Facial Expression Recognition Method Based on Convolution Neural Network Combining Attention Mechanism

Peizhi Wen[1], Ying Ding[1], Yayuan Wen[2], Zhenrong Deng[1](✉), and Zhi Xu[1]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
zhrdeng@guet.edu.cn

[2] College of Electronic Engineering, Guangxi Normal University, Guilin 541004, China
wenyy@gxnu.edu.cn

Abstract. Facial expression recognition is one of the research hotspots in the field of computer vision. Aiming at the problems that the current machine learning method extracts facial features are less robust and the traditional convolutional neural network can not fully extract the expression features, a residual network model integrating CBAM attention mechanism is proposed. Given an intermediate feature map, a attention map is generated on the channel domain and the spatial domain of the feature map respectively by our module, and multiplied by the original feature map to obtain a recalibrated feature map. In the training process, the improved loss function A-Softmax is used to generate the angular interval by manipulating the feature surface, so that the different class features learned have angular intervals in the angular space, and the similar features are more closely clustered. Experiments on FER2013 and JAFFE dataset show that the proposed method effectively improves the feature expression ability of the network, enhances the ability to distinguish different facial expression features, and achieves good recognition performance.

Keywords: Facial expression recognition · Residual network · Attention mechanism · CBAM · A-Softmax

1 Introduction

Human emotion information is mainly expressed through rich facial expressions. With the rapid development of artificial intelligence, intelligent computing technology is becoming more and more popular in people's lives. The recognition of facial expressions has more and more obvious application research value in human-computer interaction, interactive games, wisdom education and criminal investigation. Facial expression recognition can make computer understand human emotion better, make human-computer interaction not only stay at the level of instruction interaction, but also help computer to move forward to the level of intelligent interaction.

This work was supported by the National Natural Science Foundation of China (No. 61662014).

X. Sun et al. (Eds.): ICAIS 2020, CCIS 1253, pp. 136–147, 2020.
https://doi.org/10.1007/978-981-15-8086-4_13

In the traditional machine learning research, the feature extraction of expressions has always been a difficult problem. Most of them use the method of artificially extracting features, and then use the artificially extracted features to train the shallow classifier to classify the expressions. Classic expression extraction methods include Histograms of Oriented Gradients (HOG [1]), Gabor wavelet transform [2], Local Binary Pattern (LBP [3]).

However, the manual feature extraction method can't explain the expression information efficiently, and the human interference factors will directly affect the feature extraction, so the shallow classifier trained by this method will have the problem of insufficient generalization ability. And in complex environments, such as the intensity of light, whether there is occlusion or posture transformation, the traditional machine learning method would be less robust.

In recent years, neural networks technology have been developed rapidly and have shown great advantages in the field of facial expression recognition. Neural networks automatically extract and learn the characteristics of samples, and can be classified by classifiers. This not only gets rid of the cumbersomeness of the artificial extraction feature, but also greatly improves the accuracy of the recognition and the robustness of the algorithm.

After that, the neural network is also used in facial expression recognition. For example, Tang [4] proposed to combine Convolutional Neural Networks (CNN [5]) with Support Vector Machine (SVM [6]), and gave up the cross entropy loss minimization method used by ordinary CNN, instead of using standard hinge loss to minimize margin-based loss. His method achieved a 71.2% recognition rate on the private test set and won the FER2013 [7] Face Expression Recognition Challenge. Zhang [8] adopted a stacked hybrid self-encoder expression recognition method. The network structure is composed of a Denoising AutoEncoder (DAE), a Sparse Auto-Encoder (SAE), and an Auto-Encoder (AE). The feature extraction is performed by DAE. SAE are cascaded to extract more abstract sparse features. Experiments show that the average recognition rate on JAFFE [9] dataset reaches 96.7%. Pramerdorfer [10] et al. analyzed the structural defects of facial expression recognition in deep convolutional neural networks, improved the structure of the classical convolutional neural network, and extracted the features of the face with a single face image as input, in the FER2013 dataset. A 72.4% recognition rate was achieved.

In addition, the combination of attention mechanism [11] and image processing further enhances the performance of the network. Attention mechanism can ignore irrelevant information and focus on effective information. For example, Jaderberg [12] and others proposed a spatial transformer module through the attention mechanism, and the spatial domain information of the image is transformed into a corresponding space, thereby extracting the region of interest in the image. Hu Jie et al. [13] proposed a novel architecture unit Squeeze-and-Excitation (SE) module. The SE module starts with the relationship between feature channels and adopts a new feature recalibration strategy. The way to automatically obtain the importance of each feature channel, and then to enhance the useful features according to this degree of importance and suppress features that are of little use to the current task.

To sum up, this paper proposes a convolution neural network method for facial expression recognition based on attention mechanism. Firstly, the depth residual network - residual neural network (ResNet [14]) is used as the basic model to avoid the gradient disappearance, the network model is too large and the accuracy is reduced with the deepening of the network. Then, the attention mechanism module of Convolutional Block Attention Module (CBAM [15]) is introduced to generate attention on the channel domain and the spatial domain of the feature map at the same time. That is to say, for a given feature map, CBAM module will calculate attention through independent learning along channel dimension and spatial dimension. Then, the attention map and the input feature map are multiplied to get a new and more detailed feature map, which improves the feature expression ability of the network model without significantly increasing the amount of parameters and calculation. At the same time, the improved loss function Angular Softmax [16] (A-Softmax) is used to manipulate the feature surface to generate the angular interval, so that the convolutional neural network can learn the angle discrimination feature, so as to increase the distance between classes, reduce the distance within classes, and further improve the network classification effect. The experimental results show that the method can prevent the gradient from disappearing, fully extract facial expression features, and effectively improve the expression recognition rate.

The remainder of the paper is organized as follows. In Sect. 2, we introduce the main techniques used in this paper; Sect. 3 introduce the methods we proposed in detail; Sect. 4 discusses the experiment setup and results. Section 5 concludes the work.

2 Related Technology

2.1 Residual Neural Network (ResNet)

ResNet was proposed in Microsoft Research Institute in 2015 and won the title of classification task in Imagenet competition. ResNet can quickly speed up the training of neural network to prevent the gradient from disappearing or exploding due to the deepening of network layers. It is also widely used in image recognition, segmentation and other fields. ResNet's main idea is to use residual learning to add identity short connection in the network and connect the original input directly to the later network layer. The residual learning module is shown in Fig. 1. The input is *x*, the output of fast connection is also *x*, and *H(x)* is the ideal mapping. Originally, the learning is *h(x)* only obtained through convolution layer. Now, the learning is the part of the difference between the input and output, i.e. the residual $H(x) - x$, which effectively prevents information loss and loss in information transmission. ResNet's residual module is divided into two types. The first one is a residual block with two 3 * 3 convolution layers connected in series, and the second is a residual block with two 1 * 1 convolution layers and one 3 * 3 convolution layer connected in series.

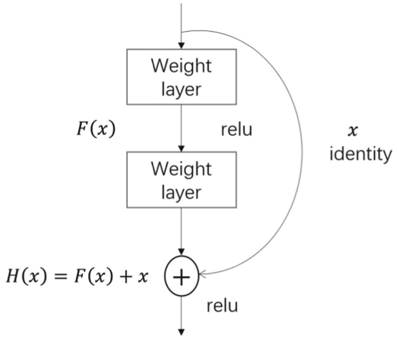


Fig. 1. Basic structure of residual network.

2.2 Convolutional Block Attention Module (CBAM)

In order to improve the feature extraction ability of the network model, this paper introduces the attention mechanism, which is to add the efficient attention module-CBAM to the network. CBAM is divided into a Channel Attention Module and a Spatial attention module. The overall structure is shown in Fig. 2.

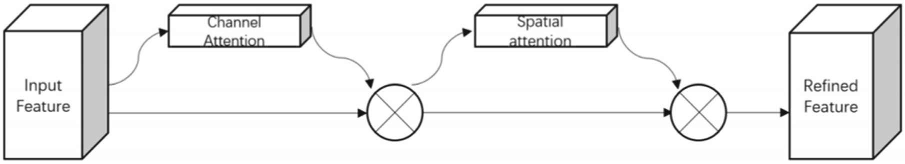


Fig. 2. The overview of CBAM.

The channel attention module is similar to the structure of SeNet, and the structure diagram is shown in Fig. 3. Firstly, the given feature map is compressed along the spatial dimension to get a one-dimensional vector and each two-dimensional feature channel becomes a real number, which has a global receptive field to some extent, and the output dimension matches the input feature channel number. It represents the global distribution of the response on the characteristic channel. The difference with senet is that the average pooling is not only considered in the spatial dimension compression of the input feature map, but also Max pooling is introduced as a supplement. After two pooling functions, two one-dimensional vectors can be obtained. Global average pooling has feedback to every pixel on the feature map, while global Max pooling only has gradient feedback in the most responsive part of the feature map, which can be a supplement to gap. The specific methods are as follows: input feature map as F, and F^c_{avg} and $F^c_{\max}$ represent the features calculated by global average pooling and global Max pooling respectively. W_0 and W_1 represent the two-layer parameters in the multi-layer perceptron model. The calculation in this part can be expressed as follows:

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$

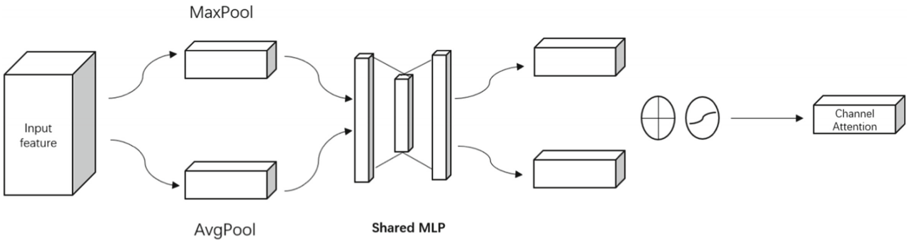


Fig. 3. Diagram of channel attention module.

$$= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max})))\tag{1}$$

Where σ represents the sigmoid activation function to obtain a normalized weight between 0 and 1. MLP stands for multi-layer perceptron, and the feature between W_0 and W_1 in the perceptron model needs to be processed using ReLU as an activation function. ReLU function can largely solve the gradient dissipation problem of back propagation algorithm in optimizing neural network. In addition, the spatial attention module extracts attention from the spatial domain. Its structure is shown in Fig. 4. First, it compresses the input feature map through average pooling and Max pooling, which is different from channel attention. Here, it compresses along the channel dimension, averaging and maximizing the input features on the channel dimension. Finally, two two-dimensional features are obtained, which are spliced together according to the channel dimensions to obtain a feature map with two channels, and then a hidden layer containing a single convolution kernel is used to convolute it, so as to ensure that the final feature is consistent with the input feature map in the spatial dimension. Define the feature map after the max pooling and average pooling operations, $F^s_{avg} \in \mathbb{R}^{1*H*W}$ and $F^s_{max} \in \mathbb{R}^{1*H*W}$. The mathematical formula of this part is as follows:

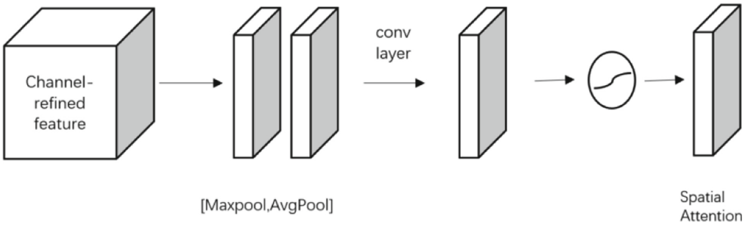


Fig. 4. Diagram of spatial attention module.

$$\begin{aligned}M_S(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F^s_{avg}; F^s_{\max}]))\end{aligned}\tag{2}$$

Where σ represents the sigmoid activation function and uses a 7 * 7 convolution kernel for extraction. CBAM is an efficient attention module that can be embedded into the mainstream CNN model. In this paper, we embed the CBAM module multiple times in the ResNet-based network model to extract facial expressions and obtain significant features.

3 Convolutional Neural Network Combining CBAM

3.1 Network Model Building

This paper takes ResNet as the basic network model, improves ResNet and introduces CBAM attention module to improve the feature extraction ability of the network. The schematic diagram of the network model is shown in Fig. 6. The network consists of 17 convolution layers, an average pooling layer and a full connection layer (FC). Among them, the convolution kernel is 3 * 3. After each convolution layer, Batch Normalization (BN [17]) and ReLU [18] are used. In addition, eight groups of residual modules (ResBlock) are set. Each residual module is composed of two 3 * 3 convolutions in series. The residual module is shown in Fig. 5.

Each two residual modules form a layer. The number of convolution channels in each layer is 64, 128, 256, 512. Each group of residual module is added with a CBAM module to re-calibrate the feature, and then the dimension is adjusted to 1 * 1 through an average pooling layer with a window size of 4 and a stride size of 4. Finally, a 7-dimensional fully connected layer is connected and seven facial expressions are classified using the A-softmax loss function.

3.2 Improved Loss Function

At present, due to the complex environment factors such as illumination, occlusion, posture deviation, etc., facial expressions of the same kind have large differences, while expressions of different classes have small differences. The traditional softmax loss function cannot solve this problem. Therefore, this paper proposes an improved loss function-Angular softmax (A-softmax) to increase the distance between classes and reduce the distance within classes, thus further improving the accuracy of expression recognition. When A-softmax was first proposed, it was applied to face recognition and won the first place in the 2017 MegaFace dataset recognition rate. Facial expression

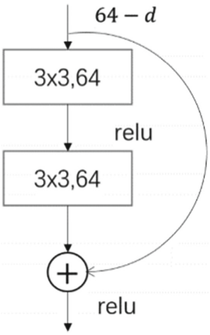


Fig. 5. ResBlock diagram.

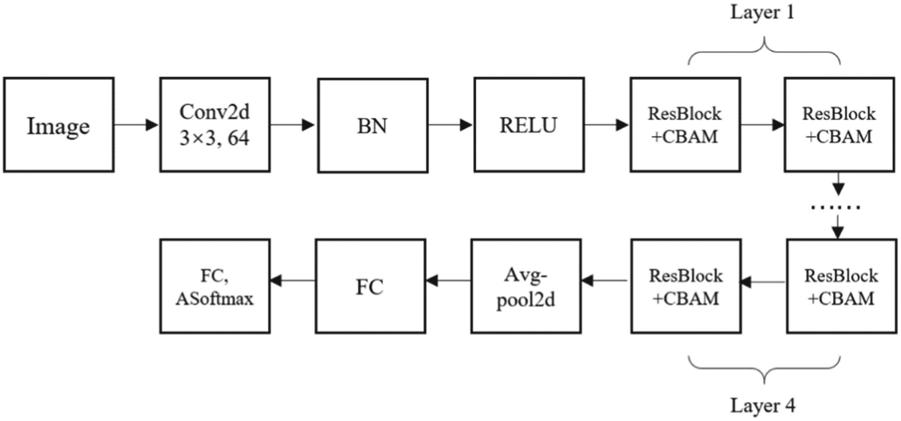


Fig. 6. The schematic diagram of the network.

recognition is essentially a multi-class problem. It is assumed that the feature vector of the last fully connected layer output is x_i, the corresponding label is y_i, and the A-softmax loss function is expressed as

$$L_{ang} = \frac{1}{N}\sum - \log\left(\frac{e^{||xi||\psi(\theta_{yi,i})}}{e^{||xi||\psi(\theta_{yi,i})}+\sum_{j\neq y_i} e^{||xi||\cos(\theta_{j,i})}}\right) \tag{3}$$

Where we define $\theta_{yi,i}$ as the angle between vector w_{yi} and vector x_i, $\Psi(\theta_{y_i,i}) = -1^k \cos(m\theta_{y_i,i}) - 2k$, $\theta_{y_i,i} \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right]$ and $k \in [0, m-1]$, m is an integer greater than or equal to 1, which is used to control the size of angle interval. On the basis of softmax, A-softmax increases the limit of angle by m times. At the same time, the weight (w) and bias (b) of full connection layer are set to 1 and 0 respectively. Therefore, the classification process depends on the angle between weight and bias. The loss function of A-softmax takes angle as the measurement standard of distance, which can learn the characteristics of angle boundary, and combine angle characteristics with learned characteristics. It improves the feature recognition ability of different facial expressions.

4 Experimental Results and Analysis

4.1 Experimental Environment

The hardware environment of this experiment is CPU E5-2698 V4, GPU Nvidia P100, memory 32G, video memory 16G. Ubuntu 16.04 is the operating system in the software environment, and python 3.6 and pytorch deep learning framework are used for programming.

4.2 Dataset Introduction

In this experiment, Fer2013 and JAFFE public datasets are used to train the network and analyze the experimental results. The Fer2013 dataset is the official dataset of the

2013 kaggle facial expression recognition competition. Because most of the pictures are downloaded from the online crawler, including pictures of different ages, different angles, partial occlusion, etc., it will be closer to the facial expression in the natural environment, but there will be some errors. Fer2013 contains 35887 pictures in total, and the dataset is divided into three parts: training set, public test set and private test set. The training set contains 28709 pictures, and both the public test set and private test set contain 3589 pictures. Fer2013 divides facial expressions into seven basic expressions: anger, disgust, fear, happiness, sadness, surprise and neutrality.

The JAFFE dataset is a dataset used by Japanese ATR institutions to study expression recognition. The dataset selects 10 Japanese women, each of which makes 7 basic expressions, including a total of 213 images. The JAFFE dataset are all face portraits. And the position in the image is roughly the same, the face size is basically the same, and it belongs to the standard expression data set.

4.3 Data Preprocessing and Data Enhancement

In the experiment, the images of the two datasets are normalized to 48 * 48 pixels. In order to enhance the robustness of the model to noise and slight transformation, data enhancement is used. In the training stage, we randomly cut the image into 44 * 44 pixels, and horizontally flip the image according to the probability of 0.5. In the test stage, we cut the image up, down, left, right and center, and then all horizontally flip, so that we can get 10 pictures, the size of the dataset will become 10 times the original, and the test accuracy is to average 10 pictures.

4.4 Experimental Results Analysis of Facial Expression Recognition

The experiment adopts the method of this paper and the existing methods to train, test and compare the two datasets respectively. Firstly, 28709 training sets in FER2013 dataset are used to train the model, 3589 public test sets are used as verification sets to adjust the super parameters of the model, and finally 3589 private test sets are used to test the model. During the training, we randomly initialize the weight and bias and perform 300 epoch training with the batchsize set to 64. The model pre-trains FER2013 and then adjusts the parameters of the JAFFE dataset, which can speed up the network convergence and improve the overall performance of the model. Due to the small amount of JAFFE data, we used a 5-fold crossvalidation method to train and test the network. The dataset is divided into five parts, four of which were taken as training set in turn, and the remaining one is used as test set validation. The average of the five results is used as the accuracy estimation of the entire dataset. All networks train only 30 epochs to prevent overfitting caused by the limited training set. In order to prove the effectiveness of the proposed method, the proposed model and a series of other models are compared on two datasets. The average recognition rate of different models on two datasets is shown in Table 1.

In this paper, we first make a comparative experiment on whether to add CBAM to ResNet. From Table 1, we can see that after adding CBAM, the recognition rate on FER2013 reach 73.1%, which is 0.7% higher than that of ResNet [10], and 1.9% higher than that of Tang [4], the champion of 2013 kaggle facial expression recognition competition, the recognition rate of JAFFE is 98.4%, 0.8% higher than that of ResNet [10],

Table 1. Comparison of average recognition rate of different models on two datasets

Model	FER2013	Model	JAFFE
Tang [4]	71.2%	Zhang [8]	96.7%
Resnet [10]	72.4%	Resnet [10]	97.6%
CBAM+ResNet+softmax	73.1%	CBAM+ResNet+softmax	98.4%
CBAM+ResNet+Asoftmax (Our)	73.5%	**CBAM+ResNet+Asoftmax (Our)**	98.9%

and higher than that of other classical models, which proves that CBAM can effectively improve the network's feature expression ability and recognition rate. In addition, this paper makes a comparative experiment between the non improved loss function and the improved loss function. It can be seen that compared with the traditional softmax loss function on the two dataset, the replacement of the Asoftmax loss function improves the recognition rate by 0.4% and 0.5% respectively, which proves that the Asoftmax loss function can effectively improve the feature recognition ability of expression and then improve the recognition rate. The method in this paper achieves 73.5% and 98.9% on FER2013 and JAFFE respectively, and achieves good recognition performance. Finally, this paper makes a comparative experiment on the sequence of channel attention module and spatial attention module in the network, and the experimental results are shown in Table 2.

Table 2. Comparison of average recognition rate of attention module sequence

Description	FER2013	JAFFE
Channel-first (Our)	73.5%	98.9%
Spatial-first	73.4%	98.7%
Channel and spatial in parallel	73.2%	98.6%

It can be seen from Table 2 that the channel-first order can achieve the best recognition effect. Table 3 and Table 4 show the confusion matrix of this model on Fer2013 and JAFFE dataset respectively. Table 3 shows that the recognition rate of happiness is the highest, reaching 91%, followed by the expression of surprise reaching 83%. However, the recognition rate of fear, sadness, neutrality and anger is lower than 70%. It can be seen from the Table 2 that the four expressions are easy to be confused. For example, anger and neutrality are easy to be confused into sadness, and fear and sadness are easy to be confused with each other.

Figure 7 shows an example of these four expressions with low recognition rate in Fer2013 dataset. It can be seen that in real life, these four kinds of expressions are very similar in themselves, especially when they don't know each other, and it is difficult to distinguish these expressions manually. As can be seen from the confusion matrix in Table 4, the recognition rate of the model in this paper has reached a high level, and the

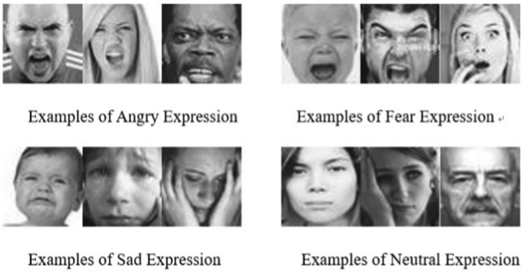
Table 3. Confusion matrix of FER2013 recognition result

Real expression categories	Predict expression categories						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.68	0.00	0.10	0.03	0.11	0.02	0.06
Disgust	0.11	0.75	0.05	0.01	0.02	0.04	0.01
Fear	0.08	0.00	0.63	0.02	0.13	0.08	0.05
Happy	0.00	0.00	0.01	0.91	0.02	0.02	0.03
Sad	0.08	0.00	0.14	0.02	0.64	0.01	0.11
Surprise	0.01	0.00	0.08	0.04	0.02	0.83	0.03
Neutral	0.05	0.00	0.06	0.06	0.15	0.01	0.67

Table 4. Confusion matrix of JAFFE recognition result

Real expression categories	Predict expression categories						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.98	0.01	0.01	0.00	0.00	0.00	0.00
Disgust	0.00	0.98	0.01	0.00	0.01	0.00	0.00
Fear	0.00	0.01	0.99	0.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sad	0.00	0.01	0.00	0.00	0.98	0.01	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Neutral	0.00	0.00	0.01	0.01	0.00	0.00	0.98

overall recognition rate has reached 98.9%, among which the recognition rate of happy and surprised has reached 100%. This is because the JAFFE is a dataset specially used to study expression recognition, expression calibration standards, and all are positive faces.



Examples of Angry Expression

Examples of Fear Expression ↵

Examples of Sad Expression

Examples of Neutral Expression

Fig. 7. Four examples of confusing expression.

5 Conclusion

In this paper, we propose a residual network model which integrates the attention mechanism of CBAM. In this model, attention mechanism is introduced into the channel domain and the spatial domain of the feature graph, which effectively improves the feature expression ability of the model. The improved loss function A-softmax is used to increase the distance between expression classes and reduce the distance within expression classes. Experiments are carried out on Fer2013 and JAFFE expression dataset to verify the effectiveness of this method. The experimental results show that the method proposed in this paper achieves better expression recognition effect. In the future work, we will further improve the basic residual network structure, so that the network has better feature extraction ability, and further improve the expression recognition rate.

References

1. Dahmane, M., Meunier, J.: Emotion recognition using dynamic grid-based HoG features. In: IEEE International Conference on Automatic Face & Gesture Recognition & Workshops. IEEE Gesture Recognition (FG 2011). Face and Gesture 2011, Santa Barbara, CA, USA, 21–25 Mar 2011, pp. 884–888. IEEE (2011)
2. Wold, S.: Principal component analysis. Chemometr. Intell. Lab. Syst. **2**(1), 37–52 (1987)
3. Guoying, Z., Matti, P.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
4. Tang, Y.: Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239 (2013)
5. Fang, W., Zhang, F., Sheng, V.S., Ding, Y.: A method for improving CNN-based image recognition using DCGAN. CMC: Comput. Mater. Continua **57**(1), 167–178 (2018)
6. Li, R., Liu, Y., Qiao, Y., Ma, T., Wang, B., Luo, X.: Street-level landmarks acquisition based on SVM classifiers. CMC-Comput. Mater. Continua **59**(2), 591–606 (2019)
7. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. In: International Conference on Neural Information Processing, pp. 117–124. Springer (2013)
8. Zhang, Z.Y., Wang, R.Q., Wei, M.M.: Stack hybrid self-encoder facial expression recognition method. Computer Engineering and Application, pp. 1–8, 09 April 2019. http://kns.cnki.net/kcms/detail/11.2127.tp.20180920.1759.010.html
9. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with Gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205. IEEE (1998)
10. Pramerdorfer, C., Kampel, M.: Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903 (2016)
11. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
12. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

15. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
16. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SphereFace: deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 212–220 (2017)
17. Ioe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)