IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2024, vol. 2, no. 3, pp: 47–57

Study on the Student Concentration in Class Based on Deep Multitasking Learning Framework

Jiangjiang Li[1], Lijuan Feng[1], and Jiaxiang Wang[1]

School of Electronic and Electrical Engineering, Zhengzhou University of Science and Technology
Zhengzhou 450064 China
1124905128@qq.com;857003841@qq.com
Corresponding author:Lijuan Feng
Received June. 26, 2024; Revised and Accepted August. 29, 2024

Abstract. In order to solve the problem of poor teaching quality caused by classroom teachers' inability to grasp students' dynamics in time, this paper designs a feedback system for classroom attention with the help of the research on expression recognition technology in deep learning. In the real-time analysis of expression, although the deeper deep learning network has more accurate recognition effect, there are drawbacks of too large model and too many parameters in the network training process. In this paper, we propose a student concentration algorithm that uses the Convolutional Block attentional module (CBAM) and the Local Binary Pattern (LBP) to reduce the number of parameters in the model by replacing the convolution with the depth-separable convolution LBP preprocessing enhances the feature validity of the input feature map and improves the training speed and accuracy of the model. The experimental results show that the algorithm has a good discriminating effect on expression recognition, and the model is small.

Keywords: Student concentration, CBAM, Local Binary Pattern, Deep learning.

1. Introduction

In daily life, facial expression can reveal people's inner activities, so the technology related to facial expression is widely used in intelligent education, companion robots, detection, medical treatment and other fields [1]. Since intelligent computers can use recognition algorithms to recognize facial expressions [2], interaction with humans will become more convenient. At present, facial expression recognition has been used as a common tool in all aspects of daily life. In school classes, students' facial expression can be captured to determine their concentration in class and monitor their learning status [3]. In terms of medical treatment, the pain degree can be judged by capturing the clinical pain expression of the elderly to achieve better surgical results [4]. In intelligent driving, the driver's expression also reflects whether it is in a tired state, if it is in a tired state 3, the computer should give timely warning and so on. In recent years, facial expression technology has attracted the attention of different research fields at home and abroad. The student expression recognition and attention analysis system based on classroom scene is one of the most eye-catching software systems in the society. It is a product based on student expression recognition technology, combined with modern database technology, and applied to primary and secondary schools [5,6].

Facial expression recognition is a branch of face recognition. Due to the proposal of industry 2.0 plan and the rapid development of computer computing power, this technology is widely used in intelligent classroom, fatigue detection, medical patient expression analysis and other fields. In the hospital, the patient's expression changes due to pain, and the doctor can not judge the pain degree easily by personal experience, so the expression recognition and analysis method of the patient's pain came into being; Parents pay more and more attention to their children's education, and the child's expression in the classroom reflects the child's knowledge absorption degree. Therefore, the application of facial expression recognition is very common and practical [7,8].

However, in a classroom setting, the large number of students presents a huge problem for the teacher to judge the concentration of the students. It is a common concern of schools to establish a perfect student expression recognition and concentration analysis system. A good student concentration analysis system based on expression recognition can not only enable teachers to better analyze students' concentration in class and make timely adjustments to class strategies, but also enable students to better explore their own concentration rules in class, improve class efficiency and indirectly improve all subjects grades [9].

Students' concentration analysis and application system based on expression recognition can easily make up for these shortcomings. Therefore, it is necessary to establish a perfect student concentration analysis system based on expression recognition. The application of artificial neural network technology, which is developing rapidly with computer power, to the student management system is an efficient and innovative way of student

performance management. Student attention analysis and application system based on expression recognition can play an immeasurable role in student management in schools [10,11].

However, there are many problems in facial expression recognition in the classroom.

1. Students' expressions are diverse, there is no unified open data set, and there is no unified standard to judge students' facial expressions in the industry.
2. Students' facial expressions may be obscured in class, resulting in inability to effectively recognize students' expressions.
3. Students' concentration can not be accurately evaluated and analyzed solely through expression recognition.

Based on the relevant research results of lightweight expression recognition framework and multi-face detection framework, this paper develops a set of students' learning concentration analysis system based on expression recognition and head posture combined with artificial neural network.

The deep learning neural network technology is used to build a student expression recognition model and a multi-face detection model, and the five key points returned by face detection are used to determine the student's head posture, and the expression results and the head posture value are combined to judge the student's concentration in class. In addition, face recognition technology based on neural network is used to identify students. The system can help primary and secondary schools to efficiently analyze and manage students' concentration, reduce teachers' class burden, provide teachers with effective student status evaluation indicators, and make school performance management more humanized and automated. The system can effectively reduce the burden of teachers in class, make up for the problems that teachers can not pay attention to each student, and achieve the effect of tracking each student's learning attention.

Learning expression [11] is an important implicit learning. Through the analysis of students' expressions, teachers can understand and grasp the overall learning state of students, and improve students' performance. Therefore, how to efficiently and accurately identify students' learning expressions in class has always been a hot topic.

In the study of students' expression recognition, in a few cases, students' psychological activities and emotional changes are known through students' self-feedback. Under normal circumstances, teachers of different disciplines listen to cross-class lectures and observe students' class effect, and analyze students' expressions and class effect. However, manual observation method will consume a lot of time and manpower, and students' self-report will be affected by subjective consciousness and the environment at that time, resulting in inaccurate judgment results. Kumar et al. [12] found that in order to study the changes of facial expressions, the face can be divided into 46 parts and each part is coded, thus creating a pioneer in expression recognition methods. However, when applied to complex practical problems, problems such as robustness and poor generalization will be generated. Therefore, in the field of wisdom education, it is urgent to develop a technology to identify students' expressions in the classroom and judge their concentration.

Kumar et al. [13] realized the detection problem by observing faces from multiple angles, and could detect faces in the case of multiple people and density, thus solving the occlusion problem to a certain extent. In 2018, in order to quickly detect faces, only convolutional neural networks with feature-free features are used to improve the detection efficiency of faces [14]. Kou et al. [15] improved the face detection algorithm, mainly by improving the design of anchor points to improve the accuracy of multi-face detection, so as to solve problems such as occlusion and Angle existing in the field of multi-face recognition. Hossain et al. [16] proposed an improved YOLO face detection model in order to overcome the uncertainties in the environment and realize the detection of multiple faces, thus realizing the results of multiple face detection in ordinary life scenes.

Many research methods are based on facial motion units to extract facial features. The attention mechanism has the property of focusing on the salient features of a picture. Therefore, a lightweight expression recognition algorithm based on attention mechanism is studied, and the faces detected in class are sent into the expression recognition algorithm model for analysis. Learn about students' facial expressions in class in order to better study students' learning status.

2. Proposed Model

While recognizing the identity of students in the classroom, it is necessary to recognize the expression recognition algorithm of students at this moment. The algorithm is a lightweight expression recognition network based on Xceptionnet [17,18], in which multiple attention mechanisms are embedded. Combined with the face detection network, the detected face is sent into the expression recognition network, and the expression state of the students is obtained. The overall network structure of expression recognition is shown in Figure 1.

In the network structure diagram, the RGB image is input, and the rough texture information of the image is extracted by LBP preprocessing method, which can reduce the time of model fitting in the later stage. This is

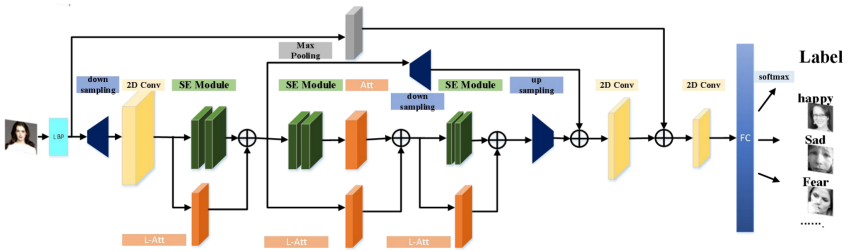

Fig. 1. Overall network structure

especially true when the number of images is large. Then the subsampling module extracts the image features. Then, a two-dimensional convolution module extracts the features of the image with Stride of 2, and scales to the appropriate size to adapt to the feature maps of local attention and global attention. Through this method of operation, the width and depth of the network will increase, and the training parameters of the network will naturally increase. In order to ensure the lightweight design of the whole network and apply to the mobile terminal, the deep separable convolutional module corresponding to local attention is designed, which greatly reduces the number of parameters during network training. In the process of network forward propagation, shallow features are obtained by maximum pooling to obtain global visual field features and deep features are combined point-to-point at pixel level. Finally, features are extracted through Dense connected layers, and then divided into 1024 categories by full connection. Softmax calculates the corresponding probabilities of seven categories of emojis.

2.1. LBP

The network front-end preprocessing adopts the traditional LBP method, which belongs to the texture extraction method of appearance features. The image texture features are calculated based on the numerical relationship between the surrounding pixel and the center pixel, which has a small improvement in recognition accuracy compared with the image normalization method. The detailed description is shown in Figure 2.

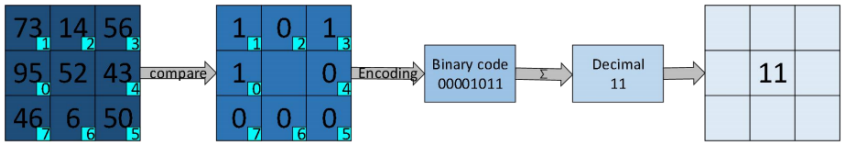

Fig. 2. LBP process

$$LBP_p = \sum_{0}^{p-1} 2^p \times s(x^k, i_p, i_c). \tag{1}$$

As shown in formula (1), taking the 3×3 grid as a region and the center of the LBP operator as a reference value, the result obtained by comparing the center value with the surrounding pixel value is the processing result of the LBP operator. Where i_c and $i_p, p = 0, 1, \cdots, 7$ represent the central pixel value of the grid and the pixel value around the central value, respectively, and a central position corresponds to eight neighboring pixels. As shown in Figure 2, the $0-th$ position is larger than the central value, resulting in "1"; the first position is smaller than the central value, resulting in "01"; the third position is larger than the central value, resulting in "011"; and so on, resulting in the binary format "00001011" around the central value, converted to decimal "19", which is the surrounding feature of the central pixel. As shown in formula (2), x^k represents the $k-th$ grayscale image. i_p represents one of the eight pixel positions around the center pixel $(p = 0, 1, 2, \cdots, 7)$. i_c indicates the center pixel position. $x^k_{i_p}$ represents the value of the pixel around the center pixel of the $k-th$ grayscale image. $x^k_{i_c}$ represents the value of the pixel in the center of the $k-th$ grayscale image. The overall formula is expressed on the $k-th$ grayscale map, and the intensity of the surrounding pixels $x^k_{i_p}$ and the center pixel $x^k_{i_c}$ of the grayscale image for a difference operation, if the difference is greater than 0, then the position will be taken as 1; Otherwise, the position is 0.

$$s(x^k, i_p, i_c) = 1, if x^k_{i_p} - x^k_{i_c} \geq 0. \tag{2}$$

2.2. Separable Convolution

Corresponding features are extracted from the input Feature map through convolution operations, and the size of pixels is changed between 0 and 1 by normalization method, and the dimension of input features is reduced by maximum pooling. After extracting features by convolution, regularization layer is used to integrate the values of output features, which is conducive to the fast convergence of the model. Multiple field of view features are obtained by stacking unsynchronously long convolution. After each two-dimensional convolution operation, image structural features are fully extracted through two convolutional modules, in which step size 2 is selected to increase the overall perception ability of the image in the early stage and provide more feature information for the later stage. It is close to the separable convolutional module, so as to reduce the capacity of the model. Separable convolution is shown in Figure 3. The separation convolution replaces the larger convolution kernel with the smaller one, which reduces the capacity of the model and improves the training speed of the whole network.

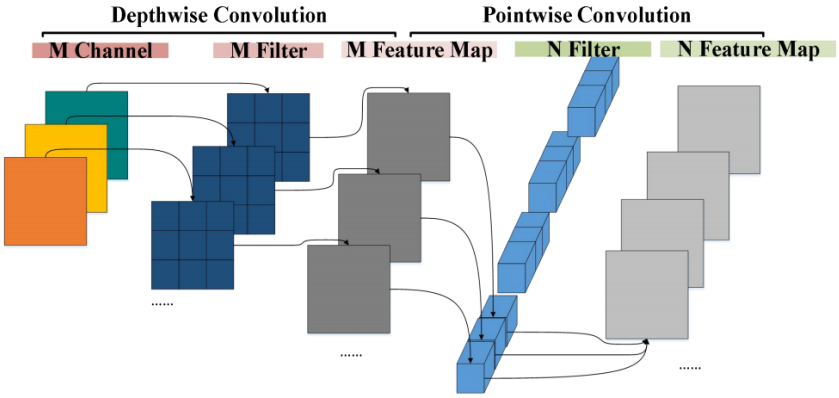

Fig. 3. Separable convolution diagram

Since expression recognition requires a high amount of information in feature extraction, it is convenient for subsequent networks to recognize more information in ordinary convolution. Therefore, cross-layer features of shallow networks are connected to deep networks through maximum pooling for fusion, which not only increases the flow of features, but also increases the acquisition degree of image information, and improves the detection performance of facial expressions in a small range. The deep stage of the whole network adopts a smaller receptive field, which is responsible for extracting richer details of the underlying feature map and further increasing the flow of feature information. The underlying information is shallow and the feature representation ability is not strong, which will cause certain classification misjudgment.

In the separation convolution diagram in Figure 3, each channel implements the convolution operation through a filter of the same size. Assuming that there are M filters, and the size of each Filter is 3×3, then the number of parameters in the convolution part is $3 \times 3 \times M = 9M$. The number of mapping channels obtained through channel convolution is the same as that obtained through input layer, which cannot effectively increase the deep dimension and does not make effective use of the feature information of different channels. Therefore, further point convolution is needed to realize the above steps. Students focus on the point convolution part in class. Assume that the size of kernel is $1 \times 1 \times M$, where M is the number of channels in the upper layer and the length of the point convolution. After M weighting coefficients are applied to each channel number, the feature information of different positions in the same space can be obtained by adding layer by layer. Assuming there are N filters, there will be N output maps. Since 1×1 convolution is used, the number of parameters involved in this step is $1 \times 1 \times 3 \times N = 3N$. Then the number of parameters obtained using separable convolution is $9M + 3N$. Suppose that the number of parameters obtained by processing the input graph with conventional convolution is $N \times 3 \times 3 \times M = 9MN$. The input in this paper is a one-dimensional grayscale graph, $M = 1$. By comparing $9N$ and $3N + 9$, it can be found that when $N > 2$, the number of parameters of ordinary convolution is much larger than the number of parameters of separable convolution.

2.3. Design of Attention Module

In the middle of the network, the attention mechanism is used to select locations with high representational ability conducive to classifying features, that is, the convolution results are input into the attention block, and the feature

information at the channel level is obtained through the channel dimension, and then the feature information at the spatial level is obtained through the spatial dimension. The features of the two dimensions are mutually utilized to make full use of the information of the feature map. The input in the attention module is the result of the output of the previous convolutional layer. This feature is first passed through a Channel Attention Module with convolution operations and activation functions, then fused with the initial feature map, and then passed through a Spatial Attention module with operations of different convolution sizes but with the same output feature map size Module), the final dot product to get the result. For the attention module in Channel dimension, the input feature Map goes through the pooled operation based on width and height respectively, Dense layer is used instead of convolution operation for point-by-point weighting and traversing of image features, and finally Sigmoid is used to activate it to generate the final Channel Map in channel dimension as shown in formula (5). Channel Map is fused with the input feature map of the previous stage to obtain the input feature of the next process. For the spatial attention module, the previous stage is characterized by the input of the spatial attention module. Before feature extraction and activation, the overall feature is obtained using a pooling operation and the two branches are spliced together through a Concat operation. Dense layers are used instead of convolution to extract features, as shown in formula (4). Finally, dot multiply the feature map and the features input into the module to get the final feature, as shown in formula (3). The module design is shown in Figure 4.

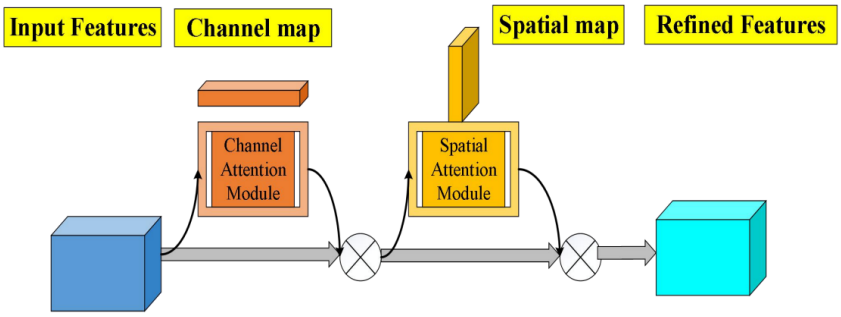

Fig. 4. Attention mechanism module

$$attention = MUL(S, C_{out}). \tag{3}$$

$$S = Dense(concat(avg(C_{out}), max(C_{out}))). \tag{4}$$

$$C_{out} = Sigmoid(Dense(add(avg(t), max(t)))). \tag{5}$$

C_{out} represents the feature map obtained after processing by the channel attention mechanism. S represents the feature map obtained after processing by the spatial attention mechanism, and t represents the image feature of the input, as shown in formulas (4) and (5). In the channel attention mechanism of the whole network, a series of operations are carried out on the input feature map, and finally activated by Sigmoid. In the spatial attention mechanism of the deep layer of the module, the features from the previous layer are pooled to extract universal features and convolution processing, and finally Dense convolution is used for feature extraction. As in formula (3), the dot-product operation of S and C_{out} is performed at last to obtain the attention diagram.

2.4. Lightweight Expression Recognition Network Design

In the backbone network using Xception, the local attention mechanism (L-Att) and global attention mechanism (Att) are embedded in the process of feature extraction, so that the main features and deep detailed features are combined to predict the final expression, which helps to improve the accuracy. The overall network structure is shown in Figure 1. The input to the network is a $48 \times 48 \times 1$ size feature map, after which the feature flow will be divided into two directions. One direction is to use the fusion of shallow information and deep information to enhance the feature display of the final feature map. The other direction is the main road direction, which extracts the deep features of the initial feature map through the convolution operation. Since a large number of convolution will increase the training burden of the network and increase the capacity of the model, in order to obtain a lightweight model, this paper replaces the convolution operation with a deep separable convolution. Experiments

show that this method can greatly reduce the capacity of the model and achieve overall lightweight. In order to enhance the ability of the whole network to capture features and acquire features with higher representativeness, the whole attention mechanism is added to the trunk of the network to strengthen the feature acquisition ability of the middle channel. The local attention mechanism is added to the bypass of the depth-separable convolution, and the features of the depth-separable convolution block are correlated to give full play to the effect of the local attention mechanism. The feature map is reduced to the size of the feature map corresponding to the deep network through the downsampling strategy, so as to complete the fusion operation, and prepare for the expression feature recognition of the deep feature map. The downsampling and upsampling strategies can be used to obtain both the reduced image and the feature from the previous feature map. This is also the method of channel fusion in this paper.

The network framework uses the local attention mechanism as the feature compensation of separable convolution, and each separable convolution module is paired with a local attention mechanism, which simultaneously broadens the width and depth of the network, avoids overfitting of data sets, improves the ability of the network to extract features, and enables the network to extend in a deeper direction. The feature addition of shallow layer and deep layer of the network adopts the form of feature graph ”Add”, and its features are linearly added. The obtained features are loaded in the deep network, and most features can be retained until the final classification stage. In this paper, the Softmax multi-classifier is used to compare the 1024-dimensional features obtained with seven categories of features. Finally, the category with higher score is selected by appropriate classifier as the expression category output. Due to the high complexity of expressions, including texture, light, etc., scanning images and extracting features using only a series of convolution layers will greatly increase the volume of the model, so this paper uses separable convolution to reduce the volume of the model. Compared with other networks, the volume can be reduced by about 10 times. The jump connection of 4 can improve the performance of the network, because the jump connection will use the initial picture information as the global information to guide the recognition direction of deep features.

3. Algorithm Experiment and Analysis

3.1. Expression Data Set Preprocessing

The data set for training the expression recognition network comes from FER2013 [19], which includes seven expressions. The content of this test set is composed of facial expressions taken by people of different ages in different angles, and the resolution is low, and many pictures are covered by hands, hair, scarves, etc., which is very challenging and very in line with the conditions of the real environment. A partial data set presentation is shown in Figure 5.

Fig. 5. FER2013 partial dataset

For the data set of students' expression, a large number of students' expression pictures are crawled from the web page of the browser by crawler technology. The search keyword is ”student classroom expression”. The pictures extracted contain a large number of cartoon pictures or fuzzy pictures, so it is necessary to manually screen the pictures containing students' faces, so that they can be used as the test data of this study.

When training the network, the input image is augmented by the data augmentation function of Keras to increase the number of images and improve the robustness of the final expression recognition network. In the pre-processing stage, the image is transformed to make the average value of each image parameter in the input network 0; The function of Keras is used to complete the normalization of the input feature map data. The white operation in Keras is used to highlight the outline and feature information of the facial expression in the picture, and to weaken other information to highlight the main features of the expression. The expression picture is rotated randomly at a certain Angle, and the maximum rotation Angle is set as 30 in this paper. Random rotation at a certain Angle; Randomly move the position of the picture by specifying the direction; Randomly scale the size of the picture; Perform random horizontal or vertical flip operations. Increase the data set by applying the above operations to achieve greater accuracy.

3.2. Experimental Parameter Setting

In the training stage of this experiment, the algorithm proposed in this paper is used to train the FER2013 dataset. The input is 48×48 gray image, and the accuracy obtained through network training is 58%, which is not good. Therefore, the attention mechanism is added to the whole network to improve the sensitivity of features, and the number of pictures from different angles is increased, which is sent into the network for training, and the Batch is set to 128, and the maximum cycle is 100 epochs. After experimental comparison, Adam optimizer was used to set epsilon value to $2e^{-8}$, and the recognition rate reached 70%. Using the ReduceLROnPlateau callback function in Keras to optimize the learning rate in the training process, the convergence speed is faster and the recognition rate is higher than that when the epoch is changed artificially. patience in the callback function is set to 50, that is, after 50 epochs, the model accuracy will not improve, and the learning rate will decrease according to the corresponding multiple.

3.3. Analysis of Experimental Results

FER2013 was used as the data set to obtain an expression recognition model after training the network. The confusion matrix based on this model is shown in Figure 6. In the visual regularization confusion matrix, it can be seen that the recognition rate of happy expression is the highest, reaching 92%, while the recognition rate of fear and sadness is lower, 60% and 61% respectively. After looking at the images of the two categories of fear and sadness in the data set, it is found that most of the features of fear and sadness are concentrated in the mouth and eyebrows. Secondly, the features of these two expressions are difficult to identify, so the recognition rate of the two expressions is low and similar.

The open source network framework models of "Vgg16", "Xception" and "LBP+CNN "combined with fine tuning were compared with the algorithm in this paper, and the resulting confusion matrix was numerical as shown in Figure 7, Figure 8 and Figure 9. It can be found that the algorithm model in this paper has a more accurate recognition rate for "happy" expressions; The accuracy of remaining expression recognition is also better than other networks, because the use of multiple attention mechanisms improves the width and depth of the network. At the same time, in terms of the accuracy of all expressions recognition, the accuracy of each expression discriminated by the algorithm in this paper is not very different. Compared with other networks, due to the addition of multiple attention mechanisms, in-depth feature extraction of various expressions can be achieved, thus obtaining higher expression recognition accuracy. Thus, the validity of the proposed framework for classification problems is verified compared with lightweight and common convolutional networks.

With FER2013 as the data set and batch size set to 128, a comparison table of training accuracy of different networks was obtained under the same optimizer and loss function, as shown in Table 1 below.

Table 1. Performance comparison of different networks

Model	Running time	Size	Accuracy/%
Xception	2h33m	25M	63
ResNet50	3h11m	366M	61
ASCNN	2h15min	280M	72.7
Proposed	43m	22M	65

In order to improve the training speed and keep the model lightweight, the algorithm takes out eight separable convolution layers of the original Xception and keeps three. The stackable residual connection of ResNet50 contributes to the feature extraction accuracy of the network, but increases the burden of network training, increases the model capacity, and lengthens the convergence time required for training the model. Therefore, in this paper, attention mechanism and cross-connection operation are added to improve the recognition accuracy of expressions while keeping the model capacity small. In addition, compared with ASCNN [20], the training time of the proposed algorithm can be greatly shortened. Although the accuracy of the original Xception network is two percentage points higher than that of the paper, the runtime and model size are much higher than that of the paper algorithm and are not suitable for lightweight practical applications.

4. Conclusion

In this paper, a student expression recognition algorithm based on depth-separable convolution combined with multiple attention mechanisms is proposed. First, the whole algorithm framework is described, and then the

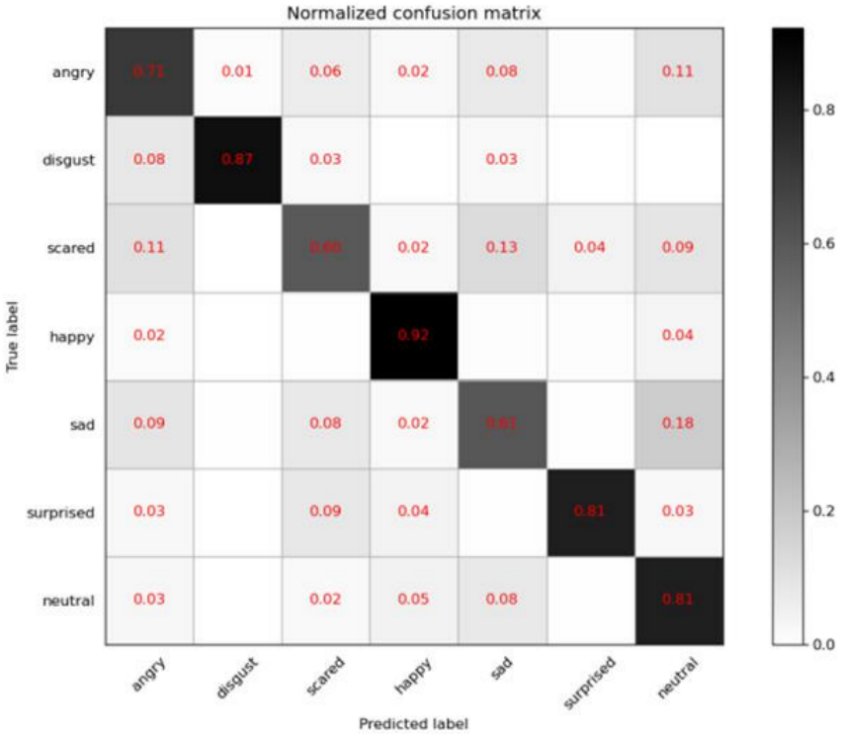

Fig. 6. Confusion matrix of FER2013

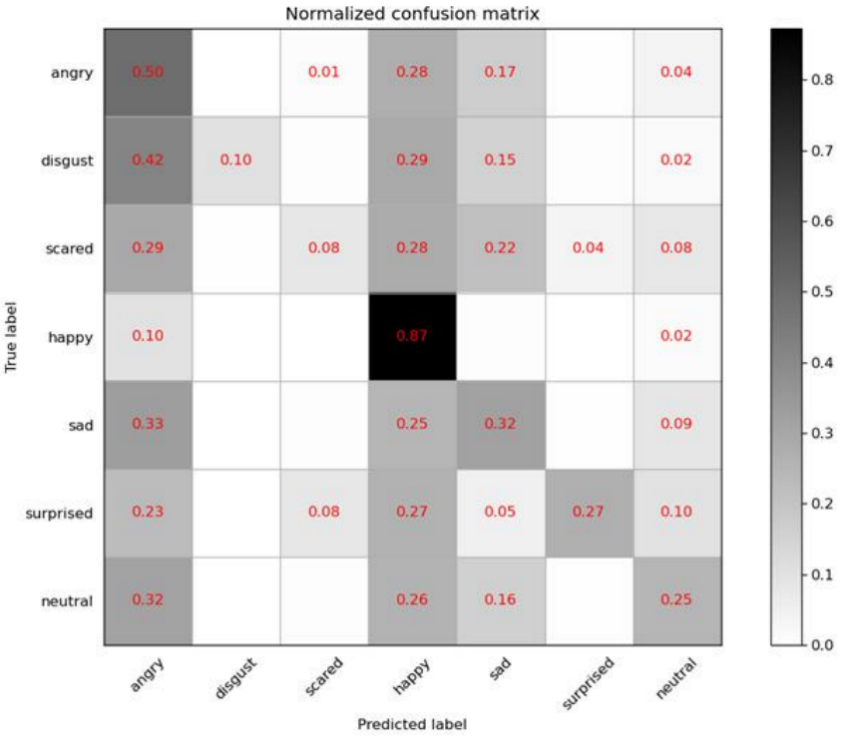

Fig. 7. Confusion matrix under Vgg16 model

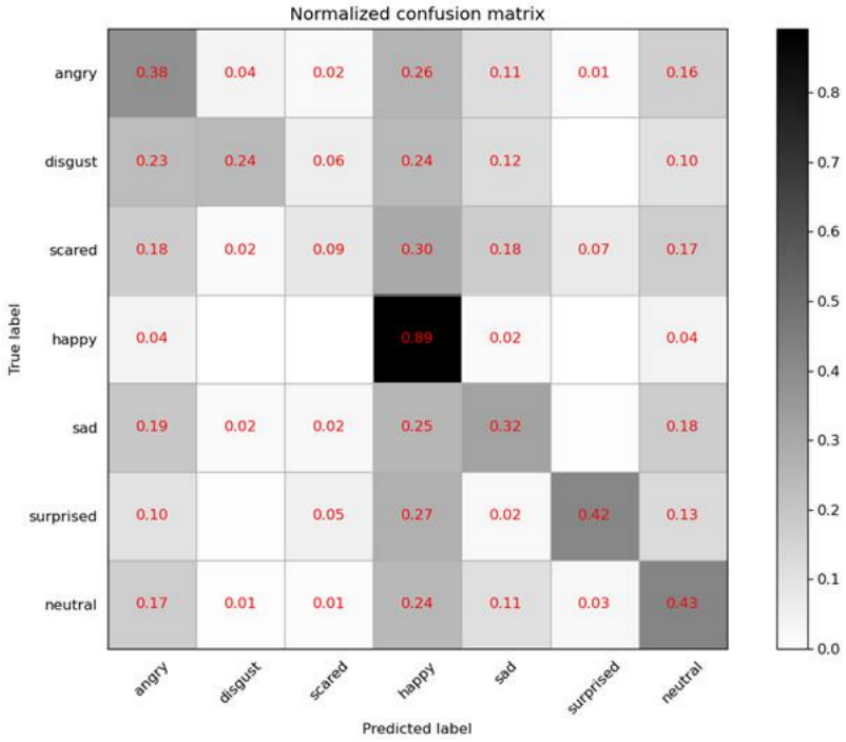

Fig. 8. Confusion matrix of Xception model

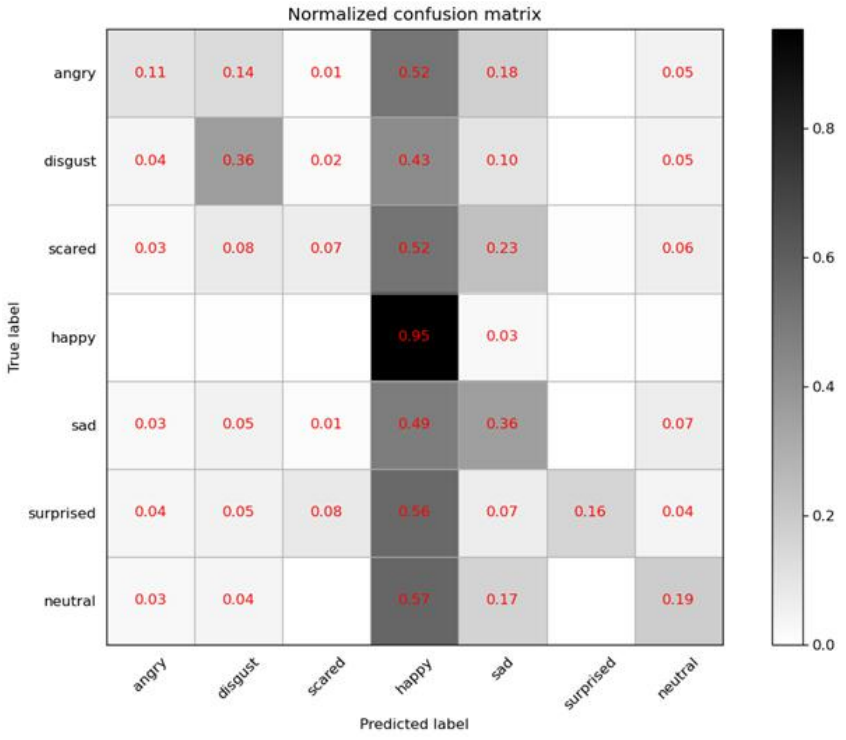

Fig. 9. Confusion matrix of LBP + CNN model

method description of each part is expanded step by step. The two modules of attention mechanism and depth-separable convolution are introduced. Then, the experimental data set used by the algorithm is introduced. Finally, the experimental comparison and effect display of the algorithm are carried out under the data set. The experimental results prove that the expression recognition algorithm proposed in this paper has the advantage of more lightweight and better recognition rate.

5. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. This paper was supported by Key research Project of higher education institutions in Henan Province (Project: Name: A Study on Students' concentration in Class Based on Deep Multi-task Learning Framework; Project No. 23B413004).

References

1. Adyapady R R, Annappa B. A comprehensive review of facial expression recognition techniques[J]. Multimedia Systems, 2023, 29(1): 73-103.
2. Karnati M, Seal A, Bhattacharjee D, et al. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-31.
3. Yin S, Wang L, Teng L. Threshold segmentation based on information fusion for object shadow detection in remote sensing images[J]. Computer Science and Information Systems, 2024. doi:10.2298/CSIS231230023Y.
4. Yin S, Li H, Sun Y, et al. Data Visualization Analysis Based on Explainable Artificial Intelligence: A Survey[J]. IJLAI Transactions on Science and Engineering, 2024, 2(2): 13-20.
5. Yin S, Li H, Laghari A A, et al. An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G internet-of-everything[J]. IEEE Internet of Things Journal, vol. 11, no. 18, pp. 29402-29411, 2024. doi: 10.1109/JIOT.2024.3353337.
6. Alhussan A A, Talaat F M, El-kenawy E S M, et al. Facial Expression Recognition Model Depending on Optimized Support Vector Machine[J]. Computers, Materials & Continua, 2023, 76(1).
7. Wu Y, Zhang L, Gu Z, et al. Edge-AI-driven framework with efficient mobile network design for facial expression recognition[J]. ACM Transactions on Embedded Computing Systems, 2023, 22(3): 1-17.
8. Singh R, Saurav S, Kumar T, et al. Facial expression recognition in videos using hybrid CNN & ConvLSTM[J]. International Journal of Information Technology, 2023, 15(4): 1819-1830.
9. Xu C, Zhu J, Zhang J, et al. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6609-6619.
10. Mao J, Xu R, Yin X, et al. POSTER++: A simpler and stronger facial expression recognition network[J]. arxiv preprint arxiv:2301.12149, 2023.
11. Jiang M, Yin S. Facial expression recognition based on convolutional block attention module and multi-feature fusion[J]. International Journal of Computational Vision and Robotics, 2023, 13(1): 21-37.
12. Kumar HN N, Kumar A S, Prasad MS G, et al. Automatic facial expression recognition combining texture and shape features from prominent facial regions[J]. IET Image Processing, 2023, 17(4): 1111-1125.
13. Liguo Wang, Yin Shoulin, Hashem Alyami, et al. A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images [J]. Geoscience Data Journal, vol. 11, no. 3, pp. 237-251, 2024. https://doi.org/10.1002/gdj3.162.
14. Kumar S, Sagar V, Punetha D. A comparative study on facial expression recognition using local binary patterns, convolutional neural network and frequency neural network[J]. Multimedia Tools and Applications, 2023, 82(16): 24369-24385.
15. Kou G, Pamucar D, Diner H, et al. From risks to rewards: A comprehensive guide to sustainable investment decisions in renewable energy using a hybrid facial expression-based fuzzy decision-making approach[J]. Applied Soft Computing, 2023, 142: 110365.
16. Hossain S, Umer S, Rout R K, et al. Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling[J]. Applied Soft Computing, 2023, 134: 109997.
17. Jiang Y, Yin S. Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment[J]. Computer Science and Information Systems, 2023, 20(4): 1869-1883.
18. Fan Y, Li H, Sun B. Cycle GAN-MF: A Cycle-consistent Generative Adversarial Network Based on Multifeature Fusion for Pedestrian Re-recognition[J]. IJLAI Transactions on Science and Engineering, 2024, 2(1): 37-44.
19. Samadiani N, Huang G, Cai B, et al. A review on automatic facial expression recognition systems assisted by multimodal sensor data[J]. Sensors, 2019, 19(8): 1863.
20. Xu C, Zhu J, Zhang J, et al. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6609-6619.

Biography

Jiangjiang Li is with the School of Electronic and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is cloud computing, computer application and AI.

Lijuan Feng is with the School of Electronic and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.

Jiaxiang Wang is with the School of Electronic and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.