Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Robot Awareness of Human Behaviors and Interacting with Humans by A Two-Step Identifying Approach combining LSTM and SSD

Chunfang Liu[1], Xiaoli Li[1], Weijia Niu[1], Qing Li[1], Muxin Chen[1]

Abstract— **Robot cooperating with humans is a hot research topic recently. And the first challenge is how to specifically identify humans' intention. In this paper, we propose a two-step approach based on LSTM and SSD for identifying humans' specific intention. A LSTM with attention mechanism is firstly presented for recognizing humans' actions based 3D skeleton data; then, the hand-held objects are detected by a SSD neural network. Finally, a human-robot interaction framework are constructed based on the proposed intention identification model and DMP model. The experiments show that the proposed method can effectively identifying humans' intention. And the robot accomplishes to interact with the human with the built motion database and DMP.**

I. Introduction

It has important application value for robots cooperating with humans. For instance, robots take care of elderly people and robot co-workers collaborate with humans in a workplace. Humans having the perception and cognitive abilities, are able to act and react with respect to a given situation. Correspondingly, the robots also have to understand the intention of human behavior and make decisions for assisting humans. Fig. 1 shows the framework of human-robot interaction in detail. We can see that there are three key problems for accomplishing human-robot interactions (HRI) [1], [2], which include (1) recognizing humans' behaviors, (2) making situationally appropriate decisions and (3) planning robot's corresponding motion for the interaction.

Thus, it is the first important task for HRI that the robot should be able to precisely aware humans' behaviours. For example, to hand a cup of tea to the human, the robot has to firstly aware that the human needs something to drink. However, in real-world environments, it is a challenging thing for robot awareness of human actions because of the complex and high dynamics of human motion, illumination variations and occlusion.

In recent years, human-action recognition is extensively studied due to its potential application fields. They recognize human actions from multiple modalities, including RGB images, depth, optical flows, and body skeletons. 2D RGB images perform to convey detail information of human action including motions and appearances (humans' clothing color or objects the human holding). However, huge amount of data processing and saving is always an important problem for directly using the whole image for motion recognition. In contrary, 3D body skeleton data use the dynamical variation of 15-30 human joint locations to express humans' actions. It is simple and effective for action recognition when the skeleton data are available. However, in the real communication situation, only skeleton data is not enough for robots determining humans' intentions. For example, the robot has to judge humans' next action by their hand-held objects. Humans' intentions may be different when the hand-held objects are different, even when the motions are same seeing from skeleton data.

In this paper, we propose a two-step identifying approach for recognizing humans' behaviors, including (1) identifying humans' motions from 3D skeleton data; (2) detecting the specific objects that the humans are holding from 2D images. The KinectV2 camera is utilized for capturing RGB images and detecting skeleton joint points. Then, the 3D joints are put into an action recognition model proposed by us, which is based on LSTM method with an attention mechanism. Near the end joint of the arm, single shot multibox detector (SSD) is utilized for detecting the holding object from 2D image. By utilizing both the recognition results of motions and hand-held objects, the robot judges humans' intention. Finally, he plans how to interact with humans and accomplishes the task by DMP (Dynamics Movement Primitive). Thus, there are mainly two contributions in our work: (1) proposed a human intention inferring model; (2) designed a human-robot interaction framework. Their effectiveness has been evaluated by human-robot real interaction experiments. Fig.2 illustrates the main human-robot interaction framework proposed by this paper. In the intention recognition stage, it consists of three sub-stages: (1) on-line detecting the important frames by the attention mechanism; (2) behavior identification under multi-tasks (action recognition and hand-held object detection); (3) human intention expectation.

II. Related Work

In this section, we review the approaches to understand human actions utilizing different kinds of data such as depth data, skeletal data, e.t. and using the attention mechanism.

Action Recogniton Based on Skeleton Data Human 3D skeleton data become easily accessible after the prevalent of KinectV2. Skeleton data are typical spatio-temporal sequences, which are invariant to locations and viewpoints. Thus, skeleton based action recognition has attracted increasing attention.

Many approaches has been proposed [3], [4], and recently they mainly focus on deep learning methods, such as Recurrent Neural Networks (RNNs) and Convolutional

[1]Chunfang Liu and Xiaoli Li are with the Department of Information, Beijing University of Technology, Beijing, China. cfliu1985@126.com lixiaolibjut@bjut.edu.cn 1225708606@qq.com 1104237004@qq.com 937773258@qq.com

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

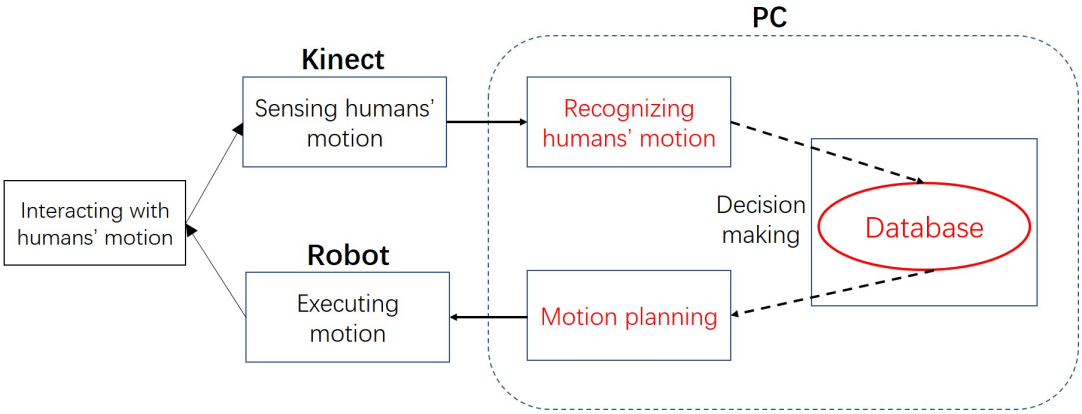


Fig. 1: Flowchart of the interaction between the human and robot

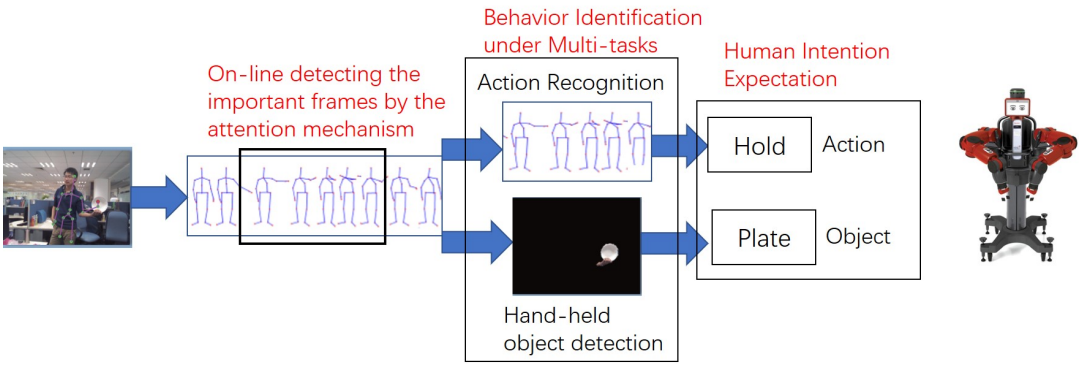


Fig. 2: The proposed human-robot interaction framework

Neural Networks (CNNs). Reference [5] proposed to represent a number of spatial skeleton features and adopted CNN to train actions. And the score fusion results made a final classification decision. In [6], the joint trajectories are encoded into images. Then, adopted ConvNets (CNN) to extract features for human action recognition. However, this method cannot distinguish some actions such as "knock" due to losing past temporal information [7]. Another approach [8] adopts both LSTM and CNN to perform the recognition task. In their approach, firstly, LSTM is utilized to model the temporal feature. Then, the CNN and LSTM channels are fused according to scores. It shows that the fusion of CNN and LSTM gets better results than only using RNN.

Attention Mechanism Attention mechanism is also a popular way recently for improving the recognition accuracy because of focusing on the data which have significant relation with the recognition task. When we identify actions, some frames are meaningless, which can be ignored; whereas, some others are important for correctly recognizing the actions. Therefore, paying more attention to the key frames is an effective way for improving the detection efficiency. Some approaches have been proposed by adding attention mechanism into the identifying process. By adopting skeleton data, [9] present a spatial and temporal attention model for human action recognition. In [10], they train the spatial and temporal attentions in different networks that the spatial attention model pays more attention to the relevant joints and the temporal model gives more importance to the relevant frames. Based on the sequential LSTM network, reference [11] develops a spatio-temporal model for 3D motion identification.

Action Recogniton Based on Multi – modal Data Multi-modal data increase the information for predicting humans' intentions. For example, the hand-held objects can be detected on 2D images for understanding humans' concrete intentions. Therefore, recent gesture/action recognition methods also deal with multi-modal data including RGB, depth data and skeleton modalities. In [12]–[14], RGB images are firstly fed into a CNN network. Then, 3D skeleton data are put into another network. Their results' fusion can be carried out in early or late layers. Reference [15] also utilizes both 3D pose data and RGB frames for human action recognition. They even develop a spatio-temporal soft-attention mechanism for building the relation between RGB images and 3D pose.

This paper is laid out as follows. Section II illustrates the main proposed human intention expectation method in which Section II-A presents the human action recognition method based on LSTM with an attention mechanism; Section II-B is the method of graspable component identification based on SSD neural network. Section IV illustrates the architecture of human-robot interaction. In Section V, the experimental

results verify the effectiveness of the proposed method and finally, Section IV gives the conclusion.

III. Human Intention Expectation

In a human-robot collaborative system, the robots should be able to specifically understand humans' intentions. By considering the computational complexity and the requirements of specific intentions, in this paper, we propose a two-step approach for the human intention expectation, including (1) identifying the humans' actions, and (2) recognizing the objects that the human are holding.

A. Human Action Recognition

Firstly, we briefly review the Long Short-Term Memory (LSTM) network. LSTM is a popular model for sequential data modeling and feature extraction. LSTM network mainly includes three layers: one input layer h^l, one hidden layer h^{l+1} and one output layer y as shown in Fig.3. The LSTM cell is the core of the LSTM network. With the LSTM cell shown in Fig.4, the information can be removed or added to the cell state by the forget gate, input gate and output gate.

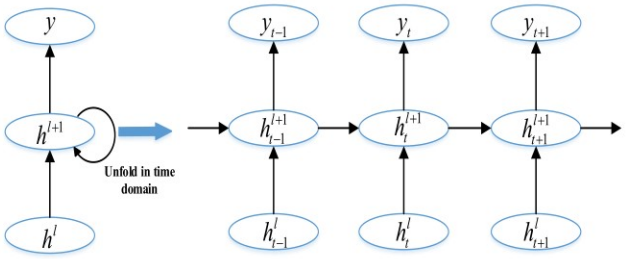


Fig. 3: Structure of LSTM neurons.

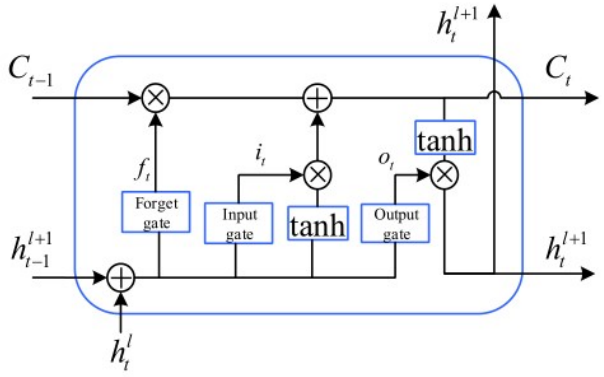


Fig. 4: Structure of LSTM cell.

In the LSTM cell, the input, the hidden layer state and the cell state at time t are denoted as h_t^l, h_t^{l+1}, and c_t, respectively. The updating rule of the LSTM cell can be summarized as,

$$f_t = \sigma(W_{xf}h_t^l + W_{hf}h_{t-1}^{l+1} + b_f) \quad (1)$$

$$i_t = \sigma(W_{xi}h_t^l + W_{hi}h_{t-1}^{l+1} + b_i) \quad (2)$$

$$c_t = f_t \times c_{t-1} + i_t \times tanh(W_{xc}h_t^l + W_{hc}h_{t-1}^{l+1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo}h_t^l + W_{ho}h_{t-1}^{l+1} + b_o) \quad (4)$$

$$h_t^{l+1} = o_t \times tanh(c_t) \quad (5)$$

where f_t,i_t,o_t are the forget gate, input gate and output gate, respectively. W_{xf}, W_{hf},W_{xi},W_{hi},W_{xc},W_{hc},W_xo,W_{ho} are the weight matrices. b_f,b_i,b_c,b_o are the bias vectors. $\sigma(\bullet)$ represents the function as follows:

$$\sigma(x) = 1/(1+e^{-x}) \quad (6)$$

After obtaining the hidden layer, a fully connected layer performs a linear transformation on the hidden state h_t^{l+1} to obtain the output of the LSTM network:

$$y_t = W_{yt}h_t^{l+1} + b_y \quad (7)$$

where (W_{yh}, b_y) are the weights and bias of the output layer.

Based on 3D skeleton data, this paper develops an approach for action recognition that LSTM network is adopted for extracting the skeleton sequence feature and an attention mechanism is presented for focusing on important frames.

It is supposed that in a motion sequence, some frames are important for discrimination while some other frames have less valuable information. To improve the identification ability, we design an attention mechanism to focus on important frames for recognition. When classify the skeleton motion sequences, in our main LSTM network, the temporal attention value β_t at each time step t are utilized for computing the class scores. We sum the scores by weights at all time steps as follows:

$$O = \sum_{t=1}^{T} \beta_t \cdot y_t, \quad (8)$$

where $O = (o_1, o_2, ..., o_C)^T$, T denotes the sequence length.

For a motion sequence X, the predicted probability of the i^{th} is

$$p(C_i|X) = \frac{e^{O_i}}{\sum_{j=1}^{C} e^{O_i}}, k = 1, ..., C. \quad (9)$$

The attention parameter β_t is computed as Equation 10, which associated with an LSTM layer, a fully connected layer, and a ReLU non-linear unit.

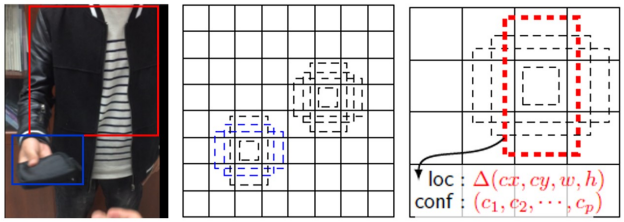
$$\beta_t = ReLU(w_{x\sim}x_t + w_{h\sim}\tilde{h_{t-1}} + b_\sim), \quad (10)$$

where x_t is the current input, $\tilde{h_{t-1}}$ is the hidden variable at time step $t-1$.

Notice that the main LSTM network and the temporal attention mechanism have been trained together for learning the temporal attention model.

B. Graspable Component Identification

In order to understand humans' specific intention, this section uses SSD approach to identify the object held by the human. Fig.5 shows the framework of SSD, which contain 13 convolution layers and three fully connected layers (See Fig.6). This network adopts VGG16 as the main network and it performs producing several bounding boxes and scores to estimate the possibility that objects are in those boxes.



(a) Image with detected boxes (b) 8*8 feature map (c) 4*4 feature map

Fig. 5: SSD framework.

In our work, the default boxes are added position constrains since the robots pay more attention to the hand-held objects. This position constrains in 2D image are obtained from the 3D positions. The end-point in the arm skeleton is decided as hand in 3D space. Then, its corresponding 2D area is considered as the position possible area. Boxes with different sizes and directions are generated for detecting the hand-held objects.

IV. The Architecture of Human-Robot Interaction

After identifying humans' intention, the robot makes decision for motion planning. Here, an action database is built for robots motion planning. Dynamic Movement Primitives (DMP) are utilized for modelling the action trajectories. It is a generic framework for motor representation based on nonlinear dynamic systems. In the model, different action trajectories are represented with fewer characteristics. Then, the learned movement trajectories can be reproduced by inputting new start points and end points. The following equations illustrate the formation of DMP.

When fitting the whole motion trajectory, a scale attribute $g - y_0$ is provided. The forced term has convergence at the stage close to the target point.

$$f_{target} = \tau^2 \ddot{y}_{demo} - \alpha_z(\beta_z(g - y_{demo}) - \tau \dot{y}_{demo}) \tag{11}$$

We can get the loss function based on the local linear regression(LWR), where $\xi(t) = x(t)(g - y_0)$:

$$J_i = \sum_{t=1}^{P} \psi_i(t)(f_{target}(t) - w_i \xi(t))^2 \tag{12}$$

Finally, we can get:

$$w_i = \frac{s^T \Gamma_i f_{target}}{s^T \Gamma_i s} \tag{13}$$

In the robot action database, some kinds of actions are described by DMP such as Greeting, Fighting, Salute, Picking, Handshake, and so on. Once an action in the database are decided as the suitable way for robot motion, the DMP model will be used again for generating the trajectories. Combining with the human intention recognition, the robot realizes interaction with the humans.

V. Experiments and Results

This section tests the models proposed for human intention identification and for human-robot interaction.

A. Human Intention Recognition

The proposed action recognition model LSTM with attention mechanism is evaluated under the SBU dataset. SBU Kinect Interaction Dataset (SBU) [16] is an interaction dataset with two subjects. It contains 8 classes of actions and totally 230 sequences. Each person has 15 joints. With 90% of data for training and 10% of data for testing, we test the model 10 times and compared it with another action recognition method on skeleton data(ST-GCN model [17]). Figure 7 demonstrates the comparison results. Its horizontal ordinate represents the test number and the vertical ordinate represents the accuracies of action recognition. The blue color bar is the proposed LSTM with attention mechanism and the yellow bar is the ST-GCN model. We can see from Fig.7 that the accuracies of action recognition by the proposed method and the comparing model are approximately same. However, the algorithm processing speed of the proposed model is shorter than the comparing model.

Then, in on-line situation, we use the proposed model for recognizing the humans' five actions such as picking, greeting and so on. Figure 8 shows the collected skeleton data of different human actions by Kinect sensor.

B. Human-Robot Interaction

In this experiment, we first build a database for robot motion planning. It is expected that by using this database, the robot could choose suitable action for interacting with humans. Notice that the motion trajectories of different actions are not directly putting into the database. They are modeled by DMP that only a few data are utilized for describing the trajectories. Then, when robots use the trajectory, by giving the start point and the end point, the trajectory can be generated using DMP with the few data in the database. Figure 9 demonstrates the original motion trajectories (the blue color) and the trajectories generated by DMP (the red color). They include the actions of Greeting, Fighting, Salute, Picking and Handshake. Figure 10 shows a complex action trajectory. We can see that the generated trajectories almost coincide with the original trajectories, even the most complex trajectory in Fig.10.

Finally, we realize that the robot interacts with a human. Figure 11 and Figure 12 demonstrate the human-robot interaction results. When the human makes a greeting action, fighting action or salute action, the robot acts like the human. Then, when the human holds a book to give the robot, the

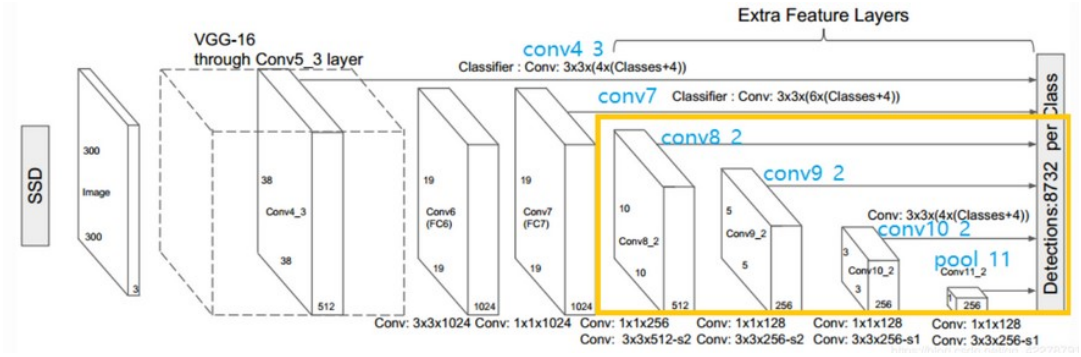


Fig. 6: The structure of SSD.

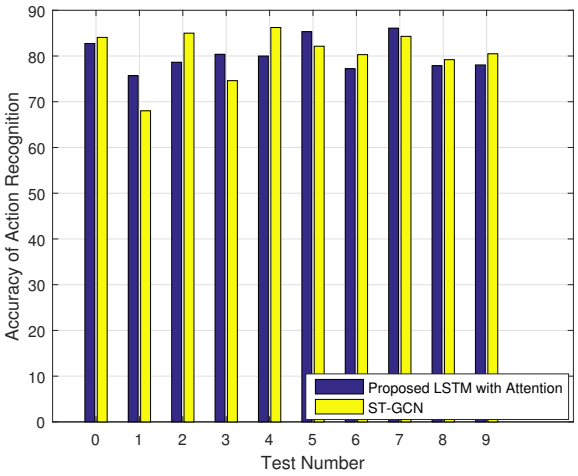


Fig. 7: Accuracy of Action Recognition

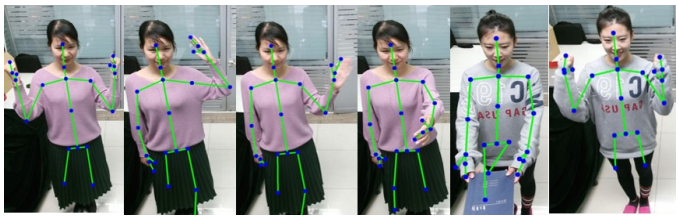


Fig. 8: Skeleton data of different human actions

robot plans to pick the book. And when the human stretches out his hand for handshake, the robot also plans to do the same thing.

VI. Conclusion

In this paper, we propose a two-step approach for identifying humans' specific intention. It includes an action recognition model by LSTM with attention mechanism and a hand-held object detection model by SSD neural network. Then, a human-robot interaction framework is constructed based on the proposed intention identification model and DMP model. The experiments show that the proposed method can effectively identifying humans' intention. And the robot accomplishes to interact with the human with the motion database and DMP.

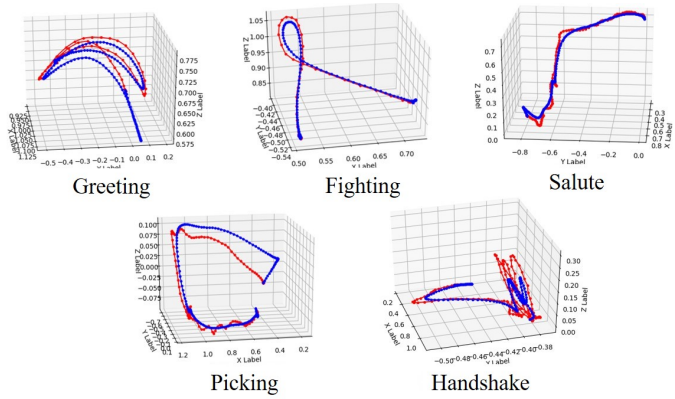


Fig. 9: The original trajectories and the trajectories generated by DMP

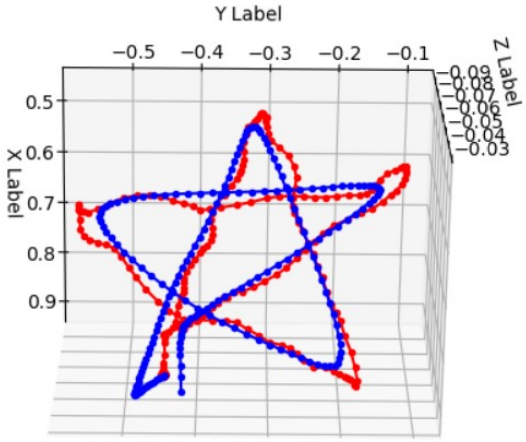


Fig. 10: A complex trajectory

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant 61703230.

References

[1] Tianmin Shu, Xiaofeng Gao, Michael S.Ryoo, and Song-Chun Zhu. Learning social affordance grammar from videos: transferring human interactions to human-robot interactions. In *ICRA*, 2017.

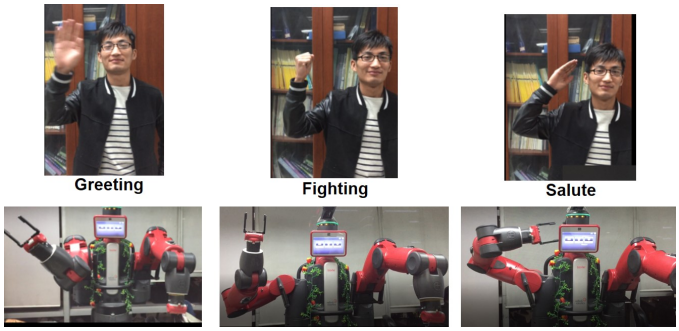


Fig. 11: Human-robot interaction by the proposed method (Greeting, Fighting, Salute)

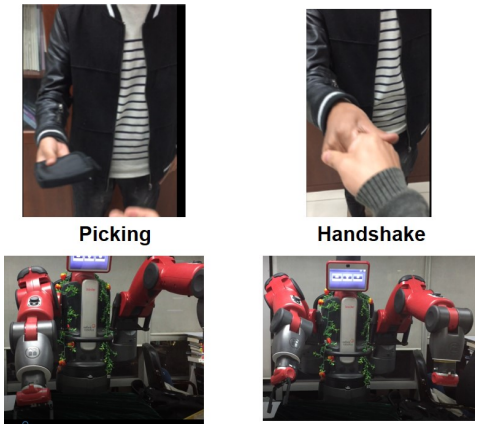


Fig. 12: Human-robot interaction by the proposed method(Picking, Handshake)

[2] Tianmin Shu, Michael S. Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. In *IJCAI*, 2016.

[3] M.E. Hussein, M. Torki, M.A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*, 2013.

[4] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, pages 588–595, 2014.

[5] Zewei Ding, Pichao Wang, Philip O. Ogunbona, and Wanging Li. Investigation of different skeleton features for cnn-based 3d action recognition. In *IEEE International Conference on Multimedia and Expo Workshops*, pages 617–622, 2017.

[6] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks. In *CVPR*, 2016.

[7] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. Differential recurrent neural networks for action recognition. In *IEEE International Conference on Computer Vision*, pages 4041–4049, 2015.

[8] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *ICMEW*, 2017.

[9] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017.

[10] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2016.

[11] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, pages 816–833, 2016.

[12] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural for human action recognition. *IEEE TPAMI*, 35(1):221–231, 2013.

[13] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE TPAMI*, 38(8):1692–1706, 2016.

[14] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE TPAMI*, 38(8):1583–1597, 2016.

[15] Fabien Baradel, Christian Wolf, and Julien Mille. Human action recognition: pose-based attention draws focus to hands. In *ICCV Workshop on Hands in Action*, 2017.

[16] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[17] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[18] P. Molchanov, X. Yang, S. Gupta, K. Kim, and S. Tyree. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *CVPR*, pages 4207–4215, 2016.

[19] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: a large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[20] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *Asian Conference on Pattern Recognition*. IEEE, 2015.

[21] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2017.

[22] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*. IEEE, 2017.

[23] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 2016.