LETTER

A Joint Convolutional Bidirectional LSTM Framework for Facial Expression Recognition*

Jingwei YAN[†], Wenming ZHENG[†a)], Zhen CUI[††], *Nonmembers*, *and* Peng SONG[†††], *Member*

SUMMARY Facial expressions are generated by the actions of the facial muscles located at different facial regions. The spatial dependencies of different spatial facial regions are worth exploring and can improve the performance of facial expression recognition. In this letter we propose a joint convolutional bidirectional long short-term memory (JCBLSTM) framework to model the discriminative facial textures and spatial relations between different regions jointly. We treat each row or column of feature maps output from CNN as individual ordered sequence and employ LSTM to model the spatial dependencies within it. Moreover, a shortcut connection for convolutional feature maps is introduced for joint feature representation. We conduct experiments on two databases to evaluate the proposed JCBLSTM method. The experimental results demonstrate that the JCBLSTM method achieves state-of-the-art performance on Multi-PIE and very competitive result on FER-2013.

key words: *facial expression recognition, convolutional neutral network, long short-term memory, shortcut connection*

1. Introduction

Facial expression recognition, which aims at recognizing human emotions via images or videos automatically, is a fundamental research topic in human-computer interaction and computer vision domains. A growing number of applications based on the facial expression recognition technology were proposed during the past several years. For example, we can design digital cameras which are capable of taking photos automatically when smile is detected. In dealing with facial expression recognition problems, feature extraction plays an important role. Robust features such as local binary patterns (LBP) and scale invariant feature transform (SIFT) are widely adopted for this purpose. In order to extract features with more discriminative ability for facial expressions, some researchers proposed to combine conventional methods with sparse representation and developed various effective algorithms. In [1], Yan *et al.* proposed sparse locality preserving projection to obtain the intrinsic manifold of facial features and select features simultaneously. Zheng [2] extracted features on image grids in a multi-scale manner and proposed group sparse reduced-rank regression (GSRRR) to model the relationship between feature vector and the sample label, in which facial regions related to expressions can be selected due to the group sparsity.

Nowadays deep learning methods, especially convolutional neural networks (CNNs) [3] and recurrent neutral networks (RNNs) [4], have dominated the pattern recognition and computer vision communities. Especially, CNNs have been widely applied in facial expression recognition field in recent years. For example, Zhang *et al.* [5] developed 1D CNN consisting of 1D convolutional and 1D max pooling layer to explore high-level semantic information from the feature matrix for multi-view facial expression recognition.

Conventional deep CNNs consist of multiple convolutional filtering and pooling layers which are stacked alternately. By progressively enlarging the receptive field, the convolutional filters are sequentially performed on local spatial regions of intermediate feature maps, so as to capture image features from low-level textural to high-level semantic. Thus CNNs may be regarded as a composite function of local spatial filtering. However, the locally performed convolutional filter can hardly model the spatial dependency relation, especially large-range spatial dependencies within the image.

The investigation of exploring the effect of spatial relation information in image classification tasks has been done in recent years. In [6] Shi *et al.* transformed 2D image to 3D tensor and developed the convolutional LSTM to capture the spatial correlation. Zuo *et al.* [7] converted CNN output into four directed sequences and utilized traditional RNN to learn the spatial dependencies of different image regions. Liang *et al.* [8] proposed a novel recurrent CNN model by integrating recurrent units into every convolutional layer. Since the receptive field of the convolutional layer grows larger as the depth of recurrent connections increases, the context information is incorporated in each convolutional layer.

It is known that the generation of facial expression involves movements of several organs such as eyebrows, eyes and lips, which are also described as action units (AUs). A group of AUs appear together when certain emotion is expressed, *e.g.*, the raised cheek (AU6) and pulled up lip cor-

Manuscript received September 21, 2017.
Manuscript revised December 1, 2017.
Manuscript publicized January 11, 2018.

[†] The authors are with the Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, Jiangsu, P.R. China.

[††] The author is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, P.R. China.

[†††] The author is with the School of Computer and Control Engineering, Yantai University, Yantai 264005, Shandong, P.R. China.

[*] The work is partly supported by National Basic Research Program of China (No. 2015CB351704) and National Natural Science Foundation of China (No. 61231002 and No. 61572009).

a) E-mail: wenming_zheng@seu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2017EDL8208

Copyright © 2018 The Institute of Electronics, Information and Communication Engineers

ner (AU12) together present the happy expression. More AUs will be involved as the facial expression becomes more complicated. In this letter, we will investigate the facial expression recognition problem by leveraging the correlations between AUs and learning the spatial dependency information among different image regions through long short-term memory (LSTM) [9]. To this end, we propose a joint convolutional bidirectional LSTM (JCBLSTM). In this method, CNN is firstly employed to extract discriminative middle-level features and then bidirectional LSTM is utilized to learn the spatial dependencies within the generated sequences from CNN feature maps. Moreover, a shortcut connection [10] is added for CNN to boost the information flow in the network.

The remainder of this letter is organized as follows: In Sect. 2, we propose the JCBLSTM method for facial expression recognition. Experiments are presented in Sect. 3. Section 4 concludes the letter.

2. JCBLSTM Method

In this section, we will propose the JCBLSTM method for facial expression recognition. Figure 1 illustrates the framework of the JCBLSTM method, which integrates CNN with LSTM to learn discriminative features and image-level spatial correlations simultaneously. In this framework, the image samples are transferred into a series of convolutional layers to extract middle-level facial texture features. Then, the rows or columns of the feature maps in the last convolutional layer are regarded as ordered sequences along two directions. Moreover, LSTM is deployed respectively to model the spatial relationship in them and the outputs of the bidirectional LSTM are added up together. Finally, the fully connected vectors from both CNN and LSTM are concatenated together as the overall feature representation to predict the expression labels. As CNN and LSTM are connected seamlessly, the whole framework can be trained in an end-to-end manner.

2.1 Transferred VGG-Face for Image Feature Extraction

As shown in Fig. 1, the goal of the convolutional layers is to learn discriminative middle-level facial texture features for the subsequent LSTM and the joint feature representation. To this end, we employ the convolutional structure of VGG-Face network [11], which was originally trained for the purpose of face recognition. As the released VGG-Face model is trained on large scale face databases, we are able to fine-tune the convolutional layers with limited sample images and training iterations while still obtain discriminative facial expression related texture features. Feature maps output from the last convolutional layer, *i.e.*, conv5_3[†], are fed in the subsequent bidirectional LSTM layers.

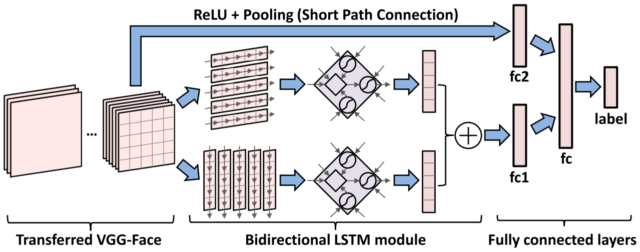

ReLU + Pooling (Short Path Connection)
...
fc2
fc1
fc
label
Transferred VGG-Face
Bidirectional LSTM module
Fully connected layers

Fig. 1 Framework of the proposed JCBLSTM (forward pass). The framework contains three parts, *i.e.*, transferred VGG-Face module, bidirectional LSTM module and fully connected layers.

2.2 Bidirectional LSTM for Image Spatial Dependency

As facial expressions are composed of different AUs, such as the movements of brows and lips, spatial relationship between different facial regions plays an important role in facial expression recognition. However, convolutional filters fail to capture such relationship due to the fact that they are performed locally on image regions. Therefore, exploring the spatial dependencies within facial expression image is very desired to improve the facial expression recognition performance. Consequently, we adopt the LSTM method for this purpose, in which we regard each row or column in feature maps as a directed sequence. Then, the generated sequence is ordered from left to right or from top to bottom separately (see Fig. 1). It is notable that each element in the sequence actually represents the receptive field of conv5_3, which is a relatively large region in the original image sample.

As is shown in Fig. 1, we employ LSTM along both directions in our method. Then, the LSTM responses of the two spatial sequence, from left to right ($\mathbf{o}_t^{\rightarrow}$) and from top to bottom ($\mathbf{o}_t^{\downarrow}$), are concatenated and added up together. The procedure is summarized as the following equations:

$$\begin{aligned}
\mathbf{o}_t^{\rightarrow} &= \sigma(\mathbf{W}_o^{\rightarrow}\mathbf{x}_t^{\rightarrow} + \mathbf{U}_o^{\rightarrow}\mathbf{h}_{t-1}^{\rightarrow} + \mathbf{b}_o^{\rightarrow}) \\
\mathbf{o}_t^{\downarrow} &= \sigma(\mathbf{W}_o^{\downarrow}\mathbf{x}_t^{\downarrow} + \mathbf{U}_o^{\downarrow}\mathbf{h}_{t-1}^{\downarrow} + \mathbf{b}_o^{\downarrow}) \\
\mathbf{o}_t &= \mathbf{o}_t^{\rightarrow} + \mathbf{o}_t^{\downarrow}
\end{aligned} \tag{1}$$

where $\mathbf{o}_t$ is the final output of bidirectional LSTM. $\mathbf{W}_o^{\rightarrow}$, $\mathbf{U}_o^{\rightarrow}$, $\mathbf{W}_o^{\downarrow}$, and $\mathbf{U}_o^{\downarrow}$ are weight matrices for corresponding directions and the activation function is $\sigma(x) = \frac{1}{1+e^{-x}}$.

2.3 Short Path for Convolutional Layers

With the CNN-LSTM structure, discriminative facial texture features and spatial region relationship information can be modeled simultaneously. Furthermore, in order to enhance the information flow in the network, especially in favor of convolutional layers to learn discriminative features for facial expression recognition, we introduce an extra short path for convolutional layers in JCBLSTM. Shortcut connections, which can be constructed in various ways, are helpful for information flow between early layers and later layers. As shown in Fig. 2, the ReLU, pooling and fully connected layers on the right side make up the short path

[†] There are 3 convolutional layers in the fifth convolutional set.

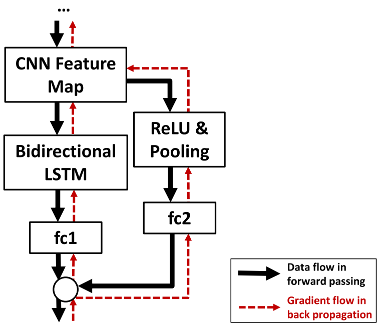

...
CNN Feature Map
Bidirectional LSTM
ReLU & Pooling
fc1
fc2
Data flow in forward passing
Gradient flow in back propagation

Fig. 2 Illustration of information flow in the short path for convolutional layers. Solid arrows indicate the forward pass procedure while dotted arrows indicate the gradient information flow in the back propagation procedure.

for VGG-Face feature maps. In the back propagation during training procedure, in addition to being passed through the complicated bidirectional LSTM module, gradient information can be easily propagated back to the convolutional layers through the shortcut connection.

Another advantage of the shortcut connection is that the facial texture features and the spatial dependency information can be concatenated in the subsequent fully connected layer. Therefore, both features are fused together so as to model facial expressions more accurately and boost the overall performance of our framework.

3. Experiments

In this section we will verify the proposed JCBLSTM by conducting experiments on two facial expression databases, namely Multi-PIE [12] and FER-2013 [13]. Both databases consist of images under various poses and illuminations.

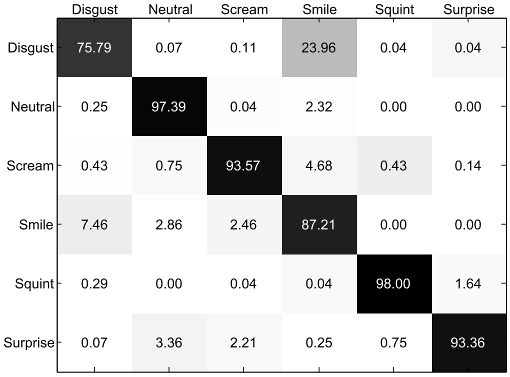
3.1 Experiments on Multi-PIE Database

Multi-PIE database contains facial expression images collected under laboratory controlled environment. For each subject, images under 15 view points and 19 illumination conditions were taken. Similar to the protocol used in [2], we select 100 subjects who performed all six facial expressions under seven views, *i.e.*, from 0° to 90°, increased by 15°. Among them 80 subjects are selected as training samples and the remaining 20 subjects are testing samples. For each subject we use all images taken under 19 illumination conditions in both training and testing phases.

We conduct experiments to verify the effectiveness of spatial dependencies and shortcut connection in facial expression recognition task. As is shown in Table 1, the recognition rate of the finetuned VGG-Face network alone is 88.63%, which has already surpassed the previous state-of-the-art method [15]. VGG-Face & BLSTM method refers to the framework without shortcut connection and joint feature representation, *i.e.*, only the output of bidirectional LSTM is utilized to classify the samples. This method performs better than the finetuned VGG-Face network which indicates that spatial relation information learned by bidirectional LSTM improves the model ability in facial expression recognition. Our proposed JCBLSTM performs best among all methods and achieves the state-of-the-art accuracy of 90.89%. This demonstrates that spatial dependencies and the shortcut connection for convolutional layers together boost the performance of the original CNN architecture. The confusion matrix of JCBLSTM is shown in Fig. 3. Except for disgust and smile, accuracies of other expressions are satisfactory.

Table 1 Comparison of facial expression recognition accuracies on Multi-PIE.

Method	Accuracy (%)
LGBP & SVM [14]	80.17
GSRRR [2]	81.7
1D CNN [5]	85.2
LLCBL [15]	86.3
VGG-Face (finetune)	88.63
VGG-Face & BLSTM	89.42
JCBLSTM	**90.89**



	Disgust	Neutral	Scream	Smile	Squint	Surprise
Disgust	75.79	0.07	0.11	23.96	0.04	0.04
Neutral	0.25	97.39	0.04	2.32	0.00	0.00
Scream	0.43	0.75	93.57	4.68	0.43	0.14
Smile	7.46	2.86	2.46	87.21	0.00	0.00
Squint	0.29	0.00	0.04	0.04	98.00	1.64
Surprise	0.07	3.36	2.21	0.25	0.75	93.36

Fig. 3 Confusion matrix of JCBLSTM on Multi-PIE.

3.2 Experiments on FER-2013 Database

Facial Expression Recognition 2013 (FER-2013) database was released in the facial expression recognition challenge of ICML 2013. It was collected using the Google image search API with a bunch of emotion-related words. Various head poses, illumination conditions and occlusions exist in the database. It consists of 35887 image samples with 7 categories. The competition organizers divided the database into training set, public test set and private test set. Here we conduct experiments to compare the model performances on private test set as the competition does.

Results of different methods are listed in Table 2. As the samples are all 48 $*$ 48 gray images under various conditions, the overall performances are not that high as Multi-PIE. From the aspect of loss status during training procedure shown in Fig. 4, compared to Multi-PIE, the loss curve of FER-2013 can not decrease to a very low status. The winning solution in the competition utilized L2-SVM loss function to train CNN and achieved an accuracy of 71.16%. JCBLSTM performs slightly better than the winner and the

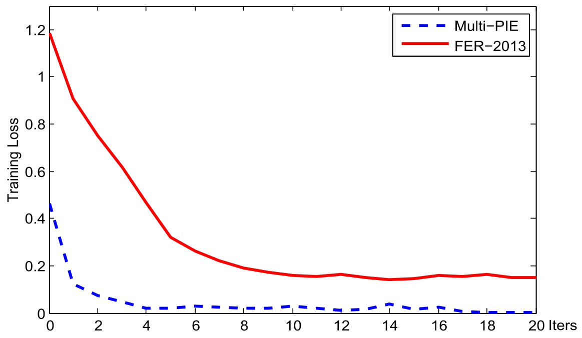
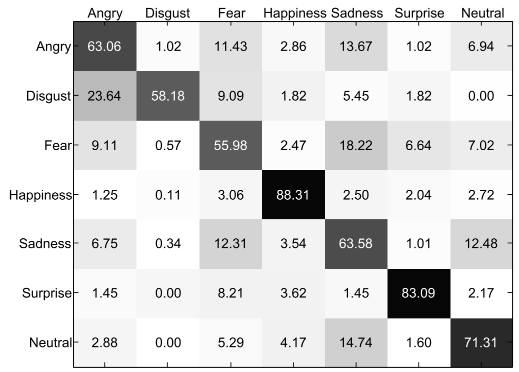


Fig. 4 Training loss w.r.t. the iteration number on two databases.

Table 2 Comparison of facial expression recognition accuracies on private test set of FER-2013.

Method	Accuracy (%)
Winner of FER-2013: CNN with L2-SVM loss [13]	71.16
First runner-up [13]	69.27
Second runner-up [13]	68.82
Multi-scale CNNs [16]	71.8
VGG-Face (finetune)	69.59
VGG-Face & BLSTM	71.04
JCBLSTM	**71.99**



	Angry	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
Angry	63.06	1.02	11.43	2.86	13.67	1.02	6.94
Disgust	23.64	58.18	9.09	1.82	5.45	1.82	0.00
Fear	9.11	0.57	55.98	2.47	18.22	6.64	7.02
Happiness	1.25	0.11	3.06	88.31	2.50	2.04	2.72
Sadness	6.75	0.34	12.31	3.54	63.58	1.01	12.48
Surprise	1.45	0.00	8.21	3.62	1.45	83.09	2.17
Neutral	2.88	0.00	5.29	4.17	14.74	1.60	71.31

Fig. 5 Confusion matrix of JCBLSTM on FER-2013.

multi-scale CNNs. From the confusion matrix shown in Fig. 5, disgust and angry, fear and sadness are easily confusing facial expressions. Given that we achieve the competitive result without using any extra data, our method demonstrates the importance and effectiveness of incorporating spatial relation information with facial texture features in facial expression recognition task.

4. Conclusions

In this letter we propose the JCBLSTM framework to address the facial expression recognition task based on the consideration that there is spatial relationship information contained in different regions of the image which are worth exploring. We employ VGG-Face convolutional module and bidirectional LSTM to extract discriminative middle-level texture features and spatial dependencies in the constructed feature sequence. Experiments on two facial databases demonstrate the effectiveness and superiority of our proposed method. In the future, we can investigate the dependency within images in other tasks such as scene recognition and image captioning. Besides that we would like to extend our method to 3D version such that dependency information can be explored between sequential video frames.

References

[1] J. Yan, W. Zheng, M. Xin, and J. Yan, "Facial expression recognition based on sparse locality preserving projection," IEICE Trans. Fundamentals, vol.E97-A, no.7, pp.1650–1653, 2014.

[2] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," IEEE Transactions on Affective Computing, vol.5, no.1, pp.71–85, 2014.

[3] G. Lyu, H. Yin, X. Yu, and S. Luo, "A local characteristic image restoration based on convolutional neural network," IEICE Trans. Inf. & Syst., vol.E99-D, no.8, pp.2190–2193, 2016.

[4] W. Han, X. Zhang, M. Sun, L. Li, and W. Shi, "An improved supervised speech separation method based on perceptual weighted deep recurrent neural networks," IEICE Trans. Fundamentals, vol.E100-A, no.2, pp.718–721, 2017.

[5] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," IEEE Trans. Multimedia, vol.18, no.12, pp.2528–2536, 2016.

[6] X. Shi, Z. Chen, H. Wang, D.Y. Yeung, W.K. Wong, and W.C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," Advances in Neural Information Processing Systems, pp.802–810, 2015.

[7] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.18–26, 2015.

[8] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3367–3375, 2015.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol.9, no.8, pp.1735–1780, 1997.

[10] G. Huang, Z. Liu, L.V.D. Maaten, and K.Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2261–2269, 2017.

[11] O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," BMVC, p.6, 2015.

[12] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi--pie," Image and Vision Computing, vol.28, no.5, pp.807–813, 2010.

[13] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," International Conference on Neural Information Processing, Neural Networks, vol.64, pp.59–63, 2015.

[14] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," Computer Vision and Image Understanding, vol.115, no.4, pp.541–558, 2011.

[15] J. Wu, Z. Lin, W. Zheng, and H. Zha, "Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition," Neurocomputing, vol.239, pp.143–152, 2017.

[16] S. Zhou, Y. Liang, J. Wan, and S.Z. Li, "Facial expression recognition based on multi-scale cnns," Chinese Conference on Biometric Recognition, vol.9967, pp.503–510, Springer, 2016.