# Convolution Neural Network for Automatic Facial Expression Recognition

Xiaoguang Chen[1], Xuan Yang[1], Maosen Wang[2],Jiancheng Zou[1]
Institute of Image Processing and Pattern Recognition, North China University of Technology[1]
No.5 Jinyuanzhuang Road,Shijingshan District[1]
School of Compute Science and Techonology, Beijing University of Posts and Telecommunications[2]
No.10 Xitucheng Road, Haidian District[2]
Beijing, China
Tel:+ 86-18813030733, Email:15261823898@163.com

**Abstract**

Facial expression recognition is a hot research direction of pattern recognition and computer vision. It has been increasingly used in artificial intelligence, human-computer interaction and security monitoring in recent years. Convolution neural network (CNN) as a depth learning architecture can extract the essential features of the image, and in the case of large changes in shooting conditions, its effect is better than the traditional methods of Support Vector Machines(SVM) and Principal Component Analysis(PCA). Therefore, an improved method of facial expression recognition based on CNN is proposed in this paper. The purpose is to classify each facial image as one of the seven facial expressions considered in this study. According to the characteristics of facial expression recognition, a new convolution neural network structure is designed which uses convolution kernel to extract implicit features and max pooling to reduce the dimensions of the extracted implicit features. In comparison to AlexNet network, we can improve the recognition accuracy about 4% higher on the CK+ facial expression database by the aid of Batch Normalization (BN) layer to our network. A facial expression recognition system is constructed in this paper for the convenience of application, and the experimental results show that the system could reach the real-time needs.

**Key words:** Facial expression recognition; CNN; Deep learning

## Ⅰ. Introduction

Facial expression as an important form of human inner emotional expression, is able to convey rich emotional information, together with words and sounds constitute the main way to express emotions. Therefore, facial expression recognition has become a popular research topic in recent years [1,2]. Automatic facial expression recognition systems is an important foundation to achieve affective computing and artificial intelligence, and the systems can be beneficial in many fields, like human-computer interaction, data mining, video conferencing, criminal interrogations, psychiatry, medical care, etc. When we are face to face, we can easily understand the meaning of the facial expressions of the others, but for the machine, understanding the human emotional accurately is still a challenge. With the improvement of face detection algorithm and the innovation of feature extraction technology, the facial expression classification effect was significantly better than before. But there are still many problems, such as low precision, slow identification, etc. So we still need to continue to study this subject.

Convolution neutral network(CNN) is a new type of neural networks, it is a combination of traditional artificial neural network and deep learning technology. With the development of deep learning, applying convolution neutral network into a classification problem has attained impressive success [3,4,5].The successes stems from that they perform the feature extraction and classification process simultaneously. Deep learning method features which include critical and unforeseen features are extracted by deep learning methods through iterative weight update by back propagation and error optimization. The method avoids the complicated feature extraction and data reconstruction process in the traditional recognition algorithms [6].

## Ⅱ. Related Work

In the mid and late twentieth century, people began to study facial expressions. In the 1970s, American psychologists Ekman and Friesen[7] made groundbreaking work on modern facial expressions. Ekman defined six basic expressions of human beings. Happy, Angry, Disgust and Sad, determine the type of recognition object. And they established the Facial Action Coding System (FACS). In the year of 2001-2006,FACS was used to classify human facial movements by their appearance on the face using Action Units (AU)[8,9], the subtle expression on the face is detected through the relationship between facial movement and expression.2003, Ira Cohen, Nicu Sebe et al.[10] introduced and tested different Bayesian network classifiers for classifying expressions from video, proposed an architecture of Hidden Markov models (HMMs) for automatically recognizing human facial expression from video sequences.2009,a method which use LBP for facial expression was proposed by Shan et al.[11].They accomplished the expression recognition by support vector machine (SVM) classifiers with boosted-LBP features.In recent years, deep learning using convolution neural networks for feature extraction of image data is becoming more popular.2014,Liu et al.[12] used 3D-CNN and a deformable facial action part model to localize facial action parts and learn part-based features for emotion classification.2015, Peter Burkert et al.[13] proposed a convolution neural network architecture for facial expression recognition. The proposed architecture is independent of any hand-crafted feature extraction. 2016, Ali et al.[14] proposed a collection of boosted neural network for multiethnic facial expression recognition. The success of each technique is dependent on pre-processing of the images and feature selection.

## Ⅲ. Proposed Method

In this paper, a new CNN structure is proposed for facial

 *ISBN 978-1-5090-4897-7*

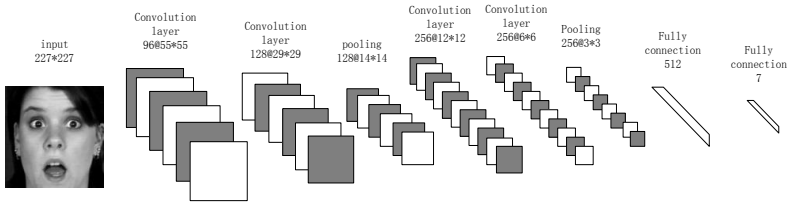


Fig. 1 Facial expression recognition network structure

expression recognition. Two publicly available databases CK+[15] and JAFFE[16] are used to carry out the experiments. In order to remove non-expression feature of a facial image, we need to carry out the pre-process of face image, which includes face detection and face image cropping. The faces are detected and cropped using OpenCV library. Then, the facial expressions features are extracted using our convolution neural network which under deep learning framework. Finally, the training model is used to classify each facial image as one of the seven facial expressions (angry, disgust, neutral, sad, fear, surprise and happy) considered in this study.

*A. Feature Extraction using CNN*

Considering that the features of facial expression recognition classification task are less, the CNN network structure for facial expression recognition is designed, and named as FENet. As fig.1 shows, the entire network includes four convolution layers, two pooling layers and two fully connection layers.

The first is the convolution layer. The convolution layer is the feature extraction layer. The convolution operation is performed on the training convolution kernel and the previous layer of all the feature maps. The output of the activation function forms the neurons of the current layer, thus forming the characteristic map of the current convolution layer. The calculation is as follows:

$$\mathrm{net}_j^l = \sum_{i\in M_j} a_i^{l-1} \otimes \omega_{i,j}^l + \omega_b \quad (1)$$

$$a_{i,j}^l = F(\mathrm{net}_{i,j}^l) \quad (2)$$

Where $\mathrm{net}_j^l$ represents the weighted input of layer l. $a_i^{l-1}$ is the characteristic map of the output of the l-1 layer.$\omega_{ij}^l$ denotes a convolution kernel matrix, it represent the connection weight between neurons.$\omega_b$ denotes the offset term of the jth feature map. In the experiment, set $\omega_b = 0$, can improve the speed of network training, while reducing the learning parameters. $a_{i,j}^l$ is the j feature graph of the convolution l layer. F() represents the activation function. This model uses 96 filters with the size of 11×11×4.This layer extracts the low-level edge features. The original image is randomly cropped to the size of 227×227, after the first convolution of the image size becomes 55×55.Fig.2 shows the sample output after first convolution filters being applied on a face image. To increase the nonlinear properties of network, we use ReLU(Rectified Linear Units) as the activation function. For any given input value x, ReLU is defined by:

$$F(x) = \max(0,x) \quad (3)$$

Where x is the input to the neuron. Using the ReLU activation function allows us to avoid the vanishing gradient problem caused by some other activation functions [17].

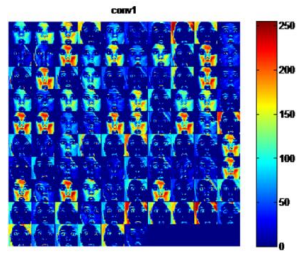


Fig. 2 first convolution layer output

With the increasing number of convolution layers, the feature dimension increases rapidly. In order to avoid such a dimension of disaster, we need to use the pooling layer to reduce the dimension. In the experiment, the down-sampling is performed by Max-pooling. The down-sampling process does not change the number of feature graphs. It reduces the number of parameters by removing unnecessary information from each feature map. The calculations are as follows:

$$\mathrm{net}_j^l = \mathrm{down}\left(a_j^{l-1}\right) + \omega_b \quad (4)$$

$$a_j^l = F(\mathrm{net}_j^l) \quad (5)$$

Where, $a_j^l$ represents the j feature map of the pool layer l. β is the offset term of the down-sampling layer. $\omega_b$ represents the offset term of the down-sampling layer. down() means the down-sampling function.

The full connection layer can also be considered as a convolution layer, but the full connection layer requires that the input must be a one-dimensional array, so that its convolution kernel size and the original data size are consistent with convolution layer. The output of each neuron is:

$$\mathrm{net}_j^l = \sum_i \omega_{i,j}^l a_j^{l-1} + \omega_b^l \quad (6)$$

$$y = F(\mathrm{net}_j^l) \quad (7)$$

Where $\mathrm{net}^l$ represents the output vector of the fully connected l layer. $\omega^l$ is the weight coefficient matrix. $a^{l-1}$ represents the input feature vector. $\omega_b^l$ denotes the offset term of the fully connected l layer. F() represents the activation function.

## Ⅳ. Experiments and Results

*A. Dataset*

The models are trained and tested on the database from the extend Cohn-Kanaded database (CK+) which includes 327 video sequences acted out by 118 participants. Each sequence which consists of approximately 10 to 30 frames is labeled with one of seven expressions categories: angry, disgust, neutral, sad, fear, surprise, happy. Every sequence starts with the neutral emotion and then the frame depicts the emotion which is for the corresponding label. We selected images of human faces with obvious facial expressions to recognize facial expression. All 327 sequences of the CK+ database are used for evaluating the proposed model. To verify the generalization of the proposed method some cross-database experiments were also performed. The experiments used the JAFFE database which consists of 213 images from 10 Japanese female subjects.

*B. Pre-processing image*

The implementation of the algorithm was done by using OpenCV, C++ and a GPU based CNN library (Caffe[18]).In the pre-process of image, the background information is removed. Then, all face images are normalize to 227×227 pixels. Pre-processed image samples are shown in Fig.3.During the training process, the super-parameter is set as follows, the initial learning rate is 0.01, and the initial weight attenuation is set to 0.0005.

Fig.3 Examples of seven facial expressions. (a)Angry,(b)Disgust, (c)Fear, (d)Happy, (e)Sad, (f)Surprise, (g)Neutral

*C. Results*

In this paper, the most commonly used stochastic gradient descent method is used for model training. The weight update formulas are as follows:

$$V_{t+1} = 0.01V_t - 0.0005\nabla L(W_t) \quad (8)$$

$$W_t = W_t + V_{t+1} \quad (9)$$

Where t is the number of iterations, V is the momentum, and $\nabla L\ (W_t)$ represents the error of the objective function's backward propagation when iterating t times. In experiment, the mean is set to 0, using a standard deviation of 0.01 Gaussian distribution for random initialization. The basic parameters of each training are the same, but in the actual training process, the parameter settings will be adjusted manually. After a certain number of iterations, if the accuracy rate does not change, the learning rate will reduce to the 1/10 of original value, until the number of iterations is equal to the maximum number of iterations to complete the network training process. In the experiment, on the basis of the above convolution neural network structure, we add BatchNormal layer, in order to speed up the training speed, improve the model accuracy. The test results are shown in the following table 1. From the table we can see that after adding the BN layer the results are significantly improved.

It can be seen from the experimental results that with the increasing of the number of iterations, the accuracy of the test set gradually increases. When the number of iterations reaches a certain number of times, the accuracy almost unchanged, that indicates the network has reached the optimal. In the vertical comparison of the two networks can be seen from table.1, adding BN layer make the network convergence faster, and the effect better.

Table.1 Comparison of the accuracy of the three network structures under different iterations

| Network | number of iteration | | | | |
|---|---|---|---|---|---|
| | 5000 | 9000 | 18000 | 25000 | 35000 |
| Alexnet[19] | 0.9158 | 0.9263 | 0.9368 | 0.9438 | 0.9468 |
| FENet | 0.8525 | 0.8528 | 0.8736 | 0.8722 | 0.8735 |
| FENet with BN | 0.9741 | 0.9800 | 0.9810 | 0.9815 | 0.9815 |

A better evaluation of the proposed method in reality is the cross-database experiments, i.e. train the method with one database and test with another. To perform the cross-database experiment, CK+ database were used to train the network and JAFFE database was used to validate. The results are shown in table.2.

Table.2 Cross-database experiment for the CK+ and JAFFE databases.

| Train | Test | accuracy |
|---|---|---|
| CK+ | CK+ | 0.991597 |
| CK+ | JAFFE | 0.831172 |
| JAFFE | JAFFE | 0.877350 |

Compared to the current top level, the list are shown in table.3.From the results of the comparison, we can see that the training results of the proposed network in this paper better than others.

Table.3 Comparison of the accuracy of different methods

| Method | Number of | Accuracy |
|---|---|---|
| LBP+SVM(2009) | 6 | 89.1% |
| CLM+SVM(2011) | 6 | 96% |
| DCNN(2016) | 7 | 96.02% |
| Proposed in this paper | 7 | 98.15% |

## Ⅴ. Facial Expression Recognition System

Face expression recognition system can detect and recognize face expression in real time, and can be used in everyday applications such as access control systems, intelligent service equipment and so on. Considering the widely use of facial expression recognition system, this paper proposes to construct facial expression recognition system. The system workflow is shown in Fig.4.

 *ISBN 978-1-5090-4897-7*

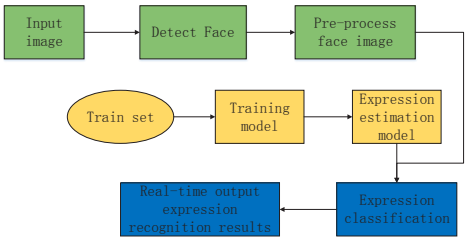


Fig.4 Face recognition system flow chart

As shown in the fig.4, we firstly need to load pre-trained models and related configuration files. In the camera real-time image acquisition process, Firstly, use the Haar classifier for face detection. Then, the cropped face area is normalized to 227×227 pixels. Finally, input image to the network and the system instantaneously output recognition results to the video image. Results are shown in Fig.5.

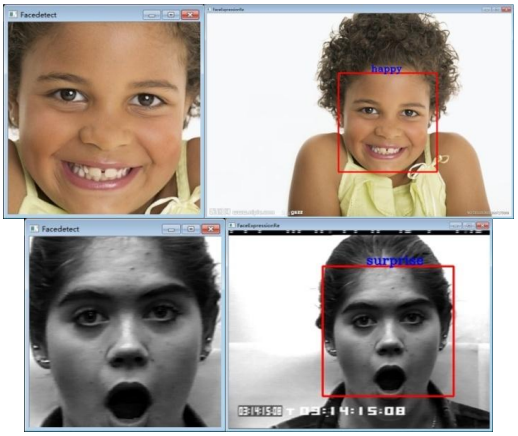


Fig.5 the results of face detect and face recognition

The system implements three data read mode, not only can detect the face image from the camera, but also can detect the face image from the picture file and the video file. In addition the system can detect and identify the multiple faces at the same time.

## Ⅵ. Conclusion

In this paper, a facial expression recognition system based on convolution neural network is proposed and the pre-process of image step is added. As we can see the experimental results in both instantaneity and efficiency is improved by adding the pre-processing step before training data and adding BN layers to the network.

We plan to investigate the network in the bigger database in order to increase the robustness of facial expression recognition algorithm in unknown environments (e.g. with varying light condition, occlusion and others) in the future. And we will add some functions on the recognition system, such as age identification and other functions, the most important thing is to further improve the recognition accuracy.

## References

[1]J.Yin et al.Face Festure Extraction Based on Principle Discriminant Information Analysis. In: Proceedings of the IEEE International Conference on Automation and Logistics, 2007, pp. 1580-1584.
[2]M. Pantic and L.J.Rothkrantz.Automatic analysis of facial expressions: The state of the art. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000, pp. 1424-1445.
[3]Ian J. Goodfellow,et al.Challenges in representation learning: A report on three machine learning contests. In Neural information processing .2013, pp. 117-124.
[4]Z.Yu and C.Zhang. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction.2015, pp. 435-442.
[5]S.E.Kahou,et al.Combining modality specific deep neural networks for emotion recognition in video. In Proceedings of the 15th ACM on International conference on multimodal interaction. 2013, pp. 543-550.
[6]S.Minchul,et al. Baseline CNN structure analysis for facial expression recognition. 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) 2016, pp. 26-31.
[7]P.Ekman and W.V.Friesen..Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press, 1978.
[8]T.Kanade et al.Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence.2001, pp. 97-115.
[9]M.S. Bartlett et al. Fully automatic facial action recognition in spontaneous behavior. In Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition, 2006, pp. 223-230.
[10]C.Ira,et al. “Evaluation of expression recognition techniques.” Image and Video Retrieval. Springer Berlin Heidelberg, 2003, pp. 184-195.
[11]C.Shan,et al. Facial expression recognition based on local binary patterns: A comprehensive study. Image Vision Computer.2009, pp. 803-816.
[12]M.Liu. Deeply learning deformable facial action parts model for dynamic expression analysis. In Computer Vision–ACCV. Springer International Publishing.2004, pp. 143-157.
[13]P.Burkert,et al.DeXpression:Deep Convolutional Neural Network for Expression Recognition.2015.
[14]G.Ali, et al.Boosted NNE collections for multicultural facial expression recognition.PatternRecognition. 2016, pp.14–27.
[15]P.Lucey,et al.The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on IEEE.2010, pp. 94–101.
[16]M.Lyons,et al.Coding facial expressions with gabor wavelets. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition.1998,pp. 200–205.
[17]A.Krizhevsky et al. Imagenet classification with deep convolutional neural networks.In Advances in neural information processing systems, 2012, pp . 1097–1105.
[18]Y.Jia,et al. Caffe: Convolutional architecture for fast feature embedding. Eprint Arxiv.2014,pp. 675-678.
[19]K.Alex,et al.Imagenet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems. 2012 , pp . 1097-1105.