IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 17, 2021, accepted March 1, 2021, date of publication March 2, 2021, date of current version March 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063493

A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network

CUIPING SHI[1], (Member, IEEE), CONG TAN[1], AND LIGUO WANG[2], (Member, IEEE)

[1]College of Communication and Electronic Engineering, Qiqihar University, Qiqihar 161000, China
[2]College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116000, China

Corresponding author: Cuiping Shi (scp1980@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 41701479 and Grant 62071084, in part by the Fundamental Research Funds in Heilongjiang Provincial Universities of China under Grant 135509136, and in part by the Scientific Research Foundation for the Heilongjiang Postdoctor under Grant SBSHJJ-001.

ABSTRACT This paper proposes a new method based on a multiple branch cross-connected convolutional neural network (MBCC-CNN) for facial expression recognition. Compared with traditional machine learning methods, the proposed method can extract image features more effectively. In addition, in contrast to single-structure convolutional neural networks, the MBCC-CNN model is constructed based on the residual connection, Network in Network, and tree structure approaches together. It also adds a shortcut cross connection for the summation of the convolution output layer, which makes the data flow between networks more smooth, improves the feature extraction ability of each receptive field. The proposed method can fuse the features of each branches more effectively, which solves the problem of insufficient feature extraction of each branches and increases the recognition performance. The experimental results based on the Fer2013, CK+, FER+ and RAF data sets show that the recognition rates of the proposed MBCC-CNN method are 71.52%, 98.48%, 88.10% and 87.34%, respectively. Compared with some most recently work, the proposed method can provide better facial expression recognition performance and has good robustness. The python code can be download from https://github.com/scp19801980/Facial-expression-recognition.

INDEX TERMS Facial expression recognition, convolutional neural network, residual connection, network in network, robustness.

I. INTRODUCTION

Facial expression recognition (FER) mainly predicts the facial expressions by facial appearance changes. Facial expression is the most direct and effective emotion recognition mode [1], [2]. FER is a task of face analysis [3]–[6]. FER has many applications in human-computer interaction, such as fatigue driving detection, psychological changes of criminals, and real-time expression recognition on mobile phones. Meanwhile, there have also been important developments in various fields, such as education monitoring and medical testing [7]–[9]. In recent years, because of its practical application value and prospects, facial expression recognition has become a research hotspot and great progress has been made.

The process of expression recognition can be roughly divided into the following steps: image selection, image preprocessing, image feature extraction, and image recognition [10], [11]. Usually, the collected face images are preprocessed, such as face detection and rotation correction. Facial detection is implemented using cascaded classifiers, for example, the Adaboost [12] and Viola-Jones frameworks [13]. Face correction can be implemented with features such as eyes and mouth. The key to facial expression recognition is the extraction of facial image features. At present, there are two types of methods for describing face images, i.e., geometric feature-based methods and texture feature-based methods. The method based on geometric features involves encoding the region of interest, that is, locating and measuring the shape and position changes of the prominent features of the face, such as the mouth, eyebrows, nose, eyes, etc. However, only use a small number of features in the region of interest can be used to describe the facial image. The texture feature method has rotation invariance and good anti-noise performance, but it can only reflect the features of the object's surface, cannot fully reflect the essential properties

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

of the object, and cannot obtain the high-level content of the image.

In recent years, some studies have been performed on facial expression recognition. However, facial expression recognition still faces great challenges. For FER, some variations unrelated to expression, such as illumination variation, occlusions, non-frontal head poses, identity bias, and the recognition of low-intensity expressions, will lead to the degradation of facial expression recognition performance. Considering that FER is a data-driven task and that training an effective deep network to capture subtle expression-related deformations requires a large amount of training data, the major challenge that deep FER systems face is the lack of training data in terms of both quantity and quality [7]. Therefore, the traditional manual feature extraction method is no longer suitable for the interference prone FER recognition. Meanwhile, the traditional expression feature extraction methods are inefficient, and the feature extraction is incomplete. With the rapid development of deep learning, great progress has been made in pattern recognition. Much work has been done on facial expression recognition based on deep neural networks, and good results have been obtained [14]–[19]. At present, developing effective expression recognition based on a convolutional neural network is still a problem worthy of study. In this paper, a new multibranch cross-connection convolutional neural network (MBCC-CNN) method is proposed, which can avoid the problem of insufficient feature extraction and improve the facial expression recognition performance. First, the expression data set is preprocessed to make it more convenient for the network to learn image features. Then, the MBCC-CNN is constructedto effectively extract the image features. Based on the residual connection, Network in Network, and tree structure approaches, the MBCC-CNN adds a shortcut cross connection for the summation of the convolution output layer, which makes the data flow between networks smoother, increases the feature extraction ability of each reception field, and avoids the problem of inadequate feature extraction. At the same time, the global mean pooling greatly reduces the network parameters and avoids over fitting. Finally, based on the features extracted by the MBCC-CNN, the SoftMax classifier is utilized for facial expression recognition.

II. RELATED WORKS AND CONTRIBUTIONS

Facial expression recognition is a popular topic in the field of computer vision. At present, the existing expression recognition methods can be roughly divided into three categories, as follows: based on traditional learning methods, based on convolutional neural network, and based on the fusion of traditional methods and convolutional neural networks.

A. TRADITIONAL MACHINE LEARNING METHODS

For expression recognition based on traditional learning methods, hand-made features, such as the Gabor wavelet coefficient [20], the local binary pattern (LBP) [21] and the histogram of oriented gradient(HOG) [22], are often used to represent a specific expression. In [23], Goyani M *et al.* proposed a facial feature extraction method based on the multistage Haar wavelet. First, the AdaBoost [12] cascade target detector was used to segment the geometric components with the most information, such as the eyes, mouth, eyebrows, etc. Then, the Haar feature of the segmented components was extracted. Finally, the OneVsAll logistic regression model was used for classification. Palermo *et al.* [24] studied a facial perception mechanism, which can realize the recalibration of the facial expression system and increase the sensitivity to facial expression changes. Meanwhile, this paper proves that the perceptual factors of adaptive coding are independent of the emotional factors, which makes an important contribution to the work on the recognition of facial expression changes. In [25], Kim S *et al.* proposed a simple multilayer perceptron (MLP) classifier to judge whether the current facial expression recognition results are reliable. In [26], Shi S *et al.* used the Gabor wavelet to extract the facial features and adopted multiscale histogram methods to extract the features in the around center instable local binary pattern(ACI-LBP). Finally, the Gabor and ACI-LBP results are fused to form a complete feature vector for facial recognition. In [27], Yan H proposed a cooperative discriminant multiscale learning (CDMML) technique for video facial expression recognition. First, multiple feature descriptors are calculated for each face video to describe the appearance and motion information of the face from different perspectives. Then, the extracted features are used to learn multiple distance measures in a cooperative manner to make use of complementary and discriminant information for recognition. The experimental results on Wild (AFEW) 4.0 and the extended Cohn- Kanada (CK+) data sets prove the effectiveness of this method.

B. CONVOLUTIONAL NEURAL NETWORK METHODS

However, the traditional expression feature extraction methods are inefficient, and the feature extraction process is incomplete. Therefore, the deep learning for facial expression recognition technique is becoming more popular. For the facial expression recognition, some works construct a convolutional neural network, which first preprocess the facial expression image data set and then uses the convolutional neural network to train and test on the expression data set to realize expression recognition [28]–[31]. In [32], Shan S *et al.* proposed a convolutional neural network with an attention mechanism (ACNN), which can sense the occlusion area of the face and focus on the most discriminating non-occlusion area. The proposed ACNN method was evaluated under both real and synthetic occlusion, including a self-collected facial expression data set with real occlusion, two large wild facial expression data sets (RAF-DB and AffectNet), and modification of synthetic occlusion. The experimental results show that the ACNN method can improve the recognition accuracy in both the non-occlusion and occlusion situations. Some works used the attention mechanism to construct the convolutional neural network [33]–[35], [61]. The attention mechanism is very interested in the important parts of the human face, such as the eyes, nose and mouth. The region of interest (ROI) in the face image is marked before the image is input into the convolutional neural network. In [35], the concept of attention is introduced to the first layer of the convolutional neural network to perform convolution calculations of ROIs. In [36], [37], a multichannel convolutional neural network was used for feature fusion. In [36], a dual-channel weighted hybrid deep convolutional neural network (WMD-CNN) based on static images and a dual-channel weighted hybrid deep long short-term memory network based on image sequences (WMCNN-LSTM) were proposed. The WMDCNN network can recognize facial expressions quickly and then provide static image features for the WMCNN-LSTM network. The LSTM network uses the static image features to further acquire the temporal features of the image sequence to realize the accurate recognition of facial expressions. In [37], a multichannel deep neural network was proposed, which can learn and fuse the temporal and spatial features of facial expressions. The basic idea of this method is to extract the optical flow from the changes between the emotional face image and the normal face image, take this change as the time information of a certain facial expression, and use the emotional face image as the space information. At the same time, a multichannel deep space-time feature fusion neural network (MDSTFN) is proposed for space-time feature extraction and fusion of still images. In [38], Li S *et al.* proposed a new facial expression database, which contains approximately 30,000 facial images with various postures and lighting from thousands of individuals of different ages and races. Meanwhile, a deep reserved convolutional neural network (DLP-CNN) method was proposed, which aims to enhance the discrimination ability of the network by maximizing the class scatter. In [39], Zhang Z *et al.* studied a deep network architecture for robust facial expression recognition, which can learn from rich auxiliary aspects and not only facial expression data. This model can mine the interaction context of the face and achieve accurate fine-grained face prediction. In [40], Alam M *et al.* proposed a biologically related sparse deep synchronous recursive network (S-DSRN) for the robust recognition of facial expressions. This method uses dropout learning to obtain feature sparsity, which can provide good classification performance and has a low computational complexity. In facial expression recognition, external illumination, occlusion and other factors greatly influence the facial expression recognition research. In [41], Liu Y *et al.* proposed a conditional convolutional neural network to enhance the random forest for expression recognition in an unconstrained environment. This method can extract deep salient features from faces robustly to reduce the effects of various distortion types, such as lighting, occlusion, and low image resolution. Meanwhile, a conditional conceptual model was designed to enhance the expressive learning ability of the decision tree, and the facial expressions from different perspectives were modeled by conditional probability learning. In addition to building a new convolutional neural network for expression recognition, transfer learning is also used in the literature [42], [37] for expression recognition. By fine-tuning the classic convolutional neural networks, such as AlexNet [43], VGGNet, and ResNet, the abilities of these networks for expression recognition are improved. In [44], a general framework of a three-dimensional convolutional neural network was proposed, including a convolution layer, maximum pool layer, leakage layer, Gabor layer, and optical flow layer. Based on this framework, four types of specific expression recognition networks are designed, and these network decisions are fused together for expression recognition.

C. THE FUSION OF TRADITIONAL METHODS AND CONVOLUTIONAL NEURAL NETWORKS METHODS

In addition to the traditional method and the convolutional neural network, there are also some works attempting to integrate the traditional method and convolutional neural network together for facial expression recognition. In [45]–[47], local binary patterns (LBP) and convolutional neural networks were fused for facial expression recognition. In [45], YAN *et al.* proposed an improved convolutional neural network model to solve the problem of the poor stability of the traditional facial expression recognition methods. The face expression and local binary image are fused, and the original image and local binary image are used as the training data set. The expression features are extracted implicitly by continuous convolution, and then the extracted features are resampled by maximum pooling. The experimental results show that the data set with the LBP feature information has high recognition accuracy and robustness. In [46], Biao Yang *et al.* proposed a weighted hybrid deep neural network, which was used to automatically extract the features effective for the FER task, and realized facial detection, rotation correction, and data enhancement preprocessing. The parameters of this network were initialized with the VGG16 [48] model trained on the ImageNet database. In [47], a network based on appearance features was used to extract the LBP face features, which utilized the geometric features to learn the changes in the points of action units (AUs), and finally, the two features were combined for classification. In [49], Wang S *et al.* proposed a new visible expression recognition method based on thermal infrared data as privileged information, which learned a deep model with visual images and thermal images and then used the learning features to train the support vector machine (SVM [50]) classifier. Finally, a good expression recognition performance was obtained. In [51], Guohang Zeng *et al.* proposed a new feature loss method, which embeds the manual feature information into the training process of the network. Then, a general framework with traditional feature information embedded was developed and tested on the CK+, JAFFE and Fer2013 data sets. In [52], based on [49], [50], a method combining the automatic features learned by a convolutional neural network (CNN) with the manual features calculated by the visual text package (BOVW) model was proposed to achieve a good FER result. First, the k-nearest neighbor model was used to select the nearest training sample of the input test image. Second, a pair of SVM classifiers was trained on the selected training samples. Finally, the SVM classifier was used to classify and predict the test images. In [53], different levels of deep learning features extracted from the SIFT and CNN models were combined, and finally, the SVM was used to classify the mixed features. In [54], the Viola-jones method [13] was used to locate the face, and contrast limited adaptive histogram equalization (CLAHE) was used to enhance the face. Then, discrete wavelet transform (DWT) was used to extract the facial features, and finally, the extracted features are used to train the CNN network. In [55], Wang X M *et al.* proposed a new method for static facial expression recognition. The main task was to use the CNN model to divide a group of static images into 7 basic emotions and then realize the expression classification automatically. First, the standard histogram equation is used to preprocess the FER data set. Then, data enhancement is used to offset and rotate the facial image to enhance the robustness of the model. Finally, the results of the SoftMax function were obtained by SVM.

Convolutional neural networks have been widely used in image recognition. However, there are still some problems in FER based on convolutional neural networks, such as low recognition rates, high complexity, and feature loss. Aiming at these problems, this paper proposes a new facial expression recognition method, i.e., a multibranch cross-connection convolutional neural network (MBCC-CNN), which integrates the residual connection, Network in Network, and multibranch tree structure techniques. In the process of constructing residual blocks, a shortcut cross connection is added to sum the output of the convolution layer, which makes the data flow between networks more smooth. The construction of the Network in Network and multibranch tree structure module is based on the idea of a network and a multibranch structure in the network. Each branch adopts the network in the network, which increases the feature extraction ability of each perception field. Each branch extracts different image features. Finally, the different features extracted from the different branches are combined to effectively avoid the problem of feature loss. Meanwhile, after the MBCC-CNN, global mean pooling is adopted to average the feature map of the last layer, and the resulting feature vector is directly input into the SoftMax layer for classification. The experimental results on the Fer2013, CK+, FER+, and RAF data sets show that the MBCC-CNN method proposed in this paper extracts the features of the image effectively and has good recognition performance.

In summary, the contributions of this paper are summarized as follows.

- The proposed MBCC-CNN model designs a module based on residual connections and a shortcut cross connection for the summation of the convolution output layer, which makes the data flow between networks more smooth and prevents the network performance degradation.
- To extract image features effectively, the MBCC-CNN model combines the Network in Network and tree-shaped multibranch structure approaches. The MBCC-CNN model utilizes the Network in Network to learn the image features and then fuses the image features of different branches together, which can effectively avoid the problem of insufficient feature extraction.
- The global mean pooling is utilized for processing the extracted features, which can greatly reduce the number of network parameters and avoid global over fitting.

III. MATERIALS AND METHODS

This section includes the following three steps: First, the facial expression data set is preprocessed. Then, a multibranch cross-connection convolutional neural network (MBCC-CNN) is constructed for extracting more facial expression features. Finally, the SoftMax classifier is adopted to classify the extracted features. The network structure of the proposed MBCC-CNN model is shown in Fig. 1. The network is mainly composed of module 1, module 2, module 3, and the GAP layer. The process of the proposed method is described in detail below.

A. DATABASE PREPROCESSING

In this paper, the Fer2013, CK+, FER+ and RAF facial expression data sets are chosen. First, the expression data sets are normalized, which can avoid the problem of the different distributions of the training samples and test samples and improve the generalization ability of the network. Then, the normalized data are enhanced. In this paper, random scaling, flipping, translation, and rotation are utilized to enhance the data.

B. THE STRUCTURE OF THE MBCC-CNN MODEL

The MBCC-CNN network proposed in this paper is mainly constructed by three modules, which are based on the residual connection, Network in Network, and tree structure approaches, respectively. The three modules are designed as follows.

1) THE STRUCTURE OF MODULE 1

The module 1 proposed in this paper is based on the residual connection approach [56]. As shown in Fig. 2, this paper utilizes the residual connection to directly transfer input information to output via a short cut. The network only needs to learn the difference between the input and output, which simplifies the goal and difficulty of learning the features. Suppose the input of the network is x and the output is $F(x)$, then the target to be fitted is $H(x)$, and the target to be trained is $F(x) = H(x)$. For module 1, the output is $F(x) + x$, then the fitting target is $H(x) - x$, and the training target is $F(x) + x = H(x) - x$. According to the ResNet approach, it is necessary to change the identity of the shallow network, i.e., $F(x) = x$. However, in module 1, the target

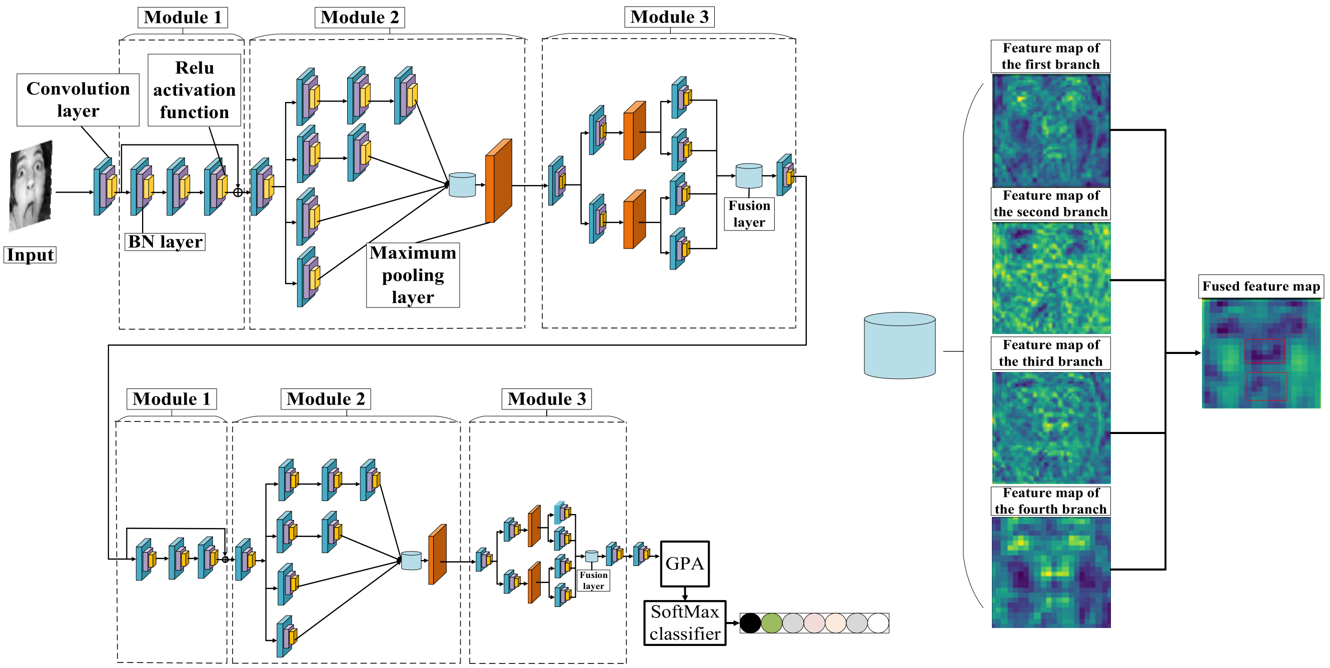


FIGURE 1. The network structure of the MBCC-CNN model.

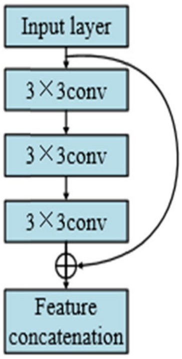


FIGURE 2. Module 1 of the MBCC network.

to be fitted becomes $F(x) + x = x$, which is equivalent to $F(x) = 0$. Obviously, this is much simpler than the original training goal. In general, the initialization of the parameters of each layer in the convolutional neural network tends to be 0. Therefore, compared with updating the parameters of the network to learn $H(x)$, the updating of the redundancy layer parameters to learn $F(x) = 0$ can converge faster. Meanwhile, the structure of the residual connection ensures that it is difficult for the gradient to be 0, and the gradient will not disappear when the updated parameters are propagated back. The residual is calculated as follows

$$Z^{[l+1]} = W^{[l+1]}a^{[l]} + b^{[l+1]} \tag{1}$$
$$a^{[l+1]} = g\left(Z^{[l+1]}\right) \tag{2}$$
$$\cdots\cdots$$
$$Z^{[l+3]} = W^{[l+3]}a^{[l+2]} + b^{[l+3]} \tag{3}$$
$$a^{[l+3]} = g\left(Z^{[l+3]} + a^{[l]}\right) \tag{4}$$

Here, $a^{[l]}$ represents the input of the residual block, l represents the current layer of the network. $Z^{[l+n]}(n = 1, 2, 3, \cdots, n)$ is the linear activation of $a^{[l]}(l = 1, 2, 3, \ldots, n)$, and $a^{[l+n]}$ $(n = 1, 2, 3, \ldots, n)$ is a non-liner activation of $Z^{[l+n]}$ $(n = 1, 2, 3, \ldots, n)$. n represents the number of layers.

2) THE STRUCTURE OF MODULE 2

The module 2 proposed in this paper is based on the Network in Network approach [57]. As shown in Fig. 3, the network in network approach is utilized for feature extraction, and more complex structures are added to each receptive field for data abstraction to enhance the discrimination ability of the model in the receptive field. Compared with a single network, the network with branches added can extract abstract features of different channels and then use the multibranch network to combine the different features, which further enhances the feature extraction ability of the network. By constructing the micronetwork Mlpconv, the micronetwork is continuously translated to cover different local areas, and different features are extracted. Meanwhile, parameter sharing is also achieved ($Mlpconv = conv + (1 \times 1)\, conv$). For the multibranch micronetwork constructed in this paper, only one 1×1 Conv is adopted, which is utilized for the full connection calculation on all features. The rest are all 3×3 Conv, which are used for

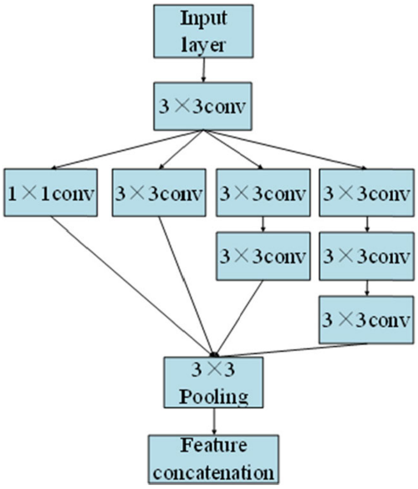


FIGURE 3. Module 2 of the MBCC network.

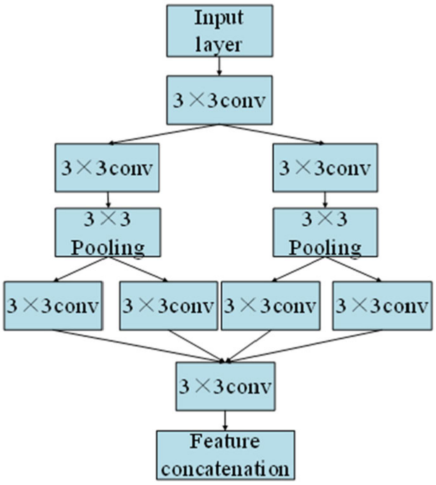


FIGURE 4. Module 3 of the MBCC network.

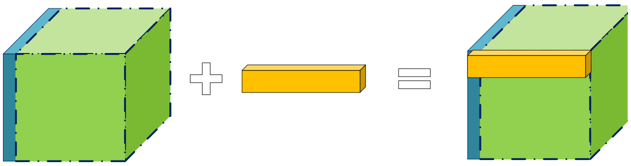


FIGURE 5. Feature fusion in the form of addition.

selective feature calculation and only extract the expression features in the field of perception. Compared with the use of multiple 1 × 1 Conv in the inception module, the redundant feature extraction that usually causes low classification accuracy is avoided. This paper constructs a multibranch micronetwork with a more complex structure to extract the features in the receiving domain, which can simulate the local network well, combine the different features extracted from different channels, and improve the effectiveness of the convolutional layer. The Mlpconv can be represented as follows:

$$f_{i,j,k_1}^{1} = \max\left(w_{k_1}^{1T}x_{i,j} + b_{k_1}, 0\right) \tag{5}$$

$$\cdots$$

$$f_{i,j,k_n}^{n} = \max\left(w_{k_n}^{nT}f_{i,j}^{n-1} + b_{kn}, 0\right) \tag{6}$$

where, (i, j) represents the position index of the image pixel, $x_{i,j}$ represents the image block in convolution window, k represents the index of the feature map to be extracted, n is the number of network layers, T represents the transpose. The first layer is linear convolution layer (the size of the convolution kernel is greater than 1), and the latter are with the convolution kernel of 1 × 1.

3) THE STRUCTURE OF MODULE 3

The Module 3 proposed in this paper is based on the idea of a tree-shaped multibranch structure. As shown in Fig. 4, the tree-branched structure is composed of multiple branches. The tree branches require a small amount of data and have a hierarchical relationship. Module 3 uses the convolutional layer and the maximum pooling layer to build a multibranch tree structure. Meanwhile, it also achieves the Network in Network effect, which is utilized to extract the image features effectively. The branch network structure can extract a variety of image features and then combine these features, which can increase the feature extraction ability of the network. The kernel size of the convolution layer and the maximum pooling layer in module 3 are all 3 × 3. In VGGNet [48], it was demonstrated that two 3 × 3 convolution kernels have the same receptive field as one 5 × 5 convolution kernel. Meanwhile, the number of parameters in two 3 × 3 convolution layers is less than that in one 5 × 5 convolution layer. Therefore, the use of a 3 × 3 convolution kernel in module 3 not only keeps the receptive field unchanged but also reduces the parameters of the convolution layer. After a layer of the convolution layer, the input data of module 3 is branched to further extract image features. Then, the different image features extracted are combined together to reduce the loss of useful features, which can improve the recognition performance of the network.

C. FUSION OF BRANCH FEATURES

For module 2 and module 3, feature fusion is carried out in the form of addition, while the number of channels remains unchanged. The process of feature fusion is shown in Fig. 5. Assuming that the image size of the input module is $W \times W \times D$, the image size becomes $W_1 \times W_1 \times D_1$ after convolution layer. Then, the size of the image is $W_1 \times W_1 \times D_1$ after the pooling layer with $F_1 \times F_1$ as the pooling core and S_1 as the step size. In this case, it is assumed that the four input channels of the module are $X_i(i = 1, 2, 3, \ldots, n)$, $Y_i(i = 1, 2, 3, \ldots, n)$, $Z_i(i = 1, 2, 3, \ldots, n)$ and $M_i(i = 1, 2, 3, \ldots, n)$ respectively. For the addition process, the fused feature is recorded as N_{add}.

$$W_1 = \frac{W + 2 \times P - F}{S} + 1 \tag{7}$$

$$D_1 = K \tag{8}$$

$$W_2 = \frac{W_1 - F_1}{S_1} + 1 \tag{9}$$

$$D_2 = D_1 \tag{10}$$

$$\begin{aligned} N_{add} &= \sum_{i=1}^{n} (X_i + Y_i + Z_i + M_i) * K_i \\ &= \sum_{i=1}^{n} X_i * K_i + \sum_{i=1}^{n} Y_i * K_i + \sum_{i=1}^{n} Z_i * K_i + \sum_{i=1}^{n} M_i * K_i \end{aligned} \tag{11}$$

where K is the number of convolution kernels, F is the size of convolution kernels, S is the step size, and P is the zero padding. K_i is the number of convolution kernels corresponding to i, '*' is the convolution symbol.

For facial expression recognition, it is very important to extract features from the key parts of human face. However, it is usually difficult to extract facial expression features accurately due to the influence of external interference factors. Moreover, some facial expression data sets have label errors, such as the Fer2013 data set. Meanwhile, compared with other image data sets, facial expression data sets are usually very small, which makes it a challenge to extract the facial features effectively and recognize facial expressions accurately. To solve these problems, this paper proposes a multiple branch cross-connected convolutional neural network, which is formed by stacking three modules. In this paper, the network makes full use of the advantages of the three modules and realizes feature extraction effectively. The three modules proposed in this paper are indispensable. The relationships and functions of the three modules in the MBCC-CNN constructed in this paper are described in detail below.

In this paper, the proposed three modules are based on the residual connection, Network in Network, and tree structure approaches. After the input image is sent to the first convolutional layer with 32 channels, the output results are sent to module 1 with 72 channels via a short cut, and the results are fused with the transmitted information that only involves learning the difference between the inputs and outputs, which reduces the difficulty of network learning. Then, the outputs of module 1 are sent to module 2 with 72 channels. Module 2 only uses a 1 × 1 convolution layer, and the others use 3 × 3 convolution layers. The information input to module 2 is extracted through multiple channels, and then the features extracted from each channel are fused. Then, down sampling is used to reduce the feature dimensions and retain the main features. Next, the outputs of module 2 are sent to module 3 with 72 channels. After one layer of convolution, the information input into the module 3 is divided into two branches. Then, through one layer of convolution, the down sampling is carried out. Finally, the information is fused after four branch convolution operations. The merged feature information is sequentially input into module 1, module 2 and module 3 with 144 channels and finally passes through the convolution layer with 288 channels.

TABLE 1. List of the network parameter settings.

Output size	The MBCC-CNN configuration	
Input:48×48×1		
46×46×32	Conv2d_1	
46×46×72	Module1	(conv2d_2, conv2d_3, conv2d_4)
22×22×72	Module2	(conv2d_5,conv2d_6,conv2d_7,conv2d_8,conv2d_9, conv2d_10,conv2d_11,conv2d_12,max_pooling2d_1)
10×10×72	Module3	(conv2d_13,conv2d_14,conv2d_15,conv2d_16,conv2d_17, conv2d_18,conv2d_19,conv2d_20,max_pooling2d_2,max_pooling2d_3)
10×10×144	Module1	(conv2d_21, conv2d_22, conv2d_23)
4×4×144	Module2	(conv2d_24,conv2d_25,conv2d_26,conv2d_27,conv2d_28, conv2d_29,conv2d_30,conv2d_31,max_pooling2d_4)
1×1×288	Convd_40	
288	GAP	
7	SoftMax	

In this paper, the whole network is mainly connected by module 1, module 2 and module 3 in sequence, and different input features are extracted from each module. The network combines the characteristics of the residual connection, Network in Network and multibranch network approaches to prevent gradient vanishing and enhance the discrimination ability of the model in the receptive field. Therefore, the proposed network can extract image features more effectively and improve the facial expression recognition ability.

The configuration of the MBCC-CNN network parameters is shown in Table 1.

Here, the convolution layer parameter is set to "strides = 1, padding = 'same' ". Only conv2d_1 in the convolution layer is set to "padding = 'valid' ". The convolution kernels of Conv2d_6 and conv2d_25 are both 1 × 1, and those of the other convolution layers are 3 × 3. The maximum pooling layer is set to "pool_size = 3, strides = 2". The weight is initialized to "he_uniform" and regularized to "l2(1e-2)". The convolution layer is followed by BN (BatchNormalization) [58] and the Relu activation function. The SoftMax loss function is adopted after the GAP layer.

D. FURTHER RESEARCH AND DISCUSSION ON MBCC-CNN

For facial expression recognition, the features to be extracted mainly include the eyebrows, eyes, nose, mouth, and facial contour, which are easily affected by illumination, occlusion, and head posture. Moreover, the commonly used face

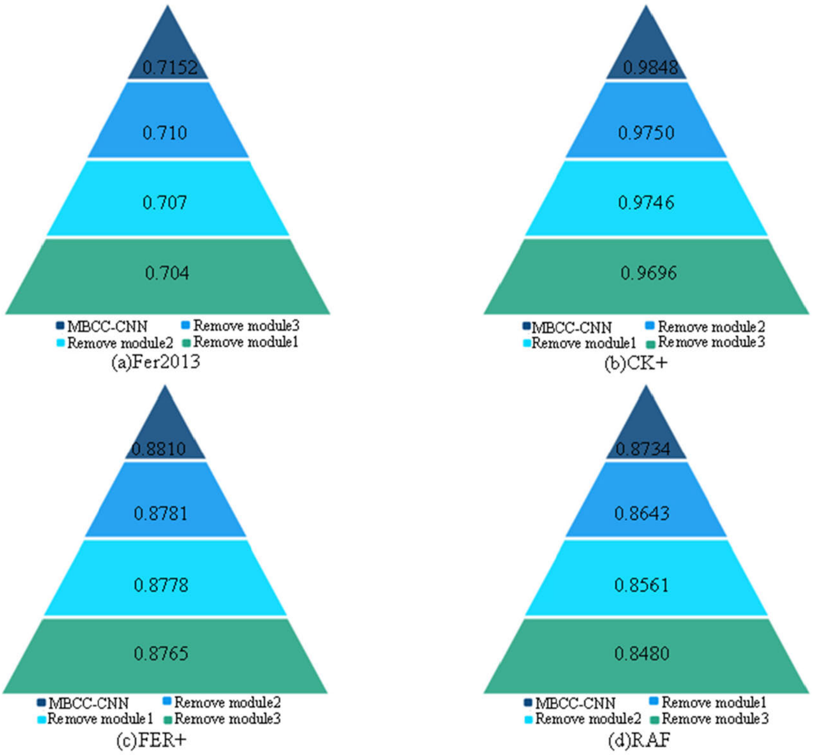


FIGURE 6. Experiments on the Fer2013, CK+, FER+ and RAF data sets.

expression data sets contain a small number of images. For example, there are 35887 images in the Fer2013 data set, 981 images in the CK + data set, 31373 images in the FER+ data set and 15339 images in the RAF data set. Compared with other image data sets, the facial expression data set is very small, which makes it a challenge to extract facial features effectively and recognize facial expressions accurately. To solve these problems, this paper proposes a multiple branch cross-connected convolutional neural network, which extracts facial expression features by different network branches first and then fuses these features together for more complete feature information. The proposed network is designed for facial expression recognition, which is a shallow network with low complexity. Although the network structure of the MBCC-CNN looks more complex than the single structure of AlexNet, the number of parameters of the MBCC-CNN model is still lower than that of AlexNet, and other network, such as VGG16 and ResNet34. With the same input image size 64 × 64 × 3, the comparison of the number of parameters of MBCC-CNN, AlexNet, VGG16, and ResNet34 is shown in Table 2. It can be seen that the parameters of MBCC-CNN are much less than those of AlexNet, VGG16 and ResNet34. The more parameters a model has, the more data it needs to train it. Facial expression recognition data set does not contain a large number of images. Therefore, the MBCC-CNN method with less parameter is more suitable for training and can avoid over fitting, which makes it more suitable for facial expression recognition.

TABLE 2. Parameter comparison of different methods.

Method(64×64×3)	Total number of parameters
AlexNet	20802951
VGG16	39917383
ResNet34	22690055
MBCC-CNN	4384751

To verify the influence of the three modules proposed in this paper on the network, some experiments are conducted. Based on the complete MBCC-CNN model, the experiments are carried out by removing module 1, module 2, and module 3, individually. The experimental results on the Fer2013, CK+, FER+ and RAF data sets are listed in Fig. 6. It can be observed from Fig. 6 that after removing module 1, module 2, and module 3, the recognition accuracies using the Fer2013, CK +, FER+ and RAF data sets are always lower than that of the MBCC-CNN. This finding shows that all three modules are indispensable and helpful to the facial expression recognition.

To further verify the effectiveness of the three modules proposed in this paper, the three proposed modules are used to replace some modules in the VGG16 and ResNet34 networks. Take module 3 for example; module 3 is used to replace the part indicated by the dotted line box of the network structures of VGG16 and ResNet34, which are shown in Fig. 7 and

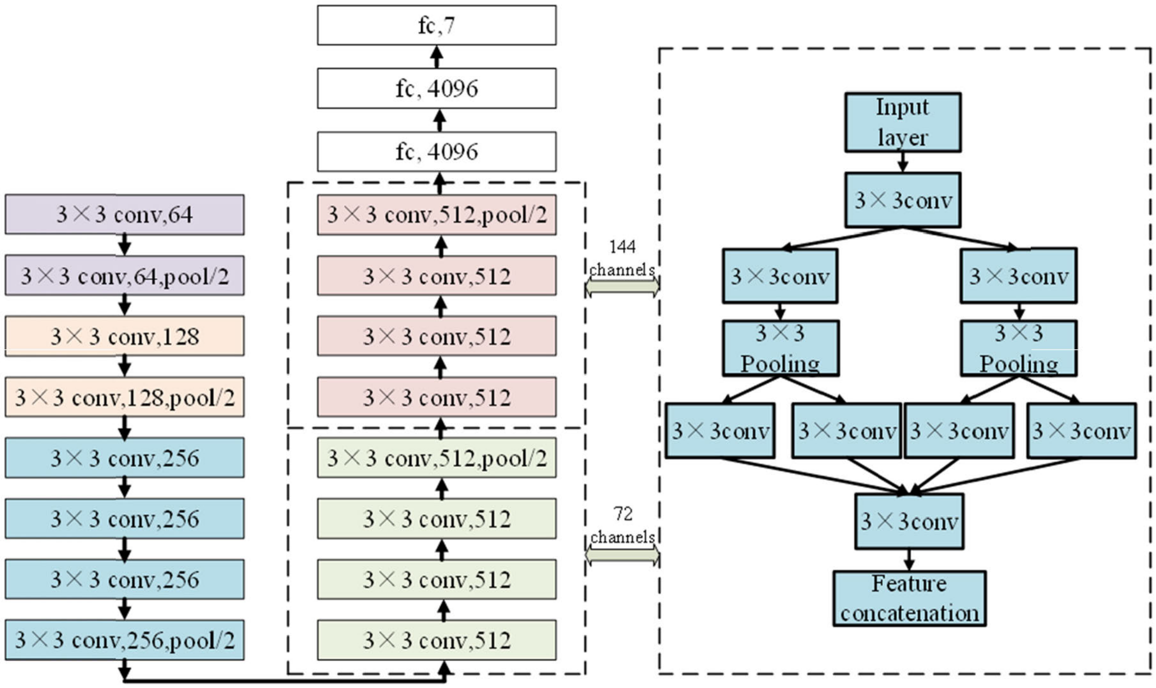


FIGURE 7. The VGG16 network structure replaced by module 3.

Fig. 8, respectively. The processes of replacing module 1 and module 2 are the same as this process, except that the dotted line box of the networks VGG16 and ResNet34 are replaced by module 1 or module 2. In this paper, the modules that partially replace VGG16 use 72 and 144 channels, and the modules that partially replace the ResNet34 use 128 channels. Some module replacement experiments are carried out on the Fer2013, CK+, FER+ and RAF data sets. The experimental results are shown in Fig. 9. Here, the VGG16_moduleX and ResNet34_moduleX ($X = 1, 2, 3$) refer to the parts of VGG16 and ResNet34 networks that are replaced by the proposed module X. As shown in Fig. 9(a), for the Fer2013 data set, after replacing the part of the VGG16 and ResNet34 networks with module 1, 2, 3, the classification accuracies are all obviously higher than those of the VGG16 and ResNet34. For the CK+, FER+ and RAF data sets, the same conclusion can be drawn. Therefore, after replacing some modules of the VGG16 and ResNet34 with the proposed modules, the recognition accuracies are obviously improved, which proves that the proposed modules have better feature extraction abilities than the existing and popular blocks, thus improving the facial expression recognition performance.

To further verify the feature extraction ability of the proposed MBCC-CNN model, the feature maps of module3 are visualized. Firstly, the feature maps of four branches of module3 are visualized separately, and then the feature maps of the fusion layer are also provided, which are shown in Fig. 10. We can see in Fig. 10 that the features extracted by each branch are different, and only a small amount of feature information can be extracted from the important parts of the face. The fuse layer can integrate and supplement the features of the four branches, which makes the final features more complete. In Fig.10, the features of fuse layer are concentrated in important parts of the face, such as eyes, nose and mouth (the important parts are marked with red box). This illustrates that the MBCC-CNN model solves the problem of insufficient feature extraction of each branch and increases the recognition performance. In [72], it has also proved that the performance of multi branch CNN is better than that of single branch CNN. In this paper, the proposed network structure fully confirms this point.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this part, first, the expression data sets used in this paper are introduced. Then, the confusion matrix analysis, recognition performance analysis, and thermodynamic diagram analysis are carried out for the proposed MBCC-CNN model. Then, the expression recognition system based on the MBCC-CNN model is designed. Finally, the proposed method is compared with some recent works with different evaluation indicators. All the experiments in this paper are performed on JetBrains PyCharm2017.1 x64, Kera2.1.4, and GeForce 940MX GPU. The parameter settings of our model are listed in Table 3.

A. EXPERIMENTAL DATA SET

To evaluate the performance of the proposed method, four facial expression data sets are chosen in this paper.

1) THE CK+ FACIAL EXPRESSION DATA SET

The CK+ data set [59] is an extension of the Cohn-Kanade data set and was published in 2010. The CK+ data set used

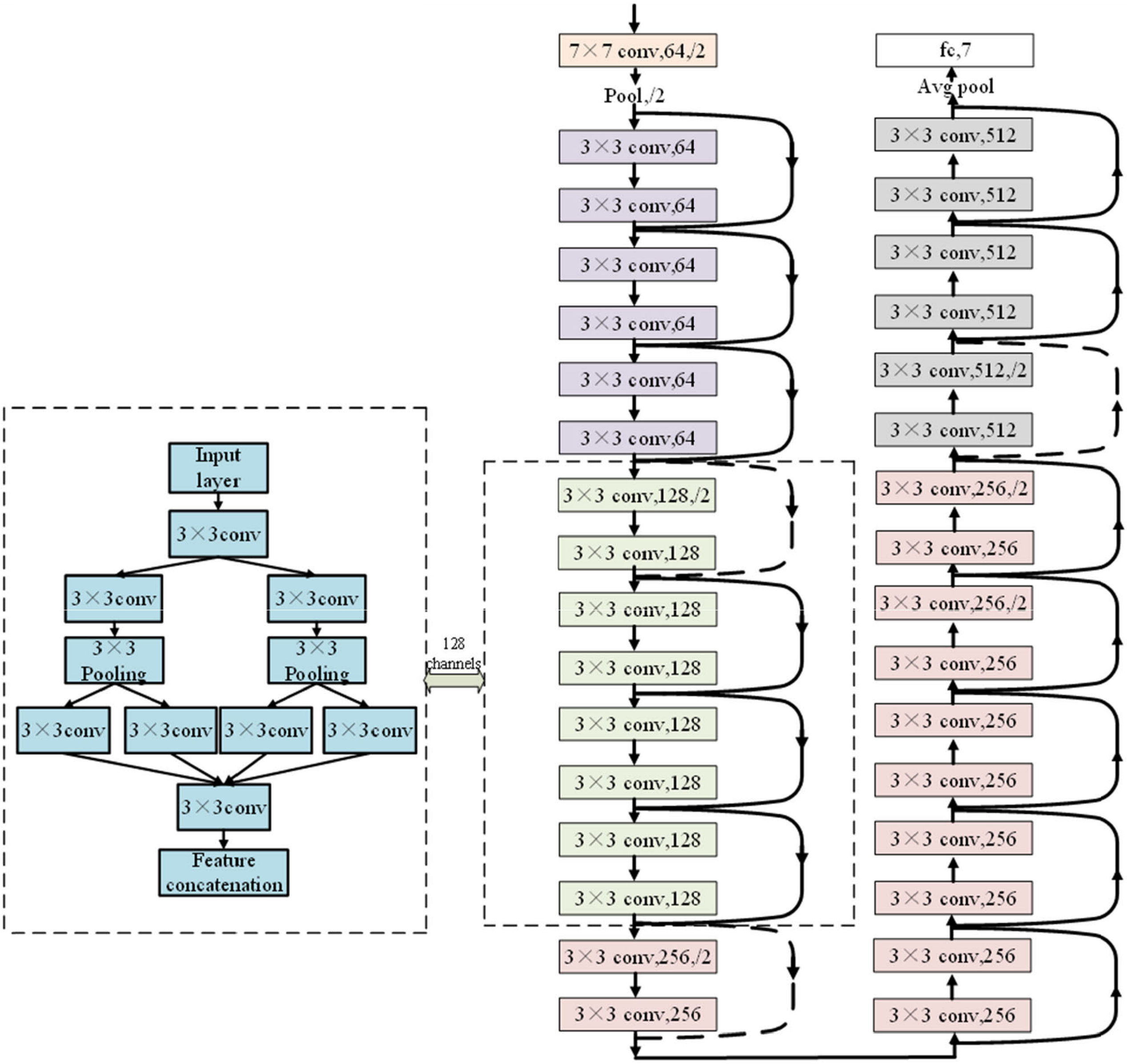


FIGURE 8. The ResNet34 network structure replaced by module 3.

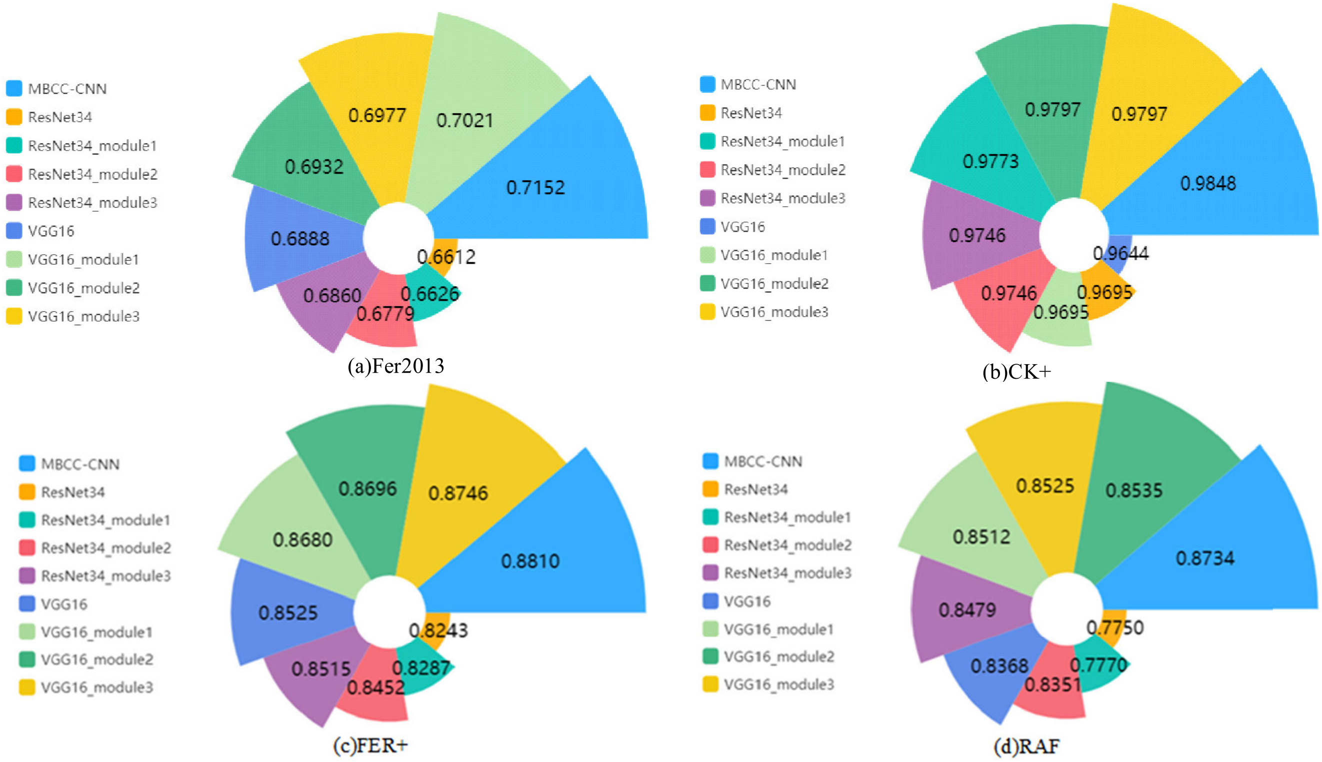
TABLE 3. The parameter settings of the model.

PROJECT	Settings
Optimizer	SGD
Momentum	0.9
Initial learning rate	0.1
Learning rate decay	ReduceLROnPlateau
L2 regularizers	1e-2

in this paper contains seven types of expressions, including 135 angry images, 54 normal images, 177 disgusted images, 75 fearful images, 207 happy images, 84 sad images, and 249 surprised images. For the CK+ data set, 10-fold cross validation is used for evaluation. The data set is randomly divided into 10 subsets, and the sampling interval is 10. Two subsets are used for testing and the rest are used for training.

2) THE FER2013 FACIAL EXPRESSION DATA SET

The Fer2013 data set [60] contains a total of 35,887 face images, including 28,709 images in the training set, 3589 images in the verification set, and 3589 images in the test set. The images in the data set are all gray images of size 48 × 48. These samples are divided into the following seven categories: 0-‘anger’, 1-‘disgust’, 2-‘fear’, 3-‘happy’, 4-‘sad’, 5-‘surprised’, and 6-‘normal’. It is important to note that there are some label errors in the test set of this data set, which lead to low test accuracy on this data set. Moreover, the recognition accuracy of human eyes on this database



(a)Fer2013 (b)CK+ (c)FER+ (d)RAF

FIGURE 9. The comparison results after module replacement on the Fer2013, CK+, FER+ and RAF data sets.

is only (65+5)%, which makes the Fer2013 data set very challenging. Although the recognition rate of this data set is not high, most researchers still choose this data set to evaluate their algorithm. In the comparison with other methods under the same conditions, this data set is also chosen in this paper.

3) THE FER+(FERPLUS) FACIAL EXPRESSION DATA SET

FER+ [69] data set is an improved version based on Fer 2013 data set, in which some original images are re-labeled, while other images (for example, without human face) are completely deleted. It is worth noting that Barsoum *et al.* [69] adds the 'contempt' as the eighth emotion. The training set of FER + data set contains 25045 images, the verification set contains 3191 images and the test set contains 3137 images.

4) THE RAF FACIAL EXPRESSION DATA SET

RAF [70] is a widely used expression data set, which contains real-world face images. These images have great changes in illumination, occlusion and background. Its 29672 images are divided into single label and double label. In this paper, the images with single label are used, which contains 15339 images, including 12271 images in the training set and 3068 images in the test set. Fig. 11 shows some image samples of the three data sets.

B. CONFUSION MATRIX ANALYSIS

The confusion matrix is mainly used to compare the actual category with the prediction category, which can evaluate the classification performance of a model from another perspective. Fig. 12 shows four confusion matrices obtained by the proposed MBCC-CNN model on the CK+, Fer2013, FER+ and RAF data sets. It can be observed from Fig. 12 that the prediction accuracy of each category is concentrated on the diagonal. The prediction accuracy for the seven categories of the CK+ data set is high, and the Fer2013 data set has low classification accuracy due to the label errors in the test set, except in the 'happy' category. Nevertheless, the Fer2013 data set is the most commonly used data set for facial expression recognition. To compare the results with other methods under the same conditions, the Fer2013 data set is also chosen for the experiment. The confusion matrix of FER+ data set and RAF data set is also shown in Fig.12. From the results of the confusion matrix, we can observe that the proposed method has good classification performance.

Fig. 13 shows some negative samples in the Fer2013 data set. In Fig. 13(a), some cartoon samples are mixed into the 'anger' category. In Fig. 13(b), some 'surprise' samples and non-facial samples are mixed into the 'disgust' category. In Fig. 13(c), some normal samples and non-facial samples are mixed into the 'fear' category. In Fig. 13(d), normal samples and non-facial samples are mixed into the category of 'sadness'. In Fig. 13(e), 'happy' sample -s and occlusion negative samples are mixed into the 'surprise' category. In Fig. 13(f), some 'happy' samples are mixed into the normal category. It is worth noting that for some categories, even the human eye can hardly distinguish the correct emotion.

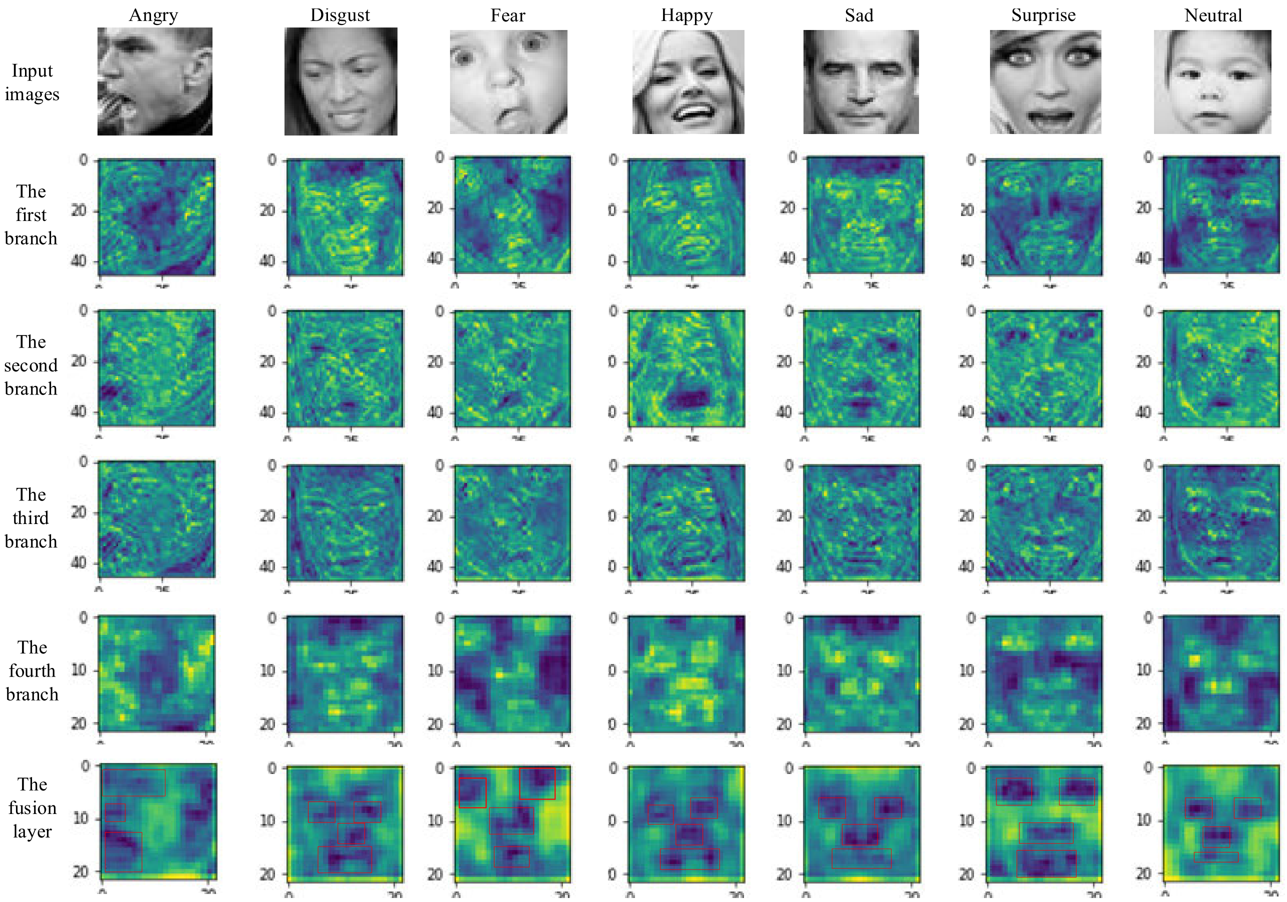


FIGURE 10. Feature maps of different branches and fusion layer.

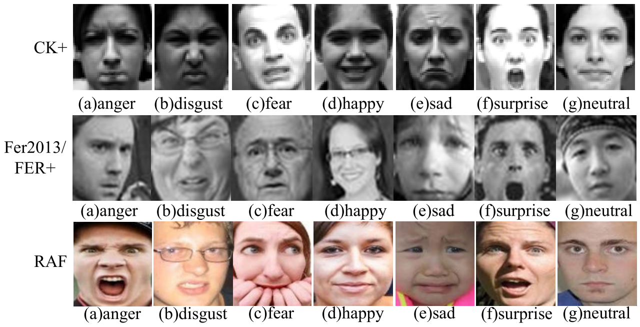


FIGURE 11. Some image samples in the CK+, Fer2013, FER+ and RAF data sets.

In general, the recognition performance of the proposed method is good, which shows that the proposed model has good generalization ability and can extract image features effectively.

C. CLASSIFICATION pERFORMANCE ANALYSIS

Fig. 14 shows the precision, recall, F1-score, and accuracy of the proposed MBCC-CNN model on the CK+, Fer2013, FER+ and RAF data sets. It can be observed from Fig. 14 that all the evaluation indicators of each category of the CK+ data set are high, and those of the Fer2013 data set are relatively low, except the 'happy' category. The reason for this phenomenon is explained in Fig. 13. Facial expression recognition based on the Fer2013 data set is very challenging because of the wrong data labels. In addition to 'disgust' category, other categories of FER+ data set are improved compared with Fer2013. All kinds of data indicators of RAF data set can also be clearly seen. Generally speaking, the experimental results show that the recognition performance of the proposed MBCC-CNN model is good. Here, the Recall, Precision, F1-score, and Accuracy are represented as follows:

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$r = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = \frac{2rP}{2 \times TP + FP + FN} \tag{14}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

TP is the number of positive samples predicted to be positive samples. FN is the number of positive samples predicted to be negative samples. FP is the number of negative samples predicted to be positive samples. TN is the number of negative samples predicted to be negative samples.

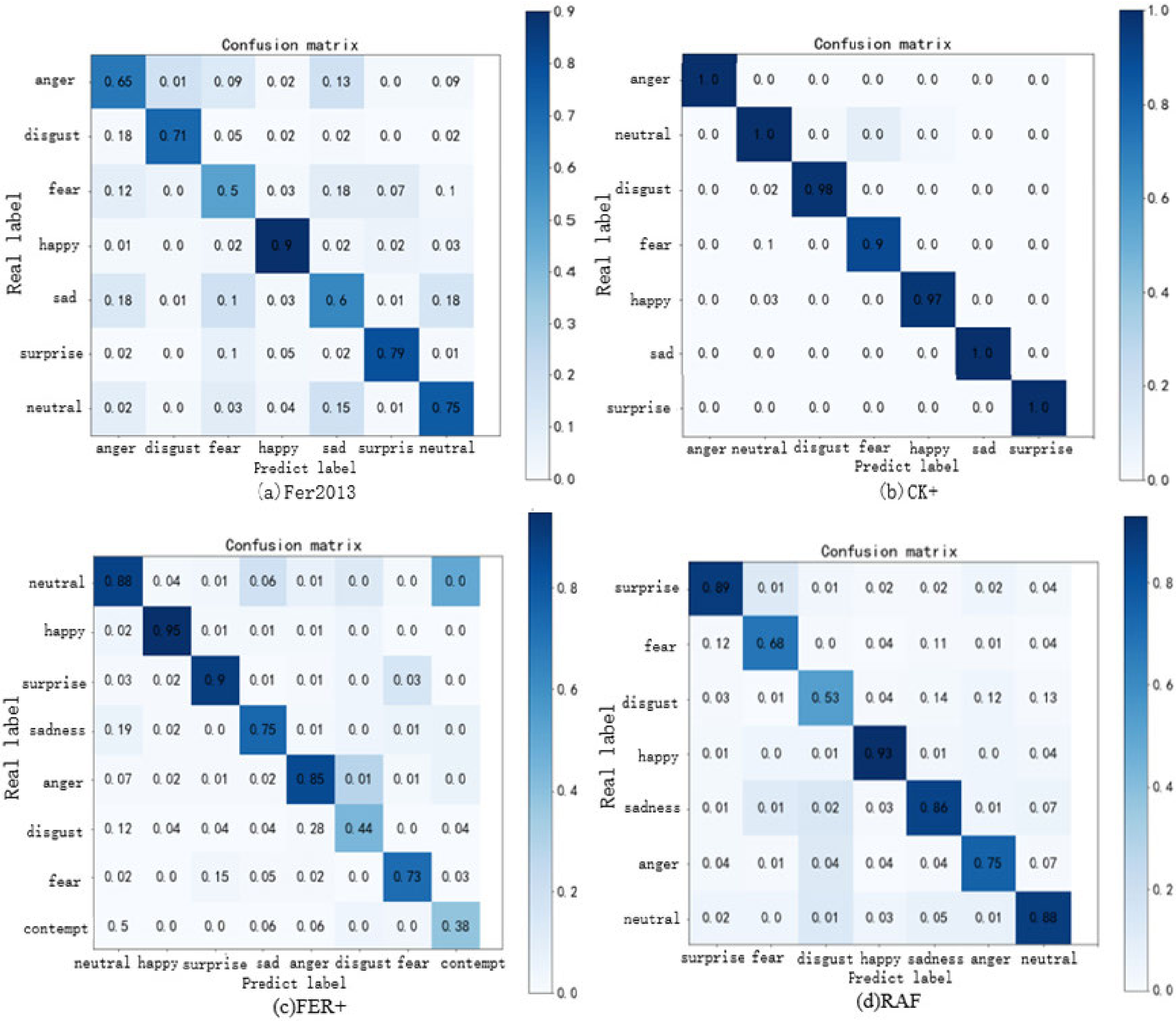
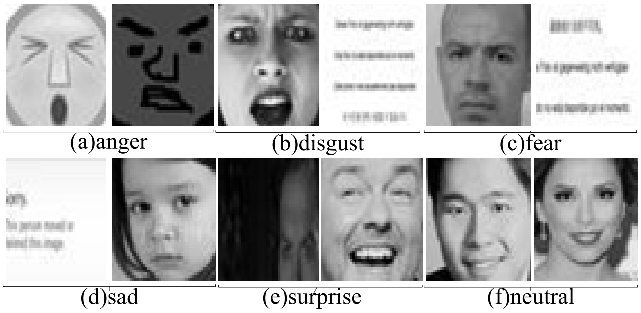


FIGURE 12. The confusion matrices obtained by the proposed method on the Fer2013, CK+, FER+ and RAF data sets.

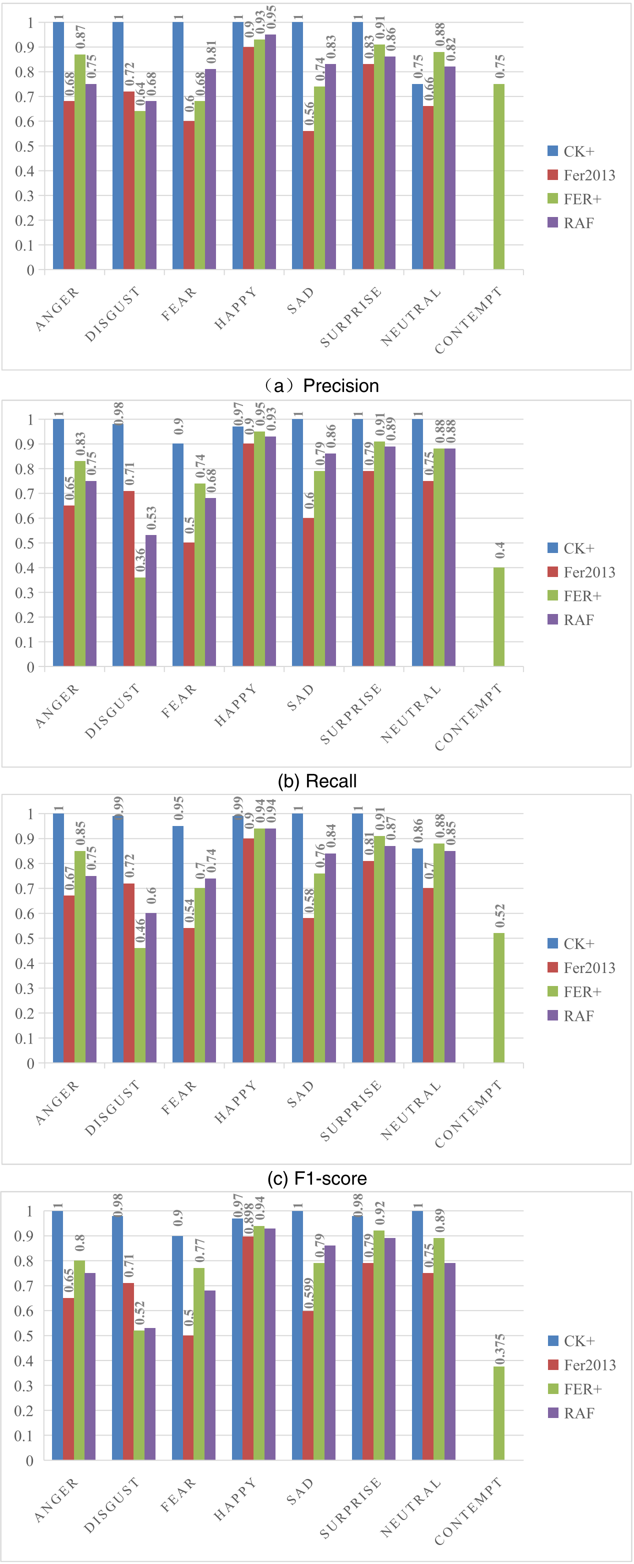

(a)anger (b)disgust (c)fear

(d)sad (e)surprise (f)neutral

FIGURE 13. Some negative samples in the Fer2013 data set.

D. HEAT MAP ANALYSIS

Heat maps are common in image recognition. In this paper, a heat map is utilized to show the features of interest on the face. A heat map can be displayed by weight, which reflects the activation values of different positions of an image. For the proposed model, the maximum output corresponds to the classification category. Starting from the nodes with the maximum probability output, back propagation is performed, and the gradient of the last convolution layer is obtained; then, the average value of each feature map is calculated. Finally, the visualization of the heat map is realized by superimposing the product of the interested part of each channel and the convolution activation values on the original image. Fig. 15 shows the comparison of the heat maps obtained by the MBCC-CNN model, the ResNet34 and VGG16 on the Fer2013, CK+, FER+ and RAF data sets. Based on these heat maps, we can clearly see the difference between the regions of interest (ROIs) extracted by each network. In Fig. 15, most of the interesting parts of the MBCC-CNN model focus on the important parts of the face, such as the nose, eyes, and mouth. By comparison, the ROIs extracted by the ResNet34 and VGG16 have interest shifts. The ROIs extracted by a network are closely related to its feature



(a) Precision

(b) Recall

(c) F1-score

(d) Accuracy

FIGURE 14. Some evaluation indicators of the proposed method on different data sets.

extraction ability, which means that the proposed model can extract facial features more accurately to achieve better facial expression recognition.

E. APPLICATION OF THE FACIAL EXPRESSION RECOGNITION SYSTEM

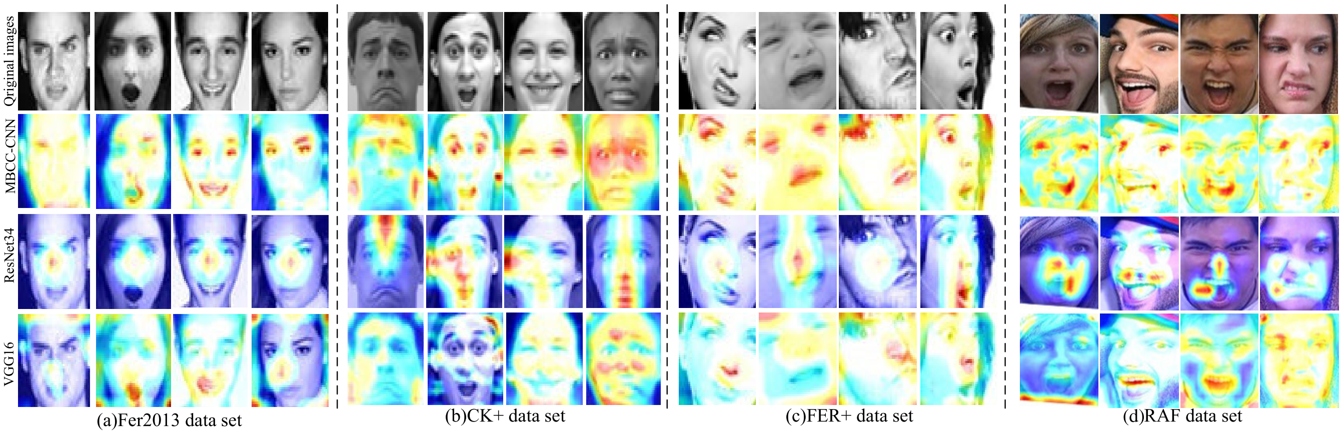
To realize real-time and intelligent expression recognition, in this paper, a facial expression recognition system is designed, which is realized by using the proposed MBCC-CNN model. Fig. 16 shows some expression recognition results on the system we designed. The first three columns show the facial expression recognition results without occlusion. The last column shows the results of the real-time facial expression recognition with occlusion test performed in our laboratory environment. In Fig. 16, we can see that, except for Fig. 16 (e), all the facial expressions can be recognized accurately, and the recognition accuracy and time are given. In Fig. 16 (e), 'disgust' is mistaken for 'angry', and it is also very difficult for human eyes to distinguish between them. In Fig. 16(d)(h)(l), the faces to be recognized are partially occluded. In this case, the proposed MBCC-CNN model can still provide good recognition performance.

The above experimental results show that the proposed model has good recognition performance and strong robustness and can accurately recognize local face images and real-time face images (even if partially occluded). The designed expression recognition system is helpful to realize the intelligent and real-time application of expression recognition.

F. COMPARISON OF RECOGNITION ACCURACY WITH OTHER METHODS

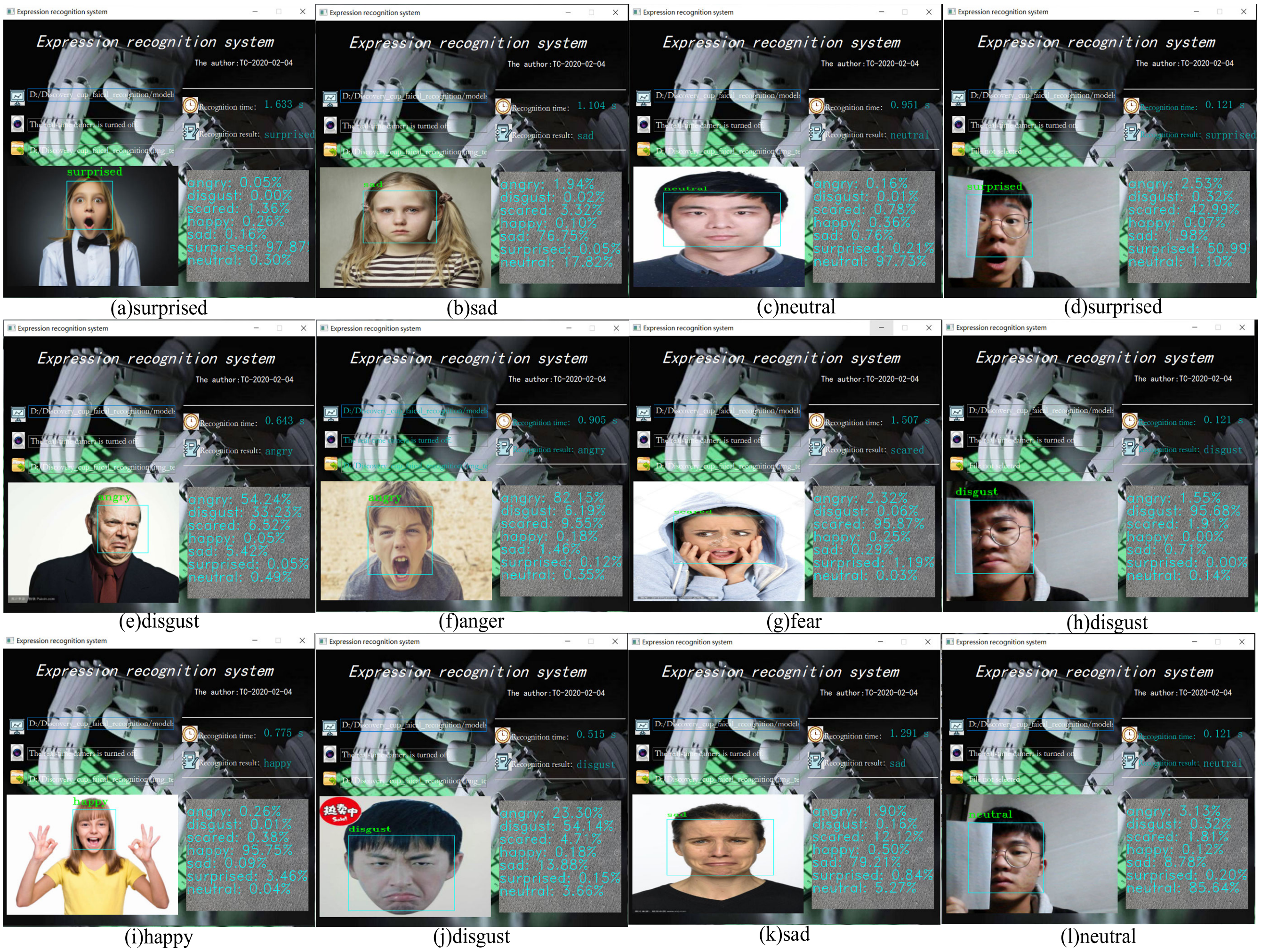
The proposed MBCC-CNN model is based on the fusion of the residual connection, Network in Network, and tree structure approaches. This model adds a data flow short-cut between networks. The Network in Network approach increases the receptive field of the convolution layer and improves the feature extraction ability of the model. In the multibranch network, the image features extracted from different branches are combined together, which effectively avoids the problem of insufficient feature extraction. The proposed MBCC-CNN method can extract image features effectively and improve the expression recognition accuracy. To fully verify the effectiveness of the proposed method, 32 related expression recognition methods are chosen for comparison under the same conditions. The comparison results of recognition accuracy on the Fer2013, CK+, FER+ and RAF data sets are listed in Table 4, Table 5, Table 6 and Table 7, respectively.

Among the above methods for comparison, some are based on traditional methods for expression recognition. Reference [23] utilized AdaBoost to segment the maximum geometric component of the face and then used the multistage Haar wavelet to extract the features of the segmented components. However, AdaBoost is sensitive to abnormal samples, which will obtain higher weights in the iterative process and, thus, affect the segmentation performance. Meanwhile, the Haar wavelet base will lead to inefficient feature extraction. For the CK+ data sets, the recognition accuracy of



(a)Fer2013 data set (b)CK+ data set (c)FER+ data set (d)RAF data set

FIGURE 15. The heat maps obtained by the MBCC-CNN model, the ResNet34, and VGG16 network methods on different data sets.



Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 1.633 s
Recognition result: surprised
surprised
angry: 0.05%
disgust: 0.00%
scared: 1.36%
happy: 0.26%
sad: 0.16%
surprised: 97.87
neutral: 0.30%

(a)surprised

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 1.104 s
Recognition result: sad
sad
angry: 1.94%
disgust: 0.02%
scared: 3.32%
happy: 0.10%
sad: 76.75%
surprised: 0.05%
neutral: 17.82%

(b)sad

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 0.951 s
Recognition result: neutral
neutral
angry: 0.16%
disgust: 0.01%
scared: 0.78%
happy: 0.36%
sad: 0.76%
surprised: 0.21%
neutral: 97.73%

(c)neutral

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned on.
File not selected
Recognition time: 0.121 s
Recognition result: surprised
surprised
angry: 2.53%
disgust: 0.32%
scared: 42.99%
happy: 0.07%
sad: 1.98%
surprised: 50.99
neutral: 1.10%

(d)surprised

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 0.643 s
Recognition result: angry
angry
angry: 54.24%
disgust: 33.23%
scared: 6.52%
happy: 0.05%
sad: 5.42%
surprised: 0.05%
neutral: 0.49%

(e)disgust

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 0.905 s
Recognition result: angry
angry
angry: 82.15%
disgust: 6.19%
scared: 9.55%
happy: 0.18%
sad: 1.46%
surprised: 0.12%
neutral: 0.35%

(f)anger

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 1.507 s
Recognition result: scared
scared
angry: 2.32%
disgust: 0.06%
scared: 95.87%
happy: 0.25%
sad: 0.29%
surprised: 1.19%
neutral: 0.03%

(g)fear

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned on.
File not selected
Recognition time: 0.121 s
Recognition result: disgust
disgust
angry: 1.55%
disgust: 95.68%
scared: 1.91%
happy: 0.00%
sad: 0.71%
surprised: 0.00%
neutral: 0.14%

(h)disgust

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 0.775 s
Recognition result: happy
happy
angry: 0.26%
disgust: 0.01%
scared: 0.38%
happy: 95.75%
sad: 0.09%
surprised: 3.46%
neutral: 0.04%

(i)happy

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 0.515 s
Recognition result: disgust
disgust
angry: 23.30%
disgust: 54.14%
scared: 4.71%
happy: 0.18%
sad: 13.88%
surprised: 0.15%
neutral: 3.66%

(j)disgust

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned off.
D:/Discovery_cup_faical_recognition/img_te
Recognition time: 1.291 s
Recognition result: sad
sad
angry: 1.90%
disgust: 0.16%
scared: 12.12%
happy: 0.50%
sad: 79.21%
surprised: 0.84%
neutral: 5.27%

(k)sad

Expression recognition system
Expression recognition system
The author:TC-2020-02-04
D:/Discovery_cup_faical_recognition/models
The real-time camera is turned on.
File not selected
Recognition time: 0.121 s
Recognition result: neutral
neutral
angry: 3.13%
disgust: 0.32%
scared: 1.81%
happy: 0.12%
sad: 8.78%
surprised: 0.20%
neutral: 85.64%

(l)neutral

FIGURE 16. The expression recognition results on our designed system.

this model is 90.48%. Among the methods for comparison, the recognition accuracy of this model is very low.

Some of the methods are based on convolutional neural networks for facial expression recognition. Reference [28]–[31] preprocessed the facial expression data set first and trained the data set with a convolutional neural network. The constructed convolutional neural network was not very effective and the recognition accuracy was low. For the Fer2013 data set, the recognition rates of these models do not exceed 68%. In reference [33]–[35], [61], the

TABLE 4. The accuracy comparison results on the FER2013 data set.

Method	Accuracy
Ref [55]	68.79%
DenseNet [25]	71.02%
GoogLeNet [25]	65.76%
VGG-Face [25]	69.18%
Ref [51]	61.86%
Ref [29]	65.03%
Ref [30]	68%
Ref [31]	66%
Ref [34]	70.02%
AlexNet [42]	66.67%
VGGNet [42]	69.41%
ResNet [42]	70.74%
MBCC-CNN(ours)	**71.52%**

TABLE 5. The accuracy comparison results on the CK+ data set.

Method	Accuracy
Ref [46]	97.02%
Ref [53]	94.82%
Ref [23]	90.48%
Ref [28]	97.38%
Ref [51]	97.35%
Ref [33]	94.67%
Ref [34]	98%
WMCNN-LSTM [36]	97.50%
Ref [35]	87.20%
Ref [54]	96.46%
Ref [37]	98.38%
Ref [44]	96.15%
Ref [47]	96.46%
Ref[61]	96.28%
MBCC-CNN(ours)	**98.48%**

TABLE 6. The accuracy comparison results on the FER+ data set.

Method	Accuracy
DenseNet [62]	86.54%
Ref [63]	85.67%
Ref [64]	87.15%
Ref[65]	85.10%
Ref[66]	82.00%
Ref[67]	84.29%
Ref[52]	87.76%
ResNet50 (transfer learning)[68]	79.90%
MBCC-CNN(ours)	**88.10%**

attention mechanism was introduced to construct the convoluted neural network for expression recognition. The effect of the attention mechanism is good, but there are some defects. When some new parameters are introduced, it may cause an over fitting phenomenon and increase the computational complexity quickly. For the CK+ data set, the recognition rate of the model mentioned in [61] is up to 96.28%, but this is at the cost of large computational complexity. In reference [36], [37], a multichannel convolutional neural network was used to fuse the features of different channels to improve the recognition performance. The recognition rates of the two methods on the CK+ data set are 97.5% and 98.38%. Reference [42] uses transfer learning to classify facial expressions. By fine-tuning classic convolutionalneural networks, including AlexNet, VGGNet, and ResNet, the feature extraction abilities of classic large networks are utilized effectively, but these networks are very complex. The recognition rates of fine-tuning AlexNet, VGGNet, and ResNet on the Fer2013 data set are 66.67%, 69.41%, 70.74%, respectively. In reference [44], a three-dimensional convolutional neural network is used for facial expression recognition. Five layers, i.e., a convolutional layer, maximum pool layer, leakage layer, Gabor layer, and optical flow layer, are defined, and four specific expression recognition networks are designed. The outputs of the four networks are integrated for facial recognition. The recognition rate of this method on the CK+ data set is 96.15%.

TABLE 7. The accuracy comparison results on the RAF data set.

Method	Accuracy
Ref[61]	85.69%
ResNet50 (transfer learning)[68]	74.76%
Ref[71]	72.46%
Ref[32]	85.07%
MBCC-CNN(ours)	**87.34%**

There are also some works that combine the traditional method with a convolutional neural network for expression recognition. Reference [46], [47] used the fusion of LBP and a convolutional neural network. In [46], a VGG16 model that trained on the ImageNet database was used for initialization. Then, the local binary patterns (LBP) features were extracted by a shallow convolution neural network based on deep identification (DeepID). Finally, the outputs of the two networks were fused in with a weighted method. The recognition rates of the method proposed in [46] and [47] on the CK+ data set are 97.02% and 96.46%, respectively. In [51], manual feature information was embedded into the training process of the network. Although this can increase some features for training, due to the incompleteness of the manual feature extraction, some features are still missing. By testing on the Fer2013 data set, a recognition rate of 61.86% is achieved. Reference [53] combines the learning features of different levels extracted from the SIFT and CNN models and finally uses SVM to classify mixed features. The recognition rate of this method on the CK+ data set is 94.82%. In reference [54], viola-jones is used to locate the face, and contrast limited adaptive histogram equalization (CLAHE) is adopted to enhance the face. Then, DWT is used to extract face features, which are used to train the CNN network. It should be noted that when the contrast of some regions of the images is too high, the CLAHE will become noisy, and some details will be lost, which will affect the classification performance. The recognition rate of this method on the CK+ data set is 96.46%.

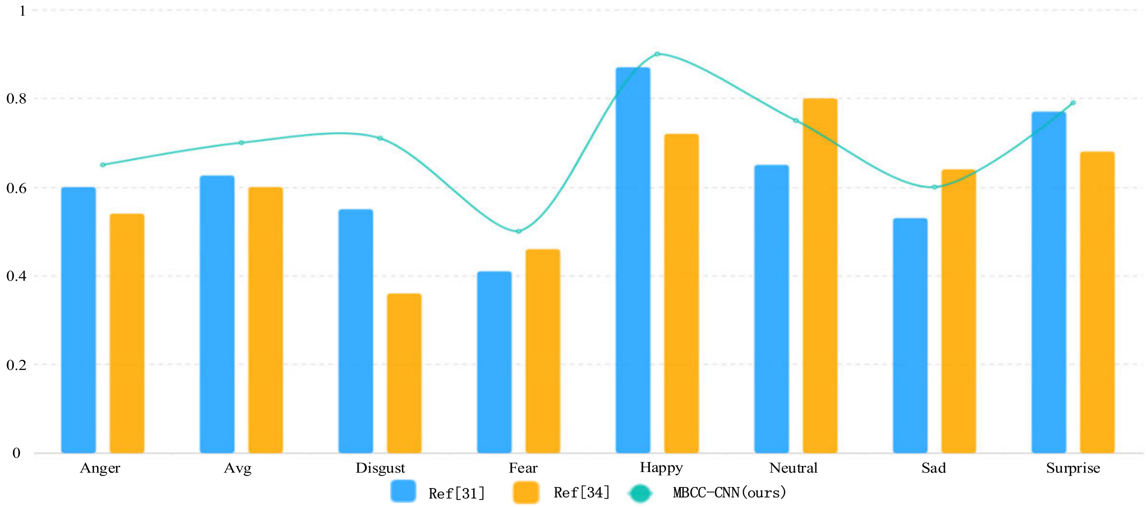


FIGURE 17. Comparison results of the confusion matrix based on the Fer2013 data set.

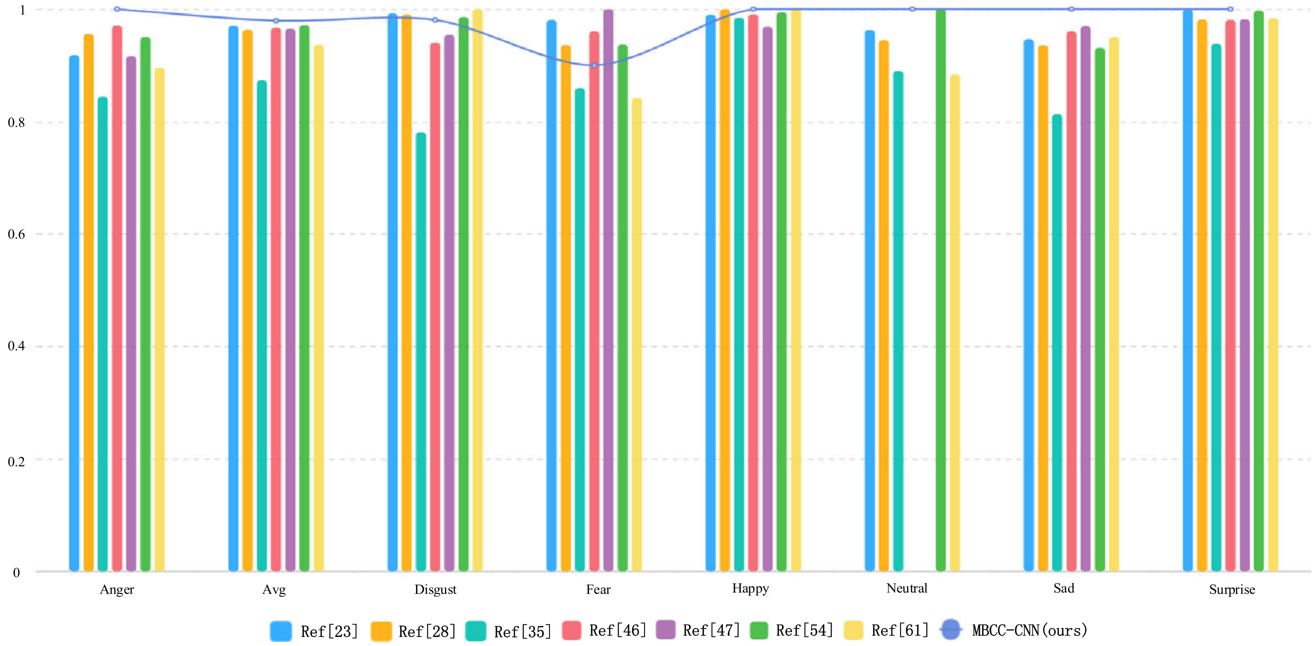


FIGURE 18. Comparison results of the confusion matrix on the CK+ data set.

The proposed MBCC-CNN model combines the residual connections, Network in Network, and multibranch tree structure approaches together. The construction of the residual block adds a shortcut cross connection for the summation of the convolution layer output, which makes the data flow between networks more smooth, ensures that the recognition accuracy does not decline, and effectively avoids the problem of gradient dissipation or explosion. For the multibranch tree structure, each branch uses the Network in Network approach, which increases the feature extraction ability of each perception field. Each branch extracts different image features and finally combines these features together, which avoids the problem of insufficient feature extraction. Meanwhile, global mean pooling is adopted, which greatly reduces the number of network parameters and avoids over fitting. The proposed MBCC-CNN method is evaluated on the Fer2013 data set and the CK+ data set. The experimental results show that the MBCC-CNN method can provide good expression recognition results on both the Fer2013 data set and the CK+ data set, with recognition rates of 71.52% and 98.48%, respectively. Compared with other related methods, the proposed

MBCC-CNN method has better facial expression recognition performance.

G. CONFUSION MATRIX COMPARISON WITH RELATED REFERENCES

Under the same conditions, the confusion matrix of the proposed MBCC-CNN method is compared with that of other expression recognition methods. Fig. 17 shows the comparison results of the three methods on the Fer2013 data set. It can be observed in Fig. 17 that, except 'neutral' and 'sad', the accuracies are slightly lower than that of reference [34], the accuracies of the other categories of the proposed method are all higher than those of the other two methods. In particular, the average accuracy of the seven categories of the proposed method is higher than that of the other two methods, which shows the effectiveness of the proposed MBCC-CNN method.

Fig. 18 shows the comparison results of the confusion matrix between the proposed MBCC-CNN method and the other seven methods on the CK+ data set. We can observe in Fig. 18 that, except for 'happy', 'fear', and 'disgust', the accuracies of the other categories obtained by the proposed method are all higher than those of the other methods. In [46], [47], the confusion matrix only gives the accuracies of six categories and lacks the accuracy of the 'neutral' category. Overall, the average accuracy of the seven categories of the proposed method in this paper is higher than those of the other seven methods, which further proves the effectiveness of the proposed method.

Fig. 17 and Fig. 18 show the comparison of the confusion matrix of the proposed method and that of the other related methods. The comparison results show that the proposed method can achieve better single class recognition results, and the overall average accuracy is the highest, which further verifies the effectiveness of the proposed method and demonstrates that the MBCC-CNN method can effectively extract the features of each category and avoid the problem of insufficient feature extraction.

H. DISCUSSION

The above experiments verify the effectiveness of the MBCC-CNN model for facial expression recognition. Compared with the traditional methods, the deep learning method based on a convolutional neural network can automatically extract image features and avoid unnecessary waste. Compared with some traditional methods and convolutional neural network fusion methods, the MBCC-CNN method has lower complexity and avoids the problem of incomplete manual information in the network, thus improving the recognition performance. The proposed MBCC-CNN model can also provide a higher recognition rate than some other network models. The reason is that the residual connection ensures a deeper network, and the multibranch network module uses the Network in Network approach to extract the image features of different branches and fuse them, and the use of global mean pooling reduces the number of network parameters and avoids over fitting. The proposed MBCC-CNN model has a good feature extraction ability, which makes it very suitable for facial expression recognition.

V. CONCLUSION

In this paper, a MBCC-CNN model is proposed for expression recognition, which combines the residual connection, Network in Network, and tree multibranch structure approaches. The proposed method first preprocesses the input images and then extracts the features of the expression images. Each feature is extracted by different network branches and then fused together, which improves the feature extraction ability of the MBCC-CNN model. Finally, global mean pooling is used to average the feature maps of the last layer, and the results are sent to the SoftMax layer directly for recognition. The experimental results show that, compared with the other methods, the proposed MBCC-CNN method always provides better expression recognition performance in terms of some evaluation indexes and has better robustness. Moreover, an expression recognition system is designed with the MBCC-CNN method. The system can recognize facial expressions quickly and accurately, which is helpful to realize the intelligent and real-time application of expression recognition. How to realize facial expression recognition in a complex environment is an issue worthy of studying. We will research this topic in the next work.

ACKNOWLEDGMENT

The authors would like to thank Ian J. Goodfellow and Patrick Lucey for providing the Fer2013 data set and CK+ data set. They also would like to thank the handling editor and the anonymous reviewers for their careful reading and helpful remarks.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] C. Darwin, P. Prodger, *The Expression of the Emotions in Man and Animals*. London, U.K.: Oxford Univ. Press, 1998.

[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 32–66, Feb. 2001.

[3] M. El Haj, P. Antoine, and J. L. Nandrino, "More emotional facial expressions during episodic than during semantic autobiographical retrieval," *Cognit., Affect., Behav. Neurosci.*, vol. 16, no. 2, pp. 374–381, Apr. 2016.

[4] A. R. Rivera, J. R. Castillo, and O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1740–1752, May 2013.

[5] H. Sadeghi and A.-A. Raie, "Approximated chi-square distance for histogram matching in facial image analysis: Face and expression recognition," in *Proc. 10th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, Isfahan, Iran, Nov. 2017, pp. 188–191, doi: 10.1109/Iranian-MVIP.2017.8342346.

[6] E. Vezzetti, S. Tornincasa, S. Moos, and F. Marcolin, "3D human face analysis: Automatic expression recognition," *Biomed. Eng., Calgary*, pp. 24–30, Feb. 2016.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[8] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, Mar. 2017.

[9] F. Ren and Z. Huang, "Automatic facial expression learning method based on humanoid robot XIN-REN," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 6, pp. 810–821, Dec. 2016.

[10] R. Gross and V. Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *Proc. 4th Int. Conf. Audio-Video-Based Biometric Pers. Authentication (AVBPA)*, Jun. 2003, pp. 10–18.

[11] S. Abe, *Feature Selection and Extraction, in Support Vector Machines for Pattern Classification*. London, U.K.: Springer, 2010, pp. 331–341.

[12] C.-R. Chen, W.-S. Wong, and C.-T. Chiu, "A 0.64 mm^2 real-time cascade face detection design based on reduced two-field extraction," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 11, pp. 1937–1948, Nov. 2011.

[13] Y.-Q. Wang, "An analysis of the Viola-Jones face detection algorithm," *Image Process. Line*, vol. 4, pp. 128–148, Jun. 2014.

[14] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1805–1812.

[15] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Washington, DC, USA, Nov. 2015, pp. 435–442.

[16] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[17] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, Aug. 2019.

[18] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognit.*, vol. 92, pp. 177–191, Aug. 2019.

[19] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, Apr. 2019.

[20] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, May 2002, pp. 229–234.

[21] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

[22] R. Girshick, J. Donahue, T. Drrell, and J. Malik, "Rich feature hierarchies for accurate object detection and demantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogit.*, Jun. 2014, pp. 580–587.

[23] M. Goyani and N. Patel, "Multi-level Haar wavelet based facial expression recognition using logistic regression," *Indian J. Sci. Technol.*, vol. 10, no. 9, pp. 1–9, Feb. 2017.

[24] R. Palermo, L. Jeffery, J. Lewandowsky, C. Fiorentini, J. L. Irons, A. Dawel, N. Burton, E. McKone, and G. Rhodes, "Adaptive face coding contributes to individual differences in facial expression recognition independently of affective factors.," *J. Experim. Psychol., Hum. Perception Perform.*, vol. 44, no. 4, pp. 503–517, Apr. 2018.

[25] T. T. D. Pham, S. Kim, Y. Lu, S.-W. Jung, and C.-S. Won, "Facial action units-based image retrieval for facial expression recognition," *IEEE Access*, vol. 7, pp. 5200–5207, 2019.

[26] S. Shi, H. Si, J. Liu, and Y. Liu, "Facial expression recognition based on Gabor features of salient patches and ACI-LBP," *J. Intell. Fuzzy Syst.*, vol. 34, no. 4, pp. 2551–2561, Apr. 2018.

[27] H. Yan, "Collaborative discriminative multi-metric learning for facial expression recognition in video," *Pattern Recognit.*, vol. 75, pp. 33–40, Mar. 2018.

[28] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, vol. 36, no. 2, pp. 391–404, Feb. 2020.

[29] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with CNN ensemble," in *Proc. Int. Conf. Cyberworlds (CW)*, Chongqing, China, Sep. 2016, pp. 163–166.

[30] V. Salunke Vibha and C. G. Patil, "A new approach for automatic face emotion recognition and classification based on deep networks," in *Proc. Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, 2017, pp. 1–5.

[31] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," 2017, *arXiv:1710.07557*. [Online]. Available: http://arxiv.org/abs/1710.07557

[32] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[33] X. Sun, P. Xia, L. Zhang, and L. Shao, "A ROI-guided deep architecture for robust facial expressions recognition," *Inf. Sci.*, vol. 522, pp. 35–48, Jun. 2020.

[34] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," 2019, *arXiv:1902.01019*. [Online]. Available: http://arxiv.org/abs/1902.01019

[35] X. Sun, S. Zheng, and H. Fu, "ROI-attention vectorized CNN model for static facial expression recognition," *IEEE Access*, vol. 8, pp. 7183–7194, 2020.

[36] H. Zhang, B. Huang, and G. Tian, "Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture," *Pattern Recognit. Lett.*, vol. 131, pp. 128–134, Mar. 2020.

[37] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognit. Lett.*, vol. 119, pp. 49–61, Mar. 2019.

[38] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[39] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, May 2018.

[40] M. Alam, L. S. Vidyaratne, and K. M. Iftekharuddin, "Sparse simultaneous recurrent deep learning for robust facial expression recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4905–4916, Oct. 2018.

[41] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, and Z. Luo, "Conditional convolution neural network enhanced random forest for facial expression recognition," *Pattern Recognit.*, vol. 84, pp. 251–261, Jul. 2018.

[42] G. Shengtao, X. Chao, and F. Bo, "Facial expression recognition based on global and local feature fusion with CNNs," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Dalian, China, Sep. 2019, pp. 1–5.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural network," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2012, pp. 1097–1105.

[44] W. Sun, H. Zhao, and Z. Jin, "A facial expression recognition method based on ensemble of 3D convolutional neural networks," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 2795–2812, Jul. 2019.

[45] Y. Yan, C. Li, Y. Lu, F. Zhou, Y. Fan, and M. Liu, "Design and experiment of facial expression recognition method based on LBP and CNN," in *Proc. 14th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Xi'an, China, Jun. 2019, pp. 602–607.

[46] B. Yang, J. Cao, and E. al, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2018.

[47] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE Access*, vol. 7, pp. 41273–41285, Mar. 2019.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[49] S. Wang, B. Pan, H. Chen, and Q. Ji, "Thermal augmented expression recognition," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2203–2214, Jul. 2018.

[50] N. O. Kadyrova and L. V. Pavlova, "Comparative efficiency of algorithms based on support vector machines for binary classification," *Biophysics*, vol. 60, no. 1, pp. 18–31, Jan-Feb. 2015.

[51] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, May 2018, pp. 423–430.

[52] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[53] X. Sun and M. Lv, "FFacial expression recognition based on a hybrid model combining deep and shallow features," *Cognit. Comput.*, vol. 11, no. 4, pp. 587–597, Aug. 2019.

[54] R. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Ahmed, "Improved facial expression recognition based on DWT feature for deep CNN," *Electronics*, vol. 8, no. 3, p. 324, Mar. 2019.

[55] X. M. Wang, J. Huang, J. Zhu, and M. Yang, "Facial expression recognition with deep learning," *Proc. 10th Int. Conf. Internet Multimedia Comput. Service*, 2018, p. 10.

[56] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. Adv. Neural Inf. Process. Syst.*, May 2016, pp. 1–9.

[57] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[59] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 94–101.

[60] I. Goodfellow, D. Erhan, P. Carrier, and A. Courville, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, 2013, pp. 117–124.

[61] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020.

[62] S. Miao, H. Xu, Z. Han, and Y. Zhu, "Recognizing facial expressions using a shallow convolutional neural network," *IEEE Access*, vol. 7, pp. 78000–78011, 2019.

[63] G. Zhao, H. Yang, and M. Yu, "Expression recognition method based on a lightweight convolutional neural network," *IEEE Access*, vol. 8, pp. 38528–38537, 2020.

[64] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," 2020, *arXiv:2001.06338*. [Online]. Available: http://arxiv.org/abs/2001.06338

[65] R. Saabni and A. Schclar, "Facial expression recognition using combined pre-trained convnets," *Comput. Sci. Inf. Technol.*, vol. 95, pp. 95–106, 2020.

[66] Z. Lian, Y. Li, J. Tao, J. Huang, and M. Niu, "Region based robust facial expression analysis," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact.*, May 2018, pp. 1–5.

[67] M. Li, H. Xu, X. Huang, Z. Song, X. Li, and X. Li, "Facial expression recognition with identity and emotion joint learning," *IEEE Trans. Affective Comput.*, early access, Nov. 9, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8528894.(49), doi: 10.1109/TAFFC.2018.2880201.

[68] B. Houshmand and N. Mefraz Khan, "Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2020, pp. 70–75.

[69] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.

[70] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.

[71] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in CNNs for facial expression recognition," in *Proc. BMVC*, 2018, p. 317.

[72] H. Zhou, X. Zhao, H. Zhang, and S. Kuang, "The mechanism of a multi-branch structure for EEG-based motor imagery classification," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2019, pp. 2473–2477.

CUIPING SHI (Member, IEEE) received the M.S. degree from Yangzhou University, Yangzhou, China, in 2007, and the Ph.D. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 2016. From 2017 to 2019, she was a Postdoctoral Researcher with the College of Information and Communications Engineering, Harbin Engineering University, Harbin. She is currently an Associate Professor with the Department of Communication Engineering, Qiqihar University. She has published two academic books about image processing and more than 50 papers in journals and conference proceedings. Her main research interests include image processing, pattern recognition, and machine learning. Her Ph.D. dissertation received the Nomination Award of Excellent Doctoral Dissertation of HIT, in 2016.

CONG TAN is currently pursuing the bachelor's degree with Qiqihar University, Qiqihar, China. His research interests include digital image processing and machine learning. He has applied for two patents of invention. His research project received two provincial Students Awards.

LIGUO WANG (Member, IEEE) received the M.S. degree and the Ph.D. degree in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively. From 2006 to 2008, he was a Postdoctoral Researcher with the College of Information and Communications Engineering, Harbin Engineering University, Harbin, where he is currently a Professor. He has published two books about image processing and more than 130 papers in journals and conference proceedings. His main research interests include image processing and machine learning.