Multiscale Convolutional Attention-based Residual Network Expression Recognition

*Fei Wang, Haijun Zhang**

School of Computer Science and Technology, Xinjiang Normal University, China
1049961711@qq.com, zhjlp@163.com

Abstract

Expression recognition has wide application in the fields of distance education and clinical medicine. In response to the problems of insufficient feature extraction ability of expression recognition models in current research, and the deeper the depth of the model, the more serious the loss of useful information, a residual network model with multi-scale convolutional attention is proposed. This model mainly takes the residual network as the main body, adds normalization layer and channel attention mechanism, so as to extract useful image information at multiple scales, and incorporates the Inception module and channel attention module into the residual network to enhance the feature extraction ability of the model and to prevent the loss of more useful information due to too deep network, and to improve the generalization performance of the model. From results of lots of experiments we can see that the recognition accuracy of the model in FER+ and CK+ datasets reaches 87.80% and 99.32% respectively, with better recognition performance and robustness.

Keywords: Expression recognition, Feature extraction, Multiscale convolution, Residual network, Channel attention mechanism Introduction

1 Introduction

Facial expressions are the most important component of the human communication system and are an important way to better reflect human mental activity [1]. Psychologist Mehrabian pointed out that when communicating face-to-face, human emotions are mainly composed of expressions, voice, and speech (speech rate, intonation), of which facial expressions account for 55%. It is clear that facial expressions are one of the most natural and common ways to communicate emotional states, so expression recognition has become a research hot topic which is of great academic and has been widely applied in fields such as distance education, safe driving scenarios etc.

Expression recognition mainly consists of several steps: face information acquisiting and preprocessing, face feature extracting and face expression identificating. Since the expression recognition task is easily influenced by conditions including lighting changes, occlusion, and non-frontal pose, feature extraction has a great effect on the overall recognition model's accuracy. Traditional feature extraction is inefficient and incomplete leading to low final accuracy, for example, Zhang et al. [2] proposed using Gabor filtering to select useful features of images, but it is difficult to find suitable parameter variables to extract image features when the image feature dimension is too large. W. L. Chao et al. [3] proposed es-LBP based on local binary patterns, which can better capture important local information of faces, but the accuracy decreases when there is more background noise in the image. With the openness of lots of tagged expression datasets and the development of deep learning, more and more scholars are relying on deep learning for expression recognition research. Tang [4] and Kahou [5] designed deep learning models with deeper network layers for facial feature extraction and won the FER2013 and EMotiw2013 expression recognition challenges respectively. Some classical neural networks have also been used to extract facial expression features, for example, VggNet, GoogLeNet, etc., and the results achieved are mostly better than traditional recognition algorithms. With the continuous research on deep learning, many deepen network layers so as to make better the model precision, which would lose some useful feature information and produce overfitting phenomenon for the dataset with small amount of data.

Facial expressions are expressed through muscle changes in the eyes, face, and mouth, and they are the most direct expressions reflecting human inner emotions. Facial expression recognition have been applied in many practical production environments, such as classroom teaching evaluation, clinical medicine, and fatigue driving, etc. Existing expression recognition methods are broadly classified into traditional learning-based methods and deep learning-based methods.

For traditional expression recognition methods, M. Goyani et al. [6] used AdaBoost cascade target detector to detect the largest geometric component of expression information and then finally used OneVsAll model to classify facial expressoins. Gupta et al. [7] used OpenCV's Haar filter to detect faces and then used SVM algorithm for classification on CK+ dataset to get 93.7% of accuracy. Pham et al. [8] trained a multilayer perceptron to retrieve similar images and to analyze if the current facial expression identification result is reasonable through the final classification output, such traditional shallow recognition algorithms suffer from low accuracy and insufficient generalization performance. As for deep learning, Zhang et al. [9] introduced a means on the basis of multi-scale global images and local images,

*Corresponding Author: Haijun Zhang; E-mail: zhjlp@163.com

DOI: 10.53106/160792642023092405015

which improved the performance of the model and achieved good results. Jinghui Chu et al. [10] proposed an attention mechanism-based face expression recognition model and introduced residual units to enhance the feature extraction capability of the attention mechanism. Y. Li et al. [11] proposed a model that adds attention to the neural network. It can perceive the face occluded area and can also focus on the non occluded area with the most discrimination ability, which can increase the precision in occluded and non occluded situations.

At present, a lot of research results have been achieved in expression recognition methods based on deep learning, but there exist still some problems to be further improved in the relevant research. For example, the low recognition rate is not enough to meet the requirements of practical application scenarios, and the model is too deep leading to the loss of useful features, etc. On one hand, the expression recognition technology should improve the expression recognition rate and increase its practical use in value. On the other hand, the effective measures should be taken to reduce the loss of effective features during model training, such as using residual structures.

In summary, this paper introduces a multi-scale convolutional attention-based expression recognition method for residual networks. Firstly, the network takes the residual network [13] as the main body, and improves the ability of feature extraction of the model by promoting the Inception module and adding BatchNormalization (BN) layer and channel attention module in parallel to it to extract important feature information from multiple scales and to prevent the loss of useful information during the network training process. Finally, comparisons with several other existing research methods shows that the method has better robustness and generalization.

2 Basic Principles And Models

2.1 Imported Attention Mechanism Module

Facial expressions are mainly caused by changes in specific facial muscles. If we do not focus on these feature channels, we could not get useful feature about the face. This paper uses Channel attention SENet network [12] to make the model concerned about the channels of important features. The SENet network framework is shown in Figure 1:

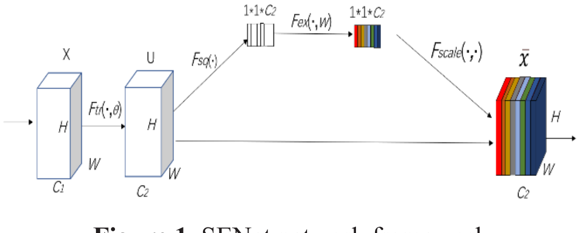


Figure 1. SENet network framework

As a result of convolution, the characteristic channel U = $\{u_1,u_2,\ldots,u_c\}$ is difficult to capture the relationship between channels, so SENet firstly carries out Squeeze operation on it, and uses global average pooling to compress and extract it in H × W dimension. Global feature, assuming that the compressed feature is z_c, the corresponding calculation formula is as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j),\ z\in R^c. \quad \textbf{(1)}$$

Since the Squeeze operation is operated on each channel, resulting in a lack of dependencies between channels, the Excitation operation is also needed to allow nonlinear interactions between channels. The specific approach is to let each channel feature go through two fully connected layers to achieve dimensionality reduction and dimensionality enhancement to limit the model complexity, and then the two fully connected layers go through ReLU and sigmoid activation function respectively, and finally get a weight s, which is calculated as the following equation:

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)). \quad \textbf{(2)}$$

Where σ and δ are the sigmoid and ReLU function respectively, the parameter $w_1 \in R^{\frac{c}{r}\times c}, w_2 \in R^{\frac{c}{r}\times c}$ is used to reduce the model complexity and to enhance the generalization performance. r is a super parameter, which can control the amount of computation of the SE module, and this paper sets it to 8.

We need to multiply the weight s_c of channels with the original feature u_c of channels to get the final feature, so that the model can enhance the feature recognition ability of each channel. The corresponding formula is as follows:

$$\tilde{x}c = Fscale(u_c, s_c) = s_c \cdot u_c. \quad \textbf{(3)}$$

During the study of facial expression recognition, the SENet module is introduced to make the model concern about the important features of different channels of the image and ignore the irrelevant features, so that the model has a stronger feature extraction ability.

2.2 Improved Multi-scale Convolution Module

In order to improve the effect of network training, such as AlexNet, VGGNet, etc., most of the traditional classical neural networks use the way of increasing the depth of the network to achieve the purpose, but the deeper the network is, the more disadvantages the model has, such as the loss of effective features, over-fitting, etc. When designing the network architecture, in order to have better expression in feature extraction, this paper improves on the basis of Inception module, and adds BN layer and SENet module in parallel after convolution layers of different scales. Those let the BN layer normalize the data to speed up convergence and let SENet focus on more channel information to speed up model training and obtain richer features. The module framework is shown in Figure 2:

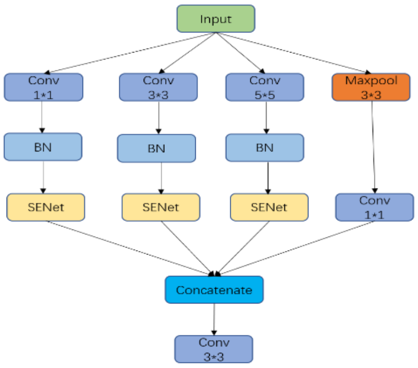


Figure 2. Multi-scale convolution module diagram

2.3 Residual Network Structure Based on Multi-scale Convolution Attention

Since the residual network is able to effectively resolve the problems of gradient dispersion and disappearance of useful information in deep network, this paper mainly applies residual network as the base network and adds the improved multiscale convolution and channel attention mechanism. The residual structure block is shown in Figure 3. After the input original data is processed by the multi-scale convolution block, the extracted features are passed through the convolution layer and BN layer to speed up the model training and to control the gradient explosion. Through the channel attention module, the network is guided to pay attention to the feature areas with prominent facial expression changes, which can extract deeper feature information in order to prevent the loss of information during the whole feature extracting process. The skip connection operation [13] is introduced to fuse the original data with the extracted information to supplement the information, so the whole residual block can extract the deep features of the image. It can also reduce the problem of useful feature loss.

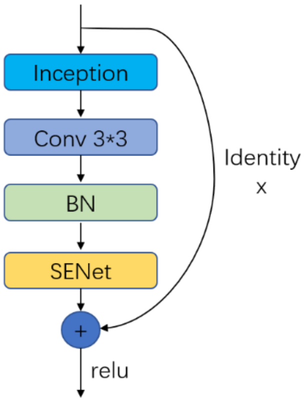


Figure 3. Resnet module structure

The multi-scale convolution attention residual network framework designed in this paper is shown in Figure 4. The residual unit which combines multi-scale convolution and channel attention mechanism is adopted, which not only increases the feature extraction ability of the network, but also prevents the network from being too deep and losing useful information during the training process. First of all, the network extracts shallow features with a convolution pooling layer, and then extracts global channel features through a channel attention mechanism module, and then puts the extracted channel feature graph into three series residual network modules. There is a maximum pooling layer in the middle to ensure the translation invariance of the image. A global average pooling is added behind the last residual block to replace the full connection layer. It can be used to reduce overfitting and to enhance the model effects. At the end of the model is a fully connected layer and softmax layer which are used to output the categories of expressions.

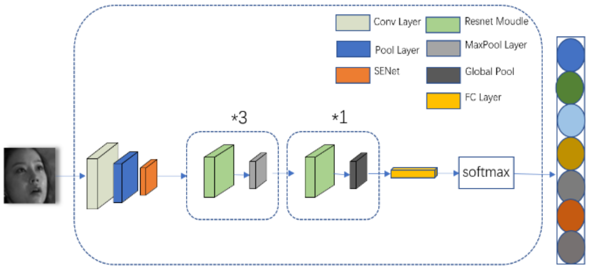


Figure 4. Multi-scale convolutional attention residual network structure

2.4 Loss Function Based on CoSine Similarity

In deep learning, model optimization is a key link. For different tasks, selecting or designing appropriate loss function plays an indispensable role in measuring the quality of the model.

If the loss function value is smaller, the performance of the corresponding model will be better, and vice versa. It can directly reflect the gap between the predicted value of the model and the real value, and is closely related to the subsequent model optimization, such as gradient decline.

In the field of visual tasks, the studied object often has a high feature dimension, and the cosine similarity still maintains the property of “1 at the same time, 0 when orthogonal, and −1 on the contrary” in the case of high dimension, which is a relatively stable loss function. In this paper, cosine similarity is selected as the loss function, shown as bellow:

$$L = -\frac{\sum_{i=1}^{n} y_{true}^{(i)} \cdot y_{pred}^{(i)}}{\sqrt{\sum_{i=1}^{n} (y_{true}^{(i)})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_{pred}^{(i)})^2}}. \quad \textbf{(4)}$$

y_{true} and y_{pred} represent real expression categories and predicted expression categories respectively, and cosine similarity is applied to measure the similarity between real and predicted categories in high-dimensional space. The cosine value is between −1 and 1. The closer the value is to −1 means the direction of the two category vectors is closer and the similarity is higher.

3 Experiment

3.1 Dataset and Preprocessing

The open datasets selected in this paper include FER+ & CK+. FER+ is extended on the basis of the dataset FER2013, which is originally labeled by crowdsourcing. The tagging accuracy of the FER+ is higher than that of Fer2013. There are 8 kinds of data tags (neutral, happy, surprise, sad, angry, disgust, fear, contempt), of which there are 28709 training images and 3589 verification images and 3589 test images. The CK+ dataset consists of 593 dynamic expression image sequences of 123 persons. It is a relatively perfect and standard public dataset with 7 kinds of tags (anger, contempt, disgust, fear, joy, sadness, surprise) at present. The example is shown in Figure 5:

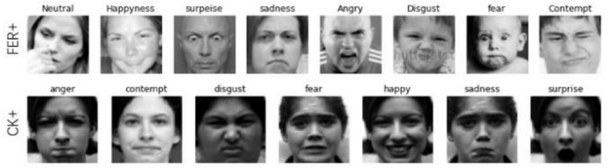


Figure 5. Sample diagram of FER+ and CK+

In view of the small number of samples in the expression dataset, in order to make the model more generalized, the expression dataset is enhanced. ImageDataGenerator technology is deployed to normalize the standard deviation of all pixels of the image, and then the training data are offset, rotated and scaled. The specific operation is to randomly offset the original data by 5% in the horizontal and vertical directions, and randomly rotate in the range of [0°, 10°] , and scale with a scale of 0.3.

3.2 Experimental Environment

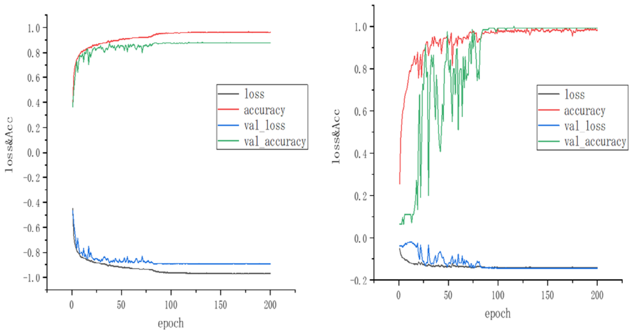
In this paper, the programming language is python3.6, and the deep learning framework is tensorflow2.4.0, and the hardware environment of the experiment is the 64-bit Microsoft Windows with 10 core CPU i7-10700 and the graphics card is NVIDIA GeForce RTX 3060 with memory 12GB.

3.3 Experimental Settings and Results

In the experiments, the batch sizes of FER+ dataset and CK+ dataset were set as 128 and 7 respectively, and the epoch was 200. The image sizes input into the model were all 48*48. The optimizer selected in the experiments was Adam, with initial learning rate(LR) 0.001 and attenuation factor 0.1. The experiments adjusted the learning rate according to the training epoch and developed a new strategy of LR, shown in Table 1:

Table 1. Learning rates of optimizers within different epochs

Epoch	0---80	81---120	121---160	161---180	>180
Learning rate	0.001	0.0001	1e-05	1e-06	5e-07



(a) Under FER+ dataset (b) Under CK+ dataset

Figure 6. Training under different datasets

Figure 6 shows that before 80 epochs, the model has not reached the fitting state, and the learning rate is 0.001 at this time. After 80 epochs, the learning rate of the model is 0.0001, and the model precision is increased. The fitting effect and the loss value are gradually stabilized and the learning rate gradually becomes smaller, so that the fitting effect of the whole model remains stable. After that, the learning rate gradually decreased, and the fitting effect of the whole model remained stable and there was no over fitting phenomenon.

3.4 Experimental Comparison

3.4.1 The Affects of Different LR

The LR is a super parameter that guides the loss function gradient to adjust the network weight in the model. The update rate of the model weight is related to the LR. the lower the LR is, the slower the corresponding loss function changes, although it will not miss the local minimum. However, the convergence speed of the model will also slow down, so setting an appropriate LR has a great influence on the performance of the training model. On the basis of the network designed in this paper, we make experiments between the LR with value 0.01, 0.001, 0.0001, 0.00001 respectively and our LR strategy in this paper.

Table 2. Accuracy rates on each dataset at different learning rates

Learning rate	Fer+'acc	CK+'acc
0.01	0.8453	0.9116
0.001	0.8676	0.9048
0.0001	0.8577	0.9524
0.00001	0.7082	0.8980
Our learning rate strategy	**0.8780**	**0.9932**

From Table 2, we can find that LR has certain effects on the model performance. The LR can neither be too large nor too small. It can be observed that when the LR is 0.0001, the training effect of the model is relatively good. In this paper, we combine the effect of higher learning rate to accelerate model convergence and a smaller learning rate to prevent missing the minimal value. It is found that the learning rate strategy used in this paper is relatively good, and the accuracy of the model is improved in different datasets compared with other learning rate trained models.

3.4.2 The Affects of Different Improvement Strategies

To enhance the feature extraction capability, the Inception module is improved in this paper by adding BN layer and SENet module. Therefore, to explore the effectiveness of this improved strategy, we conducted several groups of experiments. The ablation experiments are shown in Table 3, where × means not added and √ means added.

Table 3. The influence of different improvement strategies for multi-scale convolution module on the performance of the model

Number	SENet	BN	FER+’acc	CK+’acc
1	×	×	0.8688	0.9728
2	√	×	0.8716	0.9864
3	×	√	0.8750	0.9796
4	√	√	**0.8780**	**0.9932**

Through the comparison of Table 3, it can be found that the different improvement strategies added in this paper improve the model as a whole, among which for the dataset FER+ with large dataset, the effect of adding BN layer is better than that of adding SENet module, and for the CK+ dataset with small dataset, the effect of adding SENet module is better than adding BN layer. Therefore, this paper comprehensively considers that the two strategies are added to the Inception module to enhance the model performance, and the accuracy of 87.8% and 99.32% on FER+ and CK+ datasets are obtained respectively, which shows that the improved strategy of this paper has a certain robustness.

3.5 Analysis of Experimental Results

To further analyze the accuracy of the model proposed of this paper for all kinds of facial expression recognition, we draw a confusion matrix for the results of recognition on FER+ and CK+ datasets respectively, as shown in Figure 7 and Figure 8:

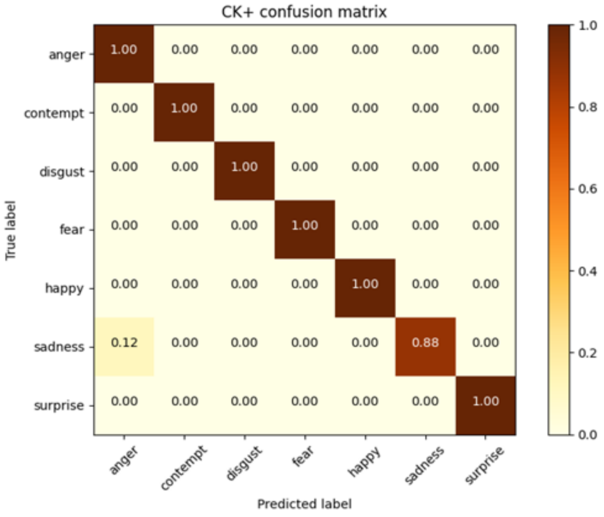


Figure 7. Confusion matrix plotted on CK+ dataset

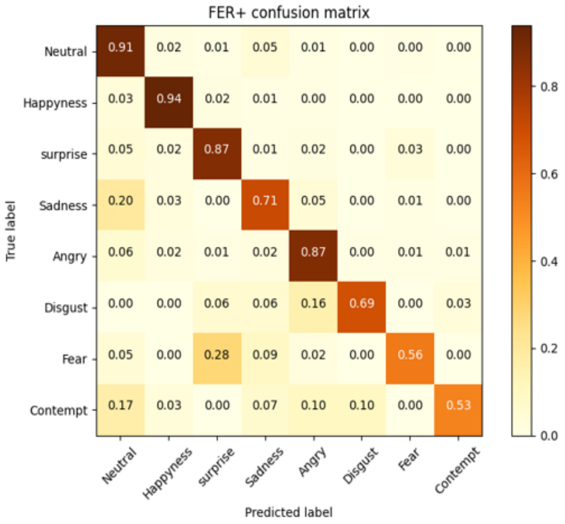


Figure 8. Confusion matrix plotted on FER+ dataset

After observing Figure 7 and Figure 8, it can be found that for the confusion matrix on the CK+ dataset, except for sad expression, the recognition accuracy of the model is very high for other categories, which is because the CK+ dataset is collected under controllable laboratory conditions, and the model is relatively accurate. There are fewer confounding factors, while there is a great similarity between sad expression and angry expression, which leads to a slightly lower accuracy of facial expression recognition in those categories. For the confusion matrix on FER+ dataset, the recognition accuracy of natural, happyness and surprise expressions is relatively high, while that of disgust, fear and contempt is relatively low. The reason is that the facial muscle changes of these expressions are very similar, and the recognition degree of features is not high. In addition, the number of these expressions is small, and there is a category imbalance, resulting in relatively low recognition rates.

For further exploring the effectiveness of the network designed in this paper, the recognition effect of the proposed method is compared with that of the methods proposed in recent years on datasets FER+ and CK+, shown in Table 4:

Table 4. Comparison with typical methods in accuracy

Datasets	Methods	Accuracy
FER+	TFE-jL [14]	0.8429
	LER [15]	0.8567
	SHCNN [16]	0.8654
	ESR-9 [17]	0.8715
	LCMA [18]	0.8740
	Our method	**0.8780**
CK+	Em-AlexNet [19]	0.9425
	CNN+DBN [20]	0.9573
	PACNN [11]	0.9703
	SCAN [21]	0.9731
	PyConv-Attention Network [22]	0.9846
	Our methods	**0.9932**

For the FER+ dataset, there are 5 comparison models. The TFE model, proposed by Li et al. [14], used independent deep convolution neural network to learn emotion and identity features, and constructed deep learning tandem facial expression features through feature cascading. Zhao et al. [15] proposed a lightweight model LER for emotion recognition to deal with the delay of the model under natural conditions, with an accuracy of 85.67%. Miao et al. [16] proposed a shallow CNN architecture SHCNN, which made use of the advantages of training samples and achieved an accuracy of 86.54%. H. Siqueira et al. [17] reduced the residual generalization error on the dataset through integration and sharing representation based on convolution network, and reached the human level performance. Pengbo Yin et al. [18] proposed a lightweight network LCMA by decomposing convolution and embedding attention mechanism, and gained an accuracy of 87.4%. Compared with the model proposed by the above researchers, our model has improved the accuracy by 3.51%, 2.13%, 1.26%, 0.65% and 0.4% respectively.

For the CK+ dataset, Xu Yang et al. [19] improved AlexNet by adding multi-scale convolution to make the network more suitable for small-size facial expression images; Linlin Wang et al. [20] used fused local features and deep confidence model into expression recognition (CNN+DBN); Li et al. [11] proposed a model (PACNN) that integrates the attention mechanism into the convolutional neural network to resolve the facial expression recognition of occlusion perception; D. Gera et al. [21] used spatial channel attention network (SCAN) to obtain local and global attention of each spatial location of each channel. Junyu Mao et al. [22] introduced a method on the base of pyramid convolution network and attention to handle the problem of multi-scale extraction of facial expression recognition. Compared with the model proposed in the above study, our model has improved the accuracy by 5.07%, 3.59%, 2.29%, 2.01% and 0.86% respectively.

Although these algorithms have achieved good results in recent years, in view of the lack of the ability to extract important features of facial expression, this paper combines multi-scale convolution and attention to allocate more weight where the expression changes are prominent, and better experimental results are obtained. Through experiments, it can also be found that, whether on FER+ or CK+ dataset, the model of this paper is improved compared with other models, and has a higher precision, which shows the feasibility of the proposed model.

4 Conclusions

In this paper, we propose a multi-scale convolutional attention-based expression recognition model for residual networks. By improving the Inception module and fusing the channel attention mechanism to extract richer features, and then incorporating both the improved Inception module and channel attention into the residual module, more levels of expression features can be learned and the gradient disappearance and overfitting problems are reduced. In addition, the learning rate of this model is set according to the training epoch, which can effectively improve the model performance. Through multiple comparison experiments and comparison with some existing algorithms, it can be found that the accuracy rate is improved, which demonstrates the usefulness of the network designed in this paper. This provides very good research ideas and directions for the research of expression recognition. The next work is to optimize the model, increase the inter-class difference of expressions for the problems of image occlusion and small inter-class variation in expression data, to improve the accuracy and adapt it to more realistic environments.

Acknowledgements

This work is Sponsored by Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2022D01A226) and the Xinjiang Key Research and Development Program (2022B01007-1).

References

[1] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, *Pattern recognition*, Vol. 36, No. 1, pp. 259-275, January, 2003.

[2] Z.-Y. Zhang, X.-M. Mu, L. Gao, Recognizing facial expressions based on Gabor filter selection, *International Congress on Image and Signal Processing*, Shanghai, China, 2011, pp. 1544-1548.

[3] W.-L. Chao, J.-J. Ding, J.-Z. Liu, Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection, *Signal Processing*, Vol. 117, pp. 1-10, December, 2015.

[4] Y.-C. Tang, Deep learning using linear support vector machines, June, 2013. https://arxiv.org/abs/1306.0239.

[5] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, Z.-Z. Wu, Combining modality specific deep neural networks for emotion recognition in video, *Proceedings of the 15th ACM on International conference on multimodal interaction*, Sydney, Australia, 2013, pp. 543-550.

[6] M. Goyani, N. Patel, Multi-level haar wavelet based facial expression recognition using logistic regression, *Indian Journal of Science and Technology*, Vol. 10, No. 9, pp. 1-9, March, 2017.

[7] S. Gupta, Facial emotion recognition in real-time and static images, *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India, 2018, pp. 553-560.

[8] T. T. D. Pham, S. Kim, Y.-C. Lu, S. Jung, C. S. Won, Facial Action Units-Based Image Retrieval for Facial Expression Recognition, *IEEE Access*, Vol. 7, pp. 5200-5207, January, 2019.

[9] C. Zhang, P. Wang, K. Chen, J. Kämäräinen, Identity-aware convolutional neural networks for facial

expression recognition, *Journal of Systems Engineering and Electronics*, Vol. 28, No. 4, pp. 784-792, August, 2017.

[10] J.-H. Chu, W.-H Tang, S. Zhang, L. Wei, An Attention Model-Based Facial Expression Recognition Algorithm, *Laser & Optoelectronics Progress*, Vol. 57, No. 12, pp. 205-212, June, 2020.

[11] Y. Li, J.-B. Zeng, S.-G. Shan, X.-L. Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, *IEEE Transactions on Image Processing*, Vol. 28, No. 5, pp. 2439-2450, May, 2019.

[12] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132-7141.

[13] K.-M. He, X.-Y. Zhang, S.-Q. Ren, J. Sun, Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[14] M. Li, H. Xu, X.-C. Huang, Z.-M. Song, X.-L. Liu, X. Lin, Facial Expression Recognition with Identity and Emotion Joint Learning, *IEEE Transactions on Affective Computing*, Vol. 12, No. 2, pp. 544-550, April-June, 2021.

[15] G.-Z. Zhao, H.-T. Yang, M. Yu, Expression Recognition Method Based on a Lightweight Convolutional Neural Network, *IEEE Access*, Vol. 8, pp. 38528-38537, January, 2020.

[16] S. Miao, H.-Y. Xu, Z.-Q. Han, Y.-X. Zhu, Recognizing Facial Expressions Using a Shallow Convolutional Neural Network, *IEEE Access*, Vol. 7, pp. 78000-78011, June, 2019.

[17] H. Siqueira, S. Magg, S. Wermter, Efficient facial feature learning with wide ensemble-based convolutional neural networks, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 4, pp. 5800-5809, April, 2020.

[18] P.-B. Yin, W.-M. Pan, H.-J. Zhang, Lightweight Facial Expression Recognition Method Based on Convolutional Attention, *Laser & Optoelectronics Progress*, Vol. 58, No. 12, pp. 245-251, June, 2021.

[19] X. Yang, Z.-H. Shang, Facial Expression Recognition Based on Improved AlexNet, *Laser & Optoelectronics Progress*, Vol. 57, No. 14, pp. 235-242, July, 2020.

[20] L.-L. Wang, J.-H. Liu, X.-M. Fu, Facial Expression Recognition Based on Fusion of Local Features and Deep Belief Network, *Laser & Optoelectronics Progress*, Vol. 55, No. 1, pp. 204-212, September, 2018.

[21] D. Gera, S. Balasubramanian, Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition, *Pattern Recognition Letters*, Vol. 145, pp. 58-66, May, 2021.

[22] J.-Y. Mao, T.-N. He, Y. Guo, A.-B. Li, Expression Recognition Based on Global Attention and Pyramidal Convolution Network, *Computer Engineering and Applications*, Vol. 58, No. 23, pp. 214-220, December, 2022.

Biographies

Fei Wang received the B.S. degree in Computer Science and Technology from Xinjiang Normal University. He is a graduate student at Xinjiang Normal University, China. His current research interests include computer vision and image processing.

Haijun Zhang received the Ph.D. degree in Computer Science from University of Science and Technology of China in 2011. As a professor, He works at the School of Computer Science and Technology of Xinjiang Normal University. His research interests include computer vision and artificial intelligence.