Multi-feature Fusion Network Acts on Facial Expression Recognition

Jingyu Li[1](✉), Weiyue Cheng[2], Jiahao Geng[1], and Kezheng Lin[1]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
920948105@qq.com

[2] Heilongjiang College of Business and Technology, Harbin, China

Abstract. In order to solve the problem of single-channel convolutional neural network feature loss in the process of facial expression recognition, a facial expression recognition algorithm based on multi-feature fusion network is proposed. The algorithm uses the dual-channel convolutional neural network model DCNN-FER (Dual-channel Convolutional Neural Network Model for Facial Expression Recognition). The pre-processed face image is input to channel one to obtain global features, and the face image that has been processed by Local Binary Patterns (LBP) is input to channel two to obtain local texture features. At the same time, it is used in part of the convolutional layer. The Convolutional Block Attention Module (CBAM) enhances the network's focus on the useful information of the image and suppresses useless features. Finally, new features are formed by weighted fusion and sent to the softmax layer for classification. This algorithm not only considers the extraction of overall facial features, but also enriches local texture features. Compared with other methods on the FER2013 and CK + facial expression data sets, the method in this paper shows good robustness.

Keywords: Facial expression recognition · CNN · Feature fusion · LBP

1 Introduction

Facial expression is an important way of communication between people. It is the significant information to understand the emotional state of a specific target. Even the smallest change on a person's face may be a different emotion signal. Ekman et al. divided expressions into six basic forms for the first time [1]. Facial expression recognition has gradually become a research hotspot in the field of computer vision. It has shown a wide range of application prospects in communication engineering, medical and health, safe driving [2], and social sentiment analysis.

Traditional feature extraction algorithms mainly include PCA [3] (Principal Component Analysis), SIFT (Scale-Invariant Feature Transformation), LBP [4, 5] (Local Binary Patterns), Gabor wavelet transform [6, 7] as well as the HOG(Histogram Of Gradient) [8], the classification algorithms mainly include SVM (Support Vector Machine),

S. Shi et al. (Eds.): 6GN 2021, LNICST 439, pp. 370–382, 2022.
https://doi.org/10.1007/978-3-031-04245-4_33

k-Nearest Neighbor [9] and so on. However, traditional facial expression recognition methods are susceptible to image noise and human interference factors, resulting in poor recognition accuracy.

Deep learning methods have shined in the field of image recognition, and DNN (Deep Neural Networks) have been applied to facial expression recognition and achieved good results [10]. CNN (Convolutional Neural Network) [11] is a machine learning model under deep supervised learning, and CNN convolutional layer uses multiple filters to extract image features, weight sharing compresses the number of parameters, and uses backpropagation to optimize parameters, which has achieved good results in facial expression recognition [12, 13]. Literature [14] added a network attention model to CNN, and automatically determined the area of interest by adjusting the channel and spatial weight, focusing on useful information, and suppressing useless information.

Studies have found that when a single-channel neural network is used to extract facial features, it may lead to insufficient focus on facial features and loss of part of effective information. ZENG et al. embedded manual facial features into a CNN and proposed a new Loss algorithm to obtain a higher recognition rate [15]. Wang et al. used CNN and SIFT to extract facial features respectively, and used SVM to realize expression classification after fusion of the two, and achieved good results [16]. Literature [17] combines the depth features extracted from CNN with the SIFT features of the image to form new features for expression classification. Rikhtegar et al. used genetic algorithm combined with CNN to extract image features, and finally used SVM to replace the last layer of CNN to complete the task of facial expression classification [18]. Gao Jingwen et al. used a three-channel convolutional neural network to focus on the face, eye, and mouth regions of the human face, and then extracted features [19]. Finally, the fusion technology based on the decision-making layer was used to fuse the features of the three channels for expression discrimination. Obtain the overall optimal recognition rate. Therefore, in the process of facial expression recognition, different algorithms or different levels of feature fusion can obtain more complex features, which to a certain extent become a significant means to solve feature loss, so this paper proposes an algorithm based on dual-channel convolutional neural network, The method of fusing global features and local texture features is used to obtain complex features.

2 Related Work

2.1 Convolutional Neural Network

CNN as a network structure of weight sharing, has stronger ability to extract image features, mainly composed of convolutional layer and pooling layer. The main structure in the convolutional layer is the filter, also known as the convolution kernel. The size of the filter is called the receptive field. The receptive field determines the size of the sliding window. The sliding window is used to extract the complex features of the image, and the activation function is used to process the complex features and output the features. The calculation process is shown in Formula (1).

$$X_j^L = f(\sum_{i \in M_j} X_i^{L-1} \cdot K_{ij}^L + b_j^L) \tag{1}$$

among them, L is the current layer; ${X_j}^L$ represents the j-th feature area of the current layer; ${X_i}^{L-1}$ represents the i-th feature area of the previous layer; K represents the convolution kernel of the two regions.

2.2 Convolutional Block Attention Module

CBAM can improve the neural network's focus on useful information. In the two dimensions of feature space and channel, it can enhance valuable features, suppress useless features, and greatly improve the feature extraction ability of the network. As shown in Fig. 1, the feature map F passes through the 1-dimensional channel attention module M_c and the 2-dimensional spatial attention module M_s to obtain the final output feature map.

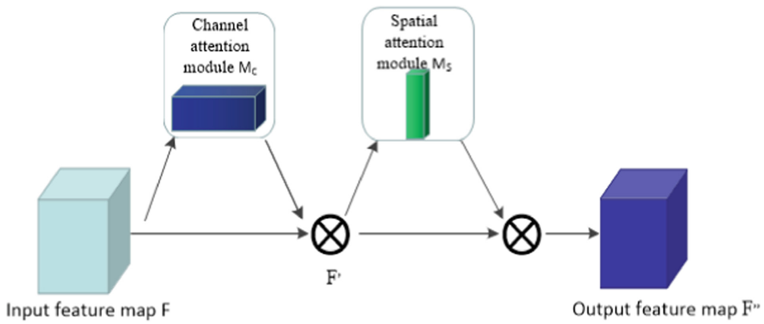


Fig. 1. CBAM structure diagram

In order to aggregate the feature information of the feature map on each channel, the average pooling and maximum pooling operations are used to generate two different channel context descriptors, then the two descriptors Don't send it into a multi-layer perceptron network (MLP) with a hidden layer, and finally combine the output feature weight parameters by element-by-element summation.

The spatial attention module first applies the average pooling and maximum pooling operations to the feature map to generate two two-dimensional feature maps, which respectively represent the average fusion feature and the maximum fusion feature on each channel. Then the attention weights in these two dimensions are sent to the convolutional layer to finally get the attention weights in the spatial dimension.

3 DCNN-FER Method

3.1 Method Structure

The DCNN-FER model has two feature extraction channels, including the HE-CNN channel for extracting global features and the LBP-CNN channel for extracting local features. In the HE-CNN channel, the original image is sent to the improved ResNet18 network to extract global features after the histogram equalization process. In the LBP-CNN channel, the original image is processed by LBP to obtain the LBP feature map, and then the LBP feature is sent Extract local features into a specific CNN, and then combine the global features and local features after dimensionality reduction. The fused new features are sent to the softmax layer to complete the classification task. The DCNN-FER algorithm flow diagram is shown in Fig. 2.

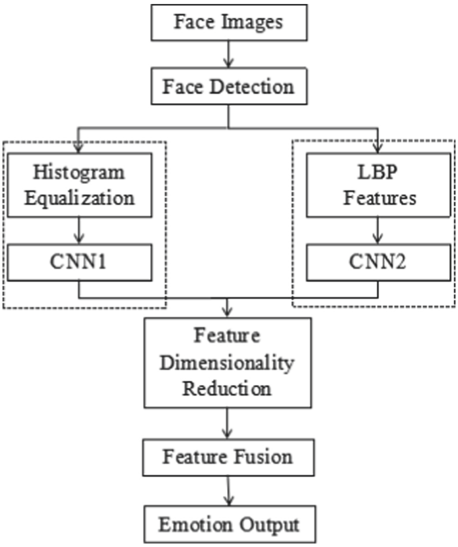


Fig. 2. Schematic diagram of DCNN-FER algorithm flow

The input layer of the HE-CNN channel is a grayscale image with pixels of 64px $\times$ 64px. In order to enhance the contrast of the image, the original image is histogram equalized. The gray levels of the original images of the FER2013 and CK + data sets are distributed in a relatively narrow interval. In order to better obtain the global features of the image, the channel uses the deep convolutional neural network ResNet18 for feature extraction. The residual unit is added to the ResNet network through the short-circuit mechanism to form residual learning, which solves the degradation problem of the deep network. As shown in Fig. 3(a), the original residual block consists of two convolutional layers and two BN layers, and each module has a fast input and output connection. In this paper, the original residual block structure is improved. As shown in Fig. 3(b), the two BN layers in the original structure are removed, and ReLU is added after the second convolutional layer to improve the residual The non-linear expression ability of the unit avoids the destruction of image space information by the BN layer. In order to enhance the performance of the residual network and reduce the burden of network training, a cross-layer integration strategy is proposed. As shown in Fig. 3(c), the output feature map of each residual unit is combined through shortcut connections. The network structure changes from global residuals to local residuals, and at the same time avoids over-fitting in model training and the disappearance of gradients in back propagation.

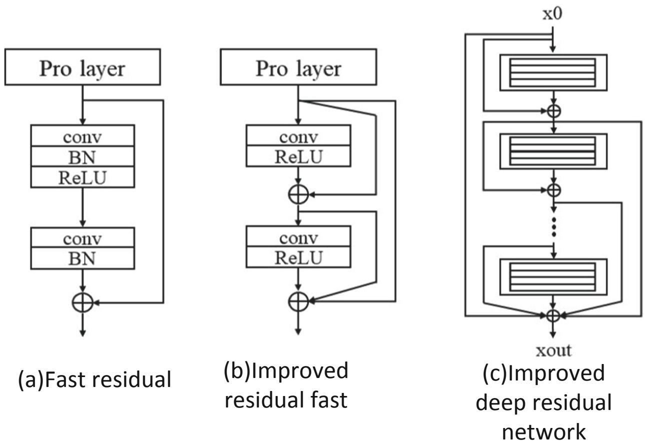


Fig. 3. Improved residual structure

The output of the network can be defined as:

$$X_{out} = \sum_{l=0}^{N} X_l = X_0 + X_1 + \cdots + X_N \tag{2}$$

among them, X_0 represents the gradual feature of the low-resolution image extracted by the feature extraction network. X_l is the output of the l-th residual unit. There are a total of N residual units, The parameters of each layer of the HE-CNN channel are shown in Table 1.

Table 1. Parameters of each layer of the HE-CNN channel network

Convolutional layer	Output size	Convolution Kernel
Conv1	$32 \times 32 \times 64$	7×7, 64, stride 2
Conv2_x	$16 \times 16 \times 64$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Conv3_x	$8 \times 8 \times 128$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
Conv4_x	$4 \times 4 \times 256$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
Conv5_x	$2 \times 2 \times 512$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
Fc1	1000	–
Fc2	6	–

According to the requirements of facial expression classification tasks, the ResNet18 network is finally connected to a fully connected layer with 6 nodes, and Softmax classification is adopted. The network attention mechanism is added to the first and last layers of the improved ResNet18 network to focus on effective features, suppress useless features, and effectively enhance the feature extraction capabilities of the network.

In the LBP-CNN channel, the original image of the data set must first be processed by the LBP algorithm. LBP is less sensitive to image gray changes and can better extract the image The local texture feature eliminates the influence of noise such as lighting. The LBP operator compares the neighboring pixels of the center pixel with the center pixel. The pixel larger than the center point is 1, and the pixel smaller than the center point is 0, so the 8-bit binary code is The LBP value of the center pixel is used to reflect the texture information of the area. The LBP operation is shown in Formula (3) and Formula (4).

$$LBP(g_c) = \sum_{i=0}^{p-1} s(g_i - g_c)2^i \tag{3}$$

$$s(x) = \begin{cases} 1, x > 0 \\ 0, else \end{cases} \tag{4}$$

among them, $s(x)$ represents the sign function, g_c represents the pixel value of the central pixel, g_i represents the pixel value of the surrounding adjacent pixels, p represents the number of adjacent pixels.

The generated LBP feature map is sent to the convolutional neural network. The LBP-CNN channel network structure is shown in Fig. 4. After the LBP feature map passes through the convolution operation, it passes through CBAM processing is sent to the pooling layer. The first layer of convolutional layer uses a filter with 32 channels and a size of 5 × 5; the second layer uses a filter with 64 channels and a size of 3 × 3; the third layer uses 128 channels and a size of 3 × 3; The fourth layer uses a filter with a channel of 256 and a size of 3 × 3; the feature map is sent to a fully connected layer with 500 nodes through the flattening layer, and finally it is sent to the feature fusion network. The parameters of each layer of the LBP-CNN channel network are shown in Table 2.

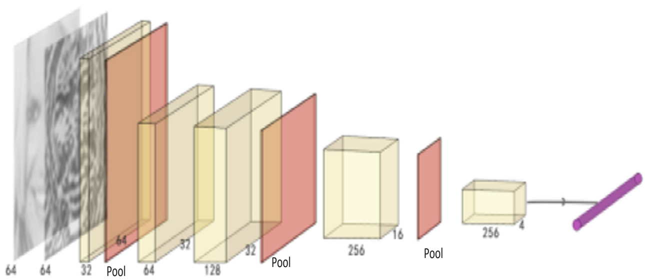


Fig. 4. LBP-CNN feature extraction process diagram

Table 2. Parameters of each layer of the HE-CNN channel network

Network layer	Input size	Convolution kernel	Output size
Conv1	$64 \times 64 \times 1$	$5 \times 5 \times 32$	$64 \times 64 \times 32$
Pool1	$64 \times 64 \times 32$	$2 \times 2 \times 32$	$32 \times 32 \times 32$
Conv2_1	$32 \times 32 \times 32$	$3 \times 3 \times 64$	$32 \times 32 \times 64$
Conv2_2	$32 \times 32 \times 64$	$3 \times 3 \times 128$	$32 \times 32 \times 128$
Pool2	$32 \times 32 \times 128$	$2 \times 2 \times 128$	$16 \times 16 \times 128$
Conv3	$16 \times 16 \times 128$	$3 \times 3 \times 256$	$8 \times 8 \times 256$
Pool3	$8 \times 8 \times 256$	$2 \times 2 \times 256$	$4 \times 4 \times 256$
Fc1	–	–	500
Fc2	–	–	6

3.2 Feature Fusion

After the original picture extracts the global feature and local texture feature feature through the dual-channel network, it is sent to the feature fusion layer, and finally the expression category is output. In order to adjust the proportion of the two channel features, the feature fusion layer adopts a weighted fusion method and sets the weight coefficient k to adjust the weight of the two channel features.

$$F_Z = k \cdot f_H + (1 - k) \cdot f_L \tag{5}$$

in the Formula (5), F_z represents the feature after fusion, f_{H} represents the features extracted by the HE-CNN channel, f_{L} represents the features extracted by the LBP-CNN channel. When the value of k is set to 0, it means that there is only the LBP-CNN channel, and when the value of k is set to 1, it means that there is only the HE-CNN channel. Choosing an appropriate value of k is the key to feature fusion. After a lot of experiments, select the most appropriate value of k through accuracy.

As shown in Fig. 5, the recognition rate of the model changes with the change of the value of k. The recognition rate reaches the highest when k is 0.6, and the recognition effect is the best.

3.3 Description of Algorithm Steps

According to the above-mentioned DCNN-FER model structure and feature fusion method, the training process of the model and the specific steps of expression recognition are described in Algorithm 1.

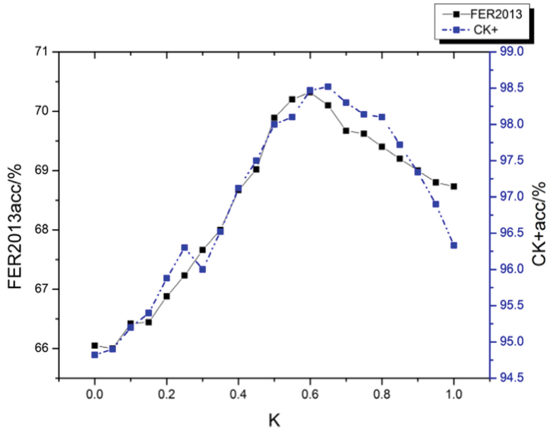


Fig. 5. Accuracy rate of accuracy rate under different k values

---

Algorithm 1: DCNN-FER algorithm

Input: The face samples in the data set are unified into 48×48 grayscale images.

Output: Emoticon category.

(1) Batch input image data {x1, x2,...} for histogram equalization.

(2) Use $LBP(g_c) = \sum_{i=0}^{p-1} s(g_i - g_c)2^i$ to find the LBP feature map of the image.

(3) The image processed by the histogram equalization is put into the ResNet18 convolutional neural network, and the 6-dimensional feature vector $f_H = \{s_1, s_2, ..., s_6\}$ is obtained after convolution operation $X_j^L = f(\sum_{i \in M_j} X_i^{L-1} \cdot K_{ij}^L + b_j^L)$, pooling operation and feature dimensionality reduction.

(4) The LBP feature is subjected to a five-layer convolution operation $X_j^L = f(\sum_{i \in M_j} X_i^{L-1} \cdot K_{ij}^L + b_j^L)$ and a three-layer maximum pooling operation, and a feature vector $f_L = \{t_1, t_2, ..., t_6\}$ with a dimension of 6 is obtained after feature dimensionality reduction.

(5) f_H and f_L are weighted and fused using formula $F_Z = k \cdot f_H + (1-k) \cdot f_L$, and the new feature after fusion is F_Z.Send it to the softmax layer for classification.

---

4 Experiments

4.1 Facial Expression Data Set

This article conducts experimental verification on two public data sets of FER2013 and CK+.

The FER2013 data set [20] has 35886 face grayscale images with a fixed size of 48px × 48px. Among them, there are 28708 images in the training set, and 3589 images in the test set and verification set. They are divided into 6 categories, corresponding to labels 0–6. For the convenience of research, this article will remove the original data set. Neutral pictures and experiment with other 6 types of pictures.

The CK + data set [21] is the extended Cohn-Kanada facial expression database (The Extend Cohn-Kanade Data, CK+), which contains a diverse group of Asian and European races covering male and female genders. It was collected and published by the University of Pittsburgh in 2010. The data set collected 593 frontal face images of 123 experimenters, including 8 expressions. This article only studies the first 6 basic expressions. 80% of the pictures in the data set are classified as the training set, and the remaining pictures are the test set.

Figure 6 is a partial sample of each expression in the FER2013 and CK + datasets.

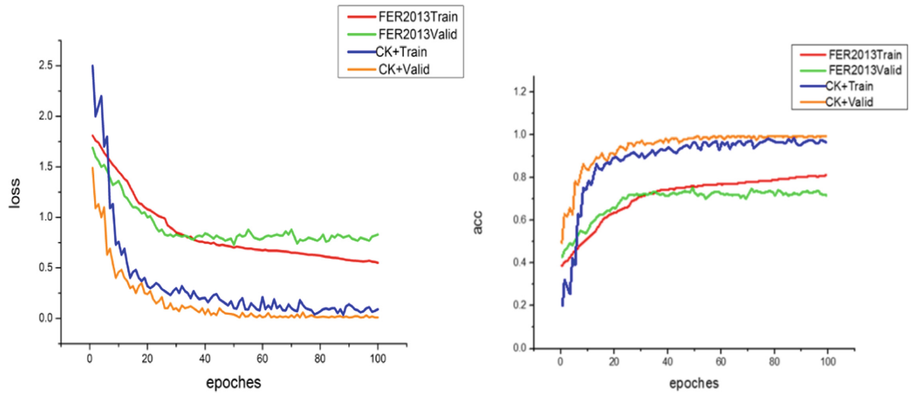
(a) Part of the image of the FER2013 (b) Part of the image of the CK+

Fig. 6. Examples of FER2013 and CK + dataset images

4.2 Algorithm Validity Verification

Use the DCNN-FER model to train on the FER2013 and CK + datasets and verify its effectiveness. In order to expand the data set on the basis of the original sample, the original image is randomly zoomed, translated, flipped, and rotated to increase the model training data. The number of data sets after data enhancement becomes 9 times the original, which greatly increases the amount of model training and reduces the risk of overfitting.

The experiment uses Tensorflow as the basic experimental framework, Python3.7 as the programming language. The learning rate is in the form of dynamic changes, and the initial value is set to le−4, with the increase of epochs Gradually decay to (le−4)/epoches. In order to prevent over-fitting, Dorpout regularization is used in the fully connected layer, some neurons are randomly discarded, the initial value of the Dropout regularization parameter is set to 0.5, and L2 regularization is added, and the regularization coefficient is 0.01. Figure 10 shows the accuracy and loss curves of the DCNN-FER model on the two data sets.

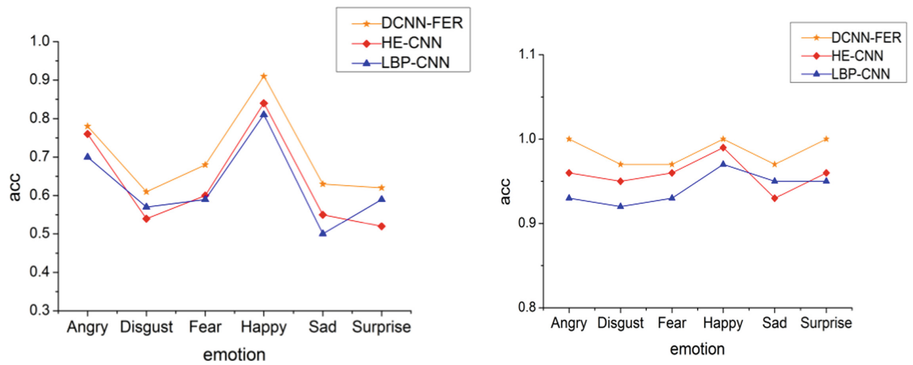


(a) Loss function curve (b) Accuracy curve

Fig. 7. Training process curve diagram

It can be seen from Fig. 7 that as the iteration period (epoches) increases, the loss of the model on the two datasets gradually decreases, and the accuracy (acc) gradually increases. The accuracy rate of the FER2013 verification set finally stabilized at about 70.49%, and the accuracy rate of the CK + verification set finally stabilized at about 98.31%, achieving the expected results.

Figure 8 is a comparison of the accuracy of DCNN-FER and HE-CNN and LBP-CNN with only a single channel on different expressions. It can be seen from the figure that the DCNN-FER model is on two data sets compared to a single channel neural

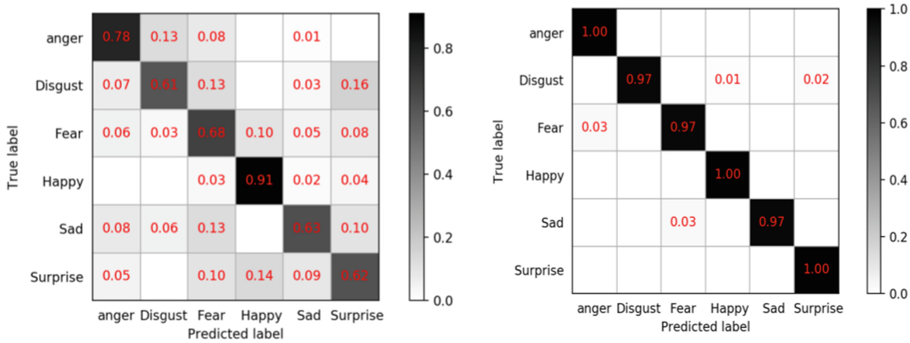


(a)FER2013 recognition rate (b) CK+ recognition rate

Fig. 8. Comparison of recognition rates under different models

network. Both achieve the desired effect. Although the recognition rate of individual models is low, the overall recognition rate is better than the other two models.

The confusion matrix of the DCNN-FER model on the FER2013 and CK + data sets is shown in Fig. 9.



(a) Confusion matrix on FER2013 (b) Confusion matrix on CK+

Fig. 9. Confusion matrix of DCNN-FER under different data sets

From Fig. 9(a), it can be seen that the model has a higher recognition rate for angry and happy on the FER2013 data set, reaching 78% and 91%, respectively. The recognition results for disgust, fear, sadness and surprise are not good, because the number of these expressions in the data set is relatively small, and the similarity between the expressions is high, which is more likely to cause classification errors. In addition, the FER2013 data set itself There are also label errors, and some images are still occluded, resulting in low classification accuracy. From the matrix in Fig. 9(b), the model has a recognition rate of 100% for happy, surprised and angry expressions on the CK + dataset. Due to the similarity of expressions, the recognition of disgust and fear is The rate is relatively low, but it also reached 97%.

Table 3 and Table 4 are the accuracy comparisons of different methods on the FER2013 and CK + data sets, respectively. The data shows that the accuracy rate of DCNN-FER on FER2013 reached 70.47%, which is 4 to 5% points higher than other models. It also showed a good recognition effect on CK+, and the accuracy rate was increased to 98.48%, which was 2 to 3% points higher than other models. Compared with DNN and FER-Net [25], although the method in this paper costs more time Many, but our method shows great advantages in recognition accuracy. Therefore, the feasibility of the algorithm in facial expression recognition is verified.

Table 3. Comparison of recognition rates of different methods on the FER2013 (%)

Method	Accuracy
Ref. [12]	65.03
DNN [13]	66.4
Ref. [15]	61.86
FER-Net [22]	69.72
DCNN-FER	70.49

Table 4. Comparison of the recognition rate of different methods on the CK + (%)

Method	Accuracy
DNN [13]	93.2
Ref. [15]	97.35
Ref. [17]	94.82
Ref. [23]	95.79
DCNN-FER	98.31

5 Conclusions

Aiming at the problem of single-channel neural network feature loss in the process of facial expression recognition, a facial expression recognition algorithm based on feature fusion network is proposed. Through the feature fusion method, this algorithm overcomes the problem of partial effective feature loss of single-channel neural network on the one hand, that is, fully extracts global features, and effectively retains local effective features. The two complement each other, complement each other, and improve expression Accuracy of recognition. This paper also uses the network attention model to automatically focus the feature area of interest in the network, suppress useless information. The next step of research will continue to improve the network model to include more detailed local features, make full use of the features of the facial features and global features to further improve the accuracy.

References

1. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
2. Lutao, G.: Research on Driver State Analysis Method Based on Facial Expression. University of Electronic Science and Technology of China (2019)
3. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemom. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
4. Jia, Q., Gao, X., Guo, H., et al.: Multi-layer sparse representation for weighted LBP-patches based facial expression recognition. Sensors **15**(3), 6719–6739 (2015)

5. Shan, C., Gong, S., Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
6. Zhou, J., Zhang, S., Mei, H., et al.: A method of facial expression recognition based on Gabor and NMF. Patt. Recognit. Image Anal. **26**(1), 119–124 (2016)
7. Gu, W., Xiang, C., Venkatesh, Y.V., et al.: Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Patt. Recogn. **45**(1), 80–91 (2012)
8. Wang, X., Jin, C., Liu, W., et al.: Feature fusion of HOG and WLD for facial expression recognition. In: IEEE/SICE International Symposium on System Integration. IEEE (2014)
9. Liu, P., Zhou, J.T., Tsang, I.W.H., et al.: Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In: European Conference on Computer Vision, pp. 151–166. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_11
10. Fulan, Q., Jianhong, L., Shu, Z., et al.: Rating recommendation based on deep hybrid model. J. Nanjing Univ. Aeronaut. Astronaut. **51**(5), 592–598 (2019)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)
12. Lopes, A.T., de Aguiar, E., De Souza, A.F., et al.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**(12), 610–628 (2017)
13. Pons, G., Masip, D.: Supervised committee of convolutional neural networks in automated facial expression analysis. IEEE Trans. Affect. Comput. **9**(3), 343–350 (2017)
14. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
15. Zeng, G., Zhou, J., Jia, X., et al.: Hand-Crafted feature guided deep learning for facial expression recognition. In: Proceedings of 2018 13th IEEE International Conference on Automatic Face&Gesture Recognition (FG 2018), Xi'an, pp. 423–430. IEEE (2018)
16. Sun, X., Lv, M.: Facial expression recognition based on a hybrid model combining deep and shallow features. Cogn. Comput. **11**(4), 587–597 (2019)
17. Wang, F., Lv, J., Ying, G., et al.: Facial expression recognition from image based on hybrid features understanding. J. Vis. Commun. Image Represent. **59**(1), 84–88 (2019)
18. Rikhtegar, A., Pooyan, M., Manzuri-Shalmani, M.T.: Genetic algorithm-optimised structure of convolutional neural network for face recognition applications. IET Comput. Vis. **10**(6), 559–566 (2016)
19. Jingwen, G.A.O., Yongxiang, C.A.I.: TP-FER: Three-channel facial expression recognition method based on optimized convolutional neural network. Appl. Res. Comput. **38**(07), 2213–2219 (2021)
20. Goodfellow, I.J., Erhan, D., Carrier, P., et al.: Challenges in representation learning: a report on three machine learning contests. In: Proceedings of International Conference on Neural Information Processing, pp. 117–124. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42051-1_16
21. Lucey, P., Cohn, J.F., Kanade, T.J., et al.: The extended Cohn-Kanade dataset (CK+) a complete dataset for action unit and emotion-specified expression. In: Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 94–101. IEEE (2010)
22. Ma, H., Celik, T.: FER-Net: facial expression recognition using densely connected convolutional network. Electron. Lett. **55**(4), 184–186 (2019)
23. Zeng, N., Zhang, H., Song, B., et al.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643–649 (2018)