CNN-MCF-ELM Model Recognizes Facial Expressions

Lin Shi[1], JiaLi Zou[2], Zhigang Li[3]
[1]School of Artificial Intelligence, North China University of Science and Technology, Tangshan, China;
[2]School of Electronic Information Engineering, Sias University, ZhengZhou, China
[3]Computer Center, TangShan College, TangShan, China
[1]Corresponding author
E-mail: [1]*12389248@qq.com*, [2]*1160575248@qq.com*, [3]*47971396@qq.com*

Abstract

In order to better solve the problems of low accuracy of facial expression recognition caused by the insufficient feature extraction of traditional neural networks, as well as the large amount of calculation of parameter adjustment, long time consumption, and weak model generalization ability in facial recognition, this paper proposes a multi-convolutional neural network Expression recognition method combining layer feature fusion and extreme learning machine (ELM). This method is to use CNN network to extract multi-layer facial expression feature maps, and then use the multi-scale pooling operation of the last three-layer feature maps extracted by CNN to merge these three feature vectors into a facial expression feature vector. The vector has the properties of multi-scale and multi-attribute, which can express the expression features well. Finally, the fused facial expression feature vector is input to the ELM classifier for expression recognition. Experimental results show that this method can effectively improve the accuracy of facial expression recognition. The average recognition accuracy on the CK+ and FER2013 data sets reaches 98.51% and 78.97%, respectively, and reduces the recognition time. At the same time, the design experiment verifies that the model has strong generalization ability.

Keywords: Facial Expression Recognition, Convolution Neural Network, Multi-scale Pooling, Multi-layer Characteristic Fusion, Extreme Learning Machine.

1. Introduction

In daily communication, facial expression recognition is a very important ability of "observing one's face" in the process of interpersonal communication. It can predict the change of inner emotion through the facial expression, and then make appropriate speech adjustment. According to a study by Merhirabian, a famous psychologist, human beings are able to communicate with each other emotionally mainly because they can express their inner feelings through language, tone and facial expressions. Facial expressions account for 55% of the total emotions. Not as easy to understand as the body gesture, facial expressions often contain a lot of internal information that can not be directly interpreted, need to be combined with psychology, physiology and other analysis. Because the research on facial expression recognition can achieve the purpose of "looking through the phenomenon to see the essence", the research in this field has become a hot spot in today's scientific research. Traditional facial expression recognition methods tend to cause the loss of original facial expression information and other problems, and the information features often cannot represent facial expression, so the accuracy of facial expression recognition is low. In addition, the traditional facial expression recognition algorithm is designed by human, so the computational power is limited and training is difficult.

Convolutional neural network can automatically learn highly specific image features according to specific classification types. Therefore, CNN has become very popular in the field of image recognition at present. Researchers in this field began to conduct a lot of research on CNN in the direction of image recognition and achieved great success. At the same time, the use of CNN for facial expression recognition has gradually become popular and achieved excellent results. Many researchers in this field have applied CNN to facial expression recognition research. Among them, Zhang et al. used DBN (Deep belief Network) [1] and BP [2] neural network to recognize facial expression and achieved good results. The deep network AUDN proposed by Liu et al. [3] combines CNN and Boltzmann Machine to extract features of facial actions, and then inputs the extracted features into SVM (Support Vector Machine) [4] classifier for classification. The selection of classifier has a great influence on improving the efficiency of facial expression recognition. The commonly used feature classification algorithms mainly include SVM and K-nearest Neighbor algorithm [5]. Compared with these commonly used feature classification algorithms, ELM (Extreme Learning Machine) [6] is a better classifier selection. In the process of ELM network training, the weights matrix between input layer and hidden layer and offset values are randomly generated, do not need to calculate and iterative update, only need to compute the weights between hidden layer and output layer matrix, which has some advantages including taking a short time, a small amount of calculation, needing simple parameter settings and fast convergence, comparing to SVM, K neighbor and some other algorithms.

The significant advantage of CNN in the field of image recognition is that it can extract high-quality features without human intervention. In order to make full use of the comprehensiveness of information after multi-layer feature fusion of CNN and the advantages of ELM classification with strong generalization ability and high classification accuracy, this paper proposes a facial expression recognition method based on CNN multi-layer feature fusion and extreme learning Machine (CNN-MCF-ELM).

2. Materials and Methods

The overall architecture of the expression recognition model (CNN-MCF-ELM) proposed in this paper is shown in Fig.1, including four stages of CNN feature extraction, multi-scale pooling, multi-layer feature fusion and ELM classification.

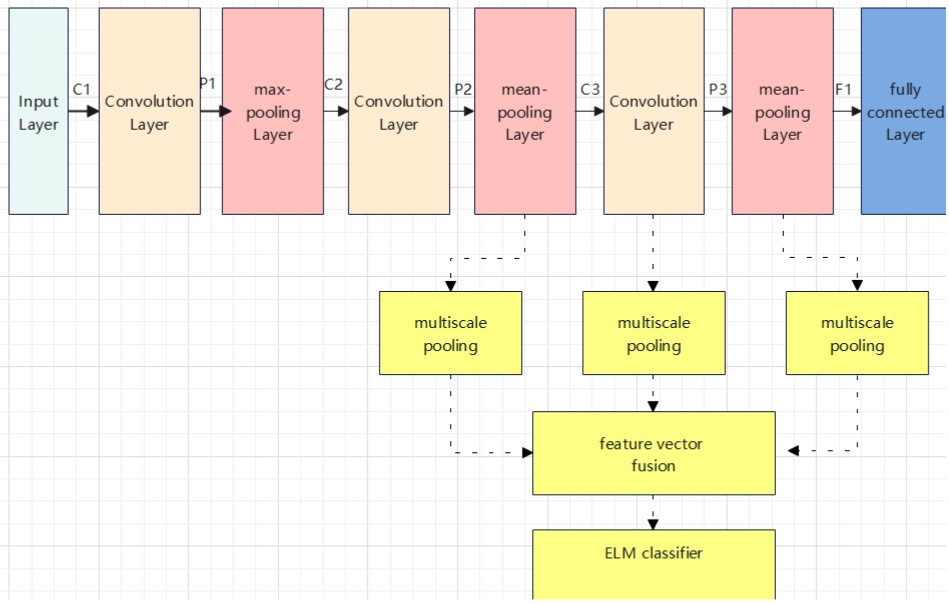

Fig.1 Structure of CNN-MCF-ELM Model

2.1. Image preprocessing

Hepatocellular carcinoma HepG2 cell line was obtained from the Cell Bank of the Chinese Academy of Science (Shanghai, China) and grown at 37 ℃ in a humidified atmosphere containing 5% CO_2. Medium used for HepG2 culture was Dulbecco's Modified Eagle Medium (DMEM; Gibco, Shanghai, China) supplemented with 10% fetal bovine serum (FBS) (v/v) (Gibco, Shanghai, China), penicillin-streptomycin (20 U/mL and 20 μg/mL, respectively) (Invitrogen, Shanghai, China). Cell number was counted by a hemocytometer. Cells at about 80% confluence were split 3 times per week to ensure health and growth as well as the cell performing assays.

Image preprocessing has big effect on CNN feature extraction. In order to prevent overfitting, this paper proposes a method for data enhancement, which is to rotate each image by θ respectively centered on the origin. The operation does not change the original image pixel values, but ensures the image resolution is exactly the same before and after the rotation. The coordinates of the pixel after rotation transformation are:

$$[x_1 \quad y_1 \quad 1] = [x \quad y \quad 1]\begin{bmatrix} cos\,\theta & sin\,\theta & 0 \\ -sin\,\theta & cos\,\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

According to Equation (1), each image is rotated by 90 °, 180 °and 270 °respectively, and the expanded data set is 3 times of the original data set.

2.2. CNN feature extraction

In the stage of feature extraction, CNN network model is adopted. CNN feature extraction is different from traditional manual feature selection; it can automatically acquire important features. The CNN model used in this paper is based on the improvement of the classic Lenet-5 model [7]. Its structure consists of three convolutional layers represented by C1, C2 and C3, and three pooling layers represented by P1, P2 and P3 respectively. Among them, P1 layer is the largest pooling layer, P2 and P3 layers are the average pooling layers, and one fully connected layer is represented by F1. The pooling operation adopts the combination of maximum pooling and average pooling. CNN is only used for feature extraction, not for classification recognition. At this time, F1 layer has no classification function. Parameters of each network layer are shown in Tab. 1.

Tab.1 Parameters of Each Layer of the Convolutional Neural Network Model

Layer	Type	feature map	Convolution kernel size	step size
0	input	48×48	——	——
C1	convolution	48×48×32	5×5	1
P1	max-pooling	24×24×32	——	2
C2	convolution	24×24×64	3×3	1
P2	mean-pooling	12×12×64	——	2
C3	convolution	12×12×64	3×3	1
P3	mean-pooling	6×6×64	——	2
F1	fully connected	1×1024	——	——

2.3. Multi-Feature Fusion

The traditional CNN network only uses the information of the feature graph in the last layer in the facial expression recognition process, and does not make full use of the features of each layer in the network. In literature [8], the author has carried out detailed visualization of the features of convolutional neural network, indicating that the representations at different levels correspond to different feature attributes of the identified objects. Therefore, the feature images extracted by multi-layer CNN are selected in this paper to represent the facial features without missing important feature information as much as possible. In this paper, the multi-feature fusion method is adopted to obtain the feature vectors which can fully represent the multi-attribute features of facial expressions.

Multiple features fusion method is as shown in figure 2. Firstly, the feature figures are extracted by the last three layers of CNN——C4,C5,P3, then they are fused through multi-scale pooling. This paper chose to fuse features extracted from the final three layers of CNN, which is inspired by LIS [9]. In the diagnosis and recognition of hepatocellular carcinoma they used the CNN last three-layer features and obtained a high accuracy, which means the last three layers contain rich feature information.

The idea of multi-scale pooling algorithm is derived from spatial pyramid pooling [10]. After several tests, the output of each layer after multi-scale pooling operation is three feature matrices with different scales of1×1×r,2×2×r and 3×3×r respectively, where r is the number of feature graphs. The three feature matrices form (13×r)×1 column vector according to the column, and finally form a feature column vector with multi-scale and multi-attributes through feature fusion as the input of ELM classifier. A schematic diagram of multi-scale pooling operation is shown in Fig. 2.

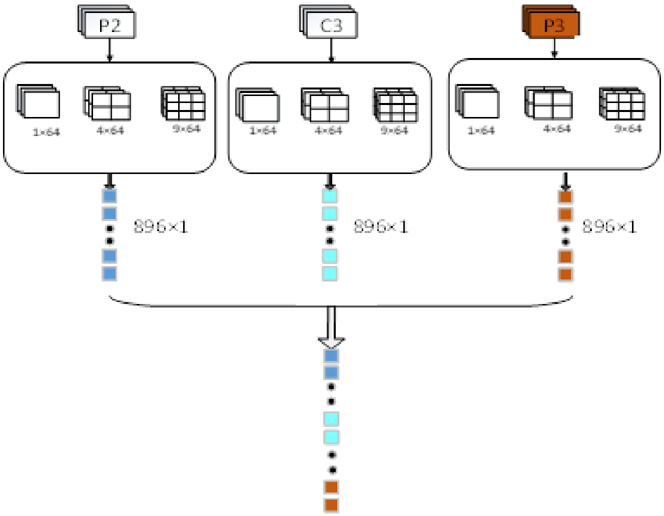

Fig.2 Schematic Diagram of Multi-scale Pooling Operation

2.4. Classification of ELM

CNN in the facial expression recognition model proposed in this paper is only used to extract features of facial expression images required for classification, and does not participate in the classification stage of facial expression recognition. Training of CNN is generally divided into two stages: In the forward propagation stage and back propagation stage, many parameters need to be calculated and adjusted in the training process, which takes a long time, so the ELM classifier with simple training process and high classification accuracy is used for facial expression recognition in the classification stage.

2.5. Introduction of ELM Classifier

Extreme learning machine is composed of input layer, hidden layer and output layer. In the process of network training, ELM only needs to calculate the weight matrix between the hidden layer and the output layer. The weight matrix and bias value between the input layer and the hidden layer are randomly generated without calculation or iterative update [7], so it can save a lot of training time and reduce the cost of calculation. The ELM network structure is shown in Figure 3. The numbers of neurons in input layer, hidden layer and output layer are D, L and M respectively.

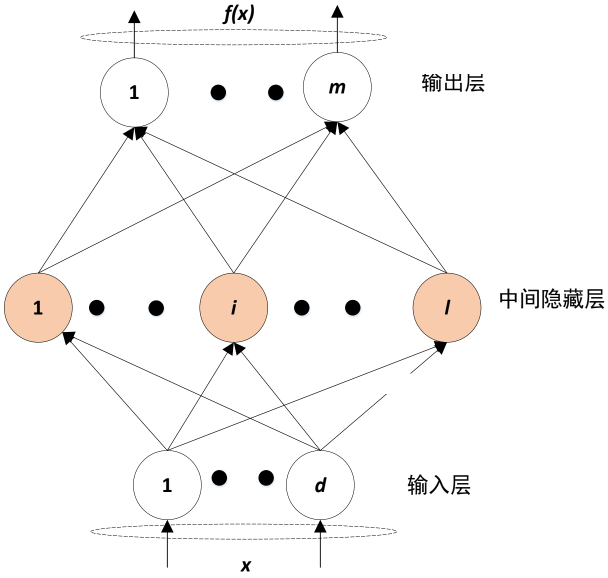

Fig.3 ELM Network Structure

Input x is a multi-scale and multi-attribute facial expression feature vector with dimension P. The output of the ith hidden node is:

$$g(x;\omega_i;b_i)=g(x\omega_i+b_i) \tag{2}$$

Where, g, ω_i, $b_i$respectively represent activation function, input weight vector between the ith hidden node and all input nodes, bias of the ith hidden node, i=1,2,…,l。 g is ReLU function as following:

$$g(x,\omega_i,b_i)=\max(0,x\omega_i+b_i) \tag{3}$$

The connection between the input layer and the hidden layer is a mapping process. As the input feature vector X is a p-dimensional space feature vector, so the connection is a process of mapping from p-dimensional space to l-dimensional space. The mapping feature vector of input vector x is:

$$h(x)=[g(x,\omega_1,b_1),g(x,\omega_2,b_2),\dots,g(x,\omega_l,b_l)] \tag{4}$$

The output layer has m output nodes, m is the number of types of expressions, and each output node corresponds to an expression. The output weight between the ith hidden node and the jth output node is expressed as β_{ij}, where j=1,2,…,m. So the value of the jth output node is:

$$f_j(x)=\sum_{i=1}^{l}\beta_{ij}\times\ g(x,\omega_i,b_i) \tag{5}$$

Therefore, input sample X, and its output vector at the hidden layer can be expressed as:

$$f(x)=[f_1(x),f_2(x),\dots,f_m(x)]=h(x)\beta \tag{6}$$

where,

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_l \end{bmatrix} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{l,1} & \beta_{l,2} & \cdots & \beta_{l,m} \end{bmatrix} \tag{7}$$

In the test phase, input test set sample X, and the corresponding expression category is expressed as:

$$label(x) = arg_{j=1,2,\dots,m} max f_j(x) \tag{8}$$

2.6. Training of ELM classifier

Suppose there are N samples, which are feature vectors with n attributes extracted from the last three layers of CNN after fusion, the input and expected output of the network are respectively represented as $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]$, $y_i = [y_{i1}, y_{i2}, \dots, y_{im}]$. The extreme learning machine described by the definition of ELM classifier can approach the training sample with zero error, namely $\sum_j^N \| f_j(x) - y_j \| = 0$, so there are ω_i， β_{ij}， b_i makes $f_j(x) = \sum_{i=1}^{l} \beta_{ij} \times \; g(x, \omega_i, b_i)$ true, which can be represented in matrix form as following:

$$H\beta = Y \tag{9}$$

$$H = \begin{bmatrix} g(\omega_1 x_1 + b_1) & \cdots & g(\omega_l x_1 + b_l) \\ \vdots & \ddots & \vdots \\ g(\omega_1 x_N + b_1) & \cdots & g(\omega_l x_N + b_l) \end{bmatrix} \tag{10}$$

$$Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,m} \\ y_{2,1} & \cdots & y_{2,m} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,m} \end{bmatrix} \tag{11}$$

β in the equation is the same as β in formula (7), which is not repeated here. H is the output matrix of the hidden layer. The connection weights between the hidden layer and the output layer can be obtained by solving the least square solution of the equations $min\|H\beta - Y\|$. The solution is:

$$\hat{\beta} = H^{+}Y \tag{13}$$

Where, H^{+} is the generalized inverse matrix of H , which is the output matrix of the hidden layer.

3. Results

3.1. The data set

Two public datasets, CK+ and FER2013, are used as the datasets of this paper. CK+ data set contains 123 facial expression images of different people, a total of 593 expression sequences and 951 image samples, which are divided into seven expressions: normal, angry, disgusted, fear, happy, sad and surprised. The schematic diagram of seven expression samples is shown in Fig.4

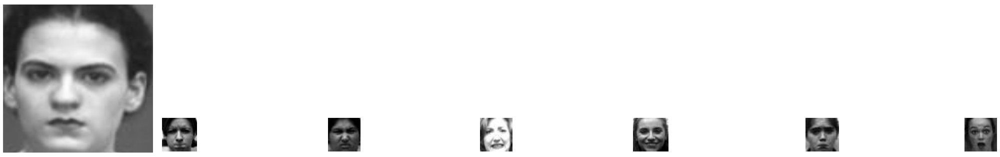

Fig.4 CK+ Expression Sample Diagram

The FER2013 dataset contains 35,887 natural facial expressions of different ages, nationalities and skin colors from real life. The seven expressions in FER2013 data set are angry, disgusted, fear, happy, neutral, sad and surprised, as shown in Fig.5.

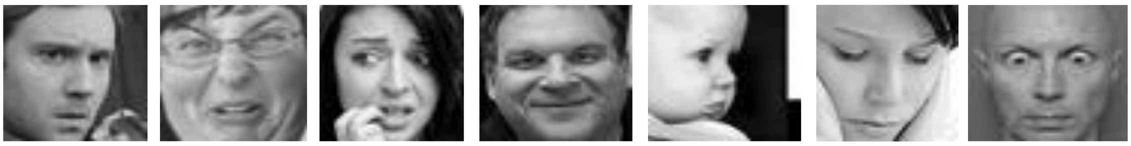

Fig. 5 FER2013 set expression sample diagram

Considering that FER2013 dataset contains more complete scenes that are more in line with the actual life of human beings, FER2013 is selected for training and testing models, and CK+ is used to verify whether the model has generalization.

3.2. Experimental Settings

This experiment was done in the PyCharm3.7 compiler and Python3.6 is used as the programming language. The deep learning framework is Google's deep learning tool, TensorFlow framework.

3.2.1. Comparison between multi-layer features and single-layer features

In order to verify that the multi-features fusion method in this paper can improve the accuracy of facial expression recognition, two groups of experiments were designed on the public FER2013 dataset and the results were compared. The experimental design of the two groups is as follows: In the first group, the single-layer features extracted from CNN network, namely the features extracted from the last average pooling layer, are input into the ELM classifier for expression recognition and the recognition accuracy is tested. In the second group of experiments, the features extracted from the last three layers of CNN were used for multi-feature fusion and then input into ELM classifier for expression recognition and its recognition accuracy is measured. The experimental results are shown in Figure 6. It can be seen from Figure 6 that the number of nodes in the hidden layer l is a factor that has a great influence on the recognition accuracy. When l <10000, the accuracy increases rapidly; when l >10000, the accuracy increases slowly. Considering the training time and calculation cost of the model, the number of hidden layer nodes l is 10000 in this experiment.

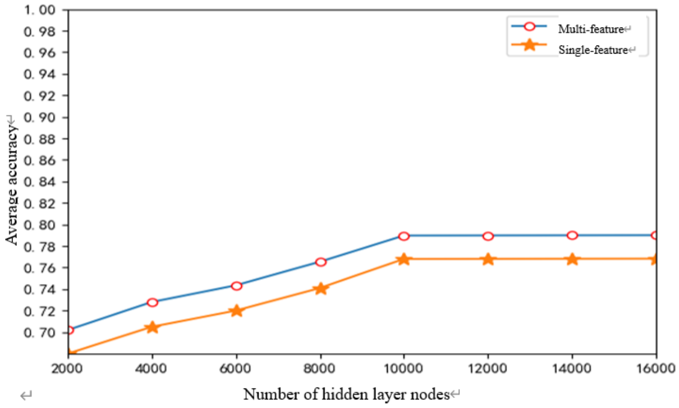

Fig.6 Accuracy of Different Number of Hidden Nodes

According to the experimental results, when l =10000, the average recognition accuracy of multi-layer features and single-layer features is 78.97% and 76.80% respectively. Regardless of the value of l, the recognition accuracy of multi-layer features is about 2 percentage points higher than that of single-layer features. Therefore, the experiment proves that multi-layer feature fusion can significantly improve the accuracy of facial expression recognition.

3.2.2. Comparison of different classifiers

Extreme learning machine is easy to implement, and has fast training speed, strong generalization ability and high classification accuracy. To validate that ELM classification also has significant advantage in the field of facial expression recognition, the paper compared ELM with the softmax classifier and the SVM classifier respectively. The comparision involve three aspects, which are training time, average recognition speed and accuracy, the experimental results are shown in Tab.2:

Tab.2 Performance Comparison of Different Classifiers

classifier	acuuracy/%	training time/min	recognition time/ms
Softmax	96.01	32.45	15.20
SVM	96.40	31.26	38.60
ELM	97.63	3.70	5.32

As can be seen from the table, the classification accuracy of ELM classifier is as high as 97.63%, significantly higher than that of Softmax and SVM. In terms of training time, ELM algorithm avoid complex iterative process, and the connection weights and bias are randomly generated, the number of parameters to be adjusted is small and the calculation cost is low, so it will obviously spend less time than other two kinds of classifier. In terms of recognition time, the recognition time of ELM classifier is greatly shortened to 5.32ms compared with the other two categories, saving a lot of time and cost.

3.2.3. Validation on other data sets

The CK+ dataset is used as the test sample to verify whether the algorithm in this paper has generalization. In order to ensure the reliability of the experiment, three times of cross-validation are conducted successively, and the average value is taken as the final result. To facilitate representation, recognition results are represented by confusion matrix, as shown in Tab. 3.

Tab.3 Confusion Matrix of Facial Expression Recognition on CK+

%	angry	disgusted	fear	happy	sad	suprised	netural
angry	98.67	0	0.96	0	0	1.23	0
disgusted	0.65	98.82	1.09	0	1.51	0.61	1.03
fear	1.68	0	96.30	0	0.15	1.01	1.09
happy	0	0	0	100	1.02	0	0
sad	0	0.57	1.04	0	98.92	0	0
suprised	1.34	1.74	0.03	0	0.54	99.92	1.00
netural	1.01	0.58	1.09	0	0	1.02	98.42

As can be seen from Table 3, the average recognition accuracy of the method in this paper is as high as 98.72% on CK+ data sets, indicating that it has good generalization ability.

4. Conclusion

It's the first time that the CNN - MCF - ELM model was used in the field of facial expression recognition, the model improved some problems of the existing facial expression recognition algorithm such as weak generalization ability, and low recognition efficiency. It made full use of some advantages of CNN and ELM, such as the comprehensiveness and diversity of CNN's multilayer features, short training time and less amount of calculation of ELM. The algorithm proposed in this paper has the following advantages:

(1) The excellent feature extraction capability of CNN network is utilized to avoid the limitation of manual feature selection and obtain features through automatic learning, which has better adaptability in complex environment.

(2) The ELM classifier is used as the classifier for the expression recognition stage, which greatly reduces the number of parameters that need to be adjusted, thus avoiding the tedious mathematical calculation process, shortening the model training time and improving the recognition efficiency.

(3) The comprehensiveness of multi-feature fusion improves the facial expression features extracted by CNN to be more representational, thus improving the facial expression recognition rate.

(4) The algorithm has good generalization ability and has certain guiding significance for practical application.

References

[1] Y Zhang, H Diao. "DBN based multi-stream models for speech" *IEEE International Conference on Acoustics IEEE*,(2003),15(2):200-204.

[2] L ZHAO, L PENG. "Ignoring Engine Misfire Based on GA-BP Neural Network". *Machinery Design & Manufacture*,(2017) 10:117-120.

[3] M LIU, S LI, S SHAN. "AU-inspired deep networks for facial expression feature learning". Neurocomputing, (2015),159:126-136.

[4] Y LI, S M MAVADATi, M H MAHOOR. "A unified probabilistic framework for measuring the intensity of spontaneous facial action units". *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, (2013):1-7.

[5] Q WANG, K JIA, P LIU. "Design and implementation of remote facial expression recognition surveillance system based on PCA and KNN algorithms". *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. (2016):314-317.

[6] L ZHANG, "Research on facial expression recognition method based on CNN-ELM". *Hunan Normal University,* (2019): 41-45.

[7] X ZHANG, L ZHOU, L ZHANG. "Facial expression recognition method based on improved LeNet-5". *Computer and Modernization*, (2019), (10): 83-87.

[8] M ZEILER, R FERGUS. "Visualizing and understanding convolutional networks". *13th European Conference on Computer Vision. Zurich :Springer*, (2014)： 818-833.

[9] S LI, J HIANG, W PANG. "Joint multiple fully connected convolutional neural network with extreme learning machine for hepatocellular carcinoma nuclei grading". *Computers in Biology and Medicine*, (2017), 84:156-167.

[10] S WU, "Research on Convolutional Neural Network Face Recognition Method Based on Multi-scale Pooling". *Zhejiang University*, (2016): 41-46