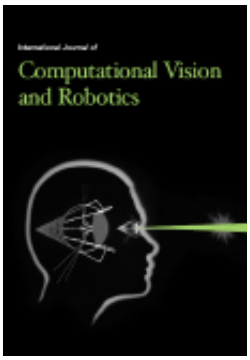


International Journal of Computational Vision and Robotics

ISSN online: 1752-914X - ISSN print: 1752-9131
https://www.inderscience.com/ijcvr

Facial expression recognition based on convolutional block attention module and multi-feature fusion

Man Jiang, Shoulin Yin

DOI: 10.1504/IJCVR.2022.10044018

Article History:

Received:	19 November 2021
Accepted:	03 December 2021
Published online:	30 November 2022

Facial expression recognition based on convolutional block attention module and multi-feature fusion

Man Jiang

Liaoning Vocational Technical College of Modern Service,
Shenyang 110034, China
Email: 459589817@qq.com

Shoulin Yin*

Software College,
Shenyang Normal University,
Shenyang 110034, China
Email: yslinhit@163.com
*Corresponding author

Abstract: In this paper, we focus on the research of facial expression recognition. A novel convolutional block attention module and multi-feature fusion method are proposed for facial expression recognition. The local feature clustering loss function is proposed, which can reduce the difference between the same classes of images and enlarge the difference between different classes of images in the training process. The convolutional block attention module is adopted to better express facial expressions in local areas with rich expressions. Experimental results show that the proposed method can effectively recognise different expressions on the RAF dataset and CK+ dataset compared with other state-of-the-art methods.

Keywords: facial expression recognition; convolutional block attention module; CBAM; multi-feature fusion; local feature clustering; LFC.

Reference to this paper should be made as follows: Jiang, M. and Yin, S. (2023) 'Facial expression recognition based on convolutional block attention module and multi-feature fusion', *Int. J. Computational Vision and Robotics*, Vol. 13, No. 1, pp.21–37.

Biographical notes: Man Jiang received his MEng from the Shenyang Normal University, Shenyang, Liaoning Province, China in 2016. Currently, he is a Lecturer in the Liaoning Vocational Technical College of Modern Service. His research interests include image processing and data mining.

Shoulin Yin received his BEng and MEng from the Shenyang Normal University, Shenyang, Liaoning Province, China in 2016 and 2013, respectively. Currently, he is a Lecturer in the Shenyang Normal University. His research interests include multimedia security, network security, image processing and data mining.

1 Introduction

Facial expression can effectively spread the human emotional states (Kim, 2021; Jiang et al., 2020a). Research on nonverbal communication shows that 55% of human emotional information is transmitted through facial expression. Facial expression recognition is an important technology in the field of computer vision (Denault and Patterson, 2021), which is widely used in medical diagnosis (Lucey et al., 2011), data analysis (McDuff et al., 2014), human-computer interaction (Yu et al., 2019), deception detection (Tsiamyrtzis et al., 2007). There are three kinds of facial expression recognition methods: geometry-based method, appearance-based method and deep learning-based method.

Geometric method is used to extract facial feature information based on shape change. Jiang et al. (2020b) developed an automatic facial analysis system, and designed a polymorphic face model and face component model to track and model various facial features. Carneiro de Melo et al. (2020) simultaneously extracted detailed parameter descriptions of facial features and recognised facial action coding system (FACS) with these parameters as inputs action unit (AU). Bao and Ma (2014) used the Bezier curve to fit facial components, accurately described the main facial information and tracked fewer feature points. However, the above methods need accurate and reliable detection, so it is difficult to track facial feature points in actual situations. In addition, the distance between facial feature points varies from person to person. Therefore, geometry-based features are not reliable.

Appearance-based methods use various local feature descriptors to extract image features, which are widely used because of their high reliability and computational efficiency. The main models include Gabor filter (Laghari et al., 2021), local binary pattern (LBP) (Yin et al., 2018), and histogram of oriented gradient (HOG) (Yin and Li, 2020). The frequency and direction of Gabor filter are similar to human visual system and suitable for texture representation and discrimination, but the feature dimension is too large to find appropriate parameter variables. LBP descriptor is a texture descriptor, which is invariant to monotonic grey change and computational efficiency, but it is sensitive to direction information. HOG feature has good illumination and geometric invariance. However, the above methods have weak adaptability to large datasets and unconstrained conditions.

In recent years, deep learning-based methods have been widely used in the field of facial expression recognition (Laghari et al., 2018, 2019; Wang et al., 2020a, 2020b). These methods can effectively overcome the shortcomings of geometry-based methods and appearance-based methods and greatly improve the recognition effect. Although deep learning has strong feature learning ability, there are still problems when they are applied to facial expression recognition. First, in order to solve this problem, researchers use additional task-oriented data to train the network from scratch or fine tune the published pre-training model to improve the performance of facial expression recognition. Jung et al. (2015) constructed two complementary small deep networks to overcome the problem of a small amount of data, and a new integration method of joint fine tuning was proposed. Jain et al. (2020) presented the novel multi-angle optimal pattern-based deep learning method to rectify the problem from sudden illumination changes, find the proper alignment of a feature set by using multi-angle-based optimal configurations. Ramya et al. (2020) proposed a facial expression recognition system in which two channels of featured images were used to represent a 3D facial scan. Features were extracted from the

local binary pattern and local directional pattern using a fine-tuned pre-trained AlexNet and a shallow convolutional neural network (CNN). The feature sets were then fused together using canonical correlation analysis. The fused feature set was fed into a multi-support vector machine (SVM) classifier to classify the expressions into seven basic categories: anger, disgust, fear, happiness, neutral, sadness and surprise.

Another problem encountered in the application of deep learning method to facial expression recognition comes from the different attributes of each individual, such as age, gender, expression range, etc. Zhao et al. (2016) proposed a facial expression recognition architecture, which implicitly embedded the natural evolution from calm expression to peak expression in the learning process, and amplified the subtle differences between expressions with weak expression, which realised the invariance of expression intensity and reduced the influence of facial identity in feature extraction. Yang et al. (2018) used conditional generative adversarial networks (cGAN) to generate neutral facial images with the same identity corresponding to expression images, and learned the components related to facial expressions retained in the generation model to classify facial expressions. However, on the other hand, some studies (Happy and Routray, 2015; Majumder et al., 2014; Wang et al., 2020c) have also confirmed the effectiveness of the local features of facial expression in solving the problem of expression recognition.

Different from the dataset under laboratory control, the real-world dataset is closer to the reality of life, and the images of the dataset have high expression polymorphism. For the same expression, there are great differences in facial angle, occlusion and expression amplitude. The deep neural network trained on softmax loss function has weak expressiveness and poor representativeness. Some previous methods have studied it and achieved good results. Wen et al. (2016) introduced the central loss of face recognition, which directly aimed at one learning goal, that is, intra class compactness, and the learning features from the same individual were more similar. Li et al. (2017) proposed local preserving (LP) loss, which aimed to aggregate local adjacent expressions and make the intraclass local clustering of each class compact. Compared with centre loss, LP loss is more flexible, especially when the class condition distribution is multimodal. However, the similarity between different classes is not considered in LP loss.

In view of the above shortcomings, this paper proposes a local feature clustering (LFC) loss function to reduce the intraclass differences and increase the differences between different images. At the same time, this paper proposes a multi-feature fusion convolutional neural network (MFF-CNN) framework. The framework consists of three separate CNNs. One extracts the overall features from the overall face image. The other two extract the local features from the cropped face image. The overall features represent the integrity of the expression, while the local features focus on the emotion rich local areas. The supervision layer with LFC loss function is introduced into the three branch networks respectively, so that the extracted features are more robust and more discriminative. Finally, the three branch networks are fused for extraction. The convolutional block attention module (CBAM) is adopted to better express facial expressions in local areas with rich expressions. Compared with a single overall feature, the extracted features are richer and more diverse. The experimental results on different datasets show the effectiveness of the new model.

2 Proposed facial expression recognition

Figure 1 is the proposed facial expression recognition framework. The proposed framework uses the VGG16 network to randomly cut the input image size into 224 × 224 × 3 for data enhancement. The network structure of VGG16 is shown in Table 1. The large convolution kernel of 5 × 5 and 7 × 7 are discarded and the convolution kernel of 3 × 3 is used to reduce the amount of data. The pooling layer adopts the maximum pooling operation with a step size of 2. After 13 convolution layers, five pooling layers and three full connection layers, finally, the output of the network is the
seven-dimensional feature vector representing different facial expressions. In this paper, the supervision layer is introduced into the VGG16 network and the LFC loss function is used, as shown in Figure 2.

Table 1 Parameter of VGG16

Layer	*Stride*	*Number of kernel*	*Output size*
Conv1_1	3 × 3	64	64 × 224 × 224
Conv1_2	3 × 3	64	64 × 224 × 224
Maxpool1	2	--	64 × 112 × 112
Conv2_1	3 × 3	128	128 × 112 × 112
Conv2_2	3 × 3	128	128 × 112 × 112
Maxpool2	2	--	128 × 56 × 56
Conv3_1	3 × 3	256	256 × 56 × 56
Conv3_2	3 × 3	256	256 × 56 × 56
Conv3_3	3 × 3	256	256 × 56 × 56
Maxpool3	2	--	256 × 28 × 28
Conv4_1	3 × 3	512	512 × 28 × 28
Conv4_2	3 × 3	512	512 × 28 × 28
Conv4_3	3 × 3	512	512 × 28 × 28
Maxpool4	2	--	512 × 14 × 14
Conv5_1	3 × 3	512	512 × 14 × 14
Conv5_2	3 × 3	512	512 × 14 × 14
Conv5_3	3 × 3	512	512 × 14 × 14
Maxpool5	2	--	512 × 7 × 7
FC6/FC7	--	--	4,096
FC8	--	--	7

2.1 LFC loss function

In real life, each individual has different expressions. For example, the expression ‘happy’ can include ‘smile’, ‘laugh’, ‘smile’ with the face covered, ‘smile’ with the side face, etc. As shown in Figure 3, although x_1 and x_3 are both ‘happy’ expressions, x_1 shows ‘laughter’ while x_3 shows ‘smile’. In addition, there may be similarities between different expressions due to expression polymorphism. As shown in Figure 3, due to the ‘open

mouth' action, 'laugh' x_1, which means 'happiness', has a high similarity with x_2, which means 'surprise'. This seriously affects the expressiveness of extracted features, as shown in Figure 3(a), where $f(x_i)$ represents extracted features.

Figure 1 Proposed method framework (see online version for colours)

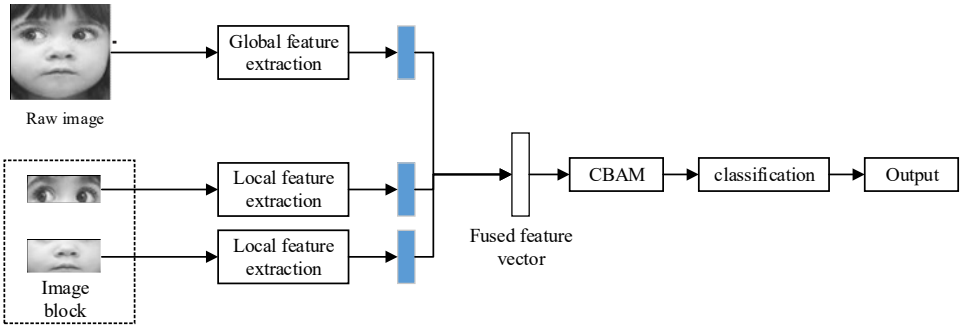


Figure 2 VGG16 network combining LFC loss layer

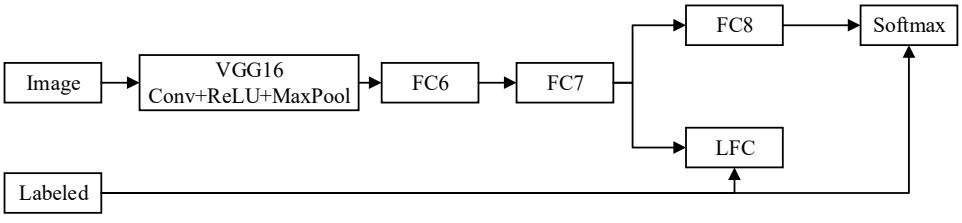
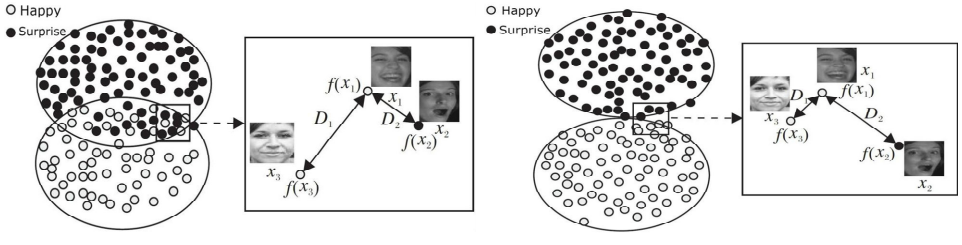


Figure 3 Feature space in deep learning training process, (a) the similiarity between different facial expression recognition (b) expected effect using LFC loss function



(a) (b)

Due to the polymorphism of the expression, the difference D_1 between $f(x_1)$ and $f(x_3)$ is greater than the difference D_2 between $f(x_1)$ and $f(x_2)$. This creates an indistinguishable 'intersection' between different expression clusters. Therefore, LFC loss function is proposed in this paper, aiming to reduce the differences between similar facial expression images and increase the differences between different kinds of facial expression images in the training process of deep neural network, which will make each kind of facial expression aggregation more compact. The final result is shown in Figure 3(b).

Local retention loss preserves the locality of each sample to make the local neighbourhood of each class more compact. The specific function is as follows:

$$L_{lp} = 0.5\sum_{i=1}^{n}\left\| x_i - \frac{1}{k}\sum_{x\in N_k\{x_i\}} x \right\|_2^2 \tag{1}$$

where $N_k\{x_i\}$ represents the set of k nearest neighbour samples with the same category as x_i. x_i represents the eigenvector of the i^{th} sample from the full connection layer before the supervision layer. $\frac{1}{k}\sum_{x\in N_k\{x_i\}} x$ represents the centre of the set of k nearest neighbour samples with the same category as x_i. n is the batch size of each processing. By minimising the local retention loss L_{lp}, each sample can be closer to the sample centre of its k nearest neighbours, making the local neighbourhood in each class more compact and reducing the intra class difference. The final loss function $L = L_s + \lambda L_{lp}$, where L_s represents softmax loss and L_{lp} represents local retention loss. The λ is used to balance the two losses.

Local retention loss reduces the local intra class loss of samples, but does not consider the similarity between classes. This paper intends to reduce the intra class difference and increase the difference between different images at the same time.

In the iterative optimisation process of CNN, sample x_i finds the nearest heterogeneous sample x_j within the k_1 nearest neighbour range. x_i and x_j are the characteristics of the extracted deep convolution network, and n is the minimum number of samples processed in batch. At the same time, in order to reduce the impact of noise on the training process, a variable x_c is introduced to represent the centres of all samples similar to x_j within the k_1 nearest neighbour range of x_i. x_c is defined as follows:

$$x_c = \frac{1}{k_2}\sum_{x\in N_{k_2}\{x_j\}} x \tag{2}$$

where $N_{k_2}\{x_j\}$ represents the sample set with the same category as x_j in the k_1 nearest neighbour set of x_i, and the number of samples in the set is k_2. We use the parameter δ to balance x_i and x_c. At this time, the difference L_a of different images is defined as follows:

$$\begin{aligned} L_a &= 0.5\sum_{i=1}^{n}\left\| x_i - \delta x_j - (1-\delta)x_c \right\|_2^2 \\ &= 0.5\sum_{i=1}^{n}\left\| x_i - \delta x_j - (1-\delta)\frac{1}{k_2}\sum_{x\in N_{k_2}\{x_j\}} x \right\|_2^2 \end{aligned} \tag{3}$$

The purpose of softmax loss function is to reduce the difference between the network output result and the real result. Similarly, the local retention loss also makes the difference within the class smaller and smaller. In the training process, the difference L_a between similar images should be larger and larger. Therefore, the LFC loss function is defined as follows:

$$\begin{aligned} L_{LPA} &= L_{lp} + \lambda_2 \frac{1}{L_a + \gamma} \\ &= L_{lp} + \lambda_2 \sum_{i=1}^{n} \left(0.5 \left\| x_i - \left(\delta x_j + \frac{1-\delta}{k_2} \right) \right\|_2^2 + \gamma \right)^{-1} \end{aligned} \tag{4}$$

where L_{lp} is the local retention loss function to reduce the intra class difference, and the rest increases the difference between different images. The introduction of parameter γ is to prevent the possible gradient explosion and infinite loss, which has no impact on the final recognition rate. In this paper, $\gamma = 2$ is set as the balance factor.

In this paper, the supervision method combined with softmax loss is adopted. softmax reflects the loss acting on the whole, The LFC reflects the local loss, which reduces the intra class difference and increases the inter class difference, so as to make the features extracted in the training process more judgmental. The loss function of the final deep neural network is as follows:

$$L = L_S + \lambda_1 L_{LFC} \tag{5}$$

where L_S represents softmax loss and L_{LFC} represents LFC loss. The two loss functions are balanced by hyperparametric λ_1. Algorithm 1 represents the forward learning and backward learning processes with LFC loss function.

Algorithm 1 Proposed method with LFC loss function

Input: training data $\{x_i\}_{i=1}^n$, batch size n, network learning rate μ, hyperparametric λ.

Output: parameter W.

Initialising $t = 0$, W, loss parameter θ of softmax loss

1 $t = t + 1$;

2 Search for the heterogeneous sample x_j closest to x_i in the k_1 nearest range of x_i;

3 In the range of k_1 nearest neighbours, the number of samples of the same kind as x_j is calculated as k_2. Calculate the k_2 nearest neighbour centre of x_j:

$$C_i^t = \frac{1}{k_2} \sum_{x^t \in N_{K_2}\{x_j\}} x^t$$

4 Update softmax loss parameter:

$$\theta^{t+1} = \theta^t - \mu^t \frac{\partial L_s^t}{\partial \theta^t}$$

5 Back propagation:

$$\frac{\partial L^t}{\partial x_i^t} = \frac{\partial L_s^t}{\partial x_i^t} + \lambda \frac{\partial L_{LFC}^t}{\partial x_i^t}$$

6 Network layer parameters:

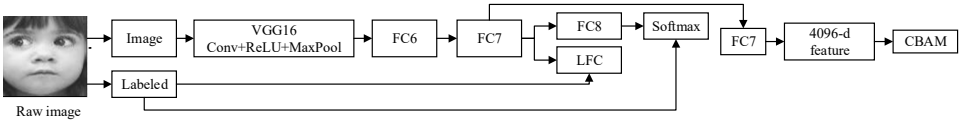
$$W^{t+1} = W^t - \mu^t \frac{\partial L^t}{\partial W^t} = W^t - \mu^t \sum_{i=1}^{n} \left(\frac{\partial L^t}{\partial x_i^t} \frac{\partial x_i^t}{\partial W^t} \right)$$

Until the convergence

2.2 Feature extraction and classification

After the proposed framework training with LFC loss function, features need to be extracted after obtaining the optimal deep learning model, as shown in Figure 4.

Figure 4 Extracting features from FC7 layer



In this paper, the whole image is input into a branch network to extract the whole feature, and the feature vector of FC7 layer is taken as the whole vector. For the other two, Face++ API is used in this paper to select appropriate feature points in the Face image as the boundary, and to cut the rectangular region that can best express emotion as the local Face image. Local images are imported into CNN network as input to extract local features. Except for some detailed configurations of CNN module, the network used and all steps of feature extraction are the same as those of overall feature extraction. Although the same CNN network is used, the extracted features have strong complementarity due to different input images. Therefore, the fusion of the two features not only increases the number of features, but also makes the extracted features more expressive.

In this paper, the overall feature is expressed as f_h^0, and the extracted local features are respectively expressed as f_1^1 and f_1^2. The fusion feature f_a is connected by the overall feature and local feature. It is expressed as follows:

$$f_a = \left(f_h^0, f_1^1, f_1^2\right) \tag{6}$$

The fusion features are imported into SVM classifier for classification, and the final results are obtained:

$$label_{out} = \arg\max f_a \tag{7}$$

2.3 CBAM

CBAM (Woo et al., 2018; Oxu et al., 2021; Wang et al., 2019) is a dual attention selection module combining channel and space in two different directions, which can be convolved in multiple directions to achieve better results. The principle is that the key features in the input data are identified by a new layer of weight assignment, so that the neural network can learn the feature areas that need attention in the input data. The principle of CBAM is shown in Figure 5.

CBAM takes the output matrix $F \in R^{C\times H\times W}$ of the original neural network convolution layer as the input matrix of this module. The module performs maximum pooling and average pooling on the channel dimension of the input matrix, and compresses and merges the two different channel descriptors. The combined descriptor generates channel weight matrix $F' \in R^{C\times H\times W}$ through the hidden layer of a single convolution kernel, as shown in equation (8). As an extension of channel attention module, spatial attention module uses maximum pooling and average pooling in spatial dimension and compresses information into a channel descriptor. The space weight matrix is obtained through the

calculation of space compression operation, as shown in equation (9). The above operations model the importance between pixels to effectively highlight the information area.

$$F' = M_s(F) \circ F \tag{8}$$

$$F'' = M_s(F') \circ F' \tag{9}$$

where $M_s(F) \in R^{C\times H\times W}$ and $M_s(F') \in R^{C\times H\times W}$ represent the three-dimensional channel compression weight matrix and space compression weight matrix respectively. ○ represents multiplication of matrix elements.

Figure 5 CBAM framework

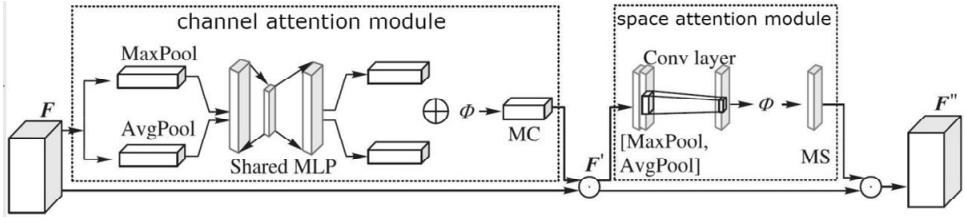
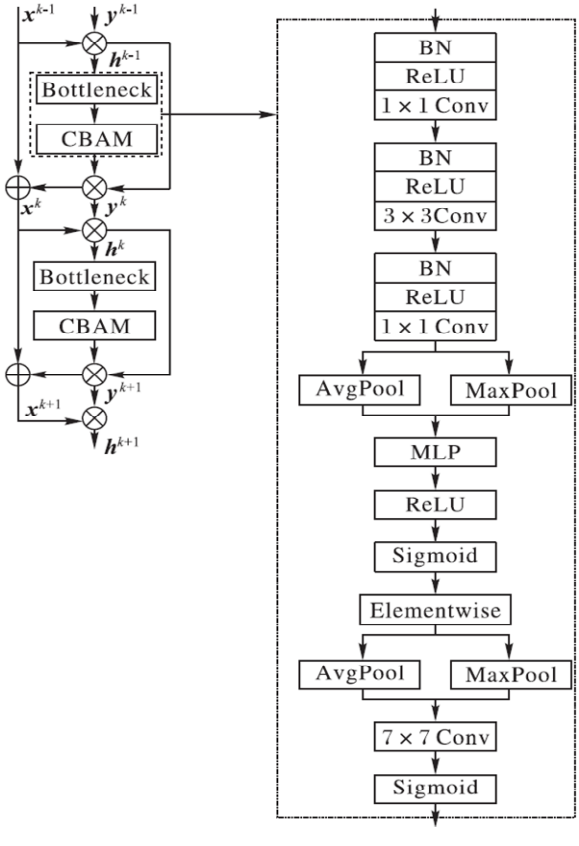


Figure 6 Calculation unit structure of single structure block



A new mirco-block structure is formed by fusing CBAM structure into each of the bottleneck layers proposed in this paper. After obtaining the shared weight of the bottleneck layer, the new CBAM structure can enhance the useful features and suppress useless features according to the importance of channels and spaces, so as to enhance the feature expression ability of the network and effectively highlight the information region in the feature matrix output by the convolutional layer. The cell structure of a single block in the new CBAM is shown in Figure 6.

In Figure 6, h^{k-1} is the feature mapping matrix of the output at $k-1$ layer, namely, the nonlinear feature mapping obtained after the convolution transformation of the previous mirco-block. The process of h^{k-1} passing through bottleneck layer and CBAM can be formulated as follows:

$$B_1^k = W_1^k * f\left(BN\left[h^{k-1}\right]\right) \tag{10}$$

$$B_2^k = W_2^k * f\left(BN\left[B_2^k\right]\right) \tag{11}$$

$$B_3^k = W_3^k * f\left(BN\left[B_3^k\right]\right) \tag{12}$$

$$U_k = W^k * f\left(\left[B_3^k\right]\right) \tag{13}$$

where W_1^k, W_2^k and W_3^k represent the weight matrices of the three convolutional layers of the Bottleneck layer respectively. B_1^k, B_2^k and B_3^k are the feature mapping matrices corresponding to the output of the three convolution layers. W^k represents the weight matrix of 7 × 7 convolution layer. BN (Campo et al., 2021) represents the batch normalisation of each eigenmatrix. $f(\cdot)$ represents ReLU activation function, and * represents the convolution operation of the matrix.

Input the feature matrix of the new CBAM single structure through the process summarised in equations (10)~(13) bottleneck formula to assign weight and obtain the new feature mapping matrix B_3^k. In order to further highlight the effective areas in the feature matrix, B_3^k enters CBAM and passes through channel and space dimensions in the module successively. The average pooling layer and the maximum pooling layer in the two dimensions can complete feature dimension reduction and achieve effective feature highlighting. Multi-layer perceptron (MLP) (Fan et al., 2018) in the channel dimension can change the dimension of the output feature matrix as needed. A new feature mapping matrix U_k is obtained by extracting spatial and channel bi-dimensional features to effectively highlight information regions.

3 Experiments and analysis

This section carries out experiments on unconstrained datasets (RAF datasets) and datasets under laboratory conditions (CK+ datasets) to verify the effectiveness of the method. The objectives are as follows:

1 study the role of LFC loss function in solving the problem of facial expression recognition

2 study the role of integrating global and local features in solving the problem of facial expression recognition

3 analyse and discuss the trade-offs of relevant parameters in this method

4 evaluate the performance of this method in facial expression recognition task through recognition accuracy, and make comparative analysis.

3.1 Experimental results on the RAF dataset

The RAF dataset used in this experiment belongs to an unconstrained dataset. RAF provides a single label expression subset and a mixed label expression subset. This paper only uses the single label expression part. The single label expression dataset contains a total of 15,339 images, including surprise (1,619), fear (355), disgust (877), happiness (5,957), sadness (2,460), anger (867) and calm (3,204), There are seven emotions in total, and the age range of participants is 0~70 years old. 52% are women, 43% are men, and 5% are still uncertain. For racial distribution, 77% of Caucasians, 8% of African Americans and 15% of Asians. RAF dataset is a real-world expression dataset, with high image resolution and label reliability. All single label expression datasets are divided into 5, 4 are used as training sets, 1 as test set on the RAF dataset, for the overall facial image, this paper directly uses the processed dataset provided in Li et al. (2017), with a scale of 100 × 100. For the local facial image, this paper selects appropriate feature points to segment the facial image according to 37 facial landmarks provided by face++ API, with scales of 40 × 90 and 55 × 90 respectively.

Proposed model is fine tuned on the pre-trained VGG16 model. The pre-training model is the best model trained on more than three million face datasets, which can effectively extract facial features and accelerate network convergence. Setting the basic learning rate = 0.01, momentum = 0.9 and training times = 20,000. Fusing the overall features extracted by CNN with two local features. The fused features are introduced into SVM classifier for classification.

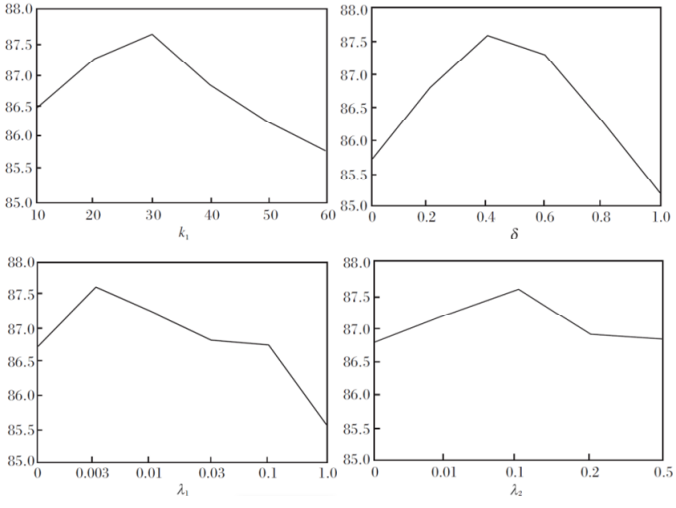
Due to the large difference in the number of various expression images in the dataset, the method of 'weighted average' can better reflect the performance of the method in the dataset. This paper constructs two other methods to undertake task evaluation: the network model for extracting overall features (abbreviated as HCNN) and the network model for extracting local features (abbreviated as LCNN). The weighted average recognition rates of the three comparison methods are as follows: proposed method is 87.72%, HCNN is 87.05% and LCNN is 82.91%. It can be seen that the proposed method has the highest weighted average recognition rate, It shows that the fused features can improve the recognition rate. Because the global features and local features can complement each other, the fused features have a stronger representation of expression.

Now, the model without fusion feature of LFC loss function is abbreviated as proposed-non-LFC. The CNN module of this new model has only VGG16 model, and also uses CGG_Face to realise the fine tuning.

The weighted average recognition rate of proposed-non-LFC is 86.84%, which is lower than 87.72% of proposed method. This can be attributed to the fact that LFC loss function reduces the difference within the same class and increases the difference between different classes during training. Therefore, LFC loss function is effective.

Next, the effects of parameters k_1, δ, λ_1 and λ_2 on the final expression recognition rate are studied and analysed through four groups of experiments. The specific experimental results of proposed method are shown in Figure 7. In the first group of experiments, fixed $\delta = 0.4$, $\lambda_1 = 0.003$ and $\lambda_2 = 0.1$, k_1 changes from 10 to 60, and the weighted average recognition rate curve is shown in Figure 7(a). When $k_1 = 30$, the best recognition rate is obtained. In the second group of experiments, fixed $k_1 = 30$, $\lambda_1 = 0.003$ and $\lambda_2 = 0.1$, δ changes from 0 to 1, as shown in Figure 7(b). When $\delta = 1$, due to ignoring the impact of noise on the training process, a lower recognition rate is obtained, and when $\delta = 0.4$, the best recognition effect is obtained. In the third group of experiments, $k_1 = 30$, $\delta = 0.4$, $\lambda_2 = 0.1$, and λ_1 changes from 0 to 1. As shown in Figure 7(c), when $\lambda_1 = 1$, a lower recognition rate is obtained, because the softmax loss function should play a leading supervisory role in the training process, at the same time, a larger λ_1 will lead to slow convergence. The best recognition effect is obtained when $\lambda_1 = 0.003$. In the fourth group, when $k_1 = 30$, $\delta = 0.4$, $\lambda_1 = 0.003$, λ_2 changes from 0 to 0.5. As shown in Figure 7(d), the best recognition rate is obtained when $\lambda_2 = 0.1$.

Figure 7 Recognition results of proposed method on RAF dataset with parameter varied



In order to verify the performance of the proposed method, comparative experiments are carried out. The comparison method is as follows: AlexNet network (Alex) + linear discriminant analysis (LDA) (Alex + LDA) (Li et al., 2017), Alex + SVM (abbreviated as Alex + SVM) (Li et al., 2017), deep locality-preserving (DLP) + LDA (abbreviated as DLP+LDA) (Li et al., 2017), DLP + SVM (Li et al., 2017), MRE-CNN (AlexNet) (Fan et al., 2018), MRE-CNN (VGG16) (Fan et al., 2018) and the IPA2LT(LTNet) (Zeng et al., 2018; Acharya et al., 2018).

The results of recognition rate are shown in Table 2. It can be seen from the table that the weighted average recognition rate and direct average recognition rate of the method in this paper are both the highest.

Meanwhile, the confusion matrix of each expression class is shown in Table 3 (the best values are ital). As can be seen from the table, among all the seven categories of expressions, the recognition rate of happy is the highest, reaching 95.72%. The recognition rate of fear is the lowest at 58.22%. Such a large difference is mainly caused

by the difference in the number of various expression categories in RAF dataset. The number of happy is 1,619 at most, and the number of afraid is 355 at least, which is far less than other expressions. Meanwhile, nearly 21.62% of fear expressions are classified as surprise, because in RAF dataset, fear and surprise have some similar movements in local facial regions, such as mouth opening and eyebrows raising. It is also because surprise had 1,619 images, which is far more than fear. Therefore, in the training process, the learning features of surprise and fear with high similarity are more inclined to surprise.

Table 2 Recognition accuracy of different methods on RAF datasets/%

Method	*Weighted average recognition rate*	*Direct average recognition rate*
Alex + LDA	67.12	47.90
Alex + SVM	68.24	55.71
DLP + LDA	69.15	71.09
DLP + SVM	84.61	74.31
MRE-CNN (AlexNet)	--	74.89
MRE-CNN (VGG16)	--	76.84
IPA2LT (LTNet)	86.88	76.91
Oxu et al. (2021)	87.11	77.52
Proposed	87.72	78.93

Table 3 Confusion matrix of seven types of expressions

	Surprise	*Fear*	*Disgust*	*Joy*	*Sad*	*Anger*	*Calm*
Surprise	*87.34*	1.71	1.11	2.63	0.80	1.41	4.45
Fear	21.51	*58.22*	0	5.30	6.65	3.94	3.94
Disgust	2.49	0	*62.61*	9.27	11.14	6.77	7.39
Joy	0.31	0	0.57	*95.50*	0.23	0.31	2.42
Sad	0	0.31	2.40	4.59	*84.21*	0.52	6.63
Anger	2.36	1.76	5.45	7.30	1.12	*76.04*	5.44
Calm	1.80	0	2.10	3.42	4.01	0	*88.35*

3.2 Experimental results on the CK++ dataset

The lab dataset is based on controlled conditions, so there are not many distractions, and the facial expression is perfectly represented. The experiment uses the Cohn-Kanade dataset, an extension of the dataset under laboratory conditions, consisting of 593 video sequences from 123 subjects, ranging in duration from 10 to 60 frames. The participants, aged 18 to 50, are 69% female, 81% European American, 13% African American and 6% other groups. This is an expanded version of the CK database, the most widely used dataset of laboratory conditions for studying facial expression recognition problems. The CK+ database provides a sequence of 327 emotion labels, containing seven emotions (anger, contempt, disgust, fear, happiness, sadness and surprise), with a size of 640 × 490 for each image. In order to compare with Liu et al. (2014, 2015), this paper eliminates the contempt expression sequence and gets 309 sequences. For each image sequence, the last

three frames with peak expression are selected for training and testing, and a dataset with 927 images is obtained.

This paper refers to the method of face alignment in Kazemi and Sullivan (2014). The face detector is used to obtain the facial region of the image, and the included angle between the symmetrical feature points around the eyes and the horizontal line is calculated. According to this angle, the image is rotated to a certain extent, and then the face region is scaled to 256 × 256 using bilinear interpolation. In this paper, histogram equalisation is used to improve contrast and grey tone changes, so that the image is clearer.

In order to obtain local images with rich emotional features, for CK+ dataset, 68 facial feature points in reference to Kazemi and Sullivan (2014) are selected to segment the facial image, and the image is divided into two parts with scales of 70 × 194 and 128 × 200 respectively. In this paper, CK+ datasets are divided into ten parts, nine are used as training sets and one as test sets. Ten times of cross-validation are carried out. For a fair comparison with the other methods, all ten subsets on the CK+ dataset are subject-independent. In order to see the generalisation ability of the trained method on a dataset (CK+ dataset) under laboratory conditions, we implement six expression classifications on CK+ dataset, but do not use fine-tuning. In the framework of deep learning, the basic learning rate is 0.001, momentum is 0.9, and training time is 20,000.

Comparative method is as follows: Happy and Routray (2015), DLP-CNN (Li et al.., 2017), boosted deep belief network (BDBN) (Liu et al., 2014), action unit deep networks (AUDN) (Liu et al., 2015; Kazemi and Sullivan, 2014; Mollahosseini et al., 2016), the recognition results are shown in Table 4. It can be seen from the table that the performance of the proposed method is equal to or even better than that of other methods on CK+ datasets under laboratory conditions.

Table 4 Recognition rates with different methods on CK+ dataset

Method	*Cross validation*	*Recognition rate/%*
Mollahosseini et al. (2016)	5	93.31
AUDN	10	93.81
Happy and Routray (2015)	10	94.80
Acharya et al. (2018)	10	95.01
DLP-CNN	5	95.89
BDBN	8	96.81
Singh and Nasoz (2020)	10	96.87
Proposed	10	97.03

4 Conclusions

In this paper, a multi-class feature fusion CNN framework is proposed to solve the facial expression recognition problem. This paper also studies the methods of image segmentation and cropping to obtain the optimal local features. In order to study facial expression recognition in the real world, reduce the influence of facial expression polymorphism and extract effective facial expression features, this paper proposes the LFC loss function to expand the differences of different kinds of facial expression on the

basis of reducing the differences within the same class in the process of deep learning training. The CBAM is adopted to better express facial expressions in local areas with rich expressions. Finally, experiments are carried out on real world RAF datasets and laboratory controlled datasets to verify the effectiveness of the proposed method.

Acknowledgements

This study was supported by the Scientific Research Funds of Education Department of Liaoning Province in 2021 (General Project) (LJKZ1311).

References

Acharya, D., Huang, Z., Paudel, D.P. and Van Gool, L. (2018) 'Covariance pooling for facial expression recognition', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.480–4807, DOI: 10.1109/CVPRW.2018.00077.

Bao, H. and Ma, T. (2014) 'Feature extraction and facial expression recognition based on Bezier curve', *2014 IEEE International Conference on Computer and Information Technology*, pp.884–887, DOI: 10.1109/CIT.2014.140.

Campo, F., Neri, M., Villegas, O. et al. (2021) 'Auto-adaptive multilayer perceptron for univariate time series classification', *Expert Systems with Applications*, Vol. 181, p.115147 [online] https://www.sciencedirect.com/science/article/abs/pii/S0957417421005881.

Carneiro de Melo, W., Granger, E. and Lopez, M.B. (2020) 'Encoding temporal information for automatic depression recognition from facial analysis', *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1080–1084, DOI: 10.1109/ICASSP40776.2020.9054375.

Denault, V. and Patterson, M.L. (2021) 'Justice and nonverbal communication in a post-pandemic world: an evidence-based commentary and cautionary statement for lawyers and judges', *Journal of Nonverbal Behavior*, Vol. 45, pp.1–10 [online] https://link.springer.com/article/10.1007/s10919-020-00339-x#citeas.

Fan, Y., Lam, J.C.K. and Li, V.O.K. (2018) 'Multi-region ensemble convolutional neural network for facial expression recognition', *Artificial Neural Networks and Machine Learning – ICANN 2018, Lecture Notes in Computer Science*, Springer, Cham, Vol. 11139, pp.84–94 [online] https://doi.org/10.1007/978-3-030-01418-6_9.

Happy, S.L. and Routray, A. (2015) 'Automatic facial expression recognition using features of salient facial patches', *IEEE Transactions on Affective Computing*, 1 January–March, Vol. 6, No. 1, pp.1–12, DOI: 10.1109/TAFFC.2014.2386334.

Jain, D.K., Zhang, Z. and Huang, K. (2020) 'Multi angle optimal pattern-based deep learning for automatic facial expression recognition', *Pattern Recognition Letters*, Vol. 139, pp.157–165 [online] https://www.sciencedirect.com/science/article/abs/pii/S0167865517302313.

Jiang, D., Li, H. and Yin, S. (2020a) 'Speech emotion recognition method based on improved long short-term memory networks', *International Journal of Electronics and Information Engineering*, Vol. 12, No. 4, pp.147–154.

Jiang, C., Wu, J., Zhong, W. et al. (2020b) 'Automatic facial paralysis assessment via computational image analysis', *Journal of Healthcare Engineering*, Vol. 2020, No. 5, pp.1–10.

Jung, H., Lee, S., Yim, J., Park, S. and Kim, J. (2015) 'Joint fine-tuning in deep neural networks for facial expression recognition', *2015 IEEE International Conference on Computer Vision (ICCV)*, pp.2983–2991, DOI: 10.1109/ICCV.2015.341.

Kazemi, V. and Sullivan, J. (2014) 'One millisecond face alignment with an ensemble of regression trees', *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1867–1874, DOI: 10.1109/CVPR.2014.241.

Kim, P.W. (2021) 'Image super-resolution model using an improved deep learning-based facial expression analysis', *Multimedia Systems*, Vol. 27, pp.615–625, https://link.springer.com/article/10.1007/s00530-020-00705-1.

Laghari, A.A., He, H., Khan, A., Kumar, N. and Kharel, R. (2018) 'Quality of experience framework for cloud computing (QoC)', in *IEEE Access*, Vol. 6, pp.64876–64890, DOI: 10.1109/ACCESS.2018.2865967.

Laghari, A.A., He, H., Shafiq, M. et al. (2019) 'Application of quality of experience in networked services: review, trend & perspectives', *Syst. Pract. Action Res*., Vol. 32, pp.501–519 [online] https://doi.org/10.1007/s11213-018-9471-x.

Laghari, A.A., Wu, K., Laghari, R.A. et al. (2021) 'A review and state of art of internet of things (IoT)', *Arch. Computat. Methods Eng*. [online] https://doi.org/10.1007/s11831-021-09622-6.

Li, S., Deng, W. and Du, J. (2017) 'Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2584–2593, DOI: 10.1109/CVPR.2017.277.

Liu, M., Li, S., Shan, S. et al. (2015) 'AU-inspired deep networks for facial expression feature learning', *Neurocomputing*, Vol. 159, No. 1, pp.126–136.

Liu, P., Han, S., Meng, Z. and Tong, Y. (2014) 'Facial expression recognition via a boosted deep belief network', *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1805–1812, DOI: 10.1109/CVPR.2014.233.

Lucey, P. et al. (2011) 'Automatically detecting pain in video through facial action units', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, June, Vol. 41, No. 3, pp.664–674, DOI: 10.1109/TSMCB.2010.2082525.

Majumder, A., Behera, L. and Subramanian, V.K. (2014) 'Emotion recognition from geometric facial features using self-organizing map', *Pattern Recognition*, Vol. 47, No. 3, pp.1282–1293.

McDuff, D., El Kaliouby, R., Senechal, T. et al. (2014) 'Automatic measurement of ad preferences from facial responses gathered over the internet', *Image and Vision Computing*, Vol. 32, No. 10, pp.630–640.

Mollahosseini, A., Chan, D. and Mahoor, M.H. (2016) 'Going deeper in facial expression recognition using deep neural networks', *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1–10, DOI: 10.1109/WACV.2016.7477450.

Oxu, Q., Qiao, X., Liu, C. et al. (2021) 'Automated ECG classification using a non-local convolutional block attention module', *Computer Methods and Programs in Biomedicine*, Vol. 203, No. 7, p.106006.

Ramya, R., Mala, K. and Nidhyananthan, S.S. (2020) '3D facial expression recognition using multi-channel deep learning framework', *Circuits, Systems, and Signal Processing*, Vol. 39, No. 2, pp.789–804.

Singh, S. and Nasoz, F. (2020) 'Facial expression recognition with convolutional neural networks', *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, pp.324–328, DOI: 10.1109/CCWC47524.2020.9031283.

Tsiamyrtzis, P., Dowdall, J., Shastri, D. et al. (2007) 'Imaging facial physiology for the detection of deceit', *International Journal of Computer Vision*, Vol. 71, No. 2, pp.197–214.

Wang, D., Gao, F., Dong, J. and Wang, S. (2019) 'Change detection in synthetic aperture radar images based on convolutional block attention module', *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pp.1–4, DOI: 10.1109/Multi-Temp.2019.8866962.

Wang, X., Yin, S., Liu, D. et al. (2020a) 'Accurate playground localisation based on multi-feature extraction and cascade classifier in optical remote sensing images', *International Journal of Image and Data Fusion*, Vol. 11, No. 3, pp.233–250, DOI: 10.1080/19479832.2020.1716862 (2020.1.15)EI(JA).

Wang, X., Yin, S., Sun, K. et al. (2020b) 'GKFC-CNN: modified Gaussian kernel fuzzy C-means and convolutional neural network for Apple segmentation and recognition', *Journal of Applied Science and Engineering*, Vol. 23, No. 3, pp.555–561.

Wang, H., Wei, S. and Fang, B. (2020c) 'Facial expression recognition using iterative fusion of MO-HOG and deep features', *The Journal of Supercomputing*, Vol. 76, No. 5, pp.3211–3221.

Wen, Y., Zhang, K., Li, Z. and Qiao, Y. (2016) 'A discriminative feature learning approach for deep face recognition', *ECCV 2016, Lecture Notes in Computer Science*, Springer, Cham, Vol. 9911, pp.499–515 [online] https://doi.org/10.1007/978-3-319-46478-7_31.

Woo, S., Park, J., Lee, J.Y. and Kweon, I.S. (2018) 'CBAM: convolutional block attention module', *ECCV 2018, Lecture Notes in Computer Science*, Springer, Cham, Vol. 11211, pp.3–19 [online] https://doi.org/10.1007/978-3-030-01234-2_1.

Yang, H., Ciftci, U. and Yin, L. (2018) 'Facial expression recognition by de-expression residue learning', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2168–2177, DOI: 10.1109/CVPR.2018.00231.

Yin, S. and Li, H. (2020) 'Hot region selection based on selective search and modified fuzzy C-means in remote sensing images', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 13, pp.5862–5871, DOI: 10.1109/JSTARS.2020.3025582.

Yin, S., Zhang, Y. and Karim, S. (2018) 'Large scale remote sensing image segmentation based on fuzzy region competition and Gaussian mixture model', *IEEE Access*, Vol. 6, pp.26069–26080, https://ieeexplore.ieee.org/document/8357569.

Yu, J., Li, H. and Yin, S. (2019) 'New intelligent interface study based on K-means gaze tracking', *International Journal of Computational Science and Engineering*, Vol. 18, No. 1, pp.12–20.

Zeng, J., Shan, S. and Chen, X. (2018) 'Facial expression recognition with inconsistently annotated datasets', *ECCV 2018, Lecture Notes in Computer Science*, Springer, Cham, Vol. 11217, pp.227–243 [online] https://doi.org/10.1007/978-3-030-01261-8_14.

Zhao, X. et al. (2016) 'Peak-piloted deep network for facial expression recognition', *ECCV 2016, Lecture Notes in Computer Science*, Springer, Cham, Vol. 9906, pp.425–442 [online] https://doi.org/10.1007/978-3-319-46475-6_27.