Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of systematic manipulation of the publication process, including compromised peer review. The Editor and publisher no longer have confidence in the results and conclusions of the article.

JT and JF agreed with retraction. YW either did not respond directly or could not be reached.

Facial expression recognition in classroom environment based on improved Xception model

Jia Tian[a,*] **Jian Fang**[a] **and Yue Wu**[b]

[a]Jilin Engineering Normal University, College of Electrical Engineering, Changchun, Jilin, China

[b]Jilin Engineering Normal University, School of Automotive Engineering, Changchun, Jilin, China

Abstract. Nowadays, there is a large amount of research on facial expression, and researchers have put forward a variety of effective methods. Now, due to the unsupervised learning function, deep learning is increasingly applied to facial expression recognition. The purpose of this paper is to study the recognition of facial expressions in a classroom environment based on an improved anomaly model. This paper proposes a new face recognition model, which is based on convolutional neural network. This paper introduces the construction and algorithm of the model and tests its performance for expression recognition through experiments. The experimental results show that the expression recognition accuracy of the facial expression recognition model proposed is 72.4%. The experiment compares this model with other models, the recognition accuracy of the proposed model architecture is 72.4%, and finds that this model has the best classification accuracy and the least parameters. © *2022 SPIE and IS&T* [DOI: 10.1117/1.JEI.31.5.051416]

Keywords: convolutional neural network; residual network; lightweight convolutional neural network; Xception model; multi-modal fusion emotion recognition technology.

Paper 220033SS received Feb. 26, 2022; accepted for publication May 6, 2022; published online Jun. 22, 2022.

1 Introduction

1.1 *Background*

The research on machine interpretation of human emotion is very meaningful. Facial expression can be understood faster than language communication,[1] because it is more intuitively displayed in front of people, but facial expression is more complex and diverse, so it will be difficult to study. Therefore, compared with face recognition technology, facial expression recognition has less research. There is a motion coding system that can get more refined expression analysis.[2] Its application can be extended to the fields of online classroom teaching, quality analysis, medicine, driving supervision, and so on. Face recognition technology includes image acquisition, feature location, identity confirmation and search, etc.

1.2 *Significance*

Facial expression recognition is now a research topic in the field of computer vision and artificial intelligence. This is an important part of artificial intelligence and computer technology, which has attracted great attention in recent years. Researchers in various fields have proposed many new methods. For example, the relevant theories of machine learning can extract the features of expression, but if the problem of expression recognition becomes complex, the network structure will become increasingly complex, the parameters will continue to increase, and the computational complexity will also increase.[3] Therefore, in this paper, we recognize facial expressions in

*Address all correspondence to Jia Tian, 126425@jlenu.edu.cn

1017-9909/2022/$28.00 © 2022 SPIE and IS&T

a classroom environment based on improved anomaly model, the data are preprocessed, facial essentials are extracted, a new convolutional neural network (CNN) model is constructed, and the new CNN model is used to extract facial expressions, and its effectiveness is studied.

1.3 *Related Work*

The CNN is a disruptive technology that breaks the most advanced algorithms in many fields from text, video, to voice. Many scholars have proposed different CNN structures. The CNN is a feedforward neural network whose artificial neurons can respond to surrounding units within a certain coverage area and has excellent performance for large-scale image processing.

For example, Burkert et al.[4] proposed a facial expression recognition model based on CNN. The network uses two parallel feature extraction modules. The final result proves that the model has achieved 99.6% and 98.63% recognition accuracy in CK+ database and MMI database, respectively.[5] Nguyen et al.[6] proposed an 18-layer CNN model similar to Visual Geometry Group. The model improves the classification task by combining feature hierarchies. The classification function has not only advanced classification but also intermediate classification (such as background, hair, and other features).

The facial expression recognition process includes three stages: (1) use a face detector to detect the face area in the image containing the face;[7] (2) extract facial features from the detected face area; and (3) facial feature analysis: analyze facial finite element movement and interpret facial expressions.[8] Feature extraction and feature analysis are the key links of facial expression recognition.[9] Feature extraction methods are mainly divided into two categories, one is the extraction method based on dynamic features. It uses the geometric feature extraction of key points of the face in the image sequence,[10] texture feature extraction,[11] optical flow method,[12] and differential image method for facial expression recognition. Texture is an important feature to express an image, it does not depend on color or brightness and reflects the homogeneity of the image, and reflects important information about the organization and arrangement of the surface structure and their connection with the surrounding environment. The second is the static feature extraction method,[13] including local binary pattern method,[14] Gabor wavelet transform method,[15] etc. In terms of facial expression feature classification and recognition, the commonly used classification algorithms include support vector machines (SVM),[16] K-nearest neighbor classification,[17] random forest,[18] and other classification methods. However, the SVM algorithm is difficult to implement for large-scale training samples, and it is difficult to solve the multi-classification problem with SVM. With the advent of deep learning, researchers have gradually shifted from using traditional methods to deep-learning methods for facial expression recognition.[19]

For the problem of facial expression samples, Yang et al.[20] proposed an edge-CNN small input network model for facial expression classification. It uses a small sample data set, which not only reduces the amount of calculation but also increases the accuracy of FER-2013 to 71.80%. Lopes et al.[21] proposed a facial expression recognition algorithm based on CNN using deep learning. In this algorithm, the facial expression parts (e.g., ears, forehead parts, etc.) that are useless for expression are cropped in the image preprocessing step. It improves the accuracy of classification, with an accuracy rate of 96.76% in the CK+ database.

1.4 *Innovation*

The innovation of this paper is: (1) the use of deep separable convolution to extract facial expressions makes the calculation cost and parameters less and more than ordinary convolution. (2) This paper proposes a new facial expression recognition model, which is designed based on CNN, and experiments verify the accuracy of the model for expression recognition.

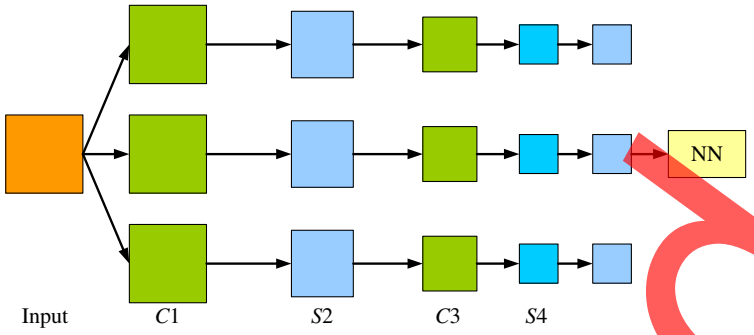

Fig. 1 CNN structure diagram.

2 Method of Facial Expression Recognition in the Classroom Environment Based on the Improved Xception Model

2.1 *Construction of CNN Model*

2.1.1 *Traditional CNN*

In a CNN, there are multiple different levels, and each plane is composed of neurons.[22] Its network structure diagram is shown in Fig. 1.

In the structure diagram of the CNN, the C layer represents the feature extraction layer, the S layer represents the feature mapping layer, and multiple feature maps form the calculation layer. This structure has an advantage, that is, the invariance of displacement.

The first feature extraction layer is to convolve the input image with the feature map and addable bias. After the convolution, a feature map is generated in the $C1$ layer. Each feature map is composed of neurons, and each neuron receives a filter. The second feature map layer is a feature map obtained by summing, weighting, and biasing the pixels of each group of images in the feature map, and then using a Sigmoid function to achieve sub-sampling and local averaging. The third feature extraction layer, which performs a second convolution (similar to how the first convolutional layer operates), consists of feature maps. Each feature map consists of neurons.

There are two training algorithms for CNNs.

The calculation formula of forward propagation process is

$$O_p = Fn(\ldots(F1(X_P W^1)\ldots)W^n). \tag{1}$$

In back propagation, let E_i be the error measure of the i'th sample and define the error measure of the neural network with respect to the entire sample set as

$$E = \sum E_i, \tag{2}$$

$$E_{\mathrm{i}} = \frac{1}{2} \cdot \sum_{j=1}^{m} (y_{pj} - o_{pj})^2. \tag{3}$$

2.1.2 *Structure of the CCN model of this article*

This article mainly considers the three elements of training speed, recognition accuracy, and memory consumption to build a new CNN model structure. The training speed reflects the timeliness of a model, the recognition accuracy reflects its accuracy, and the memory consumption reflects the performance of the model. If there are too few network layers, the ability to express data information is insufficient. If there are too many layers, the training time will be longer. Since more detailed features need to be extracted in facial expression recognition, the input image size is 48×48. In this paper, various factors are considered, and the input image size,

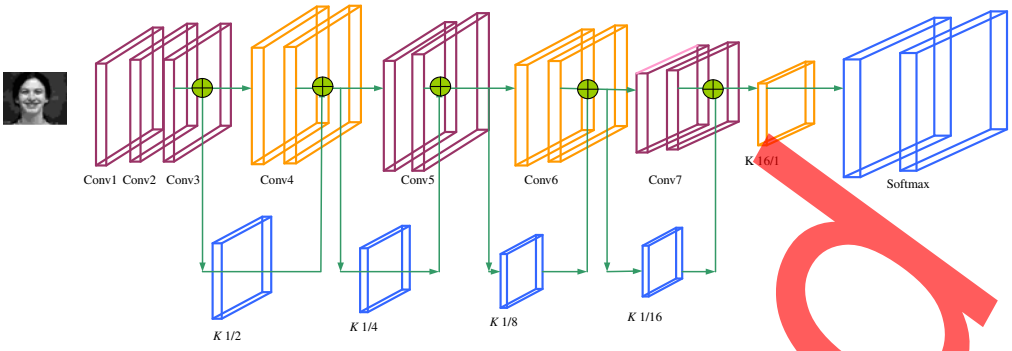

Fig. 2 The model structure of this article.

convolution kernel size, convolution step length, etc. are selected and set experimentally. Figure 2 shows the new CNN structure.

2.1.3 *Construction process of the CNN model in this article*

The image of the input layer is a 48×48 facial image, a 3×3 size convolution kernel is selected, and convolution processing is performed on the input image to obtain basic visual features. After detecting these functions, their correct position is not so important, and their relative position is not so important compared with other functions. This correct position is not only irrelevant, but the object of each formula is different, so it may cause problems. The model is predicted by the softmax activation function.

In the softmax function, let x be the input, the category is j, and the probability value is $p(y = j|x)$. Assuming that the function outputs a k-dimensional vector, the sum of the vectors is 1, and if the function is $h(x)$, then:

$$h(x) = \begin{bmatrix} p(y^{\mathrm{i}} = 1|x^{\mathrm{i}}; \theta) \\ p(y^{\mathrm{i}} = 2|x^{\mathrm{i}}; \theta) \\ \cdots \\ p(y^{\mathrm{i}} = k|x^{\mathrm{i}}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^i}} \begin{bmatrix} e^{\theta_1^T x^i} \\ e^{\theta_2^T x^i} \\ \cdots \\ e^{\theta_3^T x^i} \end{bmatrix}. \tag{4}$$

In the formula, $\theta_1, , , \theta_k$ are all parameters, and the cost function of softmax is:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1(\{y^i = j\}) \log \frac{e^{\theta_j^T x^i}}{\sum_{l=1}^{k} e^{\theta_1^T x^i}}\right] + \frac{\lambda \cdot \sum_{i=1}^{k}\sum_{j=0}^{n} \theta_{ij}^2}{2}. \tag{5}$$

In the formula, $1\{\cdot\}$ is a formula function, and $\lambda > 0$, the cost function will become a strictly convex function, and W has a unique solution at this time.

2.2 *Depth Separable Convolution*

The network uses four residual depth separable convolutions. Then the total multiplication of deep convolution is as follows:

$$\mathrm{DC} = M \times D_P^2 \times D_K^2. \tag{6}$$

The total number of multiplications for point-by-point convolution is as in Eq. (7)

$$\mathrm{PC} = M \times D_P^2 \times N. \tag{7}$$

As shown in Eq. (8), the total number of multiplications and trainable parameters is reduced by $\frac{1}{N} + \frac{1}{D_k^2}$

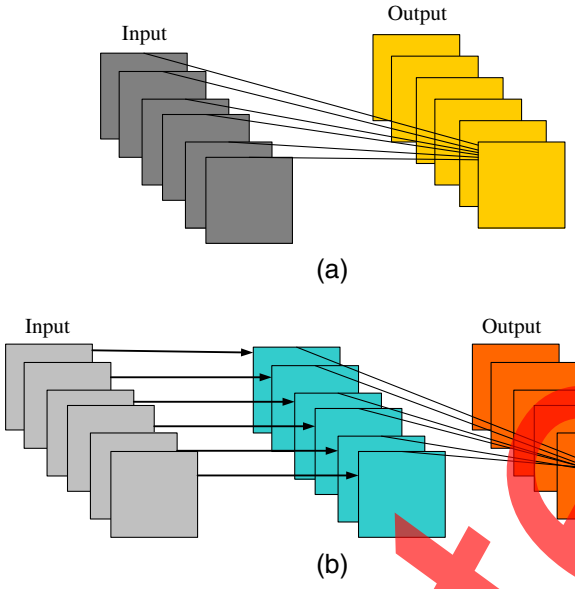

Fig. 3 The difference between (a) ordinary convolution and (b) depth separable convolution (b).

$$\frac{[\text{depthwise}][\text{separable}][\text{convolution}]}{[\text{standard}][\text{convolution}]} = \frac{M \times D_P^2 \times (D_K^2 + N)}{N \times D_P^2 \times D_K^2 \times M} = \frac{1}{N} + \frac{1}{D_K^2}. \tag{8}$$

Figure 3 shows a diagram showing the difference between depth separable convolution and ordinary convolution. The computational efficiency of depth-wise separable convolution is far superior to that of ordinary convolution.

2.3 *Residual Block and Pre-activated Residual Block*

Figure 4 shows an example of the residual module.

The structure in this paper uses a preactivated residual unit[23] when using the residual network. The preactivated residual module is shown in Fig. 5. The activation function after each hop the connection is moved to the inside of the residual block. Using batch normalization (BN) in preactivation makes training easier. At the same time, it also improves the generalization ability of the network, with ~58,000 parameters. Compared with the original CNN, it is reduced by 80 times, and compared with the Xception model, it is reduced by more than 300 times.

The study found that when the training samples approached infinity, the weights of the network after training converged to the really needed weights in probability. However, first, the number of training samples is limited, and second, if the number of training samples is too large, there will be an "overfitting phenomenon." The so-called overfitting phenomenon means that during the training process of the network, if too many special samples are input, the network will remember these cases and noises, and cannot grasp the true rules of the training samples, which will eventually lead to the failure of non-training samples. The input cannot give correct results, resulting in poor generalization ability. Therefore, when selecting the number of training

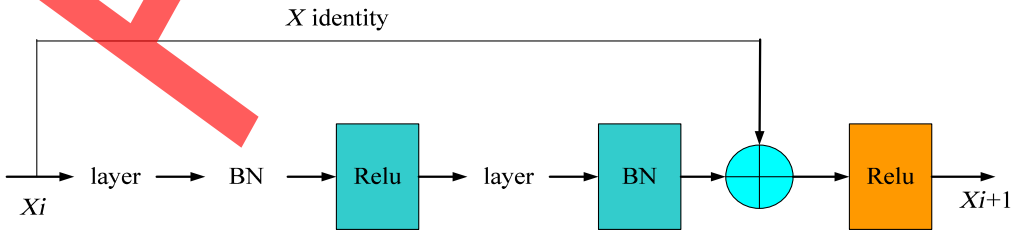

Fig. 4 Residual error module.

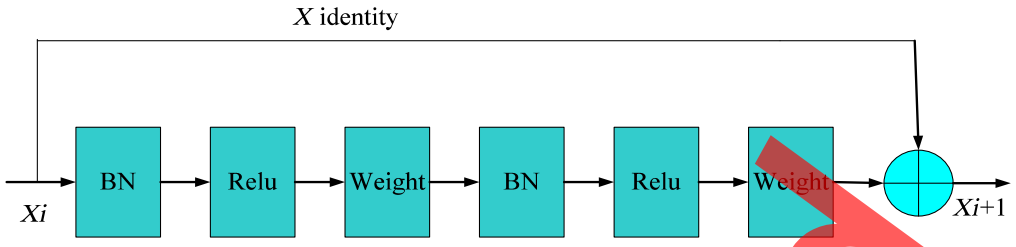

Fig. 5 Pre-activated residual unit.

samples, it is necessary to select a sufficient number so that the regularity of the samples can be found, and to prevent the occurrence of overfitting. In general, when the network structure is known, training samples that are several times larger than the number of weights can be selected to make the learning results reliable.

2.4 *Batch Standardization*

Bayesian classification algorithm is a classification method in statistics, which is a class of algorithms that use the knowledge of probability and statistics to classify. People who make the same expression may have great differences in skin color, appearance, and age, which will cause intra-class differences at will. The BN operation is applied to the network architecture in this paper, and the network focuses on more "real" images by reducing distortion.

To alleviate the phenomenon of covariate shift, this paper introduces BN.[24] For a layer $X(D))$ with d-dimensional input $x = (x(1), \dots)$, the normalization of each dimension will be

$$\hat{x}^k = \frac{x^k - E(x^k)}{\sqrt{\mathrm{Var}(x^k)}}. \tag{9}$$

Batch processing normalization reduces the internal covariance, and the impact on loss, local response standardization, and image distortion is also reduced. In the CNN in this article, each convolutional layer uses BN. The preactivated residual depth separable convolution model proposed in this paper can recognize facial expressions in a coarse to fine manner. It reduces the number of parameters and the computational cost required in identifying multiple features by means of deep separable convolution.

3 Facial Expression Recognition Experiment in the Classroom Environment Based on the Improved Xception Model

This section first discusses the database selected for the experiment in this article. Second, it describes the hardware and parameter settings for the experiment. Third, its description of the indicators for evaluating the accuracy of the system. Fourth, it gives the experimental results of this article and analyzes the performance. Finally, by comparing with the accuracy of existing algorithms, the superiority of this algorithm is proved.

3.1 *Data Set Selection*

A more famous facial expression data set is Extended Cohn-Kanada (CK+).[25] The CK+ dataset contains 123 test subjects and 327 marked expression image sequences.

The image resolution of the FER-2013 data set is 48×48. An example of the FER-2013 data set is shown in Fig. 6, and the number of images for each expression is given in Table 1. Most of the images in the database are obtained from web crawler browsers, with a certain degree of error.[26] The model in this paper will transform the data set image and perform data expansion operations such as inversion, rotation, and cutting.

Fig. 6 Sample image of FER-2013 data set.

Table 1 The number of images for each emotion in the FER-2013 database.

Emotion label	Emotion	Number of image
0	Anger	4593
1	Disgust	547
2	Fear	5121
3	Happy	8989
4	Sad	6077
5	Surprise	4002
6	Neutral	6198

3.2 *Experimental Environment Design*

TensorFlow is the second-generation artificial intelligence learning system developed by Google based on DistBelief. Its name comes from its own operating principle. Tensor (tensor) means N-dimensional array, Flow (flow) means calculation based on data flow graph, and TensorFlow is the calculation process of tensors flowing from one end of the flow graph to the other end. TensorFlow is a system that transmits complex data structures to artificial intelligence neural networks for analysis and processing.

Hardware platform Inter Xeon Bronze 3106, Nvidia Quadro P5000.

The method in this paper is based on the deep-learning algorithm model parameters shown in Table 2.

The rectified linear unit (ReLU) function greatly reduces the amount of data, reduces the amount of operation to a certain extent, and avoids the increase of the number of layers. Adam is an optimization algorithm that can be used to replace the traditional stochastic gradient descent algorithm, using a separate learning rate for each weight to update the network weights.[27] In this algorithm, the weight of the neural network is comprehensively studied using

Table 2 Model parameters.

Model parameters	Values
Total images	35,887
Activation	ReLU and softmax
Learning rate	0.1
Epoch	200
Optimizer	Adam
Loss function	Categorical cross-entropy

the first-order moment estimation and the second-order moment estimation of the gradient. The n'th order matrix of random variables is as follows:

$$m_n = E(X^n), \tag{10}$$

where m is the moment estimate and X is a random variable. Adam algorithm calculates the exponential average of the gradient as the parameter adjustment direction and calculates the exponential moving average of the square gradient to adjust the learning rate, as shown in the following formula as[28]

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{11}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \tag{12}$$

where m and v are moving averages, β_1 and β_2 are exponential decay rates (hyperparameters), and g represents the gradient. Adam algorithm adds error correction.[29] The deviation of the estimation of the first and second moments is corrected by the following formula as

$$m_t' = \frac{m_t}{1-\beta_1^t}, \tag{13}$$

$$v_t' = \frac{v_t}{1-\beta_2^t}, \tag{14}$$

$$w_t = w_{t-1} - \frac{n \cdot m_t'}{\varepsilon + \sqrt{v}}. \tag{15}$$

The loss function is used to optimize the classification model, the classification cross entropy function, is as follows:

$$L(y, y') = -\sum_{j=0}^{N} \sum_{i=0}^{N} (y_{ij} \cdot \log(y_{ij}')), \tag{16}$$

where y' is the predicted value; this function is used to compare the distribution of the predicted value and the actual value.

Select softmax as the activation function, which takes the vector Z of K as the input. The softmax function is as follows:[30]

$$y_i = S(z)_i = \frac{e^{zi}}{\sum_{j=1}^{C} e^{zj}}, \quad i = 1, 2, \ldots, C. \tag{17}$$

Among them, z is the output of the previous layer, as the input of softmax, the dimension is C, and yi is the probability that the predicted object belongs to the c'th category.

3.3 *Experimental Results*

3.3.1 *Accuracy and loss rate of the model*

The model in this paper uses the preactivated residual structure to build the model of the deep separable convolutional network. In this experiment, each experiment cycle is more than 100 times, and the average of all results is taken as the test result. The accuracy rates on the training set and validation set are shown in Fig. 7.

The loss rate of this model on the training set and validation set is shown in Fig. 8.

The proposed preactivation residual upgrade-reading separable convolution method was compared with the post-activation depth separable convolution method adopted by mini-Xception. In the experiment, it can be seen from the figure that at the beginning of training, the training error of the activated residual network decreases very slowly after use. The accuracy

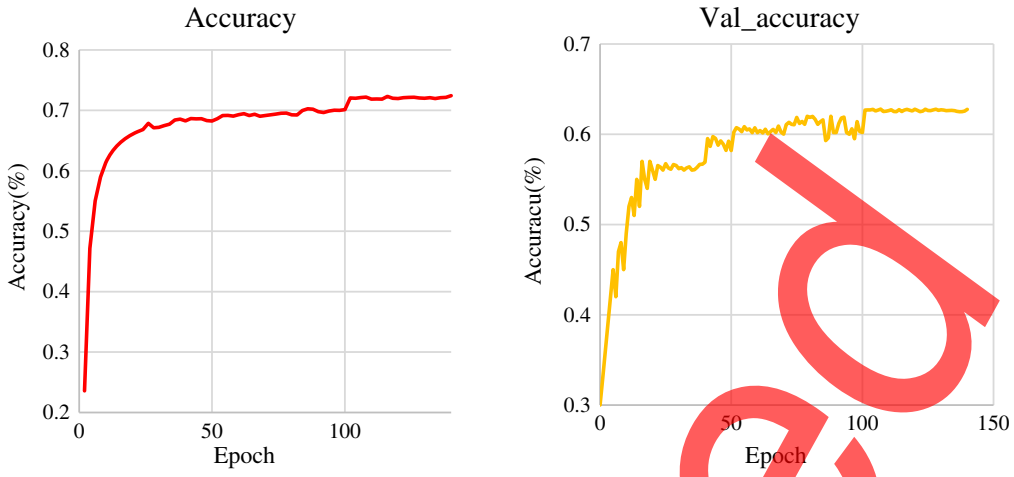

Fig. 7 Accuracy on training set and validation set.

Fig. 8 Loss rate on training set and validation set.

of the final model is low, the error of the preactivated residual network model is reduced faster during the training process, and the accuracy of the final model reaches the maximum value of 72.4%. In this paper, the preactivated network used in the improved residual depth separable convolution model can reduce the training loss very quickly to the minimum loss. The model in this paper is successful in the optimization of facial expression subrecognition for small data sets.

3.3.2 *Confusion matrix*

Among the expressions, the recognition accuracy of the model for the expression of happiness is the most accurate, with an accuracy rate of 91%. For the expression of "disst," the recognition accuracy of the model is 77%. Even if the disst data in the data set account for only 1.5%, the accuracy of the model is still so high, which reflects the effectiveness of the model. However, it can be seen from the figure that the recognition rate of the model is not always very high, and there will be many recognition errors, such as recognizing the image of sadness as the expression of "fear" and "anger" as the expression of "disgust." However, in general, the recognition accuracy of the model architecture proposed in this paper is 72.4%, which is a relatively high level, as shown in Fig. 9.

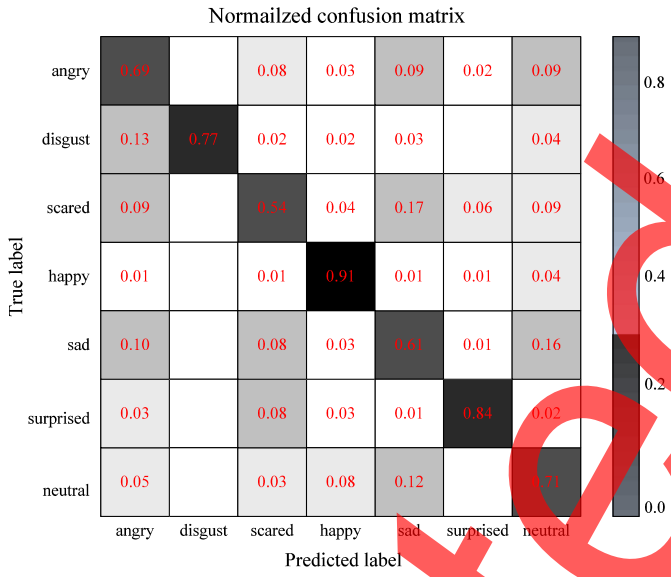

Fig. 9 Confusion matrix.

3.3.3 *Accuracy rate under different loss functions*

When a model is perfect (though it does not exist), its error is 0. When there is a problem with the model, whether the error is negative or positive, it deviates from 0. The closer the error is to 0, the better the model. The loss function selected in this article is softmax. To verify that the application of this loss function in the model can improve the accuracy of the model, we compare this loss function with the hinge function and measure the accuracy of the model. The result is shown in Fig. 10.

It can be seen from the figure that no matter how many iterations there are, the accuracy of the model under the softmax loss function is always greater than the hinge loss function. Under different iteration times, the average accuracy of the softmax loss function is 65.89%, and the average accuracy of the hinge function is 63.82%. Therefore, the softmax loss function is more suitable for the model in this paper than the hinge function.

Fig. 10 The accuracy of different loss functions.

Table 3 Comparison of accuracy and parameter values of different models.

Model	Precision	#Params (millions)
Simple CNN	0.53	0.64
Simpler CNN	0.54	0.60
Tiny Xception	0.56	0.02
Mini Xception	0.625	0.06
Big Xception	0.67	0.21
Xception	0.714	20.87
Our model	0.724	0.06

3.3.4 *Model accuracy and parameter analysis*

In this experiment, this model is compared with several other expression recognition models based on CNN to explore the number of parameters required by various models and the accuracy of expression recognition. It can be clearly seen that this model has the best classification accuracy and the least number of parameters. The results are shown in Table 3.

4 Discussion

Facial expression recognition is one of the fields with high research value. The feature extraction and classification of facial expression are used to judge human expression. At present, facial expression recognition is widely used in the fields of human–computer interaction, safe driving, intelligent monitoring, medical treatment, and so on. That proves that the understanding of facial expression has important research value. The previous expression recognition methods have low recognition accuracy and poor classification performance. However, with the development of deep-learning theory, expression recognition technology is also developing rapidly, and the recognition accuracy is also greatly improved. In this paper, a new convolution neural network model is proposed by improving the anomaly model with high accuracy of face recognition.

5 Conclusion

This paper solves the classification problem of small facial expression datasets and proposes an improved residual depth separable convolution lightweight CNN model. During data preprocessing, it first detects and extracts facial feature points and then cuts out a small data set of facial expressions to increase the training accuracy of the model. This paper also tested the performance of the model through experiments and found that the model has a good performance in recognition accuracy and loss rate. The model proposed in this paper is very effective for the recognition of small sample data sets of facial expressions. It can not only reduce the number of parameters, but also recognize facial expressions more accurately. Overall, the model achieves a certain accuracy and performance. The next step of the research is to improve and optimize the model architecture to improve the balance of sample data, reduce missing errors, and improve the classification performance.

Acknowledgments

This work was supported by the Science and Technology Development Plan Project of Jilin Province, the project name is Educational Robot Patent Information Analysis and Strategic Research (Grant No. 20190802025ZG). This work was also supported by the Education Robot Innovation team of Jilin Engineering Normal University.

References

1. R. Kaiser and K. Oertel, "Emotions in HCI: an affective e-learning system," in *Proc. HCSNet Workshop Use Vision Human-Comput. Interaction*, Vol. 56, Darlinghurst, Australia, pp. 105–106 (2006).
2. S. Saurav et al., "Hardware accelerator for facial expression classification using linear SVM," in *Advances in Signal Processing and Intelligent Recognition Systems*, S. Thampi et al., Eds., Vol. **425**, pp. 39–50, Springer, Cham (2016).
3. B. Wu and C. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access* **6**, 12451–12461 (2018).
4. P. Burkert et al., "Dexpression: deep CNN for expression recognition," arXiv:1509.05371 (2015).
5. T. Chang et al., "Facial expression recognition based on complexity perception classification algorithm," arXiv:1803.00185 (2018).
6. H.-D. Nguyen et al., "Facial emotion recognition using an ensemble of multi-level CNNs," *Int. J. Pattern Recognit. Artif. Intell.* **33**, 1940015 (2019).
7. J. Chen et al., "Automatic social signal analysis: facial expression recognition using difference convolution neural network," *J. Parallel Distrib. Comput.* **131**, 97–102 (2019).
8. X. H. Wang, A. Liu, and S. Q. Zhang, "New facial expression recognition based on FSVM and KNN," *Optik* **126**(21), 3132–3134 (2015).
9. C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: a comprehensive study," *Image Vis. Comput.* **27**(6), 803–816 (2009).
10. J. Zhou et al., "A method of facial expression recognition based on Gabor and NMF," *Pattern Recognit Image Anal.* **26**(1), 119–124 (2016).
11. P. Lucey et al., "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expressions," *in IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops*, IEEE (2010).
12. C. C. Hsieh et al., "Effective semantic features for facial expressions recognition using SVM," *Multimedia Tools Appl.* **75**(11), 6663–6682 (2016).
13. G. Fanelli et al., "Hough forest-based facial expression recognition from video sequences," in *Proc. 11th Eur. Conf. Trends and Top. Comput. Vision* (2010).
14. H. Jung et al., "Joint fine-tuning in deep neural networks for facial expression recognition," in *IEEE Int. Conf. Comput. Vision*, IEEE, (2015).
15. G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 915–928 (2007).
16. D. Acharya et al., "Covariance pooling for facial expression recognition," *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. Workshops*, Salt Lake City, UT, pp. 480–4807 (2018).
17. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
18. C. Szegedy and V. Vanhoucke, "Rethinking the inception architecture for computer vision," arXiv:1512.00567 (2015).
19. A. T. Lopes et al., "Facial expression recognition with CNNs: coping with few data and the training sample order," *Pattern Recognit.* **61**, 610–628 (2017).
20. S. Yang et al., "EdgeCNN: CNN classification model with small inputs for edge computing," arXiv:1909.13522 (2019).
21. A. T. Lopes, E. D. Aguiar, and T. Oliveira-Santos, "A facial expression recognition system using convolutional networks," in *28th SIBGRAPI Conf. Graphics, Patterns and Images*, IEEE (2015).
22. T. F. Cootes et al., "Active shape models their training and application," *Comput. Vis. Image Underst.* **61**(1), 38–59 (1995).
23. P. Liu et al., "Facial expression recognition via a boosted deep belief network," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1805–1812 (2014).
24. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *33rd Int. Conf. Learn. Represent.*, San Diego (2015).

25. D. Viet Sang, N. Van Dat, and D. P. Thuan, "Facial expression recognition using deep CNNs," in *9th Int. Conf. Knowl. and Syst. Eng.*, IEEE (2017).
26. F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1800–1807 (2017).
27. A. G. Howard et al. "MobileNets: efficient CNNs for mobile vision applications," arXiv:1704.04861 (2017).
28. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
29. K. He et al., "Identity mappings in deep residual networks," *Lect. Notes Comput. Sci.* **9908**, 630–645 (2016).
30. O. Arriaga and P. G. Ploger, "Real-time CNNs for emotion and gender classification," arXiv:1710.07557 (2017).

Jia Tian received her master's degree from Daqing Petroleum Institute. Now, she works at the College of Electrical Engineering of Jilin Engineering Normal University. Her research interests include artificial intelligence, educational robotics, and automation.

Jian Fang received his MS degree in control engineering from Jilin University, China. He is currently the dean of the Electrical Engineering School of Jilin Engineering Normal University. Now, he is studying for a PhD in the School of Mechanical and Electrical Engineering at Changchun University of Technology. His research interests includes artificial intelligence, educational robotics, and automation.

Yue Wu received her master's degree from Changchun University of Technology, China. Now, she works at the College of Automotive Engineering of Jilin Engineering Normal University. Her research interests include artificial intelligence, educational robotics, and automation.