Gesture Recognition and Localization Using Convolutional Neural Network

Fei Wang, Li Kong, Xing Zhang, Hu Chen

Faculty of Robotics Science and Engineering, Northeastern University, Shenyang 110169
E-mail: wangfei@mail.neu.edu.cn

Abstract: Gesture recognition based on computer vision, which is natural and intuitive, is getting more and more attention in the field of human-computer interaction. Due to the limitations of existing gesture recognition methods, this paper presents an efficient and effective solution, which is divided into two main parts: dataset collection and the convolutional neural network (CNN) design. Firstly, a novel method is proposed to collect the hand gesture dataset through the sliding window, which makes it easy to capture arbitrary gesture data in any background, and the whole process is simple and fast. Secondly, a novel CNN named HandNet is designed for improving the performance of the gesture recognition and localization. The bone structure of the proposed CNN is HandBlock, which has the ability of richer feature expression and adaptive weight expression between high-level features and low-level features. The experimental results show that the proposed model achieves comparable performance with the state-of-the-art methods.

Key Words: Gesture Recognition, Convolutional Neural Network, Sliding Window, Feature Extraction

1 INTRODUCTION

As an efficient and direct body language, hand gesture plays an important role in the human-computer interaction. Meanwhile, with the development of artificial intelligence and human-computer interaction technology, gesture recognition has gradually become a hot research topic. According to the different sources of gesture information, gesture recognition methods are mainly based on data glove, EMG signals, inertial sensors and computer vision. The recognition method based on computer vision has great application prospects due to its advantages of not requiring wearable devices and low cost.

A hierarchical static identification method using finger detection and gradient direction histogram (HOG) features is proposed [1]. The proposed method can achieve better gesture recognition results, while the hand region can be successfully detected. Quan describes the edge feature of the gesture by constructing a distance operator and uses the KNN classifier to complete the gesture recognition after segmenting the gesture region and the background [2]. Gao proposes a new method based on multi-space feature fusion to recognize static gesture depth images, that is, local principal component analysis of 3D point clouds, and extract local gradient information and local point cloud depth distribution that encodes the local shape of the gesture [3]. A simple 5-layer neural network is used to train the grayscale gesture samples, which incorporates gesture feature extraction and classification [4]. Gao designs a parallel convolutional neural network to improve the accuracy of static gesture recognition in the complex background and dynamic lighting environment [5]. So we train the Faster R-CNN network based on the NUS gesture dataset, recognizes 10 static gestures from real-time webcam input, and controls the VLC media player based on the recognized gestures [6]. To eliminate the influence of background noise, the Kinect v2 depth camera is used to study the static gesture recognition problem, and the depth map information can effectively solve the problem of separation of gesture and background information. This method has successfully constructed different convolutional neural network for gestures classification problem [7].

The limitations of hand gesture recognition based on computer vision can be concluded as follows: (1) The lack of gesture data sets and the singularity of gesture categories in the dataset. The classification of the gestures based on the machine learning method is often limited by the train set; (2) The input of the gesture recognition system is usually only the hand image, while the input of the system in the realistic human-computer interaction scenario includes the human body and a large amount of background information in addition to the gestures; (3) The result of gesture recognition only contains category information, while ignoring the position information of the gesture, which does not meet the needs of human-computer interaction where the same gesture may convey different information in different positions.

In view of the above limitations, this paper proposes a simple and fast data collection method, combined with skin color features for spatial modeling, to improve the sample diversity while solving the lack of dataset. In addition, our HandNet based on HandBlock has the richer feature representation, which can predict the location information of the gesture in the whole image while accurately recognizing gesture.

This work is supported in part by the Fundamental Research Funds for the Central Universities under Grant N172608005, Liaoning Provincial Natural Science Foundation of China under Grant 20180520007, Public Welfare Research Funds for Scientific Undertaking of Liaoning Province under Grant 20170021.

978-1-7281-0106-4/19/$31.00 ©2019 IEEE

2 DATASET CONSTRUCTION

Due to the lack of gesture datasets, this paper presents a simple and fast way to collect hand gesture samples. In order to improve the diversity of the sample, we preprocess the data set and finally send it to the network for training.

2.1 Data Collection

We propose a method to collect data through the sliding window. The sliding window can move across the image acquisition interface. As shown in Figure 1, the N and L on the keyboard randomly adjusts the position and size of the sliding window respectively. And the S key is used to save the image capturing window as the train set. The program will record the position and size of the sliding window and the gesture category label.

Fig 1. Sample diagram of the data acquisition process

This data collection method is simple and fast, and is not restricted by hardware devices and places. The entire image can be saved instead of the hand image, and any gesture can be collected to satisfy the diversity of gesture categories. However, the collected samples tend to have a relatively simple background. Therefore, the color space-based skin color segmentation model is used to preprocess the gesture image, which can improve the background diversity while eliminating the influence of background and illumination.

2.2 Feature Space Model

The YC_bC_r color space has the advantages of the separation of brightness and chromaticity, which couldn't be influenced by fluctuations in brightness and skin color and has good clustering and stability [8-10]. In this paper, the improved YC_bC_r space is adopted, which is the nonlinear segmented space $YC_b'C_r'$. The skin color clustering region does not change with the value of Y, and then the luminance can be separated from the chrominance information. The nonlinear chrominance function is shown as follows:

$$C_i'(Y)=\begin{cases}(C_i(Y)-\overline{C_i}(Y))\cdot\dfrac{W_{C_i}}{W_{C_i}(Y)}+\overline{C_i}(K_h) & if\ Y<K_l\ or\ K_h<Y\\ C_i(Y) & if\ Y\in[K_l,K_h]\end{cases} \quad (1)$$

The skin color region obtained by nonlinear segmentation can be approximated by the elliptical model, which is shown as the formula (2) and formula (3).

$$\frac{(x-ec_x)^2}{a^2}+\frac{(y-ec_y)^2}{b^2}=1 \quad (2)$$

$$\begin{bmatrix}x\\y\end{bmatrix}=\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}C_b'-C_x\\C_r'-C_y\end{bmatrix} \quad (3)$$

If the point (x,y) is inside the ellipse, the point is the skin color point, and the pixel value of the point is set to 255. Otherwise, the point is considered to be a non-skin color point, and the pixel value of the point is set to 0. Finally, a gesture binary image can be generated. The specific process based on feature space transformation and feature space modeling is shown in Figure 2. The final gesture image, the sliding window position, size, and gesture category labels compose the gesture dataset for the network training.

Fig 2. Data transformation and modeling process diagram

3 DEEP NEURAL NETWORK

In order to realize gesture recognition and location, this paper proposes a HandNet network model. Compared with other CNN networks, the HandBlock structure introduced by the network has richer feature expression capabilities. Through selective and adaptive weight expression of high-level features and low-level features, it can improve network learning performance, realize gesture recognition, and verify its localization ability.

3.1 HandNet's Overall Framework

The architecture of HandNet is shown in Figure 3. The backbone network is the HandBlock structure for feature extraction. In the test network, the fully connected layer will return the location and classification results of the gesture.The specific parameter settings of the HandNet network are shown in Table 1.

Table 1. Configuration of HandNet

Layer Name	Output Size	Layers
Input	224×224	
Conv1_x	112×112	7×7,1/2
Maxpooling	56×56	3×3,1/2
Conv2_x	56×56	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Conv3_x	28×28	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
Conv4_x	14×14	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$
Conv5_x	7×7	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
Avgpool	1×2048	Global
FC	1×2048	2048×2048
FC	1×637	2048×637

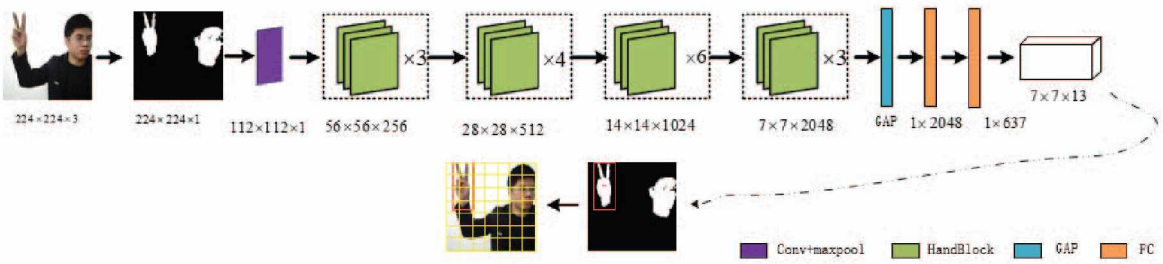

Fig 3. The architecture of HandNet

The network is mainly composed of multiple convolution blocks. Each convolution block is composed of multiple HandBlocks, and each HandBlock is composed of three convolution layers. Each convolutional layer is followed by batch normalization [11] processing.

All intermediate layers have a modified linear unit as the activation function. There is no pooling layer between the convolutional blocks. The feature dimension is reduced by the convolution operation of stride 2, and the feature scales of the convolution block are consistent. The fully connected layer divides the input image into 7x7 grids, each of which is responsible for detecting gestures that fall into the grid. Each grid outputs 2 bounding boxes containing gesture positions, and the gestures belong to probability information of each category. Finally, the bounding box is selected by non-maximum suppression, and the network outputs the position and category information of the gesture to realize gesture localization and recognition.

3.2 HandBlock Structure

3.2.1 Structural Design of HandBlock

As the depth of the CNN network increases, there will be problems such as gradient disappearance and network degradation during the training process. He Kaiming et al. propose the ResNet [12] network frame with residual structure.

The main idea of ResNet is to degenerate the depth model into a shallow model by a residual mapping, which is realized by the shortcut connection, as shown in Figure 4.

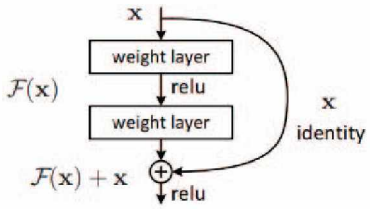

Fig 4. The structures of Blocks in ResNet

The original mapping $H(x)=F(x)$ is recast into $H(x)=F(x)+x$ on the hypothesize that it is easier to optimize the residual mapping than to optimize the original mapping. This structure is called Block. In addition, ResNet's residual network element can be decomposed into multiple network cascaded forms, as shown in Figure 5.

This structure allows ResNet to have more feature learning and presentation capabilities than other networks in the same parameters or networks of the same depth. Based on the above ideas and previous research [13], we propose the HandBlock structure:

$$H(x)=\gamma F(x)+\beta x \tag{4}$$

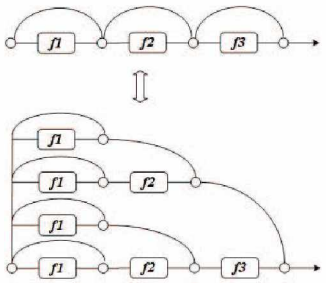

Fig 5. The cascaded structures transformed from ResNet

We call γ and β as Gate Units, which are used to control the switch of the branch and the resistance. γ and β are parameters that are automatically tuned as network parameters are updated. The proposed residual block is shown in Figure 6.

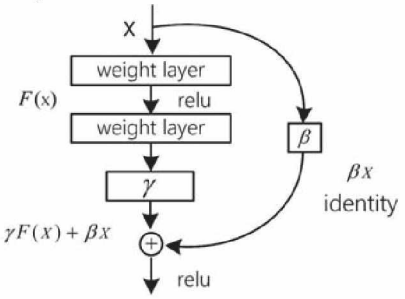

Fig 6. The structures of HandBlocks

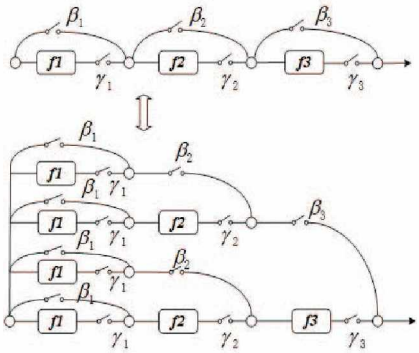

Fig 7. The cascaded structures transformed from HandNet

The gate unit used by HandBlock can achieve better feature extraction, feature selection and feature combination. The decomposition form is shown in Figure 7. Each of its branches is controlled by the gate unit, and its effects are similar to those of LSTM [14] and GRU [15]. When γ and β learn a small value (close to 0), the branch is broken and feature selection can be achieved. When γ and β are not 0, the branch is closed. The values of γ and β are the weights of the branch, which enables a richer feature representation.

3.2.2 HandBlock Optimization Strategy

The main idea of Batch Normalization (BN) is to add the standard normal distribution process in the middle layer of the deep network and constrain the network to automatically adjust the intensity of the standardization during the training process, thereby speeding up the training and reducing the cost of weight initialization. Similarly, the parameters in the HandBlock also adopt this idea, and finally achieve feature selection and feature weight adjustment. In the initial state, β=1 and γ=1 .

In the process of backpropagation, the updates of γ and β are as follows:

$$\gamma := \gamma - \alpha \frac{dL}{dH(x)} F(x) \tag{5}$$

$$\beta := \beta - \alpha \frac{dL}{dH(x)} x \tag{6}$$

Where L is $Loss$, $F(x)$ is the feature of the stack layer of each HandBlock, x is the feature of the upper HandBlock obtained by the shortcut, and $H(x)$ is the output feature of the HandBlock fusion.

3.2.3 HandBlock Verification

According to the above, the network composed of HandBlock should have two advantages: it can solve the network degradation problem more easily, which means that the learning speed is fast; the network feature expression ability is stronger.

In order to verify whether HandBlock has these two advantages, the preliminary work [13] has carried out detailed experimental proof. In our experiments, the ResNet and HandBlock networks with the same structure were used to compare the expression recognition tasks in the FER2013 dataset [16] and the NVIE dataset [17]. The network structure used is shown in Table 2.

The experimental parameters of the two groups were set consistently, and the Gradient Descent with Momentum was used to optimize the cost function. Among them, the batch size was 64, the momentum parameter was 0.9, the weight attenuation parameter was 0.0001, and the full connection layer uses Dropout with a parameter of 0.5. The initial learning rate is 0.01. When the accuracy stops increasing, the learning rate is attenuated by a factor of 10. The Gaussian distribution is used to randomly initialize the weights, and all offsets are initialized to 0. In order to avoid over-fitting problems, the diversity of training samples is increased by random cropping, flipping, and rotation.

Table 2. Configuration of ResNet and CNN Composed of HandBlock

Layer Name	Output Size	Layers
Input	96×96	
Conv1_x	48×48	$\begin{bmatrix} 1\times1,16 \\ 3\times3,16 \\ 1\times1,64 \end{bmatrix} \times3$
Conv2_x	24×24	$\begin{bmatrix} 1\times1,32 \\ 3\times3,32 \\ 1\times1,128 \end{bmatrix} \times3$
Conv3_x	12×12	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64 \\ 1\times1,512 \end{bmatrix} \times3$
FC	1×7	Average pool, 7-d fc, softmax

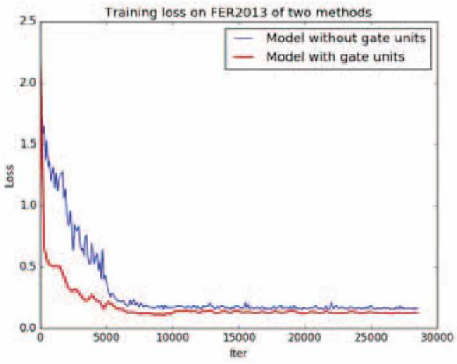

Fig 8. Training loss on FER2013 of two methods

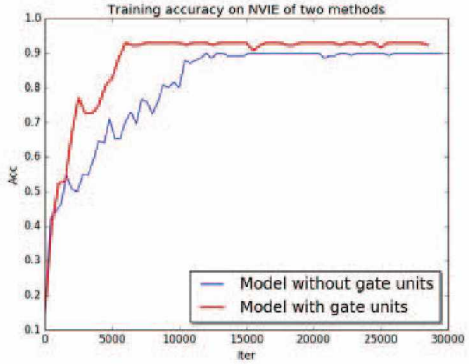

Fig 9. Training accuracy on NVIE of two methods

The comparison of Loss on the FER2013 dataset and the comparison of the accuracy on the NVIE dataset are shown in Figures 8 and 9. Model with gate units is the HandBlock structure, and Model without gate units is the Block structure.

It can be concluded from the training results of the two network architectures in the FER2013 and NVIE datasets that HandBlock has faster convergence speed and higher training accuracy than ResNet, which fully proves the advantages of HandBlock.

3.2.4 Loss Function

The loss function used in the HandNet includes coordinate error, IOU error and classification error. The specific loss formula is shown as formula (7).

Different errors have different contributions to the loss function, so the correction factors α , α_{out} are added to the loss function, and are respectively 0.5 and 0.5. M is the dimension of the grid into which the input image is divided. N is the number of bounding boxes that can be predicted for each grid. $\overline{x}、\overline{y}、\overline{\omega}、\overline{h}、\overline{L}、\overline{p}$ are the label values. $\prod_{ij}^{in}$ denotes that the j^{th} bounding box predictor in cell i is "responsible" for that prediction, and $\prod_{ij}^{out}$ is the opposite. In the HandNet, M is 7 and N is 2. Each bounding box predicts 5 values, which are center coordinate x、y , width w , height h , and confidence.

$$
\begin{aligned}
loss = & \alpha \sum_{i=0}^{M^2} \sum_{j=0}^{N} \prod_{ij}^{in} [(x_i - \overline{x_i})^2 + (y_i - \overline{y_i})^2] \\
& + \alpha \sum_{i=0}^{M^2} \sum_{j=0}^{N} \prod_{ij}^{in} [(\sqrt{\omega_i} - \sqrt{\overline{\omega_i}})^2 + (\sqrt{h_i} - \sqrt{\overline{h_i}})^2] \\
& + \sum_{i=0}^{M^2} \sum_{j=0}^{N} \prod_{ij}^{in} (L_i - \overline{L_i})^2 \\
& + \alpha_{out} \sum_{i=0}^{M^2} \sum_{j=0}^{N} \prod_{ij}^{out} (L_i - \overline{L_i})^2 \\
& + \sum_{i=0}^{M^2} \prod_{i}^{in} \sum_{l \in L} (p_i(l) - \overline{p_i}(l))^2
\end{aligned}
\tag{7}
$$

4 Experimental Results and Discussion

4.1 Data Set Description

The data sample is collected by sliding the window, and the image acquisition window screen is set to 224×224 . The gestures are set to three categories: scissors, stone, and cloth. The experiment carried out data collection for 5 people. The number of samples in each category was 2,400, 2000 as the training set and 400 as the test set.

4.2 Experimental Environment Settings

The experimental platform is the Ubuntu 16.04 system with Nvidia Titan XP, and the model uses the TensorFlow framework. The optimizer uses Adam [18] with parameters set to β_1=0.9 、 β_2=0.999 . The Batch Size is set to 128 and the initial learning rate is 0.01. The learning rate is exponentially attenuated with a decay rate of 0.9. The weight is initialized using the Xavier method [19], and its formula is as follows:

$$
W \sim U[-\frac{6}{\sqrt{n_i + n_{i+1}}}, \frac{6}{\sqrt{n_i + n_{i+1}}}] \tag{8}
$$

Where n_i is the number of neurons in the upper layer and n_{i+1} is the number of neurons in the lower layer. In addition, the input sample of the network is a picture of $224 \times 224 \times 3$. At the same time, the BN layer is connected to each layer of the convolutional layer of the network. The parameter is activated by LeakeyRelu [20], and the parameter is set to 0.1.

4.3 Experimental Results and Analysis

In this paper, the HandNet network proposed is used to perform 6000 iterations of the gesture data samples. The changes of the loss and IOU curves with the number of iterations during the training are shown in Figures 10 and 11. It can be seen from the figure that on the training set, when the number of iterations is about 1800, the values of the loss and IOU tend to be stable, and the network parameters tend to be stable.

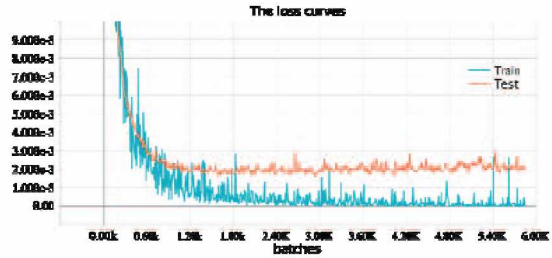

Figure 10. Loss curves of the HandNet

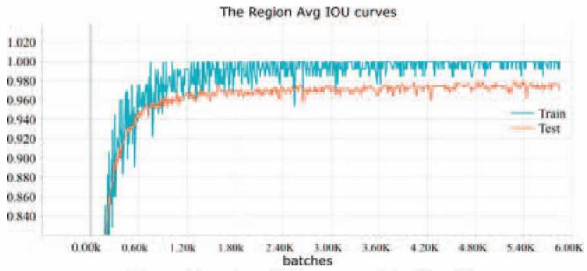

Figure 11. Avg IOU curves of the HandNet

In order to verify HandNet's advantages in gestures, it is compared with traditional gesture recognition algorithms and CNN-based gesture recognition algorithms. Table 3 shows that because our network introduces the Handblock module with richer feature expression capabilities, it has a higher recognition rate than other algorithms.

Table 3. Comparison of Accuracy of Different Methods

Method	Accuracy
Quan[2]	90%
Wang[4]	95.0%
Gao[5]	93.3%
Soe[6]	92.1%
Goga[7]	89.79%
Our method	**97.6%**

5 Conclusion

Due to the lack of existing gesture datasets and the limited types of gestures in datasets, this paper proposes a method to collect data through the sliding window, which makes it easy to capture arbitrary gesture data in any background, and the whole process is simple and fast. In addition, the proposed novel network is named HandNet, which is based on HandBlock that has the ability of the richer feature

expression and adaptive weight expression between high-level features and low-level features.

In this paper, we collect data of three types of gestures, such as scissors, stones and paper through the sliding window, and train them through HandNet. Eventually, the network showed good performance of gesture recognition and localization. The method proposed in this paper can further meet the needs of intelligent human-computer interaction.

In future research, we will extend the proposed HandBlock architecture to other tasks. The problem that the model is easily affected by skin color in the background will be further explored.

REFERENCES

[1] Liu Shuping, Liu Yu, Yu Jun, Wang Zengfu, Hierarchical static hand gesture recognition by combining finger detection and HOG features, Journal of Image and Graphics, Vol.20, No.6, 781-788, 2015.

[2] Chunying Quan, Jianning Liang, A simple and effective method for hand gesture recognition, International Conference on Network and Information Systems for Computers, 302-305, 2016.

[3] Zhe Gao, Hand gesture recognition using multiple spatial features fusion, Journal of Chinese Computer Systems, Vol.37, No.7, 1577-1582, 2016.

[4] Long Wang, Hui Liu, Bin Wang, Pengju Li, Gesture recognition method combining skin color models and convolution neural network, Computer Engineering and Applications, Vol.53, No.6, 209-214, 2017.

[5] Qing Gao, Jinguo Liu, Zhaojie Ju, Static hand gesture recognition with parallel CNNs for space human-robot interaction, 10th International Conference Intelligent Robotics and Applications, 462-473, 2017.

[6] Soe H M, Naing T M. Real-time hand pose recognition using faster region-based convolutional neural network, 1st International Conference on Big Data Analysis and Deep Learning, 104-112, 2018.

[7] Goga J, Kajan S. Hand gesture recognition using 3D sensors, International Symposium ELMAR, 181-184, 2017.

[8] Jie Li, Xiaoli Hao, Face detction using ellipse skin model, Computer Measurement and Control, Vol.14, No.2, 170-171, 2006.

[9] Dongcheng Shi, Kang Ni, Background modeling based on YCbCr color space and gesture shadow elimination, Chinese Optics, Vol.8, No.4, 589-595, 2015.

[10] Heng Yang, Zaijun Zhang, Dong Yang, Ruliang Zhang, Research on face detection algorithm based on YCbCr skin model and area marking, Software Guide, Vol.15, No.2, 41-43, 2016.

[11] Ioffe S, Szegedy C, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778, 2016.

[13] Fei Wang, Hu Chen, Li Kong, Weihua Sheng, Real-time facial expression recognition on robot for healthcare, IEEE International Conference on Intelligence and Safety for Robotics, 402-406, 2018.

[14] Hasim S, Andrew S, Francoise B, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, 15th Annual Conference of the International Speech Communication Association, 338-343, 2014.

[15] Dey R, Salem F M, Gate-variants of Gated Recurrent Unit (GRU) neural networks, IEEE 60th International Midwest Symposium on Circuits and Systems, 1597-1600, 2017.

[16] Jeon J, Park J C, Jo Y J, A real-time facial expression recognizer using deep neural network, International Conference on Ubiquitous Information Management & Communication, 1-4, 2016.

[17] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, A natural visible and infrared facial expression database for expression recognition and emotion inference, IEEE Transactions on Multimedia, Vol.12, No.7, 682-691, 2010.

[18] Kingma D P, Ba J. Adam, A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[19] Glorot X, Bengio Y, Understanding the difficulty of training deep feedforward neural networks, 13th International Conference on Artificial Intelligence and Statistics, 249-256, 2010.

[20] Andrew L M, Awni Y H, Andrew Y, Rectifier nonlinearities improve neural network acoustic models, International Conference on Machine Learning, 6-10, 2013.