Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

© 2020 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA200685

Facial Expression Analysis Based on Fusion Multi-Layer Convolutional Layer Feature Neural Network

Hao Meng[a], Fei Yuan[a,1] and Tianhao Yan[a]
[a]*Department of Automation, Harbin Engineering University, China*

Abstract. Concerning the problem that the current facial expression analysis based on convolutional neural network (CNN) only uses the features of the last convolutional layer but the recognition rate is not high, this paper proposes the use of sub-deep convolutional layer features and builds a CNN model which fuses the features of multi-layer convolutional layers. The model uses a CNN for feature extraction and saves the deepest feature vectors and sub-deep feature vectors of the expression images. The sub-deep feature vector is used as the input of the multilayer CNN established in this paper. The processed fourth convolution layer feature is fused with the deepest feature previously saved to perform facial expression analysis. Experiments are performed on FERPLUS dataset, Cohn-Kanade dataset (CK+) and JAFFE dataset. The experimental results show that the improved network structure proposed in this paper can capture richer feature information during facial expression analysis, which greatly improves the accuracy of expression recognition and the stability of the network. Compared with the original CNN-based facial expression analysis using only the last layer of convolution layer features, using multi-layer fusion features on three kinds of datasets can improve the expression recognition rate by 33. 3%, 2.3% and 22%, respectively.

Keywords. Convolutional neural network, features of sub-deep convolutional layers, multi-layer convolutional layer Feature fusion, facial expression analysis

1. Introduction

Facial expression analysis refers to the use of computers to analyze human facial expressions and changes through pattern recognition and machine learning algorithms and to judge human psychology and emotions, thereby achieving intelligent human-computer interaction [1]. Deep convolution neural network has the outstanding characteristics of unsupervised feature learning, which has been proved to have the ability to mine the deep potential distributed expression features of data in the fields of image, speech and text. It is very effective when using deeper levels (ie with many layers) to learn features with high level of abstraction [2]. Therefore, in facial expression analysis, CNNs are also often used [3-4], using the powerful learning capabilities of CNNs to learn deep feature representations of expression pictures.

[1] Corresponding Author: Fei Yuan, Harbin Engineering University, China. Email: bohelion@hrbeu.edu.cn

The CNN can be regarded as a combination of feature extraction and classifier, which maps the image layer by layer and the result is the result of feature extraction. Judging from the mapping of its various layers, it is similar to a feature extraction process and features at different levels are extracted. Through continuous interactive mapping and finally mapping to several labels, it has the function of classification. However, the features extracted by the intermediate convolutional layer also include certain information and have a certain expression ability for the image [5]. Ali Mollahosseini [6]et al. proposed a deep neural network architecture to address the facial emotion recognition problem across multiple well-known standard face datasets; Hui Ding [7]et al. proposed a novel idea for training facial expression analysis networks based on static images, because the deep features may contain redundant information from the pre-training domain. These show that the use of intermediate convolutional layer features can improve the feature representation of pictures and then improve the accuracy of deep convolutional network classification. At present, most deep learning models for facial expressions have low accuracy and weak feature representation capabilities.

Based on GoogleNet Inceptionv3 [8] network for facial expression analysis, this paper proposes a fusion neural network structure with multiple convolutional layers to improve the expression recognition rate. The feature of sub-deep convolutional layer using CNN is proposed to ensure that deeper features can be obtained if the original features are relatively complete. The model is based on the GoogleNet Inception v3 network. First save the feature vector of the deepest convolutional layer currently used by the CNN and the feature vector of the sub-deep convolutional layer proposed in this paper; Secondly, the sub-deep high-dimensional feature vector is used as the input of the multilayer CNN established in this paper for training; Finally, the processed convolutional layer 4 features are fused with the deepest feature vector previously saved to perform softmax feature classification.

The rest of the paper is organized as follows. Section II gives an overview of our proposed approach, describing the features of each layer of the CNN and the improved network structure proposed in this paper. It also includes the multi-layer CNN established in this paper. Section III provides experimental results. Section IV concludes the paper.

2. Deep neural network

In CNNs, different convolution kernels have different sizes and the receptive fields are different. CNN can be regarded as the combination of feature extraction and classifier. From the mapping of each layer of CNN, it is similar to a feature extraction process, which extracts different levels of features. The CNN can map the features to different labels, which makes the CNN have the ability of classification. In this study, CNN is regarded as a method of feature extraction. The traditional deep convolution neural network is divided into two parts: feature extraction and final classification. Among the basic CNNs for image classification, the best one is GoogleNet[9-10]. This network structure mainly uses a split-merge idea. First, it splits, makes many branches and each branch does its own convolution pooling, then the results are concatenated to form a better feature channel. Capturing multi-scale features improves multi-scale adaptability and increases the width of the network. This paper uses GoogleNet Inception v3

network for feature extraction. Figure 1 shows the last few layers of the Inceptionv3 structure seen on the tensorboard [11] visualization. Currently the Inception v3 model trained on ImageNet uses the 1×2048-dimensional vector output from the last layer poo1_3 before softmax classification, but the recognition rate is not good for expression recognition.

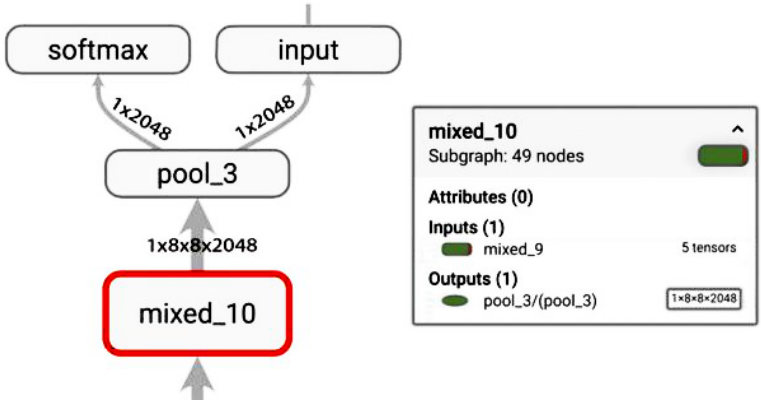


Figure 1. Tensorboard visualization node

The CNN is to map the image layer by layer and the mapping result is the result of feature extraction. How many convolution kernels are there in each convolution layer will get the characteristics of how many channels. After visualizing the convolution layer through feature map [12], we can get the characteristic map of each channel and fuse each channel according to 1:1 to get the fused characteristic map, as shown in Figure 2. Because there are many convolution layers in Inception v3, there are also many convolution kernels in each layer, that is, many channels. Figure 2 only shows the features of the first 25 channels of the first convolution layer and the features after 1: 1 fusion of all channels. Figure 3 shows the convolution feature map of all channels of different convolution layers. Through the visualization of the feature map, it can be seen that the shallow features tend to detect the edge of the image and the detected content is comprehensive. At the same time, there will be key information extracted (such as the bright eyes and mouth of the first convolution layer). With the deepening of the level, the feature image is more and more abstract and the resolution of the image is smaller and smaller. At the same time, a lot of information is ignored. Relatively speaking, the deeper the level is, the more representative the extracted features are. The current CNNs use only the features output by the last convolutional layer for classification and the intermediate feature information also has a certain ability to express images.

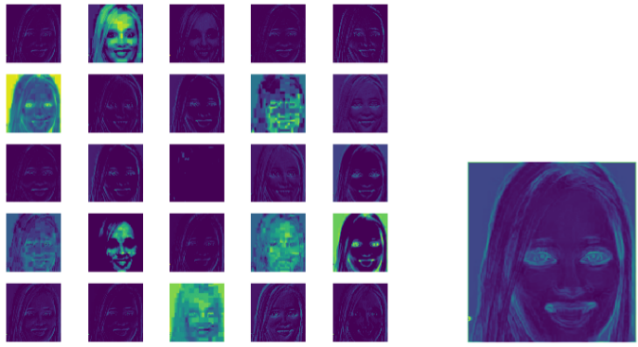


Figure 2. Feature map of the first 25 channels of the convolution layer 1 and the feature map of each channel after 1: 1 fusion

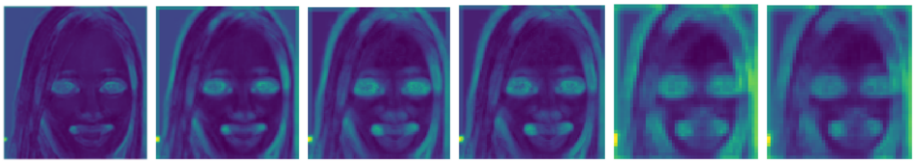


Figure 3. Feature map of each convolutional layer after 1: 1 fusion

This study proposes to use the output vector of the previous convolution layer mixed_10 of the convolution layer pool_3, that is, the input 1×8×8×2048 of pool_3 as the feature vector and extract the feature vector of the node mixed_10 and save it. The selection of sub-deep features can ensure that deeper features are obtained when the original features are relatively complete. The deeper the number of layers, the higher the level of semantic information and the more sufficient the semantic information is. For the 8×8×2048 feature vector after extracting the mixed_10 node, this study establishes a CNN structure as shown in Table 1 for training.

Table 1. Multi-layer convolutional network structure

Layer	Input(W*H*D)	Kernel_num	Kernel_size	Stride	Pad	Out(W*H*D)
Conv1	8×8×2048	2048	3	1	0	6×6×2048
Conv2	6×6×2048	2048	3	1	0	4×4×2048
Conv3	4×4×2048	2048	3	1	0	2×2×2048
Conv4	2×2×2048	2048	2	1	0	1×1×2048
Conv5	1×1×2048	1024	1	1	0	1×1×1024
Conv5	1×1×1024	512	1	1	0	1×1×512
Conv6	1×1×512	10	1	1	0	1×1×10

The network structure designed in this paper is shown in Figure 4. After inputting facial expression pictures, they are sent to the GoogLeNet Inception v3 network for feature extraction and the deepest feature vector and sub-deep feature vector are extracted. The sub-deep features are processed using the multi-layered volume neural network established in this paper (that is, Table 1) and the processed feature vector output from the con4 layer is fused with the deepest feature vector of the original CNN

to obtain a fused feature vector with a size of 1×2048. Finally, the fused 2048-dimensional feature vector is subjected to softmax classification.

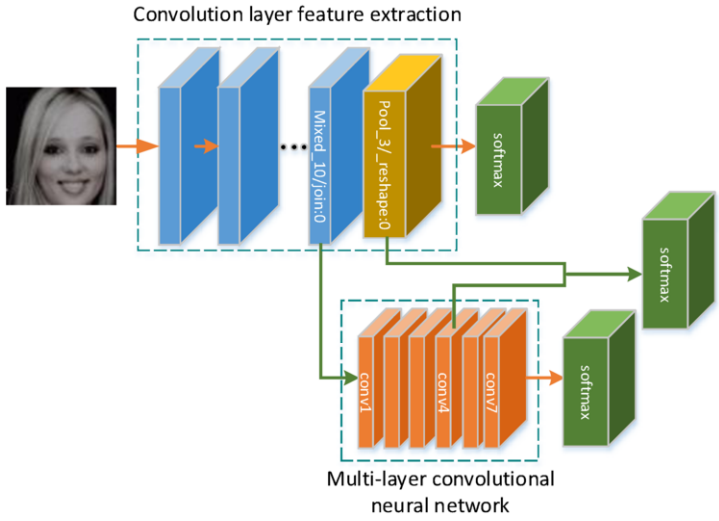


Figure 4. Designed Network Structure

3. Experiment and Results

The experiment is based on the deep learning framework tensorflow of Python3, using the operating system Windows10. Hardware configuration: the CPU is Intel (R) Xeon (R) gold 5122 CPU, the main frequency is 3.60ghz and the memory is 16.0gb; the GPU is NVIDIA geforce RTX 2080 Ti and the video memory is 16GB. This paper uses the FERPLUS dataset [13], the CK+ dataset [14] and the JAFFE dataset [15]. Among them, there are 10 categories of tags in the FERPLUS dataset: neutral, happiness, surprise, sadness, anger, disgust, fear, contempt, unknown, NF. This article removes the unknown and NF, which are a total of 8 expression categories. Both CK+ and JAFFE datasets have 7 expression categories. In this paper, the image is processed into a size of 299 pixels×299 pixels and the data set is divided into a training set, a verification set and a test set. The experimental settings train 6000 epochs, the initial learning rate is set to 0.01, the optimizer uses Adam and the batchsize is set to 100.

The experimental steps are as follows:

(1)Send the facial expression data set directly to the Inception v3 network for classification, save the feature 1×2048-dimensional vector of the bottleneck layer, that is, the feature of the deepest convolutional layer and save the 8×8×2048-dimensional feature vector output by the node mixed_10, that is, the feature of the sub-deep convolutional layer, records the final test results;

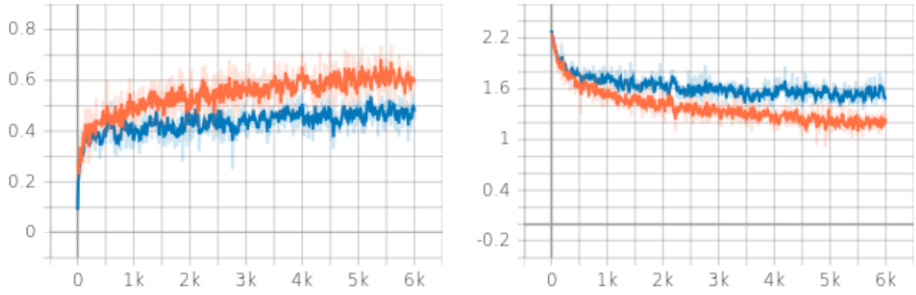
(2)Extract the 8×8×2048-dimensional feature vector output by the mixed_10 node from the network model and save it. Send these 8×8×2048-dimensional vectors to the multi-layer convolutional neural network(that is, Table 1)established by us for classification, save the 1×2048-dimensional feature vectors of the con4 layer and record the final test results;

(3)The feature vectors of the bottleneck layer layer saved in step 1 and the feature vectors of the con4 layer in step 2 are fused and sent to a CNN for classification and the final test results are recorded.

3.1 Experiment on the FERPLUS Dataset

3.1.1 Deepest Feature Training Experiment

The FERPLUS dataset is sent to the inception v3 network for migration learning and the softmax layer was changed from the original 1000 class to 8 class for training classification. Each face expression picture is sent to the network and the features of the bottleneck layer layer, that is, the 1×2048-dimensional feature vector output by the node poo1_3 are extracted and saved. Each digit is a 32-bit floating point number, which is a total of 35887 pictures, which is 35887×1×2048. And save the 8×8×2048-dimensional feature vector output by node mixed_10. Figure 5 is a graph of accuracy and loss function during training. Figure (a) shows a graph of accuracy of training and verification. Figure (b) shows a graph of loss function of training and verification. Orange represents training and blue represents verification.



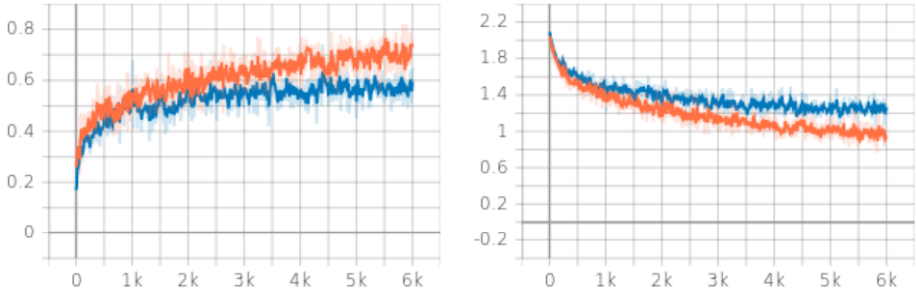
(a) Accuracy chart **(b)** loss function graph

Figure 5. Accuracy curve and loss function of train and validation

It can be seen from Figure 5 that the accuracy rate during training is stable on average at 60% and the loss function value is stable on average at 1.2; the accuracy rate during verification is stable at 46% and the loss function value is stable at 1.5 on average. The accuracy rate is relatively low, the loss function value is relatively large and the curve oscillation is relatively large and unstable.

3.1.2 Sub-deep Feature Training Experiment

In this paper, a multi-layer neural network is designed and the saved sub-deep feature 8×8×2048-dimensional feature vector is used as the input of the multi-layer CNN for training and classification. Figure 6 is a graph of accuracy and loss function during training. Figure (a) shows a graph of accuracy of training and verification. Figure (b) shows a graph of loss function of training and verification. Orange represents training and blue represents verification.



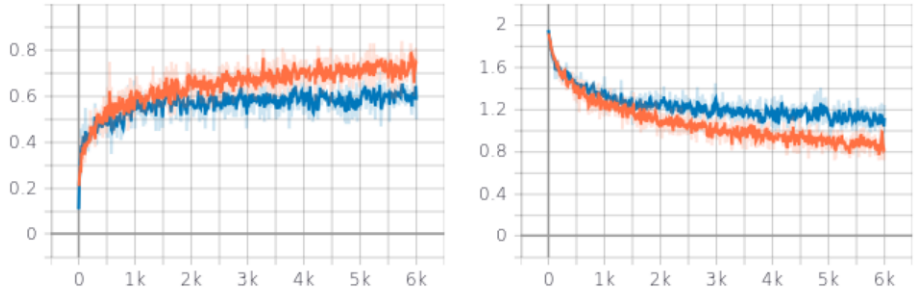
(a) Accuracy chart **(b)** loss function graph

Figure 6. Accuracy curve and loss function of train and validation

It can be seen from Figure 6 that the accuracy rate during training is stable at 66% on average and the loss function value is stable at 1; the accuracy rate during verification is stable at 58% and the loss function value is stable at 1.2 on average. Compared with the traditional method, only the deepest features are used, the accuracy is improved, the value of the loss function is reduced, the amplitude of the oscillation is reduced and the curve is smoother. The effectiveness of using the features of the sub-deep convolutional layer proposed in this paper is proved.

3.1.3 Fusion Feature Training Experiment

The 1×2048-dimensional features of the conv4 layer processed by the multilayer neural network established in this paper are fused with the 1×2048-dimensional feature vectors output by the previously saved node pool_3 and the softmax layer is changed to 8 classes for training classification. Figure 7 is a graph of accuracy and loss function during training. Figure (a) shows a graph of accuracy of training and verification. Figure (b) shows a graph of loss function of training and verification. Orange represents training and blue represents verification.



(a) Accuracy chart **(b)** loss function graph

Figure 7. Accuracy curve and loss function of train and validation

It can be seen from Figure 7 that the accuracy rate during training is stable at an average of 70% and the value of the loss function is stable at 0.9; the accuracy rate during verification is stable at an average of 62% and the value of the loss function is stable at 1.1. Compared with using only the sub-deep features proposed in this paper, the accuracy rate has been improved again, the loss function value has been reduced, the oscillation amplitude has become smaller and the curve has been smoother. It is proved that the use of the fused features can further improve the expression recognition

rate and network stability. That is, the effectiveness of the improved convolutional network structure proposed in this paper.

3.2 Final Experimental Results

We obtained three models trained and saved on the FERPLUS dataset through 3.1 and then performed the same experiments on the CK+ dataset and JAFFE dataset to train and save the models according to the above experimental steps. Finally, the test is performed on the test set and the experimental results are shown in the following table.

Table 2. Test Results

Test accuracy (%)	FERPLUS	CK+	JAFFE
Primitive deepest feature	47.7	97.2	61.9
Sub-deep features proposed	59.4	98.9	72.3
Fusion features proposed	63.6	99.4	75.5

It can be concluded that:

Test on the FERPLUS dataset, CK + dataset and JAFFE dataset. The test accuracy of the saved model trained using the original CNN on the final test set is 47.7%, 97.2% and 61.9%, respectively; The accuracy of the model trained and saved using the features of the sub-deep convolutional layer proposed in this paper is 59.4%, 98.9% and 72.3% on the final test set; Using the trained and saved model after fusing two features, the test accuracy on the final test set is 63.6%, 99.4% and 75.5%, respectively. The use of the fused features is 33.3%, 2.3% and 22% higher than using only the deepest features in the original CNN. Compared with the sub-deep features using only the CNN proposed in this paper, it has improved by 26.7%, 1.7% and 16.8%, respectively. It is proved that the convolutional network structure proposed in this paper, which integrates the features of multi-layer convolutional layers, effectively improves the expression recognition rate.

4. Summary and Discussion

The current CNN uses only the last layer of convolutional layer features for facial expression classification, but the effect on expression recognition is not high. This paper proposes a multi-layer convolutional layer feature vector neural network model, which combines the deepest feature vector and sub-deep feature vector of the CNN for facial expression analysis. The experimental results show that the improved network structure proposed in this paper can capture richer feature information, thereby improving the expression recognition rate.

But there are still many pacts need to be improved. (1)In this paper, only the deepest feature vector and the penultimate layer feature vector in the CNN are used for feature fusion. According to the use of the features of the convolutional layer by the full CNN, we can also try to use the previous layer or even the first two layers of the sub-deep feature vectors. However, at the same time, the amount of calculation will increase, which needs to be considered comprehensively. (2)Aiming at the problems of low resolution and high error rate of the data set, an expression database can be re-established to improve the expression recognition rate. (3) For the problem of low

discrimination between facial expression classes, the softmax loss function can be researched and changed, so that the CNN is more suitable for facial expression classification.

References

[1] Haines N, Bell Z, Crowell S, et al. Using automated computer vision and machine learning to code facial expressions of affect and arousal: Implications for emotion dysregulation research[J]. Development and psychopathology, 2019, 31(3): 871-886.

[2] Xin M, Wang Y. Research on image classification model based on deep convolution neural network[J]. EURASIP Journal on Image and Video Processing, 2019, 2019(1): 40.

[3] Liu Q, Zhang J, Xin Y. Face expression recognition based on improved convolutional neural network[C]//Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition. 2019: 61-65.

[4] Sun X, Lv M. Facial Expression Recognition Based on a Hybrid Model Combining Deep and Shallow Features[J]. Cognitive Computation, 2019, 11(4): 587-597.

[5] Amin S U, Alsulaiman M, Muhammad G, et al. Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification[J]. IEEE Access, 2019, 7: 18940-18950.

[6] Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks[C]//2016 IEEE Winter conference on applications of computer vision (WACV). IEEE, 2016: 1-10.

[7] Ding H, Zhou S K, Chellappa R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition[C]//2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017: 118-126.

[8] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.IEEE, 2015: 1-9.

[9] Yoo H J. Deep convolution neural networks in computer vision: a review[J]. IEIE Transactions on Smart Processing & Computing, 2015, 4(1): 35-43.

[10] Al-Qizwini M, Barjasteh I, Al-Qassab H, et al. Deep learning algorithm for autonomous driving using GoogLeNet[C]//2017 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2017: 89-96.

[11] Nguyen V, Dang T, Jin F. Predict saturated thickness using tensorboard visualization[J]. Visualization in Environmental Sciences 2018, 2018.

[12] Zou J, Rui T, Zhou Y, et al. Convolutional neural network simplification via feature map pruning[J]. Computers & Electrical Engineering, 2018, 70: 950-958.

[13] Tümen V, Söylemez Ö F, Ergen B. Facial emotion recognition on a dataset using convolutional neural network[C]//2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE, 2017: 1-5.

[14] Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]//2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010: 94-101.

[15] Lyons M J, Akamatsu S, Kamachi M, et al. The Japanese female facial expression (JAFFE) database[C]//Proceedings of third international conference on automatic face and gesture recognition. 1998: 14-16.