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/EMBC48229.2022.9871630

Multimodal Emotion Recognition in Response to Oil Paintings

Shuai Luo, Yu-Ting Lan, Dan Peng, Ziyi Li, Wei-Long Zheng and Bao-Liang Lu*, *Fellow, IEEE*

***Abstract*— Most previous affective studies use facial expression pictures, music or movie clips as emotional stimuli, which are either too simplified without contexts or too dynamic for emotion annotations. In this work, we evaluate the effectiveness of oil paintings as stimuli. We develop an emotion stimuli dataset with 114 oil paintings selected from subject ratings to evoke three emotional states (*i.e.*, negative, neutral and positive), and acquire both EEG and eye tracking data from 20 subjects while watching the oil paintings. Furthermore, we propose a novel affective model for multimodal emotion recognition by 1) extracting informative features of EEG signals from both the time domain and the frequency domain, 2) exploring topological information embedded in EEG channels with graph neural networks (GNNs), and 3) combining EEG and eye tracking data with a deep autoencoder neural network. From the experiments, our model obtains an averaged classification accuracy of 94.72% ± 1.47%, which demonstrates the feasibility of using oil paintings as emotion elicitation material.**

I. INTRODUCTION

In the last few years, various kinds of emotion elicitation materials have been utilized for emotion recognition studies. Zheng and Lu recorded EEG signals of the subjects during movie watching [1]. Li *et al.* used photographs of smile and cry facial expressions to elicit happy and sad emotions [2]. Lin *et al.* selected Oscar's film soundtracks as stimuli [3]. Emotion elicitation materials can be categorized into two types: (a) static cues, such as pictures and paintings, and (b) dynamic cues, such as films and music [4].

With the first use of movie clips for emotion elicitation in a study designed to investigate the impact of fear, anger and sexual arousal on blood pressure [5], movies have been selected by most previous experiments as stimuli to create emotional states in the laboratory for scientific purposes. However, dynamic cues such as movies may contain scenery changes, luminance variation, narrative development or other dynamic changes in the information array that can complicate interpretation of the elicited affective response [4]. For example, Yang *et al.* found that the contrast control of visual contents affects arousal on a very minute scale; the reduction of contrast induces a decrease in valence, while the enhancement of contrast induces an increase [6]. Camera movement, editing, sound and other film technical elements are typically used to emphasize plot lines and their emotional meaning. However, each of these techniques may introduce noise under experimental conditions, as it is difficult to accurately annotate emotional dynamics continuously with complex stimuli dynamics.

In contrast, static cues, including affective pictures and paintings and their physical properties, such as contrast, clarity, and color-saturation, can be uniformly adjusted as control variables so as to avoid the interference. In addition, most artworks are created to raise strong emotional responses and emotions in aesthetics are contained within the narrative [7]. It has been proven that emotions arising from aesthetic experience are different from those arising from photographs [8]. The perception, interpretation and the consequent emotion raised in the recipient by the art are subject to factors such as familiarity, complexity, curiosity and the aesthetic spirit of the viewer [9]. Therefore, emotions evoked in an aesthetic environment can be more intense or more diverse than those evoked from other static cues. Artworks could be another potential approach to study how our brain perceives and processes affective information.

In this paper, we use oil paintings as stimuli to evoke three types of emotions, namely, negative, neutral, and positive. To evaluate the effectiveness of the oil paintings as emotion elicitation material, we extract features of EEG signals in the time and frequency domains, integrate these features which are scattered across EEG channels into graph-level embedding via a simple graph convolutional network (SGC) [10], and obtain the high-level representations of EEG and eye movement signals with a deep autoencoder neural network.

II. DATA COLLECTION

A. Stimuli and Subjective Ratings

The emotional stimuli used in this paper is composed of 114 carefully selected oil paintings from the publicly available artwork dataset WikiArt[1], which covers paintings created from the mid-16^{th} century to the 19^{th} century. The paintings cover most of the major art-styles (*e.g.*, baroque, rococo, realism, post-impressionism) and 5 genres (*i.e.*, portrait, animal, still life, cityscape and landscape). All the paintings are uniformly adjusted in contrast, clarity, color-saturation and sharpness with Adobe Lightroom Classic,

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.

S. Luo, Y. T. Lan, Z. Y. Li, 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 the Brain Science and Technology Research Center, Shanghai Jiao Tong University, 800 Dong Chuan Rd., Shanghai 200240, Peoples Republic of China.

D. Peng and B. L. Lu are with the RuiJin-Mihoyo Laboratory, Clinical Neuroscience Center, RuiJin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin 2nd Rd., Shanghai 200020, Peoples Republic of China.

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

[1]https://www.wikiart.org

Fig. 1. Illustration of 114 oil paintings that are annotated as positive, neutral, and negative.

so as to eliminate the influence of these factors. For each painting, we asked at least 7 annotators to report their emotional response after observing the painting.

Specifically, we distributed questionnaires on the perception and emotional response of artworks among students of Shanghai Jiao Tong University and China Academy of Art. Each questionnaire contained 30 paintings from the previously selected 114 artworks. Participants were required to report what kind of emotion (negative, neutral or positive) was evoked by each painting, and to choose a number from 1 to 5 to describe the intensity of the emotion. In total, we collected 217 rating samples, and built a dataset with 48 negative paintings, 30 neutral paintings, and 36 positive paintings as illustrated in Fig.1.

B. Experimental Procedure

EEG data in this work was collected from 20 subjects (9 males and 11 females; age 22.35 $\pm$ 2.82) during artwork appreciation. All of them were recruited from Shanghai Jiao Tong University. The experiment has been approved by the Scientific & Technical Ethics Committee at Shanghai Jiao Tong University. Since Batt *et al.* observed different cortical activities in response to paintings in artists and in non-artists [11], we kept the recruited subjects balanced in the evaluation of artistic accomplishment. Each participant was asked to fill out an art experience questionnaire [8]. Following the criterion proposed in [8], 9 of the subjects were considered as artistically experienced, and the other 11 subjects were considered as artistically naive. As we did not expect any brain laterality effects for the objective of our study, all participants were right-handed. None of them were visually impaired or had ever been diagnosed with any affective disorders.

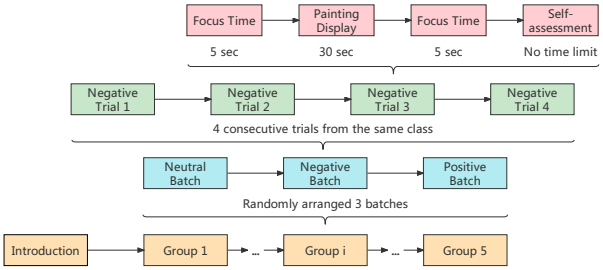

Fig. 2. The experimental protocol.

The experimental protocol is shown in Fig.2. There were 60 trials in each experiment. To prevent the subjects from changing emotions frequently, we organized 60 trials into 5 groups. Each group contained 3 batches, each batch consisted of 4 consecutive paintings that were randomly picked from the same class, and the order of the batches within each group was also randomly decided. Prior to the trials, participants were informed of the experimental protocol, and three example paintings corresponding to negative, neutral, and positive emotions were shown. This served as a tutorial and anchored the base for the emotions. For each trial, there was a 5 s focus time before the stimulus onset. Each painting was displayed for a fixed time of 30 s, and participants were asked to intently appreciate the painting.

In the whole experiment, EEG data of the subjects were recorded at a 1000 Hz sampling rate using the ESI NeuroScan System with a 62-channel module arranged according to the international 10-20 system. Eye movements of the subjects were also recorded at a sampling rate of 120 Hz using a Tobii Pro X3-120 screen-based eye tracker. After the presentation of the painting, there was a 5 s focus time, and

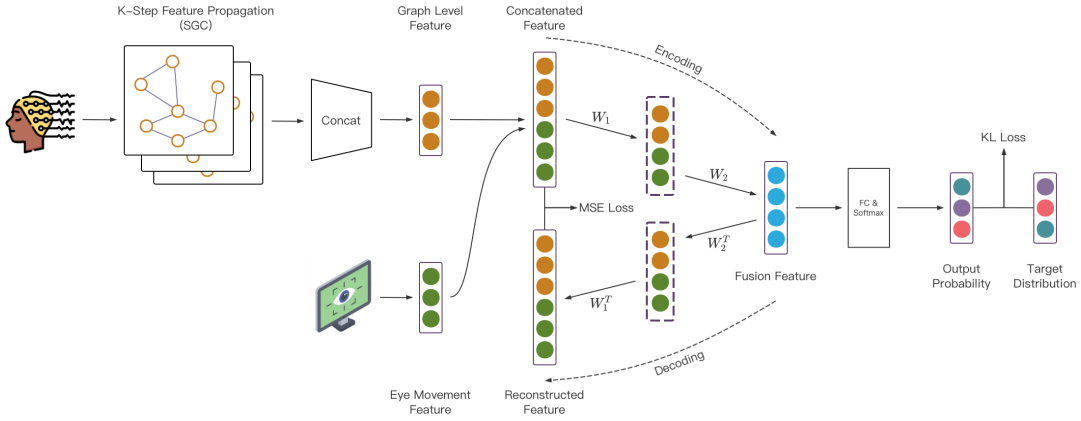

Fig. 3. The overall architecture of our proposed model. Here, concat denotes the vector concatenation operation, MSE denotes the mean squared error, and KL denotes the Kullback-Leibler divergence.

then the participants were instructed to report their emotional reaction following a 2-D valence-arousal emotion model. Emotional reaction reports with negative, zero, and positive valence ratings were labeled as negative, neutral, and positive emotions, respectively. These self-reported emotion labels were used to validate the classification results.

III. EMOTION CLASSIFICATION

A. Data Processing

For the EEG data, the raw EEG signal was downsampled to 200 Hz to reduce computation complexity, and then a bandpass filter between 1 Hz and 50 Hz was applied in order to eliminate power-line contamination and low-frequency noise. In addition, we employed independent component analysis (ICA) to remove components of muscle activity and eye movement artifacts. With visual inspection, EEG epochs that were heavily contaminated were removed. EEG epochs with an arousal rating less than 3 were marked as failing to evoke emotions in the subjects effectively and thus were also removed. We divided the 30 s length trials into 5 s short trials so as to increase the number of samples. Therefore, each experiment consisted of approximately 360 short trials.

Since Duan *et al.* attested that differential entropy (DE) performs better than other EEG features such as the energy spectrum (ES) in emotion-related tasks [12], we used the DE feature, which reflected an energy change in different frequency bands over time, to fully utilize the information of the EEG data in both the frequency domain and the time domain. Using short-term fourier transform (STFT) with 1 s Hanning window without overlapping, we extracted the DE features in the following five frequency bands: delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-49 Hz).

For eye tracking data, after eliminating the light reflex in the pupil diameter, various features, such as pupil diameter, saccade, blink and fixation, were used. Details can be found in [13].

B. Classifier

Our model structure is shown in Fig.3. The topology of the EEG electrodes can be regarded as a network with 62 nodes, and each electrode is a node within the network. For each 5 s short trial, the EEG data can be represented as a feature matrix $X \in \mathbb{R}^{n\times d}$, where n is the number of EEG channels and d is the input feature dimension. There are inter-channel relations in the EEG signals, which can be described as a weighted adjacency matrix $A \in \mathbb{R}^{n\times n}$. Here, we adopt a simple graph convolutional network (SGC) to learn a feature transformation function for input X. We define the feature transformation of each layer as follows:

$$H^{l+1} = SH^lW^l, (l = 0, 1, \ldots, L-1) \tag{1}$$

where $S = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$, $\tilde{A} = A + I_n$ is the adjacency matrix with self-loops, $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$, L denotes the number of layers, or the number of times for message passing between neighboring nodes, and W^l is a learnable weight matrix at layer l.

The output feature of the last layer can be represented as $Z \in \mathbb{R}^{n\times d'}$, where d' denotes the output feature dimension. Then, we have the following:

$$Z = H^L = SH^{L-1}W^{L-1} = S^LXW \tag{2}$$

where $W = W^0W^1 \ldots W^{L-1}$. This graph convolution operator exploits the topological information of EEG channels and outputs a learned feature representation for each node.

To maximize the retention of the information of each node, we concatenate the learned features of 62 nodes into graph-level features $\hat{Z}$, and together with the eye movement feature E, we feed them into a deep autoencoder to learn a fine-grained representation for these two modalities. The encoding network consists of two fully connected layers, and the decoding network performs the inverse process symmetrical to the encoding network. The encoded vector is finally fed into two fully-connected layers to obtain the output distribution over all classes $\hat{Y} \in \mathbb{R}^C$, where C is the number of classes. Let $G_{en}(\cdot)$, $G_{de}(\cdot)$, and $G_c(\cdot)$ be the network

functions of the encoding network, the decoding network and the label classifier, respectively. $\hat{Y}$ can be computed as follows:

$$\hat{Y} = softmax(G_c(G_{en}(\hat{Z} \oplus E))) \tag{3}$$

where $\oplus$ denotes the vector concatenation operation.

C. Model Training

Inspired by [14], we convert the self-reported label $Y \in \{0,1,2\}$ into a label distribution $\tilde{Y} \in \mathbb{R}^C$, and the conversion is defined as follows:

$$\tilde{Y} = \begin{cases} (1-\frac{2\varepsilon}{3}, \frac{2\varepsilon}{3}, 0), & Y=0 \\ (\frac{\varepsilon}{3}, 1-\frac{2\varepsilon}{3}, \frac{\varepsilon}{3}), & Y=1 \\ (0, 1-\frac{2\varepsilon}{3}, \frac{2\varepsilon}{3}), & Y=2 \end{cases} \tag{4}$$

where $\varepsilon \in [0,1]$ is a hyperparameter that controls the noise level. In this work, we empirically set $\varepsilon = 0.1$.

During model training, we minimize two kinds of losses, including the Kullback-Leibler (KL) divergence between output probability and the refined label distribution, and the mean squared error (MSE) between the concatenated features and the reconstructed features. The overall loss function is as follows:

$$\Phi = KL(\hat{Y}, \tilde{Y}) + \alpha \cdot MSE(\hat{Z} \oplus E, G_{de}(G_{en}(\hat{Z} \oplus E))) \tag{5}$$

where α is a hyperparameter that can be tuned.

IV. RESULTS AND DISCUSSION

We conduct subject-dependent experiments for emotional recognition. Following the experimental setting in [2], we randomly divide the trials into a training set and a test set with a ratio of 5:1 for each subject, and we adopt a sixfold cross validation strategy to evaluate the performance. A support vector machine (SVM) is chosen as the baseline, and Table I presents the classification accuracy and the F1-score averaged across 20 subjects.

TABLE I

THE SUBJECT-DEPENDENT CLASSIFICATION ACCURACY AND THE F1-SCORE (MEAN/STD) AVERAGED ACROSS 20 SUBJECTS

Freq. Band	SVM		Our model	
	Acc.	F1	Acc.	F1
delta	81.47/05.04	81.05/05.23	**88.38/04.83**	**84.79/07.63**
theta	82.69/05.52	82.31/05.80	**87.85/05.08**	**83.47/07.69**
alpha	83.91/04.83	**83.62/04.99**	**85.06/06.97**	79.96/09.86
beta	87.43/04.22	**87.28/04.30**	**88.90/06.30**	84.32/10.25
gamma	86.74/04.82	86.57/04.91	**92.21/04.52**	**88.57/08.27**
all bands	89.12/04.26	89.03/04.31	**94.72/01.47**	**90.22/01.61**

Both two methods achieve high prediction accuracy and F1-scores on the beta band and the gamma band, indicating that these two frequency bands are critical in the task of artwork-induced emotional classification. In addition, our model performs better than the SVM over all frequency bands with respect to prediction accuracy. For the F1-score, the SVM outperformed our model over the alpha band and the beta band. However, our model obtains the best prediction accuracy (94.72% $\pm$ 1.47%) and the best F1-score (90.22% $\pm$ 1.61%) when using data from all frequency bands, which demonstrates its effectiveness.

V. CONCLUSIONS

In this paper, we analyze the defects of widely used dynamic cues, such as films, and provide new insights into the selection of emotion elicitation material. We use oil painting as an emotion stimulus to evoke three emotional states (*i.e.*, negative, neutral and positive) in subjects. Furthermore, we propose a deep learning model to classify these three emotions based on EEG data and eye tracking data. The proposed multimodal emotion recognition model achieves an averaged prediction accuracy of 94.72% $\pm$ 1.47% across 20 subjects for 5 s length trials. In addition, our experimental results suggest that the beta band and the gamma band are two critical frequency bands in the task of emotion classification, which is consistent with the evidence in the literature [1][2].

REFERENCES

[1] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[2] M. Li and B.-L. Lu, "Emotion classification based on gamma-band EEG," in *IEEE EMBC*, 2009, pp. 1223–1226.

[3] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[4] J. Coan, *Handbook of emotion elicitation and assessment*. Oxford University Press, 2007.

[5] J. C. Scott, "Systolic blood-pressure fluctuations with sex, anger and fear." *Journal of Comparative Psychology*, vol. 10, no. 2, p. 97, 1930.

[6] H. Yang, J. Han, and K. Min, "Emotion variation from controlling contrast of visual contents through EEG-based deep emotion recognition," *Sensors*, vol. 20, no. 16, p. 4543, 2020.

[7] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe, "In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings," in *ACM MM*, 2012, pp. 349–358.

[8] A. Chatterjee, P. Widick, R. Sternschein, W. B. Smith, and B. Bromberger, "The assessment of art attributes," *Empirical Studies of the Arts*, vol. 28, no. 2, pp. 207–222, 2010.

[9] H. Hagtvedt, V. M. Patrick, and R. Hagtvedt, "The perception and evaluation of visual art," *Empirical Studies of the Arts*, vol. 26, no. 2, pp. 197–218, 2008.

[10] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *ICML*, 2019, pp. 6861–6871.

[11] R. Batt, M. Palmiero, C. Nakatani, and C. van Leeuwen, "Style and spectral power: processing of abstract and representational art in artists and non-artists," *Perception*, vol. 39, no. 12, pp. 1659–1671, 2010.

[12] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *IEEE NER*, 2013, pp. 81–84.

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

[14] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affective Computing*, 2020.