Journal of Parallel and Distributed Computing 131 (2019) 97–102

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

Automatic social signal analysis: Facial expression recognition using difference convolution neural network

Jingying Chen, Yongqiang Lv, Ruyi Xu *, Can Xu

National Engineering Laboratory for Educational Big Data, Central China Normal University, Luoyu Road 152, Wuhan, China
National Engineering Research Center for E-Learning, Central China Normal University, Luoyu Road 152, Wuhan, China

HIGHLIGHTS

- The neutral and the fully expression are picked out from the facial expression sequence by a binary CNN.
- An end-to-end DCNN learns the difference between the neutral and the fully expression for automatic FER.
- Our method obtains competitive or even better performance on 2 benchmark databases.

ARTICLE INFO

Article history:
Received 3 December 2018
Received in revised form 29 March 2019
Accepted 19 April 2019
Available online 2 May 2019

MSC:
00-01
99-00

Keywords:
Social signal analysis
Facial expression recognition
Deep learning
Difference convolution neural network

ABSTRACT

Facial expression is one of the most powerful social signals for human beings to convey emotion and intention, hence automatic facial expression recognition (FER) has wide applications in human–computer interaction and affective computing, it has attracted an increasing attention recently. Researches in this field have made great progress especially with the development of deep learning method. However, FER remains a challenging task due to individual differences. To address the issue, we propose a two-stage framework based on Difference Convolution Neural Network (DCNN) inspired by the facial expression's nonstationary nature. In the first stage, the neutral expression frame and fully expression frame are automatically picked out from the facial expression sequences using a binary Convolution Neural Network (CNN). Then in the second stage, an end-to-end DCNN is proposed to classify the six basic facial expressions using the difference information between the neutral expression frame and the fully expression frame. Experiments have been conducted on the CK+ and BU-4DFE datasets, and the results show that the proposed framework delivers a promising performance (95.4% on the CK+ dataset and 77.4% on the BU-4DFE). Moreover, the proposed method is also successfully applied to analyze the student's affective state in an E-learning environment which suggests that it has strong potential to analyze nonstationary social signals.

1. Introduction

As one of the most powerful social signals, facial expression helps to understand each other's internal emotion in communication. Psychologists conclude that people try to convey the same emotion with the similar facial expression even though from diverse race, culture and religion. This conclusion provides possibility of developing the emotional computing system by facial expression.

To further analyze the potential meanings of different facial expressions, facial expressions are divided into six basic categories: angry, disgust, fear, happiness, sadness and surprise [10]. Nowadays, automatic FER has many applications. In the field of education, the student's affective state can be understood by automatic FER. It helps teachers understand student's learning interest so as to provide appropriate teaching strategies to improve the effectiveness of teaching. For example, an intelligent computer-assisted system is developed to understand children's affective state by FER and then provide appropriate support to improve their social communication skills [2]. A hybrid intelligence-aided approach is presented to develop an affect-sensitive e-learning system that recognizes the student's affective state using multimodal information via hybrid intelligent approaches, e.g., head pose, eye gaze tracking, facial expression recognition, physiological signal processing and learning progress tracking [5]. An affect-enhanced student modeling framework is proposed to leverage facial expression tracking for game-based learning, which generates predictive models of student learning and student engagement [24]. To reduce the redundancy information caused by meaningless and uncommon facial expression in the class, Zhou et al. [32] classified learning affective state into

[*] Corresponding author.
E-mail address: xuruyi@mail.ccnu.edu.cn (R. Xu).

https://doi.org/10.1016/j.jpdc.2019.04.017
0743-7315/© 2019 Elsevier Inc. All rights reserved.

three basic classes: positive, neutral, and negative states. This paper focuses on both basic emotion recognition and learning affective state analysis based on facial expression.

Numerous methods for automatic FER have been developed and can be roughly divided into two classes: traditional method and deep learning method. Traditional method primarily consists of two steps: feature extraction and classification. In the feature extraction stage, some elaborate features are designed to describe the facial shape or texture changes caused by expression, e.g., the points detected using the Supervised Descent Method (SDM) [28] are employed to describe the shape changes of facial organ [15]; and the texture features, such as LDTP [23], Hog [1], STTM [16], and LBP [22], are introduced to describe the texture change around the area of facial muscle movement. Moreover, some works aim to analyze the temporal dependency between frames to further improve the performance, such as LBP-TOP [11]. However, the computational complexity is high in these methods because the whole sequence needs to be computed. As a compromise, several works attempt to use the difference features to describe the changes in shape and texture [6,9]. Despite a given reference frame (or neutral frame) is necessary, these methods have an advantage that can effectively eliminate individual differences.

In the classification stage, diverse machine learning methods are presented to train the classifier using the extracted features. Jiang et al. [14] proposed a Sparse Representation Classification (SRC) method for automatic FER to overcome the high-level noise interference. Du and Hu [9] improved the discriminating capability of classification and regression tree (CRT) for FER by maximizing intra class purity and inter class distance simultaneously. Dapogny et al. [7] addressed FER under a frame of Pairwise Conditional Random Forests (PCRF) that extended random forests to learn the spatio-temporal patterns upon pairs of images.

Different from the traditional methods, deep learning method, e.g., Convolution Neural Network (CNN), adopts an end-to-end mode to train a very deep network structure with millions of parameters, which adaptively learns right features from the massive data automatically without the hand-crafted features. Khorrami et al. [18] discussed the corresponding relationships between the high-level features learned by expression CNN and the facial action units (AUs). To address the problem of insufficient expression data, data augment and transfer learning are two common means, e.g., Lopes et al. [20] adopted a set of pre-processing operations to augment data and discussed the presentation order of the samples during training. Tang et al. [26] trained ten facial AUs separately by fine-tuning the Vgg-Face model [21]. To improve the performance of FER, Yang et al. [29] fused two CNNs trained separately using the gray images and LBP images. Hasani and Mahoor [13] combined the CNN with Conditional Random Fields (CRF) to capture the spatial relation within facial images and the temporal relation between the image frames. Although the deep learning method have made some progress, few deep learning methods consider the influence of the individual differences, that is, appearance dissimilarity of the different subject typically overwhelms the expression dissimilarity of the same subjects, considering the facial expression's nonstationary nature, we attempt to eliminate the individual differences using the neutral expression frame and fully expression frame based on deep learning method in this paper.

In the work [6], the deep representation features named DPND were presented to eliminate individual differences. However, the intermediate layer of pre-training CNN is used to represent the key-frames, which is not an end-to-end mode. Therefore, a two-stage framework based on Difference Convolution Neural Network (DCNN) is proposed for automatic FER in this paper. The flowchart of the proposed method is shown in Fig. 1. In the first stage, a binary Convolution Neural Network (CNN) is introduced to automatically pick out the neutral expression frame and fully expression frame; in the second stage, the two frames are respectively sent to two branch networks with the same structure that consist of a set of convolution layers and a full connected (FC1) layer. Following the FC1 layer, a difference layer is added to hold the differences information between the neutral expression frame and fully expression frame. Also, the proposed method is applied to assist the analysis of the student's affective state in the E-learning environment. The main contributions of this paper are reflected as follows:

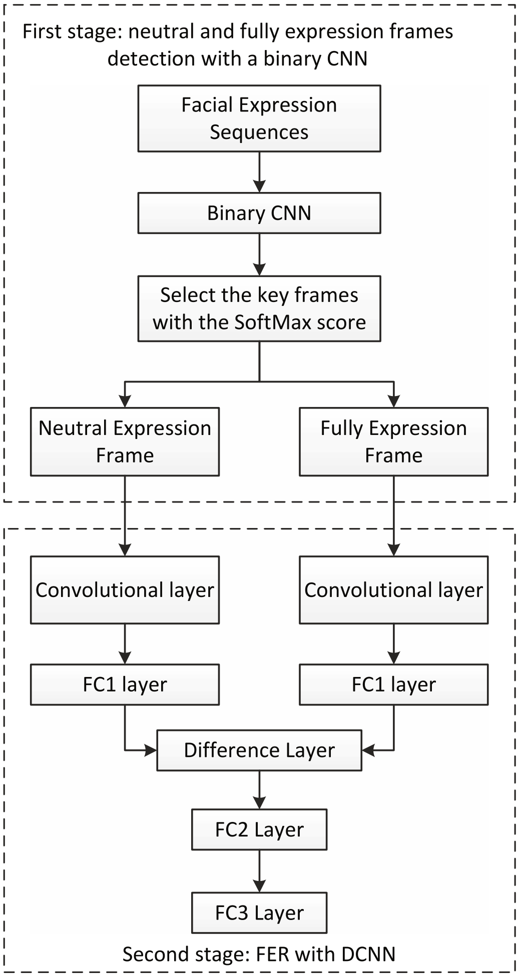

First stage: neutral and fully expression frames detection with a binary CNN
Facial Expression Sequences
Binary CNN
Select the key frames with the SoftMax score
Neutral Expression Frame
Fully Expression Frame
Convolutional layer
Convolutional layer
FC1 layer
FC1 layer
Difference Layer
FC2 Layer
FC3 Layer
Second stage: FER with DCNN

Fig. 1. The overview of the proposed method.

(1) To get reliable neutral expression frame and fully expression frame, the SoftMax score of the binary CNN is presented to pick out the neutral expression frame and the fully expression frame from the facial expression sequence.
(2) To eliminate the individual difference, an end-to-end DCNN is proposed for automatic FER. The proposed method learns the difference features between the neutral expression and fully expression from the pair-wise data automatically.

The rest of this paper is organized as follows. The details of the proposed DCNN method for automatic FER are presented in Section 2. The experimental setup is described in detail, and the experiment results are given in Section 3. Section 4 concludes the paper.

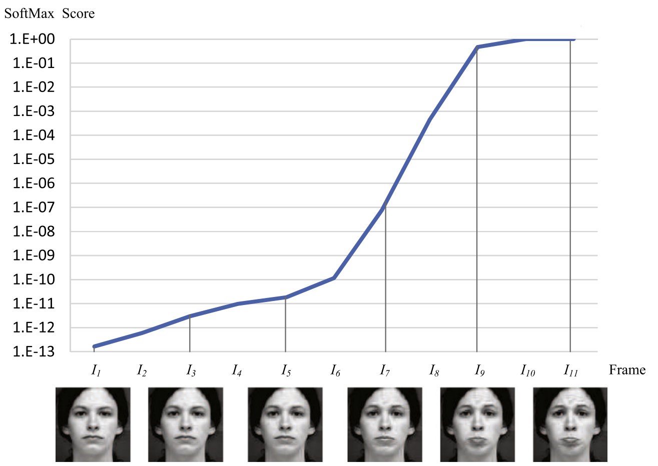


Fig. 2. An example of the test result on a facial sequence of the CK+ dataset.

2. The proposed method

In this section, we describe the proposed method in details. First, a binary Convolution Neural Network (CNN) is constructed to automatically pick out the neutral expression frame and fully expression frame from the sequences. Second, a DCNN is proposed to classify the facial expression using the difference information between the neutral expression frame and fully expression frame.

Neutral and fully expression frames detection with binary CNN. This subsection focuses on the neutral and fully expression frame selection. The neutral and fully expression can be detected according to the expression intensity. The intensity of the neutral expression is the lowest, the transient expression is middle, and the fully expression is the highest. Although there are some works [8] that divide the expression intensity into three levels, the accuracy of the frame detection is not satisfactory due to the errors from the manual annotations and existence of critical state.

Since standard annotations of the neutral and fully expression frames are available, it is convenient to train the binary classifier. Although the binary classifier can distinguish neutral expressions and fully expressions, it cannot pick out them directly from expression sequences. In the proposed method, the SoftMax score of binary CNN is used to pick out the neutral and fully expression frames. The binary CNN adopts the Vgg-Face framework [21] to train the model due to its excellent performance in face-related visual tasks. Vgg-Face has 16 weight layers including 13 convolutional layers and 3 fully connected layers (FC1, FC2 and FC3) and up to 138 million parameters. Following the weight layers, the SoftMax layer calculates the probability S_i that the sample belongs to the i-th class:

$$S_i = \frac{exp(x_i)}{\sum_{i=1}^{2} exp(x_i)} \tag{1}$$

where the x_i is the i-th input of SoftMax layer.

To address the issue of insufficient training samples, data augment is implemented with the method introduced by [20]. Then, we train the neutral/fully expression classifier by fine-tuning the Vgg-Face model. The output of the SoftMax layer is a 2-dimensional vector, where S_1 and S_2 respectively represent the probability that the sample belongs to the neutral expression and fully expression. An example of the SoftMax score (S_2) on a facial sequence of the CK+ dataset is shown in Fig. 2.

Generally, the classifier score reflects the prediction confidence of samples belonging to the specified class. Although the score cannot reflect completely the intensity changes of facial expression sequence [19], it is still practical to pick out the frame with the lowest score as the neutral expression frame and that with the highest score as the fully expression frame when the sequences only contain single neutral expression frame and single fully expression frame. For in-the-wild sequences with multiple fully expression frames (e.g., the videos recorded the students' learning processing), we suppose that the intensity level of the expression changes continuously. Hence, multiple local maximum values of the SoftMax score are detected as the fully expression frames and global minimum value of the SoftMax score is detected as the neutral expression frame.

Facial expression recognition based on DCNN. In this subsection, the details of the proposed DCNN are presented for facial expression classification. The architecture of the DCNN is shown in Fig. 3.

The network consists of two branches (F-net and N-net) with the same structure, which consists of multiple Convolution Layers and FC1 layer. The Convolution Layers of F-net are used to learn the features a^F from the fully expression frames while that of the N-net are used to learn the features a^N from the neutral expression frames. Then the FC1 layers of two branches are calculated respectively as follows:

$$\begin{aligned} z^F &= W^F a^F + b^F, c^F = f(z^F) \\ z^N &= W^N a^N + b^N, c^N = f(z^N) \end{aligned} \tag{2}$$

where W^F and W^N are the network weights; b^F and b^N are the biases; $f(\bullet)$ is the activation function. c^F and c^N are the outputs of F-net and N-net respectively. Subsequently, the difference layer is introduced to extract the difference features a^{diff}:

$$a^{diff} = c^F - c^N = f(z^F) - f(z^N) \tag{3}$$

Then the next two fully connected layers (FC2 and FC3) are used to classify the difference features. In fact, the difference layer can be added behind any fully connected layer. Follow the [6], the difference layer added behind the FC1 layer has the best performance. The weight layers except the FC3 layer have the same size as the Vgg-Face model, which is convenient for training the model by transfer learning.

In the training process, all the weight layers are initialized using the Vgg-Face model. The two branches of network have the same value at every position at the beginning of the training. Then all the weights are set to be trainable and weights of the l-th layer are updated as the formula (4):

$$\begin{aligned} W^{l+} &= W^l - \eta \frac{\partial J(W,b)}{\partial W} \\ b^{l+} &= b^l - \eta \frac{\partial J(W,b)}{\partial b} \end{aligned} \tag{4}$$

where η is the learning rate, the gradient of the FC1 layer is calculated as the formula (5):

$$\begin{aligned} \frac{\partial J(W,b)}{\partial W^F} &= \frac{\partial J(W,b)}{\partial z^{FC2}} \frac{\partial z^{FC2}}{\partial a^{diff}} \frac{\partial a^{diff}}{\partial z^F} \frac{\partial z^F}{\partial W^F} \\ &= (W^{FC2})^T \delta^{FC2} f'(z^F)(a^F)^T \\ \frac{\partial J(W,b)}{\partial W^N} &= \frac{\partial J(W,b)}{\partial z^{FC2}} \frac{\partial z^{FC2}}{\partial a^{diff}} \frac{\partial a^{diff}}{\partial z^N} \frac{\partial z^N}{\partial W^N} \\ &= -(W^{FC2})^T \delta^{FC2} f'(z^N)(a^N)^T \\ \frac{\partial J(W,b)}{\partial b^F} &= \frac{\partial J(W,b)}{\partial z^{FC2}} \frac{\partial z^{FC2}}{\partial a^{diff}} \frac{\partial a^{diff}}{\partial z^F} \frac{\partial z^F}{\partial b^F} = (W^{FC2})^T \delta^{FC2} f'(z^F) \\ \frac{\partial J(W,b)}{\partial b^N} &= \frac{\partial J(W,b)}{\partial z^{FC2}} \frac{\partial z^{FC2}}{\partial a^{diff}} \frac{\partial a^{diff}}{\partial z^N} \frac{\partial z^N}{\partial b^N} = -(W^{FC2})^T \delta^{FC2} f'(z^N) \end{aligned} \tag{5}$$

where W^{FC2} represents the weights of FC2 layer, $\delta^{FC2} = \frac{\partial z^{FC2}}{\partial a^{diff}}$ is the residual of FC2 layer.

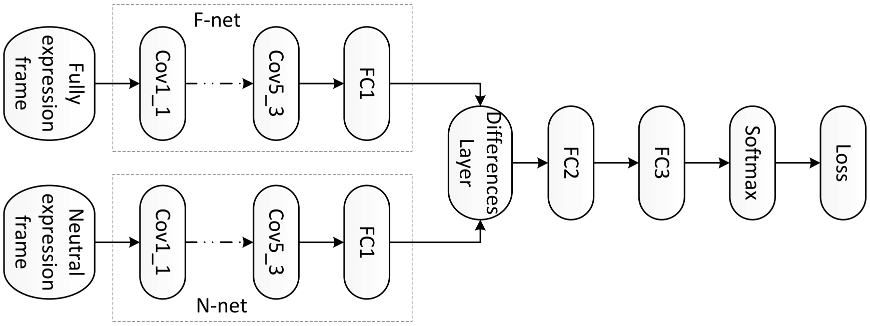


Fig. 3. The architecture of the DCNN.

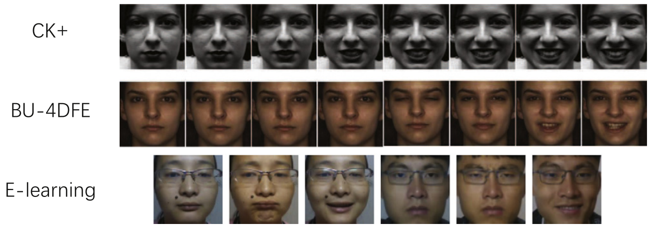


Fig. 4. Exemplar expression sequences in the CK+ and BU-4DFE datasets.

Table 1
The accuracy of fully/neutral frame detection on CK+ and BU-4DFE respectively.

Dataset	The accuracy of detection
CK+	99.02%
BU-4DFE	98.35%

Table 2
Comparison of DCNN using different pair-wise data on CK+ and BU-4DFE respectively.

Methods	CK+	BU-4DFE
DCNN(manually)	97.5%	79.4%
DCNN(automatically)	95.4%	77.4%

In conclusion, the proposed two-stage framework is equivalent to a cascaded model that recognizes the facial expression in a coarse-to-fine way, i.e., in the first stage, the neutral expression and fully expression are identified according to the expression intensity; and in the second stage, the proposed DCNN learns the expression features from the neutral frame and the fully frame simultaneously. The face identity appearance contained in the two frames is eliminated by the difference layer of DCNN.

3. Experiments

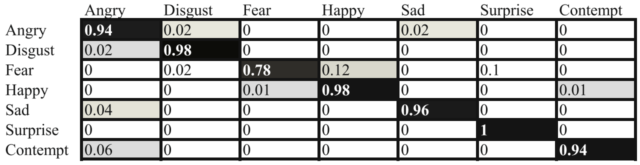
In this section, the proposed method is evaluated on two public datasets: Extended Cohn–Kanade facial expression (CK+) database [27] and Binghamton University 4D Facial Expression (BU-4DFE) dataset [31]. CK+ dataset contains 593 expression sequences from 123 subjects aged from 18 to 30 years old. Each sequence begins with a neutral expression frame and ends at a fully expression frame. BU-4DFE contains 606 3D dynamic facial sequences from 101 subjects aged from 18 to 70 years old. Each subject is requested to perform six basic expressions. But unlike CK+, not all the sequences in the BU-4DFE begin with a neutral expression frame and end at a fully expression frame. Although great progress has been made in EEG-based method [3,4,17,25] to analyze affective state, special device is needed. In this study, the student's affective state in an E-learning environment is analyzed by facial expressions. To this end, an E-learning dataset is built by ourselves, which records the students' facial sequences using a WebCam during the learning processing. The samples are annotated manually as positive, neutral or negative states. 120 sequences from 20 subjects are used in the experiment. Some sample sequences from each dataset are given in Fig. 4.

Settings. In the experiment, 327 sequences in the CK+ dataset with the labels of the seven facial expressions (i.e., six basic expressions and contempt) and 600 sequences in the BU-4DFE sequences are adopted. For BU-4DFE, the neutral and fully frames are manually labeled like CK+. We used the five-fold method for cross-validation without subject overlap between groups: divide all the sequences into five groups, the sequences from the same subject are divided into the same group, each group contains the same number of expressions of all class, of which 4 groups are used for training, the rest group is used for testing. For the fine-tuning binary CNN and DCNN, Stochastic Gradient Descent (SGD) is adopted as the optimization algorithm, the mini-bath size is 32, the total training epochs is 300, the learning rate is set as 1e-4 and decreased by 0.1 after 100 epochs, and momentum is fixed to be 0.9. All the experiments were conducted on a desktop computer with an Intel Core i7 3.6 GHz CPU, 32GB memory, and an Nvidia GeForce GTX TITAN X GPU.

Performance of the fully and neutral expression frame detection. As introduced in Section 2, the fully/neutral expression frame is picked out from the sequence by the binary CNN. The neutral expression frames and fully expression frames in the training set are used to train the binary CNN. Specifically, in the CK+ dataset, the first frames of sequences are the negative samples and the last frames are the positive samples; in the BU-4DFE, the neutral and fully frames manually labeled are as the negative samples and positive samples respectively. Then the neutral/fully expression frames detection is performed on the testing set. To test the recognition accuracy, three experts are invited to annotate the neutral/fully expressions, and the ground truth is determined by the majority of votes. The accuracy of detection is shown in Table 1.

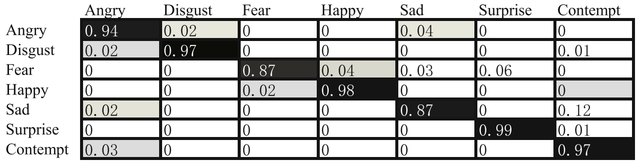
To further demonstrate the performance of the fully and neutral expression frame detection, we compare the recognition accuracy of the proposed DCNN using two different pair-wise data, one of which is detected automatically, and the other of which is labeled manually. The comparison results are shown in Table 2.

The confusion matrices of the proposed method using automatic pair-wise data and manual pair-wise data respectively on the CK+ and BU-4DFE datasets are shown in Figs. 5–8. It demonstrates that: (1) the BU-4DFE dataset has a greater challenge than the CK+ dataset; (2) the proposed method has a highest recognition rate for Happy and a lowest recognition rate for Fear both on two datasets; (3) the proposed method for neutral and fully expression frame detection is effective since the performance is comparable to the DCNN using the manual frames.



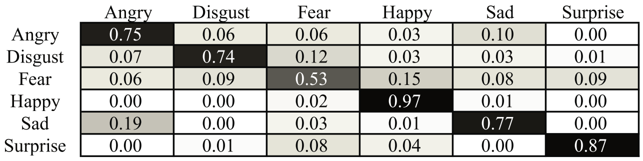
	Angry	Disgust	Fear	Happy	Sad	Surprise	Contempt
Angry	0.94	0.02	0	0	0.02	0	0
Disgust	0.02	0.98	0	0	0	0	0
Fear	0	0.02	0.78	0.12	0	0.1	0
Happy	0	0	0.01	0.98	0	0	0.01
Sad	0.04	0	0	0	0.96	0	0
Surprise	0	0	0	0	0	1	0
Contempt	0.06	0	0	0	0	0	0.94

Fig. 5. Confusion matrix of the DCNN (automatic pair-wise data) on the CK+.



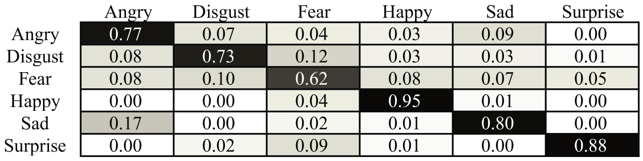
	Angry	Disgust	Fear	Happy	Sad	Surprise	Contempt
Angry	0. 94	0. 02	0	0	0. 04	0	0
Disgust	0. 02	0. 97	0	0	0	0	0. 01
Fear	0	0	0. 87	0. 04	0. 03	0. 06	0
Happy	0	0	0. 02	0. 98	0	0	0
Sad	0. 02	0	0	0	0. 87	0	0. 12
Surprise	0	0	0	0	0	0. 99	0. 01
Contempt	0. 03	0	0	0	0	0	0. 97

Fig. 6. Confusion matrix of the DCNN (manual pair-wise data) on the CK+.



	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	0.75	0.06	0.06	0.03	0.10	0.00
Disgust	0.07	0.74	0.12	0.03	0.03	0.01
Fear	0.06	0.09	0.53	0.15	0.08	0.09
Happy	0.00	0.00	0.02	0.97	0.01	0.00
Sad	0.19	0.00	0.03	0.01	0.77	0.00
Surprise	0.00	0.01	0.08	0.04	0.00	0.87

Fig. 7. Confusion matrix of the DCNN (automatic pair-wise data) on the BU-4DFE.



	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	0.77	0.07	0.04	0.03	0.09	0.00
Disgust	0.08	0.73	0.12	0.03	0.03	0.01
Fear	0.08	0.10	0.62	0.08	0.07	0.05
Happy	0.00	0.00	0.04	0.95	0.01	0.00
Sad	0.17	0.00	0.02	0.01	0.80	0.00
Surprise	0.00	0.02	0.09	0.01	0.00	0.88

Fig. 8. Confusion matrix of the DCNN (manual pair-wise data) on the BU-4DFE.

Comparison with state of the art. We compare the proposed method based on DCNN with State of the Art, including NES [12], CNN+Landmark [30], PCRF [7], Pre-CNN [20], DPND and DPND+DPR [6].

For the CK+ dataset, the experiments for six basic expression recognition and seven expression recognition are conducted respectively with the proposed method. NES method applied neutral-subtracted HOG/SIFT features to predict facial expression. PCRF modeled the dynamic process of facial expression with Pairwise Conditional Random Forests. The two methods attempted to eliminate to individual differences using traditional method. CNN+Landmark method combined CNN and landmark features to recognize 3D facial expression; Pre-CNN applied some preprocessing techniques to learn the expression model only from the fully expression frame. The two methods trained the end-to-end model by deep learning. DPND method trained the SVM classifier using the difference features extracted from the pretraining Vgg-model. And DPND+DPR method combined the deep representation features of the fully frame with DPND to train the expression model. DPND and DPND+DPR use 64 subjects in BU-4DFE, while other methods use 100 subjects. The experiment results are shown in Table 3.

From this table, one can see that the proposed method outperforms state of the art on BU-4DFE. The results on the CK+ for six expression recognition indicate that the accuracy of the proposed method (automatically) is 3.8% higher than that of the DPND, which strongly proves the excellence of the end-to-end model. The accuracy of the proposed method (manually) without any data augment is 1.6% higher than that of the pre-CNN, which validates the ability of our method to eliminate the individual differences. The performance of seven expression recognition on the CK+ is comparable to that of the existing method.

Table 3
Comparison of different methods on CK+ and BU-4DFE respectively.

Methods	CK+		BU-4DFE
	6 basic expressions	7 expressions	6 basic expressions
NES	–	96.0%	74.8%(100)
CNN+Landmark	–		75.9%(100)
PCRF	96.4%	–	76.1%(100)
Pre-CNN	96.8%	–	–
DPND	92.9%	–	76.6%(64)
DPND+DPR	94.4%	–	78.4%(64)
DCNN(manually)	98.0%	97.5%	79.4%(100)
DCNN(automatically)	96.7%	94.2%	77.4%(100)

Table 4
Learning affective state analysis for E-learning dataset with the proposed method.

	Recognition rate
Neutral state recognition	97.56%
Positive/Negative state recognition	95.00%
Aver.	96.28%

Performance of the proposed method on learning affective state recognition. The proposed method is used to analyze the student's affective state in an e-learning environment. The affective states are divided into three classes, i.e., neutral, positive and negative states, which correspond to neutral, happiness and other facial expressions. Hence, the neutral state is detected using binary CNN, then the positive and negative states are distinguished using the DCNN. The experiment results are shown in Table 4, one can see that the accuracy of the neutral state recognition and that of the positive/negative state recognition are 97.56% and 95.00%, respectively; the proposed method achieves an average accuracy of 96.28% of recognizing the affective state on the E-learning dataset, which proves that learning affective state can be obtained effectively by the proposed method.

To show the generalization ability, the learning affective states of 120 sequences selected from the E-learning dataset are estimated using the model trained with CK+ dataset. The recognition rate of cross-dataset evaluation is 88.91%, which shows that the proposed method has good generalization ability.

4. Conclusions

In this paper, we propose a two-stage framework based on DCNN to recognize facial expression for social signal analysis. Considering the facial expression's nonstationary nature, the proposed framework contains two stages: in the first stage, the neutral expression frame and fully expression frame are automatically picked out from the facial expression sequence by the SoftMax score of the binary CNN; then in the second stage, the selected neutral expression frame and fully expression frame are fed to the DCNN respectively. The proposed method can effectively eliminate the individual difference using the difference information of the neutral expression frame and fully expression frame. The proposed method has been evaluated on two facial expression sequence datasets, which achieves an accuracy of 95.4% on the CK+ dataset and an accuracy of 77.4% on the BU-4DFE dataset. Also, the proposed method achieves an accuracy of 96.28% in the student's affective state analysis in an E-learning environment. In future work, we plan to apply the proposed DCNN to expression recognition and intensity estimation jointly.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2018YFB1004504), the National Natural Science Foundation of China (No. 61772380, 61273063), Foundation for Innovative Research Groups of Hubei Province (No. 2017CFA007), Hubei Province Technological Innovation Major Project and Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU19Z02002).

Conflict of interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.jpdc.2019.04.017.

References

[1] J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, IEEE Trans. Affect. Comput. (99) (2018) 38–50.

[2] J. Chen, D. Chen, X. Li, Towards improving social communication skills with multimodal sensory information, IEEE Trans. Ind. Inf. (99) (2013) 1–8.

[3] D. Chen, Y. Hu, L. Wang, A.Y. Zomaya, X. Li, H-PARAFAC: Hierarchical parallel factor analysis of multidimensional big data, IEEE Trans. Parallel Distrib. Syst. 28 (4) (2017) 1091–1104.

[4] D. Chen, X. Li, L. Wang, S.U. Khan, J. Wang, K. Zeng, C. Cai, Fast and scalable multi-way analysis of massive neural data, IEEE Trans. Comput. 64 (3) (2015) 707–719.

[5] J. Chen, N. Luo, Y. Liu, L. Liu, K. Zhang, J. Kolodziej, A hybrid intelligence-aided approach to affect-sensitive e-learning, Computing 98 (1–2) (2016) 215–233.

[6] J. Chen, R. Xu, L. Liu, Deep peak-neutral difference feature for facial expression recognition, Multimedia Tools Appl. (2) (2018) 1–17.

[7] A. Dapogny, K. Bailly, S. Dubuisson, Pairwise conditional random forests for facial expression recognition, in: IEEE International Conference on Computer Vision, 2016, pp. 3783–3791.

[8] J.R. Delannoy, J. Mcdonald, Automatic estimation of the dynamics of facial expression using a three-level model of intensity, in: IEEE International Conference on Automatic Face & Gesture Recognition, 2008, pp. 1–6.

[9] L. Du, H. Hu, Modified classification and regression tree for facial expression recognition with using difference expression images, Electron. Lett. 53 (9) (2017) 590–592.

[10] P. Ekman, W.V. Friesen, Facial action coding system (FACS): a technique for the measurement of facial actions, Riv. Psichiatria 47 (2) (1978) 126–138.

[11] Z. Guoying, P. Matti, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[12] N. Haber, C. Voss, A. Fazel, T. Winograd, D.P. Wall, A practical approach to real-time neutral feature subtraction for facial expression recognition, in: Applications of Computer Vision, 2016, pp. 1–9.

[13] B. Hasani, M.H. Mahoor, Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields, in: IEEE International Conference on Automatic Face & Gesture Recognition, 2017, pp. 790–795.

[14] X. Jiang, B. Feng, L. Jin, Facial expression recognition via sparse representation using positive and reverse templates, IET Image Process. 10 (8) (2016) 616–623.

[15] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: IEEE International Conference on Computer Vision, 2016, pp. 2983–2991.

[16] S.K.A. Kamarol, M.H. Jaward, J. Parkkinen, R. Parthiban, Spatiotemporal feature extraction for facial expression recognition, IET Image Process. 10 (7) (2016) 534–541.

[17] H. Ke, D. Chen, T. Shah, X. Liu, X. Zhang, L. Zhang, X. Li, Cloud-aided online EEG classification system for brain healthcare: A case study of depression evaluation with a lightweight CNN, Software: Practice and Experience.

[18] P. Khorrami, T.L. Paine, T.S. Huang, Do deep neural networks learn facial action units when doing expression recognition? in: 2015 IEEE International Conference on Computer Vision Workshop, ICCVW, 2015, pp. 19–27.

[19] J.J. Lien, T. Kanade, J.F. Cohn, C.C. Li, A.J. Zlochower, Subtly different facial expression recognition and expression intensity estimation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998, pp. 853.

[20] A.T. Lopes, E.D. Aguiar, A.F.D. Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order, Pattern Recognit. 61 (2017) 610–628.

[21] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Conference, 2015, pp. 41.1–41.12.

[22] C. Qi, M. Li, Q. Wang, H. Zhang, J. Xing, Z. Gao, H. Zhanga, Facial expressions recognition based on cognition and mapped binary patterns, IEEE Access (99) (2018) 1–1.

[23] B. Ryu, A.R. Rivera, J. Kim, O. Chae, Local directional ternary pattern for facial expression recognition, IEEE Trans. Image Process. (99) (2017) 1–1.

[24] R. Sawyer, A. Smith, J. Rowe, R. Azevedo, J. Lester, Enhancing student models in game-based learning with facial expression recognition, in: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, ACM, 2017, pp. 192–201.

[25] Y. Tang, D. Chen, L. Wang, A.Y. Zomaya, J. Chen, H. Liu, Bayesian Tensor factorization for multi-way analysis of multi-dimensional EEG, Neurocomputing 318 (2018) 162–174.

[26] C. Tang, W. Zheng, J. Yan, Q. Li, Y. Li, T. Zhang, Z. Cui, View-independent facial action unit detection, in: IEEE International Conference on Automatic Face & Gesture Recognition, 2017, pp. 878–882.

[27] Y. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.

[28] X. Xiong, F.D.L. Torre, Supervised descent method and its applications to face alignment, in: Computer Vision and Pattern Recognition, 2013, pp. 532–539.

[29] B. Yang, J. Cao, R. Ni, Y. Zhang, Facial expression recognition using weighted mixture deep neural network based on double-channel facial images, IEEE Access (99) (2017) 1–1.

[30] H. Yang, L. Yin, CNN based 3D facial expression recognition using masking and landmark features, in: International Conference on Affective Computing & Intelligent Interaction, 2017, pp. 556–560.

[31] L. Yin, X. Wei, Y. Sun, J. Wang, A 3D facial expression database for facial behavior research, in: International Conference on Automatic Face and Gesture Recognition, 2006, pp. 211–216.

[32] C. Zhou, P. Shen, C. Xiong, Research on algorithm of state recognition of students based on facial expression, in: International Conference on Electronic & Mechanical Engineering & Information Technology, 2011.

Jingying Chen received bachelor's and master's degrees from the Huazhong University of Science and Technology, Wuhan, China, and Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001. She had been working as a Post-doctor in INRIA, France, and a Research Fellow with University of St. Andrews and University of Edinburgh, U.K. She is currently a Professor with the National Engineering Center for E-earning, Central China Normal University, China. Her research interests include image processing, computer vision, pattern recognition, and multimedia applications.

Yongqiang Lv received the B.S. degree in Huazhong University of Science and Technology, Wuhan, China, in 2018. He is currently pursuing the master degree at National Engineering Research Center for E-Learning, Central China Normal University. His research interests include machine learning and pattern recognition.

Ruyi Xu received bachelor's degree in automation from Wuhan University of Science and Technology, China, in 2008 and master's degrees in circuit and system from the Huazhong University of Science and Technology, Wuhan, China in 2016. He is currently a Research Assistant with the National Engineering Research Center for E-Learning, Central China Normal University, China. His research interests include computer vision and multimedia applications.

Can Xu received the B.S. degree in Wuhan University, Wuhan, China, in 2016. He is currently pursuing the master degree at National Engineering Research Center for E-Learning, Central China Normal University. Her research interests include machine learning and pattern recognition.