2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871605

A Cross-modality Deep Learning Method for Measuring Decision Confidence from Eye Movement Signals

Cheng Fei, Rui Li, Li-Ming Zhao, Ziyi Li and Bao-Liang Lu* *Fellow, IEEE*

Abstract— **Electroencephalography (EEG) signals can effectively measure the level of human decision confidence. However, it is difficult to acquire EEG signals in practice due to the expensive cost and complex operation, while eye movement signals are much easier to acquire and process. To tackle this problem, we propose a cross-modality deep learning method based on deep canoncial correlation analysis (CDCCA) to transform each modality separately and coordinate different modalities into a hyperspace by using specific canonical correlation analysis constraints. In our proposed method, only eye movement signals are used as inputs in the test phase and the knowledge from EEG signals is learned in the training stage. Experimental results on two human decision confidence datasets demonstrate that our proposed method achieves advanced performance compared with the existing single-modal approaches trained and tested on eye movement signals and maintains a competitive accuracy in comparison with multimodal models.**

I. INTRODUCTION

Currently, computer-aided decision-making has been widely applied in various fields with the rapid development of deep learning. However, decision making for tasks with high risk, such as commercial decision making and military remote sensing image interpretation, cannot exclusively rely on computers and professionals are still indispensable. Hence, it is necessary to find an objective way to measure the reliability of decisions to assist decision-makers in making sound judgments.

Decision confidence is the feeling of correctness or optimization of an individual when making a decision and can reflect the probability of being correct [1]. Various studies have demonstrated that it is feasible to use EEG signals to infer human decision confidence, and event-related potential (ERP) is used to investigate the neural mechanisms of human decision confidence [2] [3]. Very recently, Li *et al.* designed a visual perception task [4] and an object detection task [5] for measuring decision confidence, and their experimental results indicate that EEG signals recorded during the experiments are able to distinguish different levels of decision confidence and that neural patterns of EEG signals for decision confidence in the visual perception task do exist.

In the field of affective computing, many studies have indicated that the performance of single-modal models turns out to be at an inadequate low level and multimodal models can improve recognition accuracies. For example, it has been indicated that complementary representation properties exist between different modalities for emotion recognition [6], and deep multimodal fusion using data from different modalities has exhibited a clear advantage over its single-modal counterpart in emotion recognition [7] [8].

Although multimodal fusion generally leads to better results, the fact that more modalities are involved also means that the data from different modalities need to be acquired at a greater cost. In fact, the process of collecting EEG signals is very complicated in practice. In addition, for several inevitable preparations, such as wearing electrode caps and injecting conductive gel, we have to guarantee that the acquisition environment is quiet without disturbance since the signals are very subtle and sensitive to interference, impeding their use in practical scenarios. Comparatively, other physiological signals, such as eye movement signals are much easier to collect.

Based on the above discussion, our goal is to use the information of multimodalities effectively with high operational feasibility. Therefore, we propose a cross-modality method based on deep canonical correlation analysis (CDCCA) for measuring human decision confidence. The basic idea behind deep canonical correlation analysis is to transform each modality separately and coordinate different modalities into a hyperspace by using specified canonical correlation analysis constraints. In the training phase, the representations for EEG and eye movement modality under canonical correlation analysis are learned, and then a parameter sharing layer is used to learn more knowledge from both modalities. In the testing phase, only eye movement signals are used as inputs. We evaluate our proposed method on the datasets proposed in [4] [5] and find that it maintains superior performance than other single-modal models tested and trained only on eye movements and achieves a competitive accuracy in comparison with multimodal models.

II. METHOD

A. Deep Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a standard statistical technique for finding linear projections of two given vec-

This work was supported in part by grants from the National Natural Science Foundation of China (Grant No. 61976135), Shanghai Municipal Science and Technology Major Project, 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, Z. Y. Li 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)

tors to maximally correlate them. Andrew and colleagues [9] proposed deep canonical correlation analysis (DCCA) based on CCA to learn complex nonlinear transformations of two views of data such that the resulting representations are highly linearly correlated. Liu *et al.* [10] found that the features transformed by DCCA from different modalities are more homogeneous and discriminative across emotions and that emotion recognition is enhanced by multimodality fusion using DCCA tested on the SEED dataset. Inspired by these ideas, we hope to extract highly linearly correlated features representing decision confidence common between eye movement and EEG signals by DCCA.

Assume that $X_1 \in R^{N\times d_1}$ represents the data of eye movement signals, $X_2 \in R^{N\times d_2}$ the data of EEG signals, and N the batch-size, d_1 and d_2 the dimensions of the extracted features for these modalities. The outputs through deep neural networks can be represented as, $O_1 = f_1(X_1; W_1), O_2 = f_2(X_2; W_2)$, where W_1 and W_2 denote all parameters for the nonlinear transformations. $O_1 \in R^{N\times d}$ and $O_2 \in R^{N\times d}$ are the outputs of the neural networks, and d denotes the output dimension of DCCA. The goal of DCCA is to jointly learn the parameters W_1 and W_2 for both neural networks such that the correlation of O_1 and O_2 is as high as possible:

$$(W_1^*, W_2^*) = \arg\max_{W_1, W_2} \operatorname{corr}\left(f_1(X_1; W_1), f_2(X_2; W_2)\right). \tag{1}$$

We use the backpropagation algorithm to update W_1 and W_2. The solution to calculating the gradients of the objective function was developed by Andrew [9].

Let $\bar{O}_1 = O_1' - \frac{1}{N}O_1'\,\mathbf{1}$ be the centred output matrix (similar to $\bar{O}_2$). We define $\hat{\Sigma}_{12} = \frac{1}{N-1}\bar{O}_1\bar{O}_2'$, $\hat{\Sigma}_{11} = \frac{1}{N-1}\bar{O}_1\bar{O}_1' + r_1 I$, where r_1 is a regularization constant (similar to $\hat{\Sigma}_{22}$). The total correlation of the top k components of O_1 and O_2 is the sum of the top k singular values of matrix $T = \hat{\Sigma}_{11}^{-1/2}\hat{\Sigma}_{11}\hat{\Sigma}_{22}^{-1/2}$, and the total correlation is the trace of T:

$$\operatorname{corr}(O_1, O_2) = \left(\operatorname{tr}(T'T)\right)^{1/2}. \tag{2}$$

Then we calculate the gradients with the singular decomposition of $T = UDV'$,

$$\frac{\partial \operatorname{corr}(O_1, O_2)}{\partial O_1} = \frac{1}{N-1}\left(2\nabla_{11}\bar{O}_1 + \nabla_{12}\bar{O}_2\right), \tag{3}$$

where

$$\nabla_{11} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}UDU'\hat{\Sigma}_{11}^{-1/2},$$
$$\nabla_{12} = \hat{\Sigma}_{11}^{-1/2}UV'\hat{\Sigma}_{22}^{-1/2},$$

and $\partial \operatorname{corr}(O_1, O_2)/\partial O_2$ has a symmetric expression. We jointly learn the parameters W_1 and W_2 for both neural networks to make the correlation of O_1 and O_2 as high as possible by the CCA constraint. The $\mathcal{L}_{CCA}$ between O_1 and O_2 can be represented as the negative correlation between O_1 and O_2:

$$\mathcal{L}_{CCA} = -\operatorname{corr}(O_1, O_2) = -\left(\operatorname{tr}(T'T)\right)^{1/2}. \tag{4}$$

B. Sharing Parameter Layer

Tang *et al.* [11] proposed a parameter sharing strategy to enhance information sharing between the speech translation task and text translation task. Inspired by this idea, a parameter sharing layer is added to our method to encourage knowledge sharing between the two modalities. The results through the parameter sharing layer can be represented as

$$R_1 = f_s(O_1; W_s), R_2 = f_s(O_2; W_s), \tag{5}$$

where f_s represents the sharing parameter layer and W_s is the parameters shared between two modalities.

C. Classification Loss

For the classifier, we apply a Multilayer Perceptron (MLP) as the classifier. The cross-entropy loss of the classifier is as follows:

$$\mathcal{L}_{CLS}(y^i, \hat{y}^i) = -\sum_i y^i \log \hat{y}^i, \tag{6}$$

where $\hat{y}^i$ represents the output of the classifier, and y^i represents the label of the data.

D. Cross-modality deep learning method

The structure of our proposed method is shown in Fig. 1. In the training stage, separating but coordinating representations for each modality are learned by canonical correlation analysis constraints, and then a parameter sharing layer is used to learn more information from EEG signals. The test stage follows the dashed box and only eye movement signals are needed.

We minimize $\mathcal{L}_{CCA}$ to make the correlation of O_1 and O_2 as high as possible. The $\mathcal{L}_{CCA}$ between O_1 is referenced in Eq. (4). Then the extracted representations of eye movements and EEG signals O_1 and O_2 are sent to the parameter sharing layer as Eq. (5). The losses of the two classifiers $\mathcal{L}_{CLS\_eye}$ $\mathcal{L}_{CLS\_eeg}$ can be rewritten as shown in Eq. (6).

The whole training process can be expressed as minimizing

$$\mathcal{L} = \lambda_{CCA}\mathcal{L}_{CCA} + \lambda_{CLS}(\mathcal{L}_{CLS\_eye} + \mathcal{L}_{CLS\_eeg}), \tag{7}$$

where λ_{CCA} and λ_{CLS} are tradeoff parameters for each loss.

III. EXPERIMENTS

A. Datasets

We conduct experiments on two datasets developed in [4] [5] and call them SEED-VP and SEED-OD in the following sections. The two datasets are all multimodal datasets including EEG signals and eye movement signals for measuring five-level decision confidence. Both of the experiments have been approved by the Scientific & Technical Ethics Committee at Shanghai Jiao Tong University.

The SEED-VP dataset comes from a confidence experimental paradigm in which 14 subjects (7 males and 7 females, aged from 18 to 24) perform a visual perception task. The experiment consists of 135 trials, and each trial contains one image selected from the Caltech 101 dataset [12]. The subjects were required to identify the animal in the image

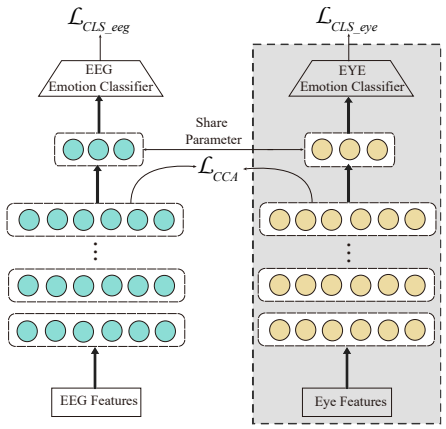

Fig. 1. The framework of our proposed CDCCA model. In the training stage, separating but coordinating representations for each modality are learned, and then a parameter sharing layer is used to learn more information from EEG signals. The test stage follows the grey area and only eye movement signals are needed.

and make decisions from three different options, including the correct answers for the other two in the same group. Eye movements and EEG signals were recorded at the same time during the entire experiment. Similarly, the SEED-OD dataset is an object detection task where subjects must find specified target objects in optical remote sensing images. The data are recorded in the same way as the SEED-VP dataset. Since the eye movement data recording for one of the subjects was incomplete, we ended up with complete eye movements and EEG signals for 13 subjects for both datasets.

The EEG signals were recorded using 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 the linear dynamic system (LDS) method is adopted to smooth features. We use the differential entropy (DE) features on all five bands, since they achieve the best performance in [4] [5]. Therefore, the dimension of EEG features is 310 per sample, calculated by 62 channels multiplied by 5 bands. The eye movement signals were recorded at a sampling rate of 120 Hz using a Tobii Pro X3-120 screen-based eye tracker. Twenty-two eye movements features, including pupil diameter, saccade, blink and fixation were extracted, which follows the work in [6].

B. Experimental settings

We follow the subject-dependent classification setting in [4] [5] and train a model for each subject. We choose two traditional classifiers, the SVM and Multilayer Perceptron (MLP), as single-modal methods trained and tested only on eye movement modality. To verify the effectiveness of our approach for each module, we remove the parameter sharing layer as the CDCCA-S method and keep the number of nodes and layers of the network constant during this process. The CDCCA-S method and our CDCCA method are all cross-modal methods trained on both modalities and tested only on eye movement modality. Further, to compare our method with the multimodal approach, we employ the DCCA method proposed in [10].

TABLE I

THE CLASSIFICATION ACCURACY AND F1-SCORE (%) (MEAN/STD) OF DIFFERENT MODELS ON TWO DATASETS.

Model	SEED-VP		SEED-OD	
	ACC	F1	ACC	F1
SVM[1]	40.76/7.61	34.49/6.14	38.07/7.09	32.93/7.52
MLP[1]	42.19/8.29	38.66/6.81	39.08/6.44	36.21/7.73
CDCCA-S[2]	47.48/9.07	43.30/7.80	43.05/6.94	39.62/6.79
CDCCA[2]	48.47/7.30	44.05/6.94	43.71/7.87	40.06/7.43
DCCA[3]	51.23/8.11	47.76/7.43	48.69/8.42	44.53/8.02

[1] Single-modal method trained and tested on eye movements.
[2] Cross-modal method trained on EEG and eye movements, but tested only on eye movements.
[3] Multimodal method trained and tested on EEG and eye movements.

IV. RESULTS

A. Comparison of single-modal methods between the two modalities

The first two rows of Table I are the results of single-modal methods trained and tested only on the eye movement modality, while EEG signals obtain an accuracy of 49.14% and F1-score of 45.07% [4] on SEED-VP and an accuracy of 47.36% and F1-score of 43.50% [5] on SEED-OD. It is apparent from the results that the EEG signals are more reliable than eye movement signals for measuring decision confidence. To further compare the classification ability of the two modalities for decision confidence discrimination, we compare the classification results of eye movements (Fig. 2 (a)) and EEG [4] using the SVM method, which is shown in Fig. 3. We observe that eye movements are superior to EEG signals in classifying low levels of confidence (1 and 2), with a mean accuracy of 56% and 42%, respectively, whereas EEG signals have more discriminative power than eye movements in recognizing the highest confidence level (67% versus 25%). Intermediate confidence levels 3 and 4 are difficult to distinguish in both modalities.

B. Comparison between single-modal and multimodal methods

In the last row in Table I, the multimodal method DCCA achieves an accuracy of 51.23% and an F1-score of 47.76% on the SEED-VP dataset and an accuracy of 48.69% and an F1-score of 44.53% on the SEED-OD dataset which performs better than any single-modal method. These results indicate that deep multimodal fusion with data from different modalities for classification is still valid on this task. Combining the findings in IV-A, it can be demonstrated that different modalities can provide complementary information for measuring decision confidence.

C. Comparison with cross-modal methods

From Table I, we can find that our method outperforms the single-modal methods with an advantage of approximately

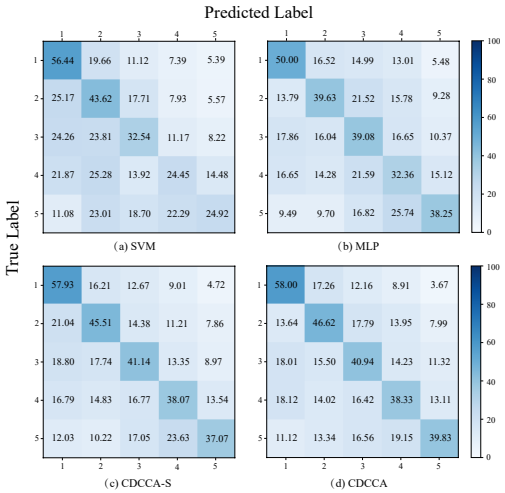

Fig. 2. The confusion matrices of SVM, MLP, CDCCA-S and our proposed method based on the SEED-VP dataset. The rows of the confusion matrices represent the target class and the columns represent the predicted class. The deeper the color is, the higher the recognition rate between different levels of decision confidence.

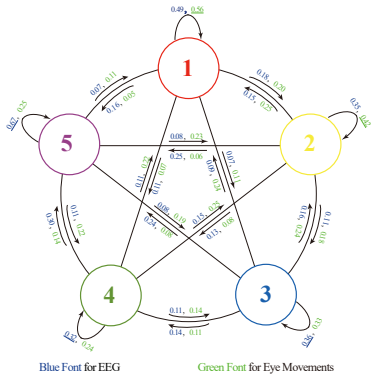

Fig. 3. Confusion graph of EEG signals and eye movements on the SEED-VP dataset. The arrow indicates the direction of state transition. The numbers denote the percentage corresponding to the confusion matrices of SVM from the two modalities. The underlined digits indicate higher values.

6.28% for the accuracy and 5.39% for the F1-score on the SEED-VP dataset and approximately 4.63% for the accuracy and 3.85% for the F1-score on the SEED-OD dataset. This is also reflected in Fig. 2, which shows that our model outperforms on all levels compared with single-modal methods. This indicates that the effective features of EEG signals that can be used to discriminate decision confidence are learned through our method. Although our method does not work as well as the multimodal method, it is worthwhile to reduce the reliance on the use of EEG signals in the test phase.

D. Ablation study

The third row of Table I represents the results of our method after removing the weight sharing layer (CDCCA-S), which are significantly superior to the results of MLP, and we also find that it achieves better accuracy and F1-score than MLP for almost all subjects. It is demonstrated that the effective features representing decision confidence common between eye movements and EEG signals are extracted by $\mathcal{L}_{CCA}$. Moreover, our method achieves relatively better performance after adding a weight sharing layer. As shown in Fig. 2, our method performs better especially on the highest level compared with CDCCA-S, which may indicate that more knowledge from EEG signals is learned after the sharing parameter layer.

V. Conclusion

In this paper, we have proposed a new cross-modality deep learning method based on DCCA for measuring human decision confidence from eye movement signals. In our method, we obtain features of each modality by transforming separately and coordinating the data of different modalities into a hyperspace using specified canonical correlation analysis constraints, and then use a parameter sharing layer to learn more information from EEG signals. In this way, knowledge from EEG signals is learned in the training stage and only eye movement signals are needed in the test stage. The experimental results on two datasets demonstrate that our proposed method has two promising characteristics: (a) the knowledge of measuring the decision confidence level from EEG signals can be learned and (b) the dependence on EEG signals can be reduced.

References

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

[2] S. Gherman and M. G. Philiastides, "Neural representations of confidence emerge from the process of decision formation during perceptual choices," *Neuroimage*, vol. 106, pp. 134–143, 2015.

[3] A. Boldt, A.-M. Schiffer, F. Waszak, and N. Yeung, "Confidence predictions affect performance confidence and neural preparation in perceptual decision making," *Scientific Reports*, vol. 9, no. 1, pp. 1–17, 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*. IEEE, 2021, pp. 946–949.

[5] ——, "Measuring human decision confidence from EEG signals in an object detection task," in *2021 10th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2021, pp. 942–945.

[6] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," *IJCAI*, vol. 15, p. 1170–1176, 2015.

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

[8] 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.

[9] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1247–1255.

[10] 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*, 2021.

[11] Y. Tang, J. Pino, X. Li, C. Wang, and D. Genzel, "Improving speech translation by understanding and learning from the auxiliary text translation task," *arXiv preprint arXiv:2107.05782*, 2021.

[12] 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.