Research on Facial Expression Recognition Method Based on Bilinear Convolutional Neural Network

Yong Zhang¹, Weikang Yi¹

¹School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China

*Corresponding author’s e-mail: yzhang@lut.cn

Abstract. Facial expressions express different emotions through different degrees of distortion of facial muscles, and the expression images spontaneously generated by people in real scenes are susceptible to interference from different illumination angles and postures. An extraction based on bilinear convolutional neural network is proposed. This is an end-to-end method for emotion classification based on second-order features of expression images. This method optimizes the network structure by adding a batch normalization layer, increasing the maximum pooling core, and replacing the fully connected layer with a global average pooling layer. Finally, the bilinear normalization method is used to convert one-dimensional features into two dimensional features complete the final classification. Through testing on the FER2013 dataset, this method obtains an accuracy of 73.2%, which is 0.5% higher than the current state-of-the-art method, and proves that extracting image second-order features is more conducive to expression classification than first-order features. In addition, compared with other mainstream methods, the proposed improved model has higher recognition accuracy.

1. Introduction

Facial expressions are one of the important ways for humans to express their inner emotions to the outside world. It is actually difficult for human to control the management of expressions. Facial expressions will change with changes in inner emotions. American psychologist Mehrabian. [1] has shown that facial expressions are the most effective way to identify the inner emotions of others. In the process of communication between people, the proportion of facial expressions conveying information is 55%, and the information contained in the speech part is only 7%, and the proportion of voice is only 38%. So far, facial expression recognition technology has been gradually applied in various fields, such as human-computer interaction, driving fatigue detection, clinical medicine, safety detection, emotional music. Also because of its wide application prospects, a large number of scholars began to study it.

Facial expressions are different from the limitations of language. Humans interact socially through emotional expressions, and expressions have become a universal language that transcends ethnic and cultural diversity. Therefore, automatic facial expression recognition has become an important research direction of human-computer interaction[2]. Zhang Xiang et al.[3] proposed a facial expression recognition method using SWA optimized cascade network with random weight averaging, and the recognition rate reached 74.478% on the FER2013 expression dataset. With the development of neural networks, deep learning has gradually shown good results in computer vision tasks. The application of deep learning methods to study expression recognition problems is mainly divided into two categories: expression recognition based on discrete frame images and continuous frame image [4,5] and the research results based on discrete frame images are the premise of continuous frame image research.
The research content of this paper is mainly based on the facial expression recognition of discrete frame images.

Convolutional Neural Network [6] (CNN) is a deep neural network that is cross-stacked by convolutional layers, pooling layers, nonlinear activation layers, and fully connected layers. It takes the pixel values of the image as direct input, through parameter sharing, Methods such as sparse connection and downsampling fully extract the local features of the input data itself for autonomous learning, acquire more abstract global features in the image non-display, and make it through data enhancement methods such as image translation, scaling, and rotation. It is robust and shows good results in computer vision related tasks [7,8]. When applying CNN to facial expression recognition, scholars have been studying how to improve the accuracy of facial expression recognition. H. Sikkandar et al.[9] have used various learning strategies such as image preprocessing, optimizing network structure, initializing weights and using external data to train deep convolutional neural networks to improve classification accuracy; Zhang Linlin et al.[10] proposed to improve the traditional single-channel deep convolutional neural network in order to reduce the problem of trainable parameters inside the network, and apply Maxout activation function and A-Softmax loss layer in the network structure.

Softmax loss function is commonly used in convolutional neural network, it is to distinguish between different classes by increasing the degree of dispersion between classes. However, the degree of compactness within the class also affects the classification performance of the network. In order to make the intra-class features more compact, Wen et al.[11] pointed out that the center loss function can minimize the intra-class distance, using the combined loss function of Softmax and center, while increasing the degree of feature dispersion between classes, it can also reduce the class inner distance.

Acharya et al.[12] proposed that the key point of facial expression recognition is the difference in the degree of distortion of facial muscles, rather than the existence of corresponding features. Second-order features can better reflect the twisted characteristics of muscles than first-order features. Inspired by this, this article believes that second-order features are more suitable for the task of facial expression recognition. Lin et al.[13] designed an end-to-end training bilinear model (Bilinear CNN), which represented the classifier as the product of two low-rank matrices, and the different features extracted by the dual-stream network could be in a second-order form. Fusion, and achieved a good result in the weakly supervised fine-grained classification direction on the CUB200-2011 dataset. Therefore, this paper proposes to use a bilinear model to extract the second-order features of facial expression images for expression recognition. In order to verify the effect of the model, experiments were performed on the FER2013 facial expression dataset.

2. Improved CNN

In recent years, deep convolutional neural networks have shined in tasks such as image classification and pattern recognition. At the same time, as the performance of GPUs and other hardware has greatly improved, the number of layers of neural networks has also been designed deeper and deeper, and the amount of calculations has also been getted higher and higher. From only 5-layer AlexNet [14] to 19-layer VGGNet[15] to 152-layer ResNet[16], ImageNet competition results are constantly being refreshed. Compared with large-scale image classification tasks such as ImageNet, expression recognition tasks can be regarded as one for 7 classification tasks with fewer classification categories, if a network with a deeper network layer is used, it is very easy to cause model overfitting. At the same time, experiments by Psramerdorfer et al.[17] show that under the same conditions, compare VGGNet, Inception and ResNet. The performance of the network, VGGNet has the best performance. Secondly, facial expressions are formed by facial muscle changes, so the image texture features extracted from the shallow layer in the convolutional neural network will be more conducive to expression classification than the abstract image features extracted from the deep layer. This article chooses the VGG-11 with fewer layers. The network structure is shown in Figure 1. Based on VGG-11, this paper improves its network structure to make it more in line with the task of expression recognition. This paper named the improved network model VGG-Emotion, and the network structure is shown in Figure2.
2.1. Batch Normalization
Since VGGNet does not have a batch normalization (BN) layer, it is difficult to converge during the training process. Because the input value of the deep neural network before the nonlinear function is trained during the training process, its overall distribution will gradually shift to the insensitive area of the activation function to the input, resulting in the gradient dispersion phenomenon of the neural network, and the network convergence slows down or even does not converge. The batch normalization layer normalizes the same batch of training data, and maps the output distribution of the convolutional layer to a standard normal distribution with a mean of 0 and a variance of 1, so that the input value of the activation layer falls on the region where the nonlinear activation function is sensitive to the input, increasing the gradient during back propagation and avoiding the problem of gradient disappearance. At the same time, the current advanced convolutional neural networks, such as Inception, ResNet, all set the batch normalization layer between the convolution layer and the activation layer, and use Conv-BN-Relu as a standard convolution module. Therefore, this article adds a batch normalization layer after each convolutional layer of VGG-Emotion.

2.2. Increase maximum pooling core
In the design of VGGNet, the maximum pooling layer with a core size of 2×2 and a step size of 2 is selected for downsampling. Inspired by the most advanced network structure, this paper intends to change the maximum pooling layer to a core with a core size of 3×3 and a step size of 2. It is believed that a larger pooling core can maintain image translation while downsampling. Denaturation is more generalized than the core size of 2×2. Especially in the task of facial expression recognition, it is more robust to a small amount of disturbance of the face position.

2.3. Global average pooling layer
The last convolutional layer of VGGNet is followed by a fully connected layer, and the parameters of the fully connected layer account for 90% of the model size. Therefore, Lin et al.[13] in the Network in Network paper presented the fully connected layer with a global average inspired by the replacement of the pooling layer, this article intends to replace the two fully connected layers of VGGNet with a global average pooling layer, which is equivalent to regularizing the entire network structure and directly removing the black box features in the fully connected layer. It directly gives practical meaning to each layer, reduces a large number of parameters of the model, alleviates the over-fitting phenomenon of the model, and pave the way for subsequent model compression and mobile terminal deployment.

2.4. Bilinear CNN
The bilinear model represents the classifier as the product of two low-rank matrices, and then merges the different features extracted by the dual-stream network in a second-order form. The bilinear model calculates the outer product of different spatial positions, and calculates the average convergence of different spatial positions to obtain second-order features. The outer product captures the pair-wise
correlation between characteristic channels and is translation invariant. Bilinear convergence provides a stronger second-order feature representation than linear features, and can be optimized end-to-end, achieving performance equivalent to or even higher than the part information used.

A bilinear convolutional neural network model consists of four parts. Among them, represents the feature extraction function, and A and B are two convolutional neural networks, representing the pooling function, and the classifier. The schematic diagram of the bilinear model is shown in Figure 3.

![Figure 3 Bilinear-CNN](image)

The feature extraction function(\(\phi()\)) is equivalent to a function mapping, which maps the input image \(I\) and the location area into a one-dimensional matrix feature, that is, the process of image feature extraction by CNN. The output of the two convolutional neural networks is combined by calculating the outer product through bilinear operation, and finally the second-order feature of the image is obtained. The calculation formula is as equation (1):

\[
\text{bilinear}(l, I, f_A, f_B) = f_A(l, I)^T f_B(l, I)
\]  

(1)

The position \(l\) covers the position and scale information, and \(I\) is the image. The dimension of \(f\) is \((K, D)\). \(D\) is the number of channels. If the feature dimensions output by A and B are \((K,M)\) and \((K,N)\) respectively, the dimension becomes \((M,N)\) after bilinear operation. The function of the pooling function is to aggregate the second-order features of all channels obtained after bilinear operation into a single-channel second-order feature. Specifically, the pooling function can select maximum pooling or additive pooling. The pooling function used by Bilinear CNN in the article is additive and pooling, that is, the second-order features of all channels are accumulated to synthesize the features of different positions to obtain a global picture, which is expressed as equation (2):

\[
\Phi(l) = \sum_{l_{\text{let}}} \text{bilinear}(l, I, f_A, f_B) = \sum_{l_{\text{let}}} f_A(l, I)^T f_B(l, I)
\]  

(2)

Since the position and dimension information of the feature is lost through the integration operation, the obtained secondorder feature is disordered. In addition, the outer product causes the dimension \(D\) of the feature to increase to the original square, and the extracted feature is a second-order feature. Finally, the second-order feature undergoes a symbol square root transformation, and L2 standardization is added, and then the image classification is completed for the fully connected layer. The bilinear normalization module is shown in Figure 4.

![Figure 4 Bilinear normalization](image)

Since the overall architecture is still a directed acyclic graph network, the back propagation (BP) algorithm can be used directly to optimize the parameters. In addition, the combination of outer products makes the gradient back propagation simple. The gradient solution for network A and network B is shown in equation (3) and equation (4):

\[
\frac{d\ell}{dA} = B \left( \frac{d\ell}{dz} \frac{dz}{dy} \frac{dy}{dx} \right)^T \]  

(3)

\[
\frac{d\ell}{dB} = A \left( \frac{d\ell}{dz} \frac{dz}{dy} \frac{dy}{dx} \right) \]  

(4)
Considering that the capacity of the model may be too large, it may cause over-fitting, so CNN adopts a fully shared parameter method, only using the previously proposed VGG-Emotion network for feature extraction, and then the VGG-Emotion network is globally averaged to pool the 512-dimensional features Matrix $X$ is used as the input of the bilinear module, as shown in Figure 2-5. Since the size of the matrix is $512 \times 1$, the size of the transposed matrix of the characteristic matrix $X$ is $1 \times 512$, and the outer product of the characteristic matrix is $512 \times 512$. The size of the second-order feature matrix eliminates the pooling operation in the bilinear model, and then the matrix is re-stretched into a one-dimensional feature vector of $1 \times 262144$ size, and then the feature vector is subjected to symbolic square root transformation and L2 standardization. Finally, the image is classified through the fully connected layer and the Softmax layer to complete the final expression classification. The bilinear model structure is shown in Figure 5.

![Figure 5 Bilinear model structure](image)

3. **Experiment**
All experiments in this article are completed in a laboratory environment, and the configuration is as follows: the system is Ubuntu16.04, the memory is 32GRAM, the GPU is Nvidia Titan X, the CUDA version is 8.0, the CPU is quad-core Intel Core i7, and the frequency is 6700KHz. At the same time, this article uses TensorFlow as a learning tool to construct the network model proposed in this article and complete the training.

3.1. **Dataset**
The traditional method of facial expression recognition is usually to manually extract features, which are susceptible to changes in lighting, posture, and occlusion. The datasets used in these methods are mainly collected in a specific environment (such as in a laboratory), with only frontal facial images. And some expressions are unnatural and exaggerated, which causes many methods to be affected by the database. If these methods are extended to other datasets or reality, the probability of obtaining good performance will be very small. The experiment in this article uses the FER2013 facial expression dataset provided by the Kaggle website, which is a representative dataset of spontaneous facial expressions, including age, race, gender, posture, background, lighting, and occlusion, as shown in Figure 6. Show.

The dataset consists of 35887 48×48 pixel grayscale images, including 7 expressions: happy, sad, disgusted, fear, surprised, angry, and neutral. The image is processed in such a way that the faces are almost centered, and each face occupies approximately the same space in each picture.

![Figure 6 FER2013 dataset](image)

3.2. **Image preprocessing**
In order to improve the accuracy and robustness of the model, make up for the shortcomings of insufficient expression data, and reduce over-fitting phenomena, three steps such as face alignment, data enhancement and image normalization are used.
First, perform face detection. Given an image containing a face, the bounding box of the face can be detected, and the original image is further cropped to obtain the facial area. Then there is the detection of key points of the face. After obtaining the general area of the face, input the detected face image to obtain the coordinates of the key points of the face. Finally, the face alignment is performed according to the key point coordinates.

- In the data reading stage of the training process, the input image is randomly enhanced, including random horizontal flip, random translation, random scaling, random rotation, random grayscale, random gamma transformation and random noise.
- Perform maximum and minimum normalization processing on all input images.

3.3. Experimental results and analysis

In order to verify the effectiveness of the algorithm in this paper, this paper uses the bilinear convolutional neural network model to conduct experiments on the FER2013 dataset. After each convolutional layer, a batch normalization layer is added and all the ReLU activation functions are used. Among them, the batch size of the batch normalization layer is 256. The cross-entropy loss function is used when training the model, the initial learning rate is 10e-4, the SGD optimizer with momentum is used, and the momentum is 0.9. Figure 7 shows the average of the model VGGNet expression convolutional neural network (VGGNet-Emotion, VE) and VGGNet bilinear expression convolutional neural network (VGG-Emotion-Bi, VE-Bi) on the FER 2013 dataset The accuracy rate changes with the increase of experimental epoch. Average accuracy (A-acc) is obtained by averaging the last 50 epoch accuracy values of each experiment, and the 50 epoch before the experiment is not averaged.

![Figure 7 VE and VE-Bi experimental results](image)

Table 1. The accuracy of mainstream methods and this method on the FER2013 dataset

| Method   | Accuracy |
|----------|----------|
| Inception| 71.6%    |
| ResNet   | 72.4%    |
| VGG      | 72.7%    |
| VE       | 72.5%    |
| VE-Bi    | 73.2%    |

Table 1 shows the experimental results (the boldface indicates the methods VE and VE-Bi in this article, and the underline indicates the highest accuracy in the evaluation index). It can be seen from the table that the FER2013 dataset is based on discrete. The accuracy of the VE-Bi single model of frames is higher than that of other mainstream methods. Among them, the accuracy of the VE-Bi method used in this article on the FER2013 dataset is 0.5% higher than the current state-of-the-art method, achieving better results than the current mainstream methods. In general, the bilinear model proposed in this paper has achieved comparable results to the current mainstream methods on the task of video facial expression recognition.
expression recognition. At the same time, it is proved that using bilinear CNN to extract second-order features of expression images is more beneficial than first-order features expression classification.

4. Conclusion

This paper proposes a facial expression recognition method based on fine-grained classification to extract the second-order features of expressions. First, the idea of fine-grained classification is elaborated, and the idea of fine-grained classification is applied to the task of facial expression recognition. Then, according to the difference of facial expression features, it is explained that the second-order information is more conducive to the task of facial expression recognition. The bilinear pooling algorithm is used to extract the characteristics of the second-order features of the image, and an end-to-end training model combining bilinear pooling is designed. Finally, we compare this model with several current mainstream facial expression recognition methods, and the results show that this method is better than most other models that currently only use a single model in terms of accuracy indicators.

References

[1] Mehrabian A. (1968) Communication without words[J]. Psychology Today, 2(4):53-56.
[2] Wang Keyao. (2019) Video facial expression recognition based on deep learning[D]. South China University of Technology.
[3] Zhang Xiang, Shi Zhicai, Chen Liang. (2020) Expression recognition method based on SWA optimized cascade network[J]. Electronic Science and Technology, 33(09):16-20.
[4] Y. Kim, B. Yoo, Y. Kwak, et al. (2017) Deep generative contrastive networks for facial expression recognition. https://arxiv.org/abs/1703.07140.
[5] J. Chen, R. Xu, and L. Liu. (2018) Deep peak-neutral difference feature for facial expression recognition, Multimedia Tools and Applications, pp. 1–1.
[6] CHANG T, WEN G, HU Y, et al. (2018) Facial expression recognition based on complexity perception classification algorithm [OL]. https://arxiv.org/abs/1803.00185.
[7] Lopes AT, de Aguiar E, De Souza AF, et al. (2017) Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order, Pattern Recognition, pp. 610–628.
[8] Pons G, Masip D. (2017) Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis. IEEE Transactions on Affective Computing, pp. 1-1.
[9] H. Sikkandar, R. Thiyagarajan. (2020) Deep learning based facial expression recognition using improved Cat Swarm Optimization. pp. 1-17.
[10] Zhang Linlin. (2019) Research on facial expression recognition based on convolutional neural network [D]. Changchun University of Technology.
[11] Wen Y, Zhang K, Li Z, et al. (2016) A Discriminative Feature Learning Approach for Deep Face Recognition [M]. Computer Vision–ECCV 2016. Springer International Publishing, pp. 499-515.
[12] Acharya D, Huang Z, Pani Paudel D, et al. (2018) Covariance pooling for facial expression recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 367-374.
[13] Lin T Y, RoyChowdhury A, Maji S. (2015) Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE international conference on computer vision. pp. 1449-1457.
[14] Shi Cuiping, Tan Cong, Zuo Jiang, Zhao Kexin. (2020) Facial expression recognition based on improved AlexNet convolutional neural network[J]. Telecommunications Technology, 60(09): 1005-1012.
[15] Chen shu. (2020) Multi-angle facial expression recognition and its application based on improved VGGNet [D]. Shenyang University of Technology.
[16] Wu Yuhao, Chen Xiaohui. (2020) Facial expression recognition system based on improved ResNet[J]. Information and Communication, (07): 37-39.
[17] Pramerdorfer C, Kampel M. (2016) Facial expression recognition using convolutional neural networks: state of the art[J]. https://arxiv.org/abs/1612.02903.