EasyChair Preprint

№ 5294

Human-Robot Interaction Method Combining Human Pose Estimation and Motion Intention Recognition

Yalin Cheng, Pengfei Yi, Rui Liu, Jing Dong, Dongsheng Zhou and Qiang Zhang

April 6, 2021

Human-robot interaction method combining human pose estimation and motion intention recognition

1st Yalin Cheng
Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education) Dalian University
Dalian, China
13513635143@163.com

2nd Pengfei Yi
Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education) Dalian University
Dalian, China
istrrr@163.com

3rd Rui Liu
Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education) Dalian University
Dalian, China
liurui@dlu.edu.cn

4th Jing Dong
Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education) Dalian University
Dalian, China
dongjing@dlu.edu.cn

5th Dongsheng Zhou
Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education) Dalian University
Dalian, China
donyson@126.com

6th Qiang Zhang
Dalian University and Dalian University of Technology
Dalian, China
zhangq@dlut.edu.cn

***Abstract*—Although human pose estimation technology based on RGB images is becoming more and more mature, most of the current mainstream methods rely on depth camera to obtain human joints information. These interaction frameworks are affected by the infrared detection distance so that they cannot well adapt to the interaction scene of different distance. Therefore, the purpose of this paper is to build a modular interactive framework based on RGB images, which aims to alleviate the problem of high dependence on depth camera and low adaptability to distance in the current human-robot interaction (HRI) framework based on human body by using advanced human pose estimation technology. To enhance the adaptability of the HRI framework to different distances, we adopt optical cameras instead of depth cameras as acquisition equipment. Firstly, the human joints information is extracted by a human pose estimation network. Then, a joints sequence filter is designed in the intermediate stage to reduce the influence of unreasonable skeletons on the interaction results. Finally, a human intention recognition model is built to recognize the human intention from reasonable joints information, and drive the robot to respond according to the predicted intention. The experimental results show that our interactive framework is more robust in the distance than the framework based on depth camera and is able to achieve effective interaction under different distances, illuminations, costumes, customers, and scenes.**

***Index Terms*—human-robot interaction; human pose estimation; intention recognition**

I. Introduction

HRI technology has become an important research hotspot in the field of robot application. In order to complete effective interaction, the robot needs to rely on its own sensors, which can sense sound, force, distance and image, etc. Compared with other sensors, image sensors are cheaper and convey richer information. Studies [1] showed that 93% of human communication is non-verbal, 55% of which is physical communication. Physical communication is the most intuitive way of communication between humans and robots. Therefore, we explore the HRI method based on human pose from the perspective of robot vision. Interaction happens if a human and a robot sharing the same workspace are communicating with each other [2]. In the HRI based on actions, the robot interacts with the human's actions as the guide. After the robot gives feedback, the human can start a new round of interaction.

At present, HRI research based on human pose mostly relies on external depth sensors, such as Microsoft's Kinect. Kinect [3] has a strict software configuration environment, the official hardware interface is not open enough, the compatibility is weak, and it is not friendly to the embedded environment, so the existing HRI frameworks lack universality. In addition, Kinect operates in range of 0.5-4.5 m, and the range of stable human body pose capture is limited to 4 m, which is not available for people at a relatively long distance. But in the actual scene, it is necessary to interact with people in the farther scene, whether for industrial robots or indoor service robots. These restrictions reduce the flexibility of HRI.

Compared with depth camera, optical camera is not sensitive to distance requirements, and has no strict software configuration environment and hardware interface. In order to improve universality and flexibility of the HRI framework based on the depth camera, we use optical camera instead of depth cameras as the acquisition device. We introduced advanced human pose estimation methods and built a set of modular interaction framework based on human motion

Correspondence Author: Pengfei Yi(istrrr@163.com)

intention, which can extract effective human joints information from RGB images and predict human intention based on the joints information, so that drive the robot to complete effective interaction.

II. Related work

For a long time, people have been accustomed to use voice, expression, body language, body contact and other information when interacting with each other, which are the solidified interaction modes formed by people for a long time. Therefore, the interaction framework for service-oriented robots is mostly based on human gestures, actions, expressions, intentions and other information as the basis of interaction.

Carli et al. [4] and Zhu et al. [5] proposed an implicit Markov model to infer user's operation intention. Wang et al. [6] established a dynamic model for intention recognition, used Bayesian theorem to estimate the probability distribution of intention from observation and infer the user's real intention. Elfring et al. [7] used a growing hidden Markov model to predict human target location. Petković et al. [8] used Markov and TOM theory to judge people's intentions. Liu et al. [9] proposed an implicit control method and used a finite state machine to recognize the human intention. Yu X et al. [10] proposed a Bayesian method to acquire the estimation of human impedance and motion intention in a human-robot collaborative task, in which the human stiffness and human motion intention are obtained by Bayesian estimation with the human prior knowledge. These intention recognition methods most use the traditional probabilistic correlation model to predict the intention of part of human body.

Sigalas et al. [11] proposed a HRI system based on gesture recognition, which identifies the tasks to be performed through the category recognition of gestures. Li X [12] proposed a robot arm interaction model based on gesture recognition and body movement recognition by DTW template matching algorithm with both RGB video frames and depth images. Luo X et al. [13] proposed a two-handed gesture recognition method based on depth camera for real-time control of McNaim wheeled mobile robot, which can perform tasks such as directional movement, grasping, and clearing obstacles based on gesture recognition. Koppula et al. [14] predicted the motion of the extracted skeleton information based on the depth image. GerardCanal et al. [15] used depth images collected by KinectTMv2 sensor for gesture recognition. Li Kang [16], a scholar of Chinese academy of sciences, and others used Kinect depth camera to recognize the directly acquired three-dimensional human skeleton, and completed the interaction task on the design of HRI rules. Mazhar et al. [17] combined Kinect V2 extraction of three-dimensional pose with two-dimensional pose estimation method to recognize the extracted hand joints, thereby enabling the robot to complete the interaction.

In summary, human intention and action play a very important role in HRI framework. In addition, the depth camera has been found to have a high usage rate in the HRI framework, while depth cameras generally use infrared for ranging, which makes the interaction frames unable to well adapt to interaction scenes at different distances due to the influence of infrared detection distances. Therefore, we will introduce advanced human pose estimation methods, extract effective joints information from RGB images, and then predict human motion intention according to the joints information, and build a set of modular interactive framework based on human motion intention.

III. Method

This section mainly introduces the design ideas of our framework to ensure that it can enable the robot to effectively interact only according to the image captured by optical cameras under different conditions. The module schematic of our framework is described in section III-A. The human joints sequence filter is presented in section III-B. A detailed description of the human intention recognition method based on joints sequences is given in section III-C.

A. Overview

In order to improve the universality and flexibility of HRI framework based on depth camera, we will replace depth camera with optical camera as acquisition device to build a HRI framework based on RGB images sequence instead of depth images. The purpose of our framework is to enable the robot to complete the corresponding interaction on different scenes over a long distance according to the human intention extracted from the RGB images sequence. For example, in a specific offensive and defensive scenario, when the robot sees a person hurling a blow at it, even at a relatively long distance, the robot can recognize the behaviour intention according to the images information and make a corresponding defence according to this intention. Because of the interaction action cycle is a short-time, we will use the human action category as the human intention.

Our framework takes the human joints information contained by the image as the intermediate data, which not only considers the interaction framework independence between modules, but also adapts to different acquisition equipment. In order to enable the robot to complete effective interaction, it is necessary to infer the information conveyed by the human joints sequence. We use multi-frame joints sequence as the intention recognition basis, which not only reduces the complexity of the intention recognition network, but also makes the framework still available when the depth camera is used as the acquisition device.

As shown in Fig. 1, when the robot sees the images of the human right arm attacking forward, it completes the action of protecting the head. Our framework involves two stages: Human pose estimation stage and Intention recognition stage. The first stage is the robot's camera capture images, using the robot system interface to obtain the images sequence recorded by the camera, and input the images acquired by the robot into the human pose estimation network to extract the three-dimensional joints features of the human body contained by the image. The second stage is to recognize the human

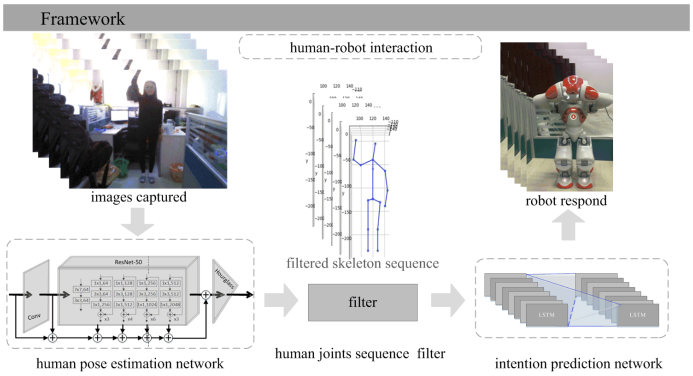

Fig. 1. Overview. Our framework enables the robot to complete corresponding interactive actions only according to the images captured by the camera.

intention according to the reasonable human joints features extracted from the image, and transmit the human intention category to the robot, so that the robot can make an expected response. In order to effectively reduce the wrong interaction caused by unreasonable joint information, we set a human joints sequence filter between the first stage and the second stage to further ensure the accuracy of human joints information. Human pose estimation stage by using image recognition technology to extract important features of human body joints, represented by convolution neural network method, this kind of method can extract 3d joints information of human body from the images obtained by robot vision sensor.

Intention recognition network stage aims at the specific interaction group of human and robot, and its purpose is to enable the robot to make corresponding feedback actions after acquiring effective human joints sequences. For example, when the robot sees the image of the man's right arm attacking forward, the trained intention recognition network can recognize the intention category of the people, and drive the robot to make corresponding feedback action to protect its head.

B. Human joints sequence filter

Based on the deep learning method, even if the network depth reaches more than a thousand layers, it is still possible to make mistakes. What's worse is that the algorithm does not know when the error occurs. It's a question of uncertainty. Human pose estimation stage using deep neural network is the intermediate stage of the HRI framework, so we need to further identify the accuracy and reliability of the prediction results. We use the length relationship between skeletons and the position relationship of the joints to roughly exclude unreasonable predictions to improve the accuracy of HRI.

As shown in Fig. 2, the joints coordinates output from the human pose estimation network are connected and displayed in the form of the skeleton. The first two human skeletons are obviously unreasonable. For example, in the first human skeleton, the left forearm length is too long because the prediction of left elbow joint's vertical dimension is much larger than the actual position, and in the second human skeleton, the left forearm length is too long because the

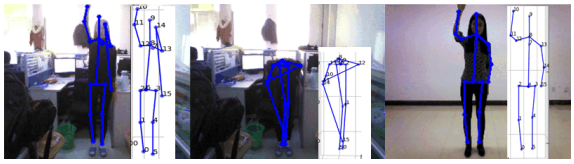

Fig. 2. The schematic diagrams of two unreasonable skeletons and a reasonable skeleton. There are skeletons extracted from three images. The first two skeletons are obviously unreasonable, and the last one is a reasonable skeleton schematic.

prediction of left elbow joint's horizontal dimension is much larger than the actual position. According to these obvious problems, we designed a human joints sequence filter based on the length relationship between skeletons and the position relationship of the joints. When similar unreasonable skeleton information is input, the human joints sequence filter will make them prohibit. Only input a reasonable skeleton similar to the last one can pass through the human joints sequence filter successfully. These unreasonable skeletons will not enter the intention recognition stage to avoid adverse effects on the latter stage.

Based on the output by human pose estimation network, according to the standard of Chinese adult body size [18], we designed the corresponding joints limitations.

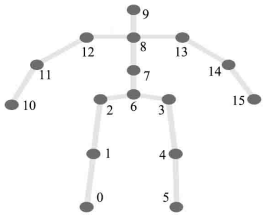

Fig. 3. The joints numbering diagram. Each joint position is marked as a fixed serial number, for example, the head joint is numbered 9, and the right and left foot joints are numbered 0 and 5.

- The height of the foot joint should not exceed the height of the head joint.
- The length of the thigh and calf does not exceed the sum of the length of the upper body and the head.
- The length of upper arm and forearm should not more than 1.5 times the length of the calf.

According to the above limitations, we designed a rationality judgment function of a single pose and used $v(p_i)$ to represent the filtering result of the i-th joints sequence.

$$v(p_i) = l_1(p_i) \times l_2(p_i) \times l_3(p_i) \tag{1}$$

l_1, l_2, l_3 are joints rationality limit functions, which are expressed as follows:

$$l_1(p) = \begin{cases} 0 & h_0 > h_9 \ or \ h_5 > h_9 \\ 1 & other \end{cases} \tag{2}$$

$$l_2(p) = \begin{cases} 0 & d_{0,1} > d_{8,9} + d_{6,8} \ or \ d_{4,5} > d_{8,9} + d_{6,8} \\ 1 & other \end{cases} \quad (3)$$

$$l_3(p) = \begin{cases} 0 & d_{11,12} > 1.5 \times d_{0,1} \ or \ d_{14,15} > 1.5 \times d_{0,1} \\ 1 & other \end{cases} \quad (4)$$

p denotes the pose information contained in an image, h_i denotes the distance between the i-th joint shown in the Fig. 3 and the ground, and $d_{j,k}$ denotes the distance between the j-th joint and the k-th joint.

$$R(X) = \begin{cases} reasonable & \sum_{i=0}^{N} v(p_i) \geq M \\ unreasonable & other \end{cases} \quad (5)$$

$R(X)$ represents the reasonable judgment function of all skeletons sequence, $X = [p_0, p_1, ..., p_N]^T$, N represents the length of the filtered skeletons sequence, and M represents the minimum number of skeletons in a reasonable skeletons sequence that conforms to the above restriction.

The human joints sequence filter is primarily for obvious errors in joints position and skeletons length. When the number of skeletons in the skeleton sequence that do not meet the above limitations exceeds the set value, we will directly regard these sequences as invalid sequences, skip the intention recognition network, drive the robot to make a voice prompt and start a new round of interaction.

C. Intention recognition network

In natural HRI, interactive intention is often abstract motion information (such as movement trend, direction) or command (gesture). Traditional intention prediction methods mainly rely on Markov model and Bayes' theorem, but we do not use the traditional probability model. Aiming at the limited category of adversarial interaction, we transform the intention prediction in a short time into the intention recognition of the current action. In this particular scenario, we build the corresponding dataset and model the relationship between human joint data and human intentions.

Because human intention to obtain directly from the image will be affected by the image background or light, a very large data set is required. To avoid this practical problem, we simplify the recognition from image to human intention into human joint coordinate sequence to human intention. In addition, a pose can't accurately reflect the human body's intention, so we use continuous multi-frame joins data as the basis for intention recognition. Therefore, we designed an intention recognition network based on the LSTM [19].

The multiple continuous images acquired by the robot are input into the pose feature extraction network. After combining the pose features of these images, they are input into the intention recognition network in this section. The size of input data is $N \times (16 \times 3)$ and N represents the number of pre-merged feature sequences. Each sequence contains the x, y, z three-dimensional features of 16 joints on the axes, and the size of each feature is 16×3 .

Assuming $N = 5$, the input data of intention recognition network is 5 frames pre-merged feature , and feature sequences can be regarded as a time series. The features learned by the intentional recognition network are composed of the hidden state h_{t-1} of the previous frame and the input of the following frame p_t . Where h_{t-1} and O_t are defined as:

$$h_{t-1} = O_{t-1} * \tanh(f_{t-1}) \quad (6)$$

$$O_t = \sigma(W_0[h_{t-1}, p_t] + b_0) \quad (7)$$

W_0 and b_0 denote output weights and bias, σ denotes the Sigmoid activation function, O_{t-1} denotes the output at time $t-1$, and f_{t-1} denotes the combination of old and new features at time $t-1$.

The human feature p_{t-1} of each frame is transformed into all neurons at time $t-1$, and the human feature p_t of the next frame is used as the input of time t . The final output $O_t = [o_0, o_1, ..., o_C]^T$ contains all the important features of the previous time, where C denotes the number of intention categories.

After two layers LSTM network is a Softmax function, which can transform to a probability $Y_{t,c}$ corresponding to the c-th class of the intention:

$$Y_{t,c} = \frac{\exp(O_t, c)}{\sum_{k=1}^{C} \exp(O_{t,k})} \quad (8)$$

where $k = 1, 2, ..., C$ and $O_{t,k}$ denotes the encoding of the confidence score on the c-th intent class. Finally, we set $Y_t = [y_1, y_2, ..., y_C]^T$ as the predicted class label vector.

IV. Experiments

In this part, we first introduced our experimental setup and experimental details. Finally, we designed two groups of experiments for five different interaction categories. One group compares the existing interaction framework and the other group verifies the adaptability of the framework of different conditions.

A. Experimental configuration

The implementation of our framework relies on Pytorch framework and uses Python language to call Naoqi interface to communicate with the robot. The pythons used in this experiment are Python 2.7 and python 3.6. Python 2.7 is used for interaction, and python 3.6 is used for posture feature extraction and human intention recognition. The experimental hardware is NVIDIA GeForce GTX-1080Ti GPU with a memory size of 11GB. In the HRI contrast experiment, the existing interactive framework uses external Kinect depth camera to simulate with Nao robot (external camera and adapter are needed).

B. Training details

The existing human action recognition datasets rely on deep camera to extract and have no clear interactive intention, and do not apply to the specific group of human and robot. Therefore, we establish the antithesis interactive intention prediction data set applicable to humans and robots.

TABLE I
HUMAN-ROBOT INTERACTIVE MECHANISM.

	Interaction actions	
	Human actions	***Robot actions***
1	Right punching	Protecting the belly
2	Pushing forward	Standing firm
3	Bowing	Standing upright with lift up
4	Right arm swing	Blocking with left arm
5	Right hand shooting forward	Protecting the head

Based on the common interaction, the human-robot interactive mechanism is designed as shown in Tab I. In order to establish human intention recognition network, we build a dataset for training intention recognition network, which including 5 males and 6 females and 41580 groups of joints sequence. Then the images of each action were organized to expand the dataset. The dataset contains a total five types of interactions, and each interaction type includes 8316 joints sequences.

In our experiment, the human pose estimation stage is based on the open method updated by Xingyi Zhou [20] on github.com, which is proved to be suitable for extracting three-dimensional joints information of a single human of a single human from the RGB image.

In the training of intention identification network, we randomly divided into 33264 training joints sequences and 8316 testing joints sequences, and the learning rate is 0.01. In order to prevent fast decay, we choose cross entropy as loss function and Adam algorithm as weight update function. We printed the accuracy of the intention recognition network within 20 generations of training, and found that the accuracy of the network was infinitely close to 100% and its performance was very stable. In the experiment, it is found that when there is unfamiliar new data input, it can still make accurate identification.

C. Experimental results

First of all, we compare with the current interaction framework at different distances, and randomly select an interaction for analysis. The video of complete comparison experiment can be found here: Youku video1.

At a relatively close distance, our framework extracts the same human skeleton information as the current interaction framework. Through the intention recognition network, the robot can correctly recognize the human intention and perform the same interaction, as shown in Fig. 4.

At a relatively long distance, the current interaction framework fails to extract the effective human skeleton due to the limitation of the working range of depth camera, but our framework can still extract the effective human skeleton and complete the expected interaction, as shown in Fig. 5.

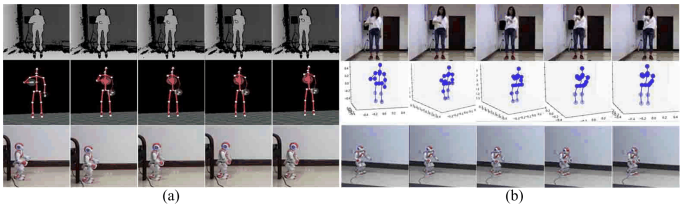

Fig. 4. (a) and (b) are short-distance interactive screenshots of the current frame and our framework respectively. The first line are depth and RGB images, the second line are skeletons extracted, and the third line are a five-frame of the robot's response action.

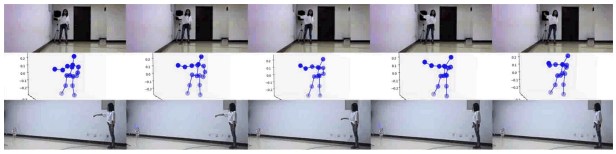

Fig. 5. Long-distance interaction with our framework.

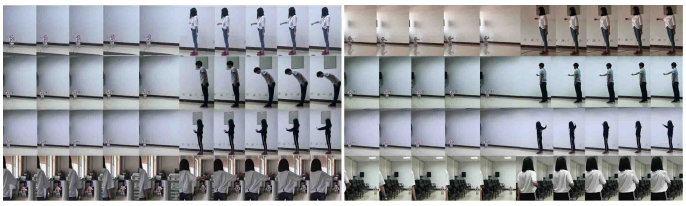

Fig. 6. Partial interactive screenshots of different scenario. (illumination, customer, costume, distance, distance, scene).

Fig. 6 shows that the robot extracts effective joints from images at different distances, lighting, clothing, customers and scene conditions and completes effective interactions through the human intention recognition network. All experimental results can be found here: Youku video2.

The experiments show that our framework can be applied to different interaction scenarios, and has certain robustness to distance, light intensity, clothing brightness, human height and background complexity, effectively reduce the limitation inside 4m and has universality in different scenes.

After a lot of experiments, we found that when the human pose estimation network output joints position are greatly different from the real position or the robot acquired image contains incomplete human body as shown in Fig. 7 (b), the joints sequences we obtained often make the intent recognition network get a wrong result. When received as shown in Fig. 7(a) the human points sequence, the intention recognition network outputs the correct human intention.

By adding the human joints sequence filter, our framework is able to reduce the impact of these unreasonable sequences. In the final experiments, we found that not all the human joints sequence can enter intention recognition network. If the human joints information extracted from the human pose estimation network is unreasonable, our framework will directly treat

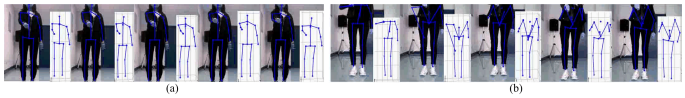

Fig. 7. (a) is a joints sequence that passes the human joints sequence filter successfully, and (b) is a joints sequence that fail to passes the human joints sequence filter. The sequences that pass successfully will be sent into the intention recognition network, and the fail sequences will be judged as invalid intention.

them as invalid sequences, and skip intention recognition network to drive the robot to make voice prompts and start a new round of interaction.

The above experiments show that compared with the framework based on depth camera, our framework not only reduces the constraints of distance and SDK, but also can complete the desired interactive actions of robots according to human body images under different conditions.

V. Conclusions and Future Work

In this paper, we introduce the human pose estimation method and build an intention recognition model, take effective human joints information as the intermediate data, improve the independence between the modules of the interaction framework, and reduce the limitations of the existing framework relies on the depth camera. The experimental results show that our framework can be implemented under the condition of different effective interaction. In addition, the framework can be applied to a wider range of interactions and robots. It is worth mentioning that the human pose estimation network introduced in the experiment is only effective for the detection of single people. Therefore, when the captured image contains multiple people during the interaction, the more important human joints information (closer to the center or with a larger background color difference) will be extracted for interaction. In addition, since we modeled for a fixed class of interactions, the interaction mechanism here has only a single interaction for actions of the same type and different magnitude. Therefore, our next work will be to consider the accurate interaction in the case of multiple people, and make the robot make corresponding response actions at different motion ranges, so as to further empower the robot with stronger intelligence.

Acknowledgment

This work was supported in part by the National Science Fund for Distinguished Young Scholars (No.61425002), the National Natural Science Foundation of China (Nos. 91748104, 61632006, 61877008, 61603066), Program for the Liaoning Distinguished Professor, Program for Dalian High-level Talent Innovation Support (No.2017RD11), the Scientific Research fund of Liaoning Provincial Education Department (No.L2019606), and the Science and Technology Innovation Fund of Dalian(No.2018J12GX036), the Liaoning Province Doctor Startup Fund(No.201601302).

References

[1] F.B. Mandal, Nonverbal Communication in Humans. Journal of Human Behaviour in the Social Environment, 2014, 24(4):417-421.

[2] L. Wang, R. X. Gao, J. Váncza, J. Krüger, X. V. Wang, S. Makris and G. Chryssolouris, "Symbiotic Human-Robot Collaborative Assembly," CIRP Annals – Manufacturing Technology, 2019, Vol.68, No.2, pp.701-726.

[3] P. Fankhauser, B. Michael, R. Diego, K. Ralf, H. Marco, Y. S. Roland, Kinect v2 for mobile robot navigation: Evaluation and modelling. In: 2015 International Conference on Advanced Robotics (ICAR). Istanbul, 2015, pp. 388-394.

[4] De Carli, D., Hohert, E., Parker, C. A., Zoghbi, S., Leonard, S., Croft, E., Bicchi, A, Measuring intent in human-robot cooperative manipulation. In: 2009 IEEE International Workshop on Haptic Audio visual Environments and Games. Lecco, 2009, pp. 159-163.

[5] C. Zhu, W. Sun, W. Sheng, Wearable sensors based human intention recognition in smart assisted living systems. In: 2008 International Conference on Information and Automation. Changsha, 2008, pp. 954-959.

[6] Z. Wang, K. Mülling, M.P. Deisenroth, H. Ben Amor, D. Vogt, B. Scholkopf, et al, Probabilistic movement modeling for intention inference in human–robot interaction. The International Journal of Robotics Research, 2013, 32(7): 841-858.

[7] J. Elfring, R. Van De Molengraft, M. Steinbuch, Learning intentions for improved human motion prediction. Robotics and Autonomous Systems, 2014, 62(4): 591-602.

[8] T. Petković, I. Marković, I. Petrović, Human Intention Recognition in Flexible Robotized Warehouses Based on Markov Decision Processes. In: Iberian Robotics conference. Seville, 2017, pp. 629-640.

[9] T. Liu, J. Wang, M. Q. H. Meng, Human robot cooperation based on human intention inference. In: IEEE International Conference on Robotics & Biomimetics. Zhuhai, 2017.

[10] X. Yu, W. He, Y. Li, C. Xue, J. Li, J. Zou, C. Yang, Bayesian estimation of human impedance and motion intention for human-robot collaboration. IEEE transactions on cybernetics. 2019.

[11] M. Sigalas, H. Baltzakis, P. Trahanias, Gesture recognition based on arm tracking for human-robot interaction. 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, 2010, pp. 5424-5429.

[12] Li X. Human–robot interaction based on gesture and movement recognition. Signal Processing: Image Communication, 2020, 81:115686.

[13] Luo X, Amighetti A, Zhang D. A Human-Robot Interaction for a Mecanum Wheeled Mobile Robot with Real-Time 3D Two-Hand Gesture Recognition. Journal of Physics: Conference Series, 2019, 1267:012056.

[14] H.S. Koppula, A. Saxena, Anticipating human activities using object affordances for reactive robotic response. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(1): 14-29.

[15] G. Canal, S. Escalera, C. Angulo, A real-time human-robot interaction system based on gestures for assistive scenarios. Computer Vision and Image Understanding, 2017, 149: 65-77.

[16] K. Li, J. Wu, X. Zhao, M. Tan, Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-mimicking Decision Mechanism. In: 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). Tianjin, 2018, pp. 498-503.

[17] O. Mazhar, B. Navarro, S. Ramdani, R. Passama, A. Cherubini, A real-time human-robot interaction framework with robust background invariant hand gesture detection. Robotics and Computer-Integrated Manufacturing, 2019, 60: 34-48.

[18] Chinese Academy of Standardization and Information Classified Coding. GB10000-88 Human dimensions of Chinese adults[S]. Beijing: Standard Press of China, 1988.

[19] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, 2017, pp. 2625-2634.

[20] X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, Towards 3d human pose estimation in the wild: a weakly-supervised approach. In:Proceedings of the IEEE International Conference on Computer Vision. Honolulu, 2017, pp. 398-407.