Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.022312

ARTICLE

Facial Expression Recognition Based on Multi-Channel Attention Residual Network

Tongping Shen[1,2,*] **and Huanqing Xu**[1]

[1]School of Information Engineering, Anhui University of Chinese Medicine, Hefei, China

[2]Graduate School, Angeles University Foundation, Angeles City, Philippines

*Corresponding Author: Tongping Shen. Email: shentp2010@ahtcm.edu.cn

Received: 04 March 2022 Accepted: 13 May 2022

ABSTRACT

For the problems of complex model structure and too many training parameters in facial expression recognition algorithms, we proposed a residual network structure with a multi-headed channel attention (MCA) module. The migration learning algorithm is used to pre-train the convolutional layer parameters and mitigate the overfitting caused by the insufficient number of training samples. The designed MCA module is integrated into the ResNet18 backbone network. The attention mechanism highlights important information and suppresses irrelevant information by assigning different coefficients or weights, and the multi-head structure focuses more on the local features of the pictures, which improves the efficiency of facial expression recognition. Experimental results demonstrate that the model proposed in this paper achieves excellent recognition results in Fer2013, CK+ and Jaffe datasets, with accuracy rates of 72.7%, 98.8% and 93.33%, respectively.

KEYWORDS

Facial expression recognition; channel attention; ResNet18; dataset

1 Introduction

Facial expression is one of the most important characteristics to show the human psychological state. Psychologist Mehrabian has shown that facial expression accounts for 55% of emotional expression, which is one of the important characteristics of emotional communication [1]. Psychologist Paul Ekman has found that the facial expressions, physiological and behavioral responses of six basic emotions: happiness, anger, surprise, fear, disgust and sadness through his research [2]. Therefore, facial expression recognition (FER) research has important social value and application value. Important research results have been continuously achieved in facial expression recognition and applied in the fields of intelligent teaching [3], human-computer interaction [4], intelligent monitoring [5], safe driving [6], medical diagnosis [7], and so on.

The facial expression recognition algorithm is mainly composed of three parts: image preprocessing, image hierarchical feature extraction and expression classification, and in which image feature extraction directly affects the accuracy of the expression classification algorithm. The traditional expression recognition methods are through manual extraction of image features, such as Gabor wavelets [8], histogram of oriented gradients [9], principal components analysis (PCA) [10], Haar features [11], and Support Vector Machine (SVM) algorithm.

Traditional expression feature extraction algorithms rely too much on manual features, and the algorithms are vulnerable to external interference. In the process of expression feature extraction, the algorithm is prone to lose important features, which affects the accuracy of expression recognition.

In 1989, LeCun et al. [12] proposed the concept of Convolutional Neural Network (CNN), which improved the recognition rate of handwritten characters. In 2012, Hinton et al. [13] designed the AlexNet deep neural network. In 2014, Simonyan et al. [14] proposed VGGNet network.

These neural network models achieve good classification accuracy by deepening the level of the network. However, with the deepening of the network level, the problem of gradient disappearance is easy to occur, which affects the recognition effect of the model. He et al. [15] proposed the residual net structure, which effectively solves the contradiction between neural network depth and recognition accuracy. In order to further improve the expression recognition effect, many researchers add attention mechanism to the facial expression recognition model and combine it with convolution neural network, such as Squeeze and Excitation Networks (SENet) [16], Convolutional Block Attention Module (CBAM) [17] and self-attention mechanism [18], etc.

Liu et al. [19] proposed an SG-DSN network structure, which introduced a two-stream network with stacked graph convolution attention block (GCAB) to automatically learn discriminant features to express facial expressions from organized graphs. Li et al. [20] proposed a special and lightweight facial expression recognition network Auto-FERNet, which is searched automatically by the differentiable neural architecture search model directly on the FER dataset. Li et al. [21] used ResNet-50 as the network infrastructure. Features are extracted by convolution neural network, and BN and activation function ReLU are used to improve the convergence ability of the model. Pham et al. [22] focused on the deep architectures with attention mechanism, combining the deep residual network with Unet-like architecture to produce a residual masking network. Lai et al. [23] increased the network depth and alleviated the problem of gradient disappearance by adding residual connections to the VGG network. Niu et al. [24] integrated the CBAM attention module into the surface part expression recognition model and improved the experimental results.

These improved and fused expression recognition algorithms further improve the expression recognition accuracy, but with the proposed deep neural network and its various variants, the neural network depth and structure become more and more complex, and the model parameters increase rapidly, it leads to some problems such as gradient explosion and long training time in model training.

Different from the above expression recognition framework, we try to design a simple self-attention mechanism module and integrate it into the convolution neural network model. In this paper, we design a channel-based multi-head self-attention mechanism module MCA, and use the transfer learning algorithm to integrate the MCA module into the ResNet18 backbone network, the model structure is MCA-Net. The model is divided into three parts: image feature pre-extraction, channel-based self-attention mechanism and local multi-head structure. In the convolution layer of the network, we use the transfer learning algorithm to pre-train the parameters of the convolution layer, and use the residual neural network to extract facial expression image features. Based on the channel self-attention mechanism, the attention mechanism highlights important information and suppresses irrelevant information by assigning different coefficients or weights. The self-attention mechanism makes the model focus on information such as location or channel, thus producing more indicative features. In this paper, the channel self-attention mechanism is used in cooperation with convolution, and the average pooling and maximum pooling are used to distinguish queries and keys, and then they are sent into a shared linear embedding layer to reduce the data dimension and complexity of the whole model.

Therefore, we propose a facial expression recognition algorithm based on the combination of Resnet18 network structure and multi-head channel attention mechanism. First, a multi-head channel attention module is designed to extract deep-level features of expression images and integrate the MCA module into the ResNet18 network structure; It uses global average pooling layer (GAP) instead of fully connected layer to simplify model parameters and prevents overfitting; finally, the expression classification is performed by a Soft-max classifier to improve the generalization ability of the model.

The main contributions of this paper include the following:

(1) The paper proposes a multi-channel attention mechanism module (MCA), which can extract deep features of facial expression images, improve the representation ability of classification feature vectors, and help the model make better decisions.

(2) The MCA module is integrated into the ResNet18 network structure, and the global average pool layer is used to replace the full connection layer in the model output phase, which simplifies the model parameters, prevents overfitting, and improves the generalization of the network.

(3) We use the channel self-attention mechanism in concert with global convolution. The attention mechanism highlights important information and suppresses irrelevant information by assigning different coefficients or weights, and a global convolutional approach for the entire numerical tensor, represents each architectural design parameter of this network in terms of a single high-order tensor pattern, significantly reducing the number of parameters.

(4) On the three public datasets, we propose MCA-Net structure. It also achieves the most advanced results on several benchmarks.

2 Related Work

2.1 Feature Extraction

The effect of facial expression image recognition depends on image feature extraction. Traditional image feature extraction methods mainly include Gabor, PCA and so on. Zhao et al. [25] combined LBP and TOP algorithm for image feature extraction, and used the SVM classification algorithm for image classification. Shan et al. [26] combined LBP and AdaBoost algorithm to extract features of facial expression images. Luo et al. [27] proposed an improved PCA algorithm, which first uses the PCA algorithm to extract the global feature information of the image, and then uses the LBP algorithm to extract the key region feature information of the facial expression image. Kumars et al. [28] proposed an optimized LBP algorithm, which uses weighted projection to extract facial expression image feature information, and achieved good recognition results. Sahaa et al. [29] fused the feature space algorithm with the PCA algorithm and achieved good results in the process of facial expression recognition. Bougurzif et al. [30] proposed a pyramid multi-level facial feature algorithm through the extraction of manual features and deep features. Qian et al. [31] proposed a facial expression recognition method based on LGRP and multi-feature fusion to solve the problem of redundant information and single feature in the Gabor filter.

Traditional image feature extraction algorithms rely too much on manual rules, so it is easy to lose the deep feature information of classified images. In the depth learning method, after image preprocessing, the depth neural network can automatically extract the deep image features of the classified image, learn more dimensional feature information, and improve the effect of image classification.

In 2012, Krizhevsky et al. [13] proposed the AlexNet model to further deepen the neural network level and learn more image feature information. In 2014, the Oxford University team optimized the AlexNet structure and proposed the VGG model to further deepen the neural network level, obtain more image feature information, and improve the classification effect [14]. In 2016, in order to solve the problem of network degradation caused by deepening network levels, He et al. [15] proposed the residual network structure, which effectively solved the contradiction between neural network depth and recognition accuracy.

The above neural network methods focus on the high-level semantic information of facial expressions and ignore the local feature information. In this paper, a local multi-head structure is proposed, which splits a large high-dimensional single head into n multiple heads, which can work on a lower dimension. When the global image is divided into smaller local images, the local images have more advantages than the global images, and pay more attention to the local feature information of facial expressions.

2.2 Attention Mechanism

In recent years, the attention mechanism has been widely used in the field of natural language processing and computer vision, and has quickly attracted the attention of researchers. The self-attention (self-attention) mechanism proposed by the Google team in 2017 has become a research hotspot of neural network attention, and has achieved good results in various tasks [32].

Subsequently, many researchers tried to introduce self-attention mechanism into computer vision, but did not achieve breakthrough results. In computer vision, attention mechanisms usually include global attention, spatial attention, channel attention, self-attention, and independent attention and so on. Different researchers integrate the attention module with the CNN network structure individually or in combination, and achieve good results. Xu et al. [33] proposed a medical image classification algorithm based on global attention module. The global attention module identifies and extracts the key regions of medical images, and then sends them to the standard convolution neural network. Hu et al. [16] proposed a new image recognition structure SE-NET, which enhances the accuracy by modeling the correlation between feature channels and strengthening important features. Woo et al. [17] proposed the CBAM module, which is composed of spatial and channel attention mechanism modules in turn. Chen et al. [34] combined spatial and channel attention modules and proposed an SCA-CNN model for image subtitle recognition with good results.

Dosovitskiy et al. [35] proposed the ViT structure, which is an independent spatial attention structure, in which the transformer's inputs are patches extracted from the tensor image. Dai et al. [36] proposed an independent model of spatial self-attention mechanism, which combines convolution and attention mechanism. ViT and CoAtNet models have achieved excellent results on ImageNet, but the complexity and cost of these two models are very high. So, it is necessary to pre-train the JFT-3B data set containing 3 billion images.

Different from the above attention mechanism research work, we used the channel self-attention mechanism in concert with global convolution. The attention mechanism highlights important information and suppresses irrelevant information by assigning different coefficients or weights. We decided to simplify the model by differentiating query and key respectively with average and max pooling in order to enhance the processing of input information at different scales. We used a global convolutional approach for the entire numerical tensor, and represented each architectural design parameter of this network in terms of a single high-order tensor pattern, significantly reducing the data dimension and complexity of the whole model.

3 The Proposed Framework

We proposed an end-to-end deep learning algorithm to classify the emotion images based on multi-channel attention network. Due to the small number of classes for facial emotion datasets, we found that using the ResNet18 residual network structure and the attention mechanism could achieve more excellent results than SOTA models for several datasets.

In the process of facial expression recognition, the entire facial picture does not need to be recognized. Based on the expression classification features, it is only necessary to recognize and classify the expression picture information in specific regions, such as the eyebrows always appearing above the eyes. Therefore, we used the attention mechanism module to obtain information about the special regions in the facial emoticon images.

The structure of the model proposed in this paper is shown in Fig. 1. We added the attention mechanism module to the residual network. By loading pre-trained ResNet18 model parameters, the training speed and effectiveness of the model in this paper could be improved.

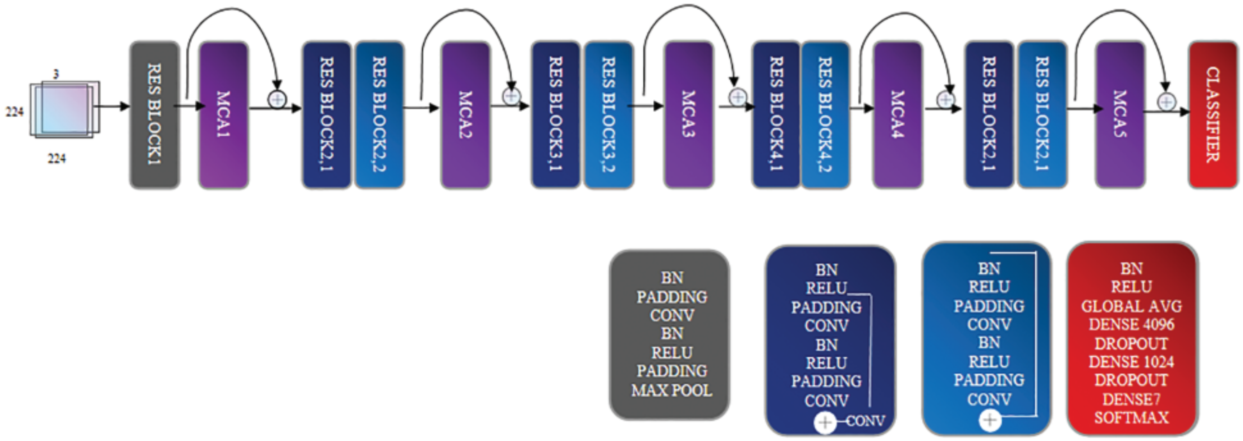

Figure 1: The proposed model architecture

3.1 Residual Module

The first proposed residual network is shown in Fig. 2, which effectively solves the contradiction between neural network depth and recognition accuracy.

When a neural network reaches a certain depth, the output x of that layer is already optimal, and further deepening the network will result in degradation. In a conventional convolutional neural network, it is difficult to ensure the weight of the next layer network $\mathrm{H}(\mathrm{x}) = \mathrm{F}(\mathrm{x}) + \mathrm{x}$. In the residual structure, when the network structure is designed as $\mathrm{H}(\mathrm{x}) = \mathrm{F}(\mathrm{x}) + \mathrm{x}$, the identity mapping $\mathrm{H}(\mathrm{x}) = \mathrm{x}$ of the next layer is changed into $\mathrm{F}(\mathrm{x}) = \mathrm{H}(\mathrm{x}) - \mathrm{x}$. The residual function $\mathrm{F}(\mathrm{x})$ only needs to update a small part of the weight of $\mathrm{F}(\mathrm{x})$. It is more sensitive to output changes, and the parameters are adjusted more widely, which can speed up the learning speed and improve the optimization performance of the network. The formula of residual structure is as follows:

$$\mathrm{y} = \mathrm{F}(\mathrm{x}, \{\mathrm{W}_i\}) + \mathrm{W}_s x \quad (1)$$

where W_s is mainly a $1*1$ convolution used to match the channel dimensions of the residual structure model input x and model output y. $F(x, \{W_i\})$ is the residual mapping that the network needs to learn. When the residual structure has the same input and output dimensions, the definition is as follows:

$$y = F(x, \{W_i\}) + x \tag{2}$$

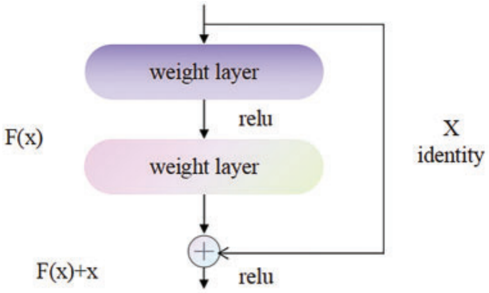

Figure 2: The residual network structure

The input information x is added to the feature calculation process, combined with the feature information of the upper layer to enrich the feature extraction of the network layer. Through the residual structure design, the degradation problem in the process of deep structure network training can be well solved without adding additional parameters and calculation. It also increases the training speed of the model and improves the training efficiency results.

3.2 Multi-Head Channel Attention Module

In the process of a large amount of input information in the neural network model, we used the attention mechanism to improve the efficiency of the neural network by selecting only some key input information for processing [37].

The attention mechanism function is calculated by first calculating the similarity or correlation between Query and each Key, getting the weight coefficient of each Key corresponding to Value, and then weighted summing the Value. The final attention value is obtained, and the structure is shown in Fig. 3.

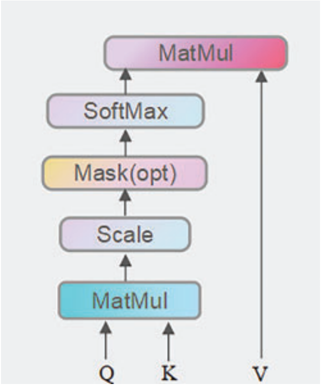

Figure 3: The self-attention network structure

$$\text{Attention}\,(\text{Q}, \text{K}, \text{V}) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V} \tag{3}$$

where Q, K, V represents three matrices calculated by the same input and different parameters. $\sqrt{d_k}$ is the k-dimensional adjustment smoothing factor to prevent the multiplication result from being too large. The SoftMax() function normalizes the result to a probability distribution and finally multiplies the matrix V to output the result.

In the MCA module proposed in this paper is shown in Fig. 4, the relevant parameters of the model are first defined: n represents the number of multiple heads, s represents the size of the convolution kernel, and d represents the embedding dimension of each head. $x \in \mathbb{R}^{H,W,C}$ is the input vector, H is the image height, W is the image width, and C is the number of channels, where HxW is required to be divisible by n.

$$\text{Q} = \text{AvgPool}_{p,1}\,(x) \in \mathbb{R}^{H,W,C} \tag{4}$$

$$\text{K} = \text{MaxPool}_{p,1}\,(x) \in \mathbb{R}^{H,W,C} \tag{5}$$

$$\text{V} = \text{AvgPool}_{3,1}\,(x) \in \mathbb{R}^{H,W,C} \tag{6}$$

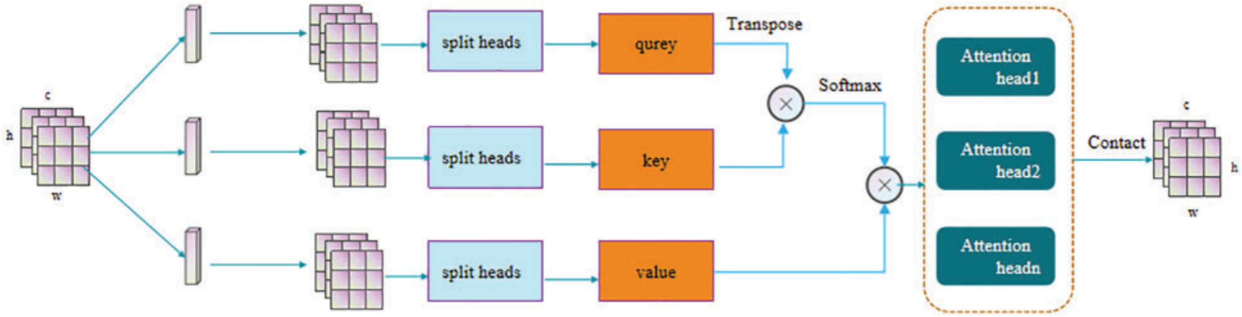

Figure 4: The proposed multi-head channel attention for the ResNet18

We slice the values of the input vectors Q, K, V into n equal parts, and reshape the segmented tensor as follows:

$$\text{q}_h = [\text{SplitHeads}\,(\text{Q})]_h \in \mathbb{R}^{C,(HxW)/n} \tag{7}$$

$$\text{k}_h = [\text{SplitHeads}\,(\text{K})]_h \in \mathbb{R}^{C,(HxW)/n} \tag{8}$$

$$\text{v}_h = [\text{SplitHeads}\,(\text{V})]_h \in \mathbb{R}^{C,(HxW)/n} \tag{9}$$

Each head operates on the (q_h, k_h, v_h) vector separately, and n full connection layers can be obtained. Among them, the query, key and value all come from the same tensor. In order to save model space, a shared linear embedding layer is constructed with weights $\text{w}_{1,h} \in \mathbb{R}^{C,(HxW)/n,d}$ and biases $\text{b}_{1,h} \in \mathbb{R}^d$.

$$\tilde{\text{q}}_h = q_h \cdot w_{1,h} + b_{1,h} \in \mathbb{R}^{C,D} \tag{10}$$

$$\tilde{\text{k}}_h = k_h \cdot w_{1,h} + b_{1,h} \in \mathbb{R}^{C,D} \tag{11}$$

The attention score is calculated by transpose and matrix product.

$$\mathrm{s}_h = \tilde{q}_h \cdot \tilde{k}_h^T \in \mathbb{R}^{C,C} \tag{12}$$

Finally, it is straightforward to calculate the final attention tensor A_h for the h head.

$$\mathrm{A}_h = s_h \cdot v_h \in \mathbb{R}^{C,(HxW)/n} \tag{13}$$

The final output o is synthesized by combining n heads using simple transpose, reshape, and connect operations.

$$\mathrm{o} = \mathrm{SplitHeads}^{-1}\left([\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n]\right) \in \mathbb{R}^{H,W,C} \tag{14}$$

3.3 Global Average Pooling and Dropout

In the facial expression picture, the expression information is mainly concentrated in the central area of the picture, such as the corners of the mouth, eyebrows and other regional features. Therefore, the global average pooling layer is used instead of the traditional full connection layer to directly sum the channel information of facial expressions to reduce the dimension and reduce the network parameters of the model. Finally, by using the Dropout function, some neurons in the neural network are discarded randomly, and the image feature information recorded by CNN is reduced so that the facial expression recognition network will not rely too much on some local features and enhance the robustness and generalization ability of the model. Therefore, the output layer of the model adopts GAP and Dropout design to further simplify the parameters and complexity of the network, improve the training speed of the network model, avoid the over-fitting phenomenon, and then improve the generalization of the network.

4 Experimental Results

In this section, we will verify and evaluate the model of three facial expression datasets. First of all, the data set used is briefly described, and the corresponding data preprocessing is carried out, including data enhancement and so on. Then compare it with other models, and finally visualize the model results, including drawing confusion matrix and ROC curve and so on.

4.1 Database

The facial expression datasets used were analyzed, including the Facial Expression Recognition 2013 (Fer2013), the extended Cohn-kanade (CK+) and Japanese female Facial Expression (Jaffe).

Fer2013: the expression data set consists of 35,668 facial expression images, including 28,709 in the test set, 3589 in the verification set and 3589 in the test set. The size of each picture is 48 × 48 gray scale image, the structure of this model is based on the ResNet pre-training model, so the original picture needs to be adjusted to 224 × 224. There are seven expressions in the data set, corresponding to the number label 0–6, which are angry, disgusted, frightened, happy, sad, surprised and neutral in turn, and the distribution of the number of each type is shown in Fig. 5 below.

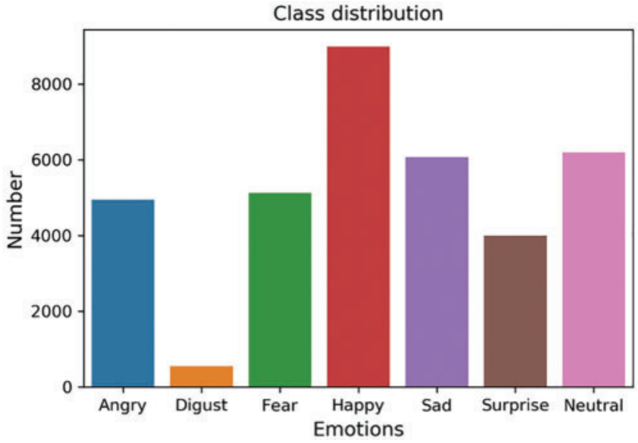

Figure 5: Distribution of Fer2013 dataset

The example of each expression in the Fer2013 dataset is shown in Fig. 6 below.

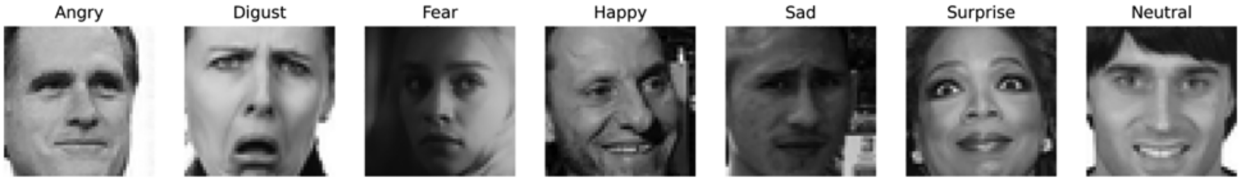

Figure 6: Seven samples of Fer2013 dataset

CK+: The dataset is extended on the basis of Cohn-Kanade dataset. There are seven kinds of emotions in the data set, corresponding to the number label 0–6, which are angry, disgusted, scared, happy, sad, surprised and neutral in turn, and the distribution of the number of each type is shown in Fig. 7 below.

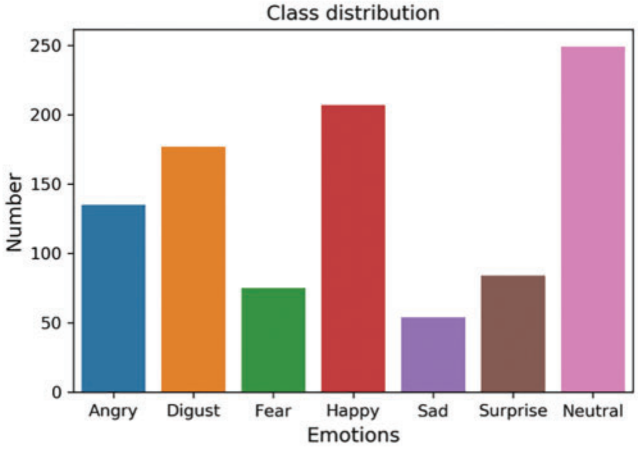

Figure 7: Distribution of CK+ dataset

The example of each expression in the CK+ dataset is shown in Fig. 8 below.

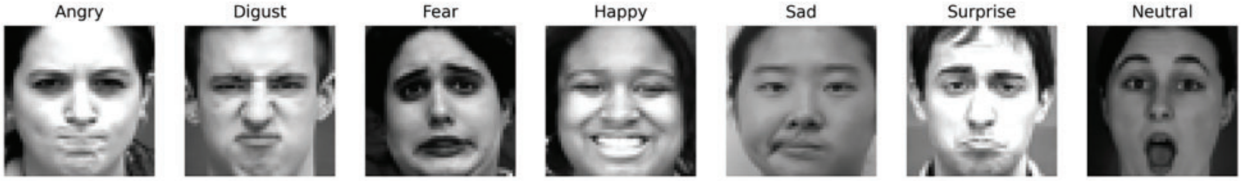

Figure 8: Seven samples of CK+ dataset

Jaffe: the database contains 213 facial expressions of 10 Japanese women. Each person makes 7 expressions, corresponding to the number label 0–6, followed by anger, disgust, fear, happiness, sadness, surprise and neutrality. The number distribution of each type is shown in Fig. 9 below.

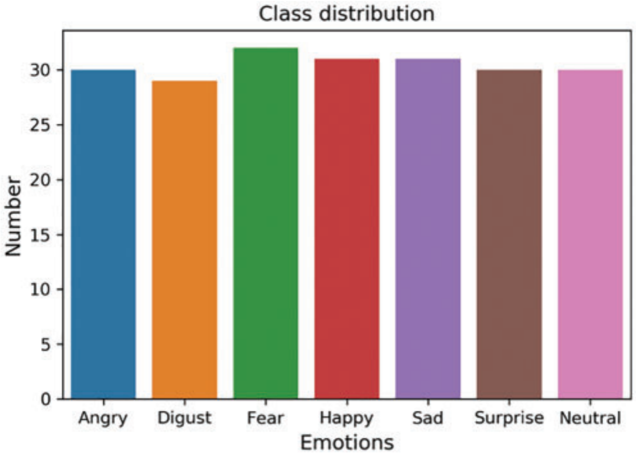

Figure 9: Distribution of jaffe dataset

The example of each expression in the CK+ dataset is shown in Fig. 10 below.

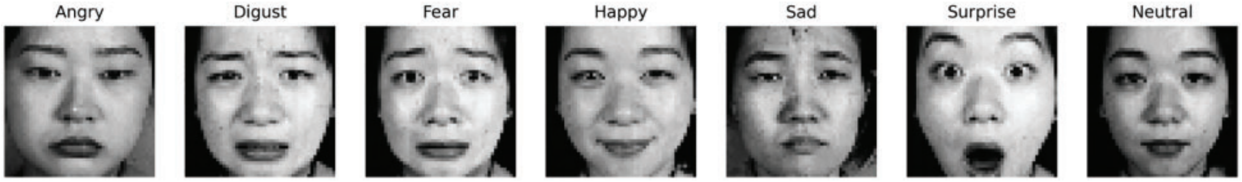

Figure 10: Seven samples of the jaffe dataset

4.2 Data Enhancement

Due to the small number of samples in CK+ and Jaffe data sets and the imbalance of expression categories in Fer2013 data sets, neural network training can easily lead to problems such as weak generalization ability and over-fitting. Therefore, it is necessary to carry out data enhancement operations on the three data sets, including random rotation, random scaling, horizontal, vertical translation and random flipping. Fig. 11 shows the effect of Jaffe data enhancement.

In addition, all experiments use 7-fold cross-validation (i.e., the images are randomly divided into 7 equal-sized subsets, 6 subsets are used for training and the remaining 1 subset is used for testing), and the final results are derived by averaging the recognition accuracy.

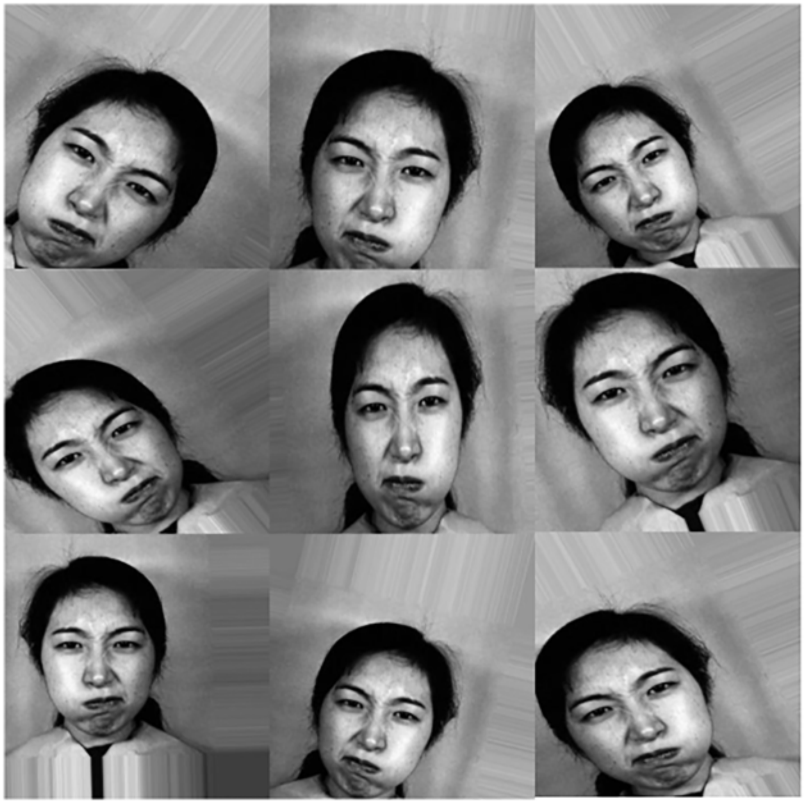

Figure 11: Example of jaffe data enhancement

4.3 Experimental Environment and Parameter Metrics

4.3.1 Implementation Details

All models in this paper were run on the open source TensorFlow platform, used the Nvida Getforce Gtx1080 for experiments. ResNet18 was used as the backbone network of MCA-Net, and the image dataset on ImageNet was used to initialize the network parameters for training. The model learning rate is set to 0.005, the dropout ratio is 0.5, the optimization algorithm is Adam, and the batch size is 64.

This model uses ResNet pre-training model, so it is necessary to rescale the images of the three expression data sets to 224 × 224. Fig. 1 describes in detail the implementation details of ResNet18 based on the MCA module, and the model parameters for each MCA model are shown in Table 1.

Table 1: MCA module parameter

Block	Heads	Dim	Pool	Scale	Ker
MCA1	8	196	3	1	3
MCA2	8	196	3	1	3
MCA3	7	56	3	1	3

(Continued)

Table 1 (continued)

Block	Heads	Dim	Pool	Scale	Ker
MCA4	7	14	3	1	3
MCA5	1	25	3	1	3

4.3.2 Metrics

We evaluate the proposed MCA-Net with accuracy metric to compare with other the performance of expression recognition models quantitatively. Accuracy is the proportion of correctly classified samples to the total number of samples.

$$\text{Accuracy} = \text{correct}/\text{total} \tag{15}$$

where correct is the number of correctly classified samples and the total is the number of total samples.

4.4 Analysis of Experimental Results

The improved model proposed in this paper is used to experiment on three facial expression data sets, and the experimental results are analyzed by drawing confusion matrix, ROC curve and model comparison experiment.

4.4.1 Confusion Matrix

The confusion matrix of the Fer2013 is shown in Fig. 12.

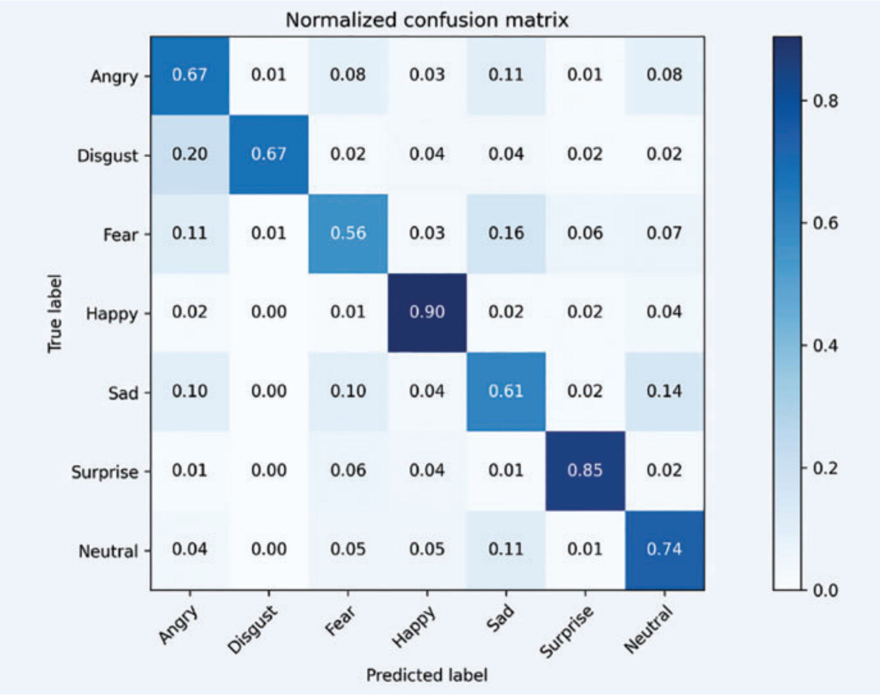

Figure 12: The confusion matrix of the Fer2013 dataset

The confusion matrix of the CK+ dataset is shown in Fig. 13.

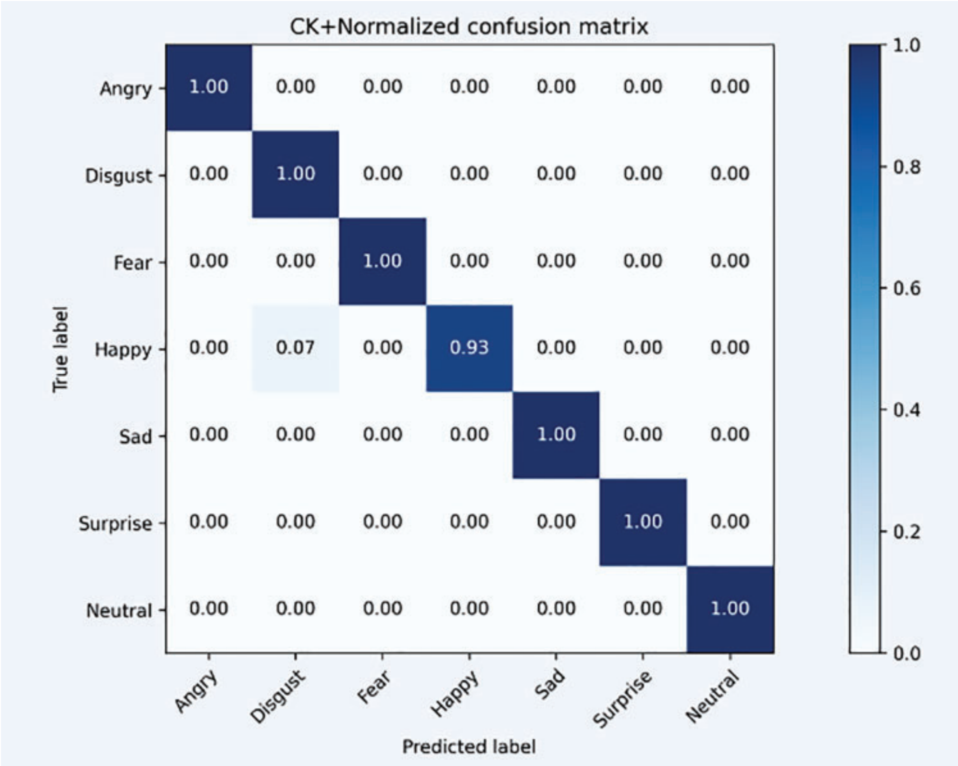

Figure 13: The confusion matrix of the CK+ dataset

The confusion matrix of the Jaffe dataset is shown in Fig. 14.

Fig. 12 shows the experimental results of the model on the Fer2013 dataset, and we found that the recognition accuracy of the model for all seven expressions is not very satisfactory; only happy expression has a recognition rate of 90%; surprise expression has a recognition rate of 85%, and the lowest fear only has a recognition rate of 56%. The main reason is that most of the expressions in the Fer2013 dataset are collected from the web with image occlusion, light blurring and label mislabeling. Therefore, the Fer2013 dataset needs to be pre-processed before model training, including filtering the emotion images and correcting emotion labeling. On the other hand, the two categories such as fear and sad are often easily confused. This is because human expressions are rich and present concomitant features, such as sadness caused by fear.

Figs. 13 and 14 show the experimental results of the model on CK+ dataset and Jaffe dataset, respectively. We found that the recognition rate of the method in this paper hits 100% on all six expressions, mainly because the expressions in the photos of CK+ dataset and Jaffe dataset originate from the laboratory environment, the image data annotation is standardized, and the expression images have high quality and easy to distinguish. The main reason for the relatively poor recognition results of the models in this paper for happy and disgust is probably because these two expressions have similar muscle deformation degrees, which leads to the situation that the models have misjudgment.

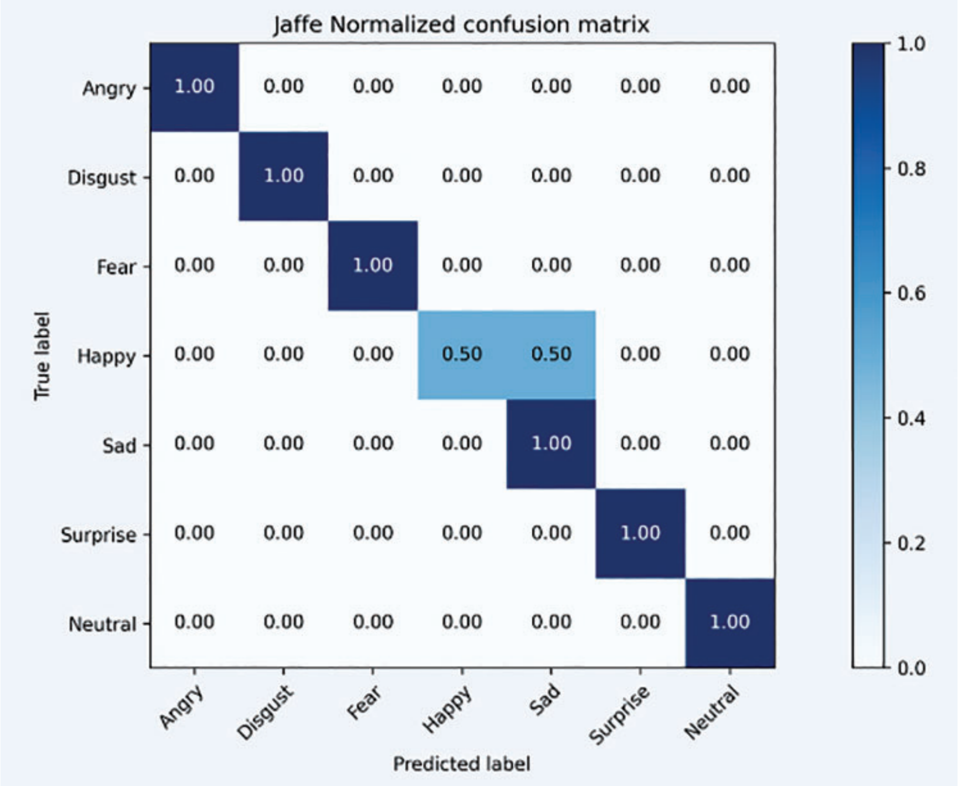

Figure 14: The confusion matrix of the jaffe dataset

Through the above analysis, we should study the facial expression recognition in real environment and consider the influence of factors such as lighting, occlusion, and expression concomitance in the study of facial expression recognition.

4.4.2 ROC Curve

The ROC curve is a composite indicator of the continuous variables of sensitivity and specificity, and is a graphical representation of the interrelationship between sensitivity and specificity.

Figs. 15 to 17 show that the ROC curves of the model on the Fer2013 dataset, CK+ dataset and Jaffe dataset, respectively. By analyzing the ROC curves, we found that the ROC curves of each dataset correspond to the results of the confusion matrix, and the higher the accuracy of expression recognition, the larger the area of the curve under the corresponding category.

4.4.3 Comparative Analysis of Model Results

To further validate the model in this paper, we compare the result of our model with previous research work, as shown in Tables 2 to 4.

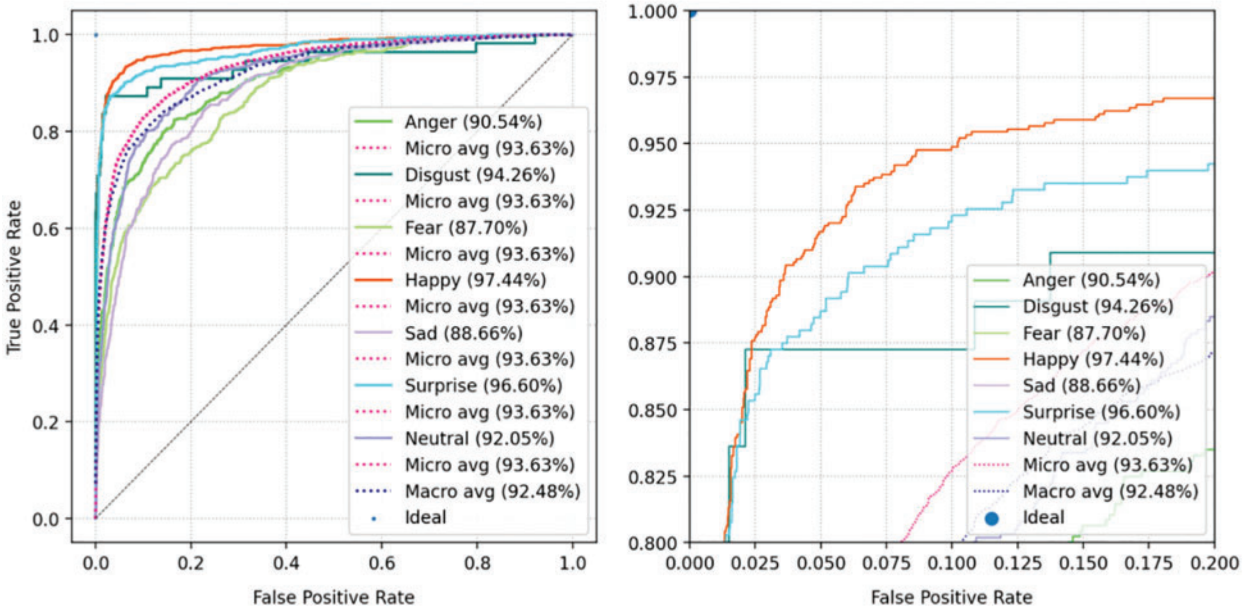

Figure 15: The ROC curve of the Fer2013 dataset

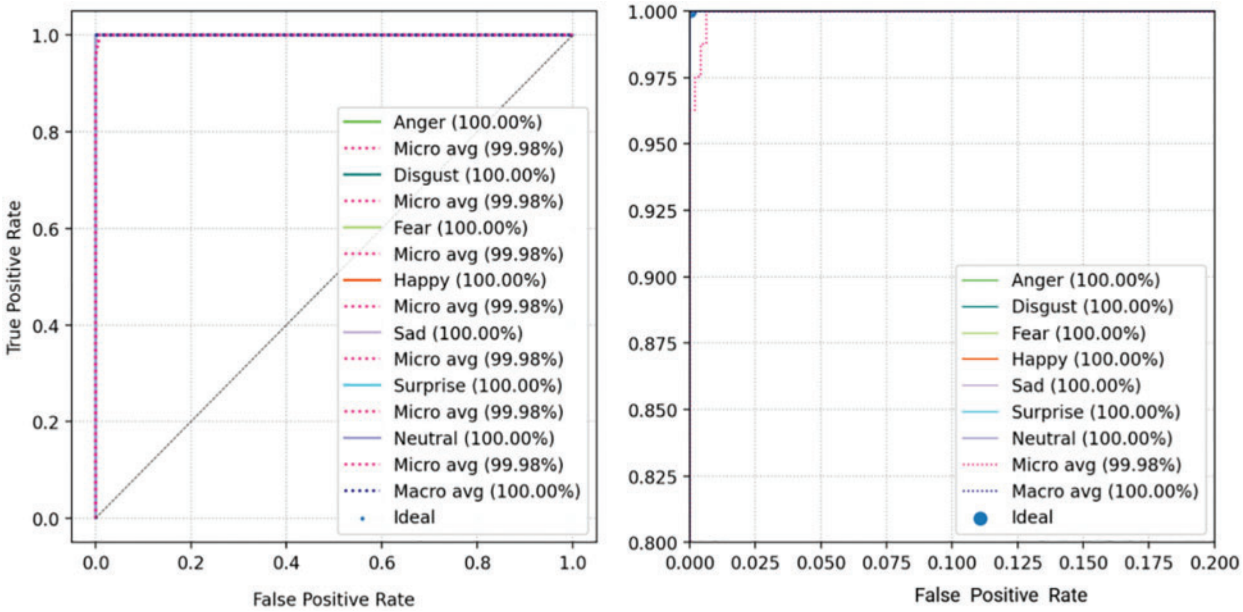

Figure 16: The ROC curve of the CK+ dataset

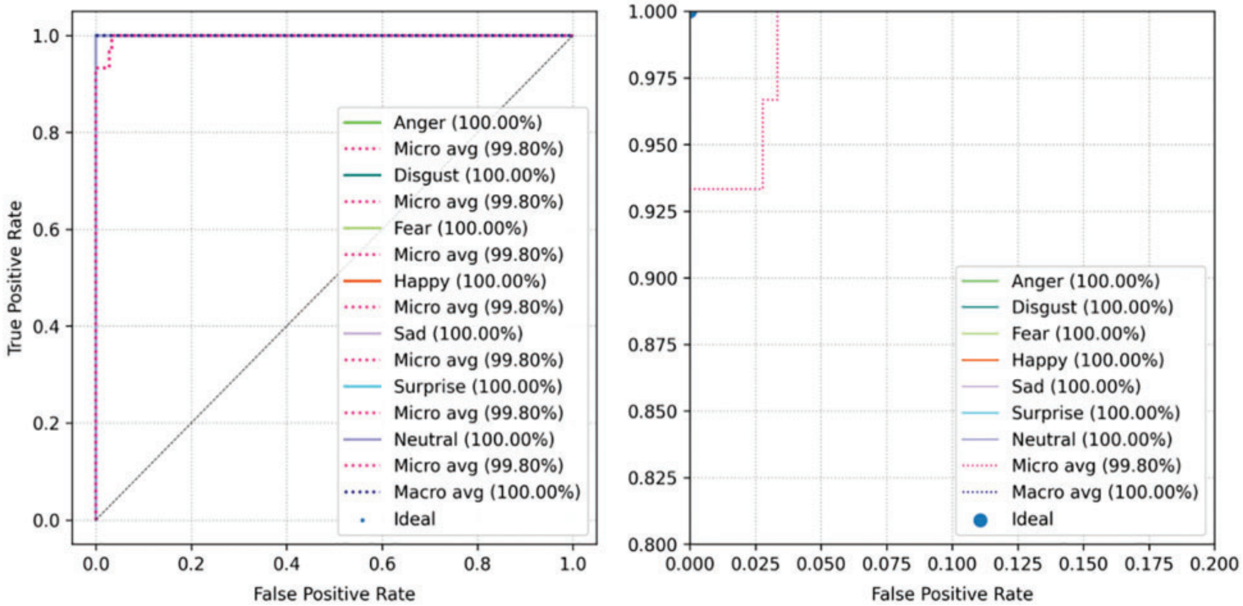

Figure 17: The ROC curve of the Jaffe dataset

Table 2 shows that the recognition results of the model in dataset Fer2013, and we find that the recognition accuracy of the model reaches 72.7%. The accuracy of LHC-Net and VGGNet models are 1.72% and 0.58% higher than the model in this paper, respectively. The LHC-Net model adopts the ResNet34 network structure, which is 16 layers more than the network structure in this paper, and the network depth and model parameters are increased accordingly. The article adopts the VGGNet architecture, rigorously fine-tune its hyper parameters, and experiment with various optimization methods.

Table 2: Classification accuracy on the Fer2013 dataset

Model	Accuracy rate
CNN [38]	72.16
ResNet [39]	72.4
VGGNet [40]	73.28
DeepEmotion [41]	70.02
SVM [42]	71.16
SE-Net50 [43]	72.7
LHC-Net [44]	74.42
ResNet18	64.8
The proposed model	72.7

Table 3 shows that the recognition results of the model in dataset CK+, and we find that the recognition accuracy of the model reaches 98.8%, which indicates the effectiveness of the model in this paper. Compared with 86.3% of the ResNet18 model, our model improves 12.5%. The result indicates that the MCA module can effectively extract the channel information of the image, and assigning different weights to different feature information to improve the model performance. The model performance is improved by giving different weights to different feature information.

Table 3: Classification accuracy on the CK+ dataset

Model	Accuracy rate
NSVT [45]	96.5
DRL [46]	89.8
CUDL [47]	96.6
CNN [48]	92.81
VGG16 [49]	94.8
HCIA [50]	96
DTAGN [51]	97.2
ST-RNN [52]	97.2
ResNet18	86.3
The proposed model	98.8

Table 4 shows that the recognition results of the model in the dataset Jaffe, and we find that the recognition accuracy of the model reaches 93.3%. The accuracy of VGG16 and ERCEC models are 0.37% and 0.17% higher than the model in this paper, respectively. The expression images in the Jaffe dataset are collected in a laboratory situation, and no specified training and test sets are provided and therefore, different segmentation results will lead to some differences in the experimental results.

Table 4: Classification accuracy on the Jaffe dataset

Model	Accuracy rate
ERCEC [47]	93.5
TIFE [53]	91.97
PCA [54]	91.3
VGG [49]	93.7
MLT [55]	89.18
LBP-ORB [56]	88.5
Deep Features-HOG [57]	90.58
ResNet18	85.7
The proposed model	93.33

We also compare the accuracy of the different models on the three datasets, as shown in Figs. 18–20. The comparative analysis of the three figures shows that the model proposed in our paper converges fast on the three datasets, and finally the model is almost converged and achieves a high accuracy rate.

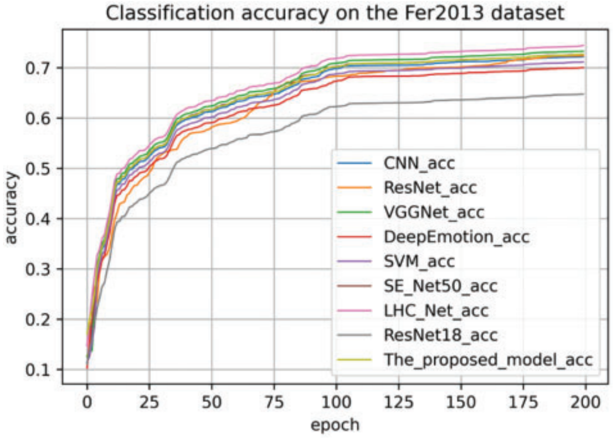

Figure 18: The classification accuracy on the Fer2013 dataset

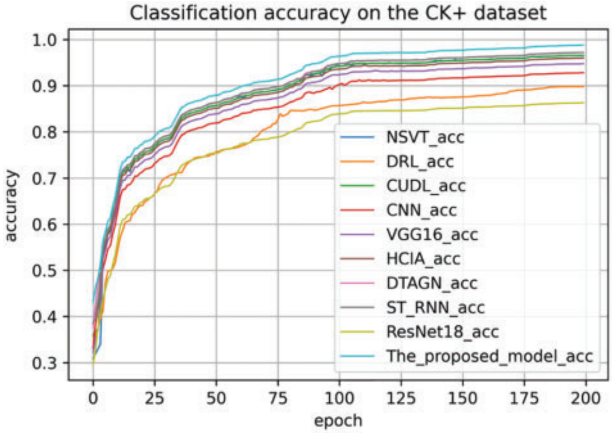

Figure 19: The classification accuracy on the CK+ dataset

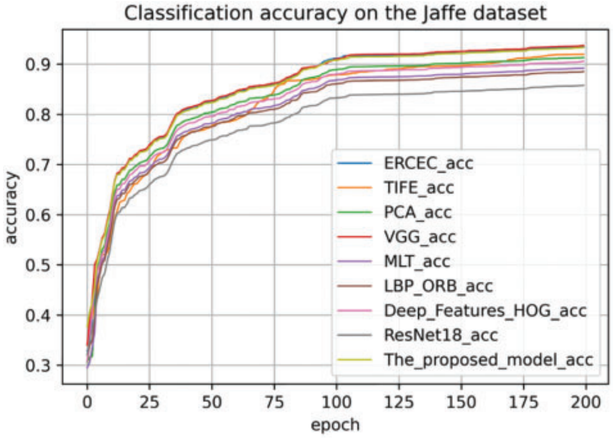

Figure 20: The classification accuracy on the jaffe dataset

From the above analysis results, we find that the recognition accuracy of our proposed method achieves good results on all three expression datasets, which verify the effectiveness of our proposed method.

5 Conclusion

The research of the facial expression recognition algorithm has important theoretical significance and practical value. Based on the analysis of the current mainstream facial expression recognition algorithms, we proposed a facial expression recognition algorithm based on the combination of ResNet18 network structure and multi-channel attention mechanism. The main purpose of this network structure is to add MCA to the ResNet18 network structure to coordinate the self-attention mechanism and the channel attention mechanism. Through the residual module output fusion to extract richer facial expression features, it can improve the network of local key parts feature extraction, and join the mainstream deep learning network structure, such as VGG, ResNet and so on. The experimental results show that the model proposed in this paper achieves excellent recognition results in Fer2013, CK+ and Jaffe datasets. Compared with the mainstream expression recognition models, the total parameters of this model are only 20 M. The complexity of the model is further reduced, the training speed of the model is further improved, and the recognition speed of facial expression recognition algorithm is further improved. However, the facial images studied in this paper are static images, which do not take into account the facial expression images under complex environments, such as the missing facial expression information and non-frontal face images, etc. In order to verify the robustness and generalization ability of the facial expression algorithm proposed in this paper, the model will be tested and evaluated using other types of facial expression datasets in the next step.

Funding Statement: The research was funded by Anhui Province Quality Engineering Project No. 2021jyxm0801; Natural Science Foundation of Anhui University of Chinese Medicine under Grant Nos. 2020zrzd18, 2019zrzd11; Humanity Social Science foundation Grants 2021rwzd20, 2020rwzd07; Anhui University of Chinese Medicine Quality Engineering Projects No. 2021zlgc046.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Mehrabian, A., Russell, J. A. (1974). *An approach to environmental psychology*. USA: The MIT Press.
2. Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48(4),* 384–392. DOI 10.1037/0003-066X.48.4.384.
3. Choi, H. J., Lee, Y. J. (2020). Deep learning based response generation using emotion feature extraction. *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 255–262. Busan, Korea (South).
4. Wu, M., Su, W., Chen, L., Liu, Z., Cao, W. et al. (2019). Weight-adapted convolution neural network for facial expression recognition in human-robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(3),* 1473–1484. DOI 10.1109/TSMC.6221021.
5. Saste, T. S., Jagdale, S. M. (2017). Emotion recognition from speech using MFCC and DWT for security system. *Proceedings of the IEEE 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA)*, pp. 701–704. Coimbatore, India.

6. Sajjad, M., Nasir, M., Ullah, F. U. M., Muhammad, K., Sangaiah, A. K. et al. (2019). Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services. *Information Sciences, 47(9),* 416–431. DOI 10.1016/j.ins.2018.07.027.
7. Ma, L., Chen, W., Fu, X., Wang, T. (2018). Emotional expression and micro expression recognition in depressive patients. *Chinese Science Bulletin, 63(20),* 2048–2056. DOI 10.1360/N972017-01272.
8. Zhang, B., Liu, G., Xie, G. (2017). Facial expression recognition using LBP and LPQ based on gabor wavelet transform. *Proceedings of the 2017 IEEE International Conference on Computer and Communications*, pp. 365–369. New York, USA.
9. Xu, F., Wang, Z. (2018). A facial expression recognition method based on cubic spline interpolation and HOG features. *Proceedings of the 2018 IEEE International Conference on Robotics & Biomimetics*, pp. 2163–2168. New York, USA.
10. Shin, M., Kim, M., Kwon, D. S. (2016). Baseline CNN structure analysis for facial expression recognition. *Proceedings of the 2016 International Symposium on Robot and Human Interactive Communication (ROMAN)*, pp. 724–729. New York, USA.
11. Jacob, W., Omlin, C. W. (2006). Haar features for FACS AU recognition. *Proceedings of the IEEE FGR 2006 7th International Conference on Automatic Face and Gesture Recognition*, pp. 5–10. Southampton, UK.
12. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86(11),* 2278–2324.
13. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097–1105. Nevada, USA.
14. Simonyan, K., Zisserman, A. (2014). A very deep convolutional networks for large-scale image recognition. *Proceedings of the 2014 Computer Vision and Pattern Recognition*, pp. 641–660. New York, USA.
15. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. Nevada, USA.
16. Hu, J., Li, S., Gang, S. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141. Utah, USA.
17. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018). Cbam: Convolutional block attention module. *Proceedings of the European Computer Vision (ECCV)*, pp. 3–19. Munich, Germany.
18. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y. et al. (2019). Dual attention network for scene segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154. California, USA.
19. Liu, Y., Zhang, X., Zhou, J., Fu, L. (2021). SG-DSN: A semantic graph-based dual-stream network for facial expression recognition. *Neurocomputing, 462,* 320–330. DOI 10.1016/j.neucom.2021.07.017.
20. Li, S., Li, W., Wen, S., Shi, K., Yang, Y. et al. (2021). Auto-FERNet: A facial expression recognition network with architecture search. *IEEE Transactions on Network Science and Engineering, 8(3),* 2213–2222. DOI 10.1109/TNSE.2021.3083739.
21. Li, B., Lima, D. (2021). Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering, 2,* 57–64. DOI 10.1016/j.ijcce.2021.02.002.
22. Pham, L., Vu, T. H., Tran, T. A. (2021). Facial expression recognition using residual masking network. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4513–4519. Milan, Italy.
23. Lai, Z., Chen, R., Jia, J., Qian, Y. (2020). Real-time micro-expression recognition based on ResNet and atrous convolutions. *Journal of Ambient Intelligence and Humanized Computing, 11(11),* 1–12. DOI 10.1007/s12652-020-01779-5.
24. Niu, R. H., Yang, J., Xing, L. X., Wu, R. B. (2021). Micro-expression recognition method based on dual-channel attention mechanism. *Computer Applications, 41(9),* 2552–2559.

25. Zhao, G., Pietikainen, M. (2007). Dynamic texture recognition u-sing local binary patterns with an application to facial expres-sions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6),* 915–928. DOI 10.1109/TPAMI.2007.1110.
26. Shan, C., Gong, S., Owan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing, 27(6),* 803–816. DOI 10.1016/j.imavis.2008.08.005.
27. Luo, Y., Zhang, T., Zhang, Y. (2015). A novel fusion method of PCA and LBP for facial expression feature extraction. *Optik Internation Journal for Light and Electron Optics, 127(2),* 718–721 DOI 10.1016/j.ijleo.2015.10.147.
28. Kumar, S., Bhuyan, M. K., Chak, B. K. (2016). Extraction of informative regions of a face for facial expression recognition. *Let Computer Vision, 10(6),* 567–576. DOI 10.1049/iet-cvi.2015.0273.
29. Saha, A., Pradhan, S. N. (2018). Facial expression recognition based on eigenspaces and principle component analysis. *International Journal of Computational Vision and Robotics, 8(2),* 190–200. DOI 10.1504/IJCVR.2018.091980.
30. Bougourzi, F., Dornaika, F., Mokrani, K., Taleb-Ahmed, A., Ruichek, Y. (2020). Fusing transformed deep and shallow features (FTDS) for image based facial expression recognition. *Expert Systems with Applications, 156,* 113459. DOI 10.1016/j.eswa.2020.113459.
31. Qian, Y. S., Shao, J., Ji, X. X. (2019). Face expression recognition based on LGRP and multi-feature fusion. *Journal of Shanghai University of Electric Power, 35(3),* 253–260.
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30,* 5998–6008.
33. Xu, L. C., Huang, J., Atsushi, N., Asaoka, R. (2020). A novel global spatial attention mechanism in convolutional neural network for medical image classification. arXiv preprint arXiv:2007.15897.
34. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J. et al. (2017). SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pp. 5659–5667. Hawaii, USA.
35. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
36. Dai, Z., Liu, H., Le, Q. V., Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. arXiv preprint arXiv:2106.04803.
37. Borji, A., Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1),* 185–207. DOI 10.1109/TPAMI.2012.89.
38. Vulpe, A., Grigore, O. (2021). Convolutional neural network hyperparameters optimization for facial emotion recognition. *12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pp. 1–5. Bucharest, Romania.
39. Ramerdorfer, C., Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art. arXiv preprint arXiv:1612.02903.
40. Khaireddin, Y., Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2105.03588.
41. Minaee, S., Minaei, M., Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors, 21(9),* 3046. DOI 10.3390/s21093046.
42. Ian, J. G., Dumitru, E., Pierre, L. C., Aaron, C., Mehdi, M. et al. (2013). Challenges in representation learning: A report on three machine learning contests. *International Conference on Neural Information Processing*, pp. 117–124. Berlin, Germany.
43. Amil, K., Bai, C., Ferhat, T. C. (2020). Facial expression recognition with deep learning. arXiv preprint arXiv:2004.11823.
44. Pecoraro, R., Basile, V., Bono, V., Gallo, S. (2021). Local multi-head channel self-attention for facial expression recognition. arXiv preprint arXiv:2111.07224.

45. She, H., Harisu, A., Will, B., Hedwig, E. (2020). Emotion categorization from video-frame images using a novel sequential voting technique. *International Symposium on Visual Computing*, pp. 618–632. Cham, Switzerland.
46. Mishra, S., Joshi, B., Paudyal, R., Chaulagain, D., Shakya, S. (2022). Deep residual learning for facial emotion recognition. In: *Mobile computing and sustainable informatics*, pp. 301–313. Singapore: Springer.
47. Muhammad, G., Hossain, M. S. (2021). Emotion recognition for cognitive edge computing using deep learning. *IEEE Internet of Things Journal, 8(23),* 16894–16901. DOI 10.1109/JIOT.2021.3058587.
48. Liliana, D. Y. (2019). Emotion recognition from facial expression using deep convolutional neural network. *Journal of Physics: Conference Series, 1193(10),* 012004. DOI 10.1088/1742-6596/1193/1/012004.
49. Dubey, A. K., Jain, V. (2020). Automatic facial recognition using VGG16 based transfer learning model. *Journal of Information and Optimization Sciences, 41(7),* 1589–1596. DOI 10.1080/02522667.2020.1809126.
50. Chowdary, M. K., Nguyen, T. N., Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications, 33(20),* 1–18. DOI 10.1007/s00521-021-06012-8.
51. Heechul, J., Lee, S., Yim, J., Park, S., Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2983–2991. Santiago, Chile.
52. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, Y. (2018). Spatial-temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics, 49(3),* 839–847. DOI 10.1109/TCYB.6221036.
53. Malik, S., Kumar, P., Raman, B. (2021). Towards interpretable facial emotion recognition. *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–9. Jodhpur, India.
54. Arora, M., Kumar, M., Garg, N. K. (2018). Facial emotion recognition system based on PCA and gradient features. *National Academy Science Letters, 41(6),* 365–368. DOI 10.1007/s40009-018-0694-2.
55. Ullah, S., Jan, A., Khan, G. M. (2021). Facial expression recognition using machine learning techniques. *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pp. 1–6. Istanbul, Turkey.
56. Ben, N., Gao, Z., Guo, B. (2021). Facial expression recognition with LBP and ORB features. *Computational Intelligence and Neuroscience, 2021,* 8828245.
57. Hao, W., Wei, S., Fang, B. (2020). Facial expression recognition using iterative fusion of MO-HOG and deep features. *The Journal of Supercomputing, 76(5),* 3211–3221. DOI 10.1007/s11227-018-2554-8.