Facial Expression Recognition Based on Multi-feature Fusion

Zhuang Miao, Jingyu Li(✉), and Kezheng Lin

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
920948105@qq.com

Abstract. In order to solve the problems of insufficient facial expression feature extraction and large parameter amount in some convolutional neural networks, a facial expression recognition algorithm based on multi-feature fusion is proposed. This method first modifies the residual block in the ResNet network, reduces the amount of network parameters and uses pre-activation to reduce the error rate. After that, the features extracted by the improved ResNet network are fused with the features extracted by the VGG network after the cut layer, and the network model P-ResNet-VGG is obtained. The loss function uses the cross entropy loss function. This model has been extensively tested on the FER2013 and JAFFE datasets. The experimental results show that this model has improved accuracy on the expression data set than other models, and it has a significant effect on the FER2013 and JAFFE data sets.

Keywords: Deep learning · Convolutional neural network · Facial expression recognition · Feature fusion

1 Introduction

Nowadays, there are many ways that humans express their emotions, but the most direct way is to judge them by their facial expressions. Guess the other person's mental activities and emotions based on their facial expressions. Facial expression recognition is one of the hot spots in the direction of computer vision. Its application field is also very extensive. Including man-machine interaction, safe driving, intelligent monitoring, assisted driving, case detection, etc.

Early facial expression recognition research is based on hand-made features [1]. In the ImageNet large-scale visual recognition competition [2], the success of the AlexNet [3] network model has made deep learning widely used in the field of computer vision. Facial expression recognition (FER) challenge [4] proposed the use of deep learning in the early days. The best performance in the 2013 FER Challenge was the deep convolutional neural network [5], and the manual feature model only ranked fourth [6]. In 2017, Tang Chuan Gao et al. [7] used deep learning to win the championship in the field of facial expression recognition. Later, the convolutional neural network has achieved good

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
S. Shi et al. (Eds.): 6GN 2021, LNICST 439, pp. 264–275, 2022.
https://doi.org/10.1007/978-3-031-04245-4_23

results in the field of facial expression recognition, but with the deepening of research, problems have also been discovered.

The network based on the attention model proposed by Chu et al. [8] achieved an accuracy of 97.45% on the CK+ data set. Lu et al. [9] designed a 7-layer CNN to perform expression recognition on CK+ and only achieved an accuracy of 81.5%. It can be inferred that too few network layers, single network model extraction features, etc. are the reasons for the lower recognition rate. Later, researchers try to combine multiple feature fusion [10–12] or multiple network models [13–15] for expression recognition In order to achieve a higher accuracy rate.

Therefore, this paper proposes a network model based on multi-feature fusion. First, the residual block in the ResNet network is improved, the number of network layers is modified, and the pre-activation method is introduced. After that, the network is combined with the modified VGG network to obtain a P-ResNet-VGG (Pre-activated residual network and Visual Geometry Group) dual-channel network structure. The fused network structure extracts more feature information and can improve The recognition rate of the sample reduces the time cost of the model training in the expression recognition training process.

2 Deep Learning Model

2.1 VGG Network Model

The VGG network uses a continuous 3 × 3 convolution kernel instead of a larger convolution kernel. For a given receptive field, it is better to use multiple small convolution kernels. Non-linear operations can be achieved through the activation function, which can train more Good network structure, and the cost will not increase. The activation function selects the R-ReLU function, which can make up for the shortcomings of the RelU function. Its function is.

$$RRelu(x) = \begin{cases} x, & x > 0 \\ ax, & x < 0 \end{cases} \tag{1}$$

The advantage of using R-ReLU compared to ReLU is that the negative values in the ReLU function are all zero. If a large gradient flows through the RelU neuron to update the gradient, this neuron will lose activation of some data, and the gradient will change. 0 leads to neuron death. The a in the RReLU function is a value randomly selected from a given uniform distribution range, and it will be fixed during the test. Solved the problem that the ReLU function may cause neuron death.

2.2 ResNet Network Model

The ResNet network adds the concept of residuals to the convolutional neural network. After several layers of convolutional layers, shortcut connections are used. This allows you to return to the previous shallow network when the increase in depth causes the accuracy of the model to decrease. This will find the optimal number of network layers

during the training process, and will not cause the model to be worse due to the complexity of the network settings.

The residual network is composed of multiple residual blocks, and the residual blocks need to be fitted to the residual mapping $f(\boldsymbol{x}) - \boldsymbol{x}$ related to the identity mapping. The residual mapping is easier to optimize in practice, and the residual block is shown in Fig. 1. The residual unit formula can be expressed as

$$\begin{cases} y_l = h(x_l) + F(x_l, W_l) \\ x_{l+1} = f(y_l) \end{cases} \tag{2}$$

Among them, x_l and x_{l+1} respectively represent the input and output of the l residual unit. F is the residual function, $h(x_l) = x_l$ is the identity mapping, and f is the RRelu activation function. Based on formula (2), the learning characteristics from shallow l to deep L can be obtained as

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{3}$$

Using the chain rule, the gradient of the reverse process can be obtained

$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, W_i)\right) \tag{4}$$

Among them, $\frac{\partial loss}{\partial x_L}$ represents the gradient of the loss function to L, and 1 can ensure that the gradient will not disappear, so that the residual network training will be easier.

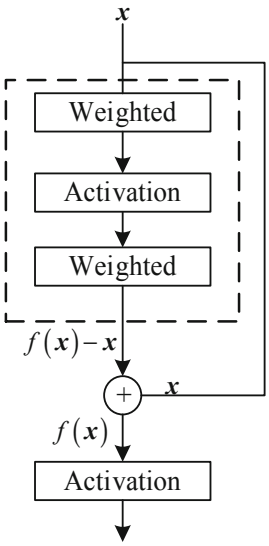

Fig. 1. Residual block

3 P-ResNet-VGG Network Structure

3.1 Improved ResNet Network

In order to extract richer facial features in the image and improve the accuracy of facial expression recognition, the residual block in the ResNet network is improved. The improvement to the network is to change the residual block to a three-layer convolutional layer, and one convolution kernel before and after is a 1×1 convolutional layer. The size of the convolution kernel of the middle convolutional layer has not changed, thus adding a convolution Operation, and the amount of network parameters is greatly reduced.

On this basis, in order to prevent the gradient from disappearing, alleviate the occurrence of overfitting, and enhance the nonlinear expression ability of the network, the RRelu activation function is added after each convolution. As shown in Fig. 2.

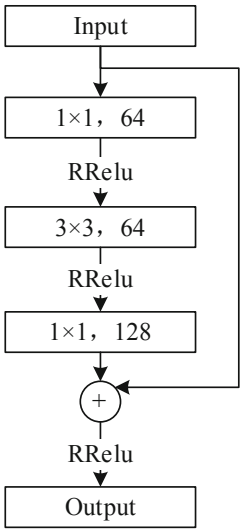

Fig. 2. Modified residual block

At the same time, the BN layer is also used for batch normalization, which can make the training of deeper neural networks easier, speed up the convergence of the network model, and improve the accuracy of the trained model. After the BN layer and the activation layer are mentioned before the convolutional layer, pre-activation can be achieved. The modified P-ResNet network will be faster than the original ResNet network in training speed, and the error will be reduced, which is more helpful for deeper networks, such as As shown in Fig. 3

Then the modified residual block is added to the original network, the input is the original picture, and the feature map is generated after multiple residual block processing, and then passed to the subsequent fully connected layer for classification processing.

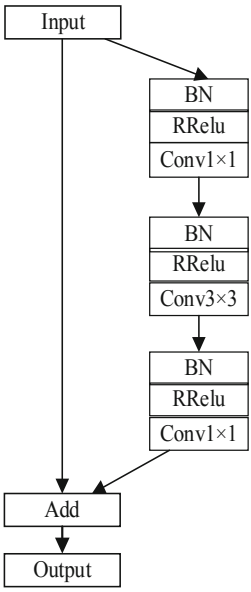

Fig. 3. Preactivation flow chart

3.2 Loss Function

The loss value of the prediction result can be obtained by bringing the feature vector and label value mentioned by the neural network into the loss function. And through the back propagation of the loss value to optimize the gradient, the commonly used multi-class loss function is the cross entropy loss function

$$L_1 = -\sum_{i=1}^{N} y^{(i)} log\ \hat{y}^{(i)} + \left(1 - y^{(i)}\right) log\left(1 - \hat{y}^{(i)}\right) \tag{5}$$

Among them, N represents the total number of samples, $y^{(i)}$ represents the output value of the i sample forward propagation, $\hat{y}^{(i)}$ represents the probability that i samples are positive samples, and L_1 represents the total loss value.

3.3 Overall Network Structure

An increase in network depth means a larger amount of parameters, longer training time, and more difficult optimization. Therefore, the overall network structure is to cut the number of VGG19 layers, and then merge with the P-ResNet network. Then the shallow information and the deep information are combined and input to the next convolutional layer, so that the extracted feature information can be more complete. Such a network structure can better obtain image features that are conducive to classification without increasing the training time. Compared with the features extracted from a single channel, the fused features are easier to match the real tags, and the recognition effect is better.

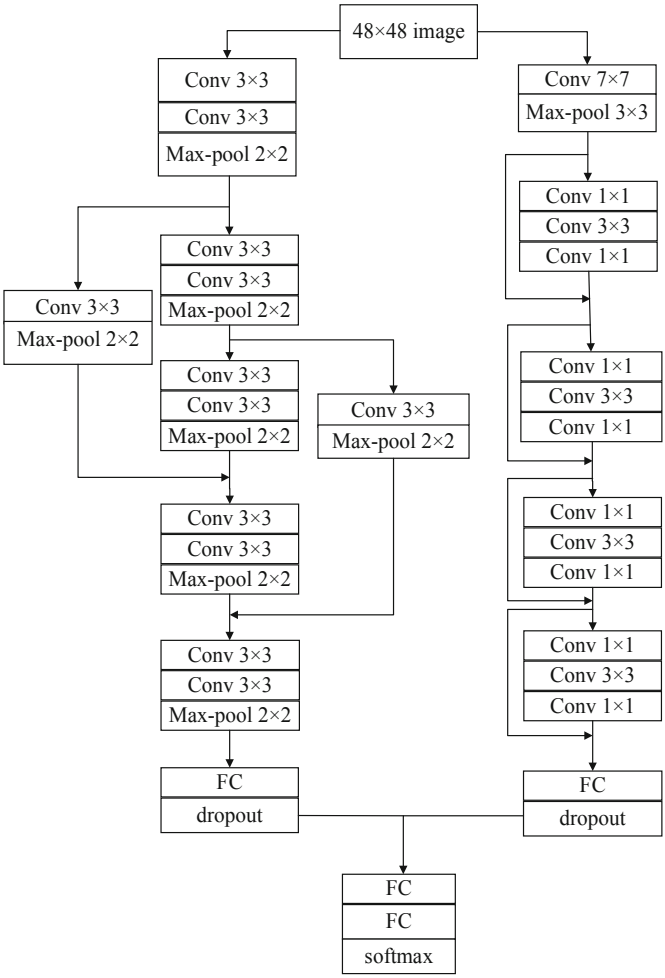

Fig. 4. Overall network structure

The 3 × 3 convolution kernel is the smallest size that can capture the eight-neighborhood information of a pixel. The limited receptive field of two 3 × 3 stacked volume base layers is 5 × 5; the receptive field of three 3 × 3 stacked volume base layers is 7 × 7, so large-size convolution can be replaced by stacking of small-size convolutional layers Layer, and the size of the field remains the same. Therefore, three 3 × 3 filters can be regarded as a decomposition of a 7 × 7 filter. The middle layer has a nonlinear decomposition and plays a role of implicit regularization. Multiple 3 × 3 volume base layers are more non-linear than a large-size filter volume base layer, making the decision function more critical. The network structure diagram after adding feature fusion is shown in Fig. 4. The input layer is a 48 × 48 single-channel picture with pixels. The VGG network is used as the basic structure on the left. The size of the convolution kernel is 3 × 3, and 0 padding is added to the periphery to ensure that the

size of the feature map obtained by the convolution kernel remains unchanged., And then pass the maximum pooling layer to reduce the feature map size to half. There are five such convolutional layers in total. The channel numbers of the five convolution kernels are 64, 128, 256, 512, 512, and there are two branches. Used as a feature fusion, the size is adjusted by the convolutional pooling layer for fusion. After the two channels pass through the fully connected layer, they become feature vectors and then fused together. In order to prevent overfitting, a dropout layer is introduced. Then it is passed to the following fully connected layer and softmax layer for classification prediction, and the prediction result is obtained. In order to introduce non-linear operations to obtain better results, the R-Relu function is selected as the activation function. The overall network structure is shown in Fig. 4.

Among them, stochastic gradient descent (SGD) is used for optimization. The objective function is the average of the loss function of each sample in the data set. The goal of stochastic gradient descent is used to minimize the loss function. The loss function selects the cross-entropy loss function, because the gradient of the last layer of weight is no longer related to the derivative of the activation function, but is only proportional to the difference between the output value and the true value. The convergence is faster at this time. Backpropagation is continuous multiplication, so the update of the entire weight matrix will be accelerated. Secondly, the derivation of multi-class cross entropy loss is simpler, and the loss is only related to the probability of the correct class. Let $f_i(\boldsymbol{x})$ be the loss function of the training data sample with index i, n is the number of training data samples, $\boldsymbol{x}$ is the model parameter vector, and the objective function is

$$f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) \tag{6}$$

The derivative of the objective function at $\boldsymbol{x}$ is

$$\nabla f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}) \tag{7}$$

In each iteration of stochastic gradient descent, a sample index $i \in \{1, \ldots, \mathrm{n}\}$ is randomly sampled uniformly, and gradient $\nabla f_i(\boldsymbol{x})$ is calculated to iterate $\boldsymbol{x}$, the formula is as shown in (8)

$$\boldsymbol{x} = \boldsymbol{x} - \eta \nabla f_i(\boldsymbol{x}) \tag{8}$$

Among them, η is the learning rate, and the computational cost of each iteration is reduced from O(n) for gradient descent to a constant O(1). And stochastic gradient $\nabla f_i(\boldsymbol{x})$ is an unbiased estimate of gradient $\nabla f(\boldsymbol{x})$, as shown in formula (9), which means that on average, stochastic gradient descent is a good estimate of gradient.

$$E_i \nabla f_i(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) \tag{9}$$

3.4 Algorithm Design

The steps of the P-ResNet-VGG network algorithm are shown in Algorithm 1:

Algorithm 1: P-ResNet-VGG algorithm

Input: facial expression image $\boldsymbol{X}$
Output: the extracted feature vector $\boldsymbol{D}$
(1)Perform data processing on the input facial expression image to generate the data format required by the neural network $\boldsymbol{x}$。
(2)Construct P-ResNet-VGG network model, input data $\{\boldsymbol{x}_i\}$， Use formula $\boldsymbol{a}_i = f(\boldsymbol{e}) = f(\boldsymbol{w} \cdot \boldsymbol{x}_i + \boldsymbol{b})$ Calculate the feature vector of the convolutional layer， $\boldsymbol{w}$ Represents the weight matrix of the convolutional layer， $\boldsymbol{b}$ Is the offset， $f(\cdot)$ Represents the activation function. Then through the pooling layer, the feature vector size is reduced to half of the original two. After multiple convolutional pooling operations, the two networks each obtain a one-dimensional feature vector $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$， Combining $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ to obtain $\boldsymbol{D}$， the class prediction is obtained through the fully connected layer and the softmax layer.
(3)Use the cross-entropy loss function to compare the result obtained in step 2 with the real label to calculate the total loss, and then use the stochastic gradient descent algorithm to update the weights $\boldsymbol{w}$。
(4)After multiple iterations of step2 and step3, until the loss value stabilizes, save the model with the highest accuracy, and then input the test image into the trained network to obtain a new feature vector， The $\boldsymbol{D}$ is classified and predicted by the Softmax method, and the prediction result is its predicted label.

4 Experiments

4.1 Experimental Environment and Data Preprocessing

The experiment in this paper uses two data sets, FER2013 and JAFFE, for training and testing, both of which are implemented on the GPU version of the Pytorch deep learning framework.

Before the experiment, the size of the pictures in the two data sets was unified to 48 × 48. In order to prevent over-fitting, the pictures were enhanced by data, and the pictures were mirrored horizontally and rotated slightly.

4.2 Data Set

The FER2013 dataset is one of the most commonly used public facial expression datasets, with a total of 35888 unique facial expression images, including faces of various poses,

light intensity and different proportions. The images are all 48 × 48 pixel grayscale images, as shown in Fig. 5. Each image in the data set is labeled for each of the seven categories: 0 anger, 1 disgust, 2 fear, 3 happiness, 4 sadness, 5 surprise, and 6 neutral.

Fig. 5. Examples of seven expressions in the FER2013 dataset

The JAFFE dataset consists of 213 pictures composed of facial expressions of ten Japanese women, each with 7 facial expressions (Angry, Disgust, Fear, Happy, Sad, Surprise and Normal). Figure 6 is a sample drawn from the JAFFE database.

Fig. 6. Example of seven expressions in Jaffe dataset

4.3 Experiment and Result Analysis

For the FER2013 data set, 28,000 sheets are selected as the training set and 3,500 sheets as the test set. During the network training process, the epoch is set to 200, the batch size is 128, and the learning rate is set to 0.01. The learning rate of the first 50 iterations remains unchanged. When the number of iterations exceeds 50, the learning rate begins to decay, once every 5 rounds. The learning rate becomes 0.9 times the original.

In order to prevent over-fitting, the JAFFE data set adopts a cross-validation method. The data set is divided into 5 parts. Each time 4 parts are used as the training set and 1 part is used as the test set. Each network is trained for 30 epochs. The model with the lowest test loss value is used as the final training model. In the experiment, two models,

P-ResNet and P-ResNet-VGG, were used for experiments on the FER2013 and JAFFE datasets. Tables 1 and 2 show the results of the model on the two data sets.

Table 1. Accuracy of P-ResNet on two data sets %

Emotion	FER2013	JAFFE
Angry	71.27	95.31
Disgust	69.43	97.01
Fear	71.53	97.66
Happy	75.71	98.48
Sad	74.51	98.41
Surprise	72.38	96.83
Normal	72.45	97.10
Average	72.47	97.25

Table 2. Accuracy of P-ResNet-VGG on two data sets %

Emotion	FER2013	JAFFE
Angry	73.25	97.66
Disgust	70.63	98.51
Fear	72.47	99.22
Happy	77.38	100.00
Sad	75.42	99.21
Surprise	74.61	98.41
Normal	71.91	98.55
Average	73.67	98.79

As shown in Table 3, the accuracy of several models in the experiment is compared with the overall accuracy of facial expression recognition of other models.

On the JAFFE data set, the P-ResNet-VGG model has a higher accuracy rate, Table 4 shows the comparison of the accuracy of several models on the JAFFE data set. In the comparison experiment, the method of Zhang et al. [19] is a stacked hybrid autoencoder based on deep learning. The method of Kommineni et al. [20] uses a hybrid feature extraction technique. Kola et al. [21] used a local binary pattern based on adaptive windows for facial expression recognition. The comparison experiment has a more detailed accuracy rate comparison, which can intuitively indicate that the model in this paper is relatively stable and has a higher accuracy rate.

It can be seen from the experimental results that the P-ResNet-VGG model is better than several other models. Improved accuracy on both data sets. When the data set is

Table 3. Comparison of accuracy rates with other models on the FER2013 dataset %

Model	Accuracy
Ref. [16]	70.86
Ref. [17]	71.80
Ref. [18]	71.91
P-ResNet	72.47
P-ResNet-VGG	73.67

Table 4. Comparison of accuracy rates of several models on Jaffe dataset %

Model	Accuracy
Ref. [19]	96.70
Ref. [20]	98.14
Ref. [21]	92.81
P-ResNet	97.25
P-ResNet-VGG	98.79

large, the effect of the parallel network model fusion is better and more accurate. In addition, the features after deep and shallow layer fusion are more complete, and the recognition rate is higher; fewer network layers are used for training, which makes up for the slow model training problem of the CNN network due to the large network of training data. Requires advanced problems, so while reducing training time, a higher recognition rate is obtained.

5 Conclusion

The method based on parallel network feature fusion combines the advantages of the two network structures of VGG and ResNet. On the one hand, it takes full advantage of the fast training speed and fewer parameters of the ResNet network to improve network training efficiency; on the other hand, the fusion of different networks and the fusion of deep and shallow features of the network itself can have a certain recognition rate for expressions in big data. Improved, and at the same time, compared with other deep network training time has been effectively reduced. From the experimental results, it can be seen that the P-ResNet-VGG model has certain advantages compared with the compared models. It plays a role in training time and accuracy, and the effect of increasing the amount of data is more obvious. In terms of feature extraction, the use of better and more accurate methods to distinguish the gaps between different expressions will be the next step in research work to further reduce training time and improve recognition accuracy.

References

1. Logie, R.H., Baddeley, A.D., Woodhead, M.M.: Face recognition, pose and ecological validity. Appl. Cogn. Psychol. **1**(1), 53–69 (2015)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, New York, pp. 1097–1105 (2012)
3. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
4. Goodfellow, I.J., Erhan, D., Carrier, P.L., et al.: Challenges in representation learning: a report on three machine learning contests. Neural Netw. **64**, 59–63 (2015)
5. Xu, M.Y., Tang, Z.M., Yao, Y.Z., et al.: Deep learning for person reidentification using support vector machines. Adv. Multimed. **2017**, 11–18 (2017)
6. Wang, Y., Su, W.J., Liu, H.L.: Facial expression recognition based on linear discriminant locality preserving analysis algorithm. J. Inf. Comput. Sci. **9**(11), 4281–4289 (2013)
7. Tang, C., Zheng, W., Yan, J., et al.: View-independent facial action unit detection. In: Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition, Los Alamitos, CA, USA, pp. 878–882 (2017)
8. Chu, J., Tang, W., Zhang, S.: Facial expression recognition algorithm based on attention model. Laser Optoelectron. Prog. **57**(12), 121015 (2020)
9. Lu, G., He, J., Yan, J.: A convolutional neural network for facial expression recognition. J. Nanjing Univ. **36**(1), 16–22 (2016)
10. Li, X., Niu, H.: Facial expression recognition based on feature fusion based on VGG-NET. Comput. Eng. Sci. **42**(03), 500–509 (2020)
11. Li, M., Li, X., Wang, X., et al.: Real-time face expression recognition based on multi-scale kernel feature convolutional neural network. J. Comput. Appl. **39**(09), 2568–2574 (2019)
12. Mishra, G., Vishwakarma, V.P., Aggarwal, A.: Face recognition using linear sparse approximation with multi-modal feature fusion. J. Discrete Math. Sci. Crypt. **22**(2), 161–175 (2019)
13. Wang, H.: Enhanced forest microexpression recognition based on optical flow direction histogram and deep multiview network. Math. Probl. Eng. **2020**(8), 1–11 (2020)
14. Li, D., Zhao, X., Yuan, G., et al.: Robustness comparison between the capsule network and the convolutional network for facial expression recognition. Appl. Intell. **51**(4), 2269–2278 (2020)
15. Zhang, T., Zheng, W., Cui, Z., et al.: A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimed. **18**(12), 2528–2536 (2016)
16. Guo, Y.N., Tao, D.P., Yu, J.: Deep neural networks with relativity learning for facial expression recognition. In: Proceedings of the 2016 IEEE International Conference on Multimedia and Expo Workshop, Washington, pp. 166–170 (2016)
17. Zhou, S., Liang, Y., Wan, J., Li, S.Z.: Facial expression recognition based on multi-scale CNNs. In: You, Z., et al. (eds.) CCBR 2016. LNCS, vol. 9967, pp. 503–510. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46654-5_55
18. Hua, W., Dai, F., Huang, L., et al.: HERO: human emotions recognition for realizing intelligent Internet of Things. IEEE Access **7**, 1 (2019)
19. Zhang, Z.Y., Wang, R.Q., Wei, M.M.: Stack hybrid self-encoder facial expression recognition method. Comput. Eng. Appl. **55**(13), 140–144 (2019)
20. Kommineni, J., Mandala, S., Sunar, M.S., et al.: Accurate computing of facial expression recognition using a hybrid feature extraction technique. J. Supercomput. **77**(11), 1–26 (2020)
21. Kola, D.G.R., Samayamantula, S.K.: A novel approach for facial expression recognition using local binary pattern with adaptive window. Multimed. Tools Appl. **80**(12), 1–20 (2020)