Facial expression recognition based on multi branch structure

Yuqing Xie[1], *Haichao* Huang[1], *Jianguang* Hong[1], *Xianke* Zhou[2,*], *Shilong* Wu[3], and *Peng* Lu[4]

[1]Information & Telecommunication Branch, State Grid Zhejiang Electric Power Company, Hangzhou, China
[2]Institure of Computing Innovation, Zhejiang University, Hangzhou, China
[3]University of California Santa Cruz, San Francisco, USA
[4]Zhejiang University, Hangzhou, China

Abstract. Facial expression recognition (FER) is an important means for machines to perceive human emotions and interact with human beings. Most of the existing facial expression recognition methods only use a single convolutional neural network to extract the global features of the face. Some insignificant details and features with low frequency are easy to be ignored, and part of the facial features are lost. This paper proposes a facial expression recognition method based on multi branch structure, which extracts the global and detailed features of the face from the global and local aspects respectively, so as to make a more detailed representation of the facial expression and further improve the accuracy of facial expression recognition. Specifically, we first design a multi branch network, which takes Resnet-50 as the backbone network. The network structure after Conv Block3 is divided into three branches. The first branch is used to extract the global features of the face, and the second and third branches are used to cut the face into two parts and three parts after Conv Block5 to extract the detailed features of the face. Finally, the global features and detail features are fused in the full connection layer and input into the classifier for classification. The experimental results show that the accuracy of this method is 73.7%, which is 4% higher than that of traditional Resnet-50, which fully verifies the effectiveness of this method.
Keywords: Multi branch; Convolution neural network; Global feature; Local features; Facial Expression Recognition.

1 Introduction

Facial expression is a form of non-verbal communication, and it is the main means to express people's physiological and psychological reactions in social communication. Through the analysis of facial expressions, we can roughly reflect people's current emotions and potential intentions, which plays an important role in the social interaction between people. In the past decade, with the rapid development of computer technology and the increasing maturity of image technology, facial expression recognition combined with

* Corresponding author: xkzhou@zjuici.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

computer technology and image technology has become a new research trend, and has achieved good results in various fields. For example, the application of facial expression recognition technology in the field of transportation can determine whether the driver is in the abnormal driving state such as fatigue driving, drunk driving and drug driving by monitoring the driver's expression; in the field of medical treatment, the psychiatrist can master the patient's psychological state through expression recognition and assist the psychiatrist in diagnosis and treatment; in the field of network education, According to the different learning ability of different students, through the expression recognition, analyze the learning situation of students, and feedback to the teacher, so as to facilitate the teacher to adjust the teaching plan in time; applied in the field of criminal investigation, it can help the police to analyze the psychological state and psychological changes of suspects, and assist in handling cases. Therefore, facial expression recognition has shown high application value in various fields. But at present, facial expression recognition is still facing great challenges. Deep network focuses on the global feature extraction of facial contour information and facial spatial distribution, and ignores the detail feature extraction such as texture, which makes great contribution to facial expression recognition, resulting in low accuracy of facial expression recognition.

In order to solve the above problem, this paper proposes a multi branch structure expression recognition method. The multi branch network uses Resnet50 as the backbone network, and the part after Conv Block3 is divided into three branches with different structures. The first branch is used to extract global features of human face, while branch 2 and branch 3 divide the face into two parts and three parts longitudinally to extract local features. Finally, the weighted fusion of global and local features is input to the classifier for classification, and softmax loss is used to calculate the network loss in the loss layer.

2 Related work

2.1 Facial expression recognition process

Facial expression recognition is an important biometric recognition technology, which extracts the facial expression features from the detected face by computer, and classifies the facial expressions according to the way of human thinking, so as to realize automatic and intelligent human-computer interaction. According to the principle of facial expression recognition, the recognition process can be roughly divided into four steps: face detection, image preprocessing, feature extraction, expression classification. The specific process is shown in Figure 1.

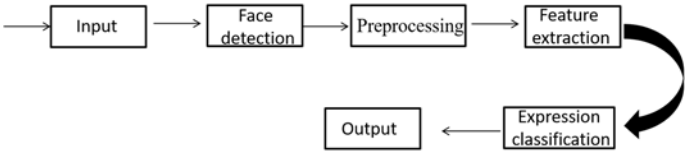

Fig. 1. Facial expression recognition process.

Face detection is the use of computer knowledge in the detection area of the existing photos or facial features in the location, shape, size and other scanning detection to determine whether there is a face in the scanning area. If there is a face, the region where the face exists is extracted. For example, the mosaic method proposed by Yang et al[1] establishes the gray distribution rules of the face region, and filters the qualified faces according to these rules. The subspace method proposed by Pentland et al[2]divides the space into principal subspace and secondary subspace. The principal subspace is used for face

recognition, and the secondary subspace is used for face detection. The projection energy of the area to be detected in the statistical secondary subspace is calculated. The smaller the distance is, the greater the probability of face recognition is. R-CNN[34] uses the second-order algorithm. In the first stage, it uses the selective search method to generate a large number of candidate regions that may be the detection object. In the second stage, it uses the convolution neural network to identify whether these candidate regions have faces. MTCNN[56] is a second-order network for face detection, which is usually composed of p-net, r-net and o-net. P-net is the recommendation network to generate candidate regions, R-Net is the perfection network to improve candidate regions, and o-net is the final output of face detection region.

Preprocessing is to input the detected image into the computer to reduce noise, eliminate interference and recover useful information. It usually includes image normalization and data enhancement. The main normalization operations are gray normalization, size normalization, pixel normalization and so on. Gray normalization is mainly used to compensate the illumination of the image to overcome the influence of illumination changes in image detection and improve the recognition efficiency. Size normalization is to adjust the image to a uniform size according to a certain proportion. Pixel normalization is to place the distribution of image pixels between 0-1, which makes it easier to converge in the subsequent network training process. Data enhancement can effectively alleviate the over fitting problem by changing the image spatial geometric position and image information while keeping the image content unchanged, generating new images and adding them to the expression database to expand the sample size. Data enhancement generally includes random clipping, random mirror, angle rotation, image radiation transformation，adding Gaussian noise to the original image, modifying the original image brightness, contrast, saturation, sharpness and other information and so on.

Feature extraction is to transform the facial features in images into numerical forms that can be recognized by computer. Lanitis et al[7] analyzed a series of facial feature points, established deformable models of global parameters for seven facial expressions, and then compared the calculated position and shape of facial feature points with the models of seven facial expressions, so as to recognize facial expressions. Coote[8] proposed ASM based on the idea of statistics, and then added texture information to ASM to propose AAM. SIFT[9], LBP, HOG[11], Haar[12] are all classic texture feature-based extraction modes. Considering that the traditional face recognition research usually inputs the whole original image into the network model, which easily leads to the loss of important face information, such as uniform and regular texture, as well as the invariance of image illumination, occlusion, scaling, rotation and so on, some researchers try to use the manually extracted face features as input to solve this problem. For example, LBP features of face are extracted as input to improve the robustness of face recognition to illumination[13], and SIFT features of face are extracted as input to improve the robustness of face pose[14]. Based on the classic CNN architecture, some researches have designed good auxiliary modules or improved the network layer to enhance the ability of network feature learning. For instance, in view of the low inter class discrimination of expressions, inspired by the center loss, two variants are proposed: Island loss (regularized to increase the center distance between different classes), locality preserving loss (LP loss) (making the local group of each class compact) to assist softmax loss to obtain more expression features[15]. According to the experience and previous studies, it is shown that using the diversity and complementarity of different networks and effective integration methods to integrate multiple networks can also improve network performance. There are many ways to generate the diversity of networks. Different training data, different training methods, different preprocessing methods, different number of neurons and different network models can generate different networks. Wen et al divided the pooled feature map of pooling layer into two paths for convolution, extracted high-level

features and low-level features of each layer, and then input them into the classifier for classification. The most commonly used integration method is to connect the features obtained from different network learning, and fuse them into a new feature to represent the image.

Expression classification is the last step in the whole process of facial expression recognition, and it is also the ultimate goal of expression recognition. It selects the appropriate classifier according to the extracted expression features, and divides the expression image into corresponding categories. Commonly used expression classification methods include SVM classification method, softmax classification method and so on. Freeman et al[17]proposed a feature extraction method for face region, and then used SVM for classification. Olson combines PCA and SVM in face recognition, which improves the accuracy of face recognition. Researchers have also proposed improved SVM methods, such as combining k-nearest neighbor method with SVM, integrating the nearest neighbor information into the construction of SVM, and proposing SVM classifier based on local; or combining CSVMT model with SVM and tree module to solve the classification sub problem with lower algorithm complexity.

2.2 Convolutional neural network

Convolutional neural network[18,19,20,21,22] is an improvement of artificial neural network. It is the most widely used network structure in the field of facial expression recognition. It mainly consists of three different structures: convolution layer, pooling layer and full connection layer.

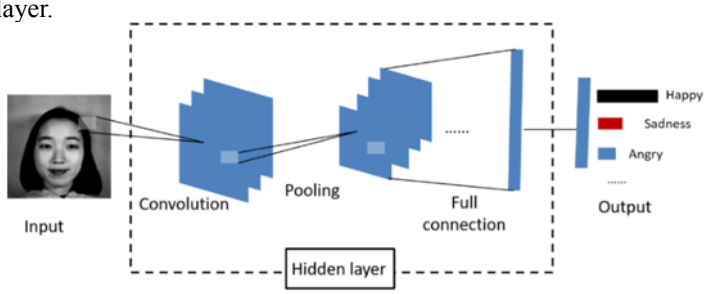

Fig. 2. Convolutional neural network structure.

Convolution layer is the key component of convolution neural network, which is usually directly connected with the input image. By processing the image pixel value, the input image is transformed into the form that convolution network can understand, and then the image features are extracted and the feature map is output. Convolution layer parameters usually include convolution kernel size, stride and padding. By moving the fixed stride of the convolution kernel, the local features in the receptive field and the weight coefficients of the convolution kernel are summed and superimposed to obtain the convolution value of the local region. After moving, various specific types of activation feature graphs are generated.

According to the output feature map of the convolution layer, the pooling layer effectively reduces the size of the image, further reduces the amount of parameters, speeds up the operation, and prevents over fitting. At the same time, the output sensitivity to displacement, tilt and other forms of deformation is reduced, and the generalization ability of the model is enhanced. Pooling layer can reduce the dimension of feature and retain the original feature information.

After the data goes through the pooling layer, it will enter the next convolution layer, extract deeper features, and then pool again. This is repeated until a certain level of features

are extracted. After feature extraction, it enters the full connection layer. The fully connected layer is usually at the end of the network to ensure that all neurons in the layer are fully connected with the active neurons in the previous layer, and the 2D feature mapping can be converted into 1D feature mapping for further feature representation and classification.

3 Facial expression recognition method based on multi branch structure

3.1 Method overview

The multi branch network designed in this paper uses Resnet-50 as the backbone network, and the network structure is shown in Figure 4 (where conv represents convolution operation, maxpool represents maximum pooling, avgpool represents average pooling, IN represents input, OUT represents output, s represents stride). After a convolution and maximum pooling, the input image enters into four groups of Conv Blocks, each of which is composed of 3, 4, 6 and 3 ResBlocks. The input and output dimensions of the first ResBlock of each group of Conv Blocks are inconsistent, so the first ResBlock uses the method on the left of ResBlock to use 1*1 convolution operation to reduce image dimension. When the dimensions of the input image and output image are consistent, the convolution operations can be concatenated. The input and output dimensions of the later ResBlocks are the same. You can directly use the method on the right side of ResBlock to concatenate convolution operations. According to the calculation formula of characteristic graph size, when the convolution kernel is 3*3 , the stride is 2 and the padding is 1, the size of the feature map after convolution is halved, so the size of the feature map after each Conv Block is halved.

Specifically, the part after Conv Block3 of Resnet-50 is divided into three branches in the multi branch network. The structures of the three branches are similar, but the down sampling rates are different. The parameters of each network layer are shown in Table 1. The leftmost branch is called global branch, which is used to extract global features. After the network layer of Conv Block5, the convolution of stride = 2 is used for down sampling, and the global maximum pooling is used to generate 2048 dimensional eigenvectors of the obtained feature graph, then compressed into 256 dimensional eigenvectors by 1*1 convolution. The middle (Part-2 branch) and right (Part-3 branch) branches are used to extract local features. Different from global branch, the image is segmented vertically from top to bottom after Conv Block5 network layer. Part-2 is divided into two blocks and Part-3 is divided into three blocks without down sampling operation. The purpose is to retain more information of the image, which is more conducive to network learning more detailed features. After segmentation, two maximum poolings are used to generate 2048 dimension feature vectors, then compressed into 256 dimension feature vectors by 1*1 convolution.

Finally, the global features and the segmented local features are input into the full connection layer to calculate the corresponding loss. The whole network not only calculates the loss of the global region of the face to fit the global spatial features of the face, but also calculates the loss of the local region of the face to fit the features of the local organs. Global features and local features complement each other to enhance the network's ability to learn expression information.

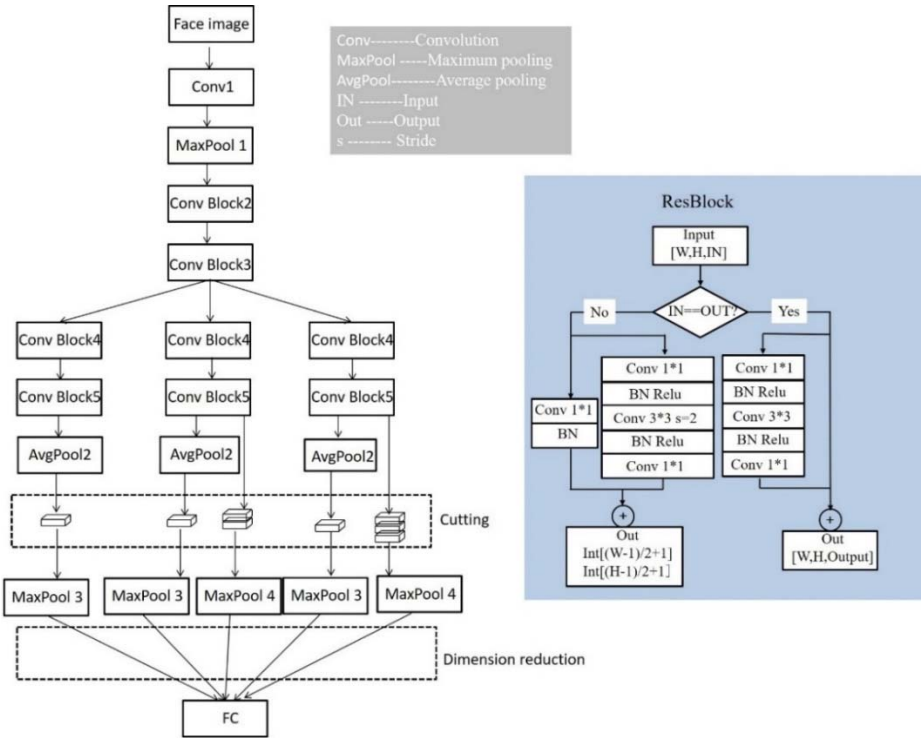

Fig. 3. Network structure diagram.

Table 1. Network parameters of each layer.

Network layer name	Nuclear parameters	Stride& Padding	Output size
Conv1	7*7	s=2,p=3	32*32*64
MaxPool 1	3*3	s=2,p=1	16*16*64
Conv Block2	[1*1; 3*3; 1*1] *3	[s=1,p=0; s=1,p=1; s=1,p=0] *3	16*16*256
Conv Block3	[1*1; 3*3; 1*1] *4	[s=1,p=0; s=2,p=1; s=1,p=0] [s=1,p=0; s=1,p=1; s=1,p=0] *3	8*8*512
Conv Block4	[1*1; 3*3; 1*1] *6	[s=1,p=0; s=2,p=1; s=1,p=0] [s=1,p=0; s=1,p=1; s=1,p=0] *5	4*4*1024
Conv Block5	[1*1; 3*3; 1*1] *3	[s=1,p=0; s=2,p=1; s=1,p=0] [s=1,p=0; s=1,p=1; s=1,p=0] *2	2*2*2048
AvgPool2	3*3	s=2,p=1	1*1*2048
MaxPool 3	1*1	-	1*1*2048
MaxPool 4	2*2	-	1*1*2048

3.2 Feature fusion method

Feature fusion is a way of fusing different features extracted from the network to form a new target. The fusion of target features is to combine the advantages of different feature

extraction, in order to achieve the purpose of complementary advantages. In this paper, the weighted fusion method is used to fuse the global and local features. As shown in the following formula 1, different features of the same position (i,j) and the same channel d are weighted and summed by setting the weighting coefficient α。Among them, the value of α is obtained by experimental verification, and different weight values are set for different features, which can better fuse different features and increase the accuracy of image classification.

$$z_d(i,j) = \alpha x_d(i,j) + (1-\alpha) y_d(i,j) \tag{1}$$

3.3 Loss function

In this method, softmax loss is used to calculate the network loss. The principle of softmax loss function is to use cross entropy to calculate the loss function of the network, and use gradient descent algorithm for error back propagation to update the weight (w) and bias (b) of the network. So the loss function can be reduced as much as possible to increase the probability of output target. As shown in the following formula 2, m represents the total number of input samples, n represents the number of categories, and w represents the network weight. $p\left\{y^i = j\right\}$ is used to calculate whether the real label and the output label are equal. If they are equal, then $p\left\{y^i = j\right\} = 1$; otherwise, the value is 0.

$$LOSS = -\frac{1}{m}\sum_{i=1}^{m}\left[\sum_{j=1}^{n} p\left\{y^i = j\right\} \log \frac{e^{w_j^T x^{(i)}}}{\sum_{z=1}^{n} e^{w_z^T x^{(i)}}}\right] \tag{2}$$

4 Experimental results and analysis

All the experiments in this paper are run on the system of Ubuntu 16.04. The CPU is Intel (R) core (TM) i5-7500, the hard disk is 1T, and the memory is 8GB. The multi branch network model designed in Section 3 is constructed and trained by using the deep learning tool of Python. The training parameters are shown in Table 2. The whole test process is implemented in Python language.

Table 2. Training parameter setting of multi granularity network in fer2013.

Parameter	Parameter size
Learning rate	0.01
Batch_size	32
Momentum	0.9
Gamma	0.1
Weight_decay	0.0003
Decay_type	step
Lr_ decay	150

4.1 Database

Fer2013 database is introduced in ICML 2013 representative learning challenge, which

contains 26190 images. All the images in the database are automatically collected by Google Image Search API. After correcting the wrong labels in the database, they are adjusted and cropped to 48 *48 gray scale images, so the resolution of the images is relatively low. There were 7 expressions corresponding to the data tag 0-6:0 anger; 1 disgust; 2 fear; 3 happy; 4 sad; 5 surprised; 6 neutral.

4.2 Preprocessing

This paper mainly uses pixel normalization to preprocess the image. After that, data enhancement techniques such as translation transformation, rotation transformation, mirror transformation and affine transformation are used to expand the sample size of data set, which can effectively alleviate the over fitting problem in the training process of deep convolution neural network.

4.3 Experimental results and analysis

Figures 4 and 5 show the confusion matrix of expression recognition results of traditional Resnet-50 and multi branch network on Fer2013 database. The diagonal of the confusion matrix is the average recognition rate of each expression, and the other results represent the degree of confusion with other expressions.

From the experimental results above, the recognition accuracy of Resnet-50 is 69.7%, while the recognition accuracy of multi branch network is 73.7%. In Resnet-50 network, the recognition accuracy of happy, fear and angry expressions is higher, which is 84.6%, 80% and 77.1% respectively. In comparison, same three expressions have 86.4%, 82.3% and 81.1% recognition accuracy using multi branch network. The main reason is that the texture changes of these three expressions are more obvious and easier to distinguish.

For the error recognition rate of different expression recognition, the accuracy of disgust is the lowest in Resnet-50 and multi branch network, which are 51.6% and 57.5% respectively. By observing the confusion matrix, we conclude that the low recognition rate of disgust is mainly due to the high proportion of identifying disgust as sad and angry. The proportion of misidentified of disgust in multi branch network was 8.94%, and the degree of confusion was the highest. This is mainly because the changes of the key parts of anger, contempt and disgust are similar. After careful observation of disgust, anger and contempt expression face image, we found that the texture changes of these three expressions near the mouth are relatively similar, and the mouth region belongs to the region which contributes a lot to expression recognition, which leads to the confusion between them.

In order to verify the effectiveness of the proposed algorithm, comparisons with other four algorithms are run on Fer2013 dataset. The results are shown in Table 3. Liu et al trained three subnets with different structures, including three to five convolution layers. Each subnet is a compact CNN model trained separately, and the output of the subnet is combined to make the integrated CNN pay more attention to face features. In our experiment, the accuracy of the best single subnet is 62.44%, and the accuracy of the whole model is 65.03%. Guo proposed a deep learning method based on deep neural network and relative learning (DNNRL), which directly learns the mapping from the original image to Euclidean space, where the relative distance corresponds to the measurement of facial expression similarity. The accuracy is 71.33% on Fer2013 dataset. Wang et al designed an architecture combining local and global information to extract different scale features from the middle layer, considering that the mouth, nose, eyes and eyebrows of human face contain representative expression information. The architecture consists of four modules (conv1, conv2, conv3 and conv4). Each module contains two or three convolution layers, and there is a maxpooling layer between each module. After conv4 module, three layers

extracted from different convolution layers are connected. Then another convolution layer is added to reduce the dimension (1*1 filter) and add two fully connected layers to get the classification results. The experimental results on Fer2013 show that the performance of this architecture is better than other methods. Through the comparison, the recognition accuracy of the proposed method in Fer2013 is 73.7%, which is 0.88% - 8.67% higher than other algorithms. It proves that the method of extracting global and local features through multiple branches can effectively improve the accuracy of facial expression recognition.

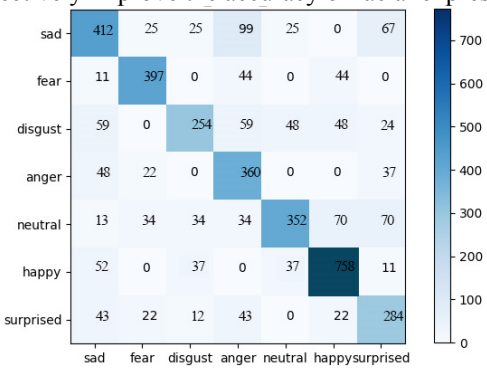

Fig. 4. Confusion matrix of traditional Resnet-50 network in fer2013 (73.7%).

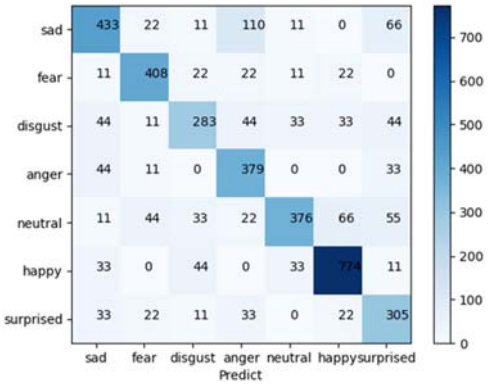

Fig. 5. Confusion matrix of multi branch network in Fer2013 (73.7%).

Table 3. Comparison of recognition accuracy of different algorithms in fer2013.

Network Model	CK+(%)
Liu, et al[27]	65.03
Guo, et al[28]	71.33
Wang, et al[29]	72.82
our work	73.7

5 Summary and Prospect

On the basis of previous studies, this paper proposes a facial expression recognition method based on multi branch structure. Three branches are designed in the network. One global branch is used to extract the global features of face. Two local branches are divided into different parts to extract the detailed features of face respectively. Finally, the global features and local features are fused in the full connection layer and output to the classifier

for classification. The experimental results on Fer2013 public dataset show that the proposed method can effectively improve the facial expression recognition rate.

Considering that facial expression recognition in real environment is vulnerable to the interference of illumination, posture, occlusion and other external environment, the accuracy of facial expression recognition is greatly reduced. Therefore, in the future work, we can collect facial expression images in the real environment for training. In addition, the network structure proposed in this paper is mainly for the training of static face images, while in the real environment, facial expressions in the form of video stream or multiple sequences are more common. Therefore, in the future, we can continue to optimize the network structure of this paper, and further study the real-time recognition and classification of dynamic face sequences.

Acknowledgement

This work was supported by the Science Foundation of State Grid (grant number 5211XT190033).

Reference

1. Yang G Z, Huang T S. Human face detection in a complex back2 ground[J]. Pattern Recognition, 1994, 27(1): 53 - 63.
2. Moghaddam B, Pentland A. Probabilistic visual learning for object recognition[J]. Pattern Analysisand Machine Intelligence, 1997, 19(7): 696-710.
3. Ren S, He K, Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
4. Girshick R. Fast R-CNN[J]. Computer Science, 2015.
5. Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.
6. Zhang Y, Lv P, Lu X, et al. Face detection and alignment method for driver on highroad based on improved multi-task cascaded convolutional networks[J]. Multimedia Tools and Applications, 2019, 78(18).
7. Lanitis A, Taylor C J, Cootes T F. Automatic interpretation and coding of face image using flexible models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 743-756.
8. Cootes T F, Wheeler G V, Walker K N, et al. View-based active appearance models[J]. Image & Vision Computing, 2002, 20(9-10): 657-664.
9. Lekdioui K, Messoussi R, Ruichek Y, et al. Facial decomposition for expression recognition using texture/shape and SVM classifier[J]. Signal Processing: Image Communication,2017.
10. Olson. Design and improvement of face recognition system based on SVM[J]. Network security technology and application, 2019 (12)
11. Lowe D G. Object Recognition from Local Scale-Invariant Features//Proceedings of the International Conference on Computer Vision.Corfu,GREECE,1999:1150-1157.
12. Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions//Proceedings of the 12th International Conference on Pattern Recognition. Beijing, China, 1994: 582-585.

13. Levi G, Hassner T. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns[C]. ACM on International Conference on Multimodal Interaction. New York, USA, 2015: 503-510.
14. Zhang T, Zheng W, Cui Z, et al. A deep nural network driven feature learning method for multi-view facial expression recognition[J]. IEEE Transactions on Multimedia, 2016:1-1.
15. Zhang C,Wang P, Chen K, et al. Identity-aware convolutional neural networks for facial expression recognition[J]. Systems of Engineering and Electronics Journal, 2017, 28(4):784-792.
16. Wen Y M, OU YW, Ling YQ. Expression recognition oriented dual channel convolutional neural network [J]. Computer engineering and design, 2019, 40 (7): 46-52
17. Freeman W T, Roth M. Orientation histograms for hand gesture recognition//Proceedings of the International workshop on automatic face and gesture recognition. Zurich, Switzerland, 1995: 296-301.
18. Huang A, Abugharbieh R, Tam R. A novel rotationally invariantregion-based hidden Markov model for efficient 3-D image segmentation[J]. IEEE Trans on Image Processing, 2010, 19(10): 2737-2748.
19. Choy S K, Tong C S. Statistical wavelet subband characterization based on generalized gamma density and its application in texture retrieval[J]. IEEE Trans on Image Processing, 2010, 19(2): 281-289.
20. Ren Jianfen, Jiang Xudong, Yuan Junsong. Noise resistant local binary pattern with an embedded error-correction mechanism[J]. IEEE Transon Image Processing. 2013, 22(10): 4049-4060.
21. Tan X Y, Bill T. Enhanced local texture feature sets for face recognition under difficult lighting conditions[J]. IEEE Trans on Image Processing, 2010, 19(6): 1635-1650.
22. Akata Z, Perronnin F, Harchaoui Z, et al. Good practice in large-scale learning for image classification[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2014, 36(3): 507-520.
23. Lai C. Analysis of activation function in convolutional neural networks [J]. Science and technology innovation, 2019 (33): 35-36
24. Zhang S, Gong Y H, Wang J J. Development of deep convolutional neural network and its application in computer vision. Acta computa Sinica, 2019, 42 (3): 453-482
25. Zhou F Y, Jin L P, Dong J. review of convolutional neural networks. Acta computa Sinica, 2017, 40 (6): 1229-1251
26. Gulcehre C,Moczulski M, Denil M , et al. Noisy Activation Functions[J]. 2016.
27. Liu K, Zhang M, Pan Z. Facial expression recognition with CNN ensemble[C]. International Conference on Cyberworlds. Chongqing: IEEE Computer Society, 2016: 163-166.
28. Guo Y, Tao D, Yu J, et al. Deep Neural Networks with Relativity Learning for facial expression recognition[C]. IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Seattle:IEEE, 2016:1-6.
29. Wang J, Yuan C. Facial expression recognition with multi-scale convolution neural network[C]. Pacific Rim Conference on Multimedia. Xi'an: Springer, 2016: 376-385.