A Cross-Culture Study on Multimodal Emotion Recognition Using Deep Learning

Lu Gan[1], Wei Liu[1], Yun Luo[1], Xun Wu[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, Shanghai, China
{ganlu_paristech,liuwei-albert,angeleader,stephanie_wx,bllu}@sjtu.edu.cn
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognition Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai, China

Abstract. In this paper, we aim to investigate the similarities and differences of multimodal signals between Chinese and French on three emotions recognition task using deep learning. We use videos including positive, neutral and negative emotions as stimuli material. Both Chinese and French subjects wear electrode caps and eye tracking glass while doing experiments to collect electroencephalography (EEG) and eye movement data. To deal with the problem of lacking data for training deep neural networks, conditional Wasserstein generative adversarial network is adopted to generate EEG and eye movement data. The EEG and eye movement features are fused by using Deep Canonical Correlation Analysis to analyze the relationship between EEG and eye movement data. Our experimental results show that French has higher classification accuracy on beta frequency band while Chinese performs better on gamma frequency band. In addition, EEG signals and eye movement data of French participants have complementary characteristics in discriminating positive and negative emotions.

Keywords: Emotion recognition · EEG · Eye movement · Deep learning · Cross-culture · Chinese · French

1 Introduction

Facial expressions, speech and non-verbal vocalizations are often used as input to recognize different emotions. Recent research found that facial expressions of emotion are not culturally universal [1]. People from different cultures can reach an agreement on the most intense emotion in judging facial expressions. However, culture differences are found when people judge the absolute level of emotional intensity [2]. Differences of non-verbal emotion cognition between western culture and remote tribe were also studied [3]. Cross-cultural similarities and differences

T. Gedeon et al. (Eds.): ICONIP 2019, CCIS 1142, pp. 670–680, 2019.
https://doi.org/10.1007/978-3-030-36808-1_73

appear in music mood perception as well. Research and the experimental results showed that listeners from different cultural backgrounds behaved differently in their selection of mood clusters and agreement ratio in each mood cluster. The similar result was found in Shuar hunter-horticulturalists from Amazonian Ecuador and American native English speakers [4]. However, it is widely agreed that cross-cultural agreement levels are lower than intra-cultural one [5,6].

With the quick development of brain-computer interface (BCI), many researches start to use neural signals to study the relationship between emotion and brain activities. EEG signals are proved to be effective in the field of emotion recognition. Recent researches indicated that there exists a stable neural pattern of EEG signals for positive, neutral and negative emotions [7]. Researchers also used EEG to investigate the differences of neural patterns between Chinese and Germans [8]. Combining EEG modality with other modalities provided an efficient way to recognize human emotions [9].

Eye movements have been widely used in studying attention, perceptions and emotion. Eye tracking data allow researchers to find users' areas of interest, attention track and subconscious behaviors. Therefore, more and more studies start to focus on the relationship between emotion and the movements of eyes. It was proved that higher trait emotional intelligence was associated with more attention to positive emotional stimuli [10]. The increase of gaze to eye region in children with autism spectrum disorders led to higher emotion recognition accuracy [11]. Furthermore, the characteristics of eye movements and EEG are complementary to emotion recognition [12]. Using modality fusion methods can significantly enhance the accuracy on emotion recognition task [13].

In this paper, we focus on investigating the similarities and differences of EEG and eye movement signals between Chinese and French on emotion recognition task using deep learning. The task is to classify positive, neutral and negative emotions. We evaluate the performance of emotion classification with different features and different frequency bands. Functional brain connectivity patterns are adopted to visualize the similarities and differences between Chinese and French. Since the complementary characteristics of EEG and eye movements in Chinese subjects have already been proven [12], we focus on the results for French participants. Multi-modality fusion algorithm is also used to reveal the relationship between EEG signals and eye movement data.

2 Methods

2.1 Functional Brain Connectivity Patterns

Functional brain connectivity patterns are used to visualize the neural patterns of Chinese and French participants instead of focusing on single-channel analysis [14]. Each EEG channel represents one node and the connections between pairs of channels are the links. To construct the functional brain network, we use spectral coherence to calculate the connectivity indices between two EEG channels under different frequency bands. Thus, one connectivity matrix can represent

one sample's brain network. Then we use critical subnetwork selection to choose the emotion-related subnetworks.

Critical subnetwork selection can be divided into several steps. Firstly, we calculate the average matrices for each emotion. The brain connectivity matrix of subjects under the same culture background are used to calculate the mean connectivity matrix. Secondly, we sort each mean connectivity matrix based on the absolute value of the connection weights. Since some weak connections between electrodes are not relevant to emotion and they may obscure the profile for the network topology, we discard the connections based on a proportional threshold. The connectivity matrices of positive, neutral and negative emotions are processed respectively. The intersection of connections under three emotions is considered to be less relevant to the specific emotion. Hence, these connections are removed from brain connectivity matrix in the visualization. The choice of threshold is based on the performance of classification. The topological feature strength is extracted from three critical subnetworks of each subject with different thresholds and then fed into a classifier. The threshold who can obtain the highest accuracy is considered to have remained the most emotion-related connections.

2.2 Augmentation of EEG and Eye Movement Data

To overcome the problem of lacking training data for deep neural network, we use Conditional Wasserstein Generative Adversarial Network (CWGAN) to generate both EEG and eye movement data [15]. CWGAN consists of two components. The generator G produces realistic-like data X_g by giving real data distribution X_r and generated data distribution X_g. The objective of generator is to confuse discriminator D which tries to distinguish whether a sample comes from X_r or X_g. The target is to solve the minimax problem during the adversarial training procedure. The formula is defined as:

$$\begin{aligned} \min_{\theta_G} \max_{\theta_D} L(X_r, X_g) = \mathbb{E}_{x_r \sim X_r}[log(D(x_r))] \\ + \mathbb{E}_{x_g \sim X_g}[log(1 - D(x_g))] \end{aligned} \tag{1}$$

where θ_g and θ_d represent the parameters of the generator and discriminator, respectively.

In CWGAN, the Earth-Mover distance (EMD, also known as Wasserstein-1 distance) is used to replace Jensen-Shannon divergence to calculate the distance between probability distribution of real data and generated data. Compared with Jensen-Shannon divergence, EMD is continuous and differentiable almost everywhere, which ensures the convergence of GAN and avoids the problem of mode collapse. To make training procedure more stable and convergence faster, a gradient penalty is added instead of using weight clipping [16].

In order to generate samples for multiple classes, label information is used. An auxiliary label Y_r is fed into both generator and discriminator. In the generator, X_z is concatenated with Y_r. In discriminator, X_r and X_g are concatenated

with Y_r to construct a hidden representation. The final objective function is defined as:

$$\begin{aligned}\min_{\theta_G}\max_{\theta_D} L(X_r, X_g, Y_r) = \\ \mathbb{E}_{x_r\sim X_r, y_r\sim Y_r}[D(x_r|y_r)] - \mathbb{E}_{x_g\sim X_g, y_r\sim Y_r}[D(x_g|y_r)] \\ - \lambda\mathbb{E}_{\hat{x}\sim\hat{X}, y_r\sim Y_r}[||\nabla_{\hat{x}|y_r} D(\hat{x}|y_r)||_2 - 1)^2]\end{aligned} \tag{2}$$

where λ is a hyperparameter controlling the trade-off between the original objective and gradient penalty, and $\hat{x}$ is defined as:

$$\hat{x} = \alpha x_r + (1-\alpha)x_g, \alpha \sim U[0,1], x_r \sim X_r, x_g \sim X_g \tag{3}$$

The loss of discriminator is the maximum term, and the loss of generator is the minimum term. They are optimized simultaneously. The discriminator loss is updated for critic times in each adversarial training iteration.

2.3 Multi-modality Fusion Approach

To analyze the characteristics of eye movements and EEG data, Deep Canonical Correlation Analysis (DCCA) is used [13]. For each modality, a neural network is constructed to realize nonlinear feature transformation which aims to represent original modality features in another feature space supposed to be related with emotion. The layer sizes for both modalities are the same. Then Canonical Correlation Analysis (CCA) is used to calculate the correlation between transformed features of two modalities. The back-propagation algorithm is adopted to update parameters of network in order to get higher correlation in CCA layer. The extracted features are fused by using the formula defined as follows:

$$F_{fusion} = \alpha M_1 + \beta M_2 \tag{4}$$

where M_1 and M_2 represent the extracted features for each modality, respectively, and α and β are the parameters to control the weight of each modality. Since we consider that EEG and eye movement features have an equivalent importance here, we choose $\alpha = \beta = 0.5$.

3 Experiment Setup

3.1 The SEED Dataset

The SEED[1] dataset is a public dataset for emotion recognition. Fifteen Chinese healthy subjects participated in the experiments to watch 15 Chinese film clips. Each subject was invited to participate in 3 sessions of experiments. The stimuli material contains positive, neutral and negative emotions. During the experiment, subjects were demanded to watch film clips attentively. 62-channel EEG signals based on international 10–20 system and eye movement signals were recorded at the same time.

[1] http://bcmi.sjtu.edu.cn/~seed/index.html.

3.2 Experiment for French Participants

To compare the results of Chinese with those of French, we have to keep consistency in experiment design and data collection. Thus, we choose film clips as stimuli material as well. Since French participants may not understand the expressions of emotion in Chinese films, film clips used in the experiments for French subjects are chosen from a large database of emotion-eliciting films developed by Schaefer *et al.* [17]. All the film excerpts were nominated by 50 experts and evaluated by 364 Belgian French-speaking undergraduates. We add film clips with highest Positive And Negative Affect Schedule (PANAS) into our stimuli material. Due to the lack of neutral excerpts, extra neutral excerpts are chosen from calm landscape films, which are consistent with SEED dataset. Finally, 21 film excerpts are chosen.

Six healthy subjects aged from 22 to 41 participated in the experiments. All of the subjects come from France and their native language is French. Since all the subjects are exchange students and professors on the campus, the number of subjects are limited. Each participant was required to perform the experiments for two sessions. During experiment, participants were asked to immerse in the film clips. 62-channel EEG signals based on international 10–20 system and eye movement signals were recorded simultaneously.

3.3 Feature Extraction and Classification

To keep balance between the number of Chinese subjects and French subjects, we randomly choose 6 subjects from the SEED dataset. In order to keep consistency with the number of sessions each French subject participated, two sessions of a Chinese subject are chosen. We apply the same data preprocessing and feature extraction methods on Chinese and French subjects.

The EEG data are downsampled to 200 Hz and transformed by a Short-Term Fourier Transform (STFT) with an 1-s Hamming window. By using a band-pass filter from 1 to 50 Hz, it allows us to filter out a large part of artifacts. Power Spectral Density (PSD), Differential Entropy (DE), Rational Asymmetry (RASM), Differential Asymmetry (DASM), Asymmetry (ASM) and Differential Causality (DCAU) features are extracted from five frequency bands: δ: 1–3 Hz, θ: 4–7 Hz, α: 8–13 Hz, β: 14–30 Hz, and γ: 31–50 Hz. The data recorded from the same film excerpt are labeled as the same label. The features extracted from EEG usually contain noises which cannot be thoroughly filtered. Therefore, we use linear dynamic system (LDS) approach to filter out the unrelated features for emotion recognition.

As the eye movement data contain different parameters, every eye movement parameter is processed separately. We adopt the same extracted features of eye movement in the work of Lu *et al.* [12] since these features were proven to be effective in emotion recognition. We also apply LDS to filter out unrelated features for eye movement data. The total number of dimension of eye movement features is 33. The details of eye movement features are presented in Table 1.

We use an SVM with linear kernel as a classifier. All the results are obtained by a 5-fold cross validation. The parameter c is searched from 2^{-10} to 2^{9}.

Table 1. Details of extracted features from Eye Movement

Eye movement parameters	Extracted features
Pupil diameter (X and Y)	Mean, standard deviation and DE in four bands: 0–0.2 Hz, 0.2–0.4 Hz, 0.4–0.6 Hz, 0.6–1 Hz
Dispersion (X and Y)	Mean, standard deviation
Fixation duration (ms)	Mean, standard deviation
Blink duration (ms)	Mean, standard deviation
Saccade	Mean, standard deviation of saccade duration (ms) and saccade amplitude (°)
Event statistics	Blink frequency, fixation frequency, fixation duration maximum, fixation dispersion total, fixation dispersion maximum, saccade frequency, saccade duration average, saccade amplitude average, saccade latency average

4 Experiment Results

4.1 Comparison on Features

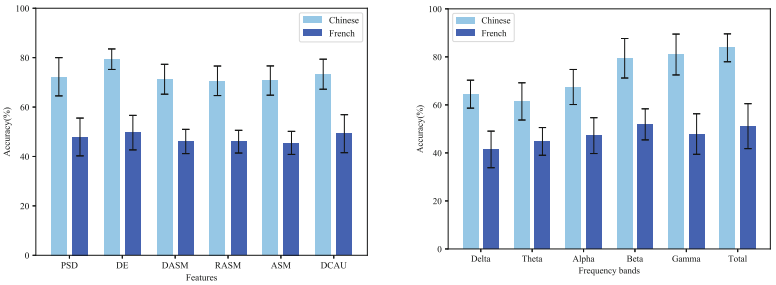
In this part, we compare the performance of emotion classification on different features. Figure 1(a) shows the classification accuracy for Chinese and French subjects.

We can see that the mean accuracy of Chinese reaches 72.93%, which is much higher than the mean accuracy of French (47.39%). The gap of accuracy between Chinese and French shows that the emotions of Chinese have been stimulated effectively while the emotions of French are relatively difficult to stimulate. The unfamiliar environment may make French subjects feel difficult to relax and immerse in the films. The standard deviation (SD) of Chinese (5.98) is close to the SD of French (6.08), indicating that the individual differences exist on both datasets. We also use two-way analysis of variance to study the statistical significance of nation and features. The p-values for the nation (0.0000), the features (0.0000), and the interaction between nation and features (0.3695) indicates that the nation and features affect the accuracy, but there is no evidence of an interaction effect of the two.

Among different features, DE feature achieves the highest classification accuracy on both datasets, 79.37% with Chinese subjects and 49.65% with French subjects. DE feature gets the lowest SD on Chinese dataset which means DE feature is a relatively stable feature for emotion recognition for Chinese subjects.

4.2 Comparison on Frequency Bands

We also compare the classification accuracy on five non-overlapping frequency bands. The results are shown in Fig. 1(b). The mean accuracy of Chinese subjects achieves 72.92% (SD = 7.23) and that of French is 47.38% (SD = 7.52).



(a) On different features (b) On different frequency bands

Fig. 1. Classification accuracies on different features and bands

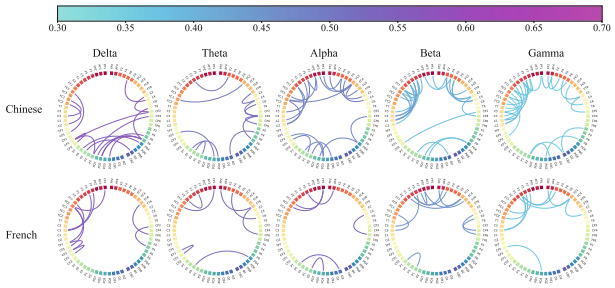
By using two-way analysis of variance, the p-values for the nation (0.0000), the frequency bands (0.0000), and the interaction between nation and frequency bands (0.2887) indicates that the nation and the frequency bands affect the accuracy, but there's no interaction between nation and the frequency bands. For Chinese, the performance on higher frequency bands, such as beta and gamma, is better than that of lower frequency bands. The finding is consistent with the existing work [18]. Total frequency band, which means to concatenate all frequency bands together, gets the highest accuracy (83.77%) with regards to Chinese subjects. For French, we find that on beta frequency band the best result (51.89%) is obtained. Unlike Chinese subjects, gamma frequency band has a relatively poor for French subjects performance (47.86%) compared with that of Chinese subjects (80.98%).

4.3 Functional Brain Connectivity Patterns

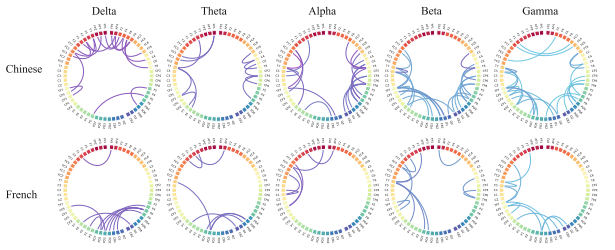
Figure 2 shows the functional brain connectivity patterns of Chinese and French with three emotions in five frequency bands. There are more connections of Chinese than those of French. It is because that French has a larger number of intersections shared by three emotions, which have been removed from visualization. Here, we choose 0.2 as threshold, which means 20% of total connections have been discarded. We get the highest mean accuracy for Chinese (71.24%) and French (44.25%) when threshold equals to 0.2.

For both Chinese and French, we can observe higher coherence connectivity of frontal lobes in positive emotion on alpha, beta and gamma frequency bands. The connectivity patterns on neutral and negative emotions are relatively similar on beta and gamma frequency bands. For Chinese, we find higher coherence connectivity on temporal and occipital lobes. For French, the higher coherence is found especially on left hemisphere. Watching positive film excerpts, the temporal and occipital sites of Chinese subjects show higher coherence while French subjects show higher coherence at frontal and temporal sites. Watching neutral film excerpts, higher coherence connectivities are located at frontal sites for

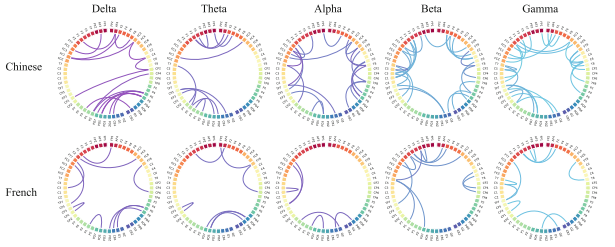
Chinese but at occipital sites for French. For both Chinese and French, higher coherence is found on lower frequency bands. However, unlike Chinese subjects, who have a relatively symmetry distribution of connectivities, French are relatively asymmetry and higher coherence connectivities appear in left hemisphere.



(a) positive



(b) neutral



(c) negative

Fig. 2. The functional brain connectivity patterns for three emotions in five frequency bands with coherence as the connectivity index. The text on each node means the name of electrode. The nodes from top to bottom represent EEG channels from the frontal, temporal, parietal to the occipital lobes. Here, the maps from first row of each emotion represent the results of Chinese and those from second row represent the results of French.

4.4 Multi-modality Fusion

Considering the lower sample rate of eye tracking glass, an STFT with a 4-second non-overlapping window is used to compute both EEG and eye movement features. Because of lack of data, we use CWGAN as data augmentation method. Since DE feature has the best performance, we use DE on total frequency band as input to the network. When it comes to eye movement data, all features have been concatenated to input into the network. Both EEG and eye movements data have been generated.

Both networks for generator and discriminator have 4 layers. We use grid search to find the optimized number of nodes for each layer. As a result, the hidden layers of the generator and discriminator networks have 512 nodes for EEG data and 64 nodes for eye movement data, respectively. ReLU (Rectified Linear Unit) is used for all hidden layers. The networks are optimized by Adam optimizer. We choose learning rate as 10^{-3}. The critic value is set to 5 and λ is set to 10. The generated data are sampled from a uniform distribution $U[-1, 1]$. During the training, the discriminator loss quickly converges to a value close to 0, which indicates that the distribution of real data and generated data are very similar. Therefore, the generated DE data and eye movement data have high quality.

Table 2. Performance of Data Augmentation

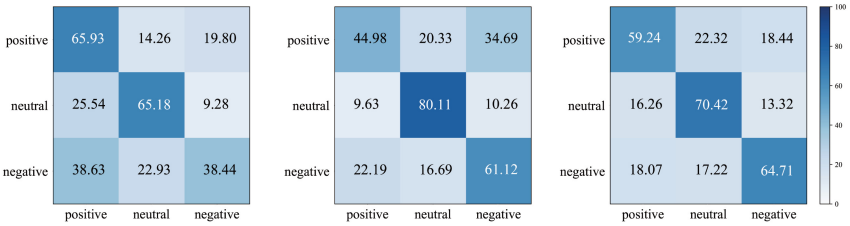
	$0\times$ dataset	$1\times$ dataset	$2\times$ dataset	$3\times$ dataset	$4\times$ dataset
EEG	0.4997	0.5160	0.5155	**0.5206**	0.5202
Eye	0.6381	**0.6603**	0.6448	0.6595	0.6504

Table 2 shows the performance of data augmentation. The generated data are appended to each 5-fold training data and an SVM with linear kernel is used. There are augmentations of classification accuracy to different extent depending on the number of generated data appended to the original dataset. Since triple generated data appended to the real dataset has the highest mean accuracy, we use the dataset including triple generated data and real data as EEG and eye movement dataset in the following part of this paper. The generated data are only used in training set.

DCCA is used to figure out whether the characteristics of eye movements are complementary with EEG. Each modality is constructed by three full connected layers. We use random search between 50 and 200 to find the optimal number of layer nodes. The learning rate is set to 10^{-3}. Batch size is set to 100 and regulation parameter is set to 10^{-7}. We choose the output dimension of features for each modality as 20.

The mean accuracy by using EEG data only is 55.35% and the mean accuracy by using eye movements only is 60.98%. When we combine two modalities and project them into another feature space with lower dimension, we get an

augmentation of classification accuracy to 64.22%. Figure 3 shows the confusion matrices of classification results. From Fig. 3, we have found that eye movement and EEG modalities have complementary characteristics. By using EEG features solely, it's very likely to confuse negative emotion with other two emotions while using eye movements alone shows a better performance. When it comes to discriminate positive emotions, using EEG features solely shows a better performance. After combining two modalities, we find that the negative emotion can be recognized with higher accuracy (64.71%).



(a) EEG features (b) eye movements (c) Two-modality

Fig. 3. The confusion matrices of classification results by using different features.

5 Conclusions and Future Work

In this paper, we have compared the neural patterns between Chinese and French on a task of recognizing three emotions (positive, neutral and negative). We have found that French has higher mean accuracy on beta frequency band while Chinese tends to perform better on gamma frequency band. The functional brain connectivity patterns indicate the coexistence of similarities and differences of neural patterns between Chinese and French subjects. The results of classification by using DCCA reveal that EEG and eye movement data of French subjects are complementary in discriminating positive and negative emotions.

As future work, we will recruit more number of subjects to participate in the experiments and use different multi-modality fusion methods to investigate the relationship between EEG signals and eye movement data.

Acknowledgements. This work was supported in part by the grants from the National Key Research and Development Program of China (Grant No. 2017YFB100 2501), the National Natural Science Foundation of China (Grant No. 61673266), and the Fundamental Research Funds for the Central Universities.

References

1. Jack, R.E., Garrod, O.G.B., Yu, H., Caldara, R., Schyns, P.G.: Facial expressions of emotion are not culturally universal. Proc. Natl. Acad. Sci. **109**(19), 7241–7244 (2012)
2. Ekman, P., et al.: Universals and cultural differences in the judgments of facial expressions of emotion. J. Pers. Soc. Psychol. **53**(4), 712–717 (1987)

3. Sauter, D.A., Eisner, F., Ekman, P., Scott, S.K.: Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc. Natl. Acad. Sci. **107**(6), 2408–2412 (2010)
4. Bryant, G., Barrett, H.C.: Vocal emotion recognition across disparate cultures. J. Cogn. Culture **8**(1–2), 135–148 (2008)
5. Elfenbein, H.A., Ambady, N.: On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychol. Bull. **128**(2), 203 (2002)
6. Hutchison, A.N., Gerstein, L.H.: The impact of gender and intercultural experiences on emotion recognition. Revista De Cercetare Si Interventie Sociala **54**, 125 (2016)
7. Zheng, W.-L., Zhu, J.-Y., Lu, B.-L.: Identifying stable patterns over time for emotion recognition from EEG. IEEE Trans. Affect. Comput. (2017)
8. Wu, S., Schaefer, M., Zheng, W.-L., Lu, B.-L., Yokoi, H.: Neural patterns between Chinese and Germans for EEG-based emotion recognition. In: 8th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 94–97. IEEE, Shanghai (2017)
9. Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M.: Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Trans. Affect. Comput. **7**(1), 17–28 (2016)
10. Lea, R.G., Qualter, P., Davis, S.K., Pérez-González, J.C., Bangee, M.: Trait emotional intelligence and attentional bias for positive emotion: an eye tracking study. Pers. Individ. Differ. **128**, 88–93 (2018)
11. Bal, E., Harden, E., Lamb, D., Van Hecke, A.V., Denver, J.W., Porges, S.W.: Emotion recognition in children with autism spectrum disorders: relations to eye gaze and autonomic state. J. Autism Dev. Disord. **40**(3), 358–370 (2010)
12. Lu, Y., Zheng, W.L., Li, B., Lu, B.L.: Combining eye movements and EEG to enhance emotion recognition. In: IJCAI 2015, pp. 1170–1176 (2015)
13. Qiu, J.-L., Liu, W., Lu, B.-L.: Multi-view emotion recognition using deep canonical correlation analysis. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11305, pp. 221–231. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04221-9_20
14. Wu, X., Zheng, W.-L., Lu, B.-L.: Identifying functional brain connectivity patterns for EEG-based emotion recognition. In: 9th International IEEE/EMBS Conference on Neural Engineering. IEEE, San Francisco (2019)
15. Luo, Y., Lu, B.-L.: EEG data augmentation for emotion recognition using a conditional wasserstein GAN. In: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2535–2538. IEEE, Honolulu (2018)
16. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems, NIPS, Long Beach, pp. 5767–5777 (2017)
17. Schaefer, A., Nils, F., Sanchez, X., Philippot, P.: Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. Cogn. Emot. **24**(7), 1153–1172 (2010)
18. Zheng, W.-L., Lu, B.-L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)