Multimodal Emotion Recognition using EEG and Eye Tracking Data

Wei-Long Zheng, Bo-Nan Dong and Bao-Liang Lu* *Senior Member, IEEE*

***Abstract*— This paper presents a new emotion recognition method which combines electroencephalograph (EEG) signals and pupillary response collected from eye tracker. We select 15 emotional film clips of 3 categories (positive, neutral and negative). The EEG signals and eye tracking data of five participants are recorded, simultaneously, while watching these videos. We extract emotion-relevant features from EEG signals and eye tracing data of 12 experiments and build a fusion model to improve the performance of emotion recognition. The best average accuracies based on EEG signals and eye tracking data are 71.77% and 58.90%, respectively. We also achieve average accuracies of 73.59% and 72.98% for feature level fusion strategy and decision level fusion strategy, respectively. These results show that both feature level fusion and decision level fusion combining EEG signals and eye tracking data can improve the performance of emotion recognition model.**

I. INTRODUCTION

In the past few decades, an increasing number of researches on emotion recognition have been done since emotion recognition has great significance and wide applications, especially its crucial role in human-machine interaction systems. Possible applications of emotion recognition cover a vast scope, whether at a professional, a personal or a social level. For driving safety, we can design an affective user interface to monitor drivers' emotional and cognitive states and response to drivers to regulate their emotions.

In the practice of emotion recognition, various signals have been adopted, which can be roughly classified into two categories: non-physiological and physiological signals. The early works are more based on non-physiological signals, such as text, facial expression, speech and gesture [1]. Recently, more researches are done based on physiological signals since the physiological signals, such as electroencephalography (EEG), pupillary diameter (PD), electromyogram (EMG), and electrocardiogram (ECG), seem to be more effective and reliable. Among them, electroencephalograph (EEG) which record brain activities in central nervous system, has been proved providing informative characteristics in responses to the emotional states [2]. Numerous researchers studied on emotion recognition using EEG [3] [4] [5]. Furthermore, previous studies [6] [7] have shown that pupil size discriminates during and after different kinds of emotional stimuli, which implies that the measurement of pupil size variation may be a potentially useful input signal. Some early attempts [8] of adopting pupillary response in emotion recognition have also arisen recently. Soleymani *et al.* [2] presented a user-independent emotion recognition method using EEG, pupillary response and gaze distance, which achieved the best accuracies of 68.5 percent for three labels of valence and 76.4 percent for three labels of arousal.

Emotion representations studied in affective computing do not always match emotions defined by psychologists (e.g. Ekman's six basic emotions) and need more quantitative information to describe. Therefore, the aim of our study is how to model and recognize emotions using various sensor technologies and physiological signals. In this paper, we investigated the relation between subjects EEG and pupillary response in response to multimedia and three categories of emotions, namely, positive, neutral and negative. Then we develop a multimodal method for emotion recognition where both EEG and pupillary responses are used together as input signals. From the experiment results, we show that the performance of the fusion model combining EEG and eye tracking features outperform previous methods based on unimodal signals.

II. EXPERIMENTS

A. Stimuli

In our experiment, 15 emotional film clips were selected to elicit three emotions: positive, neutral and negative. Each emotion had 5 video clips for a session and each clip lasted for around 4 minutes. In order to elicit the emotions of Chinese subjects efficiently, we chose these video clips from Chinese movies which were representative and popular, including Tangshan Earthquake, Just Another Pandora's Box, Lost in Thailand, Flirting Scholar and World Heritage In China.

B. Subjects

In our experiment, there were total 5 subjects participated in the emotion experiments (two females and three males whose ages range from 22-24). All of them were students from Shanghai Jiao Tong University, who had normal hearing and normal vision and were right-handed. They all were informed the destination of this experiment before the experiment started and in good spirits when performing experiment. Most participants performed the experiments three times and some participants for twice with an interval time of one week or longer.

Wei-Long Zheng, Bo-Nan Dong and Bao-Liang Lu are with the Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University and Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China.

*Corresponding author (bllu@sjtu.edu.cn). This work was partially supported by the National Natural Science Foundation of China (Grant No. 61272248), the National Basic Research Program of China (Grant No. 2013CB329401) and the Science and Technology Commission of Shanghai Municipality (Grant No.13511500200).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

C. Procedure

The experiments were arranged in the morning or early afternoon. Before experiments, subjects were asked to fill in a form containing basic information and sleep quality. The EEG signals were recorded using an ESI NeuroScan System at a sampling rate of 1000 Hz from a 62-channel electrode cap according to the international 10-20 system. The eye tracking data was recorded using SMI eye track glasses with 30 Hz of temporal resolution to collect pupillary information including pupil diameter. We also recorded the frontal face videos in the experiments. Fig. 1 shows the protocol of the EEG experiment. There are totally fifteen sessions in one experiment. The movie clips are played with a fixed order. There is a 5s hint before each clip and 45s for self-assessment and 15s for rest after each clip in one session. For self-assessment, participants were told to report their emotional reactions to each film clip by completing the questionnaire for the feedback.

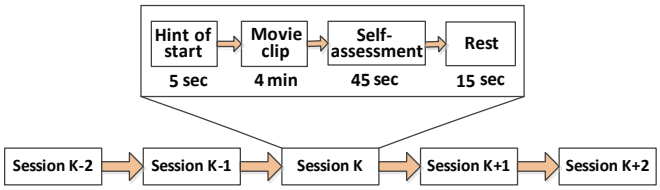

Fig. 1. The protocol of the experiment

III. METHODS

The framework of our experiment processing is shown in Fig. 2. For EEG data, we extracted different features from five frequency bands. For eye tracking data, we extracted mean values, standard deviations and spectral powers of frequency bands from pupil responses. We applied fusion methods of feature level fusion and decision level fusion combining features from EEG signals and eye tracking data.

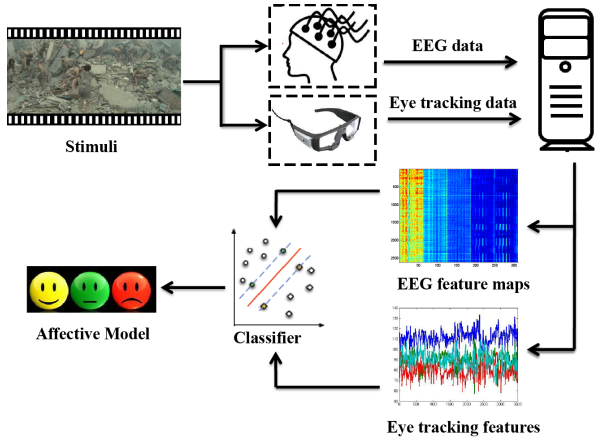

Fig. 2. The framework of our experiment processing

A. Preprocessing

In order to filter the noise and remove the artifacts, the EEG data was then processed with a bandpass filter between 0.5Hz to 70Hz. And in order to accelerate the computation, raw EEG data were downsampled to 200Hz and segmented into different trials.

B. Feature Extraction

1) EEG Signals: In order to transform the raw sequence signals into frequency domain features, which are highly correlated with emotion relevant processing, we used short-time fourier transform with a non-overlapped Hanning window of 4s. Four different features, power spectral density (PSD), differential entropy (DE), differential asymmetry (DASM) and rational asymmetry (RASM) were extracted and compared.

According to five frequency bands: delta (1-3Hz); theta (4-7Hz); alpha (8-13Hz); beta (14-30Hz); gamma (31-50Hz), we computed the traditional PSD features. Differential entropy feature is defined as follows [9],

$$h(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2} \log\frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2} dx = \frac{1}{2}\log 2\pi e\sigma^2 \quad (1)$$

where X submits the Gauss distribution $N(\mu, \sigma^2)$, x is a variable and π, and e are constant. According to [10], in a certain band, DE is equivalent to the logarithmic power spectral density for a fixed length EEG sequence. DASM and RASM are defined as:

$$DASM = h(X_{left}) - h(X_{right}) \quad (2)$$

$$RASM = h(X_{left})/h(X_{right}) \quad (3)$$

where X_{left} and X_{right} are DE features of left-hemisphere and right-hemisphere shown in Fig. 3. The electrodes of left-hemisphere and right-hemisphere are shown in blue and red color, respectively, and data from the middle eight electrodes are not included for the asymmetry features.

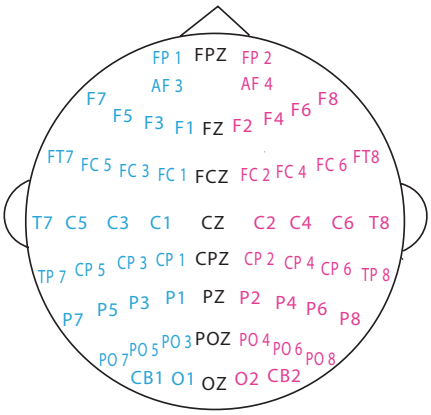

Fig. 3. The electrode distribution used in asymmetry features

2) Eye Tracking Data: In our experiment, pupil diameter was chosen as the feature for emotion classification. Pupil diameter has been discovered to change in different emotional states [11]. However, the pupil diameter is highly dependent on the luminance of the video clip. So it couldn't be used for emotion recognition directly. In this paper, we first builded

a light reflex model to approximately remove the luminance influences. The major changes of pupil diameter comes from lighting. And we assumed that the pupil in response to the lighting follows similar patterns in the experiments due to the controlled light environment. Here, we used principal component analysis (PCA) to build the light reflex model.

Suppose Y is the $M \times N$ matrix representing pupil diameters to the same video clip from N subjects and M samples. Then $Y = A + B + C$, where A is luminance influences which is prominent, and B is emotional influences which we want and C is the noises. We use principal component analysis to decompose Y. We extract the first principal component from PCA to approximate the pupil response for the lighting changes during the experiments. Let Y_{rest} be the emotion relevant pupil response. We define $Y_{rest} = Y - Y_1$. Then we extracte PSD and DE features from four frequency band (0-0.2Hz; 0.2-0.4Hz; 0.4-0.6Hz; 0.6-1Hz) [2] of the preprocessed pupil responses using short-time fourier transform with a non-overlapped Hanning window of 4s.

C. Feature Smoothing

When recording EEG signals, noises are easily introduced during the experiments. Here, we assume that the emotional state is defined in a continuous space and emotional states change slowly. In order to filter out noises and emotion unrelevant features, we applied a moving average filter with window length of five for eye tracking feature smoothing and linear dynamic system (LDS) approaches for EEG feature smoothing.

D. Classification

To evaluate the model, we divided data from the same experiment into two parts, the first nine sessions as training data and the rest six sessions for testing containing totally more than 800 samples each experiment. In this study, we employed support vector machine (SVM) as classifier. For SVM, we employed linear kernel and searched the parameter space to find the optimal value.

E. Multimodal Fusion

Signals from different modalities can be fused at both the feature level and the decision level. Here, we applied these two fusion strategies and evaluated their performance. For feature level fusion, the feature vectors from different approaches were concatenated to form a larger feature vector. In our experiment, We selected differential entropy features from EEG data and pupil response, and train the fusion model combining EEG features and eye tracking features.

For decision level fusion, two classifiers were trained with different features, respectively, and were fused to generate a new classification using some principles or learning algorithms. We applied two principles in decision level fusion in our studies. One was called max strategy which selected the higher probabilistic outputs of classifiers trained with a single modality separately as final results. Another was called sum strategy which summed up probabilities of same emotions from different frequency bands and selected higher one.

IV. RESULTS

A. EEG Based Classification

Table I shows the performance of different kinds of features on Delta, Theta, Alpha, Beta and Gamma frequency bands. The features were smoothed by linear dynamic system (LDS) and SVM was trained to classify the emotional states (positive, neutral, and negative). For each experiment, first nine sessions were used for training data and the rest six sessions for testing. In Table I, ASM features are a concatenation of DASM and RASM. As we can see, Delta and Gamma frequency bands perform better than Theta and Alpha frequency bands, and total frequency band has a stable and prominent accuracy. Also we can find that, differential entropy (DE) features get best accuracies in almost all frequency bands except theta band (47.98% of DE features is less than 51.87% of PSD features). This result makes it reasonable that we select DE feature to fuse with pupil diameter feature.

TABLE I

THE PERFORMANCE (%) OF CLASSIFIERS USING DIFFERENT KINDS OF FEATURES ON DELTA, THETA, ALPHA, BETA, GAMMA AND TOTAL FREQUENCY BANDS

Feature		Frequency Bands					
		Delta	Theta	Alpha	Beta	Gamma	Total
PSD	Mean	51.60	51.87	54.74	53.23	51.36	**59.04**
	Std	19.56	14.48	16.58	18.06	16.10	20.31
DE	Mean	70.51	47.98	60.18	64.29	68.73	**71.77**
	Std	12.18	15.19	12.94	23.05	20.30	12.03
DASM	Mean	61.08	43.42	49.98	46.96	64.12	**68.37**
	Std	22.45	19.45	15.59	15.21	22.94	23.86
RASM	Mean	61.44	44.90	48.69	48.18	62.71	**66.03**
	Std	22.90	12.14	14.62	15.93	21.11	24.62
ASM	Mean	65.18	44.78	50.29	45.19	63.92	**67.91**
	Std	22.32	13.87	15.91	12.77	22.19	24.45

B. Pupil Diameter based Classification

Fig. 4 shows the average pupil diameter of each experiment. From the results, we can see that pupil diameter changes in different emotional states. In most experiments we can find that the pupil diameter is biggest during sorrow sessions and smallest during calm sessions except experiment 1. The pupil diameters of positive and negative emotions are larger than those of neutral emotion. which is consistent with previous psychology literatures [11]. This result shows a correlation between pupil diameter and emotion and therefore we can extract emotion relevant features from pupil diameter to classify different emotions.

Table II shows the performance of using different features from pupil diameter. As we can see, DE features perform much better than PSD features because DE features have the balance ability of discriminating patterns between low and high frequency energy. Although the average accuracy achieved with pupil response features (58.90%) is not better than EEG features (71.77%), the dimensionality of eye tracking features is much less than EEG features. Comparing to EEG data, the dimensionality of eye tracking was only 8

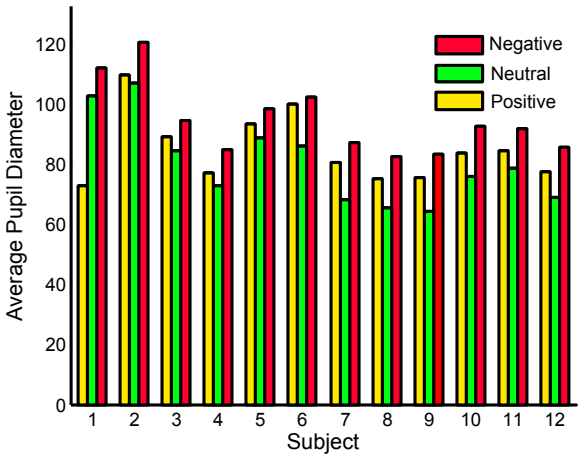

Fig. 4. The average pupil diameter of each experiment

for each sample (pupil diameter has 2 dimensions and each of dimension has 4 frequency bands) while the dimensionality of EEG features was 310 (EEG has 62 channels and each channel has 5 frequency bands). Therefore this result is considered acceptable and is potential to improve.

TABLE II
THE PERFORMANCE % OF USING DIFFERENT FEATURES FROM PUPIL DIAMETER

Exp	Feature	Accuracy	Exp	Feature	Accuracy
1	PSD	65.43	7	PSD	33.95
	DE	**86.42**		DE	**61.73**
2	PSD	56.79	8	PSD	46.91
	DE	**70.37**		DE	**50.62**
3	PSD	54.94	9	PSD	43.83
	DE	**56.79**		DE	**59.88**
4	PSD	60.49	10	PSD	36.42
	DE	**63.58**		DE	**59.88**
5	PSD	37.65	11	PSD	44.44
	DE	**48.77**		DE	**50.62**
6	PSD	33.95	12	PSD	34.57
	DE	**47.53**		DE	**50.62**
Mean	PSD	45.78	Std	PSD	11.03
	DE	**58.90**		DE	**10.25**

C. Fusion Results

Table III showed the results of the EEG based DE feature, decision level fusion using max strategy, decision level fusion using sum strategy and feature level fusion. From Table III, we see that decision level fusion using max strategy and feature level fusion performed better than single modality like EEG or pupil diameter, which achieved average accuracies of 72.98% and 73.59%, respectively.

V. CONCLUSION

In this paper, we designed an emotion experiment and collected EEG signals as well as eye tracking data of total 12 experiments, simultaneously, while subjects were watching emotional film clips (positive, neutral and negative). Here,

TABLE III
THE ACCURACIES (%) OF 12 EXPERIMENTS USING FUSION STRATEGIES FROM DIFFERENT MODALITIES

	EEG (DE)	Max Strategy	Sum Strategy	Feature Fusion
1	83.09	83.09	83.09	**93.59**
2	68.22	68.22	51.31	**78.72**
3	68.22	67.93	51.02	**68.22**
4	85.13	68.22	**85.13**	83.97
5	51.31	51.31	51.31	**77.55**
6	83.09	**83.09**	**83.09**	**83.09**
7	51.31	**68.22**	**68.22**	58.02
8	83.09	83.09	83.09	**83.38**
9	68.22	**83.09**	68.22	63.56
10	68.22	68.22	68.22	**69.10**
11	**68.22**	**68.22**	**68.22**	40.82
12	**83.09**	**83.09**	65.89	**83.09**
Mean	71.77	72.98	68.90	**73.59**
Std	12.03	**10.09**	12.85	14.43

we extracted different features including PSD, DE, DASM, RASM and ASM features for EEG signals and PSD, DE features for eye tracking data. From the results, we showed that EEG and pupil diameter were efficient cues to recognize emotions. Then we employed two fusion strategies (feature level fusion and decision level fusion) to build emotion recognition models which achieved the best classification accuracies of 73.59 % and 72.98 %, respectively.

REFERENCES

[1] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[2] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[3] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks," to appear in *2014 IEEE International Conference on Multimedia & Expo*.

[4] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "EEG-based emotion recognition during watching movies," in *2011 5th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2011, pp. 667–670.

[5] Y.-P. Lin, C.-H. Wang, T.-L. Wu, S.-K. Jeng, and J.-H. Chen, "EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.* IEEE, 2009, pp. 489–492.

[6] E. Granholm and S. R. Steinhauer, "Pupillometric measures of cognitive and emotional processes," *International Journal of Psychophysiology*, vol. 52, no. 1, pp. 1–6, 2004.

[7] T. Partala, M. Jokiniemi, and V. Surakka, "Pupillary responses to emotionally provocative stimuli," in *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 2000, pp. 123–129.

[8] N. A. Harrison, T. Singer, P. Rotshtein, R. J. Dolan, and H. D. Critchley, "Pupillary contagion: central mechanisms engaged in sadness processing," *Social cognitive and affective neuroscience*, vol. 1, no. 1, pp. 5–17, 2006.

[9] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2013, pp. 6627–6630.

[10] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2013, pp. 81–84.

[11] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.