MULTIPLE FEATURE FUSION FOR AUTOMATIC EMOTION RECOGNITION USING EEG SIGNALS

Ningjie Liu, Yuchun Fang, Ling Li, Limin Hou, Fenglei Yang, Yike Guo

School of Computer Engineering and Science, Shanghai University
School of Computing, University of Kent
School of Communication and Information Engineering, Shanghai University
Department of Computing, Imperial College London
liuningjie@i.shu.edu.cn, ycfang@shu.edu.cn, c.li@kent.ac.uk
lmhou@staff.shu.edu.cn, flyang@shu.edu.cn, y.guo@imperial.ac.uk

ABSTRACT

Automatic emotion recognition based on electroencephalographic (EEG) signals has received increasing attention in recent years. The Deep Residual Networks (ResNets) can solve vanishing gradient problem and exploding gradient problem well in computer vision and can learn more profound semantic information. And for traditional methods, frequency features often play important role in signal processing area. Thus, in this paper, we use the pre-trained ResNets to extract deep semantic information and the linear-frequency cepstral coefficients (LFCC) as features from raw EEG signals. Then the two features are fused to improve the emotion classification performance of our approach. Moreover, several classifiers are used for our fused features to evaluate the performance and it shows that the proposed approach is effective for emotion classification. We find that the best performance is achieved when use k-nearst neighbor (KNN) as classifier, and we provide a detailed discussion for the reason.

Index Terms— emotion recognition, EEG, Residual Networks, cepstral coefficients

1. INTRODUCTION

Affective computing is a new research hotspot in human-computer interaction (HCI) system. Emotion recognition plays an important role in affective computing [1], which includes speech recognition, facial expression recognition, text recognition and physiological signal recognition. Currently, there are numerous studies measuring the emotional states by analyzing physiological signals under the emotional stimuli [2]. The most common physiological signals used in emotion studies are electroencephalogram (EEG), electrocardiogram (ECG), respiration and skin conductance. Among them, EEG signals provide a direct and comprehensive way for emotion recognition by measuring immediate response to emotional stimuli in good temporal resolution [3] [4] [5]. Thus, automatic EEG-based emotion recognition has received increasing attention.

EEG signals are always susceptible to noise and artifacts. In recent years, the frequency feature does well in emotion recognition. The most common frequency feature is the frequency band power feature [6]. And mel-frequency cepstral coefficient (MFCC) [7] is often used for speech recognition but is beginning to find there way into EEG research [8]. In addition, deep learning can automatically derive features from the raw signals without the expert knowledge. Recent studies developed different kinds of emotion recognition models and some deep learning models obtained comparable performance in comparison with other traditional methods. For example, Zheng [9] and Xu [10] trained the Deep Belief Network (DBN) to classify emotions from EEG data, and Jirayucharoensak [11] implemented a sparse auto-encoder whose input features are from 32-channel EEG signals. For cross validation, the k-fold cross validation may be more suitable for machines to learn and predict the emotional state of a particular object so that they can provide better service for a particular person. The LOO cross validation is more suitable for universal emotion prediction, has nothing to do with the user identity.

In this paper, we choose two features, one is extracted by the pre-trained Residual Networks (ResNets) called "ResNet-50" using 32 channel EEG signals while another is the linear-frequency cepstral coefficients (LFCC) extracted from 2 channels. And then we classify users' emotions by several classifiers and discuss the results of the proposed model in details.

The paper is organized as follows: Section 2 overview the ResNet-50 and LFCC and propose our whole model architecture; in Section 3 we present experimental results to evaluate the proposed approach and analyze the performance in detail. Finally in Section 4 we conclude the paper.

The work is funded by the National Natural Science Foundation of China (No.61371149, No.61170155), Shanghai Innovation Action Plan Project (No. 16511101200) and the Open Project Program of the National Laboratory of Pattern Recognition (No. 201600017).

2. MODEL

In this paper, the raw EEG signals are preprocessed at first. Then LFCC and ResNet-50 features are extracted from preprocessed EEG signals. Eventually, all features are fed into several different classifiers to recognize emotions.

2.1. ResNet-50

Deep Residual Networks (ResNets) [12] lead a dramatic increase in both depth and accuracy of CNNs, facilitated by constraining the network to learn residuals. ResNets are built up by stacking residual units, which is shown in Figure 1. For residual unit i, x and y represent the input and output vectors of layers considered, respectively. The $\mathcal{F}(\cdot)$ represents the trainable non-linear residual mappings. The output of residual unit i can be expressed as:

$$y = \mathcal{F}(x, \mathcal{W}_i) + x \tag{1}$$

where $\mathcal{W}_i$ denotes the trainable parameters of i-th residual unit. ResNets can be intuitively understood by regarding residual functions as paths that information can propagate easily. In each layer, a ResNet learns more complex feature combinations with the shallower representation from the previous layer. The network architecture allows the construction of deeper networks.

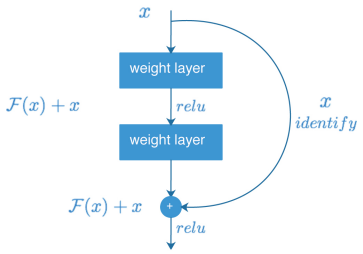


Fig. 1. Residual Networks block illustration.

2.2. LFCC

The MFCC [7] is a classical speech feature used for speech recognition. It exploits nonlinear frequency scale and the property of cepstrum. The cepstrum provides parameter concentration and it helps reduce dimensionality. The human audio system can be considered as a nonlinear system, however, since there is no evidence that a log scale is also meaningful for EEG signals [13], we change the mel-scale filters in MFCC to linear-scale filters. The modified-MFCC for EEG signals is named LFCC in this paper. The LFCC is employed in this study as features from EEG signals and the extracting process is shown in Fig. 2.

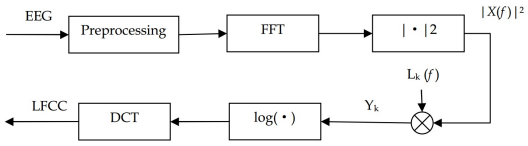


Fig. 2. Extracting LFCC block diagram.

In Fig. 2, preprocessing includes framing and windowing. In EEG signal analysis, frame length is 1s. The 1s Hamming window was shifted at a 1/3s frame interval. Then obtain the spectrum of each frame presented as $X(f)$ using Fast Fourier Transform (FFT). After that, calculate the power spectrum $|X(f)|^2$ and gain Y_k by (2). The spectrum is smoothed and the main frequency components in the spectrum is highlight through (2), which also facilitate the extraction of the cepstrum:

$$Y_k = \sum_{f_{kl}}^{f_{kh}} L_k(f)|X(f)|^2, 1 \le k \le K \tag{2}$$

Where $L_k(f)$ is the frequency response of the kth hann shaped filters in linear frequency domain while f_{kl} and f_{kh} are the lowest frequency and the highest frequency for the kth filter. The filter number K is set to 24.

Next, calculate LFCC by (3):

$$C_{LFCC}(i) = \sum_{k=1}^{K} log(Y_k) cos(\frac{(2k-1)i\pi}{2K}), 1 \le i \le I \tag{3}$$

Where I is the dimension of LFCC that is set to 12.

Finally, we obtain a 12-dimension feature vector for a frame.

2.3. Model structure

For our approach, the raw EEG signals are preprocessed for ResNet-50 and LFCC in different way at first. Then LFCC and ResNet-50 features are extracted from preprocessed EEG signals and fused by channel. Eventually, all fused features are fed into the several classifiers to recognize emotions. And the architecture of proposed approach is shown in Fig. 3.

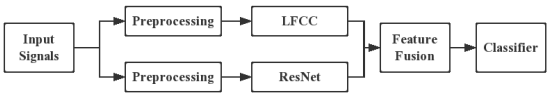


Fig. 3. The architecture of proposed approach.

For classification, we use 7 different classifiers to evaluate the features from ResNet-50 and LFCC: k-nearst neighbor(KNN), support vector machines (SVM), logical regression (LR), random forest (RF), naive Bayesian (NB), decision

tree (DT) and a fully-connected neural network (FC) with 3 Dense layers and 2 Dropout layers.

3. EXPERIMENTS

3.1. Database

DEAP, the open database for emotion analysis from EEG signals, is used in this work [14]. 32 participants watched a subset of 40 of one-minute music videos. Their EEG and other physiological signals were recorded. Each trial includes 63s signal where the first 3s is baseline signal. At the end of each video, each participant performed self-assessment (SAM) of arousal, valence, liking and dominance on a scale of 1 to 9 for each video. Moreover, the database contains a preprocessed version of the original EEG signals, which were down-sampled to 128Hz and removed the EOG artifacts, and a bandpass frequency filter from 4.0Hz to 45.0Hz was applied. We use the preprocessed version database to evaluate the proposed model.

This paper mainly takes valence-arousal (VA) model [15] into account. We construct 3 classification tasks based on VA model: low/high valence (task1) and low/high arousal (task2) and low arousal low valence/high arousal low valence/low arousal high valence/high arousal high valence (task3). Moreover, the SAM-ratings value ranging from 1 to 5 is low and the value ranging from 5 to 9 is high.

We first normalize our database to a Gaussian distribution and use 32 channel EEG signals from one trial as a unit to reconstruct the database. We convert our data into 2D image format so the pre-trained ResNet-50 can learn to classify them effectively. Eventually, we get 1280 (32 participants $\times$ 40 videos) signal images with the shape of $224 \times 384 \times 3$ (32 channels with 8064 data). For LFCC features, we choose 2 channels, Fp1 and C4, which with the largest average sample entropy. And we get 189 feature vectors with 12 dimensions for each signal and we flatten it as a one-dimension vector with the length of 2268. Then two features are fused. The feature from ResNet-50 is fused to 2 different channels for one video of one subject. And finally we obtain the features with the shape of 1280×8632, where 8632 represents 4096 (ResNet-50 features) + 2268 (LFCC features) $\times$ 2 (channels).

3.2. Experimental Results

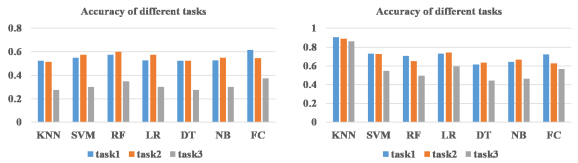
Both 10-fold cross validation and LOO cross validation are used to evaluate the classification performance in experiments. And different classifiers are used in our experiments.

3.2.1. Results for 10-fold cross validation

For task1 (high/low valence) and task2 (high/low arousal), the best accuracy of our proposed approach can reach 93.75% and the average accuracy is 89.72%. For task3 (low arousal low valence/high arousal low valence/low arousal high valence/high arousal high valence), the best accuracy of the proposed approach is 90.21%. And the performance with different classifiers is shown in Fig. 4(a). Fig. 4(a), we can see that our approach reaches the best performance when we use KNN as classifier. And the average accuracies of all classifiers are 72.18%, 70.59% and 56.75% for 3 different tasks respectively. Koelstra et al. [14] only has the average accuracy 59.72% of high/low valence and high/low arousal. It is obvious that our approach with different classifiers is effective for emotion recognition using EEG signals. Moreover, the performance of our proposed method is compared to other methods with deep learning networks and using k-fold cross validation on DEAP database, which is shown in Table 1.

Table 1. The comparison of our model with previous studies.

research	method	average accuracy		cross validation
		valence	arousal	
Li et al. [16]	C-RNN	72.06%	74.12%	5-fold
Al-Nafjan et al. [17]	PSD+DNN	82.00%	82.00%	10-fold
Liu et al. [18]	Multimodal Deep Learning	85.20%	80.50%	10-fold
our model	ResNet+LFCC+KNN	**90.39%**	**89.06%**	10-fold



(a) LOO cross validation (b) 10-fold cross validation

Fig. 4. The accuracies of different tasks. Among them, task1 (high/low valence), task2 (high/low arousal), task3 (low arousal low valence/high arousal low valence/low arousal high valence/high arousal high valence)

From Table 1, it can be seen that our average accuracy for task1 and task2 is 89.72% which is about 16.63% and 7.72% and 6.87% higher than in Ref. [16] and Ref. [17] and Ref. [18], respectively. Nevertheless, Ref. [16] used 5-fold cross validation to split data while we use 10-fold. It is not as comparable as other researches with 10-fold validation. And the average accuracy is 86.05% for task3. It is obvious that our best proposed approach evidently outperforms the comparison methods on classification performance using k-fold cross validation.

3.2.2. Results for LOO cross validation

For task1 (high/low valence) and task2 (high/low arousal), the best accuracy of our proposed approach can reach 82.5% and the average accuracy is 58.03%. For task3 (low arousal low valence/high arousal low valence/low arousal high valence/high arousal high valence), the best accuracy of the proposed approach is 37.5%. And the performance with different

Table 2. The comparison of our model with previous studies.

research	method	average accuracy		cross validation
		valence	arousal	
Shu et al.[19]	restricted Boltzmann machine (RBM)	60.70%	64.60%	leave-one-video-out
Xu et al.[10]	Deep Belief Networks	66.88%	69.84%	LOO
Zhong et al.[20]	transfer recursive feature elimination (T-RFE)	78.75%	78.67%	LOO
our model	ResNet+LFCC+FC	61.55%	54.53%	LOO

classifiers is shown in Fig. 4(b).

It can be seen that our method reach the best performance when we use FC as classifier. For FC, we build the network to fine-tune the ResNet-50 parameters, making it better for dealing EEG signals. So its effect is better than the other classifiers to some extent. And the average accuracies of all classifiers are 54.93%, 55.49% and 31.07% for 3 tasks respectively. It is obvious that our model achieves better performance than random classification performance for emotion recognition using EEG signals.

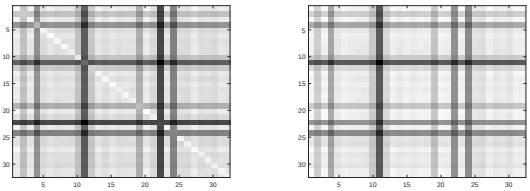
Moreover, the performance of our proposed method is compared to other methods with deep learning algorithm and using LOO cross validation on DEAP database, which is shown in Table 2. From Table 2, it can be seen that our average accuracy for task1 and task2 is not as comparable as other researches with LOO cross validation.

3.3. Discussion

From the experimental results, it is obvious that our model achieves better performance with 10-fold cross validation than LOO. The situation is caused by several factors. There are two main factors: personal emotional specificity [21] and the huge difference between people in their self-assessment of their own emotional state. Among them, the second factor has less effects of the experiments by setting thresholds for emotion labels. However, it may be hard to predict an unknown person's emotion state by learning or analyzing the information contained in EEG signals of other persons, especially when the current DEAP database contains only 32 subjects.

For 10-fold cross validation method, it is obvious that the performance of KNN is better than other classifiers. The main reason we suspect is that, by the same person, the similarity of EEG signals produced in similar emotions is higher than by different persons, and the difference would not disappear with the feature extraction of LFCC and ResNet-50. And we verify the ideas by calculating the average Euclidean distance of EEG signals between different people by the same stimulus. One of the distance array is shown in Fig. 5(a).

In Fig. 5(a), the lighter the color means the smaller the distance value is, and also means the more similar the EEG signals is. It can be easily seen that the distance on the diagonal is significantly smaller than the other values, indicating that the degree of similarity between multi-channel EEG signals by the same stimulus for the same person is higher than for different persons. And the main basis of KNN is the similarity between features, so its effect is significantly better than other classifiers. Moreover, we also calculate the average European distance between different subjects by different videos. The result is shown in Fig. 5(b). It further illustrates our suspect though the result is not so obvious as Fig. 5(a), due to video differences. To the 11nd subject, his produced EEG signals are significantly different from other subjects, taking the data of other 31 subjects to predict his emotional state is difficult theoretically. And with the LOO cross validation method, his emotion recognition accuracy is indeed the lowest of all, which only achieve 40% by KNN classifier.



(a) the same video (b) all 40 videos

Fig. 5. The average Euclidean distance of multi-channel EEG signals between 32 subjects.

4. CONCLUSION

This paper proposes an automatic approach to address the emotion recognition problem of EEG signals using fused ResNet-50 and LFCC features and several classifiers. We also discuss the performance of proposed approach with 10-fold cross validation and LOO cross validation. Our results show that the our model is effective for emotion classification. Moreover, we find that KNN achieves the best performance in different classifiers, and we provide an easy understanding explanations that by the same person, the similarity of EEG signals produced in similar emotions is higher than by different persons. In the future, our work will focus on the model that performances better both on LOO cross validation and k-fold cross validation.

5. REFERENCES

[1] Hatice Gunes and Massimo Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[2] Jonghwa Kim and Elisabeth André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[3] Michela Balconi and Claudio Lucchiari, "Eeg correlates (event-related desynchronization) of emotional face elaboration: a temporal analysis," *Neuroscience letters*, vol. 392, no. 1, pp. 118–123, 2006.

[4] Marni YV Bekkedal, John Rossi, and Jaak Panksepp, "Human brain eeg indices of emotions: delineating responses to affective vocalizations by measuring frontal theta event-related synchronization," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 9, pp. 1959–1970, 2011.

[5] Paul R Davidson, Richard D Jones, and Malik TR Peiris, "Eeg-based lapse detection with high temporal resolution," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 832–839, 2007.

[6] Robert Jenke, Angelika Peer, and Martin Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[7] Joseph W Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[8] A Temko, G Boylan, W Marnane, and G Lightbody, "Speech recognition features for eeg signal description in detection of neonatal seizures," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 3281–3284.

[9] Wei-Long Zheng, Hao-Tian Guo, and Bao-Liang Lu, "Revealing critical channels and frequency bands for emotion recognition from eeg with deep belief network," in *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*. IEEE, 2015, pp. 154–157.

[10] Haiyan Xu and Konstantinos N Plataniotis, "Eeg-based affect states classification using deep belief networks," in *Digital Media Industry & Academic Forum (DMIAF)*. IEEE, 2016, pp. 148–153.

[11] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena, "Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *The Scientific World Journal*, vol. 2014, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[13] A Harati, M Golmohammadi, S Lopez, I Obeid, and J Picone, "Improved eeg event classification using differential energy," in *Signal Processing in Medicine and Biology Symposium (SPMB), 2015 IEEE*. IEEE, 2015, pp. 1–4.

[14] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[15] JA Ressel, "A circumplex model of affect," *J. Personality and Social Psychology*, vol. 39, pp. 1161–78, 1980.

[16] Xiang Li, Dawei Song, Peng Zhang, Guangliang Yu, Yuexian Hou, and Bin Hu, "Emotion recognition from multi-channel eeg data through convolutional recurrent neural network," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 352–359.

[17] Abeer Al-Nafjan, Manar Hosny, Areej Al-Wabil, and Yousef Al-Ohali, "Classification of human emotions from electroencephalogram (eeg) signal using deep neural network," .

[18] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu, "Emotion recognition using multimodal deep learning," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 521–529.

[19] Yangyang Shu and Shangfei Wang, "Emotion recognition through integrating eeg and peripheral signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2871–2875.

[20] Yin Zhong and Zhang Jianhua, "Subject-generic eeg feature selection for emotion classification via transfer recursive feature elimination," in *Control Conference (CCC), 2017 36th Chinese*. IEEE, 2017, pp. 11005–11010.

[21] Randy J Larsen, "Toward a science of mood regulation," *Psychological Inquiry*, vol. 11, no. 3, pp. 129–141, 2000.