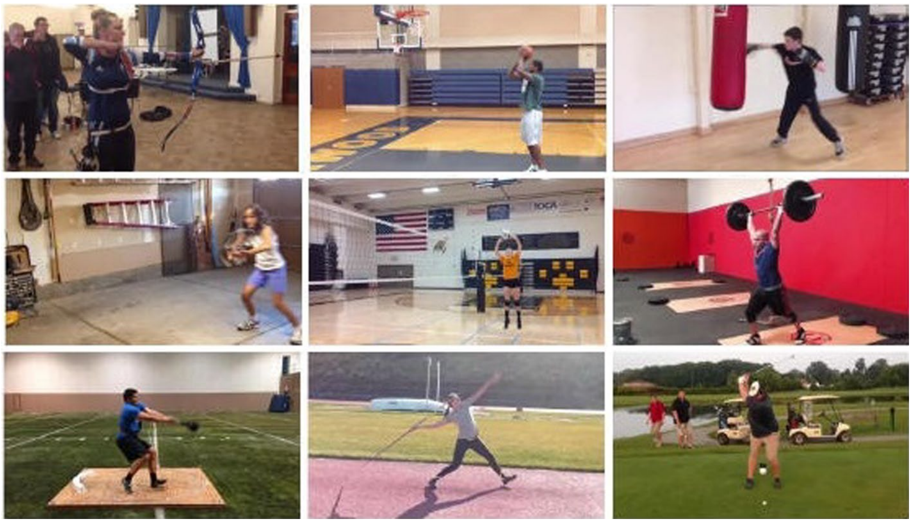

Fig. 8 Image frames from UCF sport Dataset

Fig. 9 Image frames from MSR Daily Activity 3D Dataset

Table 1 The accuracy of with TLR and without TLR on three datasets

	Hand Gesture (%)	UCF sport (%)	Daily Activity 3D (%)
Without TLR	90.7	83.5	84.1
With TLR	98.2	92.4	96.9

Table 2 The comparison between dense optical flow and FGOF on datasets

	Hand Gesture	UCF sport	Daily Activity 3D
Dense OF			
Computation time	188 ms	402 ms	386 ms
Accuracy	91.6%	80.7%	89.1%
FGOF			
Computation time	49 ms	77 ms	68 ms
Accuracy	98.2%	92.4%	96.9%

Horse, Running, Skate-Boarding, Swing-Bench, Swing-Side and Walking shown in Fig. 8.

MSR Daily Activity 3D dataset consists of 16 activity types which are drinking, eating, reading, calling, writing, using laptop, using vacuum cleaner, cheering up, siting still, tossing paper, playing game, laying downing, walking, playing guitar, standing up, sitting down. There are 320 activity sequences in total and part of activities are shown in Fig. 9.

To validate the efficiency of the TLR module, we compared the accuracy between "with TLR module" and "without TLR module" on three datasets. Also, we record the computation time and accuracy between "with FGOF" and "without FGOG" to prove the availability of FGOF. The results are shown in Tables 1 and 2.

In Table 1, the accuracy between "with TLR" and "without TLR" has only a difference of 4.5% in Cambridge Hand Gesture dataset but more than 25% in UCF sport dataset and MSR Daily Activity 3D dataset. This is due to the reason that gesture data between each class in Cambridge Hand Gesture dataset are with great differentiation and can be recognized from several isolated frames, but in UCF sport dataset and MSR Daily Activity 3D dataset are with small differentiation and strong temporal logical relation which can't be easily recognized only from isolated frames. Thus, the recognition rate without the TLR module is low on such strong temporal logical relation datasets. From Table 2, we can easily find that the computation time is reduced by more than 70% and the accuracy improves by at least 6.6% and even 11.7% on UCF sport dataset. Because only the optical flow of gird points needs to be calculated in FGOF, the computation cost is greatly reduced compared with typical dense optical flow. Also, FGOF restricts the calculation region of optical flow vectors, and the influence of useless information is reduced. The results from Tables 1 and 2 confirm the reliability of the TLR module and FGOF.

We also compared gesture recognition algorithm with other methods: PLSA (Wong et al. 2007), STCD (Sanin et al. 2013), DT + HS (Baraldi et al. 2014) and IT (Zhao and Elgammal 2008). The results are shown in Table 3. The results indicate that our algorithm achieves the best accuracy and outperforms all the compared methods on three datasets.

5.2 Wire bending training experiments

ARAWTS is built on Mixed Reality platform Microsoft HoloLens (1st Generation). Task videos of ARAWTS are captured by cameras on HoloLens. The resolution of the cameras is 1268 × 720 with 30 frames per second (fps). We invite 20 people to take part in wire bending training experiments. 20 people are divided into two groups: participants of one group are all orthodontists who have experience with wire bending; participants of the other group are all without experience in wire bending. Each person performs 4 types of training and each training is conducted 5 times which creates 400 videos in total. The demonstration scenario is shown in Fig. 10.

Table 3 The recognition accuracy on three datasets

Method	Hand Gesture (%)	UCF sport (%)	Daily Activity 3D (%)
PLSA	91.5	82.5	84.6
STCD	93.6	88.3	92.1
DT + HS	94.3	90.5	94.6
IT	96.2	90.8	96.3
Proposed	98.2	92.4	96.9

In ARAWTS, each training is divided into several steps. The system gives instructions at the beginning of each step, and evaluations at the end of each step according to the trainee's gestures. At the end of the whole training, the system will give an overall evaluation of the training and tell the trainee which step may be nonstandard. ARAWTS can remember the data of each type's last training, and in the next time of the same type's training, the nonstandard step will be further divided into several steps to refine the gestures of the trainee.

To evaluate the performance of wire bending training, we give an achievement evaluating indicators denoted by *Ach* which is defined as:

$$\text{Error} = \frac{1}{2}\left(\frac{\left|H_{\text{Result}} - H_{\text{Std}}\right|}{H_{\text{Std}}} + \frac{\left|W_{\text{Result}} - W_{\text{Std}}\right|}{W_{\text{Std}}}\right) \times 100\% \tag{11}$$

$$\text{Ach} = 1 - \text{Error} \tag{12}$$

where Error is the error between final loop and standard loop, H_{Result} and W_{Result} are the height and weight of the final loop, and H_{Std} and W_{Std} are the height and weight of the standard loop, respectively.

We compared the proposed algorithm with the traditional method (Zhao and Elgammal 2008) which is without TLR module and FGOG. Each group's average achievement for each time is given in Table 4. We can see that the evaluations of the proposed algorithm are all showing an upward trend for both two groups, which prove that ARAWTS can improve the wire bending skill of trainees. For trainees in the no experience group, the evaluation increases from 30.6% at the first time to 80.1% at the fifth time which is nearly two times of improvement. Since the initial evaluation of the orthodontist group is high, though the evaluations of the orthodontist group only rise by 4.2%, the results are still convincing in improving the wire bending skill. However, the results of the traditional methods are far from satisfactory. The average accuracy of the no experience group is only 10.4%. Although the orthodontist group obtains the average accuracy of 70.4%

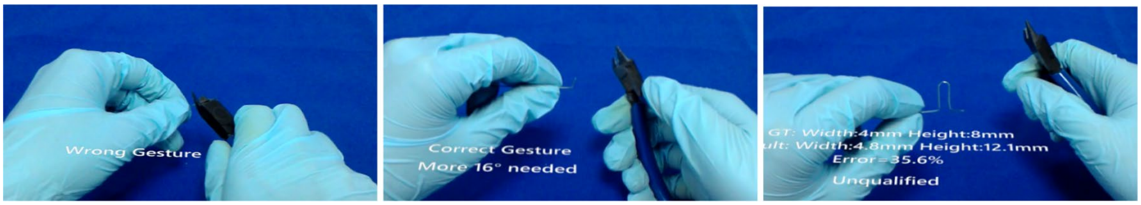

Fig. 10 Experimental scenario of ARAWTS

Table 4 The achievement of each group for five times

	Group	First (%)	Second (%)	Third (%)	Fourth (%)	Fifth (%)	Average (%)
Traditional method	Orthodontists	70.2	69.8	71.6	70.5	70.1	70.4
	No experience	10.9	10.3	9.5	11.4	9.7	10.4
Proposed	Orthodontists	94.1	94.7	96.6	97.5	98.3	96.2
	No experience	30.6	50.7	68.2	75.3	80.1	61.0

because of the previous professional skills, the whole results are far lower than that of the proposed algorithm. The results of the traditional method don't show an upward trend but also have a decrease at some times. This is because the traditional method without the TLR module has the great recognition errors to elaborate and vague gestures, and the recognition results mislead the trainee's wire bending operations. The results in Table 4 also prove the importance of TLR module to elaborate and vague gesture recognition.

6 Conclusion

In this paper, we proposed a human–machine integration-based augmented reality assisted wire-bending training system (ARAWTS). In ARAWTS, the trainee can choose the 4 typical wire bending training tasks. The system can give feedback and improvement advice to the trainee through gesture recognition and prediction during the training. We develop a temporal logical relation (TLR) to learn the temporal causal relations efficiently for the elaborate wire bending gesture recognition. We also introduce the action unit and develop a new type of sparse optical flow called Focus Grid Optical Flow (FGOF) to reduce the computational cost and time. From the results of experiments compared with the other algorithms, the proposed algorithm implemented on an AR device (HoloLens) achieves a high recognition rate with a low computational complexity which improves the efficiency and feasibility of the proposed algorithm and ARAWTS.

Acknowledgements This study was funded by the National Natural Science Foundation of China (U2013205).

Declarations

Conflict of interest The author declare that they have no conflict of interest.

Ethical statements Ethical review and approval were waived for this study, due to the nature of data collected, which does not involve any personal information that could lead to the later identification of the individual participants.

References

Ballesté F, Torras C (2013) Effects of human–machine integration on the construction of identity. In: Luppicini R (ed) Handbook of research on technoself: identity in a technological society. IGI Global, pp 574–591

Baraldi L, Paci F, Serra G et al (2014) Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 688–693

Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308

Cheng W, Sun Y, Li G et al (2019) Jointly network: a network based on CNN and RBM for gesture recognition. Neural Comput Appl 31(1):309–323

Dong J, Xia Z, Yan W et al (2019) Dynamic gesture recognition by directional pulse coupled neural networks for human-robot interaction in real time. J Vis Commun Image Represent 63:102583

Hilliges O, Kim D, Izadi S et al (2017) Grasping virtual objects in augmented reality: U.S. Patent 9,552,673. 2017-1-24.

Hughes CE, Stapleton CB, Hughes DE et al (2005) Mixed reality in education, entertainment, and training. IEEE Comput Graph Appl 25(6):24–30

Karpathy A, Toderici G, Shetty S et al (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732

Kay W, Carreira J, Simonyan K et al (2017) The kinetics human action video dataset. arXiv:1705.06950

Kethman W (2021) Human–machine integration and the evolution of neuroprostheses. In: Atallah S (ed) Digital surgery. Springer, Cham, pp 275–284

Kim TK, Wong SF, Cipolla R (2007) Tensor canonical correlation analysis for action classification. In: 2007 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–8

Kono H, Kikuchi M (2020) Analysis of orthodontic wire springback to simplify wire bending. Orthod Waves 79(1):57–63

Lau MN, Kamarudin Y, Zakaria NN et al (2021) Comparing flipped classroom and conventional live demonstration for teaching orthodontic wire-bending skill. PLoS ONE 16(7):e0254478

Lee SH, Cui J, Liu L et al (2021a) An evidence-based intelligent method for upper-limb motor assessment via a VR training system on stroke rehabilitation. IEEE Access 9:65871–65881

Lee SH, Yeh SC, Cui J et al (2021b) Motor indicators for the assessment of frozen shoulder rehabilitation via a virtual reality training system. Electronics 10(6):740

Lo YC, Chen GA, Liu YC et al (2021) Prototype of augmented reality technology for orthodontic bracket positioning: an in vivo study. Appl Sci 11(5):2315

Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision, vol 81, pp 674–679

Mehta D, Sridhar S, Sotnychenko O et al (2017) Vnect: real-time 3d human pose estimation with a single RGB camera. ACM Trans Graph: TOG 36(4):1–14

Nyre-Yu MM (2019) Determining system requirements for human-machine integration in cyber security incident response. Purdue University Graduate School, West Lafayette

Osti F, de Amicis R, Sanchez CA et al (2021) A VR training system for learning and skills development for construction workers. Virtual Real 25(2):523–538

Rios H, Hincapié M, Caponio A et al (2011) Augmented reality: an advantageous option for complex training and maintenance operations in aeronautic related processes. In: International conference on virtual and mixed reality. Springer, Berlin, Heidelberg, pp 87–96

Rodriguez MD, Ahmed J, Shah M (2008) Action Mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–8

Sanin A, Sanderson C, Harandi MT et al (2013) Spatio-temporal covariance descriptors for action and gesture recognition. In: 2013 IEEE workshop on applications of computer vision (WACV). IEEE, pp 103–110

Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. arXiv:1406.2199

Sivarajan S, Soh EX, Zakaria NN et al (2021) The effect of live demonstration and flipped classroom with continuous formative assessment on dental students' orthodontic wire-bending performance. BMC Med Educ 21(1):1–12

Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402

Tang YM, Ng GWY, Chia NH et al (2021) Application of virtual reality (VR) technology for medical practitioners in type and screen (T&S) training. J Comput Assist Learn 37(2):359–369

Tran D, Bourdev L, Fergus R et al (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497

Vakaliuk TA, Pochtoviuk SI (2021) Analysis of tools for the development of augmented reality technologies. In: CEUR workshop proceedings

Wang J, Liu Z, Wu Y et al (2012) Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 1290–1297

Wang L, Xiong Y, Wang Z et al (2016) Temporal segment networks: towards good practices for deep action recognition. In: European conference on computer vision. Springer, Cham, pp 20–36

Waters NE, Stephens CD, Houston WJB (1975) Physical characteristics of orthodontic wires and archwires—part 1. Br J Orthod 2(1):15–24

Wong SF, Kim TK, Cipolla R (2007) Learning motion categories using both semantic and structural information. In: 2007 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–6

Wu XY (2020) A hand gesture recognition algorithm based on DC-CNN. Multimed Tools Appl 79(13):9193–9205

Wu Y, Zheng B, Zhao Y (2018) Dynamic gesture recognition based on LSTM-CNN. 2018 Chinese Automation Congress (CAC). IEEE, pp 2446–2450

Zhao Z, Elgammal AM (2008) Information theoretic key frame selection for action recognition. In: BMVC, pp 1–10

Zhou J (2021) Virtual reality sports auxiliary training system based on embedded system and computer technology. Microprocess Microsyst 82:103944

Zhu G, Zhang L, Shen P et al (2017) Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access 5:4517–4524