Combining Eye Movements and EEG to Enhance Emotion Recognition

Yifei Lu[1,*]**, Wei-Long Zheng**[1,*]**, Binbin Li**[1]**,** and **Bao-Liang Lu**[1,2,†]

[1]Department of Computer Science and Engineering
[2]Key Laboratory of Shanghai Education Commission for
Intelligent Interaction and Cognitive Engineering
Shanghai Jiao Tong University, Shanghai, China
{luyifei0715,weilong,libinbin,bllu}@sjtu.edu.cn

Abstract

In this paper, we adopt a multimodal emotion recognition framework by combining eye movements and electroencephalography (EEG) to enhance emotion recognition. The main contributions of this paper are twofold. a) We investigate sixteen eye movements related to emotions and identify the intrinsic patterns of these eye movements for three emotional states: positive, neutral and negative. b) We examine various modality fusion strategies for integrating users external subconscious behaviors and internal cognitive states and reveal that the characteristics of eye movements and EEG are complementary to emotion recognition. Experiment results demonstrate that modality fusion could significantly improve emotion recognition accuracy in comparison with single modality. The best accuracy achieved by fuzzy integral fusion strategy is 87.59%, whereas the accuracies of solely using eye movements and EEG data are 77.80% and 78.51%, respectively.

1 Introduction

Emotion is a subjective, conscious experience when people are faced with internal or external stimuli, which plays an essential part in natural communication among humans. In recent years, there has been a rising tendency in Human-Computer Interaction (HCI) research to enhance the systems with the ability to detect, process, and respond to users emotional states [Tanguy *et al.*, 2007]. Besides logical intelligence, emotional intelligence (EI) is also considered as an important part of human intelligence, which was firstly proposed by Salovery and Mayer [Salovey and Mayer, 1989]. Emotional intelligence refers to the ability to perceive emotions, understand emotions and regulate emotions. In particular, the introduction of emotional intelligence into computer technologies has been rapidly developed as an interdisciplinary field called Affective Computing [Picard, 2000] due to the wide range of potential applications. For example, an

---

*The first two authors contributed equally to this work and should be considered as joint first authors

†Corresponding author

affective intelligent car interface can enhance driving safety [Nasoz *et al.*, 2010] and computer agents use emotions to enhance decision-making [Antos and Pfeffer, 2011]. In the framework of EI, emotion recognition is the first critical phase because computers can never respond to users emotional states without recognizing emotions. Recent work [Ptaszynski *et al.*, 2009] reports the need to apply contextual analysis to emotion processing.

Since emotion contains many nonverbal cues, various studies apply different modalities such as facial expressions, speech and gestures as indicators of emotional states [Calvo and D'Mello, 2010]. However, these methods usually base on the external behaviors, and ignore the 'inner' cognitive states of users and semantic contexts of emotion, which limits the usability and reliability in real world applications. In contrast, methods based on physiological signals are considered more reliable ways to interpret emotions for their objective measure of the central nervous system and autonomic nervous system [Chanel *et al.*, 2011]. Among these approaches, EEG-based emotion recognition has attracted increasing interest and various studies have shown its suitability and effectiveness [Lin *et al.*, 2010; Nie *et al.*, 2011; Wang *et al.*, 2014].

Eye movement signals have become widely used in HCI research for usability analysis and assessment since they can provide a natural and efficient way to observe the behaviors of users. Most previous work uses eye movements to analyze interest of users, visual search processes, and information processing [Rayner, 2009]. Eye movement signals allow to find out what is attracting users attention and observe their subconscious behaviors. They can be important cues for context-aware environment, which contain complementary information for emotion recognition. However, limited studies have developed effective features of eye movements for emotion recognition so far [Bradley *et al.*, 2008; Soleymani *et al.*, 2012; Zheng *et al.*, 2014], where most researchers focus on pupillary responses to different emotions. In this paper, we systematically evaluate sixteen different features extracted from eye movement signals and investigate the intrinsic patterns associated with different emotions.

Since emotions are complex psycho-physiological phenomena associated with many nonverbal cues, it is difficult to build robust emotion recognition models using just a single modality. Signals from different modalities represent differ-

ent aspects of emotion and the complementary information from different modalities can be integrated to build a more robust emotion recognition model compared to the existing unimodel approaches [Calvo and D'Mello, 2010]. For multimodal emotion recognition, most studies focus on the combinations of audio-visual features [Chen *et al.*, 1998] or different physiological signals [Verma and Tiwary, 2014]. There has been a tendency of combining external behavior activities and internal physiological changes [Soleymani *et al.*, 2012].

In this paper, we adopt a multimodal emotion recognition framework by combining eye movements and EEG for three emotions (postive, negative and neutral). We explore the efficient features of eye movements and EEG, and utilize the advantages of their complementary information for emotion recognition from different modality fusion strategies. Results show that modality fusion could significantly enhance emotion recognition accuracy compared with single modality. Fusion of eye movement and EEG could better model both the subconscious behaviors and cognitive states of users simultaneously under different emotion elicitation.

2 Methods

2.1 Data Preprocessing

Eye movement data provides different detailed parameters, such as pupil diameters, fixation details, saccade details, blink details and event details statistics. In order to align eye movement time series with EEG time series, we further re-sample the eye movement data. It should be noted that although pupil diameter has been shown to change in different emotional states [Bradley *et al.*, 2008; Partala and Surakka, 2003], the major cause to the change of pupil diameter is the lighting. It is essential to remove the light reflex if we want to obtain the emotional information in the pupil diameter. Based on the observation that the pupil responses of different subjects to the same video clips have similar patterns, we adopt a nonparametric method to estimate pupillary light reflex using principle component analysis [Soleymani *et al.*, 2012]. The first principal component of observation matrix containing pupil diameter data of the same video clip from different subjects is used to estimate the light reflex. After subtracting the light reflex from the original data, the residual part contains emotional pupil response in addition to noise. For EEG signals, a band-pass filter between 1 and 75 Hz is applied to reduce the artifacts and drift. After filtering, we down-sample the EEG signals to 200 Hz to reduce the computational complexity.

2.2 Feature Extraction

EEG Signals

Here we extract and compare two kinds of efficient features from EEG, power spectral density (PSD) and differential entropy (DE). The spectral power of EEG signals in different frequency bands have been shown to be highly correlated with emotions [Lin *et al.*, 2010]. Besides, we extract the differential entropy features [Duan *et al.*, 2013]. According to [Duan *et al.*, 2013], for a fixed length EEG sequence, DE feature is equivalent to the logarithm of PSD in a certain frequency band. Short-term Fourier transform (STFT) with a 4s

non-overlapping window is used to compute the power spectral density in five frequency bands, i.e., delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz) and gamma (31-50 Hz) for each channel. The total dimension of EEG features for a sample of 62 electrodes is 310.

Eye Movement Signals

After eliminating the light reflex in the pupil diameter, the PSD and DE features are computed for the pupil diameter in X and Y axes using STFT in four frequency bands (0-0.2 Hz, 0.2-0.4 Hz, 0.4-0.6 Hz, and 0.6-1 Hz) [Soleymani *et al.*, 2012]. Moreover, conventional features (mean and standard deviation) are also extracted from pupil diameter. The PSD (or DE) feature dimension of pupil diameter is 12.

Eye movement parameters	Extracted features
Pupil diameter (X and Y)	Mean, standard deviation and PSD (or DE) in four bands: 0-0.2 Hz, 0.2-0.4 Hz, 0.4-0.6 Hz, 0.6-1 Hz
Dispersion (X and Y)	**Mean, standard deviation**
Fixation duration (ms)	**Mean, standard deviation**
Blink duration (ms)	Mean, standard deviation
Saccade	**Mean, standard deviation of saccade duration (ms) and saccade amplitude** ($°$)
Event statistics	Blink frequency, **fixation frequency, fixation duration maximum, fixation dispersion total, fixation dispersion maximum, saccade frequency, saccade duration average, saccade amplitude average, saccade latency average**[1].

Table 1: The details of features extracted from eye movement signals. (To our best knowledge, these features indicated by bold type are new eye movement features, which are firstly studied for emotion recognition in this paper.)

In addition to pupil diameter which has been studied in [Soleymani *et al.*, 2012; Zheng *et al.*, 2014], we systematically investigate other fifteen new eye movements. For eye fixation, dispersion in X and Y axes (which is the small deviation of the fixation point) and fixation duration (ms) are extracted. Blink is also shown to have relation with emotional states [Soleymani *et al.*, 2012], so blink duration (ms) is used as a useful feature. Saccade interpreted as the fast movement of eye when it makes a sudden change of fixation point, is generally dependent on the content of interest and has not been used for emotion recognition yet. We extract saccade duration (ms) and amplitude (degree) to examine whether they are relevant to emotions. The mean and standard deviation are computed from all extracted features mentioned above. Another nine event detailed statistics for each trial, such as blink frequency, and saccade frequency, are added to

_____

[1]Saccade latency average is the average value of the next saccade start time minus the last saccade end time.

the feature set, as well. Ultimately, the total number of different features from eye movement signals is 33. A detailed summary of the extracted features is presented in Table 1.

The features we extract from eye movement and EEG signals usually have strong fluctuations. Since emotions change gradually in general, we apply the linear dynamic system (LDS) approach [Shi and Lu, 2010] with the window of 20s to filter out the unrelated features for emotion recognition.

2.3 Emotion Recognition

For emotion recognition based on single modality, we adopt support vector machine with linear kernel as classifier. For evaluation, we use the data from the first nine trials as training data and the data from remaining six trials as testing data in the whole experiment. After two classifiers on eye movement and EEG data are trained, different modality fusion strategies are used to combine them. The fusion strategies could be divided to two main categories: feature level fusion (FLF) and decision level fusion (DLF). At feature level, the eye movement feature vector and the EEG feature vector are concentrated into a larger feature vector. At decision level, the classification outcomes from the two classifiers are combined to obtain the final decision. In our work, we adopt FLF and DLF to compare the performance of different fusion strategies.

For DLF, maximal rule and sum rule are often used due to their simplicity: no need of training. Given the outputs of each classifier, the maximal (sum) rule is to compute the maximal (sum) values of all the probabilities that a sample belongs to each category in all classifiers and choose the class label with the highest probability.

The rules mentioned above rely on the assumption that all the classifiers are mutually independent, which is inconsistent with the real situation. Thus, the predicted result of these rules are inaccurate to some degree. Therefore, we adopt an advanced fusion strategy called fuzzy integral [Murofushi and Sugeno, 1989]. The fuzzy integral is integrals of a real function with regard to fuzzy measures.

Definition 1. A fuzzy measure μ defined on a finite index set $X = \{x_1, x_2, \ldots, x_n\}$ is a set function $\mu : \mathcal{P}(X) \to [0, 1]$ ($\mathcal{P}(X)$ is the power set of X) satisfying:

1. $\mu(\emptyset) = 0, \mu(X) = 1$,

2. $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$.

In this paper, we adopt the discrete Choquet integral [Murofushi and Sugeno, 1989].

Definition 2. Let μ be a fuzzy measure on X. The discrete Choquet integral of a function $f : X \to \mathbb{R}^+$ with respect to μ, is

$$
\begin{aligned}
&\mathcal{C}_\mu(f(x_1), f(x_2), \ldots, f(x_n)) \\
&:= \sum_{i=1}^{n} [f(x_{(i)}) - f(x_{(i-1)})] \mu(A_{(i)}),
\end{aligned}
\tag{1}
$$

where $\cdot_{(i)}$ presents the permuted indices to satisfy $0 \leq f(x_{(1)}) \leq f(x_{(2)}) \leq \cdots \leq f(x_{(n)}) \leq 1$. Also $f(x_{(0)}) = 0$ and $A_{(i)} := \{x_{(i)}, x_{(i+1)}, \ldots, x_{(n)}\}$.

Let $C_1, C_2, \ldots, C_m$ be m classes and $X^T = [x_1 \ldots x_n]$ be a n-dimensional vector. There are n classifiers, one for each attribute x_i, which provide a confidence value denoted

by $\Phi_i^j(X^\circ)$ for an unknown sample X° in the statement "X° belongs to class C_j", for all C_j.

To integrate all the confidence values of n classifiers, a fuzzy integral is used. The global confidence value in the statement "X° belongs to class C_j" is given by

$$
\Phi_{\mu^j}(C_j; X^\circ) := \mathcal{C}_{\mu^j}(\Phi_1^j, \Phi_2^j, \ldots, \Phi_n^j),
\tag{2}
$$

where μ^j ($j \in \{1, 2, \ldots, m\}$) are defined on the set of attributes (or classifiers) and represent the importance of the classifiers. Ultimately, X° is predicted to be in the class with the highest confidence value.

The goal is to learn the fuzzy measure μ, which has $m(2^n - 2)$ coefficients. Suppose the number of classes is 2 (i.e. $m = 2$) for the sake of simplicity. Then there are $l = l_1 + l_2$ training examples labelled $X_1^j, X_2^j, \ldots, X_{l_j}^j$, $j = 1, 2$. We can compute μ by minimizing error J,

$$
\begin{aligned}
J = &\sum_{k=1}^{l_1} (\Phi_{\mu^1}(C_1; X_k^1) - \Phi_{\mu^2}(C_2; X_k^1) - 1)^2 \\
&+ \sum_{k=1}^{l_2} (\Phi_{\mu^2}(C_2; X_k^2) - \Phi_{\mu^1}(C_1; X_k^2) - 1)^2.
\end{aligned}
\tag{3}
$$

This reduces to a quadratic optimization problem with $2(2^n - 2)$ variables and $2n(2^{n-1} - 1)$ constraints which can be written in the following form:

$$
\begin{aligned}
&\text{minimize } \frac{1}{2} u^T D u + \Gamma^T u \\
&\text{under the constraint } Au + b \geq 0
\end{aligned}
\tag{4}
$$

where u is a $2(2^n - 2)$ dimensional vector including all of the fuzzy measures μ^1, μ^2, i.e. $u := [u_1^T u_2^T]^T$, with

$$
\begin{aligned}
u_j := &[\mu^j(\{x_1\}) \mu^j(\{x_2\}) \ldots \mu^j(\{x_n\}) \\
&\mu^j(\{x_1, x_2\}) \ldots \mu^j(\{x_{n-1}, x_n\}) \ldots \\
&\mu^j(\{x_2, x_3, \ldots, x_n\})]^T.
\end{aligned}
\tag{5}
$$

After computing an appropriate set of fuzzy measures, we can use them to represent the importance of each classifier and the relative importance between any subset of the classifiers. Fuzzy integral can implement maximal and sum rule with certain fuzzy measures. Furthermore, it can learn an optimized set of fuzzy measures according to different individuals, which is much more precise and reasonable.

3 Experiment Setup

Previous studies have already tested the reliability of film clips to elicit emotions [Schaefer *et al.*, 2010]. In our work, since all the subjects are Chinese students, we use popular Chinese movie clips to elicit emotions effectively. Three emotional states are designed to be induced, i.e., positive, negative and neutral. We select the clips with highly emotional contents and ensure the integrity of the plot within the clips to avoid subjects' confusion. A preliminary study is conducted to select the movie clips, where twenty participants are asked to assess the materials with five point scales. Finally, fifteen movie clips (five clips per emotion) with the average

score all higher than or equal to 3 points and ranked top 5 in each category, are chosen. Each movie clip lasts about four minutes. The finally chosen sources of movie clips include *Tangshan Earthquake, Back to 1942, Just Another Pandora's Box, Flirting Scholar, and World Heritage in China.*

Fifteen video clips are totally used for each experiment. There are a 5s hint for starting, a 45s self-assessment and a 15s rest in each trial. During self-assessment, subjects are asked to report their emotional reactions in questionnaires by scoring between 1 and 5 for each trial. The trials with score below 3 points should be discarded because the subjects fail to elicit the corresponding emotion or the aroused emotion is not strong enough.

Nine health, right-handed subjects (5 females) participate in the experiment. Each of them takes part in the experiment for three times at an interval of about one week and there are totally 27 experiments evaluated here. All the subjects are undergraduate or graduate students aged between 20 and 24 years old with normal or corrected-to-normal vision, none of whom have any history of mental disease or drug use.

Before experiment, they are informed of the purpose, procedure of the experiment and the harmlessness of the equipment. We also advise subjects to sit comfortably and still in order to reduce the interference of artifact on EEG signals. The experiments are performed in the morning or early in the afternoon to avoid sleepiness. Eye movement signals are recorded using SMI ETG eye tracking glasses[2]. EEG signals are recorded with a 1000 Hz sampling rate using ESI NeuroScan System. A 62-channel electrode cap is placed on the scalp of the subject according to the international 10-20 electrode system. The dataset used in this paper will be freely available to the academic community via the website[3].

4 Experiment Results

4.1 Eye Movement Based Emotion Recognition

We first compare the performance of the PSD and DE features of pupil diameter in order to evaluate their discrimination ability for different emotional states. The average accuracy of the DE features (M = 57.40%, SD = 17.92) is slightly higher than that of PSD features (M = 52.75%, SD =21.59), which achieve a comparative accuracy for three emotions. Moreover, we investigate the relationship between pupil diameters and different emotions. We find that PD is largest in negative state and second largest in positive state, while smallest in neutral state for a majority of experiments, which are consistent with previous findings [Partala and Surakka, 2003; Zheng *et al.*, 2014]. In addition, we use one way analysis of variance (ANOVA) to study the statistical significance of PD in different categories of emotions. It shows that the difference between the means of pupil diameters in different emotion categories is significant ($p < 0.05$). These results indicate that PD is a stable measure of emotional activation.

The one way ANOVA is also used to study the statistical significance of the feature group: dispersion (*X* and *Y*), fixation duration, blink duration, saccade and event statistics. We

find that except for blink duration, the difference between the means of these features in different categories is significant ($p < 0.05$). The box plots of four features, namely, dispersion of *X*, saccade amplitude, saccade duration, and fixation duration, are shown in Figure 1. The dispersion is lower in neutral category, while saccade amplitude and duration are higher in neutral category. Fixation duration is the lowest in positive state and the highest in negative state.
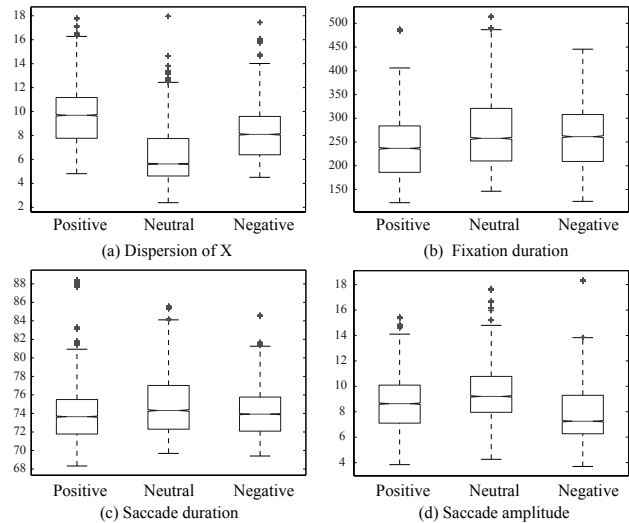

Figure 1: Box plots of four eye movement features for three emotional states. The differences between the means of these features in different categories are found significant ($p < 0.05$) with one way ANOVA.

The difference between the means of blink duration in different categories is not found significant, which means blink duration does not have significant difference in different emotional states. In addition, the average recognition accuracies of the total 33-dimensional eye features with blink duration (77.06%) is slightly lower than that of the 31-dimensional eye features without blink duration (77.80%). This indicates that blink duration does not have much contribution to improving recognition accuracy, so we do not utilize this kind of feature in the following data analysis.

The recognition performance of dispersion (*X* and *Y*), fixation duration, saccade and event statistics are shown in Table 2. As we can see, the accuracies of all the feature groups are higher than 45%, above the random level 33.3%, a lot. This indicates that all these features have potential emotion discrimination ability, considering the low dimensionality of each feature. The classification accuracies of event statistics, saccade and PD are the first three highest. These results indicate that they have a relatively higher discrimination ability compared to the other two feature groups. Finally, we concentrate all the eye movement features listed in Table 2 and study its discrimination ability. The best average classification accuracy of the total 31-dimension feature achieves 77.80% (SD = 14.61). This promising result suggests that these eye movement features can be used to discriminate different emotions effectively.

	PD	Dispersion	Fixation	Saccade	Event	FLF
Ave.	57.40	50.17	47.32	57.95	62.26	**77.80**
Std.	17.92	16.99	16.79	21.51	23.82	**14.61**

Table 2: Classification accuracies (%) of different eye movement features. ('FLF' means feature level fusion by combining all the eye movement features)

4.2 EEG Based Classification

Table 3 shows the performance of different features from different frequency bands. From Table 3, we can see that the average accuracies of beta and gamma bands for each feature are significantly higher than the other bands, which indicates that beta and gamma bands are more informative and suitable for emotion recognition than the other bands. Furthermore, the DE features achieve higher average classification accuracies and lower standard deviations than the PSD features. This implies that the suitability of the DE features for EEG-based emotion recognition. As DE features from total frequency bands have the most prominent and stable performance among all the features, we choose the DE features of EEG data from five total frequency bands to fuse with eye movement data.

Feature		Delta	Theta	Alpha	Beta	Gamma	FLF
PSD	Ave.	60.62	60.18	61.69	**72.10**	69.46	65.35
	Std.	17.23	17.72	19.04	**15.50**	18.53	18.41
DE	Ave.	69.14	63.39	68.07	78.33	77.48	**78.51**
	Std.	14.45	16.94	16.94	**13.11**	16.60	14.32

Table 3: Classification accuracies (%) of different features and their feature level fusion from EEG.

4.3 Performance of Modality Fusion

In this section, we combine eye movement signals and EEG data to enhance emotion recognition accuracy. The performance of single modality and different modality fusion strategies is shown in Figure 2. The performance obtained by all the models with modality fusion outperforms that based on single modality, which indicates that modality fusion can combine complementary information in each single modality and effectively enhance the performance of emotion recognition. Fuzzy integral achieves the best performance with the average accuracy of 87.59%, which is nearly ten percent higher than the single modality. The average accuracy of FLF is ranked second, followed by sum and max rules. Fuzzy integral computes an optimal set of fuzzy measures to fuse two single modality for each subject, so it is much more precise compared with max and sum rules. Directly concatenating the two feature vectors of eye movements and EEG into a larger feature vector can also achieve comparative performance.

The confusion matrices of each modality are shown in Table 4, which gives details of the strength and weakness of each modality. Each row of the confusion matrix represents the target class and each column represents the predicted class. The element (i, j) is the percentage of samples in class i that is classified as class j. From all the sub-tables in Table 4, we can see that the positive class is generally recognized with
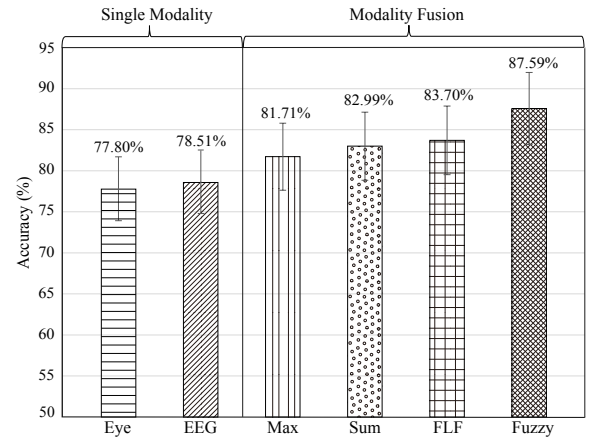

Figure 2: Performance of each single modality and different modality fusion strategies.

very high accuracy of ninety percent, while negative emotional state is the most difficult class to be recognized with the lowest accuracy. Figure 3 demonstrates that eye movement and EEG modalities have important complementary characteristics. For eye movement modality, positive state is confused with negative state (15%) and vice versa (17%), and negative state is confused with neutral state (15%). For EEG modality, negative state is often confused with neutral state (34%) and vice versa (15%). Neutral and negative states can be recognized with higher accuracy in eye movement modality than in EEG modality. Thus, it is expected that these two modalities can be complementary to improve the performance for recognizing each emotional state.
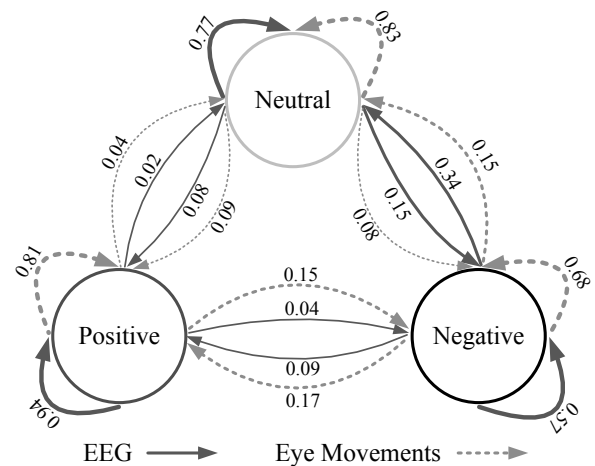

Figure 3: Confusion graph of eye movements and EEG, which shows their complementary characteristics for emotion recognition. (The numbers is the percentage of samples in class (arrow tail) that is classified as class (arrow head). Bolder lines mean higher values.)

From Tables 4(c), 4(d), 4(e), and 4(f), it can be observed that the classification accuracies of multimodal systems are

	Positive	Neutral	Negative
Positive	**0.81**	0.04	0.15
Neutral	0.09	**0.83**	0.08
Negative	0.17	0.15	**0.68**

(a) Eye Movements

	Positive	Neutral	Negative
Positive	**0.94**	0.02	0.04
Neutral	0.08	**0.77**	0.15
Negative	0.09	0.34	**0.57**

(b) EEG

	Positive	Neutral	Negative
Positive	**0.90**	0.02	0.08
Neutral	0.07	**0.84**	0.09
Negative	0.10	0.15	**0.75**

(c) FLF

	Positive	Neutral	Negative
Positive	**0.84**	0.03	0.13
Neutral	0.06	**0.87**	0.07
Negative	0.13	0.16	**0.71**

(d) Max Rule

	Positive	Neutral	Negative
Positive	**0.95**	0.02	0.03
Neutral	0.05	**0.88**	0.07
Negative	0.10	0.26	**0.64**

(e) Sum Rule

	Positive	Neutral	Negative
Positive	**0.96**	0.00	0.04
Neutral	0.10	**0.81**	0.09
Negative	0.02	0.13	**0.85**

(f) Fuzzy Integral

Table 4: Confusion matrices of each single modality and different modality fusion strategies.

higher than those of unimodal systems in most cases. For max and sum rules, they confuse negative state with neutral state (16% and 26% respectively), but they can classify neutral state with higher accuracy. FLF strategy makes fewer mistakes when classifying negative state compared to max and sum strategies. For fuzzy integral strategy, since eye movement and EEG modalities both misclassify negative state with neutral state a lot, fuzzy integral strategy also misclassify these two states (13%), but much less than that of single modality. The accuracy of negative state is much higher than others. As a result, the overall classification accuracy of fuzzy integral strategy is the highest among all the fusion strategies.

5 Discussion

The experimental results show that the performance of multimodal system is better than unimodal system with an improvement of almost 10 percent. it indicates that fusing different kinds of signals to recognize emotion is feasible and promising. The confusion matrices reveal that the ability to recognize different emotions is different between eye movements and EEG. Positive emotion is confused with negative emotion in eye movement domain while positive emotion can be classified with higher accuracy in EEG domain. Negative and neutral emotions are usually misclassified in EEG domain, while eye movements are quite good at recognizing neutral emotion. Therefore, it is reasonable to expect that there are complementary information for eye movements and EEG. This is why fusion of the two modalities can achieve higher accuracies than single modality.

For multimodal emotion recognition based on eye movements and EEG, a similar recent work is reported by [Zheng *et al.*, 2014]. However, in their work, they extracted only pupil diameter features and utilized simple fusion methods to combine two modalities. In contrast, in addition to pupil diameter, our work investigate fifteen new eye movement features such as eye saccade, fixation and dispersion and analyze the intrinsic patterns of these eye movement features for different emotions. Moreover, we introduce more advanced fusion strategies to improve the performance and achieve a significantly improved recognition accuracy (87.59%). The experimental results indicate the efficiency of the extracted eye movement features and the superiority of the multimodal methods for emotion recognition.

In this paper, we study three discrete categories of emotion and label the EEG data of each trial as a discrete steady emotion. However, emotion in real world is much more complex and it is a function of time, context, space, culture, and person [Kim and André, 2008]. In other words, emotion recognition indeed is a regression problem, instead of a classification problem. Here, we simplify the problem with certain restricted conditions. To recognize emotional states more precisely, emotion recognition from regression perspective should gain considerable research attention. Due to the fuzzy boundaries of emotion, the challenging problem is how to obtain the 'ground truth' of emotion.

6 Conclusion and Future Work

This paper has shown that combining eye movements and EEG can considerably improve the performance of emotion recognition systems. The experiment results demonstrate that pupil diameter, dispersion, fixation duration, saccade duration, saccade amplitude and nine event statistics are distinguishable for three emotions, which could be used as efficient features for emotion recognition. We have revealed that the characteristics of eye movements and EEG are complementary to emotion recognition. Modality fusion can significantly enhance the emotion recognition accuracy in comparison with single modality. The best accuracy achieved by the fuzzy integral fusion strategy is 87.59%, whereas the accuracies of solely using eye movements and EEG are 77.80% and 78.51%, respectively. The promising accuracy shows the advantages of combining eye movements and EEG.

With the fast development of wearable dry EEG sensors [Chi *et al.*, 2012], it is now practical to implement brain computer interfaces from laboratory to real-world environments. Therefore, as future work, we will consider experiment scenarios for real-world applications, instead of virtual scenarios and simulated stimuli. For example, we are going to develop a novel approach for estimating students feelings in real-time from EEG and eye movements while they attend MOOC.

References

[Antos and Pfeffer, 2011] Dimitrios Antos and Avi Pfeffer. Using emotions to enhance decision-making. In *IJCAI*, volume 22, pages 24–30, 2011.

[Bradley *et al.*, 2008] Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45:602–607, 2008.

[Calvo and D'Mello, 2010] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[Chanel *et al.*, 2011] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 41(6):1052–1063, 2011.

[Chen *et al.*, 1998] Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. Multimodal human emotion/expression recognition. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 366–371. IEEE, 1998.

[Chi *et al.*, 2012] Yu Mike Chi, Yu-Te Wang, Yijun Wang, Christoph Maier, Tzyy-Ping Jung, and Gert Cauwenberghs. Dry and noncontact EEG sensors for mobile brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(2):228–235, 2012.

[Duan *et al.*, 2013] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering*, pages 81–84. IEEE, 2013.

[Kim and André, 2008] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.

[Lin *et al.*, 2010] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.

[Murofushi and Sugeno, 1989] Toshiaki Murofushi and Michio Sugeno. An interpretation of fuzzy measures and the choquet integral as an integral with respect to a fuzzy measure. *Fuzzy sets and Systems*, 29(2):201–227, 1989.

[Nasoz *et al.*, 2010] Fatma Nasoz, Christine L Lisetti, and Athanasios V Vasilakos. Affectively intelligent and adaptive car interfaces. *Information Sciences*, 180(20):3817–3836, 2010.

[Nie *et al.*, 2011] Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu. EEG-based emotion recognition during watching movies. In *5th International IEEE/EMBS Conference on Neural Engineering*, pages 667–670. IEEE, 2011.

[Partala and Surakka, 2003] Timo Partala and Veikko Surakka. Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, 59(1):185–198, 2003.

[Picard, 2000] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[Ptaszynski *et al.*, 2009] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka, and Kenji Araki. Towards context aware emotional intelligence in machines: computing contextual appropriateness of affective states. In *IJCAI*, pages 1469–1474, 2009.

[Rayner, 2009] Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

[Salovey and Mayer, 1989] Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1989.

[Schaefer *et al.*, 2010] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7):1153–1172, 2010.

[Shi and Lu, 2010] Li-Chen Shi and Bao-Liang Lu. Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning. In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6587–6590. IEEE, 2010.

[Soleymani *et al.*, 2012] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2012.

[Tanguy *et al.*, 2007] Emmanuel Tanguy, Philip J Willis, and Joanna Bryson. Emotions as durative dynamic state for action selection. In *IJCAI*, volume 7, pages 1537–1542, 2007.

[Verma and Tiwary, 2014] Gyanendra K Verma and Uma Shanker Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102(0):162–172, 2014.

[Wang *et al.*, 2014] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.

[Zheng *et al.*, 2014] Wei-Long Zheng, Bo-Nan Dong, and Bao-Liang Lu. Multimodal emotion recognition using eeg and eye tracking data. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5040–5043. IEEE, 2014.