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 electroencephalo-
graphic (EEG) signals has received increasing attention in re-
cent years. The Deep Residual Networks (ResNets) can solve
vanishing gradient problem and exploding gradient problem
well in computer vision and can learn more profound se-
mantic 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 classi-
fication performance of our approach. Moreover, several
classifiers are used for our fused features to evaluate the per-
formance and it shows that the proposed approach is effective
for emotion classification. We find that the best performance
is achieved when use k-nearest neighbor (KNN) as classifier,
and we provide a detailed discussion for the reason.

Index Terms— emotion recognition, EEG, Residual Net-
works, 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. Cur-
rently, 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), electrocar-
diogram (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 fre-
quency 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 (DB-
N) 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 emo-
tion prediction, has nothing to do with the user identity.

In this paper, we choose two features, one is extract-
ed by the pre-trained Residual Networks (ResNets) called
"ResNet-50" using 32 channel EEG signals while another
is the linear-frequency cepstral coefficients (LFCC) extract-
ed 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 archite-
cture; 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.
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 $F(\cdot)$ represents the trainable non-linear residual mappings. The output of residual unit $i$ can be expressed as:

$$ y = F(x, W_i) + x \quad (1) $$

where $W_i$ 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.](image)

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.

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:

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

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

Next, calculate LFCC by (3):

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

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. 2. Extracting LFCC block diagram.](image)

![Fig. 3. The architecture of proposed approach.](image)
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/high arousal low valence. 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 × 40 videos) signal images with the shape of 224 × 384 × 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) × 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 our 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.

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>
<thead>
<tr>
<th>research</th>
<th>method</th>
<th>average accuracy</th>
<th>valence</th>
<th>arousal</th>
<th>cross validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li et al. [16]</td>
<td>C-RNN</td>
<td>72.06%</td>
<td>74.12%</td>
<td>5-fold</td>
<td></td>
</tr>
<tr>
<td>Al-Nafjan et al. [17]</td>
<td>PSD+DNN</td>
<td>82.00%</td>
<td>82.00%</td>
<td>10-fold</td>
<td></td>
</tr>
<tr>
<td>Liu et al. [18]</td>
<td>Multimodal Deep Learning</td>
<td>85.20%</td>
<td>80.50%</td>
<td>10-fold</td>
<td></td>
</tr>
<tr>
<td>our model</td>
<td>ResNet+LFCC+KNN</td>
<td>90.39%</td>
<td>89.68%</td>
<td>10-fold</td>
<td></td>
</tr>
</tbody>
</table>

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)
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 11th 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.

### 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.

---

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

<table>
<thead>
<tr>
<th>Research Method</th>
<th>Average Accuracy</th>
<th>Cross Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shu et al [19]</td>
<td>60.70% 64.60%</td>
<td>leave-one-video-out</td>
</tr>
<tr>
<td>Xu et al [10]</td>
<td>66.88% 69.84%</td>
<td>LOO</td>
</tr>
<tr>
<td>Zhong et al [20]</td>
<td>78.75% 78.67%</td>
<td>LOO</td>
</tr>
<tr>
<td>Our Model</td>
<td>61.55% 54.53%</td>
<td>LOO</td>
</tr>
</tbody>
</table>
5. REFERENCES


