Affective Computation Based on Multi-modality Learning and Fusion

Xiang Zhang
Department of Computer Science, State University of New York at Binghamton, NY, USA
zxiang4@binghamton.edu

***Abstract*— Affective computing aims to enable intelligent systems to recognize, feel, interpret, and simulate human affects. It is an interdisciplinary field crossing computer science, psychology, and cognitive science. Multimodal human affect analysis has been studied for decades. With various data representations, researchers are looking for an efficient way to unify their strength in pursuit of more generalized features. Recently, we developed a new database(BU-EEG) by collecting EEG signals and facial action videos. The database was evaluated through the experiments on both posed and spontaneous emotion recognition with images alone, EEG alone, and EEG fused with images, respectively. The result validate the peripheral information e.g., EOG-like [1] and EMG-like [2] artifacts can be used as complementary features for benefiting both posed facial expression and spontaneous emotion analysis. Meanwhile, the two-modality feature fusion performs better than the single-modality feature alone in terms of the facial expression classification.**

My Ph.D. research will continue the expansion of the BU-EEG dataset, and will design a deep model for EEG-based late fusion, as well as explore more modalities fusion, e.g., texture, audio, 3D, thermal. In multimodal learning and fusion, 3D domain, a better representation of dynamic facial surfaces, is not trivial to fully utilize. Considering 3D face reconstruction as a powerful tool, I plan to design a robust and extendable algorithm to tackle this challenge.

I. Summary of existing work

A. EEG-Based Multi-Modal Emotion Database

Recently, I worked on a project of analyzing EEG signal and facial images and co-authored a paper [3] that has been accepted by FG 2020. In this paper, we present a new EEG-based multi-modal emotion database with posed expressions, action units, and spontaneous emotions. 29 participants were asking to imitate 6 facial expression, 10 action units. The spontaneous emotion trail were designed by one meditation trail and one pain task, which is stimulated by ice water. One front camera and one EEG collection machine are recording the participants' data collection.

Fig. 1 shows the data collection at work, corresponding EEG electrodes with a frontal view, and EEG electrodes location information, respectively.

As a baseline, the database has been evaluated through the experiments on both posed and spontaneous emotion recognition with images alone, EEG alone, and EEG fused with images, respectively. In order to represent the EEG features in a two-dimensional format which is compatible to the 2D images, we take the following steps to process the EEG signals for feature extraction and feature map generation. First, we apply a band-pass filter on the EEG data. Second, we extract the features and generate the feature map. Third, we apply the Kalman filter to smooth the extracted features map. Last, we normalize the extracted features and save them to 2D gray images as the extracted feature maps. After extracting the DE feature, based on different filter, we generate three different feature maps (FP), namely is Feature A, Feature B, and Feature C, respectively. Fig. 2 shows the pipeline from data acquisition to the subsequent experiment. The specification of EEG features is shown in Table I.

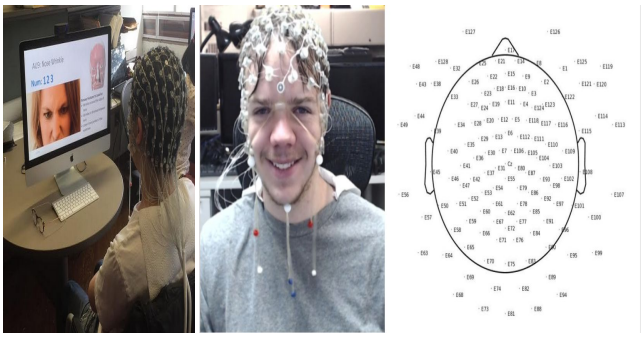

Fig. 1: The Experiment scene and EEG electrodes location information.

Fusion strategy Feature level fusion (FLF) was employed to fuse the features from two modalities (facial expression images and EEG signals). Multimodal feature fusion is expected to bring more considerable performance improvement of recognition for the spacial and temporal information they carry. We concatenate the facial expression and their corresponding EEG feature map directly to form a fused feature map before feeding them into the model. We employed three types of fusion features, which are the combinations of EEG Feature A, B and C. The concatenated fusion features were resized to the same 2D dimension (128×128). Fig. 3 shows an example of EEG Feature C and facial image fusion.

We used the same DANN model for a fair evaluation of the performance of both fused features and single modality features.

Table II shows the performance of single modal features and image-EEG fused features for posed expression recognition using DANN [4]. First, it clearly shows that EEG feature B and EEG feature C perform better than EEG Feature A because the high-frequency EEG signals (over 50HZ) are included in the feature maps. Such high-frequency EEG signals provide necessary complementary information associating with individual facial expressions to improve the classification performance significantly. Second,

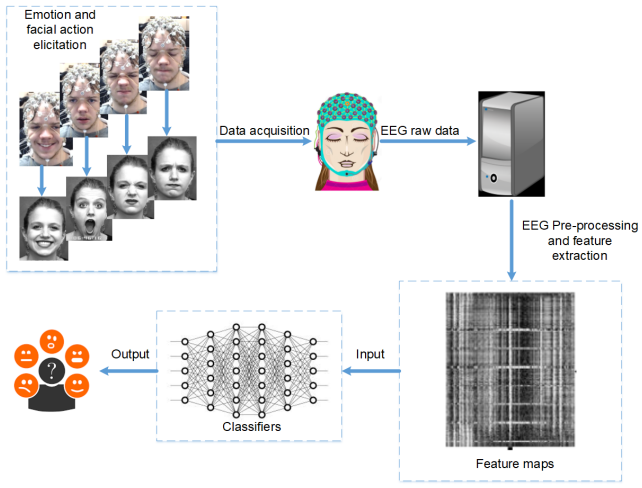

Fig. 2: EEG data processing, feature extraction and expression/emotion recognition.

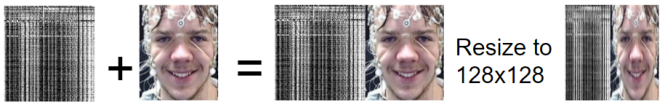

Fig. 3: Early fusion of EEG Feature C and facial image.

the performance based on the facial expression images is superior to the the performance of EEG-based single modal features by using DANN. Finally, the two-modal fusion based method generally outperforms the single modal feature based method for posed facial expression recognition.

Table III shows the performance of single modal features and two-modal fused features for spontaneous emotion recognition using DANN.

Table IV illustrates the confusion matrix of the fusion method using DANN.

The validation result shows that the peripheral information e.g., EOG-like [1] and EMG-like [2], [1] artifacts can be used as complementary features for benefiting both posed facial expression and spontaneous emotion analysis. Our validation experiments shows that the two-modality feature

TABLE I: Detailed description of 3 types of Extracted EEG Features.

Feature Map Type	Feature A	Feature B	Feature C
Feature Map Size	5×128	7×128	100×128
Description	5 frequency bands (0.1∼4 Hz, 4∼8 Hz, 8∼14 Hz, 14∼31 Hz, 31∼50 Hz)	7 frequency bands (0.1∼4 Hz, 4∼8, 8∼14 Hz, 14∼31 Hz, 31∼50 Hz, 50∼75 Hz, 75∼100 Hz)	100 frequency bands (0.1∼1 Hz, 1∼2 Hz, 2∼3 Hz...,99∼100 Hz)

fusion performs better than the single-modality feature alone in terms of the facial expression classification. This work gives rise to a new investigation on how to utilize EEG signal frequency to correlate the facial behavior and emotion, with an attempt to improve the emotion analysis.

II. Working Plan

Purpose

Affect computation on EEG based fusion. Explore the relation between brain EEG signal and facial images.

Goals

Improve the result by EEG and image fusion. Find an efficient fusion method.

Strategy

- Extract facial image by CNN, or VGG, or Resnet...
- Extract EEG feature from the raw signal through RNN/LSTM
- Design network for features fusion
- Validate network on FER, AU detection, emotion detection and so on

III. Future Plans and Challenges

First, I plan to expand the data size to a larger scale. Right now, it only contains 29 subjects. Second, EEG signals is very useful and the way extraction features could be explored more. Consider the power of deep feature, which is widely used nowadays, design a neural network on EEG raw data directly may surpass the handcraft feature we applied.Third, explore more modalities. 2D, 3D, audio, thermal, text, physiology signals and so on. Especially study in modality representation.

My goal is to improve the state-of-the-art in AU detection, FER, and emotion detection through the multimodal learning and fusion. Meanwhile, I will consider 3D face reconstruction as a powerful tool to achieve the goal. In addition, 3D domain exploration for better representation of dynamic facial surfaces will require the integration of knowledge from computer vision, graphics, and machine learning. I plan to design a robust and extendable algorithm to tackle the challenge, as well as develop a new approach with fusion of physiological signals in an attempt to achieve a more reliable affection recognition.

The challenges in multi-modality learning are in representation, fusion, transfer knowledge, and so on.

- How to find a way to represent each modality and combine them?
- How to fuse the same info from each modality and how the unique info could be helpful?
- When one modality has limited resources, how another modality can help?

Moreover, another challenge on multi-modalities learning and fusion is the overfitting problem, which is largely due to the issue caused by data unbalance (e.g. Action Units). Data augmentation could be a potential remedy, thus give rise to another task to address this challenge as one of my Ph.D. research directions.

TABLE II: Comparison result of single modal features and fusion features for posed expression recognition using DANN. For the specification of EEG Feature A, B and C, please refer Table I, Fusion Features A, B and C means the concatenation of facial expression with EEG Feature A, B and C. (ACC and STD means accuracy and standard deviation, respectively)

Evaluation method	Evaluation criteria	EEG Feature A	EEG Feature B	EEG Feature C	Facial expression image	Fusion Feature A	Fusion Feature B	Fusion Feature C
LOOCV	ACC	77.12%	80.51%	82.82%	88.15%	67.32%	85.96%	**95.02%**
	STD	0.1146	0.1484	0.1255	0.0757	0.1719	0.0909	**0.0546**
4 fold CV	ACC	61.85%	66.98%	69.68%	72.85%	74.28%	72.59%	**76.68%**
	STD	0.0689	0.0749	0.0819	0.0837	**0.0521**	0.0759	0.0769

TABLE III: Evaluation of binary classes (pain versus neutral) spontaneous emotion recognition of ours database by using DANN. For the specification of EEG Feature A, B and C, please refer Table I, Fusion Features A, B and C means the concatenation of facial expression with EEG Feature A, B and C. (ACC and STD means accuracy and standard deviation, respectively.)

Evaluation method	Evaluation criteria	EEG Feature A	EEG Feature B	EEG Feature C	Facial expression image	Fusion Feature A	Fusion feature B	Fusion feature C
LOOCV	ACC	94.24%	95.69%	97.15%	92.13%	93.38%	92.90%	**98.60%**
	STD	0.0959	0.1013	0.0860	0.0963	0.0890	0.1009	**0.0481**
4 fold CV	ACC	91.71%	**94.28%**	90.54%	86.90%	87.02%	92.44%	91.00%
	STD	0.0738	0.0443	**0.0424**	0.0984	0.1028	0.0663	0.0637

TABLE IV: The confusion matrices of fusion feature C based facial expression recognition using DANN in 4 fold cross-validation.

	Neutral	Sadness	Fear	Happy	Anger	Disgust	Surprise
Neutral	**84.51%**	5.94%	1.12%	0.88%	2.33%	0.00%	5.22%
Sadness	0.71%	**71.71%**	3.00%	1.12%	17.88%	1.12%	4.47%
Fear	0.42%	3.19%	**65.37%**	6.94%	4.53%	4.18%	15.37%
Happy	0.11%	2.56%	8.35%	**86.12%**	2.21%	0.00%	0.65%
Anger	0.99%	11.82%	2.31%	0.92%	**67.99%**	8.84%	7.13%
Disgust	0.58%	4.67%	4.90%	6.57%	15.74%	**64.42%**	3.11%
Surprise	0.36%	3.17%	9.43%	0.04%	2.21%	0.72%	**84.07%**

REFERENCES

[1] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch. EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, 118(3):480–494, 2007.

[2] I. Goncharova, D. McFarland, T. Vaughan, and J. Wolpaw. EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, 114(9):1580–1593, 2003.

[3] X. Li, X. Zhang, H. Yang, W. Duan, W. Dai, and L. Yin. An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis. In *the IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2020.

[4] G. Yaroslav et al. Domain-adversarial training of neural networks. *Advances in Computer Vision and Pattern Recognition*, 2017.