EEG-Eye Movements Cross-Modal Decision Confidence Measurement with Generative Adversarial Networks

Cheng Fei, Rui Li, Li-Ming Zhao, Wei-Long Zheng and Bao-Liang Lu* *Fellow, IEEE*

2023 11th International IEEE/EMBS Conference on Neural Engineering (NER) | 978-1-6654-6292-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/NER52421.2023.10123730

***Abstract*— Decision confidence is an individual's feeling of correctness or optimization when making a decision. Various physiological signals, including electroencephalography (EEG) and eye movements have been studied extensively in measuring levels of decision confidence in humans. While multimodal fusion generally performs better than single-modal approaches, it requires data from different modalities at a greater cost. In particular, collection of EEG data is more complicated and time consuming while eye movement signals are much easier to acquire. To tackle this problem, we propose a cross-modal method based on generative adversarial learning. In our method, the intrinsic relationship between eye movement and EEG features in a high-level feature space can be learned in the training phase, and then we can obtain multimodal information during the test phase when only eye movements are available as inputs. Experimental results on the SEED-VPDC dataset demonstrate that our proposed method outperforms single-modal methods trained and tested only on eye movement signals with an improvement of approximately 5.43% in accuracy, and maintains competitive performance in comparison with multimodal methods. Our cross-modal approach requires only eye movements as inputs and reduces reliance on EEG data, making the decision confidence measurement more applicable and practicable.**

I. Introduction

Decision confidence is an individual's feeling of correctness or optimization when making a decision and can reflect the probability of being correct [1]. As a common psychological phenomenon in real life, decision confidence is probably one of the most basic components of the decision making process, and it is also an important bridge between cognition and emotion in the decision-making process.

Studies have found that physiological signals such as eye movements and EEG can be used to estimate the confidence level of an individual during the decision-making process. Lempert *et al.* conducted a study on eye movement signals and confidence value in an auditory task and observed that the pupil diameter had a direct correlation with the levels of confidence [2]. Shooshtari *et al.* [3] designed a random dot motion (RDM) task to evaluate the correlation between confidence levels and EEG and eye movement signals and obtained eight features from these methods relative to confidence levels. Recently, Li *et al.* [4] designed a visual perception task for measuring decision confidence, and their experimental results indicate that EEG signals recorded during the experiments can distinguish different levels of decision confidence and that neural patterns of EEG signals for decision confidence in the visual perception task do exist. Sadras *et al.* [5] found that EEG classification is accurate enough to build a simulated BCI framework and that the decoded confidence could be used to improve decision making performance particularly when the task difficulty and cost of errors are high.

Many studies have shown that multimodal methods outperform single-modal methods because of the complementary features among different modalities. In the task of emotion recognition, it has been proven that complementary representation properties exist between eye movement and EEG signals [6], and multimodal data are more conducive to building a reliable and accurate emotion recognition model than single-modal data [7]. In the vigilance estimation task, Zheng *et al.* found the complementary information between forehead electrooculography (EOG) and EEG features for vigilance estimation, and demonstrated that the multimodal model has a higher recognition rate than the single-modal method [8].

While multimodal fusion generally performs better than single-modal approaches, it requires data from different modalities, which means that data acquisition is more expensive. For EEG signals, the preparation before acquisition is also time-consuming, including the correct wearing of the electrode caps and the injection of conductive gel. In contrast, eye movement signals can be collected simply by wearing an eye tracking device.

Based on the above discussion, our goal is to use information from both modalities to enhance the performance in the training stage, and to simplify the process in actual practice using only eye movements in the test stage. Inspired by generative adversarial learning [9], we propose a generative adversarial learning method to extract the intrinsic relationship between both modalities. Different from the method proposed by Cai *et al.* [9], we extract high-level representations of both modalities before generative adversarial learning. Specifically, in the training stage, high-level

This work was supported in part by grants from STI 2030-Major Projects+2022ZD0208500, the National Natural Science Foundation of China (No. 61976135), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), SJTU Global Strategic Partnership Fund, Shanghai Marine Equipment Foresight Technology Research Institute 2022 Fund (No. GC3270001/012), SJTU Global Strategic Partnership Fund (2021 SJTU-HKUST), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

C. Fei, R. Li, L. M. Zhao, W. L. Zheng and B. L. 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 Brain Science and Technology Research Center, Shanghai Jiao Tong University, 800 Dongchuan Rd., Shanghai 200240, People's Republic of China.

B. L. Lu is with the RuiJin-Mihoyo Laboratory, Clinical Neuroscience Center, RuiJin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin 2nd Rd., Shanghai 200020, People's Republic of China.

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

representations of both modalities are learned for decision confidence by a deep autoencoder, and a generator, used to generate EEG features from corresponding eye movement features is trained through generative adversarial networks (GANs). In the test stage, only eye movement signals are available as inputs, the EEG representations can be generated from the corresponding eye movement features. We evaluate our proposed method on the SEED-VPDC dataset proposed in [4] and find that it achieves superior performance compared to other single-modal models tested and trained only on eye movements.

II. Method

In our method, the training stage can be divided into two parts as high-level feature extraction and EEG high-level feature generation. More specifically, high-level features of each modality used to identify the levels of decision confidence are learned by deep auto encoder (DAE), and then GANs are trained to generate features of the EEG modality from eye movements. In the test stage, only eye movement signals are needed, and the corresponding EEG features are generated from them.

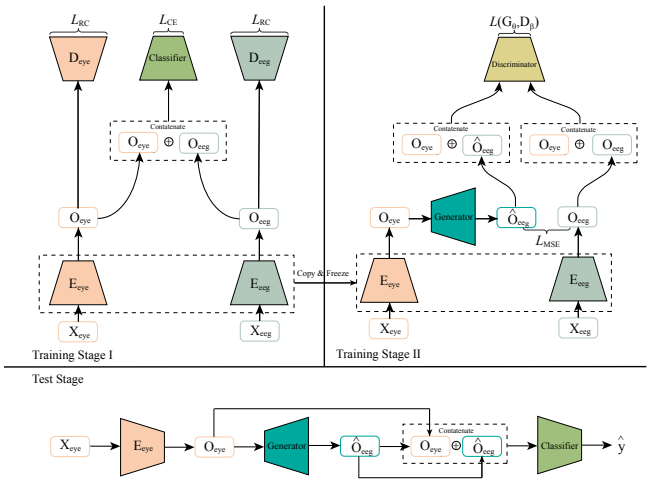

Fig. 1. The framework of our method. The training stage can be divided into two parts as high-level feature extraction and EEG high-level feature generation. Only eye movement signals are needed in the test stage.

1) Training Stage I: Feature Extraction: Assume that $X_{eye} \in \mathbb{R}^{N\times d_1}$ represents the data of eye movement signals, $X_{eeg} \in \mathbb{R}^{N\times d_2}$ represents the data of EEG signals, and N is batch-size, d_1 and d_2 are the dimensions of the extracted features for these modalities. E_{eye}, D_{eye} represent the encoder and decoder of eye movements while E_{eeg}, D_{eeg} represent the encoder and decoder of EEG signals, and u_{eye}, v_{eye}, u_{eeg}, v_{eeg} denote their respective parameters. The outputs through encoders can be represented as

$$O_{eye} = E_{eye}\left(X_{eye}; u_{eye}\right), O_{eeg} = E_{eeg}\left(X_{eeg}; u_{eeg}\right). \quad (1)$$

Correspondingly, the outputs of decoders are

$$\hat{X}_{eye} = D_{eye}\left(O_{eye}; v_{eye}\right), \hat{X}_{eeg} = D_{eeg}\left(O_{eeg}; v_{eeg}\right). \quad (2)$$

The loss of DAE can be represented as the reconstruction loss of eye movement and EEG features,

$$\mathcal{L}_{RC} = \mathcal{L}_{MSE}(X_{eye}, \hat{X}_{eye}) + \mathcal{L}_{MSE}(X_{eeg}, \hat{X}_{eeg}). \quad (3)$$

Then, we choose aggregation-based fusion as the multimodal fusion strategy, which concatenates O_{eye} and O_{eeg} directly. The classification result can be represented as

$$\hat{y} = CLS(\hat{X}_{eye}, \hat{X}_{eeg}), \quad (4)$$

where CLS denotes the multimodal classifier.

The high-level features of eye movements and EEG signals representing the level of decision confidence can be learned by minimizing the following loss:

$$\mathcal{L} = \lambda_{RC}\mathcal{L}_{RC} + \lambda_{CLS}\mathcal{L}_{CLS}, \quad (5)$$

where λ_{RC}, λ_{CLS} are the tradeoff parameters for each loss.

2) Training Stage II: EEG High-level Feature Generation: After training stage I, we obtain high-level features of eye movements O_{eye} and EEG O_{eeg}, respectively. Then, we take O_{eye} as the inputs of the GANs to guide the generator to produce the corresponding EEG features

$$\hat{O}_{eeg} = G(O_{eye}; \theta), \quad (6)$$

where θ denotes the parameters of the generator G.

The discriminator D is a binary classifier to distinguish the true modality pairs from the predicted modality pairs. We give the true multimodal data (O_{eye}, O_{eeg}) a label of 1 and the predicted multimodal data $(O_{eye}, \hat{O}_{eeg})$ a label of 0. We minimize the following cross-entropy loss to train the discriminator,

$$\begin{aligned}\mathcal{L}_D = &\mathcal{L}_{CE}(D((O_{eye}, O_{eeg}); \beta), 1) \\ &+ \mathcal{L}_{CE}(D((O_{eye}, \hat{O}_{eeg}); \beta), 0),\end{aligned} \quad (7)$$

where β denotes the parameters of discriminator D.

The generator is optimized to estimate the generated data in order to make it difficult for the discriminator to distinguish from the true data. Therefore, we train the generator with the following objective function $\mathcal{L}_G$ by fixing the parameter β in the discriminator,

$$\mathcal{L}_G = \mathcal{L}_{CE}(D((O_{eye}, \hat{O}_{eeg}); \beta), 1). \quad (8)$$

In addition, a content loss function is employed to encourage $\hat{O}_{eeg}$ to be close to O_{eeg}. This can be achieved by minimizing the Euclidean distance between them, resulting in a mean squared error (MSE) loss L_{MSE} defined as

$$\mathcal{L}_{MSE}(O_{eeg}, \hat{O}_{eeg}) = \|O_{eeg} - G(O_{eye}; \theta)\|_2^2, \quad (9)$$

where L_{MSE} encourages the learning of detailed information for completing the EEG modality. Therefore, the overall loss function of generator G can be described as

$$\mathcal{L} = \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_G\mathcal{L}_G. \quad (10)$$

The algorithm for optimizing the problem is given in Algorithm 1.

Algorithm 1: The cross-modal method based on generative adversarial learning

Data: Eye movement data X_{eye}, EEG data X_{eeg} and labels Y. Divide training and test sets according to cross-validation.

Result: Predicted labels on test data.

Training Stage I:
Initialize encoders E_{eye}, E_{eeg}, decoders D_{eye}, D_{eeg} and classifier C;
while *not converged* **do**
  Optimize E_{eye}, E_{eeg}, D_{eye}, D_{eeg} and C by minimizing Equation (5) ;
end
Training Stage II:
Initialize the generator G and discriminator D;
while *not converged* **do**
  Update the discriminator D by minimizing Equation (7);
  Update the discriminator G by minimizing Equation (10);
end
Test Stage:
Obtain eye movement features using trained feature extractor E_{eye};
Generator G generates EEG features from eye movement features;
Send concatenated data with eye movement features and EEG features to the classifier C;
Return predicted labels;

III. Experiment

A. Datasets

We conduct the experiments on the SEED-VPDC dataset [7]. The dataset is a multimodal dataset including EEG signals and eye movements for measuring five-level decision confidence. The experiment consists of 135 trials, where each trial contains one image, which corresponds to one decision. The stimuli materials contain three types of similar animails chosen from the Caltech 101 dataset [10]. Fourteen subjects participate in the experiments, and eye movements and EEG signals are recorded at the same time during the entire experiment. Since the eye movement data for one of the subjects is incomplete, we conclude with complete eye movements and EEG signals for thirteen subjects.

For eye movement signals, 22 features including pupil diameter, fixation duration, blink duration, and saccade duration are extracted by a Tobii Pro X3-120 screen-based eye tracker. The EEG signals are recorded by a 62-channel active AgCl electrode cap with an ESI NeuroScan System at a sampling rate of 1000 Hz according to the international 10-20 system. For data preprocessing, a bandpass filter between 0.3 and 50 Hz is applied to each channel to filter the noise and a linear dynamic system (LDS) method is adopted to smooth features. Different entropy (DE) features [4] are extracted within a nonoverlapping one-second time window from 5 frequency bands (namely δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, and γ: 31-50 Hz) of every sample, which has been proven to have the best performance for the classification of decision confidence [4].

TABLE I

The classification accuracy and F1-score (%) (Mean/Std) of different models on the five-category SEED-VP dataset

Method	Training data		Test data		Score	
	EYE	EEG	EYE	EEG	F1-Score	Accuracy
SVM [4]	✓		✓		34.49/6.14	40.76/7.61
	✓	✓	✓	✓	40.94/7.50	46.51/7.91
DNNS [4]	✓		✓		39.11/6.70	44.09/7.49
	✓	✓	✓	✓	46.62/6.98	50.15/8.14
DAE	✓	✓	✓	✓	48.79/6.85	52.23/7.83
DAL [9]	✓	✓	✓		39.38/8.32	42.42/9.47
Our Method	✓	✓	✓		44.54/7.35	48.22/8.64

B. Experimental settings

We adopt a five-fold cross-validation method and the subject-dependent classification setting which follows the work in [4]. We choose two classifiers, support vector machine (SVM) and deep neural network with shortcut connections (DNNS) as the baselines, which were employed to investigate the capability of EEG signals for measuring human decision confidence in [4]. we use the radial basis function kernel and search the parameter space from $2^{[-5:10]}$ for C in SVM. The DNNS method employs four hidden layers and one output layer, the size of the hidden layers is searched from 16 to 256, and the learning rate is set to 0.001. The two classifiers are tested and trained on eye movements and multimodal data, respectively. To further validate the performance gap between our cross-modal model and the multimodal model, we choose aggregation-based fusion as the multimodal fusion strategy, which concatenates the high-level eye movement and EEG features extracted from DAE. Finally, to verify the performance of our cross-modal method, we test the model based on deep adversarial learning (DAL) proposed in [9] on the SEED-VPDC dataset, which generates EEG information from primary eye movement features directly.

IV. Results

The experimental results including the accuracy and F1-score of different methods are listed in Table I. The mean accuracies and standard deviations of our model are compared with those of other methods.

A. EEG vs. Eye Movement

From Table 1, we find that the DNNS method significantly outperforms the SVM method on either modality, which demonstrates the superiority of neural networks. In addition, the ability of EEG signals to classify confidence decisions is stronger than that of eye movements, regardless of whether the SVM method or the DNNS method is used, which indicates that EEG signals are more reliable than eye movements in the task of confidence decision recognition.

From the confusion matrix shown in Fig. 2, we find that eye movements have a relatively high recognition rate for low decision confidence levels (1 and 2), while EEG signals have a stronger ability to discriminate the extreme confidence levels (1 and 5) [4]. This indicates that complementary representations exist between eye movement and EEG signals for measuring the decision confidence.

B. Cross-modal vs. Single-modal methods

Our proposed cross-modal method outperforms the DNNS method trained and tested only on eye movement signals with an improvement of approximately 5.43% in accuracy and 4.13% for the F1-score, see Table I. In addition, it is demonstrated from the confusion matrices that our proposed method has a relatively large improvement on the recognition rate on all levels, especially extreme confidence levels, by up to 11.55% on the first level and 8.82% on the fifth level. We believe that EEG knowledge which has stronger ability to identify extreme confidence levels has been learned through our method even with eye movements as the only inputs.

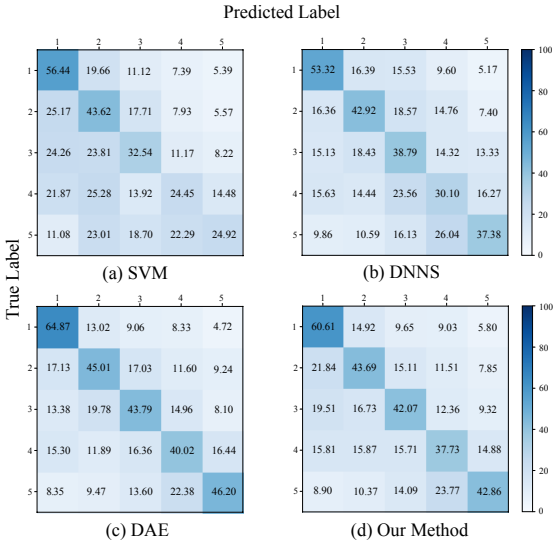

Fig. 2. The confusion matrices of SVM, DNNS trained and tested only on eye movement modality, DAE multimodal method and our proposed cross-modal method. The rows of the confusion matrices represent the target class and the columns represent the predicted class. The numbers from 1 to 5 represent weak to strong confidence in decision making.

C. Cross-modal vs. Multimodal methods

Our method is as competitive as the DNNS multimodal method, but there is still performance gap in comparison with the DAE multimodal approach. As seen from the confusion matrix, our approach mainly performs worse in the identification of extreme confidence levels. This can be explained by the fact that generated EEG information cannot completely replace the real EEG signals. However, the moderate decrease in accuracy compared to multimodal methods is considered acceptable because our method only tested on eye movement signals, thereby reducing the dependence on EEG signals which makes decision confidence measurement more applicable and practicable.

D. Comparison of cross-modal methods

Compared with the DAL method, which generates EEG information from primary eye movement signals without going through the high-level feature extraction process, our method obtains an improvement of approximately 5.80% in accuracy and 5.16% for the F1-score. The DAE method achieves better performance than multimodal DNNS method, which indicates that the high-level features extracted from the first training phase are more conducive to decision confidence classification. More importantly, it is difficult to generate primary high-dimensional EEG features from low dimensional eye movement features.

V. Conclusion

In this paper, we propose a cross-modal approach based on generative adversarial learning for the task of decision confidence measurement. In our method, the intrinsic relationship between eye movement and EEG features in a high-level feature space can be learned in the training stage. Experimental results on the SEED-VPDC dataset demonstrate that our proposed method outperforms the single-modal methods trained and tested only on eye movement signals. This indicates that EEG features can be generated from the eye movement features without EEG, which to some extent complements the information of the EEG modality.

References

[1] A. Pouget, J. Drugowitsch, and A. Kepecs, "Confidence and certainty: distinct probabilistic quantities for different goals," *Nature Neuroscience*, vol. 19, no. 3, pp. 366–374, 2016.

[2] K. M. Lempert, Y. L. Chen, and S. M. Fleming, "Relating pupil dilation and metacognitive confidence during auditory decision-making," *PLoS One*, vol. 10, no. 5, p. e0126588, 2015.

[3] S. V. Shooshtari, J. E. Sadrabadi, Z. Azizi, and R. Ebrahimpour, "Confidence representation of perceptual decision by EEG and eye data in a random dot motion task," *Neuroscience*, vol. 406, pp. 510–527, 2019.

[4] R. Li, L.-D. Liu, and B.-L. Lu, "Discrimination of decision confidence levels from EEG signals," in *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2021, pp. 946–949.

[5] N. Sadras, O. G. Sani, P. Ahmadipour, and M. M. Shanechi, "Post-stimulus encoding of decision confidence in eeg: Toward a brain-computer interface for decision making," *bioRxiv*, 2022.

[6] L.-M. Zhao, R. Li, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from EEG and eye movement signals: complementary representation properties," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 611–614.

[7] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 715–729, 2021.

[8] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *Journal of Neural Engineering*, vol. 14, no. 2, p. 026017, 2017.

[9] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1158–1166.

[10] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, 2004, pp. 178–178.