Paper:

A Multi-Scale Feature Fusion Network for Facial Expression Recognition

Jian Wang, Kaoru Hirota, Yaping Dai, Zhiyang Jia

School of Automation, Beijing Institute of Technology
E-mail: wang_jian@bit.edu.cn

A multi-scale feature fusion network is proposed for facial expression recognition (FER). The three-branch architecture is employed to extract abundant facial features from various depths and scales. Our main contributions are twofold. First, the feature extraction module is designed to effectively extract multi-scale image information and enhance network ability of feature representation. Second, the feature fusion module is presented to adaptively evaluate the importance of different cross-branch features and fully emphasize the role of distinguishable features for expression classification. In addition, the basic feature extraction unit cascaded in feature extraction module is proposed with different receptive field filters to extract multi-scale features. Experimental results on two public datasets, FER2013 and CK+, demonstrate that our proposed method outperforms the previous methods with the accuracies of 72.67 % and 96.00 % respectively.

Keywords: Facial expression recognition, Multi-scale feature, Feature fusion, Deep neutral network

1. Introduction

Facial expression is one of the most direct signals to convey emotions and transmit messages. Therefore, Facial expression recognition (FER) is of great significance in human-computer interaction [1], medical treatment [2], distance education [3] and other fields related to facial expression, and has gradually become an important research direction.

Previous traditional FER methods generally extract hand-crafted features for expression classification, which can be separated into two categories: 1) geometric-based methods, such as active appearance model (AAM) [4]; 2) appearance-based methods, such as local binary pattern (LBP) [5] and gabor wavelet representation [6]. These hand-crated features achieve satisfactory classification performance on the high-resolution datasets collected in laboratory-controlled conditions. In recent years, deep convolutional neural network (CNN) has been successfully applied to facial expression recognition and other computer vision tasks. It can automatically learn facial features from various images for classification. Moreover, the fusion features extracted by multi-branch CNN can effectively remedy that the features extracted from single branch network might perform unsatisfactory for classification. For example, Hamester et al. [7] employ a two-channel network for feature extraction, and features extracted from these two channels are fused and then used for classification. Jung et al. [8] independently train two CNNs to extract geometric and appearance features, and fuse these features through a fully connected layer by joint fine-tuning. Shao et al. [9] and Li et al. [10] respectively construct two sub-networks to extract global and texture information from raw images and LBP features, which is helpful to distinguish expressions with subtle difference. More recently, Liu et al. [11] present a hybrid feature extraction network to extract the expressional features from static face images and explore the potentials of facial landmarks for expression recognition. Shi et al. [12] propose a multiple branch cross-connected convolutional neural network to fuse the features of each branches and increase the recognition performance. However, each branch of these methods usually extracts facial features with a single receptive field in the same size input images, and element addition or channel concatenation is used to fuse features from different branches. These manners might bring out insufficient feature extraction from facial images and fail to emphasize the importance of different branch features in the fusion process.

In this paper, a multi-scale feature fusion network is proposed for recognizing facial expressions. A convolution operation is firstly employed to increase the number of feature channels. Then the feature maps are upsampled and downsampled by bicubic interpolation to capture facial features at different scales, because input feature maps with various sizes show different feature information. Three branches with diverse depths are employed to extract multi-scale features, and various number of feature extraction modules are cascaded in three branches. The downsampled, original and upsampled feature maps are respectively fed into the top, middle and bottom branches to sufficiently extract facial shallow and deep features. Specifically, a feature fusion module is presented to automatically evaluate the importance of features from different branches, and a basic feature extraction unit with different receptive field filters is introduced to enhance the ability of feature representation. In addition, residual and recursive learning strategies are employed to reduce the amount of parameters and speed up the network fitting

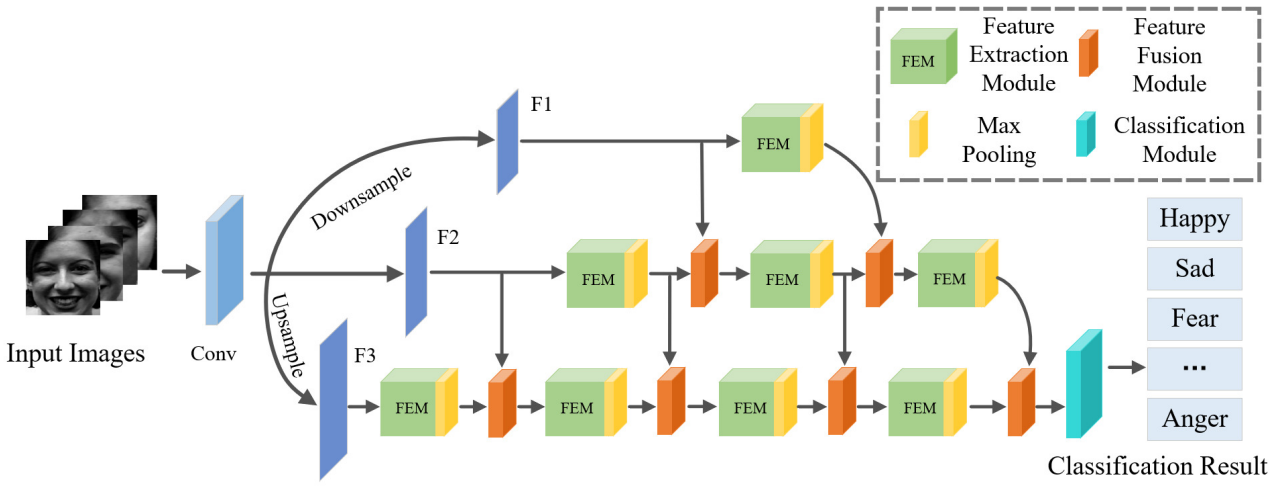


Fig. 1. : The framework of multi-scale feature fusion network

during training.

In summary, the main contributions are as follows: 1) We propose a novel multi-scale feature fusion network, in which the three-branch structure contributes to enhancing the ability of feature representation for facial expression recognition. 2) A feature fusion module is introduced to adaptively evaluate the importance of features from different depth branches. Meanwhile, a feature extraction module is presented to extract abundant multi-scale features for classification. 3) Extensive experimental evaluations on two public datasets show that the proposed method outperforms the previous methods in facial expression recognition task.

The remainder of this paper is organized as follows. The framework of multi-scale feature fusion network and the details of the proposed method are described in Section 2. The experimental process and results on different datasets are given in Section 3. We provide conclusions and future work in Section 4.

2. Multi-scale Feature Fusion CNN for FER

In this section, we first present the overall framework of the proposed multi-scale feature fusion network with three branches. Then the facial feature extraction module is introduced with cascaded basic feature extraction unit. Finally, the feature fusion module is described in detail.

2.1. Overall Network Framework

Our proposed multi-scale feature fusion network, as shown in **Fig. 1**, employs three-branch structure for feature extraction. When given the facial image input, the shallow feature is extracted by a 3×3 convolution layer. Then the shallow feature maps are upsampled and downsampled by bicubic interpolation to obtain facial features at different scales. The three branches are used to extract various depth features from the original, downsampled and upsampled feature maps. Different numbers of feature extraction modules and feature fusion modules are cascaded in each branch to ensure that each branch has different depth of network structure. The feature fusion module adaptively evaluates the importance of features from different branches by endowing self-training weights. The facial features with same dimension from different branches are fused and then transported to the lower branch for deep feature extraction. The extracted deep features of bottom branch are fused with the top and middle features for final classification. At last, two fully connected layers with Softmax function are used in classification module. In addition, the batch normalization and ReLu function are employed to speed up the convergence of the network. In order to reduce the difficulty of training and speed up the network fitting, we adopt the recursive learning strategy to decrease parameters when cascading feature extraction modules in different branches. Therefore, the feature maps with same size from three branches are fed into the feature extraction module with same function.

2.2. Feature Extraction Module

It can be seen from **Fig. 2**, m basic feature extraction units (BFEUs) are cascaded and linked together by residual learning in the multi-scale feature module. The residual learning strategy contributes to alleviate the phenomenon of vanishing and exploding gradients when cascading basic feature extraction units. The process can be formulated as:

$$\begin{aligned} F_{MF}^{k} &= G_{mf}^{k}\left(F_{MF}^{k-1}\right) \\ &= R_{bf}^{m}\left(R_{bf}^{m-1}\left(\cdots R_{bf}^{1}\left(F_{MF}^{k-1}\right)\right)\right)+F_{MF}^{k-1}, \end{aligned} \tag{1}$$

where $G_{mf}^{k}(\cdot)$ and $R_{bf}^{r}(\cdot)$ represent the k-th $(k=1,2,...,t)$ feature extraction module function and the r-th $(r=1,2,...,m)$ BFEU function, respectively. The F_{MF}^{k-1} and F_{MF}^{k} denote the input and output of the k-th feature extraction module. The element-wise addition operation is

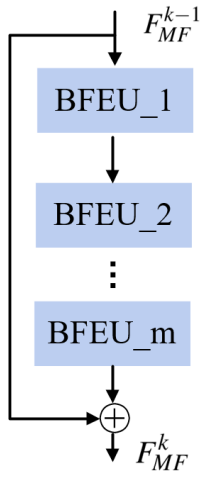


Fig. 2. : The architecture of feature extraction module

adopted to fuse the features of two consecutive stages and provide more abundant information. The output F_{MF}^{k} is applied to 2×2 max pool layer for feature dimensionality reduction to improve the classification efficiency.

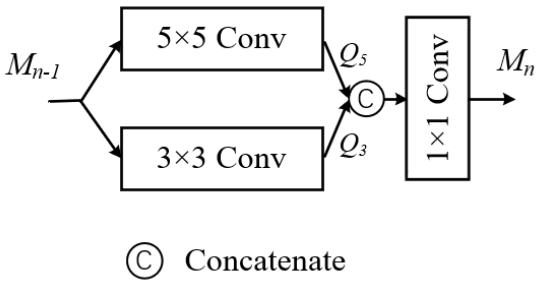


Fig. 3. : The architecture of basic feature extraction unit.

The most of FER methods use 3×3 convolutions alone as minimum feature extraction unit, which easily leads to single receptive field of network. As shown in **Fig. 3**, the basic feature extraction unit (BFEU) is employed with 3×3 and 5×5 filters to enlarge receptive field for extracting multi-scale features. Each BFEU also includes a 1×1 filter for fusing the extracted features and ensuring the same numbers of input and output channels. The process of BFEU is described as:

$$Q_5 = \sigma\left(C_{5\times5}\left(M_{n-1}\right)\right), \tag{2}$$

$$Q_3 = \sigma\left(C_{3\times3}\left(M_{n-1}\right)\right), \tag{3}$$

$$M_n = C_{1\times1}\left(\left[Q_5, Q_3\right]\right), \tag{4}$$

where $\sigma(\cdot)$ is the ReLu function. $C_{5\times5}(\cdot)$, $C_{3\times3}(\cdot)$ and $C_{1\times1}(\cdot)$ are 5×5, 3×3 and 1×1 convolution layers, respectively. M_{n-1} and M_n $(n = 1, 2, ..., m)$ are the input and output of the n-th BFEU, respectively. The features Q_5 and Q_3 are concatenated and then sent to 1×1 convolutional layer for maintaining the same number channels of input and output.

2.3. Feature Fusion Module

The element-wise addition and concatenation operations are often employed to fuse facial features with same dimension in the adjacent network modules. Because of equally integrating features from different network modules, these manners ignore that the features from various perspective and depth modules have different contributions to expressing classification task.

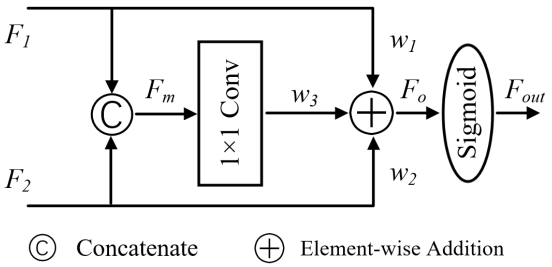


Fig. 4. : The architecture of feature fusion module.

A feature fusion strategy is introduced in this module to adaptively evaluate the importance of different branch features by assigning self-training weights. As shown in **Fig. 4**, the two inputs F_1 and F_2 from two branches are concatenated and then fed into 1×1 convolution layer to obtain the middle fusion features F_m. The 1×1 convolution layer simply fuses the concatenated features and ensures the coincident number of input and output channels. The features F_1, F_2 and F_m are multiplied by self-training weights w_1, w_2 and w_3, respectively, and then added up. The output F_o is fed to Sigmoid() function to obtain the final output F_{out}. The main process of feature fusion module is formulated as follows:

$$\begin{aligned} F_{out} &= G_{fm}\left(F_1, F_2\right) \\ &= Sigmoid\left(\sum_{i=1}^{2} w_i F_i + w_3 C_{1\times1}\left(\left[F_1, F_2\right]\right)\right), \end{aligned} \tag{5}$$

where $G_{fm}(\cdot)$ represents the function of feature fusion module, and $C_{1\times1}(\cdot)$ denotes the 1×1 convolution layer. As a result, the self-training weights allow the critical branch to provide information with larger proportion, so as to improve the ability of feature selection for network.

3. Experiments on Expression Recognition

In this section, we first introduce two facial expression datasets and the experimental parameters. Then the classification results with the corresponding analysis are presented. Finally, we compare proposed method with several previous methods on two datasets.

3.1. Datasets Description

We conduct on two public facial expression recognition dataset, the FER2013 dataset and the Extended Cohn-Kanade (CK+) dataset, to evaluate the proposed network performance.

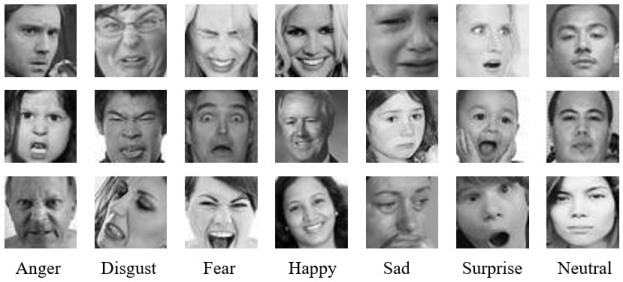


Fig. 5. : The samples from FER2013 dataset

FER2013 dataset: The FER2013 is a large-scale and unconstrained dataset, which contains a total of 35,887 face images, including 28,709 training images, 3,589 validation images, and 3,589 test images. All images are registered and resized to 48 $\times$ 48 pixels grayscale face images after adjusting the cropped region. These images are divided into the following seven categories: 0-'anger', 1-'disgust', 2-'fear', 3-'happy', 4-'sad', 5-'surprised', and 6-'neutral'. The samples from FER2013 dataset are shown in **Fig. 5.**

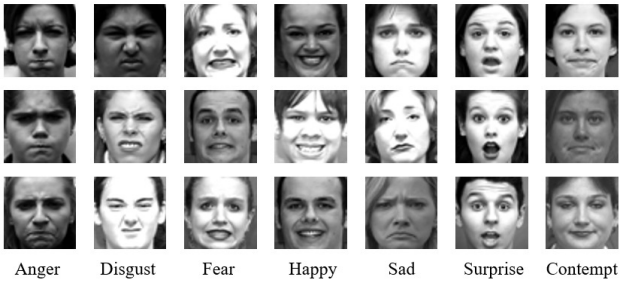


Fig. 6. : The samples from CK+ dataset

CK+ dataset: The CK+ dataset is the most extensively used laboratory-controlled dataset for evaluating FER systems. CK+ dataset contains 593 image sequences from 123 subjects. The image sequences vary in duration from 10 to 60 frames, and show a shift from the neutral facial expression to the peak expression. In total, 327 image sequences from 118 subjects are labeled with seven basic expression labels: anger, contempt, disgust, fear, happiness, sad, and surprise. For static-based FER methods, the most common data selection method is to extract the last three frames with peak formation of each sequence. The samples from the CK+ dataset are shown in **Fig. 6.**

3.2. Experimental Parameters

For image preprocessing, a large amount of redundant information is contained in CK+ images with resolution of 640 $\times$ 490. The Dlib toolkit is used for face detection to remove the noisy data. The last three frames with peak expressions of face sequences are selected for experiments, and the face region of each frame are clipped and rescaled. The resolution of 48 $\times$ 48 is selected as the final image input for proposed network.

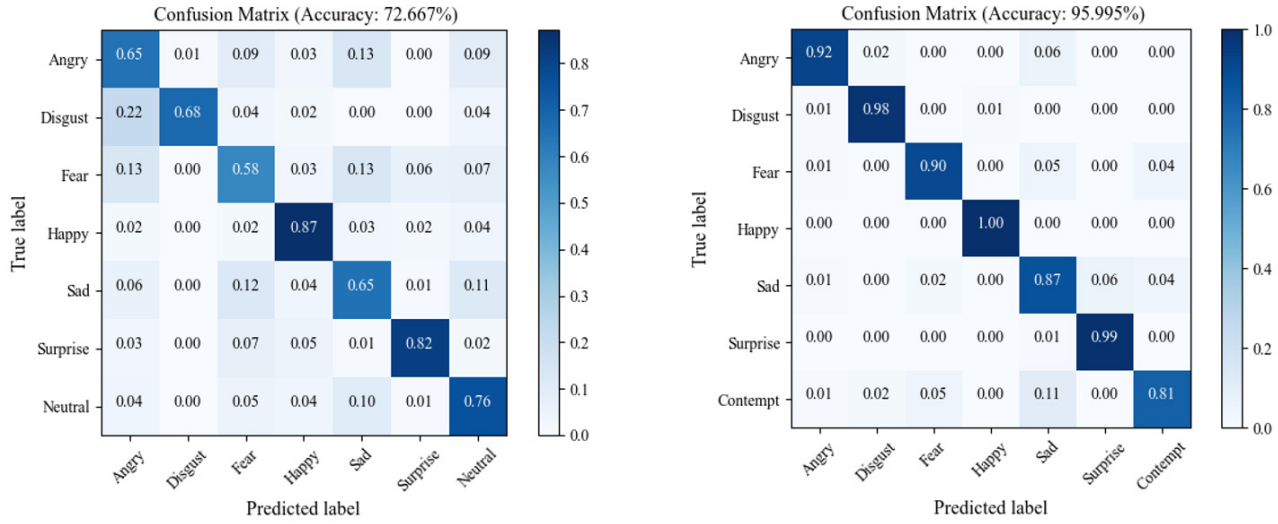
For the training, the 16 image patches are randomly chose as network inputs. The stochastic gradient descent (SGD) algorithm is employed for weights update with the initial learning rate $lr = 1\mathrm{e}-4$, momentum $m = 0.9$ and weight decay $wd = 5\mathrm{e}-4$. For FER2013 dataset, the total number of epochs is set to 200. The learning rate begins to decrease after 50 epochs, and is decreased by multiplying with 0.95 every 20 epochs. The 60 epochs are set for the CK+ dataset. The learning rate begins to decrease after 20 epochs, and is decreased by multiplying with 0.98 every 5 epochs. In addition, dropout is adopted in fully connected layer to relieve overfitting, and the dropout ratio is set to 0.5. For data argumentation, the training images are randomly flipped horizontally and cut to 48 $\times$ 48 after 4 zero-padding for images edges. The PyTorch framework is employed to implement our method on Linux system with Nvidia GTX 1080Ti.

3.3. Results on Two Datasets

The performance of multi-scale feature fusion network is evaluated on the unconstrained FER2013 dataset and constrained CK+ dataset for FER task. According to the performance in extensive ablation experiments, the number of basic feature extraction units is set to 3, which achieves a fine balance between accuracy performance and parameter quantity. In addition, the confusion matrix is employed to intuitively compare the actual category with the prediction category.

Results on Fer2013 dataset: The confusion matrix of proposed network on FER2013 dataset is shown in **Fig. 7(a)**, and the average recognition accuracy achieves 72.67%. Because of clearly distinguished facial muscle movements and shape deformations, the 'happy' and 'surprised' categories are easier to be classified with high recognition accuracies of 87% and 82%, respectively. However, it can be observed that the facial expressions of 'fear', 'angry' and 'sad' are undesirably misclassified due to the similar facial deformations. The similarity among these expressions is difficult to distinguish for the proposed network.

Results on CK+ dataset: The ten-fold cross validation is implemented on the CK+ dataset to obtain reliable experimental results. The facial images are equally divided into ten groups, of which nine groups are used for training, and the remaining one group is employed for testing. The final experiment result is the average of the ten various testing accuracies. The confusion matrix of proposed network on the CK+ database is shown in **Fig. 7(b)**, and the average recognition accuracy achieves 96.00%. The 'disgust', 'happy' and 'surprised' categories are recognized with almost 100% average accuracies. Similar to the results on FER2013, the 'fear', 'angry' and 'sad' are



(a) Confusion matrix on FER2013 dataset

(b) Confusion matrix on CK+ dataset

Fig. 7. : The confusion matrix of proposed network on FER2013 and CK+ datasets

easily misclassified. In addition, the recognition accuracy of 'contempt' is relatively hard to distinguish, which may be caused by the lack of training samples in prototypical facial expressions.

3.4. Comparison Experiments with Other Methods

The multi-scale feature fusion network is compared with several previous methods including single branch methods and multiple branch methods on FER2013 dataset and CK+ dataset.

Table 1. : Performance comparison on FER2013 dataset.

Method	accuracy (%)
Guo et al. 2016 [13]	71.44
Munasinghe et al. 2017 [14]	71.10
Xie et al. 2019 [15]	66.20
(light-CNN) Shao et al. 2019 [9]	68.00
(dual-branch CNN) Shao et al. 2019 [9]	54.64
(pre-trained CNN) Shao et al. 2019 [9]	71.14
Sun et al. 2020 [16]	72.50
Liu et al. 2020 [17]	72.56
Minaee et al. 2021 [18]	70.02
Shi et al. 2021 [12]	71.52
Proposed method	**72.67**

Comparison on FER2013 dataset: The accuracies of the proposed method and several existing expression recognition methods are listed in **Table 1**. The proposed method obtains an average accuracy of 72.67% on the unconstrained expression dataset, which is higher than the listed methods. Besides, Liu et al. [17] achieve the second highest recognition accuracy of 72.56% among the listed methods by employing the two-channel network to extract features from original and LBP images. The similar strategy of combining original and LBP images is adopted in dual-branch CNN [9], but the recognition accuracy is only 54.64% due to the simple feature extraction module and feature fusion strategy. Attention mechanism has been successful applied into some listed FER methods [15–18], which achieves a satisfactory performance by focusing on salient areas for FER task. Compared with the listed methods, it is evident that our proposed method achieves outstanding performance by extracting multi-scale features that combine various perspective fea-

Table 2. : Performance comparison on CK+ dataset.

Method	Accuracy (%)
Lopes et al. 2017 [19]	92.73
Cai et al. 2018 [20]	94.35
Turan et al. 2018 [21]	95.90
Zhang et al. 2019 [22]	92.35
Jain et al. 2019 [23]	93.24
Xie et al. 2019 [15]	95.88
(light-CNN) Shao et al. 2019 [9]	92.86
(dual-branch CNN) Shao et al. 2019 [9]	85.71
(pre-trained CNN) Shao et al. 2019 [9]	95.29
Sun et al. 2020 [16]	87.20
Shahid et al. 2020 [24]	94.90
Liu et al. 2020 [17]	94.24
Liu et al. 2021 [11]	95.15
Proposed method	**96.00**

tures from three different depth branches. The feature fusion module contributes to selecting more discriminative features for classification.

Comparison on CK+ dataset: The performance comparison on CK+ dataset is illustrated in **Table 2**. The proposed method is superior to all the other listed methods, and achieves a recognition rate of 96.00%. In addition, Turan et al. [21] achieve the second highest recognition accuracy of 95.90% with higher image resolution and more sub-regions. Similar to the results on FER2013 dataset, attention mechanism in network is helpful to achieve outstanding recognition accuracies. The pre-trained CNN [9] with accuracy of 95.29% demonstrates that pre-trained deep network structures may solve the lack of training samples and over-fitting problems. However, the dual-branch CNN with feature concatenation fusion operation still performs imperfect for expression recognition task. Through contrastive analysis, it is clear that our proposed method extract and select more discriminative features from original facial images for expression classification.

4. Conclusions and Future work

A multi-scale feature fusion network is proposed for recognizing facial expressions. The three-channel structure is employed to extract multi-scale features from different depth receptive fields. Specifically, the feature extraction module is introduced to enhance the ability of feature representation, and the feature fusion module is presented to automatically evaluate the importance of features from different branches. The experiments are carried on two public datasets, FER2013 and CK+, and achieve average accuracies of 72.67% and 96.00% respectively. The confusion matrices of the proposed network on two datasets demonstrate that the 'happy' and 'surprised' categories are easier to be classified with high recognition accuracies, while the facial expressions of 'fear', 'angry' and 'sad' are relatively hard to distinguish. The proposed method is compared with several single-branch and multiple-branch methods, which verifies the superiority of our model.

Our future work will be focused on improving the low accuracies of facial expression categories due to lack of training samples. Furthermore, we will make effort to study the feature fusion module with more multi-scale inputs to explore its potentials of future application.

Acknowledgements

This work is supported by the National Talents Foundation under Grant No.WQ20141100198.

References:

[1] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," Inf. Sci. (Ny), vol. 428, pp. 49–61, 2018.

[2] M. S. Hossain, "Patient state recognition system for healthcare using speech and facial expressions," J. Med. Syst., vol. 40, no. 12, p. 272, 2016.

[3] J. Khalfallah and J. B. H. Slama, "Facial expression recognition for intelligent tutoring systems in remote laboratories platform," Procedia Comput. Sci., vol. 73, pp. 274–281, 2015.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, 2001.

[5] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, 2007.

[6] S. Zhang, L. Li, and Z. Zhao, "Facial expression recognition based on Gabor wavelets and sparse representation," in 2012 IEEE 11th International Conference on Signal Processing, 2012.

[7] D Hamester, P. Barros , and S. Wermter . "Face expression recognition with a 2-channel Convolutional Neural Network." International Joint Conference on Neural Networks IEEE, 2015.

[8] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

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

[10] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," Neurocomputing, vol. 411, pp. 340–350, 2020.

[11] C. Liu, K. Hirota, J. Ma, Z. Jia, and Y. Dai, "Facial expression recognition using hybrid features of pixel and geometry," IEEE Access, vol. 9, pp. 18876–18889, 2021.

[12] C. Shi, C. Tan, and L. Wang, "A facial expression recognition method based on a multibranch cross-connection convolutional neural network," IEEE Access, vol. 9, pp. 39255–39274, 2021.

[13] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep Neural Networks with Relativity Learning for facial expression recognition," in 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2016.

[14] S. Munasinghe, C. Fookes, and S. Sridharan, "Deep features-based expression-invariant tied factor analysis for emotion recognition," in 2017 IEEE International Joint Conference on Biometrics (IJCB), 2017.

[15] 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, 2019.

[16] 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.

[17] C. Liu, K. Hirota, B. Wang, Y. Dai, and Z. Jia, "Two-channel feature extraction convolutional neural network for facial expression recognition," J. adv. comput. intell. intell. inform., vol. 24, no. 6, pp. 792–801, 2020.

[18] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," Sensors (Basel), vol. 21, no. 9, 2021.

[19] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," Pattern Recognit., vol. 61, pp. 610–628, 2017.

[20] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. OReilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018.

[21] C. Turan and K.-M. Lam, "Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study," J. Vis. Commun. Image Represent., vol. 55, pp. 331–341, 2018.

[22] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal recurrent neural network for emotion recognition," IEEE Trans. Cybern., vol. 49, no. 3, pp. 839–847, 2019.

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

[24] A. Raza Shahid, S. Khan, and H. Yan, "Contour and region harmonic features for sub-local facial expression recognition," J. Vis. Commun. Image Represent., vol. 73, p. 102949, 2020.