LETTER

Unconstrained Facial Expression Recognition Based on Feature Enhanced CNN and Cross-Layer LSTM*

Ying TONG[†], **Rui CHEN**[†a)], ***Nonmembers**, **and*** **Ruiyu LIANG**[†], ***Member***

SUMMARY LSTM network have shown to outperform in facial expression recognition of video sequence. In view of limited representation ability of single-layer LSTM, a hierarchical attention model with enhanced feature branch is proposed. This new network architecture consists of traditional VGG-16-FACE with enhanced feature branch followed by a cross-layer LSTM. The VGG-16-FACE with enhanced branch extracts the spatial features as well as the cross-layer LSTM extracts the temporal relations between different frames in the video. The proposed method is evaluated on the public emotion databases in subject-independent and cross-database tasks and outperforms state-of-the-art methods.

key words: *facial expression recognition, video sequence, long short-term memory, feature extraction*

1. Introduction

Automatic facial expression recognition has made significant progress in the past two decades, and most previous databases and studies are limited to posed facial behavior under controlled conditions. But in real world, a large amount of images from different events and social gatherings in unconstrained environments have been captured [1], [2]. It brings challenges and opportunities for Facial Expression Recognition (FER).

Traditional FER algorithms use handcrafted features for feature extraction, such as LBP (Local Binary Patterns), HOG (Histogram of Oriented Gradients), LPQ (Local Phase Quantization), PCA, and etc. Since the handcrafted features are extracted for specific application, they often lack required generalizability in cases where there is high variation in lighting, views, resolution, subjects' ethnicity, etc. Fortunately, the emerging deep learning techniques have advanced unconstrained FER to a new state-of-the-art [3]. Moez et al. presented the first deep Convolution Neural Network (CNN) to automatically recognize facial expression by learning features [4]. Then, Yao et al. presented several CNN models to illustrate the correlation between facial expression features and facial expression recognition [5]. Connie T et al. [6] used a hybrid CC-SIFT network to improve the accuracy of expression recognition. By combining CNN and SIFT, the hybrid classifier is established, which can have a good recognition effect on small samples and the recognition rate achieves 99.4% on the CK+ database. To improve the recognition speed, Jeon et al. [7] used the HOG feature to detect the face, and used CNN to extract the deep feature, achieving 70.7% recognition rate and 6.5fps speed on FER-2013 database.

The above methods mainly consider still images independently while ignore the temporal relations of the consecutive frames in a video sequence which are essential for recognizing subtle changes in the appearance of facial images especially in transiting frames between emotions. PPDN (Peak-Pilot Deep Network) [8] was presented to supervise the intermediate feature responses for a sample of non-peak expression (hard sample) of the same type and from the same subject. Based on PPDN, Yu et al. [9] proposed a deeper cascaded peak-piloted network to enhance the discriminative ability of the learned features and employed an integration training method called cascade fine-tuning to avoid overfitting. Jung et al. [10] proposed a joint fine-tuning network method based on two different models to improve the recognition accuracy. One is used to extract time-varying features from the video sequence, the other is used to extract geometric shape changing features from the facial key points of a single frame image.

Recently, a hybrid network which combines CNN and long short term memory (LSTM) network is applied to model the temporal and spatial changes of facial expression in video. By combining the powerful perceptual vision representations learned from CNNs with the strength of LSTM for variable-length inputs and outputs, Jain et al. [11] proposed a both spatially and temporally deep model which cascades the outputs of CNNs with LSTMs for various vision tasks involving time-varying inputs and outputs. However, the algorithm was only validated on laboratory-controlled databases, such as CK+ and MMI. In this paper, we propose an end-to-end framework which includes an enhanced CNN and a cross-layer LSTM for unconstrained FER, named as ECNN-LSTM. To avoid heavy computing overhead and improve the recognition rate, we widen the CNN width instead of increasing the network depth. The cross-layer LSTM network is used to obtain temporal features which helps to reduce the risk of gradient disappearance. The cross-layer structure can ensure the effective transmission of relevant information between video frames,

Manuscript received April 30, 2020.
Manuscript publicized July 30, 2020.

[†] The author is with Dept. of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing, 211167, China.

[*] This work is financially supported by the Natural Science Foundation of China (61703201), Natural Science Foundation of Jiangsu Province (BK20170765), Key Laboratory of meteorological detection and information processing in Jiangsu Province (KDXS1503), and Nanjing Institute of Technology fund for Young Scholar (CKJB201602).

a) E-mail: chenrui@njit.edu.cn

DOI: 10.1587/transinf.2020EDL8065

Copyright © 2020 The Institute of Electronics, Information and Communication Engineers

and obtain the accurate inter-frame temporal features.

2. Proposed Framework

2.1 Framework

The overall framework of our proposed ECNN-LSTM is depicted in Fig. 1, including a feature-enhanced CNN module and a cross-layer LSTM module. The former is a deep hierarchical spatial feature extractor and the latter is a temporal module that characterizes temporal information. These two modules are cascaded for end-to-end training which can effectively improve the recognition ability of unconstrained facial expression features. Finally, the learned deep semantic features are mapped into the sample tag space by the fully connected layer for classification.

2.2 Enhanced CNN

The enhanced CNN We use VGG-16 [12] network as the backbone network of CNN. Due to the limited layers of VGG-16, it has poor recognition rate when processing the unconstrained facial expression data. The samples in the training set are interfered by many factors, such as illumination, posture changes, occlusions, accessories, and so on. Moreover, the degree of the same emotion of the subjects is also different due to the individual culture. Considering of complexity, VGG-16 network is widened instead of increasing the network depth. Specifically, we introduce an enhanced branch into the backbone CNN network to integrate different level features. The structure of the enhanced branch is depicted in Fig. 2.

As we can see from Fig. 2, the enhanced branch includes 5 layers. The first convolution layer uses 7 × 7 kernel for larger receptive field to obtain more spatial features. The second convolution layer uses 1 × 1 kernel to compress the high-dimensional features for further integrating the features and reduce the complexity. The batch normalization layer is used to normalize the features to improve the stability of the feature distribution and speed up the learning of the model. The flatten layer is to quantize multi-dimensional features in one dimension for connection with the full connection layer. The detail parameter setting of each layer is shown in Table 1.

The enhanced CNN uses large input data dimension for spatial learning, and utilizes transfer learning with pretrained weights from VGG-Face model [13] which was trained on LFW database.

2.3 Cross-Layer LSTM Network

To preserve the temporal dimension as its dynamics is crucial for recognizing facial movements, we use a cross-layer LSTM to learn the sequential input, which is shown in Fig. 3.

According to the structural characteristics of LSTM, the input needs to be sequence information. At the same time, the end-to-end training needs CNN model provide at least two consecutive facial features. Then, we input n consecutive facial images at a time. Each facial image shares the same CNN weights for feature extraction. Because the complexity and parameters of the model are increased by processing multiple face images at a time, we set n = 10 to avoid memory overflow. Furthermore, the video segment in the same video overlaps with the front and back segments by 5 frames to augment the training data and strengthen the model learning ability. As shown in Fig. 2, the CNN outputs a feature vector with length of 4,096, then the input data dimension of the first LSTM layer is 104,096.

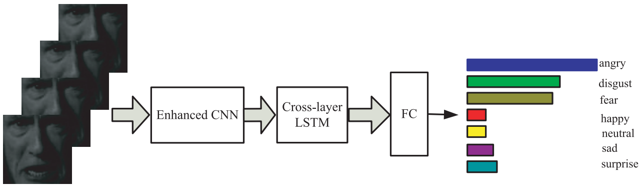

Fig. 1 Proposed ECNN-LSTM framework

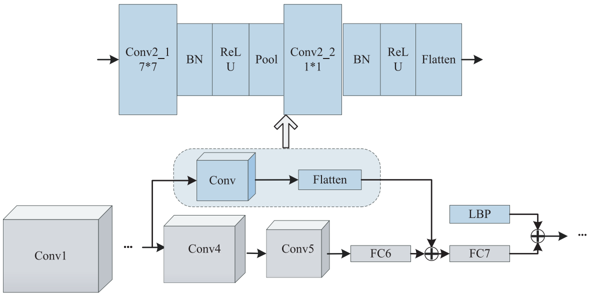

Fig. 2 The structure of feature enhanced CNN

Table 1 Parameters setting of the enhanced branch

Layer	Output size	Kernel size
Conv2_1	28×28×1024	7×7×1024
BN1	28×28×1024	-
Pooling	13×13×1024	-
Conv2_2	13×13×14	1×1×14
BN2	13×13×14	-
Flatten	2366	-

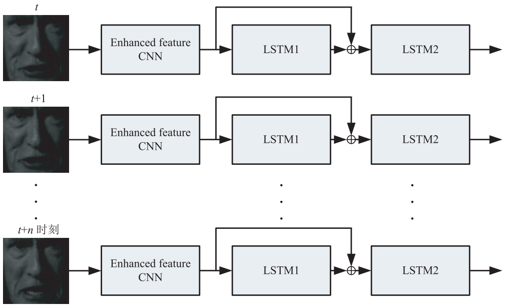

Fig. 3 Cross-layer LSTM network

3. Experiments

3.1 Databases

CK+ [14]: The Extended CohnKanade (CK+) database is the most extensively used laboratory-controlled database for evaluating FER systems. It contains 593 video sequences from 123 subjects, where 327 sequences from 118 subjects are labeled with seven basic expression labels.

AFEW [15]: The Acted Facial Expressions in the Wild (AFEW) database contains video clips collected from different movies with spontaneous expressions, various head poses, occlusions and illuminations. Samples are labeled with seven expressions: anger (An), disgust (Di), fear (Fe), happiness (Ha), sadness (Sa), surprise (Su) and neutral (Ne). The AFEW 7.0 is divided into three data partitions in an independent manner in terms of subject and movie/TV source: Train (773 samples), Val (383 samples) and Test (653 samples), which ensures data in the three sets belong to mutually exclusive movies and actors.

SFEW [14]: The Static Facial Expressions in the Wild (SFEW) was created by selecting static frames from the AFEW database by computing key frames based on facial point clustering. The SFEW 2.0 is divided into three sets: Train (958 samples), Val (436 samples) and Test (372 samples).

3.2 Cross-Layer LSTM Network Experiment

The CNN-LSTM network is trained on Keras platform and VGG-16 model preloads VGG-16-FACE weights. The training data is from the video frame provided by AFEW database. Since AFEW is from the EmotiW challenge, the test set does not have the corresponding expression label. So we divide the training set into training and validation sets according to 8 : 2, and take the validation set as the test set during the training process.

We use the classic VGG-16 network, and the results of LSTM with different layers and parameters on AFEW database are shown in Table 2. The first two lines are the results of single layer LSTM, and the last three lines are the results of two layers of LSTM. The values in brackets represent the output eigenvector dimensions outputting of each LSTM. As we can see, the maximum F1-score of two-layer LSTM network is 0.3279, which is better than that of single layer LSTM network (i.e. 0.2954). For two-layer LSTM, we set different output parameters respectively, and the experimental results show that CNN-LSTM (2048, 2048) has the highest F1-score. So, we adopt LSTM (2048, 2048) for the cross-layer LSTM network.

For the cross-layer LSTM, we carried out end-to-end training and experiments on AFEW and CK+. The results are shown in Table 3. It can be seen that the end-to-end training outperform than non-end-to-end training (i.e. independent training). The cross-layer LSTM network can further improve the accuracy of unconstrained FER.

Table 2 Performance of LSTM with different layers and parameters on AFEW database

LSTM layer	model	F1-score	Accuracy
Single-layer LSTM	CNN-LSTM(2048)	0.2895	33.69%
	CNN-LSTM(3000)	0.2954	32.88%
Two-layer LSTM	CNN-LSTM(3000,3000)	0.3069	**34.77%**
	CNN-LSTM(2048,2048)	**0.3279**	34.50%
	CNN-LSTM(2048,1024)	0.2950	34.23%

Table 3 Performance of cross-layer LSTM on AFEW and CK+ databases

Database	Model	Accuracy
AFEW	Non-end-to-end CNN-LSTM	34.50%
	End-to-end CNN-LSTM	38.57%
	End-to-end CNN & cross-layer LSTM	39.89%
CK+	Non-end-to-end CNN-LSTM	95.14%
	End-to-end CNN-LSTM	95.71%
	End-to-end CNN & cross-layer LSTM	95.92%

Table 4 Confusion matrix of the end-to-end CNN & cross-layer LSTM network on CK+ (%)

	An	Di	Fe	Ha	Ne	Sa	Su
An	**97.32**	2.31	0.00	0.00	2.07	3.16	0.00
Di	0.00	**91.54**	0.00	0.00	0.00	0.00	0.00
Fe	2.68	2.31	**97.00**	0.00	0.00	0.00	0.00
Ha	0.00	0.00	0.00	**94.70**	0.00	3.16	0.00
Ne	0.00	3.85	0.00	2.65	**97.93**	0.00	0.00
Sa	0.00	0.00	0.00	2.65	0.00	**93.67**	0.00
Su	0.00	0.00	3.00	0.00	0.00	0.00	**100**

Table 5 Confusion matrix of the end-to-end CNN & cross-layer LSTM network on AFEW (%)

	An	Di	Fe	Ha	Ne	Sa	Su
An	**54.68**	22.50	36.36	6.35	18.33	16.67	40.00
Di	1.56	**12.50**	2.27	4.76	8.33	3.33	2.22
Fe	9.38	7.50	**18.18**	1.59	10.00	8.33	4.44
Ha	7.81	20.00	15.91	**80.95**	15.00	21.67	20.00
Ne	3.13	17.50	11.36	6.35	**35.00**	16.67	13.33
Sa	15.63	17.50	13.64	0.00	10.00	**30.00**	2.22
Su	7.81	2.50	2.27	0.00	3.33	3.33	**17.78**

The confusion matrices of the proposed method on CK+ and AFEW are shown in Table 4 and Table 5, respectively. It can be seen that the results on CK+ are better than on AFEW. For “surprise”, only 17.78% are classified correctly, while 40% are classified incorrectly as “angry”. This is due to the people’s mood is a mixture of many emotions, such as “anger”, “disgust” and “sadness” are usually accompanied by each other, and the facial morphological changes of “fear”, “surprise” and “happiness” have some similarities. The multimodal expression recognition method can help improve the accuracy.

3.3 ECNN and Cross-Layer LSTM Experiment

To evaluate the proposed ECNN & cross-layer LSTM network, we have done experiments on CK+ and AFEW respectively, and the results are shown in Table 6. (FC1, 5 × 5)

Table 6 Performance of ECNN & cross-layer LSTM on CK+ and AFEW

Database		F1-score	Accuracy
CK+	ECNN & cross-layer LSTM (FC1, 5×5)	0.9425	94.74%
	ECNN & cross-layer LSTM （FC1,7×7）	**0.9718**	**97.47%**
	ECNN & cross-layer LSTM (FC2, 5×5)	0.9551	95.68%
	ECNN & cross-layer LSTM (FC2, 7×7)	0.9623	96.53%
AFEW	Baseline	—	38.81%
	ECNN & cross-layer LSTM (FC1, 5×5)	0.3733	40.16%
	ECNN & cross-layer LSTM （FC1,7×7）	**0.3816**	**41.25%**
	ECNN & cross-layer LSTM (FC2, 5×5)	0.3514	39.34%
	ECNN & cross-layer LSTM (FC2, 7×7)	0.3763	40.44%

Table 7 Performance comparison on CK+ and SFEW

Database	Method	Accuracy (%)
CK+	3DCNN-DAP[17]	92.35
	STC-NLSTM [18]	93.88
	DTAGN[10]	96.43
	The proposed	**97.47**
SFEW	3DCNN-DAP[17]	24.7
	STC-NLSTM [18]	31.73
	DTAGN[10]	26.14
	Inception[12]	47.7
	The proposed	**54.37**

represents the enhanced branch is fused with FC1 layer, and the convolution kernel size of conv1 in enhanced branch is $5 \times 5 \times 1024$ (shown in Fig. 2 and Table 1). Similarly, (FC2, 7×7) represents the enhanced branch is fused with FC2 layer, and the convolution kernel size of conv1 in enhanced branch is $7 \times 7 \times 1024$. It can be seen that ECNN & cross-layer LSTM (FC1, 7×7) has the best performance, outperform the official baseline 2.44% on AFEW.

The performance comparisons of recognition accuracy with the state-of-art on CK+ and SFEW databases are shown in Table 7.

4. Conclusion

In this letter, we present a hybrid framework for facial expression recognition of video sequence. By combining the enhanced CNN and cross-layer LSTM, we obtain better performance in terms of F1-score and Accuracy. Due to the samples in AFEW and SFEW databases are gathered in unconstrained environments, our future work will focus on optimizing the proposed framework and image preprocessing to further improve recognition accuracy and speed.

References

[1] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.12, pp.1424–1445, 2000.

[2] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.1, pp.39–58, 2009.

[3] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, arXiv: 1804.08348v2 [cs.CV] 22 Oct. 2018.

[4] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification," Procedings of the British Machine Vision Conference 2012, pp.124.1–124.12, 2012.

[5] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing AU-Aware Facial Features and Their Latent Relations for Emotion Recognition in the Wild." Proc. 2015 ACM on International Conference on Multimodal Interaction - ICMI '15, pp.451–458, 2015.

[6] T. Connie, M. Al-Shabi, W.P. Cheah, and M. Goh, "Facial Expression Recognition Using a Hybrid CNN-SIFT Aggregator," Multidisciplinary Trends in Artificial Intelligence, Lecture Notes in Computer Science, vol.10607, pp.139–149, Springer International Publishing, Cham, 2017.

[7] J. Jeon, J.-C. Park, Y. Jo, C. Nam, K.-H. Bae, Y. Hwang, and D.-S. Kim, "A Real-time Facial Expression Recognizer using Deep Neural Network." Proc. 10th International Conference on Ubiquitous Information Management and Communication - IMCOM '16, pp.1–4, 2016.

[8] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted Deep Network for Facial Expression Recognition." Computer Vision – ECCV 2016, Lecture Notes in Computer Science, vol.9906, pp.425–442, Springer International Publishing, Cham, 2016.

[9] Z. Yu, Q. Liu, and G. Liu, "Deeper Cascaded Peak-piloted Network for Weak Expression Recognition." The Visual Computer, vol.34, no.12, pp.1691–1699, 2018.

[10] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint Fine-tuning in Deep Neural Networks for Facial Expression Recognition." 2015 IEEE International Conference on Computer Vision (ICCV), pp.2983–2991, 2015.

[11] D.K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," Pattern. Recogn. Lett., 2017.

[12] A. Mollahosseini, D. Chan, and M.H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.1–10, 2016.

[13] O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," Proc. British Machine Vision Conference 2015, pp.41.1–41.12, 2015.

[14] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp.94–101, 2010.

[15] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," IEEE Multimedia Mag., vol.19, no.3, pp.34–41, 2012.

[16] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp.2106–2112, 2011.

[17] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," Computer Vision – ACCV 2014, Lecture Notes in Computer Science, vol.9006, pp.143–157, Springer International Publishing, Cham, 2015.

[18] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," Neurocomputing, vol.317, pp.50–57, 2018.