ICIC Express Letters
Part B: Applications
Volume **8**, Number **4**, April **2017**

ICIC International ©2017 ISSN 2185-2766
pp. **767–772**

DEEP CONSECUTIVE CONVOLUTIONAL NEURAL NETWORK MODEL FOR FACIAL EXPRESSION RECOGNITION

Lianqiang Niu[1,2], Xiangzhen Chen[2] and Shengnan Zhang[2]

[1]School of Software
[2]School of Information Science and Engineering
Shenyang University of Technology
No. 111, Economic and Technological Development Zone, Shenyang 110870, P. R. China
niulq@sut.edu.cn; chen_xiang_zhen@163.com; zsncjr@sina.com

Received September 2016; accepted December 2016

Abstract. *General convolutional neural network is not able to accurately express the facial expression features and to generalize effectively. A kind of facial expression recognition model based on deep consecutive convolutional neural network is designed in this paper. In the model, small-scaled kernels are employed to extract local features more detailedly, and to increase non-linear representation ability of the model with the help of two consecutive convolutional layers. After this, the dropout technique is introduced to optimize the model. As a result, the concurrent dependence of updated weight on the implicit nodes with fixed relation in combination is decreased. Experimental results show that the proposed model has obvious advantages in the recognition accuracy and generalization performance compared with other approaches, and is able to provide good performance and practicability in the face expression analysis and recognition.*
Keywords: Deep learning, Convolutional neural network, Network structure, Facial expression recognition

1. **Introduction.** Since 2006, the deep learning, as a new machine learning theory, has been successfully applied to signal processing, computer vision and other fields, and got better effect especially in the aspects of speech recognition, computer vision, natural language processing and information retrieval. In terms of structure, DBN (Deep Brief Network) [1], proposed by Hinton et al., can be considered as a superposition of multiple RBM (Restricted Boltzmann Machine) [2], similar to the traditional multilayer perceptron, but it needs unsupervised training before supervised training, and then makes the learned parameters as the initial values of supervised learning. It is exactly the change of learning method that makes the deep structure can remedy the previous BP network's deficiency. In 2007, Bengio et al. [3] proposed that the auto encoder can better initialize the weights of all layers to reduce optimized difficulty of the deep network. After this, some improved techniques and approaches (Specially, CNNs (Convolutional Neural Networks) and DCNNs (Deep Convolutional Neural Networks)) are proposed in [4-12]. Although the proposed architecture has become a success for computer vision, these models were designed to keep a large computational budget, and could be put to real world use at a reasonable cost. So how to improve the network performance and reduce the size of the network model is also a challenge for our research.

Facial expression is an important research topic in emotional computing, intelligent control, computer vision, image processing and pattern recognition. The static facial expression recognition is based on static images to recognize facial expression. Commonly, the basic human expression is defined as seven categories: Happy, Angry, Surprise, Fear, Disgust, Sad and Neutral. The expression recognition approaches can fall into two categories which are called global approach and local approach respectively. The former includes PCA (Principal Component Analysis), ICA (Independent Component Analysis)

and LDA (Linear Discriminate Analysis), and the latter contains Gabor wavelet transform, LBP (Local Binary Patterns), and so on [13-18].

This paper discusses a deep neural network (DNN) model for facial expression recognition constructed with the combination of consecutive convolution, max-pooling and dropout, with the purpose of improving the property of DNN to extract local features better and enhance the network's ability for non-linear representation. By comparing and analyzing two different kinds of data sets, it showed a strong ability to recognize facial expression. The rest of this paper is organized as follows. In Section 2, typical CNNs named LeNet-5 is reviewed. Section 3 presents the DCNNs model with two consecutive convolutional layers, and parameters selection and other efficient techniques, while in Section 4 the experiments performed are explained. Finally, Section 5 concludes the paper.

2. **Construction of Consecutive Convolutional Network.** As typical CNNs, LeNet-5 [6] stacks two convolutional layers and sampling layers before connecting one or more fully connected layers. Hereinto, the convolutional layers use a small convolutional kernel (such as 5×5) as a feature detector to execute a convolution with the original large resolution image to obtain the feature activation values of any positions on the image. Therefore, the convolutional layer can be described by the number of feature maps and the size of the kernels. Each convolutional layer is composed of several feature maps with the same size, and each feature is extracted by one of its own kernels.

The most direct way to improve the performance of the deep neural network is to increase its depth and width. However, the two simple solutions come with two major drawbacks. One is that a larger network size usually means a number of parameters, which makes the network more prone to producing over-fitting, especially for the use of finite number of training sets of labeled samples. The other is that a large network size will greatly increase consumption of computing resources. In order to improve the representation ability of the network rather than a massive increase in size, we design a network structure with two consecutive convolutional layers which takes sample for feature maps by the pooling layer after consecutive convolutional operations, as shown in Figure 1.

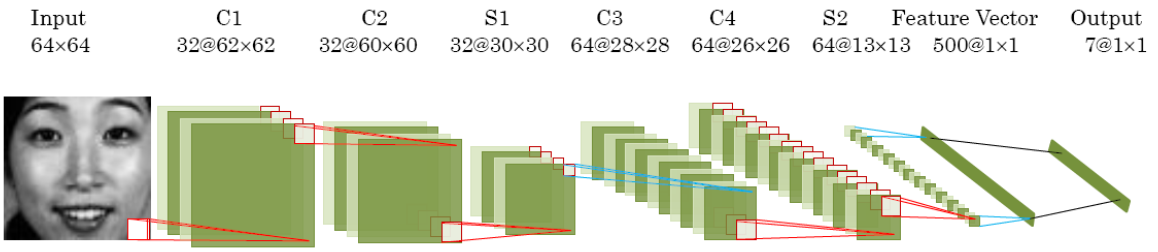

FIGURE 1. Structure of CNNs with two consecutive convolutional layers

Our consecutive convolutional structure consists of 4 convolutional layers and 2 pooling layers. The original input image firstly is executed by a convolution operation, and then we refine the features by two consecutive convolutional layers and reduce dimension by pooling layer. Subsequently, we use double feature maps to extract more features and execute 2 convolutional operations and one pooling operation. Finally, a 500 dimensional feature vector is generated.

3. **Analysis and Parameter Optimization.** When a classification problem is not linearly separable by a simple function, traditional linear convolution is not sufficient to abstract the features for classification, and a highly nonlinear function should be required

to enhance the abstraction ability of local model. It is generally believed that, in the conventional CNNs, by means of an over complete set of filters which is able to cover all the changes of potential features, a linear filter can be learned to detect the different changes of the same features. Filters need to consider the combination of all changes which transmit from the previous layer. The high-level features are formed by the high-level filters with combination of low-level features. As a result, in the composition of the higher level features, it is more advantageous to make a better abstraction of them in each local area.

3.1. **Consecutive convolution with small convolutional kernels.** In CNNs, an output feature map is obtained by a nonlinear activation function applying an input set. An input set is the sum of a convolution to the input image or feature map of previous layer with a linear filter (convolutional kernel) and a bias term. The kth feature map in l layer executes a convolution with selected subset M_j. Assuming that x_i^{l-1}, k_{ij}^l, b_j^l, and $f(\cdot)$ denote input data, weight, bias, and active function respectively, the corresponding output feature x_j^l can be given as follows:

$$x_j^l = f\left(\sum\nolimits_{i\in M_j} x_i^{l-1} k_{ij}^l + b_j^l\right) \tag{1}$$

Since the neurons of specific feature maps share their weights, the parameters needed to be learned are reduced, and the algorithm can be run in parallel. That improves the efficiency of the algorithm and generalization ability.

While two consecutive convolutional layers are introduced, the feature map output of previous layer is executed by a convolution operation again, and then a new feature map output can be obtained as follows:

$$x_j^l = f\left(\sum\nolimits_{i\in M_j} f_i^{l-1} k_{ij}^l + b_j^l\right) \tag{2}$$

From the perspective of network structure design, around the premise that the demand of expression ability can be met, the network should use scale as small as possible to reduce the network learning time and complexity. As we can see, the consecutive convolution can improve the representation ability despite controlling the network in an appropriate size. By Formula (2), it can be found that the image is calculated by two nonlinear activation functions, and the representation of the complex degree and the nonlinear ability of the function is enhanced.

A convolutional kernel represents a local receptive field of a neuron. Considering the effect range, while smaller convolutional layers are directly connected, we can get a same receptive field size as that gained by a convolutional layer with a larger convolution kernel. The existing applications show that the convolutional kernel with 5×5 size is a kind of appropriate receptive field. Therefore, in the consecutive convolutional layers, we use the smaller 3×3 convolutional kernels. It is because that a 5×5 convolutional layer can be formed by two consecutive 3×3 convolutional layers in the receptive field equivalently. Since one non-linear rectified layer is replaced by two nested ones, the non-linear ability of the network is enhanced. This makes the decision function to be more discriminative, and have better ability to deal with complex images. Besides, the structure can reduce the parameters of the network model. If the number of channels in a consecutive two-layer 3×3 convolution stack is C, the weight number will be $18C^2$, but $25C^2$ in a single 5×5 convolutional layer. Thus, more complex networks can be constructed with the same number of parameters.

3.2. **Pooling.** In order to be able to describe some subtle changes of features, 2×2 max-pooling is employed as a pooling approach, shown as Formula (3):

$$x_j^l = f\left(\beta_j^{l-1} \textit{max-pooling}\left(x_j^{l-1}\right) + b_j^l\right) \tag{3}$$

The *max-pooling*$(\cdot)$ denotes the pooling function, and β_j^{l-1} and b_j^l denote the weights and the bias respectively for the jth output feature map. The *max-pooling*$(\cdot)$ can find the maximum of a consecutive $n \times n$ local region in the input image of current layer, and the size of the output image is $1/n$ of the size of the input image.

3.3. **ReLU and dropout.** The common activation function used by neural network is $f(x) = (1 + e^{-x})^{-1}$ or $f(x) = \tanh(x)$. We treat neurons with the nonlinearity as $f(x) = \max(0, x)$ (Rectified Linear Units, ReLUs) suggested by Nair and Hinton [7]. At the same time, the dropout technique [8] is also introduced to control the node weights. With the dropout, each weight of hidden neuron will stop to work in a 0∼0.5 random probability. Since that each time the sample of the input network is updated, the hidden nodes are random in a certain probability, it avoids the occurrence of every 2 hidden nodes. Therefore, the updating of the weights is no longer dependent on the common function of the implicit nodes with fixed relationship. Some features can be avoided, which can effectively restrain the fitting problem and enhance the generalization ability of the network.

4. **Experiments and Results Analysis.** In order to compare performances of CNNs with two consecutive convolutional layers (DCCNNs), LeNet-5 and general 3-layer CNNs (DCNNs), three experiments and their results are described here.

Two typical facial expression databases are employed. The first database is the JAFFE (Japanese Female Facial Expression) [9] which contains 213 female facial expression images. The second one is the Cohn-Kanade, which is released in 2010, namely CK+ [10]. For JAFFE, according to the approach [11], the 3/4 or 2/3 images are selected as the training samples, and the other 70 images as test samples. For CK+, according to the approach [12], the first piece of each image sequence is selected as the natural expression sample, and the last three as other expression sample, and the total of 1308 images are divided into training sample and test sample according to the ratio of 1:1.

In order to remove the influence of the background, the face detection algorithm Adaboost is employed to detect the face area, namely pure face. After detecting, all pure face images are separated and normalized to 64×64 size. This is able to reduce the network parameters and the difficulty of training.

The structure of the DCNNs is shown in Table 1, three convolutional layers, three pooling layers and two fully connected layers. Each convolution kernel size is 5×5, and the pooling kernel size is 2×2. Table 2 demonstrates the average recognition rates of LeNet-5, DCNNs and DCCNNs on the JAFFE and CK+ database.

Table 1. Structure parameters of DCNNs

Layer	Type	Out Maps	Kernel Size	Pooling Size
0	Input	64 3 64 64		
1	Conv1	64 64 60 60	5	
2	Pool1	64 64 30 30		2
3	Conv2	64 64 26 26	5	
4	Pool2	64 50 13 13		2
5	Conv3	64 128 9 9	5	
6	Pool3	64 128 5 5		2
7	ip1	64 500 1 1		
8	ReLu1	64 500 1 1		
9	ip2	64 7 1 1		

TABLE 2. Accuracies of three approaches on JAFFE and CK+ expression database

Approach	Database	Accuracy
LeNet-5	JAFFE	94.11%
DCNNs	JAFFE	97.14%
DCCNNs	JAFFE	100%
LeNet-5	CK+	92.85%
DCNNs	CK+	94.28%
DCCNNs	CK+	100%

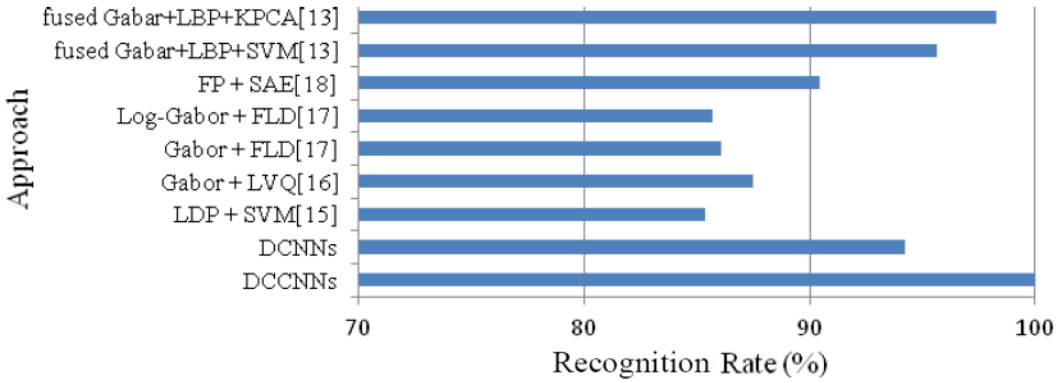

FIGURE 2. Recognition rates of several approaches on CK+

By observing the picture which was recognized wrongly by DCNNs, we can see that it is easy to recognize the Sad expression as Disgust. These two expressions have some similar features, such as mouth down, frown, and so on. For the fine local features, the neural network with two consecutive convolutional layers abstracts the features in two consecutive abstractions, and is able to extract or represent the features better, so as to realize correct classification. At the same time, the use of dropout technique can effectively avoid the joint effect between nondependent features and learn better network parameters. The recognition rate was about 1 percentage higher than that of the model without dropout.

Figure 2 demonstrates the ability differences among DCNNs, DCCNNs and other approaches which extract features manually. It shows that the consecutive convolutional structure has certain advantages in the extraction of more complex features.

5. **Conclusion.** By twice feature extraction and twice consecutive nonlinear activations with two consecutive convolutional layers, the capabilities of our model on the complex function's fitting and nonlinear degree are improved and is in favor of extracting the locally complex features. In the process of convolution, the local detailed features of images can be effectively extracted by means of applying small-sized kernel consecutively, and network parameters can also be reduced. In addition, the dropout technique is able to effectively prevent against the combined action of feature detector, and improve the performance of neural network and restrain over-fitting.

The further work of this paper is to investigate the mechanism of consecutive convolution and the influence on the performance with different numbers of consecutive convolutional layers.

REFERENCES

[1] R. Sarikaya, G. E. Hinton and A. Deoras, Application of deep belief networks for natural language understanding, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol.22, no.4, pp.778-784, 2014.

[2] A. Fischer and C. Igel, Training restricted Boltzmann machines: An introduction, *Pattern Recognition*, vol.47, no.1, pp.25-39, 2014.

[3] Y. Bengio, P. Lamblin, D. Popovici et al., Greedy layer-wise training of deep networks, *Advances in Neural Information Processing Systems*, p.153, 2007.

[4] L. Wan, M. Zeiler, S. Zhang et al., Regularization of neural networks using dropconnect, *Proc. of the 30th International Conference on Machine Learning*, pp.1058-1066, 2013.

[5] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, vol.25, no.2, pp.1097-1105, 2012.

[6] Y. Lecun, L. Bottou, Y. Bengio et al., Gradient-based learning applied to document recognition, *Proc. of the IEEE*, vol.86, no.11, pp.2278-2324, 1998.

[7] V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, *Proc. of the 27th International Conference on Machine Learning*, pp.807-814, 2010.

[8] G. E. Hinton, N. Srivastava, A. Krizhevsky et al., Improving neural networks by preventing coadaptation of feature detectors, *Computer Science*, vol.3, no.4, pp.212-223, 2012.

[9] M. J. Lyons, J. Budynek and S. Akamatsu, Automatic classification of single facial images, *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol.21, no.12, pp.1357-1362, 1999.

[10] P. Lucey, J. F. Cohn, T. Kanade et al., The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pp.94-101, 2010.

[11] Y. Zheng, Q. Liu, E. Chen et al., Time series classification using multi-channels deep convolutional neural networks, *Lecture Notes in Computer Science*, vol.8485, pp.298-310, 2014.

[12] M. Liu, S. Li, S. Shan et al., Au-aware deep networks for facial expression recognition, *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pp.1-6, 2013.

[13] S. Liu, Y. Tian and C. Wan, Facial expression recognition method based on Gabor multi-orientation features fusion and block histogram, *ACTA Automatica Sinica*, vol.37, no.12, pp.1455-1463, 2011.

[14] C. Shan, S. Gong and P. W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, *Image and Vision Computing*, vol.27, no.6, pp.803-816, 2009.

[15] T. Jabid, M. H. Kabir and O. Chae, Robust facial expression recognition based on local directional pattern, *ETRI Journal*, vol.32, no.5, pp.784-794, 2010.

[16] S. Bashyal and G. K. Venayagamoorthy, Recognition of facial expressions using Gabor wavelets and learning vector quantization, *Engineering Applications of Artificial Intelligence*, vol.21, no.7, pp.1056-1064, 2008.

[17] N. Rose, Facial expression classification using Gabor and log-Gabor filters, *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, pp.346-350, 2006.

[18] Y. Lv, Z. Feng and C. Xu, Facial expression recognition via deep learning, *Proc. of IEEE International Conference on Smart Computing*, Hong Kong, China, pp.303-308, 2014.