Multimodal Emotion Recognition from Eye Image, Eye Movement and EEG Using Deep Neural Networks

Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao and Bao-Liang Lu* *Senior Member, IEEE*

***Abstract*— In consideration of the complexity of recording electroencephalography(EEG), some researchers are trying to find new features of emotion recognition. In order to investigate the potential of eye tracking glasses for multimodal emotion recognition, we collect and use eye images to classify five emotions along with eye movements and EEG. We compare four combinations of the three different types of data and two kinds of fusion methods, feature level fusion and Bimodal Deep AutoEncoder (BDAE). According to the three-modality fusion features generated by BDAE, the best mean accuracy of 79.63% is achieved. By analyzing the confusion matrices, we find that the three modalities can provide complementary information for recognizing five emotions. Meanwhile, the experimental results indicate that the classifiers with eye image and eye movement fusion features can achieve a comparable classification accuracy of 71.99%.**

I. INTRODUCTION

With the current boom of Human-Computer Interaction (HCI), recent research has been devoted to enhancing computers with multiple abilities such as action recognition, emotion recognition, and natural language processing to build better interactions between computers and users. Emotion recognition plays an important role in HCIs because of its great potential for applications. For instance, a car can monitor the driver's emotional state by analyzing electroencephalography (EEG) and eye movements [1].

EEG is widely used to recognize human emotions. Nie *et al.* obtained an accuracy of 89.22% while classifying two emotions elicited by film clips [2]. Emotional states elicited by music videos have also been studied. Researchers tried various EEG preprocess methods for classification. Differential entropy features are found suitable for emotion recognition tasks [3]. Zheng and Lu classified three kinds of emotions with EEG signals in diverse frequency bands and found that gamma band was suitable for emotion classification [4]. In addition to EEG, eye movements can also reflect emotions because they represent users' external features. Researchers analyzed eye movements in a variety of experiments and demonstrated the existence of correlations between eye movements and emotional states [5]. Partala *et al.* evaluated pupillary responses to emotionally provocative sound stimuli, including negative, positive and neutral clips [6].

In order to improve the performance of emotion recognition, Zheng *et al.* combined EEG signals and pupillary response using feature level fusion strategy and decision level fusion strategy [7]. Lu *et al.* further extracted 33 different features from eye movement data, fused 62 channel EEG signals and achieved 87.59% accuracy on the classification of three emotions [8]. In addition, multiple fusion strategies using multimodal deep learning have been developed with the combination of EEG and eye movement features [9].

Many researchers also attempt to extract emotional features from facial images by using deep learning [10], which has been a well-studied topic in the domain of computer vision. However, its real-world application scenarios are limited since the system requires a fixed camera to record the users' full facial images. In contrast, Hickson *et al.* used portable eye-tracking cameras to record images of eyes and surrounding areas as input for expression recognition [11]. Their results prove that eye image, which can be collected easily in real time, is an effective modality for emotion recognition.

In this paper, we integrate eye images, eye movements and EEG into different combinations to investigate their properties for emotion recognition. Our main contributions to multimodal emotion are as follows: (1) we adopt a deep model of combining Convolutional Neural Network (CNN) [12] and Long Short-Term Memory Network (LSTM) [13] to extract high-level features from eye images; (2) we evaluate two fusion features on five-class emotion classification; (3) we explore the feasibility of a portable emotion recognition device based on eye movements and eye images. (4) we record eye images to form a subset of SEED V [14];

II. METHOD

A. Feature Extraction

For SEED V dataset, Differential Entropy (DE) features are extracted from EEG signals using a 256-point Short-time Fourier transform (STFT) with 4 s non-overlapping Hanning window [3]. These features are divided into five frequency bands: δ (1-4 Hz), θ (4-8 Hz), α (8-14 Hz), β (14-31 Hz), and γ (31-50 Hz). In this way, at every time step, we have DE features of 62 channels, each of which contains data in 5 frequency bands. Finally, the features extracted from EEG contains 310 dimensions in total. As for eye movement features, the same method as in [8] is applied to extract thirty-three-dimension features, including pupil

This work was supported in part by grants from the National Key Research and Development Program of China (Grant No. 2017YFB1002501), the National Natural Science Foundation of China (Grant No. 61673266), and the Fundamental Research Funds for the Central Universities.

Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao and Bao-Liang Lu are with the Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering and the Brain Science and Technology Research Center, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China.

* Corresponding author (bllu@sjtu.edu.cn)

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

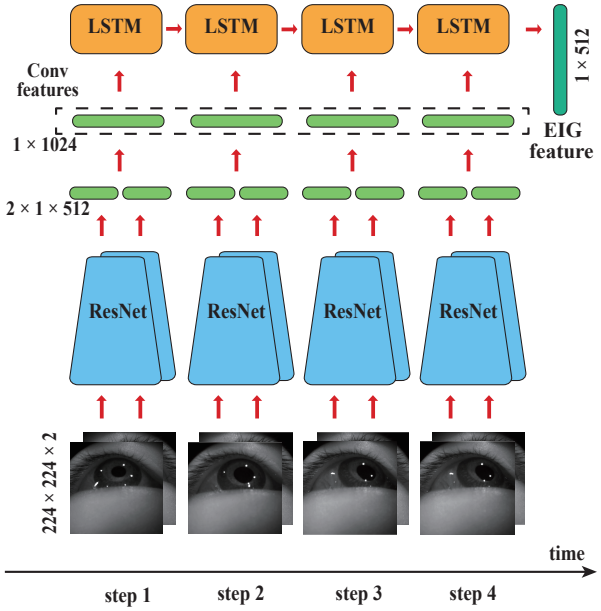

Fig. 1. The deep neural network model for extracting high-level features (EIG features) from eye images by combining ResNet and LSTM. For each equivalent step of LSTM, two ResNets are used to extract Conv features from left and right eye images.

diameter, dispersion, and so on. All the extracted features are rescaled to (0-1) and we have one sample every 4 seconds after processing.

We adopt a deep neural network model of combining CNN and LSTM networks as shown in Fig. 1 to extract high-level features from eye images. This method has been proposed in action recognition research based on image sequence, where CNN is used to do dimensionality reduction on images and LSTM is used to extract temporal features and recognize the exact action [15]. Inspired by its high performance on processing image sequence, we adopt this network structure to extract temporal features of the eye image sequence and adjust its sampling rate.

We first apply two deep residual networks (ResNet) [16] pre-trained on ImageNet to reduce the dimensions of both left and right eye images, respectively. The input size of a ResNet is fixed to 224×224 while the original size of each image is 320×240. Hence, there is a need for image resizing. We integrally rescale the images from 320×240 to 298×224 to fit the bill for height. And for the width, we randomly crop each one to get a 224×224 image in the training process and chose the middle part in the testing process. In this way, models trained with variously located images can get better generalization ability to face the variation.

Since we only use ResNets to do dimension reduction, we take out of the output layers to get two efficient feature extractors. Therefore, we get two 1×512 vectors (according to the structure of ResNet) after feeding the preprocessed images to ResNets. The two vectors are concatenated as Conv features (1×1024) for each pair of eye images.

Nevertheless, we have one pair of eye images every 1 second while one sample of EEG and eye movement features are obtained every 4 seconds, which means we need to standardize the frequency. It is inefficient to choose one of the four samples of eye images since the information of the others will be lost. So we introduce a two-layer LSTM network to implement, of which the time step is set to 4. The input size is 1×1024 while the output size is 1×512.

The whole process can be divided into the training phase and the feature extracting phase. In the training phase, we add a fully connected layer to the LSTM as an output layer and use the exact labels of eye images in training set to train the LSTM and fine-tune the ResNets, with four vectors of Conv features fed to it in sequence. After that, in the feature extracting part, the output layer is discarded and the final output of LSTM represented the high-level features extracted from four pairs of eye images. Therefore, we can adjust the samples of eye image features to be synchronous with eye movements and EEG, using the same labels.

This model mentioned above has the following two main features. First, it can flexibly adjust the sampling rate of eye image features without losing too much information. A simple reduction of the sampling rate is not an ideal method since the samples discarded might also contain important information. In contrast, LSTM networks can smoothly combine the information of samples together and output high-level features. Second, it can extract features that contain temporal information. Eye image samples may show emotional expression in the form of sequence, hence LSTM is a good way to extract temporal features of them.

B. Fusion Methods

In order to combine three different modalities, we adopt two fusion strategies, feature level fusion (FLF) and BDAE.

FLF is a simple method to fuse features. For each sample, we directly concatenate features of different modalities at the feature level. For example, EEG features contain 310 dimensions and eye movement features contain 33 dimensions. If we use the FLF approach to fuse them, we will get fusion features with 343 dimensions in total.

BDAE denotes Bimodal Deep AutoEncoder to extract high-level representation of features [9]. The training process contains an encoding part and a decoding part. In the encoding part, one Restricted Boltzmann Machine (RBM) is trained for each modality features. After that, hidden layers of RBMs are concatenated and used as input of an upper auto-encoder. In the decoding part, a symmetric structure is built to reconstruct the input features. As a result, the high-level features can approximately represent the original features of all the modalities.

C. Classification

Support vector machine (SVM) with a linear kernel is used as a baseline classifier to compare the performance of different modality groups and two different fusion methods. Parameter C is tuned in space $2^{[-10:10]}$. For single modality features, FLF fusion features and high-level fusion features generated by BDAE, SVM is trained directly to classify different emotions.

Fig. 2. The experimental scene in the emotion experiments.

III. EXPERIMENT AND RESULTS

A. The SEED V Datasets

We apply our methods on the SEED V dataset. The recording scenario of it is shown in Fig. 2, in which participants were sitting comfortably and the EEG, eye movements and eye images data were recorded by EEG cap and eye tracking glasses simultaneously.

- **SEED V.** The SEED V is a five-class (happy, sad, fear, disgust and neutral) dataset collected and used in [14][17]. There are totally sixteen healthy subjects (6 males and 10 females) aging from 19 to 28 are selected in total. For each emotion, 9 emotional movie clips are chosen according to the ratings of elicitation effect that subjects gave after watching clips. The durations of the video clips range from two to four minutes. The experiments contain 3 sessions, each of which consists of 15 random placed clips. But for the convenience of subsequent three-fold cross-validation, each fold is guaranteed to contain 5 clips with different labels. The EEG signals, containing 62 channels, are recorded with ESI NeuroScan System at a sampling rate of 1000 Hz. During the experiment, we additionally sample eye movement signals simultaneously. SMI ETG eye tracking glasses are used to record them, which contain information such as pupil diameter and blink. Three-fold cross-validation is adopted so that the 15 segments in each session are split into 3 parts equally. For every subject, three different SVMs are trained, each of which uses permutations of two out of three parts as training data and uses the remaining third as testing data. As a result, each SVM is trained on about 1200 samples and tested on about 600 samples.
- **Eye images.** During collecting eye movements and EEG data of the SEED V, eye images, which are recorded by eye-tracking cameras on SMI ETG glasses at a sampling rate of 1 Hz, form a subset of it. But they only show eyes and surrounding areas, which are different from full facial images. Since the data recording is simultaneous, we can use the same labels as other modalities. Eventually, we have data of three modalities for every trial in SEED V dataset.

B. Experimental results

TABLE I shows the classification results on the SEED V dataset. In the first line, EIG (eye images)+EYE (eye movements), EIG+EEG, EYE+EEG and ALL are listed as 4 combinations of modalities in total. In the first column, FLF and BDAE are listed as 2 fusion methods. Take FLF for instance, it is obvious that classification with three-modality fusion features can get the highest accuracy (73.96%). As for high-level fusion features generated by BDAE, SVM performs even better, getting the highest accuracy (79.63%). Moreover, we find that the more modalities we use in SEED V, the better result we will get since the accuracies of EIG, EYE and EEG single modalities are 58.67%, 59.66%, and 68.57%, successively.

TABLE I mainly shows two obvious phenomena. First, the more modalities we use, the better results we will get, which means two-modality features are generally better than single modality features while three-modality features perform best. This indicates that EEG, EYE, and EIG are effective modalities for emotion recognition. Also, modality fusion can combine complementary information in every single modality and effectively enhance the performance of emotion recognition. Second, classification with fusion features generated by BDAE invariably achieve better results than with FLF fusion features. This comparison clearly illustrates that BDAE can effectively combine complementary information of different modalities with deep learning algorithms, which is also proved in [9].

Fig. 3 (a)-(d) shows the confusion matrices of EIG, EYE, EEG, and ALL (generated by FLF) on SEED V, respectively. Among them, ALL features basically perform better, especially for the Disgust, Fear, and Neutral. EIG is particularly adept in recognizing Fear and Neutral emotions. In addition, the main phenomena shown in the confusion matrices of EEG and EYE are coincident with the results of [17].

The values in the diagonal lines of confusion matrices shown in Fig. 3 are the accuracies of each emotion using single modalities. In order to explore complementary characteristics of EIG, EYE, and EEG, the accuracies in diagonal lines for each emotion are evaluated. Through comparison of the results of each pair of modalities, we find that they all show strong complementary characteristics for five emotions. This means two-modality fusion features can still achieve considerable results. This observation can be also obtained from TABLE I. Although EIG is not the best modality for each emotion, it can provide additional information for

TABLE I
AVERAGE ACCURACIES (%) AND STANDARD DEVIATIONS OF DIFFERENT FEATURES AND METHODS IN SEED V.

Method		EIG+EYE	EIG+EEG	EYE+EEG	ALL
FLF	Ave.	64.34	70.54	73.47	**73.96**
	Std.	11.37	11.59	11.41	**10.94**
BDAE	Ave.	71.99	73.98	79.37	**79.63**
	Std.	7.91	9.84	7.03	**6.93**

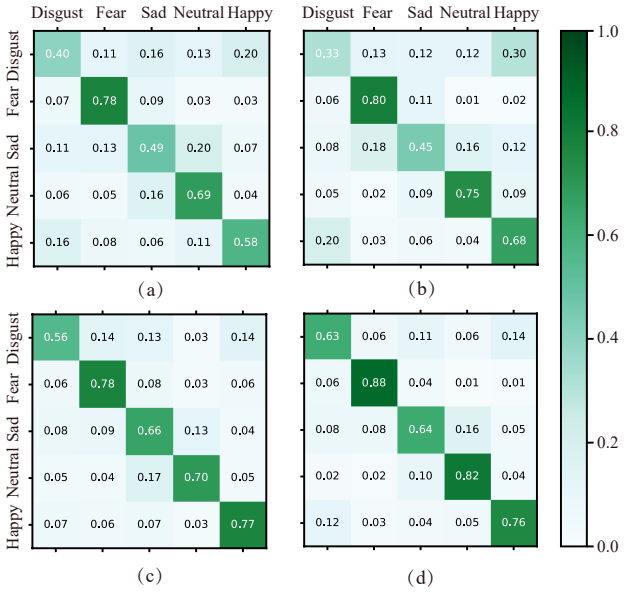

Fig. 3. The confusion matrices of single modality and combination of three modalities using SVM in SEED V: (a) EIG, (b) EYE, (c) EEG, (d) ALL.

emotion recognition by adding EIG to EEG and EYE, which means EIG can provide additional useful information for emotion recognition.

We also investigate the emotion recognition ability of newly added eye movements which focus on the classification results of EIG features. Compared with EYE and EEG, accuracies achieved with EIG features is relatively lower, which indicates that the recognition ability of this single modality is limited at present. However, by fusing with EIG, we can get more accurate results of EYE and EEG and get the best result with ALL features. Therefore, we believe EIG can be an effective assistant modality of EYE and EEG, directly providing with external features of facial expression. Although emotion recognition based on EEG has been well researched and achieves the best result among the three modalities, the recording process of EEG features used in this experiment is fairly complex. Conversely, only a pair of eye-tracking glasses is required to record eye movements and eye images, making it possible to collect data almost anywhere at any time. Moreover, we focus on the comparison of the accuracies of EYE and EIG and find that EIG, which enhances the shortage emotions of EYE, can be a great assistant modality for EYE. TABLE I shows the same results that the accuracy of EYE+EIG features generated by BDAE is 71.99%, which is fairly competitive compared with the highest accuracy of ALL features generated by BDAE (79.63%). So we believe that a real-time model can be applied on a pair of eye-tracking glasses to build a wearable emotion recognition device by using eye movements and eye images.

IV. CONCLUSIONS

In this paper, we have introduced a modality of eye image taken from eye tracking glasses into multimodal emotion recognition. We have evaluated all kinds of combinations of eye images, eye movements, and EEG for classifying five emotions. By analyzing the confusion matrices of each single modality, we have observed that every modality is adept at respective emotional recognition so that fusion features can get better accuracy with more emotional information. The experimental results indicate that the combination of eye image and eye movement data from eye tracking glasses with the deep neural network can reach a comparable accuracy, and the eye tracking glasses seem to be a promising convenient wearable device for emotion recognition in real-world applications.

REFERENCES

[1] H. Leng, Y. Lin, and L. Zanzi, "An experimental study on physiological parameters toward driver emotion recognition," in *International Conference on Ergonomics and Health Aspects of Work with Computers*. Springer, 2007, pp. 237–246.

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

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

[4] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[5] L. Nummenmaa, J. Hyönä, and M. G. Calvo, "Eye movement assessment of selective attentional capture by emotional pictures." *Emotion*, vol. 6, no. 2, p. 257, 2006.

[6] T. Partala, M. Jokiniemi, and V. Surakka, "Pupillary responses to emotionally provocative stimuli," in *Symposium on Eye Tracking Research & Applications*. ACM, 2000, pp. 123–129.

[7] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 5040–5043.

[8] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition." in *International Joint Conference on Artificial Intelligence*, vol. 15, 2015, pp. 1170–1176.

[9] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 521–529.

[10] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[11] S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. Essa, "Eyemotion: Classifying facial expressions in VR using eye-tracking cameras," *arXiv preprint arXiv:1707.07204*, 2017.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[13] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[14] T.-H. Li, W. Liu, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from EEG and eye movement signals: Discrimination ability and stability over time," *in 9th International IEEE EMBS Conference on Neural Engineering*, 2019.

[15] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Action classification in soccer videos with long short-term memory recurrent neural networks," in *International Conference on Artificial Neural Networks*. Springer, 2010, pp. 154–159.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] L.-M. Zhao, L. Rui, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from EEG and eye movement signals: Complementary representation properties," *in 9th International IEEE EMBS Conference on Neural Engineering*, 2019.