Journal of Network Intelligence ©2022 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **7**, Number **1**, February **2022**

A Facial Expression Recognition Method based on Residual Separable Convolutional Neural Network

Xiaoyu Xu

College of Computer and Information Engineering
Xiamen University of Technology
Xiamen 361024,China
xiaoyxu@foxmail.com

Jianfeng Cui *

College of Software Engineering
Xiamen University of Technology
Xiamen 361024,China
Corresponding Author: jfcui@xmut.edu.cn

Xuhui Chen*

[1] Fujian Key Laboratory of Internet of things application technology
[2]Xiamen City University of Technology
Xiamen 361024, China
Corresponding Author:xhchen@xmut.edu.cn

Chin-Ling Chen*

[1] School of Computer and Information Engineering
Xiamen City University of Technology
Xiamen 361024,China
[2] School of Information Engineering
Changchun Sci-Tech University
Changchun 130600, China
[3] Department of Computer Science and Information Engineering
Chaoyang University of Technology
Taichung 41349, Taiwan
Corresponding Author:clc@mail.cyut.edu.tw

Received March 2021 Revised Augustl 2021

Abstract. *Facial expression recognition is a typical application of artificial intelligence. It has practical value and commercial prospects in application scenarios such as safe driving, medical assistance, online classrooms, and human-computer interaction. Traditional convolutional neural networks have low accuracy in facial expression recognition. To solve this problem, we propose a residual separable convolutional neural network, which combines depth-wise separable convolution with an improved residual network. When transmitting feature information, it fuses facial expression features through skip connection and utilizes global average pooling to reduce dimension, thus preventing gradient disappearance. We conduct experiments based on the data set FER2013. The results show that it can improve the accuracy of facial expression recognition, reduce model parameters and raise the utilization efficiency of the parameters effectively.*

Keywords: Facial expression recognition, convolutional neural network, depth-wise separable convolution, residual network.

1. **Introduction.** In recent years, with the rapid development of artificial intelligence technology, human-computer interaction has begun to enter the public's field of vision. As an important part of the field of computer vision, facial expression recognition has also become a research hotspot [1]. Facial expression is one of the most significant characteristics of human emotion recognition. We can judge a person's emotion by observing his face. The purpose of facial expression recognition research is to classify the human emotions obtained from facial images into one of seven basic emotions: happiness, sadness, fear, disgust, anger, surprise, and neutral. It plays a great role in safe driving, intelligent monitoring, online teaching, criminal interrogation, and other application scenarios. In facial expression recognition and classification, using traditional feature extraction and classification methods to achieve a high recognition rate is computationally complex and challenging. Convolutional neural network shows great potential because of their automatic feature extraction ability and computational efficiency [2]. This paper aims to improve the prediction accuracy and robustness of facial expression and proposes a new residual separable convolution neural network model. Traditional facial expression recognition methods are mainly based on hand-crafted feature extraction. Shan et al. [3] proposed a boosted local binary pattern (LBP) model combined with a support vector machine (SVM) classifier, which is robust to low-resolution face images. Deng et al. [4] put forward an algorithm based on improved principal component analysis (PCA). It features higher adaptability to expression variation, but when the environment changes greatly, the performance of recognition will drop significantly. Xie et al. [5] presented the Turbo-Boost method based on Haar features [6] that can extract the main information from face images and classify facial expressions. Hand-crafted feature extraction, which depends on manual operation, not only takes a lot of time but often fails to identify some important features, thus affecting the final classification effect.

With the development of deep learning in recent years, convolutional neural networks [7] play an increasingly important role in image processing and facial expression recognition. Its advantage is that it uses characterization learning and hierarchical feature extraction algorithms to overcome feature engineering problems [8]. Lecun et al. [9] put forward the first convolutional neural network model — LeNet-5. It integrates convolution into a

neural network, making it possible to automatically extract image features. In addition, it further shortens the learning time and improves the efficiency of recognition, laying a foundation for the development of a convolutional neural network. Zou et al. [10] proposed a LeNet network with deepened layers, which can improve the accuracy of facial expression recognition on small data sets. However, when processing complex and large-scale facial expression data, its accuracy and robustness need to be improved. Lu et al. [11] introduced a facial expression recognition method including seven layers of convolutional neural networks, however, without taking the relationship of each layer's feature images into account. The large and deep convolutional neural network model — AlexNet — proposed by Krizhevsky et al. [12] has raised the training efficiency of the whole network and achieved success in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13] making it possible to train a convolutional neural network to process massive data. However, the generated image features lack diversity, and feature extraction is insufficient, which further affects the recognition performance. Subsequently, some famous network model structures with higher performance were put forward, including VggNet by Oxford University [14] and Inception [15]. With the deepening of network layers, problems such as gradient disappearance and over-parameterization occur, exerting an impact on network training and robustness [16]. The deep residual network — Restnet — presented by He et al. [17] involves a residual structure, which can not only train deeper networks but also mitigate the problem of network degradation to some extent. Due to the introduction of the residual structure, a slight variation in network weight can lead to a significant change in the output. Therefore, the output structure is unsatisfactory. Dachapally et al. [18] put forward an 8-layer convolutional neural network and introduced automatic encoders. The model works well in the recognition of random images, but it ignores the structural integrity of the images. Moreover, it has millions of parameters, which makes training difficult. Arriaga et al. [19] proposed an improved network model — Xception [20] — with fewer parameters. But in this model, information fusion only occurs in local areas during feature extraction, and the global information among channels is ignored.

In response to the above problems, this paper presents a new model that combines depth-wise separable convolution with an improved residual network to achieve the following goals:
1) To extract richer and more complete feature information and raise the efficiency of feature extraction.
2) To solve the problems of gradient disappearance and over-parameterization, reduce the difficulty of network training, and prevent over-fitting.
3) To improve the accuracy of recognition and decrease the number of model parameters.

2. **Algorithm description.**

2.1. **Improved residual structure.**

Resnet has become a widely popular model of convolutional neural network (CNN) in recent years. Compared with the traditional ones, it adds jump connection [21] and batch normalization (BN) [22] to the network. It can help low-dimensional features transfer to high-dimensional ones, maintain a certain proportion of output, reuse image feature information and solve the problem of gradient disappearance. In this paper, the traditional residual structure is upgraded. We use the 1×1 convolution to increase dimension first, and then in a 3×3 network structure, dimension reduction is achieved through the 1×1 convolution. In other words, the channel is expanded first and then compressed. In addition to retaining the jump connection which exists in the basic residual structure, the improved model also adopts convolution kernel, depth-wise separable convolution, and activation module. The Basic residual structure is shown in Figure 1 and the improved residual structure is shown in Figure 2.

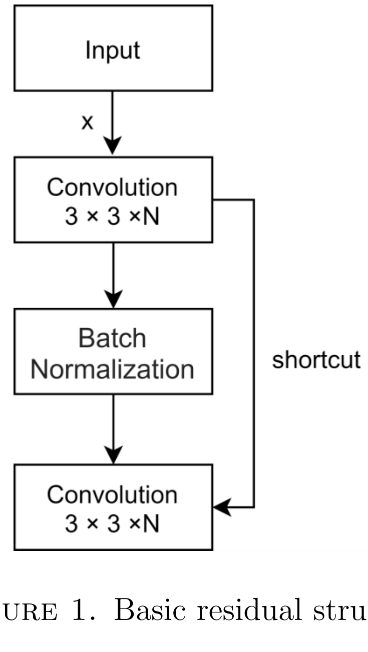


FIGURE 1. Basic residual structure

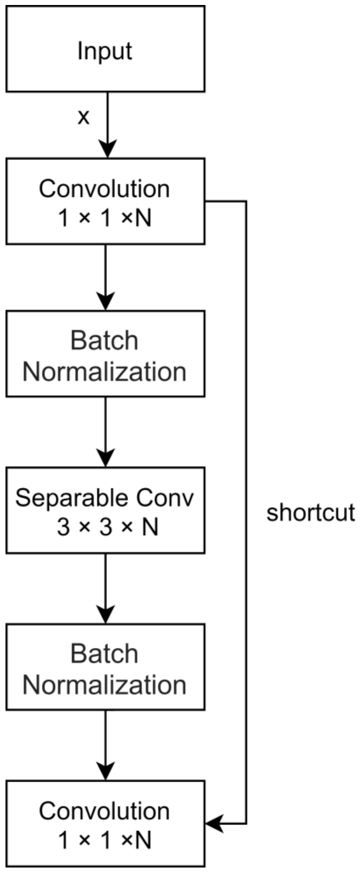


FIGURE 2. Improved residual structure

2.2. **Nonlinear activation function.** The Swish function is a new type of activation function (AF) proposed by Google, and the original equation is as follows:

$$swish\left(x\right) = x.sigmoid\left(x\right) = \frac{x}{1+\exp\left(-x\right)} \tag{1}$$

It is a smooth and non-monotonic function with a lower bound but no upper bound, which outperforms the ReLU function in deep models [23]. Due to the complexity of the Sigmoid function, this paper adopts the nonlinear activation function h_swish [24] that replaces Sigmoid with ReLU6(x+3)/6. The ReLU6 function can be used in almost all software and hardware frameworks. It runs faster and can eliminate the potential loss of numerical accuracy caused by the different implementations of approximate Sigmoid in specific modes.

$$h\_swish\left(x\right) = x\frac{RELU6\left(x+3\right)}{6} = \begin{cases} 0, x \leq 3 \\ \dfrac{x^2}{6}, -3 \leq x \leq 3 \\ x, x > 3 \end{cases} \tag{2}$$

2.3. **Lightweight attention model (SE module).** Attention mechanism has become the mainstream model component in practical application because of its good effect and expandability . [25]. In the previous models, information fusion only occurs in local areas during feature extraction and the global information among channels is ignored. To solve this problem, we adopt the Squeeze-and-Excitement Module (SE-Module) [26] to model the dynamic and nonlinear relationship between network channels, recalibrate different channels in the original convolutional layer and change the weight ratios between them.

By doing so, the network can integrate global features into the channel information. We use global average pooling (GAP) [27] to compress input image features. When the global feature information between the channels is obtained, the model begins to learn. The first fully connected layer (FC) is linked to reducing the dimension, and the RELU (Rectified Linear Unit) activation function is used to enhance the nonlinearity of the network and better fit the complex correlation between the channels. Then, the number of channels will grow to the initial state through the second fully connected layer. The h_swish function is introduced to filter the channels and change their weights. The more important a channel is, the greater its weight is. After that, the feature recalibration layer — Scale — begins to weigh the recalibrated weights to the previous features, to complete the screening of the original image features.
The mechanism of the lightweight attention model is shown in Figure 3.

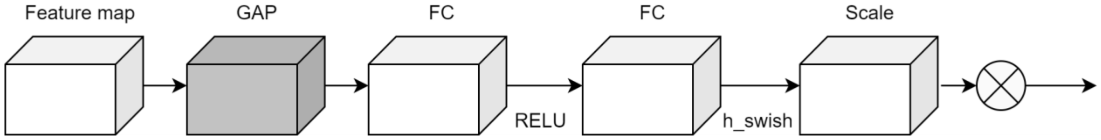


Figure 3. Lightweight attention model

3. **Residual separable convolutional neural network model.** The model proposed in this paper is developed from Xception and ResNet. To increase the accuracy of recognition and reduce model parameters, an improved residual neural network model based on separable convolution is designed. In this model, the attention mechanism is adopted, the RELU function is replaced with the h_swish function, and the input size is 48 × 48 × 1. The network structure is shown in Figure 4 and the parameters of each layer are in Table 1. The model has three convolutional layers and six improved residual modules based on separable convolution. In the first layer, there are 16 3 × 3 filters with a step size of 2. Batch normalization and the h_swish function are introduced, however, without pooling. In IRes modules, the dimension is raised by 1×1 convolution, then more complex features are extracted by 3 × 3 depth-wise separable convolution, and a lightweight attention mechanism is added. After that, the dimension is reduced by 1×1 convolution, and finally, the input is connected to the output by a shortcut. When the construction of the first convolutional layer is completed, it begins to stack IRes modules. In the first module, 16 3×3 filters are set with a step size of 1. Then another two modules are constructed. The first one has a step size of 2 and 32 channels where the input feature layer will be compressed and the number of the channels will grow. The step size of the second IRes is 1 and the input feature layer will not change. When completing the construction of the two modules, it begins to stack three IRes modules and introduce the attention

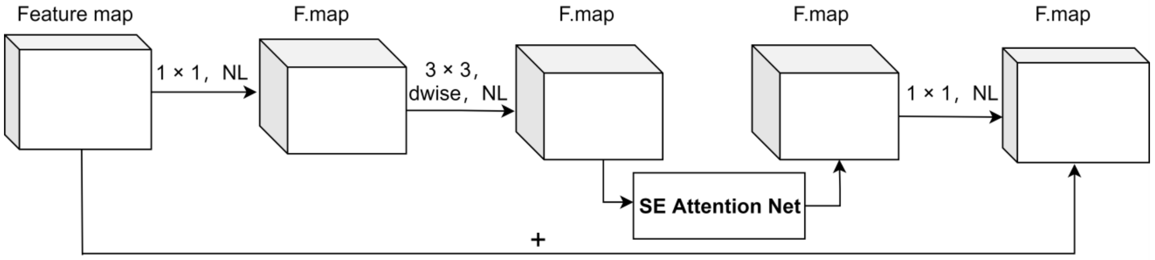


Figure 4. Improved residual separable convolutional neural network model

Table 1. Improved residual separable convolutional neural network model structure

Ipunt Size	Layer type	Exp size	Patch Size/ Stride	Output	SE	AF
48×48×1	Convolution	-	3×3 / 2	24×24×16	-	HS
24×24×16	Ires, 3 × 3	16	3×3 / 1	24×24×16	-	RE
24×24×16	Ires, 3 × 3	32	3×3 / 2	12×12×24	-	RE
12×12×24	Ires, 3 × 3	64	3×3 / 1	12×12×24	-	RE
12×12×24	Ires, 5 × 5	64	3×3 / 2	6×6×40	✓	RE
6×6×40	Ires, 5 × 5	72	3×3 / 1	6×6×40	✓	RE
6×6×40	Ires, 5 × 5	88	3×3 / 1	6×6×40	✓	RE
6×6×40	Conv2d,1 × 1	-	3×3 / 2	6×6×256	-	HS
6×6×256	Globalpooling	-	-/1	-	-	-
1×1×256	Conv2d,1 × 1	-	-/1	1×1×7	-	-

mechanism. A convolutional layer is added after the IRes modules, which is provided with 256 1×1 filters and adopts the h_swish function.

In this paper, we substitute GAP for the conventional FC, to achieve dimension reduction and reduce over-fitting. After global average pooling, the softmax function is introduced to classify facial expressions. And in all layers, we adopt RELU and h_ swish as the activation functions.

4. Experiment and results.

4.1. Preparation.

The experiment configuration is Microsoft Windows 10 64-bit Operating System, i7-6700HQ CPU (2.60 GHz), and NVIDIA GTX965M (2 GB). The experiment is based on the TensorFlow Deep learning framework (Python Language).

4.2. Data sets.

This study, to evaluate the effectiveness of the method, employs the data set Fer2013 [28] provided by the Kaggle website, which was obtained by searching Google keywords (Pierre and Aaron). The data set collected 35,887 facial expression images (48 × 48 pixels), including 28,708 training sets, 3,589 validation sets, and 3,589 test sets. The face in each image is almost centered and occupies about the same space in each image. Each image is classified into one of seven expression categories, and each category corresponds to a digital label of 0-6: 0-angry, 1-disgust, 2-fear; 3-happy, 4-sad, 5-surprised, 6-neutral. Figure 5 depicts an example of each type. Table 2 shows the distribution of each category, including the expressions corresponding to tags 0-6, and the number of training sets, validation sets, and test sets contained in each expression.

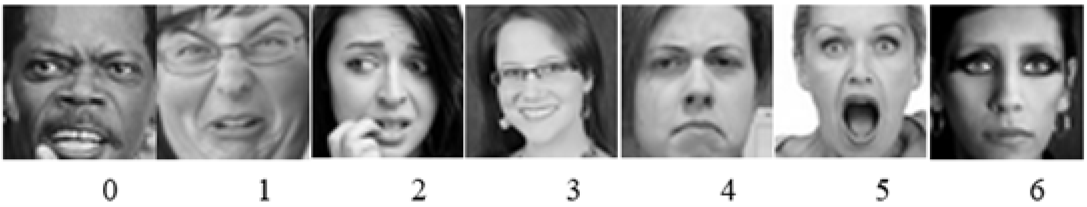


Figure 5. Examples of seven expressions

Table 2. Quantity distribution of facial emotion categories

Type	Training set	Validation set	Test set
0-Angry	3,995	467	491
1-Disgust	436	56	55
2-Fearful	4,097	496	528
3-Happy	7,215	895	879
4-Sad	4,830	653	594
5-Surprised	3,171	415	416
6-Neutral	4,965	607	626

4.3. **Pre-processing of data.** In order to avoid network over-fitting and improve the generalization ability of the model, some data enhancement operations, such as flipping, rotating, and cutting, are performed on the image in this paper. They expand the amount of data in the database and make the training network more robust. In this paper, 48×48 images were randomly cut out during training and then fed back into training. 80

4.4. **Results.** In this experiment, the accuracy is used to evaluate the identification effect of a neural network, in which TP (True Positive) refers to the number of outcomes that are correctly predicted to be positive samples, FP (False Positive) refers to the number of outcomes which are incorrectly predicted as positive samples, TN (True Negative) refers to the number of outcomes which are correctly predicted as negative samples, and FN (False Negative) refers to the number of outcomes which are incorrectly predicted as negative samples. The accuracy equation is shown in Eq. (3).

$$Accuracy = \frac{TP \ + \ TN}{TP + TN + FP + FN} \tag{3}$$

In order to compare the classification effect of the improved model in facial expression images, this paper experimented on the Fer2013 data set and trained the data for 100 epochs (Table 3). Table 3 shows the accuracy comparison of this experimental method and the method based on the convolutional neural network on the Fer2013 data set. In the traditional CNN model structure, the accuracy of the model on the data set is 60.71 Figure 6 and Figure 7 show the training and validation accuracy and the training and validation loss obtained by training the improved network model on the data set, respectively. It can be seen that the improved network has good learning ability. With the superposition of training times, the validation loss value will decrease with the reduction

Table 3. Comparison results of facial expression images recognized by each model

Approach	Accuracy rate	Parameter quantity
Simple_CNN	60.71%	642,900
Simper_CNN	61.84%	559,700
LeNet	63.84%	1,587,000
Xception	64.04%	58,400
Resnet50	64.71%	10,646,600
This paper	66.29%	54,900

of training loss value until it is flat. The whole training process does not have the phenomenon of good accuracy on the training set, but poor accuracy on the validation set, that is, over-fitting phenomenon [29], which indicates that the improved network model in this paper has good generalization ability.

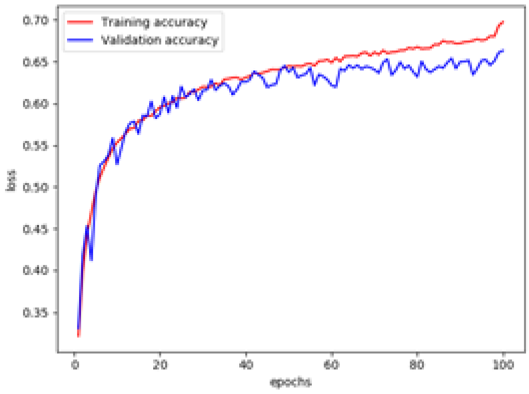


Figure 6. Training and validation accuracy

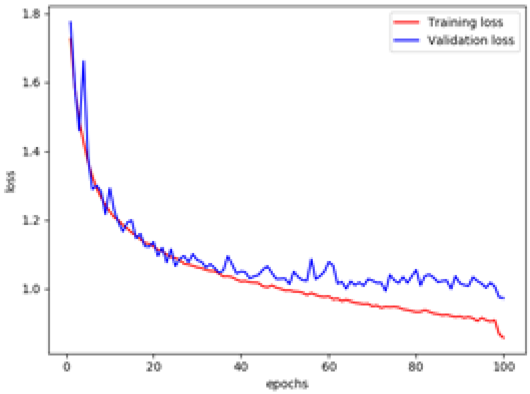


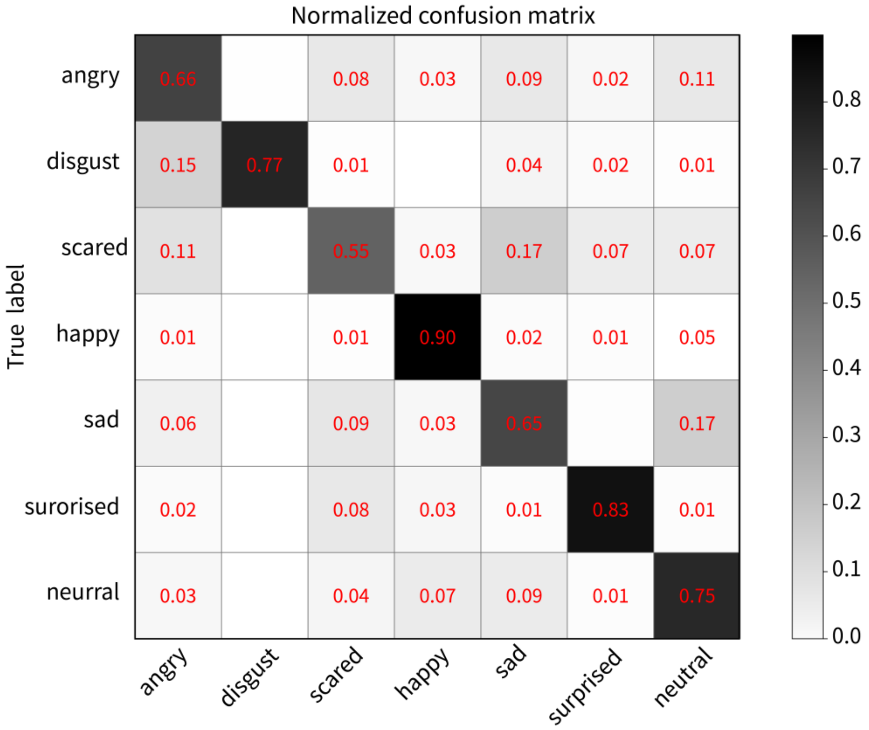
Figure 7. Training and validation loss

The improved model is used in the experiment, and the results are shown in Figure 8. The accuracy of the confusion matrix in the figure indicates that of each category's prediction (77% in "fearful", 83% in "surprised" and even 90% in "happy"). On the Fer2013 data set, several other expression identification methods were used to test. Compared with the methods in this study, we got the results (Table 4). It can be seen from Table 4 that compared with other methods, this model can get better expression identification results. Ref. [30] adopts a convolutional neural network based on FPGA, which can only extract some features of images, and the accuracy rate is only 60.3

Table 4. Comparison of recognition rate based on Fer2013 data set

Method	Accuracy
Ref. [30]	60.30%
Ref. [31]	61.86%
Ref. [32]	62.4%
Ref. [33]	64.4%
Ref. [34]	65.03%
Method of this paper	66.29%

5. **Conclusion.** Aiming at the challenges of insufficient feature extraction in facial expression recognition and low recognition accuracy, this paper proposes a residual separable



	angry	disgust	scared	happy	sad	surprised	neutral
angry	0.66		0.08	0.03	0.09	0.02	0.11
disgust	0.15	0.77	0.01		0.04	0.02	0.01
scared	0.11		0.55	0.03	0.17	0.07	0.07
happy	0.01		0.01	0.90	0.02	0.01	0.05
sad	0.06		0.09	0.03	0.65		0.17
surorised	0.02		0.08	0.03	0.01	0.83	0.01
neurral	0.03		0.04	0.07	0.09	0.01	0.75

FIGURE 8. Accuracy of correct prediction for each category

convolutional neural network model. By comparing with representative algorithms, combining deep separable convolution and improved residual network structure, not only can the convolution calculation be reduced, but also feature fusion can be carried out while transferring feature information, thereby extracting richer features. The convolutional layer uses a lightweight attention mechanism to adjust the weights between channels and adds global features between channels to make the information of the feature map more complete. The experimental results show that the accuracy of expression identification is improved and the network parameters are reduced. However, the identification effect on sad expressions is relatively poor, because human expressions are very complex, and the random combination of various expressions will greatly increase the identification difficulty. For example, the combination of sadness and consternation will produce disgust, and that of happiness and surprise will produce surprises. In the future, we will further extract the global information and edge information of images, improve the identification rate of each expression, and realize the real-time identification of expressions.

Acknowledgment: This research was supported by the industry-university research innovation fund of Chinese universities(2020ITA03003), CERNET Innovation Project (NGII20160201), and Informatization Project of Fujian Education Department.

REFERENCES

[1] C. M. Tang, H. B. Zhao, X. Y. Zhang, East Asians facial expression identification algorithm based on manifold learning 2D-LDLPA, *Computer Engineering and Application*, vol. 54, no. 17, pp.146-150, 2018.

[2] Y. Zhang, L. Guo, C. N. Du, Extraction of English Drug Names Based on Bert-CNN Mode, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 11, no. 2, pp.70-78, 2020.

[3] C. Shan, S. Gong, P. W. Mcowan, Facial expression recognition based on Local Binary Patterns: A comprehensive study, *Image and Vision Computing*, vol. 27, no. 6, pp.803-816, 2009.

[4] W. Deng, J. Hu, J. Lu, Transform-invariant PCA: a unified approach to fully automatic face alignment, representation, and recognition, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 6, pp.1275-1284, 2014.

[5] E. Xie, S. Luo, L. Pan, Turbo-Boost expression recognition using Haar-like features, *Journal of Computer Aided Design and Computer Graphics*, vol. 23, no. 8, pp.1442-1446, 2011.

[6] P. A. Viola, M. J. Jones, Rapid object detection using a boosted cascade of simple features, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.511-518, 2001.

[7] F. Yang, J. P. Li, X. Li, Salient object detection algorithm based on multi-task deep convolution neural network, *Journal of Computer Applications*, vol. 38, no. 1, pp.91-96, 2018.

[8] J. M. T. Wu, M. H. Tsai, Applying an ensemble convolutional neural network with Savitzky-olay filter to construct a phonocardiogram prediction model, *Applied Soft Computing*, vol. 78, pp.29-40, 2019.

[9] Y. Lecun, L. Bottou, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, vol. 86, no. 11, pp.2278-2324, 1998.

[10] J. C. Zou, X. L. Cao, A facial expression identification method based on improved convolutional neural network, *Journal of North China University of Technology*, vol. 32, no. 2, pp.39-44, 2020.

[11] G. M. Lu, J. L. He, J. J. Yan, Convolutional neural network for facial expression recognition, *Journal of Nanjing University of Posts and Telecommunications*, vol. 36, no. 1, pp.16-22, 2016.

[12] A. Krizhevsky, I. Sutskever, ImageNet Classification with deep convolutional neural networks, *Neuro Computing*, vol. 60, no. 6, pp.84-90, 2017.

[13] O. Russakovsky, J. Deng, ImageNet large scale visual recognition challenge, *International Journal of Computer Vision*, vol. 115, no. 3, pp.211-252, 2015.

[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint* arXiv:1409.1556, 2014.

[15] C. Szegedy, W. Liu, Y. Jia, Going deeper with convolutions, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-9, 2015.

[16] K. Wang, P. Xu, C. M. Chen, Neural architecture search for robust networks in 6G-enabled massive IoT domain, *IEEE Internet of Things Journal*, vol. 8, no. 7, pp.5332-5339, 2020.

[17] K. He, X. Zhang, S. Ren, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[18] P. R. Dachapally, Facial emotion detection using convolutional neural networks and representational autoencoder units, *Neural computation*, vol. 22, no. 2, pp.511-538, 2017.

[19] O. Arriaga, M. Valdenegro-Toro, P. Plöger, Real-time convolutional neural networks for emotion and gender classification, *arXiv preprint* arXiv:1710.07557, 2017.

[20] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1800-1807, 2017.

[21] R. Yasrab, SRNET: A Shallow Skip Connection Based Convolutional Neural Network Design for Resolving Singularities, *Journal of Computer Science and Technology*, vol. 34, no. 4, pp.924-938, 2019.

[22] M. Ji, J. C. Chun, An Improved Image Classification Using Batch Normalization and CNN, *Journal of Internet Computing and Services*, vol. 19, no. 3, pp.35-42, 2018.

[23] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, *Proc. of the 14th International Conference on Artificial Intelligence and Statistics*, pp.315–323, 2011.

[24] P. Ramachandran, B. Zoph, Q. V. Le, Searching for Activation Functions, *arXiv preprint* arXiv:1710.05941, 2017.

[25] E. K. Wang, X. Zhang, F. Wang, Multilayer Dense Attention Model for Image Caption, *IEEE Access*, vol. 7, pp.66358-66368, 2019.

[26] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, *Proc. of the IEEE Conference on Applications of Computer Vision*, pp.7132-7141, 2018.

[27] M. Lin, Q. Chen, S. C. Yan, Network in network, *arXiv preprint* arXiv:1312.4400, 2013.

[28] I. J. Goodfellow, D. Erhan, P. L. Carrier, Challenges in representation learning: A report on three machine learning contests, *Proc. of the International Conference on Neural Information Processing*, pp.117–124, 2013.

[29] J. M. T. Wu, M. H. Tsai, S. H. Xiao, A deep neural network electrocardiogram analysis framework for left ventricular hypertrophy prediction, *Journal of Ambient Intelligence and Humanized Computing*, pp.1-17, 2020.

[30] H. Phan-Xuan, T. Le-Tien, S. Nguyen-Tan, FPGA platform applied for facial expression recognition system using convolutional neural networks, *Procedia Computer Science*, vol. 151, pp.651-658, 2019.

[31] G. Zeng, J. Zhou, J. Xi, Hand-crafted feature guided deep learning for facial expression recognition, *Proc. of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp.423-430, 2018.

[32] S. Alizadeh, A. Fazel, Convolutional Neural Networks for Facial Expression Recognition, *arXiv preprint* arXiv:1704.06756, 2017.

[33] A. Mollahosseini, D. Chan, M. H. Mahoor, Going deeper in facial expression recognition using deep neural networks, *Proc. of the IEEE Winter Conference on Applications of Computer Vision*, pp.1-10, 2016.

[34] K. Liu, M. Zhang, Z. Pan, Facial Expression Recognition with CNN Ensemble, *Proc. of the 2016 International Conference on Cyberworlds*, pp.163-166, 2016.