A Two-stream CNN Framework for American Sign Language Recognition Based on Multimodal Data Fusion

Qing Gao[1,2], Uchenna Emeoha Ogenyi[3] Jinguo Liu[2*], Zhaojie Ju[1,3], and Honghai Liu[3]

[1] State Key Laboratory of Robotics, Shenyang Institute of Automation, Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110016, China,
liujinguo@sia.cn,
[2] University of Chinese Academy of Sciences, Beijing, 100049, China,
[3] School of Computing, University of Portsmouth, PO1 3HE, Portsmouth, UK

Abstract. At present, vision-based hand gesture recognition is very important in human-robot interaction (HRI). This non-contact method enables natural and friendly interaction between people and robots. Aiming at this technology, a two-stream CNN framework (2S-CNN) is proposed to recognize the American sign language (ASL) hand gestures based on multimodal (RGB and depth) data fusion. Firstly, the hand gesture data is enhanced to remove the influence of background and noise. Secondly, hand gesture RGB and depth features are extracted for hand gesture recognition using CNNs on two streams, respectively. Finally, a fusion layer is designed for fusing the recognition results of the two streams. This method utilizes multimodal data to increase the recognition accuracy of the ASL hand gestures. The experiments prove that the recognition accuracy of 2S-CNN can reach 92.08% on ASL fingerspelling database and is higher than that of baseline methods.

Keywords: hand gesture recognition, CNN, Multimodal Data Fusion

1 Introduction

At present, with the development of computer technology and artificial intelligence (AI), the HRI has evolved from robot-centered methods to human-centered methods. The purpose is to achieve more convenient, natural and coordinated interaction between humans and robots, and give full play to the advantages of people and robots to achieve higher work efficiency. The new methods of HRI mainly include hand gestures, voice and EEG [1]. Among them, the vision-based hand gesture interaction is very suitable for HRI because of its intuitive, natural, and non-contact characteristics. So, it has been paid much attention and in-depth research by many scholars.

But the vision-based hand gesture interaction is not yet mature now. It has mainly three difficulties: (1) better and more natural interactive hand gestures;

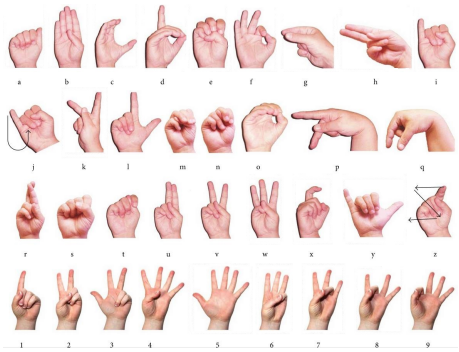

Fig. 1. ASL hand gesture dataset

(2) data processing to remove interference from noise and background; (3) the designs of hand gesture feature extractor and classifier.

The design of hand gesture sets in HRI is crucial. Simple and natural interactive hand gestures make it easier for operators to interact with robots. We chose the ASL hand gesture set [2], which contains 26 different hand gestures, including 24 static hand gestures and 2 dynamic hand gestures. These hand gestures are shown in Fig.1. We can choose several hand gestures from them for HRI and use different alphabetic hand gestures to represent the corresponding interactive hand gestures. It is very convenient for the operator to learn and use.

Illumination changes and background can seriously affect the recognition accuracy of vision-based hand gesture recognition. The hand segmentation can effectively remove the background. For RGB images, skin color-based hand segmentation is often used, but the objects in the background with similar skin colors can seriously affect the effect of the method. The depth image can avoid the influence of illumination changes, and it is more convenient to remove the background. So, the depth images are utilized in this paper.

Traditional hand gesture recognition methods mainly extract the shallow layer features of hand gestures. These features are handcrafted, such as SIFT and HOG [3]. But they cannot get good performance for complex hand gesture recognition. With the development of deep learning, more and more works have focused on applying deep neural networks to hand gesture recognition. Among the deep learning methods, CNN is mainly used for image recognition [4]. Typical CNN are mainly VGG, GoogLeNet, Inception, ResNet and MobileNet. Each of these methods has its own characteristics and combining these methods can improve the efficiency of image recognition.

In this paper, a two-stream CNN framework is designed for the recognition of ASL hand gestures. The contributions are mainly as follows:

– A two-stream CNN framework is proposed. It uses two CNN feature extractors to extract gesture RGB and depth features and a fusion layer to fuse the recognition results. Multimodal data features are utilized to increase the hand gesture recognition accuracy.

– In the data processing part, a simple and convenient data enhancement method is proposed by combining the advantages of RGB and depth images.
– In the result fusion part, a fusion layer is designed and connected to the network. The weights of the fusion can be obtained through training.

The rest of this paper is organized as follows. Related work of vision-based hand gesture recognition is introduced in Chapter 2. The 2S-CNN proposed in this paper is introduced in Chapter 3, which includes network framework, data preparation, data enhancement, CNN feature extractor, fusion method. The experiments and analyzes are shown in Chapter 4. Chapter 5 is conclusion remark and future work.

2 Related work

Traditional vision-based hand gesture recognition algorithms include dynamic time warping (DTW), artificial neural network (ANN) and hidden Markov model (HMM) [5]. But each of them has some shortcomings. With the development of deep learning in the field of image recognition, more and more scholars have applied deep learning methods to the research of vision-based hand gesture recognition and have achieved certain research results. For example, Oyebade K used the CNN to identify 24 ASL hand gestures and achieved a high recognition accuracy. But the gesture database used is too small and the image background is single [6]. Jawad Nagi used a CNN with Max-Pooling layers to classify six hand gestures and used them to control a robot. He wore gloves to segment and classify hand gestures simply. But this method is not universal [7]. Youngwook Kim used DCNN to identify 10 gestures of radar-acquired Micro-Doppler Signatures and achieved a high recognition accuracy. But the disadvantage is that the image background is single [8]. Takayoshi Yamashita designed a deep convolutional neural network with bottom-up structure for hand gesture recognition. However, this method used two-dimensional grayscale images, which may not be suitable for some complicated three-dimensional hand gestures [9]. There are a lot of interference information for hand gesture recognition in images, such as complex background, illumination changes and noise. In-depth researches are still needed to achieve robust hand gesture recognition.

Depth images can effectively reduce the effects of background and illumination changes. At present, with the development of depth sensors like Kinect, many scholars introduce depth images into hand gesture recognition, or fuse depth images and RGB images to achieve a higher recognition accuracy. For example, Qing Gao used a parallel CNNs to fuse depth information with RGB information, which improved the recognition accuracy of the ASL hand gesture [10]. C. Jose L. Flores used two CNN architectures with different amounts of layers and parameters per layer to classify 24 alphabet hand gestures of sign language of Peru (LSP) [11]. Zhenyuan Zhang proposed a HandSense network, which extracts features of RGB and depth hand gesture images through two

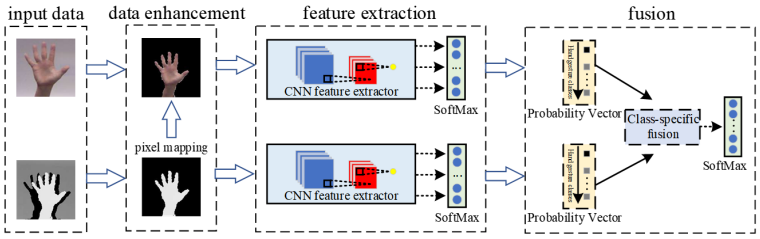

Fig. 2. 2S-CNN framework

parallel 3D-CNNs, and then fused the features and classified hand gestures by a SVM classifier [12]. Although there are many work to study multi-modal information fusion for hand gesture recognition, some work still need further research, such as how to make full use of RGB and depth information, which CNN network is more effective, and how to efficiently fuse information.

The recognition of ASL hand gestures belongs to a fine-grained image recognition task. This paper makes full use of the advantages of hand gesture RGB and depth information by fusing the two kinds of information. Efficient hand gesture segmentation, feature extraction and recognition methods are designed. The proposed 2S-CNN can effectively improve the recognition accuracy of ASL hand gestures.

3 Two-stream CNN framework (2S-CNN)

3.1 Network Framework

The RGB images of hand gestures can express various performance information such as the colors and shapes of the hands. While the depth images of hand gestures can express the spatial information of the hands. So, combining RGB with depth images can take advantage of more hand gesture information to increase hand gesture recognition accuracy. The proposed 2S-CNN adopts a two-stream CNN framework [13]. One channel is used to process the RGB images, and a CNN is used to extract the performance features of the hand gestures. The other channel is used to process the depth image, and another CNN is used to extract the 3D space features of the hand gestures. Finally, the outputs of the two channels are fused using class-specific fusion method to achieve the final prediction of hand gestures. It can get more feature information of hand gestures to achieve a robust prediction and better performance. Its network framework is mainly divided into four parts, which are input data, data enhancement, feature extraction, and fusion. Its framework is shown in Fig. 2.

The main ideas of the 2S-CNN are shown as follows:

– Augment the amount of input data to prevent over-fitting when train the network and improve the generalization ability of the network.

– In the data enhancement part, a simple and efficient hand segmentation method is designed by using the characteristics of the depth images to remove background and noise from the RGB and depth images.
– In the feature extraction part, adopt a more efficient CNN feature extractor.
– In the fusion part, a fusion layer is designed and connected into the network. Then, the fusion weight can be obtained through training.

The details of the method are described below.

3.2 Data Preparationk

In this paper, the ASL fingerspelling database is used as training and testing database. It contains both ASL hand gesture RGB and depth images. There are 24 static hand gestures, representing 24 English letters (except J and Z because they are represented by dynamic hand gestures). Each hand gesture includes 5000 hand gesture images which are 2500 RGB and 2500 depth images collected by 5 objects in different backgrounds. In order to more efficiently segment hand and train the data, the RGB and depth images should be aligned firstly, and then, make the data augmentation.

Image alignment The depth and the RGB images in the ASL database are not aligned, which affects the subsequent hand segmentation operation. So, we need to align the RGB and depth images first. The alignment equations are shown as follows [14]:

$$Z_{rgb} * p^{rgb} = R * Z_d * p_d + T \tag{1}$$

$$R = K_{rgb} * R_{d2rgb} * K_d^{-1} \tag{2}$$

$$T = K_{rgb} * T_{d2rgb} \tag{3}$$

where rgb means RGB image, d means depth image. $Z * p$ represents the mapping relationship of the homogeneous 3D points ($P = [XYZ1]^T$) in the respective camera coordinate systems to the pixel coordinates ($p = [uv1]^T$) on the respective images. K_{rgb} and K_d are internal parameters for color camera and depth camera of Kinect. Set R_{rgb} and T_{rgb} as the external parameters of the color camera, and R_d and T_d as the external parameters of the depth camera. So, the equations of rigid body transformation matrices (R_{d2rgb} and T_{d2rgb}) of the two cameras can be get as follows:

$$R_{d2rgb} = R_{rgb} * R_d^{-1} \tag{4}$$

$$T_{d2rgb} = T_{rgb} - R_{d2rgb} * T_d \tag{5}$$

Data augmentation The increase of training data can help to avoid overfitting and improve the generalization of the network. For data enhancement, we use the following enhancement methods: (a) flip horizontal; (b) translation transformation: pan the image by ±5 pixels; (c) rotation transformation: rotate the image by ±10°.

3.3 Data Enhancement

There is some interference information in the data of ASL fingerspelling database, such as background in RGB images and noise in depth images, which will affect the accuracy of hand gesture recognition. Because the depth images can effectively distinguish hands and background, depth images are utilized to segment hands first, and then segment RGB images by mapping the segmentation pixels. Suppose the pixel with the closest distance value in the depth image appears in the hand area (it's true in ASL fingerspelling database). Therefore, the following equation can be used to segment the depth images.

$$d_{w,h} = \begin{cases} d_{w,h} \; if d_{w,h} \geq d_{min} - \triangle d \\ 0 \quad if d_{w,h} < d_{min} - \triangle d \end{cases} \tag{6}$$

where w and h represent the width and height of the image, respectively. $w, h \in [0, 200]$. $d_{w,h}$ represents the depth value of the image at the coordinate point (w, h). d_{min} is the minimum value in the depth image. $\triangle d$ represents the difference between the depth value and the minimum depth value d_{min}, set $\triangle d = 2$. So, the pixel whose depth value is in the range $[d_{min}, d_{min} - \triangle d]$ defaults to the hand gesture. The remaining pixels are regarded as background and their values are set to 0. Then, the pixel coordinates of the hand gesture in the depth map are mapped onto the color image, and the segmented hand gesture on color image is obtained.

Since the sizes of the ASL fingerspelling database images are different, all image sizes are converted into 299×299 pixels, which can be performed at the input layer of the network.

3.4 CNN feature extractor

Feature extraction of hand gesture images is the key to hand gesture recognition. A more efficient CNN feature extractor can better extract hand gesture features and help to improve the recognition accuracy of hand gestures. In this part, the Inception-ResNet v2 is chosen as the CNN feature extractor by analyzing a variety of typical CNN structures [15]. Inception-ResNet v2 combines the advantages of Inception and ResNet. It can train deeper networks with better feature extraction and avoid overfitting. Its network structure is shown in Fig. 3. where the network structure of the three Inception-resnets in Fig. 3 are shown in Fig. 4.

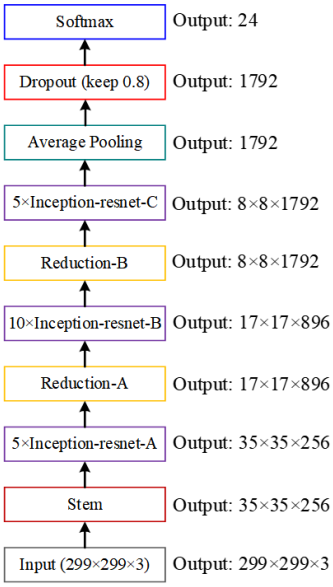

Fig. 3. Inception-ResNet v2 structure

3.5 Fusion method

It can be seen from Fig. 2 that in the proposed 2S-CNN, the hand gesture predicted probability scores can be obtained from the input RGB and depth images through CNN feature extractors and softmax classifiers. After that, the results from the two streams need to be fused. A fusion layer is designed and connected after the network. So, the weight of the fusion can be obtained through training. Firstly, the form of outputs is converted into probability vectors. Set P_1 and P_2 denote the probability vectors obtained from the RGB and depth streams, respectively. Where $P_1, P_2 \in R^{1\times N}$, and N denotes the number of classifications of the ASL hand gestures. Then, set $\omega_i = \omega_1^i, \omega_2^i, \cdots, \omega_N^i (i = 1, 2)$ indicates the corresponding fusion weight. Where ω_j^i represents the $j-$th class fusion weight in the $i-$th stream. Finally, set P_f represents the vector through the fusion of P_1 and P_2, then P_f is expressed as

$$P_f \cdot \omega = P_1 \odot \omega_1 + P_2 \odot \omega_2 = [P_1^1, \cdots, P_1^N, P_2^1, \cdots, P_2^N] \begin{bmatrix} \omega_1^1 & 0 & \cdots & 0 \\ 0 & \omega_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_1^N \\ \omega_2^1 & 0 & \cdots & 0 \\ 0 & \omega_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_2^N \end{bmatrix} \quad (7)$$

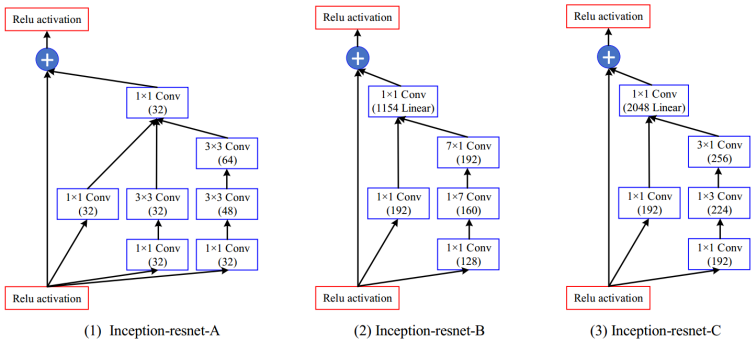

Fig. 4. Inception-resnet structure

where $\omega \in R^{2N \times N}$ is a matrix composed by the hand gesture class fusion weights. $\odot$ represents the element-wise product.

The purpose of equation (7) is to learn the optimal hand gesture class fusion weights to achieve the best fusion effect. The network is finally classified by a softmax layer. To avoid over-fitting of the training, a rectified linear unit (ReLU) is integrated into the loss function of the softmax layer. The loss function is given as follows:

$$\omega = \min_{\omega} loss(p, y; \omega) + \max(0, \omega) \tag{8}$$

where $loss(\cdot)$ is the original loss function of the softmax layer, y represents the ground-truth labels, and $\max(0, \omega)$ is the ReLU result of .

Finally, the recognized hand gesture c is:

$$c = \arg\max P_f \tag{9}$$

4 Experiment results and discussion

4.1 Train

The training process uses Transfer Learning, which can effectively reduce training time and steps. The Inception_resnet_v2 model trained under ImageNet database is used as the pre-training model, freeze all network layer parameters except softmax layer. Then, change the neuron number in the softmax layer to 24. After that, retrain the model on the ASL fingerspelling database.

In the process of training, the data is divided into batches, which can improve the efficiency of training. Set the batch size N to 24. The gradient solution method in this experiment uses Stochastic gradient descent (SGD). In the SGD, ω is updated by the linear combination of the negative gradient and the last weight update value V_t. The iteration equation is shown as follows:

$$V_{t+1} = \mu V_t - \alpha \bigtriangledown L(\omega_t) \tag{10}$$

$$\omega_{t+1} = \omega_t + V_{t+1} \tag{11}$$

where α is the learning rate of the negative gradient. μ is the weight of the last gradient value and it is used to weight the effect of the previous gradient direction on the current gradient direction. These two parameters can get the best results through tuning. t represents the current number of iterations. The adjustment method of the learning rate selects the step uniform distribution strategy, which can make the network quickly converge in the early stage and reduce the oscillation in the later stage. Its calculation equation is shown as follows:

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

where α_0 is the initial learning rate and is set to 0.001. γ is the adjustment parameter and is set to 0.1. s represents the iteration length of the adjustment learning rate and is set to 10000. That is, when the current iteration number t reaches an integral multiple of 10000, the learning rate is adjusted. The total number of training steps is 30000.

4.2 Validation

The verification process of the experiment adopts the "half-half" method. That is, half of the ASL fingerspelling database is used as training data and the other half is used as testing data. Data augmentation is performed only on the training data, and data alignment and data enhancement are performed on the test data.

To verify the superiority of the Inception-ResNet v2, the experimental results are compared with that of other typical networks. Four networks, VGG-19, ResNet152, Inception v4 and Inception-Resnet v2, are used as CNN feature extractors for 2S-CNN. After training, 4 ASL gesture recognition models are obtained. These four models are used to verify their accuracies and speeds on the testing data. The results are shown in Table 1. The experiments are carried out under the Ubuntu16.04 system. The deep learning framework uses Tensorflow and the GPU chose GTX1060.

Table 1. The accuracies and speeds of four models

CNN Feature Extractor	Speed(ms)	Accuracy(%)	GPU
VGG-19[16]	36	80.86	GTX1060
ResNet152[17]	38	83.80	GTX1060
Inception v4[15]	**30**	85.32	GTX1060
Inception-ResNet v2[15]	33	**92.08**	GTX1060

As can be seen from Table 1, the speeds of the four ASL gesture recognition models are all within 40ms, and the recognition accuracies are all above 80%.

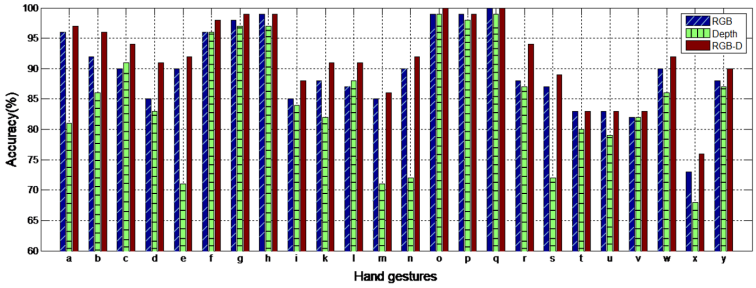

Fig. 5. ASL hand gesture recognition accuracy comparison chart

Where the model that uses ResNet152 as the CNN feature extractor gets the slowest speed (38ms). The model that uses Inception v4 as the CNN feature extractor gets the fastest speed (30ms). The model that uses Inception-ResNet v2 as the CNN feature extractor gets the highest accuracy (92.08%). And the model that uses VGG-19 as the CNN feature extractor gets the lowest accuracy (80.86%). Therefore, it can prove that the Inception-ResNet v2 can extract hand gesture features efficiently and help improve the recognition accuracy of the model. And the recognition speed of the model can be real-time. So, it can be applied as a gesture recognition model to the real-time HRI system.

In addition, in order to verify the effectiveness of the proposed 2S-CNN network for fusing RGB and depth features, we compared the hand gesture recognition results of the 2S-CNN with single-stream frameworks using RGB and depth, respectively. These three models all use Inception-ResNet v2 as the CNN feature extractors. The comparison of the hand recognition accuracies of the three models for each ASL gesture is shown in Fig. 5.

As can be seen from Fig. 5, the average recognition accuracies of the ASL hand gestures by the RGB-stream framework and depth-stream framework are 89.7% and 84.8%, respectively. The ASL hand gesture recognition accuracy of the 2S-CNN is 92.1%. Its average recognition accuracy increases by 2.4% and 7.3% compared to the RGB-stream and depth-stream, respectively. And its recognition accuracy of each ASL hand gesture exceeds 75%. Therefore, it can be proved that the proposed 2S-CNN framework can effectively improve the hand gesture recognition accuracy.

To further verify the superiority of our method, the experimental result is compared with some baseline methods. The comparison results are shown in Table 2.

It can be seen from Table 2 that the proposed 2S-CNN has a higher recognition accuracy for ASL hand gestures than the main baseline methods. Therefore, the validity and superiority of the method are proved.

Table 2. Comparison of the recognition accuracy

Method	Accuracy (%)
GF-RF[18]	75
ESF-MLRF[19]	87
RF-JP[20]	59
RF-JA+C[21]	90
2S-CNN	**92**

5 Conclusion remark and future work

In this paper, a two-stream CNN framework combining RGB and depth information is proposed for increase the recognition of ASL hand gestures. The framework extracts the RGB and depth features of the hand gesture through two CNN streams, respectively, and fuses the recognition results. The validity and superiority of the proposed method are verified by comparison of experimental results. The contributions of this paper mainly include as follows:

- A two-stream CNN framework is designed to fuse multiple hand gesture information to improve the recognition accuracy of hand gestures.
- Effective processing and enhancement of ASL data facilitates subsequent hand gesture feature extraction and recognition.
- The fusion layer is designed and connected to the network. By doing so, the optimal fusion weight can be obtained through network training.

Currently, this method is only applied to the recognition of static hand gestures. We know that the application of dynamic hand gesture recognition is more important in HRI, so the application of the idea in this paper can be transferred to dynamic hand gesture recognition in the future work.

Acknowledgment

Resrach supported in part by the Research Fund of China Manned Space Engineering under Grant 050102, in part by the Key Research Program of the Chinese Academy of Sciences under Grant Y4A3210301, in part by the Natural Science Foundation of China under Grant 51775541, 51575412, 51575338 and 51575407, in part by the EU Seventh Framework Programme (FP7)-ICT under Grant 611391, in part by the Research Project of State Key Lab of Digital Manufacturing Equipment & Technology of China under Grant DMETKF2017003, in part by National Key R&D Program Projects 2018YFB1304600.

References

1. M. A. Goodrich, A. C. Schultz.: HumanCrobot interaction: a survey. Foundations and Trends in HumanCComputer Interaction, vol.1, no.3, pp. 203-275 (2008)

2. J. Liu, Y. Luo, Z. Ju.: An interactive astronaut-robot system with gesture control. Computational intelligence and neuroscience, vol.2016. (2016)
3. S. S. Rautaray, A. Agrawal.: Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review, vol.43, no.1, pp.1-54 (2015)
4. Y. LeCun, Y. Bengio, G. Hinton.: Deep learning. nature, vol.521, no.7553, pp.436 (2015)
5. T. Wang, Y. Li, J. Hu, A. Khan, L. Liu, C. Li, M. Ran.: A Survey on Vision-Based Hand Gesture Recognition. In International Conference on Smart Multimedia, pp. 219-231, Aug (2018)
6. O. K. Oyedotun, A. Khashman.: Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications, vol.28, no.12, pp.3941-3951 (2017)
7. J. Nagi, F. Ducatelle, A. G. Di Caro, D. Cire?an, U. Meier, A. Giusti, L. M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. 2011 IEEE International In Signal and Image Processing Applications (ICSIPA), Conference on, pp. 342-347 (2011)
8. Y. Kim, B. Toomajian.: Hand gesture recognition using micro-Doppler signatures with convolutional neural network. IEEE Access, vol. 4, pp.7125-7130 (2016)
9. T. Yamashita, T. Watasue.: Hand posture recognition based on bottom-up structured deep convolutional neural network with curriculum learning. In Image Processing (ICIP), 2014 IEEE International Conference on, pp. 853-857 (2014)
10. Q. Gao, J. Liu, Z. Ju, Y. Li, T. Zhang, L. Zhang.: Static Hand Gesture Recognition with Parallel CNNs for Space Human-Robot Interaction. In International Conference on Intelligent Robotics and Applications, pp. 462-473 (2017)
11. C. J. L. Flores, A. G. Cutipa, R. L. Enciso.: Application of convolutional neural networks for static hand gestures recognition under different invariant features. In Electronics, Electrical Engineering and Computing (INTERCON), 2017 IEEE XXIV International Conference on, pp. 1-4 (2017)
12. Z. Zhang, Z. Tian, M. Zhou.: HandSense: smart multimodal hand gesture recognition based on deep neural networks. Journal of Ambient Intelligence and Humanized Computing, pp.1-16 (2018)
13. S. Hao, W. Wang, Y. Ye, T. Nie, L. Bruzzone.: Two-stream deep architecture for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, vol.56, no.4, pp.2349-2361 (2018)
14. Z. Zhang.: Microsoft kinect sensor and its effect. IEEE multimedia, vol.19, no.2, pp.4-10 (2012)
15. C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi.: Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, Vol. 4, pp. 12 (2017)
16. K. Simonyan, A. Zisserman.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. K. He, X. Zhang, S. Ren, J. Sun.: Identity mappings in deep residual networks. In European conference on computer vision, pp. 630-645 (2016)
18. N. Pugeault, R. Bowden.: Spelling it out: Real-time ASL fingerspelling recognition. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 1114-1119 (2011)
19. A. Kuznetsova, L. Leal-Taix, B. Rosenhahn.: Real-time sign language recognition using a consumer depth camera. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 83-90 (2013)
20. C. Keskin, F. K?ra?, Y. E. Kara, L. Akarun.: Real time hand pose estimation using depth sensors. In Consumer depth cameras for computer vision, pp. 119-137 (2013)

21. C. Dong, M. C. Leu, Z. Yin.: American sign language alphabet recognition using microsoft Kinect. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 44-52 (2015)