A Cross-subject and Cross-modal Model for Multimodal Emotion Recognition

Jian-Ming Zhang[1], Xu Yan[6], Zi-Yi Li[1], Li-Ming Zhao[1], Yu-Zhong Liu[7], Hua-Liang Li[7], and Bao-Liang Lu[1,2,3,4,5(✉)]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
{jmzhang98,liziyi,lm_zhao,bllu}@sjtu.edu.cn
[2] Center for Brain-Machine Interface and Neuromodulation, RuiJin Hospital Shanghai Jiao Tong University School of Medicine, Shanghai 200020, China
[3] RuiJin-Mihoyo Laboratory, RuiJin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200020, China
[4] Key Laboratory of Shanghai Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[5] Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China
[6] Department of Linguistics, University of Washington, Seattle, WA 98195, USA
xyan3@uw.edu
[7] Key Laboratory of Occupational Health and Safety of Guangdong, Power Grid Co., Ltd, Electric Power Research Institute of Guangdong Power Grid Co., Ltd., Guangdong, China

Abstract. The combination of eye movements and electroencephalography (EEG) signals, representing the external subconscious behaviors and internal physiological responses, respectively, has been proved to be a dependable approach with high interpretability. However, EEG is unfeasible to be put into practical applications due to the inconvenience of data acquisition and inter-subject variability. To take advantage of EEG without being restricted by its limitations, we propose a cross-subject and cross-modal (CSCM) model with a specially designed structure called gradient reversal layer to bridge the modality differences and eliminate the subject variation, so that the CSCM model only requires eye movements and avoids using EEG in real applications. We verify our proposed model on two classic public emotion recognition datasets, SEED and SEED-IV. The competitive performance not only illustrates the efficacy of CSCM model but also sheds light on possible solutions to dealing with cross-subject variations and cross-modal differences simultaneously which help make effective emotion recognition practicable.

Keywords: Cross subject · Cross modality · EEG · Eye movements · Multimodal emotion recognition · Transfer learning

T. Mantoro et al. (Eds.): ICONIP 2021, CCIS 1517, pp. 203–211, 2021.
https://doi.org/10.1007/978-3-030-92310-5_24

1 Introduction

Emotional intelligence (EI) has become the spotlight in artificial intelligence since it is a promising way to perfect user experience in human-computer interfaces. EI contains three phases, namely emotion recognition, emotion understanding, and emotion regulation, among which the first step is the most critical [1] for its huge potential to be applied in broad scenarios such as entertainment, smart gadgets, education, and even medical treatment.

Researchers have dived into various modalities to seek an effective way to measure emotions. It has been proved that the combination of eye movements and EEG signals, representing the external subconscious behaviors and internal physiological responses, respectively, is a more dependable approach with high interpretability [12]. However, although this complementary collocation delivers decent performance, it is unfeasible to put it into real-life practice due to the restrictions of EEG in both extrinsic and intrinsic sides. The extrinsic obstacles are unavoidably caused by the equipment like injecting conductive gel, which leads to high cost and operational difficulty when using in daily life. Comparatively, the intrinsic limitation is related to the property of physiological signals. EEG data is highly subject-dependent and susceptible to the structural and functional differences between subjects [7], which brings great challenges to the construction of practical EEG-involved affective models.

To overcome the impediments raised by EEG, scholars have attempted to find solutions from diverse aspects, such as cross-modal transfer. Palazzo *et al.* [6] combined GAN with RNN to generate corresponding images from EEG signals recorded when subjects were watching images. Our previous work has verified the complementary characteristics of EEG and eye movements [5], supporting that eye movement analysis can be an accessible, simple, and effective method to study the brain mechanism behind cognition. As for subject dependency, transfer learning provides a practical solution to diminish the variability of data distribution between individuals. Zheng *et al.* [11] first applied several basic domain adaptation (DA) methods to EEG-based emotion recognition task, including transfer component analysis (TCA) [8], etc. Furthermore, combining DA with deep networks is an alternative way. Ganin *et al.* [3] proposed the domain-adversarial neural network (DANN) to extract the shared representations between the source domain and the target domain. Li *et al.* [4] first applied deep adaptation network (DAN) to EEG-based cross-subject emotion recognition. To reduce the huge demand of the target domain data in the test stage, Zhao *et al.* [9] proposed a plug-and-play domain adaptation (PPDA) method and achieved a trade-off between performances and user experience by using only a few target data to calibrate the model.

However, all these models are either cross-subject or cross-modal. The CSCM model that we propose eliminates the structural variability between individuals while learning the shared features of EEG and eye movements so that the latter can be an alternative modality to their combination. As a consequence, the test phase only requires data from the single eye movement modality with no need for

data from the new subject in advance. In this way, we guarantee the model with both maximum generalization ability and good feasibility in real applications.

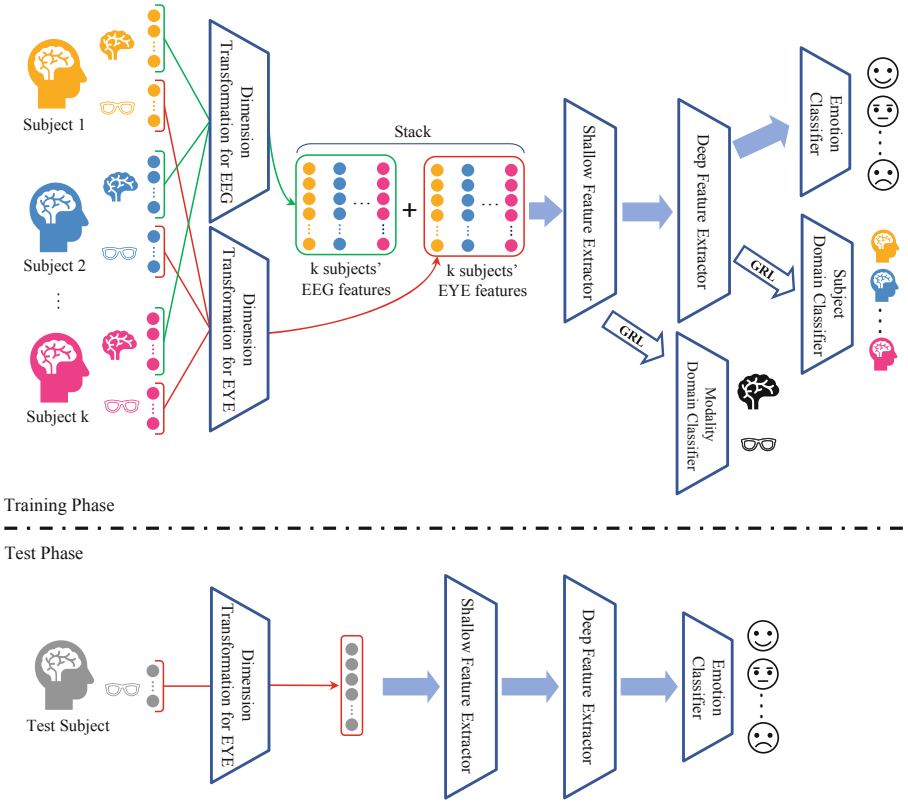

Fig. 1. The framework of our proposed CSCM model. The whole structure contains the training phase and the test phase. The training phase has a main chain in blue and two subchains to bridge the modality differences and eliminate the subject variation, respectively. The test phase only requires eye movement signals as input to predict emotions.

2 Methods

2.1 Overview

To make emotion recognition more generalizable and practicable, we propose a cross-subject and cross-modal model called CSCM model that gets over the inter-subject variability and modal restrictions caused by EEG signals. The framework of CSCM model is depicted in Fig. 1. The whole process can be divided into the

training phase and the test phase. In the training phase, both eye movements and EEG signals from source subjects are required as input. There is a main chain together with two subchains aiming to bridge modality differences and eliminate subject variation, respectively. In the main chain, dimension transformation layers are applied first to both types of signals separately to unify the dimensions. Then they are fed to a shallow feature extractor where the first subchain stretches out. A gradient reversal layer [3] connects the modality domain classifier in the subchain to the shallow feature extractor in order to generate domain-invariant features. The deep feature extractor follows the shallow one with the second subchain to make subject domains indistinguishable. An emotion classifier ends the training phase at last. For the test phase, our model only demands eye movement signals from the target subject.

2.2 Training Phase

Dimension Transformation. We use $\boldsymbol{X}'_{EEG} \in \mathbb{R}^p$ to annotate EEG feature vectors, where p is the feature dimension. Each dimension represents information from a specific channel of a frequency band. Similarly, $\boldsymbol{X}'_{EYE} \in \mathbb{R}^q$ stands for eye movement feature vectors with dimension q. Since p is much larger than q due to the information sufficiency of EEG, we conduct dimension transformation via a specific layer separately first to unify the feature dimensions so that $\boldsymbol{X}_{EEG}, \boldsymbol{X}_{EYE} \in \mathbb{R}^r$ where r is the dimension of mapped features.

Modality Reduction. After dimension transformation, we feed the mapped features to the shallow feature extractor $\boldsymbol{E}_s$ with parameters θ_s where s represents shallow. It is connected to the modality domain classifier $\boldsymbol{C}_{md}$ via a specially designed layer called Gradient Reversal Layer (GRL) $\boldsymbol{L}_{md}$, where md represents a modality domain. In the forward propagation, $\boldsymbol{L}_{md}$ functions in the typical way but in the backpropagation period it takes the gradient from $\boldsymbol{C}_{md}$ and multiplies by a certain negative number before passing back to $\boldsymbol{E}_s$, i.e. reversing the sign of the gradient. Optimization process is integrated as follows:

$$\begin{aligned}(\hat{\theta}_s, \hat{\theta}_y) &= arg\min_{\theta_s, \theta_y} E(\theta_s, \theta_y, \hat{\theta}_{md}) \\ (\hat{\theta}_{md}) &= arg\max_{\theta_{md}} E(\hat{\theta}_s, \hat{\theta}_y, \theta_{md}),\end{aligned} \tag{1}$$

where y stands for a label, i.e. a modality of EEG or eye movements. By this ingenious gradient reversal mechanism, features from EEG and eye movement modalities are gradually indistinguishable until $\boldsymbol{C}_{md}$ cannot tell them apart. Thus, modality-invariant features have been extracted out.

Inter-subject Variability Elimination. Since subject differences are determined by more factors and vary a lot, it is harder to minimize this variability than those between modalities. Hence, deep features are generated by $\boldsymbol{E}_d$, where d means deep, from the output of $\boldsymbol{E}_s$. We adopt the same structure with $\boldsymbol{L}_{md}$ to

build $\boldsymbol{L}_{sd}$ where *sd* is a subject domain linking $\boldsymbol{E}_d$ to the subject domain classifier $\boldsymbol{C}_{sd}$. Different from $\boldsymbol{L}_{md}$, $\boldsymbol{L}_{sd}$ tries to obtain subject-invariant features guided by the reversed gradient. Features after this step are expected to no longer contain modal or subject information. The prediction given by the emotion classifier $\boldsymbol{C}_e$ is based on both modality-invariant and subject-invariant features.

Learning Loss. In the training phase, EEG data and eye movement data from source subjects are utilized to train the model to minimize the following loss:

$$\mathcal{L} = \alpha \mathcal{L}_{\boldsymbol{C}_e} + \beta \mathcal{L}_{\boldsymbol{L}_{md}} + \gamma \mathcal{L}_{\boldsymbol{L}_{sd}}, \tag{2}$$

where α, β and γ are trade-offs that control the synergy of the three loss terms. We minimize the cross-entropy loss of the emotion classifier as:

$$\mathcal{L}_{\boldsymbol{C}_e} = -\sum_i y_i \log \hat{y}_i, \tag{3}$$

where $\hat{y}_i$ is the prediction of $\boldsymbol{C}_e$, and y_i is the ground truth label for the input x_i. $\boldsymbol{C}_{md}$ and $\boldsymbol{C}_{sd}$ also use cross-entropy as the loss.

2.3 Test Phase

Since two pre-trained feature extractors $\boldsymbol{E}_s$ and $\boldsymbol{E}_d$ have learned the knowledge from EEG signals and subject-invariant components, the whole model does not need any calibration by the new data and only needs the eye movement signals of the new subject as input in this phase.

3 Experiments

3.1 Datasets and Protocols

We verify the performance of our model on two public datasets, SEED [13] and SEED-IV[1] [10]. For SEED, we take data from 9 subjects who have multimodal signals while all 15 subjects in SEED-IV are qualified to be tested. Each subject participated in the experiments three times to watch videos that evenly cover every emotion in each experiment on different days. The EEG signals and the eye movement signals were recorded with a 62-channel electrode cap together in ESI Neuroscan system and SMI wearable eye-tracking glasses, respectively.

The recorded EEG signals are downsampled 200 Hz and then processed with a bandpass filter of 0–75 Hz and baseline correction as well. After the preprocessing, different entropy (DE) [2] features are extracted with non-overlapping 1-second and 4-second time windows for SEED and SEED-IV, respectively, in the five frequency bands (δ: 1–3 Hz, θ: 4–7 Hz, α: 8–13 Hz, β: 14–30 Hz, and γ: 31–50 Hz) from every sample. The eye movement signals are extracted by SMI BeGaze software, including pupil diameter, dispersion, etc. [5]

[1] The SEED and SEED-IV are available at https://bcmi.sjtu.edu.cn/home/seed/index.html.

Table 1. Results of different methods running on SEED and SEED-IV.

Methods	Training modalities	Test modality	SEED		SEED-IV	
			Avg.	Std.	Avg.	Std.
SVM	EYE	EYE	0.5223	0.0724	0.5231	0.1088
SVM	EEG	EEG	0.5690	**0.0594**	0.5567	0.0899
MLP	EYE	EYE	0.6646	0.0762	0.5508	0.0951
MLP	EEG & EYE	EYE	0.6837	0.0975	0.6110	0.1243
CSCM-SM	EYE	EYE	0.7030	0.1316	0.6836	**0.0590**
CSCM	EEG & EYE	EYE	**0.7618**	0.0761	**0.7222**	0.1123

3.2 Experimental Results

We select two popular methods, namely support vector machine (SVM) and multilayer perceptron (MLP), to make comparisons. The average accuracies (avg.) and standard deviations (std.) are reported in Table 1. Results for each subject of each method on SEED and SEED-IV are depicted in Fig. 2.

Comparison Under Single Modality. To examine the performance of CSCM model, we first compare models all trained and tested with a single modality. Since our motivation is to avoid using EEG in the test stage, here we mainly focus on the eye movement modality. It is necessary to mention that CSCM-SM model on line 5 refers to a transformed model of CSCM model that does not have the modality domain classifier, allowing it to be trained with a single modality (SM) in the cross-subject emotion recognition task.

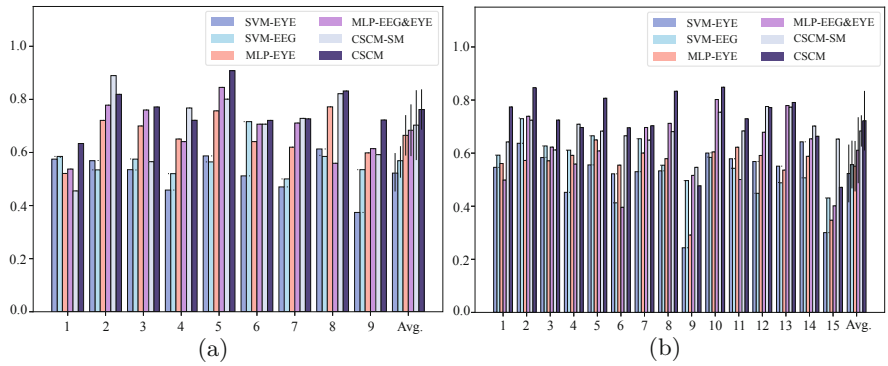

Fig. 2. The accuracies for each subject of each method and the averages on SEED (a) and SEED-IV (b).

Line 1, 3, and 5 display the results of training and testing with only the eye movement modality. We tested on data from 1 subject and used all remained data as training sets for each dataset. SVM is taken as the baseline. For three-class emotion recognition on SEED, CSCM model slightly outperforms MLP

by 4%, and these performances are drastically higher than the baseline ranging from 14% to 18% as shown above. For SEED-IV when the number of emotion categories increases, CSCM model surpasses both SVM and MLP for at least 13% with a quite decent standard deviation around 0.06, which demonstrates that our proposed method is solid and has great potential to perform better when complicated emotional states are involved.

Previous studies have confirmed that EEG signals contain much more useful information than eye movement signals [5,10]. In other words, for the same method to be working on a single modality, the EEG modality is supposed to have better performance than the eye movement modality which has been proved by the comparison between the first two lines in Table 1. However, our model with eye movement modality acts much better than SVM with EEG modality as presented in line 2. This surprising fact convinces us that the utilization of information in CSCM model is of high efficiency.

Verification of Two GRL Layers. We corroborate the effect of the two specially designed GRL separately as follows. The first GRL $\boldsymbol{L}_{md}$ is set to minimize the modality differences. However, whether it really employs the knowledge from EEG remains a question. Therefore, we compare the standard CSCM model with CSCM-SM model. Specifically, these two differ from modalities used in the training phase that CSCM model takes both while the other only uses eye movement modality. The comparable results are listed on the last two lines in Table 1. The CSCM model shows great superiority over CSCM-SM model, especially in three-label classification about 6% with outstanding stability, proving that $\boldsymbol{L}_{md}$ works well in helping capture useful information from EEG and further provides a solid foundation for disposing of EEG in real applications.

The second GRL $\boldsymbol{L}_{sd}$ aims to reduce the subject discrepancies. For line 1, 3, and 5 in Table 1, all can be regarded as cross-subject methods with different implementation ways as mentioned in the previous subsection above. The better results on both datasets indicate that using $\boldsymbol{L}_{sd}$ to diminish subject variation is more reliable and effective than the traditional ways. This evidence supports the idea that CSCM model provides a new solution to tackle subject variability.

Comparison with Multimodal Models. Few existing models are cross-modal and cross-subject at the same time in the emotion recognition task. In order to examine our model, we adapt a classic model MLP to the same task as our baseline, i.e. training with multimodalities and testing with only eye movement signals. Particularly, in order to be consistent with CSCM model, we also transform the dimensions of EEG and eye movement features at first. Compared with MLP, it is evident that our model fits the properties of data well from the significantly higher accuracies and less standard deviations, especially on SEED-IV. Our proposed CSCM model affords researchers a novel idea to realize cross-subject and cross-modal simultaneously with respectable performances and more state-of-the-art methods need to be further explored in the future.

4 Conclusions

In this paper, we have devised a novel multimodal learning model, CSCM model, to make emotion recognition as practicable as possible beyond lab environment by applying cross-subject and cross-modal techniques simultaneously. It successfully outperforms baselines and is substantiated from both sides individually according to the comprehensive experiment results. Besides the effective model itself, this pioneering thought sheds light on further attempts at making affective computing systems more practicable and bringing tangible benefits to daily life.

Acknowledgments. This work was supported in part by grants from the National Natural Science Foundation of China (Grant No. 61976135), SJTU Trans-Med Awards Research (WF540162605), the Fundamental Research Funds for the Central Universities, the 111 Project, and the China Southern Power Grid (Grant No. GDKJXM20185761).

References

1. Brunner, C., et al.: Bnci horizon 2020: Towards a roadmap for the BCI community. Brain-Comput. Interf. **1**, 1–10 (2015)
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. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2030 (2017)
4. Li, H., Jin, Y.M., Zheng, W.-L., Lu, B.-L.: Cross-subject emotion recognition using deep adaptation networks. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11305, pp. 403–413. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04221-9_36
5. Lu, Y.F., Zheng, W.L., Li, B., Lu, B.L.: Combining eye movements and EEG to enhance emotion recognition. In: Yang, Q., Wooldridge, M.J. (eds.) Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, pp. 1170–1176. AAAI Press (2015)
6. Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Shah, M.: Generative adversarial networks conditioned by brain signals. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3410–3418 (2017)
7. Samek, W., Meinecke, F.C., Müller, K.R.: Transferring subspaces between subjects in brain-computer interfacing. IEEE Trans. Biomed. Eng. **60**(8), 2289–2298 (2013)
8. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Trans. Neural Netw. **22**(2), 199–210 (2010)
9. Zhao, L.M., Yan, X., Lu, B.L.: Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence (2021)
10. Zheng, W.L., Liu, W., Lu, Y.F., Lu, B.L., Cichocki, A.: Emotionmeter: a multimodal framework for recognizing human emotions. IEEE Trans. Cybern. **49**(3), 1–13 (2018)
11. Zheng, W.L., Lu, B.L.: Personalizing EEG-based affective models with transfer learning. In: Kambhampati, S. (ed.) Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 2732–2739. IJCAI/AAAI Press (2016)

12. Zheng, W.L., Dong, B.N., Lu, B.L.: Multimodal emotion recognition using EEG and eye tracking data. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5040–5043. IEEE (2014)
13. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Mental Dev. **7**(3), 162–175 (2015)