
Hand gesture recognition using multimodal data fusion and multi-scale parallel convolutional neural network for human-robot interaction

Qing Gao^{1,2} | Jinguo Liu^{1*} | Zhaojie Ju^{1,3}

¹State Key Laboratory of Robotics, Shenyang Institute of Automation, Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110016, China

²University of Chinese Academy of Sciences, Beijing, 100049, China

³School of Computing, University of Portsmouth, PO1 3HE, Portsmouth

Correspondence

Jinguo Liu, State Key Laboratory of Robotics, Shenyang Institute of Automation, Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110016, China
Email: liujinguo@sia.cn

Present address

[†]Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China

Funding information

Funder One, National Key R&D Program of China, Grant No. 2018YFB1304600; Funder Two, Natural Science Foundation of China, Grant No. 51775541 and 51575412; Funder Three, CAS Interdisciplinary Innovation Team, Grant No. JCTD-2018-11; Funder Four, the EU Seventh Framework Programme (FP7)-ICT, Grant No. 611391.

Hand gesture recognition plays an important role in human-robot interaction (HRI). The accuracy and reliability of hand gesture recognition are the keys to gesture-based HRI tasks. To solve this problem, a method based on multi-modal data fusion and multi-scale parallel convolutional neural network (CNN) is proposed in this paper to improve the accuracy and reliability of hand gesture recognition. First of all, data fusion is conducted on the sEMG signal, the RGB image, and the depth image of hand gestures. Then, the fused images are generated to two different scale images by downsampling, which are respectively input into two sub-networks of the parallel CNN to obtain two hand gesture recognition results. After that, hand gesture recognition results of the parallel CNN are combined to obtain the final hand gesture recognition result. Finally, experiments are carried out on a self-made database containing 10 common hand gestures, which verify the effectiveness and superiority of the proposed method for hand gesture recognition. In addition, the proposed method is applied to a 7-degree-of-freedom bionic manipulator to achieve robotic manipulation with hand gestures.

KEYWORDS

hand gesture recognition, multimodal data fusion, parallel CNN, sEMG signal

1 | INTRODUCTION

In gesture-based space HRI: Malima et al. (2006); Raheja et al. (2010), reliability and security are the key to ensuring the normal operation of HRI: Liu et al. (2016b). The traditional methods of acquiring hand gesture information mainly include collecting RGB images of hand gestures with a color camera, collecting depth information of hand gestures with a depth sensor, and collecting electromyography information of hand gestures with a sEMG device. Each of these methods has its advantages and disadvantages. For example, the RGB image of the hand gesture has rich performance features, but it cannot show the 3D information of the hand gesture. The depth image of the hand gesture contains 3D features, but it lacks sufficient performance features. Moreover, the RGB image and the depth image cannot be recognized at the most of time in the case where the hand is severely blocked. While the use of SEMG device does not need to consider the occlusion problem of gestures, but the noise and interference of the device are large, and it is often impossible to obtain a high hand gesture recognition accuracy. Various information about the hand gesture can be utilized and the recognition accuracy of the hand gesture can be improved by fusing multi-modal gesture information. For example, the reference: Chen et al. (2015) combines depth camera data and inertial sensor data for the recognition of 27 human body movements, which improves the recognition accuracy. Reference: Miao et al. (2017) combines RGB, depth and optical flow information to identify dynamic gestures of the human upper bodies. The reference: Kopuklu et al. (2018) combines the color map and the optical flow graph of dynamic hand gestures, and achieves good results on the Jester, ChaLearn LAP IsoGD and NVIDIA Dynamic Hand Gesture Datasets. For the way of multimodal data fusion, according to the stage of fusion, it is mainly divided into data level fusion: Kopuklu et al. (2018), feature level fusion: Miao et al. (2017), decision level fusion: Simonyan and Zisserman (2014). Among them, data level fusion can achieve the highest fusion efficiency. It has two advantages: (1) training only requires a single channel network; (2) automatically establish pixel-wise correspondence between different modalities. Built on the above analysis, the data level fusion of RGB information, depth information and sEMG information is proposed to improve the recognition accuracy of hand gestures in this paper.

At present, twork has achieved great achievement in the field of image recognition: Gao et al. (2017a). In addition, the use of multi-scale input in parallel networks can also improve the recognition accuracy of images effectively. For example, the reference: Karpathy et al. (2014) proves that better experimental results can be obtained by a dual-stream CNN with raw and spatially clipped video streams. In reference: Molchanov et al. (2015), a parallel 3DCNN is designed, and the original data and the data after down-sampling are input into two parallel sub-networks respectively, which realizes the improvement of the dynamic hand gesture recognition accuracy. Take into account these, a parallel CNN structure is proposed. And the fused data and the downsampled data are taken as input to the two parallel sub-networks. Finally, the output results are brought together to obtain the final hand gesture recognition accuracy.

The contributions of this paper are concluded as follows.

- a The RGB image, the depth image and the SEMG signal of hand gesture are combined to deal with hand gesture recognition in the case of hand occlusion and to improve the reliability and safety of gesture-based HRI.
- b A data-level fusion method is designed to convert the SEMG signal into an image and then fuse it with the RGB image and the depth image.
- c A multi-scale parallel convolutional neural network (MPN) framework is designed to improve the recognition accuracy of hand gestures.
- d A set of hand gesture database containing 10 common HRI hand gestures is made. This database contains RGB images, depth images and sEMG information, which can be used for verification of the proposed method.
- e The proposed method is applied to the control of a 7-degree-of-freedom bionic manipulator to realize gesture-based

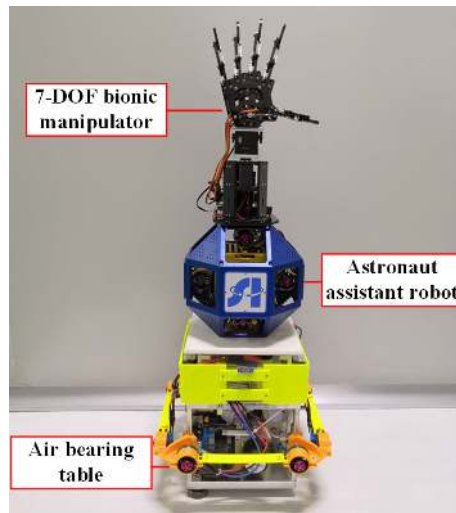


FIGURE 1 AAR platform

space HRI.

The rest of this paper is structured as follows. Chapter 2 introduces the gesture-based space HRI system. Chapter 3 presents the data fusion method. Chapter 4 introduces the MPN framework. Chapter 5 is the experimental results and discussion, and the conclusion remark and future work are adopted in Chapter 6.

2 | GESTURE-BASED SPACE HRI SYSTEM

Security and reliability play important roles in space HRI tasks. This chapter is aimed at a bionic manipulator on an astronaut assistant robot (AAR): Gao et al. (2017b); Liu et al. (2016a), using hand gestures to control and operate it.

2.1 | Space robot

In the space station, due to the heavy workload and the limited number of astronauts, the astronaut assistant robot has an obligation to assist the astronauts in completing some space missions. For example, it can assist astronauts in conducting space experiments, helping astronauts take some tools, and manage the safety of astronauts. Therefore, we design an AAR for using in the space station cabin. As shown in FIGURE 1, its platform consists of an AAR, a bionic manipulator and a simulated air bearing table.

AAR This space robot is capable of free flight in the cabin and is equipped with 12 ducted fans as its drives for six-degree-of-freedom motion in microgravity environments.

Air bearing table In the ground experiment, the simulated air bearing table can help the AAR to realize simulated micro-gravity movement in the horizontal direction.

Bionic manipulator A Bionic manipulator is mounted on the AAR and it can be utilized to grab some tools or objects. Its structure consists of fingers, palm and wrist with seven degrees of freedom (5 degrees of freedom in fingers

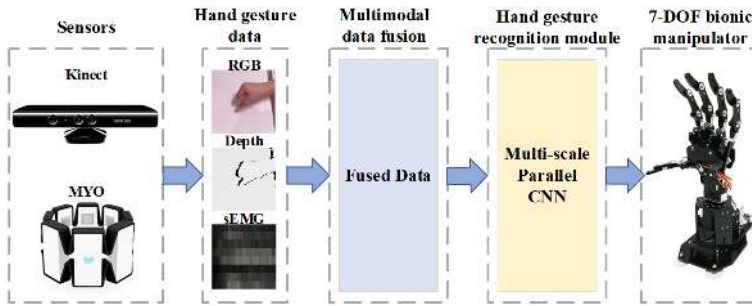


FIGURE 2 The pipeline of HRI method

and 2 degrees of freedom in the wrist). It is extremely important to control the motion of the bionic manipulator. Traditional control methods, such as handles, joysticks, and consoles, are complicated and inconvenient to operate. Since the structure of the manipulator is similar to a human hand, it is very convenient to directly control it with hand gestures.

2.2 | Gesture-based HRI











Hand gesture recognition technology is very important in gesture-based HRI: Gao et al. (2019). Current hand gesture recognition methods are mainly based on wearable devices and vision: Rautaray and Agrawal (2015); Smith et al. (2000). Both methods have their pros and cons. For example, wearable-based hand gesture recognition is limited by the device, and the interference is large. While vision-based hand gesture recognition is susceptible to occlusion. Security and reliability are the priority in space missions. Therefore, in this paper, these two methods are combined to improve the recognition accuracy of hand gestures and to cope with various interference, such as occlusion and signal interference.

For the acquisition of hand gesture signals, we use the MYO armband to collect the sEMG signals of hand gestures: Benalcázar et al. (2017). The sensors on the MYO band capture the bioelectrical changes that occur when the user's arm muscles move, thus can judge the wearer's hand gestures, and then send the recognition results to the robot via Bluetooth. And use the Kinect to collect the RGB and Depth images of hand gestures: Ren et al. (2013). Kinect is a 3D sensor that includes a color camera and a TOF depth camera. It can capture hand gesture movements in 3D space in real time. Then, combine the three kinds of information and transmit them to the hand gesture recognition model based on deep neural network to identify the corresponding hand gestures. After that, the recognized result is transmitted to the AAR. Thereby enabling the human hand to control the bionic manipulator so that it can simulate the human hand gestures. The specific method flow chart is shown in FIGURE. 2.

2.3 | HRI hand gesture dataset

Since the bionic manipulator has 7 degrees of freedom, it can simulate the movement of human fingers and wrist, 10 common static hand gestures are designed, including the gesture of the fingers or the gesture of fingers and wrist. These hand gestures and their semantics are shown in TABLE 1. Among these hand gestures, hand gesture 1 and hand gesture 2 can control the initial action and stop action of the manipulator. Hand gestures 3-6 can control the movement direction of the manipulator, that is, left, right, upward and downward. The manipulator can learn different ways of grasping from hand gesture 7-10, which help to grasp different objects. By simulating these hand gestures, the manipulator can

TABLE 1 HRI hand gesture dataset

Hand gesture number	Hand gesture semantic	Hand gesture diagram
Hand gesture 1	Relax	
Hand gesture 2	Fist	
Hand gesture 3	Left	
Hand gesture 4	Right	
Hand gesture 5	Downward	
Hand gesture 6	Upward	
Hand gesture 7	Grab a cylinder	
Hand gesture 8	Grab a ball	
Hand gesture 9	Pinch	
Hand gesture 10	Buckle	

perform some conventional directional motion operations and grab operations.

2.4 | Stability analysis of HRI

For a complete gesture-based HRI system, the stability of the system should be also considered. When astronaut hand gestures incorrect or at the transition process between different hand gestures, or when the system just starts and the hand gesture changes suddenly, the HRI system maybe unstable. Therefore, it is necessary in order to design a method to map astronaut's hand gestures to the bionic manipulator, to filter out unstable hand gesture information, and to use stable hand gesture output for controlling the manipulator.

Finite State Machine (FSM) is a mathematical model employed to represent finite states and transitions and actions between these states. It is primarily used for the parsing of programming languages. The gesture-based HRI can also be taken into account as a language form that expresses semantic commands through hand gestures. Therefore, it is very suitable to use FSM to model the semantics of different hand gestures.

For each predefined hand gesture, a FSM model needs to be created. As shown in FIGURE 3, in the process of interaction between the astronaut and the bionic manipulator, firstly, each frame data containing a hand gesture is collected by Kinect and MYO. Secondly, the hand gesture recognition algorithm is utilized to recognize the type of the hand gesture, and then the processed hand gesture information is input to the corresponding FSM model. This can output the predefined hand gesture.

Taking hand gesture 1 as an example, the state transition relationship of the hand gesture model is established by

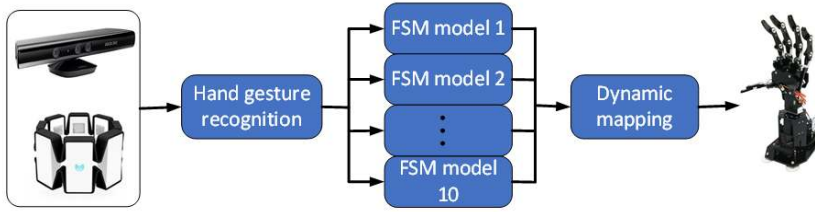


FIGURE 3 Hand gesture interaction flow chart

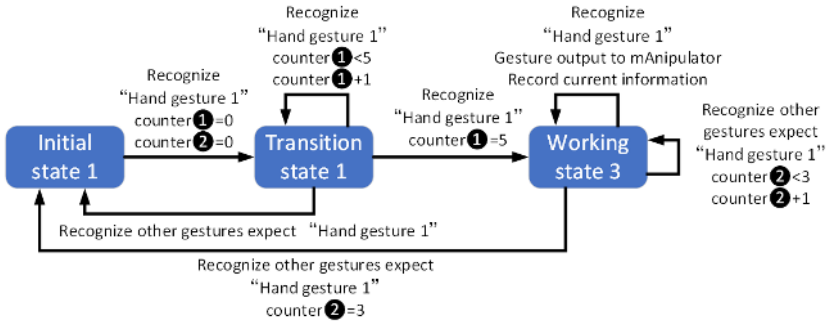


FIGURE 4 FSM model diagram of "Hand gesture 1"

using the FSM method, and it is illustrated in FIGURE 4. The implementation principle of hand gesture 1 FSM model is shown as follows: the initial time of the system is in state 1. At this time the bionic manipulator is not controlled. When the astronaut enters a hand gesture and the system recognizes that the hand gesture is "Hand gesture 1", it transitions to transition state 2 and clears counter 1 and counter 2. At this time, the bionic manipulator has not yet been controlled. If a hand gesture other than "Hand gesture 1" is recognized in this state, it returns to the initial state 1. Keep the "Hand gesture 1" for more than 5 consecutive frames, then enter the working state 3. If the "Hand gesture 1" is recognized in this state, the hand gesture is output to control the bionic manipulator.

3 | DATA FUSION METHOD

Data fusion method plays an important role in multi-modal hand gesture recognition tasks: EL-SAYED (2015); Liu et al. (2014). According to the order of data fusion, its methods can be subdivided into data level fusion, feature level fusion and decision level fusion Kopuklu et al. (2018). In this paper, data level fusion is utilized to fuse the RGB, depth and sEMG signals of hand gestures. Compared with feature level fusion and decision level fusion, data level fusion has the following advantages: (a) The fused data can be extracted by a single-channel deep neural network, which can effectively reduce the number of parameters and improve the speed of the algorithm. (b) Achieve pixel-wise correspondence between multi-modal data to improve data fusion efficiency. This chapter introduces the data fusion method of RGB, depth and sEMG signals of the mentioned 10 hand gestures.

3.1 | Data correspondence

Fusion of multimodal data needs to maintain the correspondence of the fused data structure and sampling time. The RGB and depth images acquired by Kinect is 30 fps with a resolution of 640480. The sEMG signal collected by MYO armband is 16 channels and the frequency is 1000Hz: Boyali et al. (2015). The data of the images and the sEMG signal are different in both spatial structure and sampling frequency, so, they cannot be directly fused. Therefore, it is necessary to convert these three kinds of data into a consistent structure and then fuse them. The process of conversion is shown as follows:

3.1.1 | Convert sEMG signals into images

Data collected by the MYO armband in each second is filtered to obtain a matrix M which indicates the strength of the myoelectric signal. The size of M is 161000. Convert it to a grayscale image with a pixel size of 161000 using equation (1).

$$s_{x,y} = 255 \times \frac{m(x,y)}{m_{max}(x,y) - m_{min}(x,y)} \quad (1)$$

where $s(x,y)$ is the pixel value of the coordinate (x,y) in the image S . $s(x,y) \in [0,255]$, $x \in [1,1000]$, $y \in [1,16]$. $m(x,y)$ is the value of the sEMG signal with the coordinate (x,y) in the matrix M . $m(x,y) \in [m_{min}(x,y), m_{max}(x,y)]$, $m_{min}(x,y)$ and $m_{max}(x,y)$ are the minimum value and maximum value in the matrix M .

3.1.2 | Cut the image S

The image S is cut into several of small images with a pixel size of 16×16 . Then, 10 images are uniformly extracted from them, and they are converted into images with a size of 160×160 by upsampling. That is, 10 frames of grayscale images with a pixel size of 160×160 are obtained from the MYO armband per second.

3.1.3 | Sampling and cutting hand gesture images in RGB and depth images

Taking the RGB images as an example, 10 frames of images are uniformly sampled in 30 frames of images acquired by the Kinect per second. These 10 images are cut into images containing hand gestures with a pixel size of 160×160 . That is, 10 RGB images with a pixel size of 160×160 can be obtained every second from Kinect's color camera. The process of depth images is the same as that of RGB images, 10 depth images with a pixel size of 160×160 can be obtained every second from Kinect's depth camera.

3.1.4 | Guarantee the time consistency of the acquired signals

After the signals are handled as mentioned above, these three types of data are collected every 100ms. As shown in FIGURE 5, the hand gesture sEMG image, RGB image, and depth image acquired at the same time t are S_t , R_t , and D_t , respectively. Where

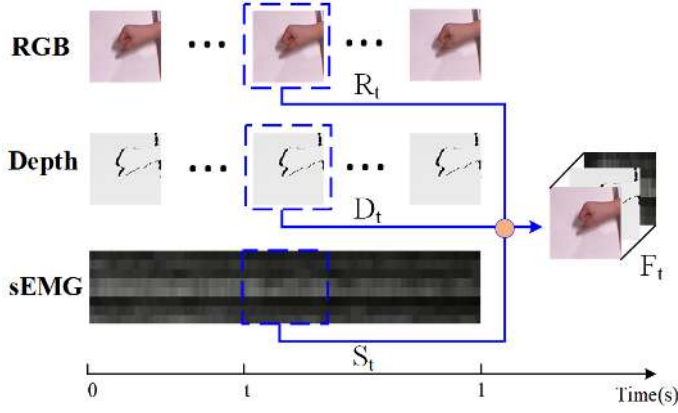


FIGURE 5 Data fusion process

$$\begin{cases} S_t \in \mathbb{R}^{w \times h \times c_s} \\ R_t \in \mathbb{R}^{w \times h \times c_{rgb}} \\ D_t \in \mathbb{R}^{w \times h \times c_d} \end{cases} \quad (2)$$

where $c_s = 1$ is the channel number of the sEMG image, $c_{rgb} = 3$ is the channel number of the RGB image and $c_d = 1$ is the channel number of the depth image.

3.2 | Data Fusion

The way of data fusion is indicated in Fig. 3. The depth image D_t and the sEMG image S_t are sequentially attached to the RGB image R_t as additional channels. The equation is as follows:

$$\Lambda : (\mathbb{R}^{w \times h \times c_{rgb}}, \mathbb{R}^{w \times h \times c_d}, \mathbb{R}^{w \times h \times c_s}) \rightarrow \mathbb{R}^{w \times h \times c_f} \quad (3)$$

$$\text{where } F_t = \Lambda(R_t, D_t, S_t) \quad (4)$$

$$c_f = c_{rgb} + c_d + c_s \quad (5)$$

where F_t is the merged image, $c_f = 5$ is the number of channels. The three types of data collected at the same time can be converted into fused data by M . The fused hand gesture image F_t contains (a) performance features contained in RGB channels; (b) 3D space features contained in a depth channel, and (c) myoelectric features contained in sEMG

channel. Finally, F_i is input as the fused data into the deep neural network.

4 | MULTI-SCALE PARALLEL CONVOLUTIONAL NEURAL NETWORK (MPN)

Aiming at feature extraction and recognition of the hand gesture fusion data, we propose a MPN framework. This chapter mainly introduces hand gesture database, network framework and training method.

4.1 | Hand gesture database

Since there is currently no public hand gesture database containing RGB images, depth images, and sEMG signals, we have to create a set of such database containing the above 10 hand gestures. The Kinect sensor is used to capture the RGB and depth images of the hand gestures, the MYO armband is used to capture the sEMG signal of the hand gestures, and the above data processing method is utilized to convert all these three data into images with size 160×160 . Hand gesture data of 6 subjects are collected, and each subject is collected 625 images for each hand gesture. In addition, one of these subjects is collected with object occlusion. Therefore, our hand gesture database contains a total of 112,500 images, of which the number of RGB, depth and sEMG images are all 37,500. Then, by combining these three kinds of images collected at the same time using the above data fusion method, 37500 images with a size of 160×160 and a channel number of 5 can be obtained.

4.2 | Network framework

The proposed MPN framework is based on the following ideas: (1) in the current CNN models for image feature extraction, the Residual connection structure: He et al. (2016) can train deeper networks. It is implemented by using shortcut connections. Its building block is shown in Figure 4 and the residual mapping formula is

$$F(x) = H(x) + x \quad (6)$$

Where x is the unit map, $H(x)$ is the optimal solution near the unit map, and $F(x)$ is the residual map between the unit map and the optimal solution.

The Inception structure Szegedy et al. (2016) can avoid computational explosion and extract features from multiple scales. So, we add these two structures to the MPN. (2) The traditional Inception v4 or Inception-ResNet: Szegedy et al. (2016) models are mainly for large-size (299×299) images. Because of our small data size (160×160), the network structure of the Inception v4 network is redesigned for using in the feature extraction of our hand gesture recognition database. (3) Inspired by the reference: Karpathy et al. (2014), the fusion of image information at different spatial scales can increase the recognition rate of images. Therefore, a parallel deep neural network structure is designed to fuse two kinds of image data with spatial scales of 160×160 and 80×80 . The specific network framework is shown in FIGURE 6. And its function modules are presented in FIGURE 7.

As shown in FIGURE 4, the MPN framework is mainly divided into two channels: a high-resolution network (HRN) and a low-resolution network (LRN). The input to the network is the fused data and downsampled data from the fused data, and the output is a probability vector of the 10 HRI hand gestures. The fused data is downsampled to obtain data with size $80 \times 80 \times 5$, and the two channels process image data of these two spatial scales in parallel. Among them, HRN

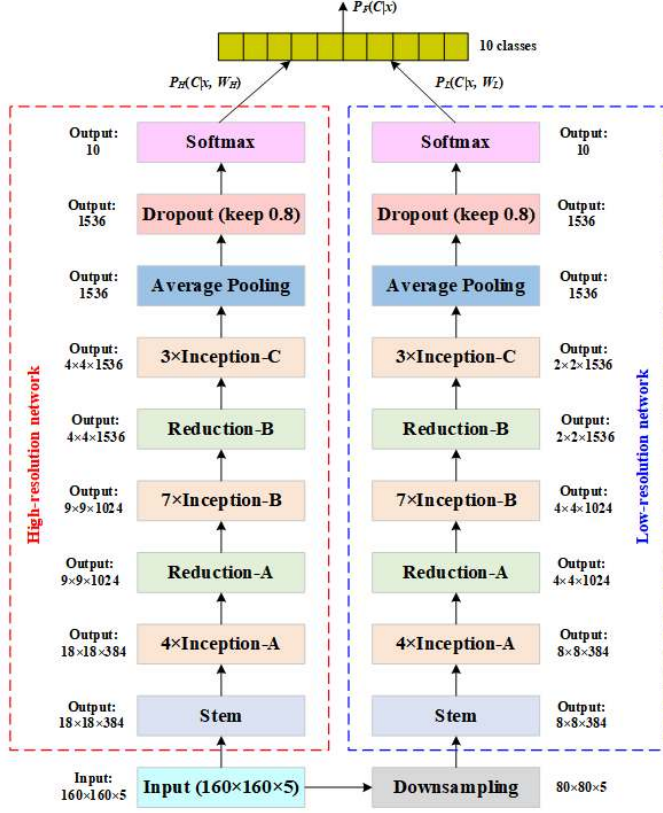


FIGURE 6 MPN framework

mainly extracts and classifies hand gesture data with a spatial size of 160×160 , and the LRN mainly extracts and classifies hand gesture data with a spatial size of 80×80 . The structures of Stem module, Inception module, and Reduction module in the MPN are the same as the structures of the corresponding modules in the Inception-v4. $P_H(C | x, W_H)$ is the classification result of HRN, and $P_L(C | x, W_L)$ is the classification result of LRN. Then, the two results are fused to obtain the final hand gesture classification result $P_F(C | x)$. The fusion process uses the element-wise method. Its equation is:

$$P_F(C | x) = P_H(C | x, W_H) * P_L(C | x, W_L) \quad (7)$$

where $P_F(C | x), P_H(C | x, W_H), P_L(C | x, W_L) \in R^{10 \times 1}$. The predicted label c_h^* chooses the maximum value of the vector $P_F(C | x)$, its equation is:

$$c_h^* = \operatorname{argmax} P_F(C | x) \quad (8)$$

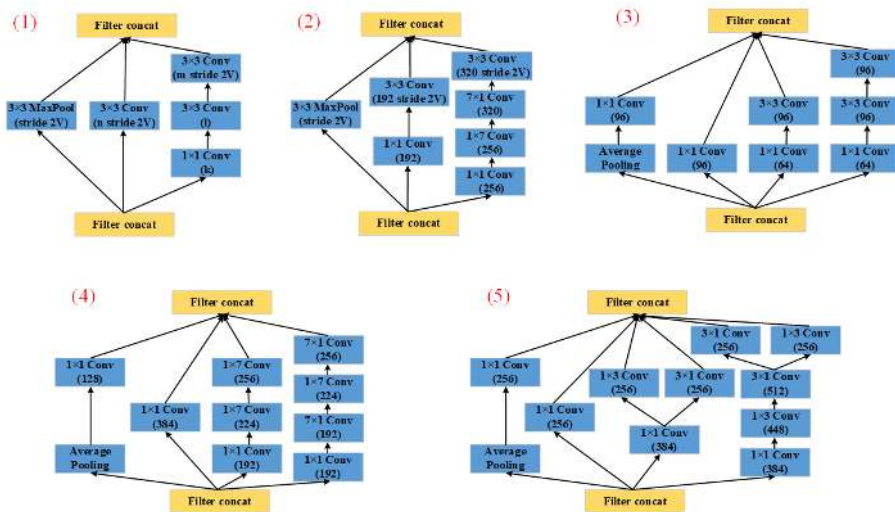


FIGURE 7 Function modules. (1) Reduction-A (2) Reduction-B (3) Inception-A (4) Inception-B (5) Inception-C

4.3 | Train

Select negative log-likelihood: Norouzi et al. (2016) as the loss function. Its equation is:

$$L(W, D_H) = -\frac{1}{D_H} \sum_{i=0}^{|D_H|} \log(P_F(C^{(i)} | x^{(i)}, W)) \quad (9)$$

where D_H is the hand gesture database, and W is the weight function.

The training process uses the steep gradient descent (SGD), and its iterative function is shown as follows:

$$V_{t+1} = \mu V_t - \alpha \nabla L(W_t) \quad (10)$$

$$W_{t+1} = W_t + V_{t+1} \quad (11)$$

where t is the current number of iterations. The weight value W_{t+1} of the $t + 1$ time depends on the weight W_t of the t time and the weight increment V_{t+1} of the $t+1$ time. The value of V_{t+1} is updated by a linear comparison of the last value V_t and negative gradient. α is the learning rate of the gradient, μ is the momentum of the last gradient value. We are required to adjust the values of α and μ to get the best training results.

Adjust the value of the learning rate α by the step method: LeCun et al. (2015). Its equation is:

$$\alpha = \alpha_0 \times \gamma^{(t/s)} \quad (12)$$

TABLE 2 Accuracy and speed

Input data	Accuracy(%)	Speed(ms)	GPU
RGB	55.07	19	GTX1060
Depth	47.26	18	GTX1060
sEMG	75.52	18	GTX1060
RGB+Depth+sEMG	88.89	21	GTX1060

Among them α_0 is the initial learning rate, γ is the adjustment parameter, s represents the iteration length of the adjustment learning rate. That is to say, when the current number of iterations reaches an integral multiple of s , the learning rate is adjusted.

5 | EXPERIMENTAL RESULTS AND DISCUSSION

The proposed data fusion method and MPN method are verified under the above hand gesture database, which proves the feasibility and superiority of the proposed methods.

5.1 | Verification of data fusion method

The data collected by a subject with occlusion in the hand gesture database is used as the verification data, and the data collected by the other five subjects is used as the training data to verify the data fusion method. The RGB images, the depth images, the EMG signal images, and the fused images of the hand gestures are separately trained and verified using the HRN, and the experimental results are compared.

Set the parameters during the training process. The values of the parameters μ and α_0 are mainly based on experience. We set the value of μ to 0.9 and the value of α_0 to 0.001. The value of γ is set to 0.1, the value of the learning rate iteration length s is set to 20000, and the total number of training steps is 60000. That is to say, when the number of training steps is 20,000 and 40,000, the learning rate becomes 0.0001 and 0.00001, respectively.

All experiments are performed in GTX1060, 6GB memory, and the deep learning framework selects Tensorflow.

Through experiments, the average accuracies and time spent of hand gesture recognition using RGB, depth, sEMG data alone and using fused data can be obtained as shown in Tab. 2. The recognition accuracy is calculated based on the ratio of the correct hand gesture images to the total hand gesture images, and the average accuracy is the average of the 10 hand gesture accuracies. The accuracy comparison of the 10 hand gestures corresponding to these four methods is shown in FIGURE 8.

It can be viewed in Tab. 2 that among the experimental results of the four different data, the accuracies of using only RGB and depth data are very low (RGB: 55.07%, Depth: 47.26%). This is because all the hand gesture images in the training process are unoccluded, but some of the hand gesture images in the verified data are partially occluded or totally occluded. Therefore, it can be observed that in the case of occlusion, the effect of using vision for hand gesture recognition is not ideal. Recognition accuracy of using sEMG data alone is 75.52%, which indicates that the sEMG signal is less affected by hand gesture occlusion. However, as can be seen from FIGURE 5, the recognition accuracy of the hand gesture 8 is only 6%, which may be because the sEMG signals of the hand gesture 8 (grip a ball) and the hand gesture 7 (grip a cylinder) are very close, resulting in most of the sEMG images of the hand gesture 8 are recognized as hand gesture 7. While these two hand gestures differ greatly in RGB images and depth images and can be easily distinguished.

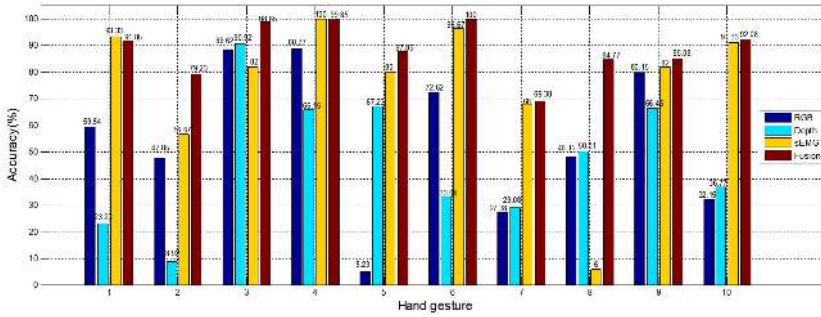


FIGURE 8 The comparison of 10 hand gesture recognition accuracy with four different input data. The blue bar indicates the accuracy of hand gesture recognition obtained by inputting RGB images separately. The cyan bar indicates the accuracy of hand gesture recognition obtained by inputting depth images separately. The yellow bar indicates the accuracy of hand gesture recognition obtained by inputting sEMG images separately. And the red bar indicates the accuracy of hand gesture recognition obtained by inputting fusion images.

TABLE 3 Accuracy and speed

Network	Accuracy(%)	Speed(ms)	GPU
High-resolution network (HRN)	88.89	21	GTX1060
Low-resolution network (LRN)	87.32	16	GTX1060
Multi-scale parallel CNN (MPN)	92.45	32	GTX1060

Recognition accuracy using the fusion data is the highest, reaching 88.89%. And we can see from Fig. 5 that the fusion data has a high recognition accuracy for each hand gesture, wherein the recognition accuracy of hand gesture 7 is the lowest, reaching 69.38%, and the recognition accuracy of hand gesture 6 is the highest, reaching 100%. Therefore, it can be proved that the use of fused data can effectively improve the recognition accuracy of hand gestures. In addition, as can be observed in Tab. 1, the speed of using the fused data is the slowest in the four methods, reaching 21 ms, but it can also achieve high real-time performance.

5.2 | Verification of MPN method

In order to verify the superiority of the proposed MPN method, we use the HRN, LRN and MPN to train and verify the fused hand gesture data. And the results are compared. In order to maintain the fairness of the method comparison, each training parameter is set to be consistent with the above, and the average accuracy and speed of the hand gesture recognition obtained are shown in Tab. 3. The accuracy comparison of the 10 hand gestures corresponding to these three methods is shown in FIGURE 9.

As can be seen from Tab. 3, the MPN has the highest hand gesture recognition accuracy of 92.45%. And we can see from Fig. 6 that the recognition accuracies of all 10 hand gestures obtained by using the MPN are high, where the lowest hand gesture recognition accuracy is 80.85% for the hand gesture 7, and the highest hand gesture recognition accuracy for the hand gesture 4 and the hand gesture 6 are both 100%. It's proposed that the MPN method can effectively improve the hand gesture recognition accuracy. In addition, it can be seen from Tab. 3 that the LRN method is the fastest, reaching 16ms, this is because the parameters of the LRN network are relatively few. The MPN method is the slowest,

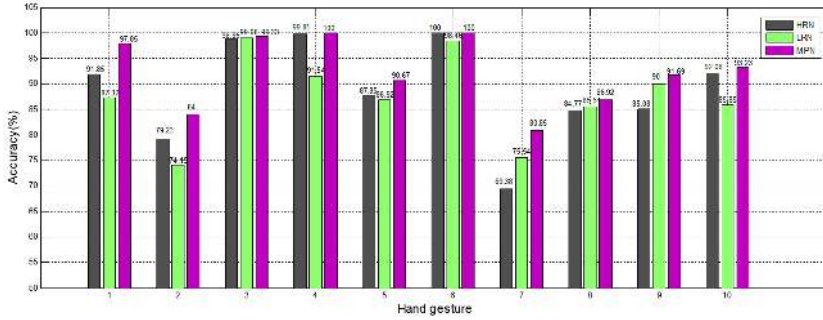


FIGURE 9 The comparison of 10 hand gesture recognition accuracy with three different network frameworks. The gray bar indicates the accuracy of hand gesture recognition using only the HRN. The green bar indicates the accuracy of hand gesture recognition using only the LRN. And the purple bar indicates the accuracy of hand gesture recognition using the MPN.

TABLE 4 Accuracies of the manipulator corresponding to the 10 hand gestures

Hand gesture number	1	2	3	4	5	6	7	8	9	10
Accuracy (%)	96	82	97	98	89	97	78	85	90	91

reaching 32ms, but at this speed, the system can still achieve real-time performance. In a word, it is proved that the MPN method proposed in this paper cannot only effectively improve the recognition accuracy of hand gestures. But also can be applied to the real-time control system of the above-mentioned bionic manipulator.

5.3 | Verification of MPN method

Applying the above MPN method to the recognition of the 10 HRI hand gestures can realize online recognition of these hand gestures. Convert the recognized hand gesture into an action instruction, and then transmit the action instruction to the 7-degree-of-freedom bionic manipulator according to the FSM model proposed above, this can realize the control of the bionic manipulator by the hand gesture operation.

The hand gesture input and the corresponding motion state corresponding to the bionic manipulator are shown in FIGURE 10.

In the real-time system, the movement of the bionic manipulator is controlled by hand gestures, and the above 10 kinds of hand gestures are collected 100 times by Kinect and MYO band, and the manipulator response action is recorded in each time, then the response accuracies of the manipulator corresponding to the 10 hand gestures are obtained and which are shown in Tab.4.

It can be seen from Table 2 that the accuracies of the manipulator operation corresponding to the above 10 HRI hand gestures are more than or equal to 78%. Among them, the accuracy of hand gesture 7 is the lowest (78%) which corresponds to the lowest recognition accuracy of hand gesture 7 in Figure 6. Hand gesture 4 has the highest accuracy of 98%, which corresponds to the highest recognition accuracy of hand gesture 6 in FIG.6. The average accuracy of the manipulator operation is 90.3%, which is 2.15% lower than the hand gesture recognition accuracy of 92.45% by using MPN. This explain that there will be an error of about 2.15% from hand gesture recognition to the manipulator response, which is acceptable in practice. In addition, through experimental records, it is found that almost all erroneous response

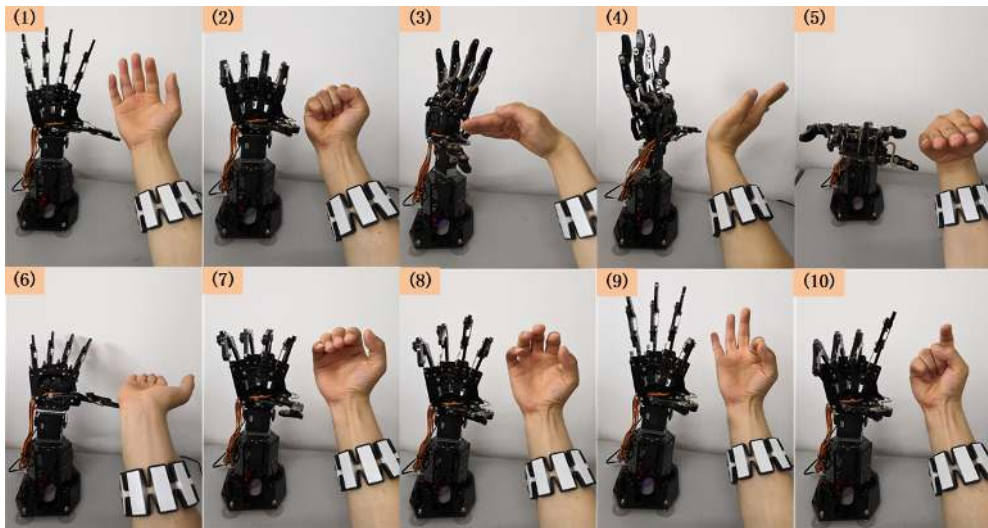


FIGURE 10 Hand gesture input and manipulator motion state diagram. (1) Hand gesture 1 (2) Hand gesture 2 (3) Hand gesture 3 (4) Hand gesture 4 (5) Hand gesture 5 (6) Hand gesture 6 (7) Hand gesture 7 (8) Hand gesture 8 (9) Hand gesture 9 (10) Hand gesture 10.

operations keep the last action, indicating that the proposed FSM method is helpful for stability in the manipulator interaction control. In this case, even if the hand gesture recognition is wrong, as long as the manipulator keeps the motion unchanged, it will not be affected by the operation error. Therefore, this paper proves that the proposed hand gesture recognition method and the HRI method are effective and superior to gesture-based robot control.

6 | CONCLUSION REMARK AND FUTURE WORK

In this paper, focus on the gesture-based HRI for a bionic manipulator on the astronaut assistant robot, a method using data fusion and multi-scale parallel neural network is proposed to improve the recognition accuracy of 10 HRI hand gestures. The contributions and innovations of this paper are summarized as follows: (a) for the control of the seven-degree-of-freedom bionic manipulator, 10 commonly used HRI hand gestures are designed, and a corresponding hand gesture database is made for these 10 hand gestures. The database contains RGB, depth and sEMG data. (b) A data fusion method is proposed to fuse RGB, depth and sEMG signals with different scales to achieve consistency of these three data sizes and sampling time. (c) A multi-scale parallel convolutional neural network framework is proposed to improve the recognition accuracy of hand gestures.

In the next step, our research will be performed on dynamic hand gestures. Since the recognition of dynamic hand gestures is more practical, the data fusion method and recognition are more difficult. In the future, the method proposed in this paper needs to be improved to make it applicable to the recognition task of dynamic hand gestures.

REFERENCES

Benalcázar, M. E., Jaramillo, A. G., Zea, A., Páez, A., Andaluz, V. H. et al. (2017) Hand gesture recognition using machine learning and the myo armband. In *2017 25th European Signal Processing Conference (EUSIPCO)*, 1040–1044. IEEE.

- Boyali, A., Hashimoto, N. and Matsumoto, O. (2015) Hand posture and gesture recognition using myo armband and spectral collaborative representation based classification. In *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*, 200–201. IEEE.
- Chen, C., Jafari, R. and Kehtarnavaz, N. (2015) Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, 168–172. IEEE.
- EL-SAYED, A. (2015) Multi-biometric systems: a state of the art survey and research directions. *IJACSA International Journal of Advanced Computer Science and Applications*, **6**.
- Gao, Q., Liu, J., Ju, Z., Li, Y., Zhang, T. and Zhang, L. (2017a) Static hand gesture recognition with parallel cnns for space human-robot interaction. In *International Conference on Intelligent Robotics and Applications*, 462–473. Springer.
- Gao, Q., Liu, J., Ju, Z. and Zhang, X. (2019) Dual-hand detection for human-robot interaction by a parallel network based on hand detection and body pose estimation. *IEEE Transactions on Industrial Electronics*.
- Gao, Q., Liu, J., Tian, T. and Li, Y. (2017b) Free-flying dynamics and control of an astronaut assistant robot based on fuzzy sliding mode algorithm. *Acta Astronautica*, **138**, 462–474.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.
- Kopuklu, O., Kose, N. and Rigoll, G. (2018) Motion fused frames: Data level fusion strategy for hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2103–2111.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *nature*, **521**, 436.
- Liu, J., Gao, Q., Liu, Z. and Li, Y. (2016a) Attitude control for astronaut assisted robot in the space station. *International Journal of Control, Automation and Systems*, **14**, 1082–1095.
- Liu, J., Luo, Y. and Ju, Z. (2016b) An interactive astronaut-robot system with gesture control. *Computational intelligence and neuroscience*, **2016**.
- Liu, K., Chen, C., Jafari, R. and Kehtarnavaz, N. (2014) Fusion of inertial and depth sensor data for robust hand gesture recognition. *IEEE Sensors Journal*, **14**, 1898–1903.
- Malima, A. K., Özgür, E. and Çetin, M. (2006) A fast algorithm for vision-based hand gesture recognition for robot control.
- Miao, Q., Li, Y., Ouyang, W., Ma, Z., Xu, X., Shi, W. and Cao, X. (2017) Multimodal gesture recognition based on the resc3d network. In *Proceedings of the IEEE International Conference on Computer Vision*, 3047–3055.
- Molchanov, P., Gupta, S., Kim, K. and Kautz, J. (2015) Hand gesture recognition with 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1–7.
- Norouzi, M., Bengio, S., Jaitly, N., Schuster, M., Wu, Y., Schuurmans, D. et al. (2016) Reward augmented maximum likelihood for neural structured prediction. In *Advances In Neural Information Processing Systems*, 1723–1731.
- Raheja, J. L., Shyam, R., Kumar, U. and Prasad, P. B. (2010) Real-time robotic hand control using hand gestures. In *2010 Second International Conference on Machine Learning and Computing*, 12–16. IEEE.
- Rautaray, S. S. and Agrawal, A. (2015) Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, **43**, 1–54.

- Ren, Z., Yuan, J., Meng, J. and Zhang, Z. (2013) Robust part-based hand gesture recognition using kinect sensor. *IEEE transactions on multimedia*, **15**, 1110–1120.
- Simonyan, K. and Zisserman, A. (2014) Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, 568–576.
- Smith, A. V. W., Sutherland, A. I., Lemoine, A. and Mcgrath, S. (2000) Hand gesture recognition system and method. US Patent 6,128,003.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.



QING GAO Qing Gao was born in Tangshan, China. He received his B.S. degree in automation from Electrical Engineering and Automation School, Liaoning Technology University, China in 2013. Currently, he is working toward the Ph.D. degree in the State Key Laboratory of Robotics, Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS), Shenyang, China. His research interests include space robot, artificial intelligence, machine vision and human-robot interaction.

Mr. Gao has authored or co-authored over 10 publications in journals and conference proceedings in above areas and received one outstanding paper award.



JINGUO LIU Jinguo Liu received his Ph.D. degree in robotics from Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS) in 2007. His research interests include modular robot, rescue robot, space robot, and bio-inspired robot. Since January 2011, he has been a Full Professor with SIA, CAS. He also holds the Assistant Director position of State Key Laboratory of Robotics (China) from March 2008.

Prof. Liu has authored and coauthored over eighty papers and thirty patents in above areas. He is the member of IEEE, the Senior Member of China Mechanical Engineering Society, the lead guest editor of international journal of *Advances in Mechanical Engineering*.



ZHAOJIE JU Zhaojie Ju received the B.S. in automatic control and the M.S. in intelligent robotics both from Huazhong University of Science and Technology, China, in 2005 and 2007 respectively, and the Ph.D. degree in intelligent robotics at the University of Portsmouth, UK, in 2010. His research interests include machine intelligence, pattern recognition, and their applications on human motion analysis, human-robot interaction and collaboration, and robot skill learning.

He is currently a Senior Lecturer in the School of Computing, University of Portsmouth. He previously held research appointments at University College London and University of Portsmouth, UK.

Dr. Ju is an Associate Editor of the *IEEE TRANSACTIONS ON CYBERNETICS*. He has authored or co-authored over 100 publications in journals, book chapters, and conference proceedings and received four best paper awards.

--