Generating Multimodal Features for Emotion Classification from Eye Movement Signals

Huang-Fei Jiang[1], Xi-Ya Guan[1], Wei-Ye Zhao[4]
Li-Ming Zhao[1], and Bao-Liang Lu[1,2,3⋆]

[1] Center for Brain-like Computing and Machine Intelligence,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China
[2] Key Laboratory of Shanghai Education Commission for
Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University,
800 Dong Chuan Road, Shanghai 200240, China
[3] Brain Science and Technology Research Center, Shanghai Jiao Tong University,
800 Dong Chuan Road, Shanghai 200240, China
[4] Robotic Institute, Carnegie Mellon University, PA, USA
{gratin, gxy, lm_zhao, bllu}@sjtu.edu.cn, weiyezha@andrew.cmu.edu

Abstract. In recent years, multimodal signals such as EEG and eye movements have been widely used and achieved a better performance than single modality in emotion classification. However, inconvenience in EEG signals collecting procedure still remains a key problem in practical applications. In this paper, we propose a novel method for generating multimodal features from eye movement signals. As a result, we could simplify classification procedure and use EEG and eye movement signals in training stage and eye movement signals in test stage. We evaluate the effectiveness of our proposed method on classification of five human emotions, which include disgust, fear, sad, neutral and happy. The experimental results indicate that our proposed method could really generate reliable multimodal representations and has a nearly comparable performance (72.80%), with multimodal models (79.70%), while it also outperforms using single eye movement signals (59.66%) and EEG signals (68.58%).

Keywords: Emotion classification, Multimodal fusion, Regressor, EEG, Eye movements

1 Introduction

Emotion classification is warmly welcomed in normal life and industrial communities. For example, it can help people realize their mental health status in daily life [8] and proper medical treatment is facilitated if emotion can be accurately recognized [7]. In recent research, emotion classification approaches widely utilize the EEG signals, since EEG signals are directly collected from our brains, and it is not easy to disguise. Li and Lu classified two kinds of emotions using gamma band of EEG signals, and their results showed that gamma band was suitable for emotion classification [4]. Wang *et al.* compared three different kinds of EEG features and a simple approach was proposed to track the trajectory of emotion changes with time [11]. Zheng and Lu classified EEG signals using a deep neural network and examined critical bands and channels of EEG signals for emotion classification [16]. To fully use the information from different modalities, Yang *et al.* proposed an auxiliary information regularized machine, which treats different modalities with different strategies [12]. With the rapid development of commercial eye tracking glasses in recent years, it is very easy to acquire eye movement signals. Various studies demonstrated that EEG and eye movements can be utilized for effective emotion classification [9] [6].

Note that providing EEG and eye movements, hidden emotion representations can be extracted from those multi-modal measurements. The multi-modal hidden representations reflect more comprehensive emotion status information, and can greatly enhance the emotion classification accuracy. Previously, complementary characteristics of EEG and eye movements have been studied, various multimodal fusion strategies have been developed and multimodal emotion classification approaches have well performance [14] [9] [15] [6] [13]. However, the procedures of collecting EEG signals are really inconvenient in a way since there are sundry steps of preprocessing such as wearing electrode cap and painting conductive paste. Therefore, it is desirable to put forward a novel method which can enhance emotion classification performance and is easy to use.

In this paper, we propose to construct a regression model to generate the multimodal representations directed from eye movement data. Providing simple eye movement signals, our method can generate more comprehensive emotion status representations, since it also utilizes the benefits from EEG signals. The intuition behind this approach is that EEG and eye movements are both the representations from people's physiological status. Therefore, they can both illustrate it when one is under certain emotions and there must be an overlapping

⋆ corresponding author

area on certain high-dimensional space. With this method, we could just collect both EEG signals and eye movement signals during the training stage and collect single eye movement signals in classification stage. And emotion classification can be done with much more portable equipment and with much higher classification accuracy compared to single modality. As comparison and supplementation to the regression model, we also implement another method by similarity analysis between eye movements in test stage and training stage [1]. In this paper, we mainly focus on discrete emotion model and perform five emotion classification tasks, including happy, sad, fear, disgust, and neutral. To the best of our knowledge, there are limited studies reported in the literature dealing with building a multimodal representation generator of projecting eye movement features into multimodal emotion representations.

2 Experiments Setting and Model Constructing

2.1 Experiment Procedure Description

The dataset we use is sourced from the work in [5]. In the experiment, sixteen subjects (6 males and 10 females, aged from 19 to 28) who turned out to be stable extroverts recognized by Eysenck Personality Questionnaire (EPQ) are selected before experiments. Meanwhile, 45 video clips are carefully selected with high possibilities of emotional arousal. Subjects are required to watch these video clips wearing EEG cap and eye tracking glasses alone in a quiet room in three sessions while each session has an interval of about one week.

2.2 Data Preprocessing

The EEG signals with Curry 7 is preprocessed and a baseline correction is carried out. Then a band pass filter (1-50 Hz) is carried out to eliminate low-frequency and high-frequency noise. In the end, the signals from EOG and FPZ channels are used to detect and remove eye movement artifacts. Before feature extraction, EEG signals are downsampled from 1000 Hz to 200 Hz, in order to speed up the data analysis procedure.

2.3 Feature Extraction

As for EEG-based emotion classification, differential entropy (DE) features have been proved to be rather effective [9] [2]. Therefore, the DE features are extracted in five frequency bands: delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-50 Hz). And a Short-time Fourier transform (STFT) is also used with 4 s non-overlapping Hanning window to extract the DE features. To smooth features, the linear dynamic system algorithm is applied, which makes features more reliable [15]. As for eye movement features, 33-dimension features are selected including the statistics of pupil diameter, dispersion, fixation duration, blink duration, saccade and other eye-movement related events [9].

2.4 Emotion Classification

Note that in experimental protocals, each subject is required to watch 15 video clips in each experiment session, we evenly divide these 15 clips into 3 parts. Each part contains 5 video clips, which involves 5 different emotion types respectively. To evaluate the classification performance, we adopt three-fold cross-validation. This cross validation follows a rule that two clips parts are concatenated as training set and the remaining third clips part is treated as testing set by turns.

In our approach, a linear SVM has been used as the emotion classification model. To find the classifier model with highest classification performance, we carefully tune the parameter C in SVM. The tuning range for C falls in a space within -10 to 10 and tuning step size is set as 1. Note that all emotion classification experiments use linear SVM. And in different classification approaches, SVM is applied on different features, which include (1) eye movement features; (2) EEG features; (3) BDAE features; (4) regression features; (5) similarity analysis features.

2.5 Multimodal Representation Fusion

Here we utilize the modality fusion method from our previous work [6] [13], where we adopt a bimodal deep auto-encoder (BDAE) to extract the high dimensional emotion representations from both EEG and eye movement data. The process of BDAE can be summarized that two Restricted Boltzmann Machines (RBMs) are built for EEG and eye movement features at encoding stage. Then we put together two hidden layers from EEG and eye movements to the auto-encoder to generate high level multi-modal representations. The decoding network is a mirror structure of encoding network, and it tries to reconstruct original EEG and eye movement data which are fed into auto-encoder.

Note that there are many multi-modal encoder options, the intuition behind using BDAE to extract multi-modal features is that, the success of high-dimensional feature decoding process demonstrates the extracted representation is in a good quality and mutually separable in feature space. Therefore, the multimodal features can be used to construct preferable emotion classification system. So it can also be regarded as an ideal multimodal regression target for our regressor.

2.6 Multimodal Representation Generator from Eye Movement Signals

As mentioned in above sections, we want to utilize the comprehensive information of multimodal features for emotion classification. However, multimodal representation extraction requires EEG signals as input, which is time-consuming in collecting. Here we propose a regression model from low-dimensional eye movement features to multimodal representations. Since the mutual features mapping between eye movement signals and multi-modal representations are already encoded into the pre-trained regression model, the generated representations' ability in reflecting human emotion hidden status will not be compromised.

We note that recently deep learning architectures have overwhelmingly outperformed the state-of-the-art in many traditional tasks. Among several methods implementing regression, deep regression neural networks (DRNNs) are more and more frequently used. Generally, DRNNs often use fully connected regression layer with linear or sigmoid activation functions instead of softmax layer [3]. Considering the five-emotion classification dataset scale, we choose a 4-layer linear neural networks as our regression architecture.

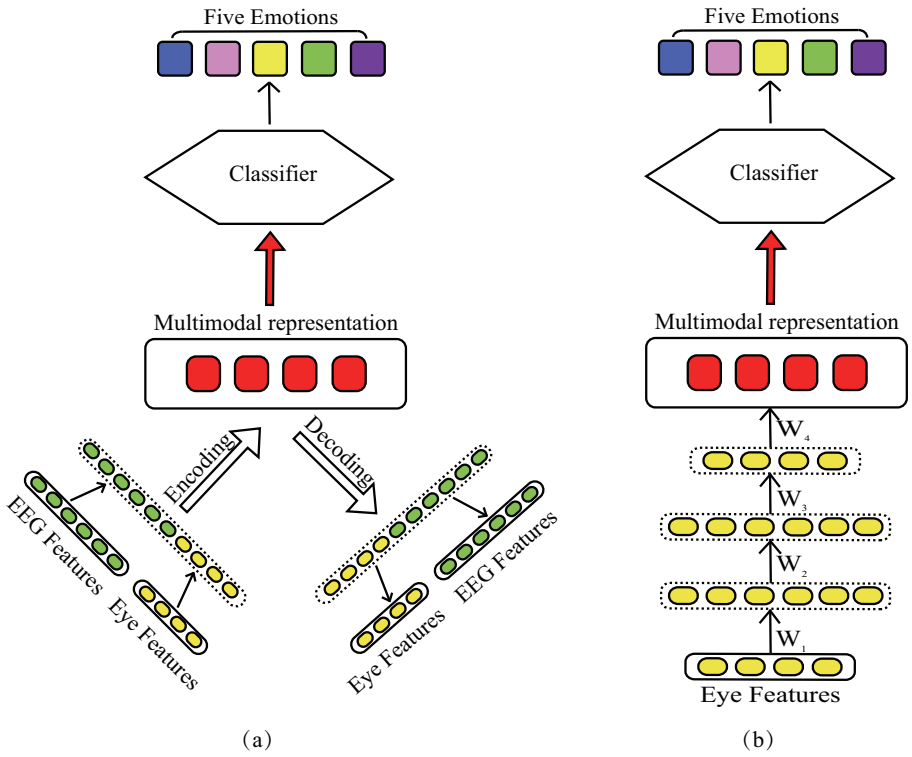

Fig. 1. Regression model. The yellow one represents eye features, the green one represents EEG features and the red one represents high-level representations. Five emotions are represented by five different colors. (a) shows procedures of reimplementation and regressor training datasets preparation. (b) shows regressor training and test procedure.

The description for the architecture is illustrated in Fig.1. Specifically, our regression procedures can be summarized that: 1) **Reimplementation.** We reimplement BDAE models by training the feature auto-encoder and SVM classifier simultaneously by feeding EEG features, eye movement features and corresponding emotion labels during training stage. 2) **Prepare training datasets in regressor.** As illustrated in Fig. 1 (a), multi-modal information is encoded into high level representations through the pre-trained BDAE auto-encoder, and we store the representations and their corresponding eye movement features as our regressor training dataset. 3) **Regressor training.** We use this constructed training dataset for regressor training, where multi-modal representation is the target and eye movement features are the input (shown in Fig. 1 (b)). What we should notice is that the classification work is directly sourced from the SVM classifier trained on the first stage. 4) **Regressor testing.** After the regressor neural network parameters converge, we put the eye movement features from testing dataset into the regressor, and get the predicted high level representation from which we get final

emotion type using trained SVM classifier. The algorithm of training and test procedure of regression model is shown in Alg. 1. From above description, it is notable that our regression model is totally independent of the

Algorithm 1: Regressor neural network training and test algorithm

Input : BDAE model, linear SVM classifier, eye movement features X_{Eye}, EEG features X_{EEG}, emotion labels Y
Output : Regressor predicted labels, classification accuracy in test stage
Variables : Test dataset size n, loss function criterion C, learning rate η, momentum α
Initialization: Regressor network parameters W, parameter update stepsize v, epoch number p, correct classification number $\kappa = 0$

Training Procedure:
for $s = 0, 1, 2, ..., p$ **do**
  Get target representation $R_{groudtruth}$ from BDAE model using both X_{Eye} and X_{EEG}.
  Get generated representation R_s with regressor parameters W and X_{Eye}.
  Construct loss $\mathbb{L}$ using $R_{groudtruth}$ and R_s with respect to C.
  Calculate network gradient $g = \nabla\mathbb{L}$.
  Use SGD with momentum α to update neural network hyperparameter:
  $v = \alpha \cdot v + \eta \cdot g$ and $W = W + v$.
end
Test Procedure: (n represents the length of test dataset)
for $i = 0, 1, 2, ..., n$ **do**
  Get sample X_{Eye_i} from test dataset.
  Generate representation R_i using trained regression model.
  Put R_i into the linear SVM classifer and get the classification result $y_{predicted_i}$.
  if $y_{predicted_i} == y_i$ **then**
    $\kappa = \kappa + 1$;
  end
end
Calculate test classification accuracy: $\mathbb{A} = \frac{\kappa}{n}$.

Table 1. Details of parameters searching methods

Hyperparameters	Description
Regressor layer1	from 500 to 5600, step size: 300
Regressor layer1	from 500 to 5600, step size: 300
Regressor layer1	from 500 to 5600, step size: 300
Loss criterion	choose from 'L1', 'MSE', 'Smooth'
Learning rate in SGD	from 0.008 to 0.014, step size: 0.001
Momentum in SGD	from 0.7 to 1.2, step size: 0.1

classifier models (SVM) and modality fusion strategies (BDAE). Therefore, our method has the advantage of high portability, where we can easily adopt it with other types of emotion classification models or multi-modal feature fusion methods.

Since eye movements that are collected from different people during different period have large variance, eye movement signals have the characteristics of individual difference. Moreover, the BDAE model also has data-specified characteristics. Therefore, we should build a regression model for each dataset. In our experiments, we find optimal hyperparameters with grid searching in parameters. Taking several experiment results and computational efficiency into accounts, we conduct hyperparameters tuning process regarding hidden neuron numbers, loss criterion method, learning rate and momentum in SGD. Detail parameters are shown in Table 1.

2.7 Similarity Analysis

Here we introduce similarity analysis method as comparison and supplementation. Since it analyzes the similarity between the eye movement feature in test stage and that in training stage which is labeled from the results in BDAE, this method takes the advantages of fusion features as well as the regressor approach. In training stage of SVM emotion classifier: 1) We first record emotion classification results as well as the corresponding eye movement features. 2) We divide eye movement features into 5 groups according to their corresponding

predicted emotion types. 3) Finally, we calculate the mean vector in each group, and catenate these 5 vectors together as the Eye-Feature-to-Emotion-Type matrix. During testing stage, we find the most similar vector A from this Eye-Feature-to-Emotion-Type matrix for currently tested eye movement feature B. So we can use the corresponding emotion type of A as the classification result of eye movement feature B.

The similarity distance between two vectors is evaluated using Minkowski Distance as below to calculate distance:

$$A = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$$
$$B = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$$
$$Distance(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

Note that when p=2, the distance becomes the Euclidean distance and it becomes city block distance when p=1. Chebyshev distance is a variant of Minkowski distance where p=∞. We define the space of parameter P, where 55 values are evenly distributed from 0.5 to 6.0 with step length of 0.1.

3 Experimental Results and Discussion

In this paper, we use single EEG signals, eye movement signals and adopt BDAE models reported from our previous work [13]. We implement these models on the five-emotion dataset and statistical results are shown as follows.

Table 2. Average accuracies (%) and standard deviations (%) of different models in classifying five emotions

Measurement	Eye	EEG	BDAE	Regressor	Similarity Analysis
Accuracy	59.66	68.58	**79.70**	72.80	50.84
Std.	8.95	10.27	7.05	**5.07**	6.43

In summary, the accuracy mean scores and standard deviations of different methods are illustrated in Table 2. We can observe that the regressor improves the classification accuracy by 13.14% compared to using single eye movement signals, and improves the classification accuracy by 4.22% compared to using single EEG signals. Besides, regressor has a comparable performance with BDAE (72.80% versus 79.70%). And it also outperforms similarity analysis method with a significant margin by 21.96%, where both methods take advantage of multimodal information.

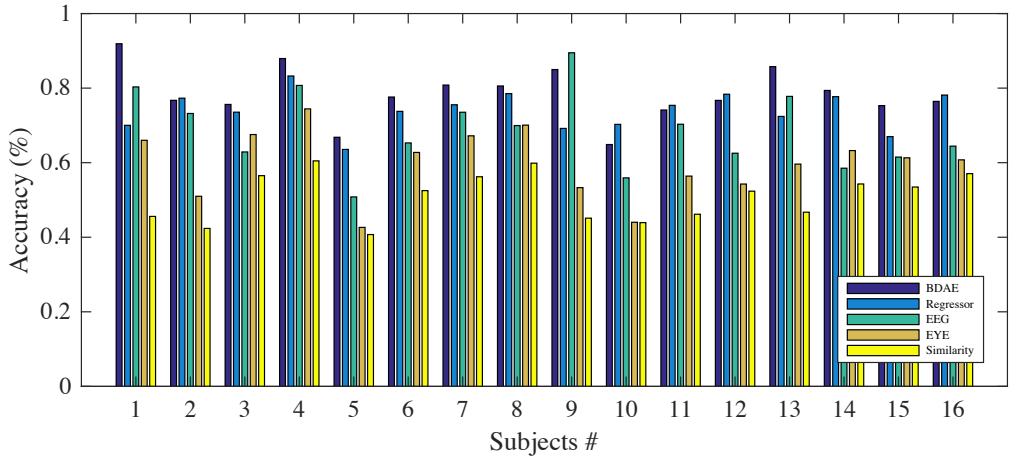

Fig. 2. Classification accuracy bar graph. Each subject has five types of accuracies corresponding to five models. From left to right shows performance of features from regressor, BDAE, similarity analysis, eye single modality and EEG single modality.

As shown in Fig. 2, regression model outperforms using single eye movement signals on all of the subjects. To be specific, the classification accuracy of Subject 2 and Subject 10 are enhanced by 26.31% and 26.26%, respectively. We also can observe that regression model's good performance is stably maintained (above 60% classification accuracy) across all subjects. However, the performance of single modality model is rather unstable. Noteworthy examples include classification accuracy of using single EEG signals (89.48% in Subject 9, but falls to 50.82% in Subject 5). Therefore, Fig. 2 serves as an intuitive representation of the low standard deviation of regression model. Moreover, accuracy in regressor also reaches BDAE model in many subjects such as Subject 8 (78.53% versus 80.58%), Subject 14 (77.71% versus 79.38%). In some subjects, regressor even slightly outperforms than BDAE model, such as Subject 2 (77.30% versus 76.73%), Subject 16 (78.14% versus 76.44%). As for the similarity analysis method, the performances are fluctuating from 46.96% ($P = 5.9$) to 50.84% ($P = 1.8$). Although this classification accuracy is much higher than random classification of five emotional states (20.00%), its performance is not satisfactory at all comparing to the regressor.

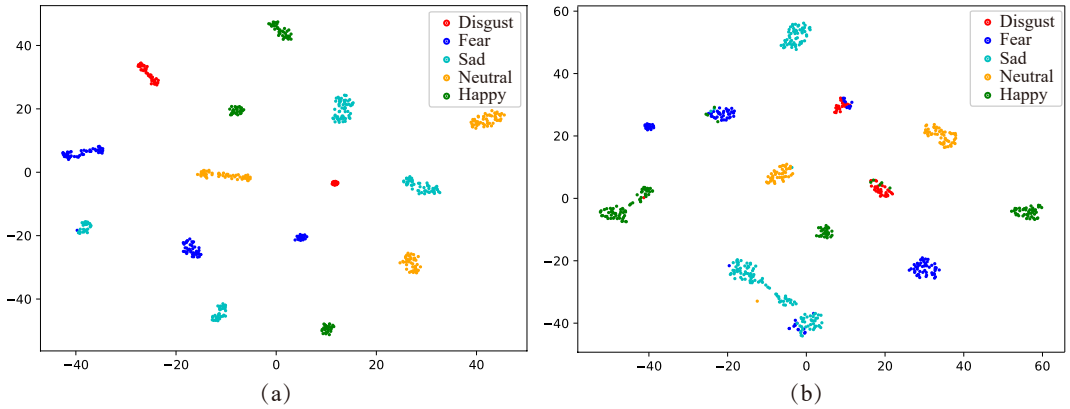

Fig. 3. Representation clusters generated from regression model with t-SNE 2D visualization. (a) BDAE representations; (b) Regressor representations. Those figures indicate the inner connections between features from different emotions, where red for disgust, blue for fear, wathet for sad, orange for neutral and grass green for happy.

To examine the reliability of representations generated from regressor, we collect the high-level representations from each cross validation in each person, which means 48 (16*3) feature matrices in all. Then we put these features into a 2D t-SNE model [10], which is able to make dimensionality reduction of the high-dimensional extracted features into two dimensions. Therefore, we can project these 2D vectors into a plane picture. Each point represents one eye movement feature from one 4s clip in a video. By unsupervised learning brought from t-SNE model, the points indicated similar features will gather together.

Fig. 3 shows two t-SNE model generated from one high level representation in 48 samples. Fig. 3 (a) shows the t-SNE graph of representations in BDAE while Fig. 3 (b) shows graph in regressor. As can be seen in these figures, although there are some slight overlapping clusters in regressor comparing to BDAE model, different representations are still generally divided into different zones. It further proves the reliability of our proposed method.

To give an intuitive cognition about classification performance between each emotion, we apply confusion matrices. The vertical axis represents the emotion label type and the horizontal axis represents the predicted type. As can be seen from the confusion matrix, regression model outperforms using single eye movement signals in all types of emotion. Accuracies of classifying disgust and neutral emotion have been improved by 10% (43% versus 33%) and 8% (83% versus 75%). Besides, it is worth noting that recognizing sad by using single eye movement signals is really unsatisfactory. But situation meets change when it comes to regression model where 34% rate has been improved.

On the other hand, we can also notice that regressor's performance reaches BDAE model from these two confusion matrices. The accuracy gap between regression model and BDAE model in fear and neutral emotion is only 2% and 6%. In happy emotion, regressor even outperforms BDAE model by 4%.

However, disgust is still an emotion always mixed with others. In eye movement modality and regression model, happy emotion is really easily to be erroneously identified as disgust emotion. It might reveal that the eye tracking features like pupil diameters or saccade details might be similar when subjects are in happy or disgust emotion condition.

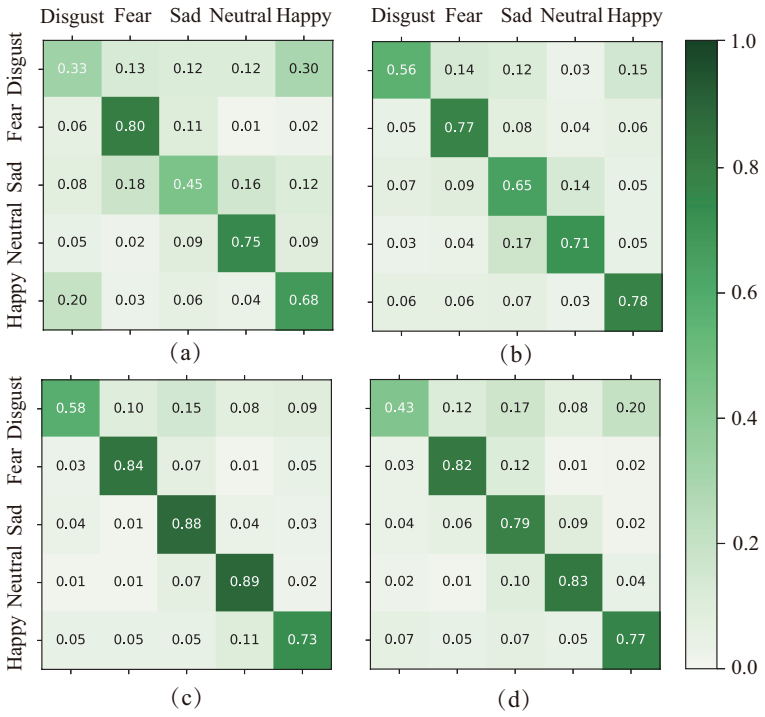

Fig. 4. Confusion matrices between five emotions: (a) Eye movements; (b) EEG; (c) BDAE; (d) Regressor. Deeper color represents higher possibilities of classification. The vertical axis represents the emotion label type and the horizontal axis represents the predicted type.

4 Conclusion and Future Work

In this paper, we have proposed an effective generator from eye movement modality to multimodal representations for five emotions classification. Since this representation also utilizes the benefits from EEG signals, it indicates a much more emotional comprehensive status. The regressor we proposed performs well with a high accuracy which incontrovertibly beats eye movements for 13.14%, and beats EEG signal for 4.22%. Regression model also has the advantage of high portability through the discussion part written above. Moreover, we also adopt one more method using eye movement signals during test stage as comparison and supplementation. And it proves the superiority of the regression model in methods under common situation.

In our future work, we could adopt our regression model into other datasets and get wider applications. Moreover, we could also concatenate it with different fusion strategies and classifiers because of the advantage of high portability. Last but not least, Generative Adversarial Networks (GAN) could also be introduced in our future work for enhancing the accuracy and reliability of our model.

5 Acknowledgement

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.

References

1. Du, C., Du, C., Huang, L., He, H.: Reconstructing perceived images from human brain activities with bayesian deep multiview learning. IEEE Transactions on Neural Networks and Learning Systems (2018)
2. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: 6th International IEEE/EMBS Conference on Neural Engineering. pp. 81–84. IEEE (2013)
3. Lathuilière, S., Mesejo, P., Alameda-Pineda, X., Horaud, R.: A comprehensive analysis of deep regression. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
4. Li, M., Lu, B.L.: Emotion classification based on gamma-band EEG. In: Annual International Conference of the IEEE Engineering in Medicine and Biology society. pp. 1223–1226. IEEE (2009)
5. Li, T.H., Liu, W., Zheng, W.L., Lu, B.L.: 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. pp. 607–610. IEEE (2019)

6. Liu, W., Zheng, W.L., Lu, B.L.: Emotion recognition using multimodal deep learning. In: International Nonference on Neural Information Processing. pp. 521–529. Springer (2016)
7. Liu, Y., Sourina, O., Nguyen, M.K.: Real-time EEG-based human emotion recognition and visualization. In: International Conference on Cyberworlds. pp. 262–269. IEEE (2010)
8. Liu, Y., Sourina, O., Nguyen, M.K.: Real-time EEG-based emotion recognition and its applications. In: Transactions on Computational Science XII, pp. 256–277. Springer (2011)
9. Lu, Y., Zheng, W.L., Li, B., Lu, B.L.: Combining eye movements and EEG to enhance emotion recognition. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
10. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(Nov), 2579–2605 (2008)
11. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing 129, 94–106 (2014)
12. Yang, Y., Ye, H.J., Zhan, D.C., Jiang, Y.: Auxiliary information regularized machine for multiple modality feature learning. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
13. Zhao, L.M., Li, R., Zheng, W.L., Lu, B.L.: Classification of five emotions from EEG and eye movement signals: Complementary representation properties. In: 9th International IEEE/EMBS Conference on Neural Engineering. pp. 611–614. IEEE (2019)
14. Zheng, W.L., Dong, B.N., Lu, B.L.: Multimodal emotion recognition using EEG and eye tracking data. In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5040–5043. IEEE (2014)
15. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: A multimodal framework for recognizing human emotions. IEEE Transactions on Cybernetics (99), 1–13 (2018)
16. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development 7(3), 162–175 (2015)