Multi-modal Emotion and Facial Actions Analysis

Xiaotian Li
Department of Computer Science, Binghamton University, NY, USA
xli210@binghamton.edu

***Abstract*— Automatic affect recognition is a challenging task due to the complexity and variety of emotion generation and exhibition. This investigation could range from multiple modalities, various representations, and different signs from faces, gestures, verbal and physiological signals. In the past decade, research on facial expression analysis has shifted its focus from posed behavior to spontaneous behavior. This shift has increased the difficulty of its analysis, as well its ecological validity and practical utility. A similar shift occurs from single modality to multi-modal analysis, resulted from the acquisition and integration of 2D and 3D videos, temperature dynamics, and physiological responses. As my Ph.D. research, I continue the efforts on data collection as well as continue the study of emotion modeling and analysis through the investigation on techniques, including domain adaptation, feature representation and fusion, multi-model learning, etc.**

I. Summary of Existing Work

The research will be carried out in 2 phases that dedicates to data acquisition and facial action analysis.

A. Data Acquisition

3D facial models have been extensively used for face recognition and animation. This motivates us to explore the usefulness of such data for 3D facial expression recognition. During the last 10 years, we have published 3 databases including BU-3DFE [9], BU-4DFE [8], BP4D+ [10], containing 3D facial expression models from a large number of subjects with different background. Last year, I had the honor to participate constructing the database which incorporates many different modalities like 3D facial models as well as their corresponding textures, 2D thermal images, 1D physiology signal (electrodermal activity, heart rate, respiration rate and etc.) and Kinect depth images. Fig. 2 illustrates sample data sequences of four modalities from a subject. In addition, the metadata are also generated, including manually labeled action units (both occurrence and intensity) on four tasks, automatically tracked head poses, and 3D/2D/IR facial landmarks. Detailed annotations.To my best knowledge, this is the first database to include such massive volume of subjects and data sources which is the main topic of the affection analysis community.

After that, we collected an EEG-based multi-Modal emotion database[3], with both posed and authentic facial actions for emotion analysis: emotion is an experience associated with a particular pattern of physiological activity along with different physiological, behavioral and cognitive changes. One behavioral change is facial expression, which has been studied extensively over the past few decades. Facial behavior varies with a person's emotion according to differences in terms of culture, personality, age, context, and environment. In recent years, physiological activities have been used to study emotional responses. A typical signal is the electroencephalogram (EEG), which measures brain activity. Most of existing EEG-based emotion analysis has overlooked the role of facial expression changes. There exits little research on the relationship between facial behavior and brain signals due to the lack of dataset measuring both EEG and facial action signals simultaneously. To address this problem, we propose to develop a new database by collecting facial expressions, action units, and EEGs simultaneously. We recorded the EEGs and face videos of both posed facial actions and spontaneous expressions from 29 participants with different ages, genders, ethnic backgrounds. Differing from existing approaches, we designed a protocol to capture the EEG signals by evoking participants' individual action units explicitly.There are three sessions in the experiment for simultaneous collection of EEG signals and facial action videos, including posed expressions, action units, and spontaneous emotions, respectively. A total of 2,320 experiment trails were recorded, which is a considerably sized database for research. The whole procedure of experiment is shown in the Fig. 1. We also investigated the relation between the EEG signals and facial action units. 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. The database will be released to the research community to advance the state of the art for automatic emotion recognition.

As a member of our research team, I participate in developing multi-modal emotion database and conducting data collection. The BP4D and BP4D+ is a large-scale emotion database with more than hundred subjects including. We have diverse and complex sensor modalities including high-definition 3D geometric facial sequence, 2D facial videos, thermal videos, physiological data sequence and their corresponding metadata like face landmarks and head pose tracking points. Currently, we're preparing to release an extended version of this project. The data capture system is shown in Fig. 4. The physiological systems capture vital sign signals in a very high sample rates, including blood pressure, respiration rate, heart rate and electrodermal activity (EDA). Note that the system synchronization is critical for data collection from various modality sensors. Due to each sensor has its own machine to control, we developed a program to trigger the recording from the start to the end across all three sensors simultaneously. This is realized through the control of a master machine by sending a trigger signal to three

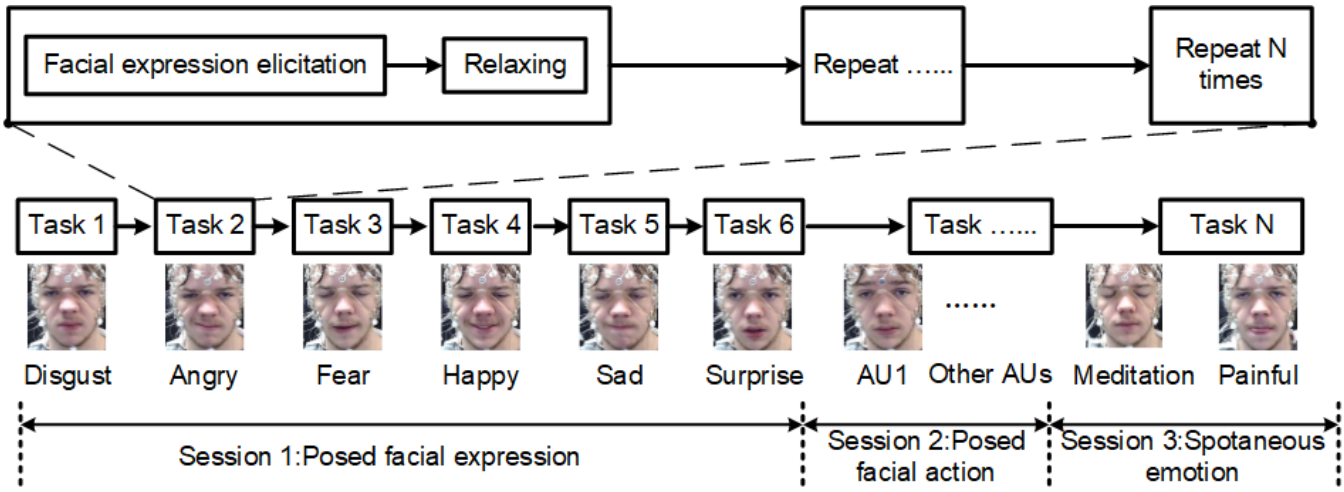


Fig. 1: Protocol of the data acquisition (Procedure of experiment).

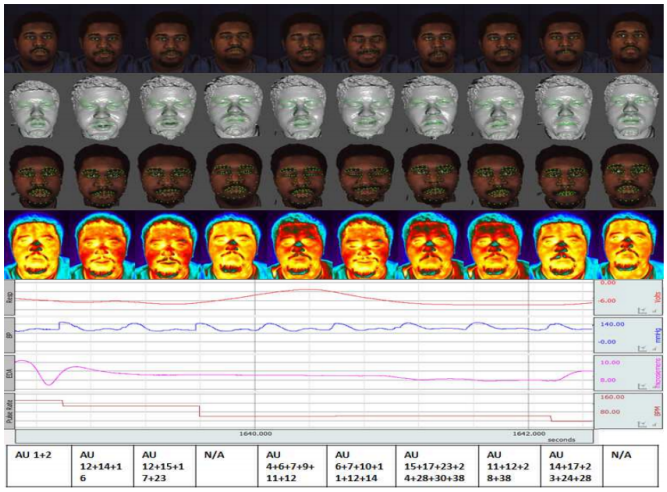


Fig. 2: Sample data sequences from a participant including original 2D texture (first row), shaded model (second row), textured model (third row), thermal image (fourth row), physiology signal(fifth row: respiration rate, blood pressure, EDA, heart rate) and corresponding action units(last row).

sensors concurrently.

B. *Facial Action Analysis*

Automatic analysis of facial action [4] is crucial in many areas: mental and physical health, education, and human-computer interaction among others. The most comprehensive method to annotate facial expression is the anatomically based Facial Action Coding System (FACS) [1]. Facial action unit recognition is still a challenging task for facial action analysis. Existing AU-labeled spontaneous facial expression datasets are either in a small-scale due to labor-intensive annotations, or lack of sufficient variety in terms of amount, ethical background, age ranges, and facial appearance variations of subjects, thus limiting the learning effectiveness. To mitigate the issue of high redundancy and low level of variants existing among image frames of facial video sequences with respect to both subject identities (ID) and facial action units (AU), [11] propose a novel learning process with convolutional neural networks (CNNs), named Adversarial Training Framework (ATF) [6]. [5] also propose a weakly supervised AU recognition method from expression-annotated facial images and domain knowledge through adversarial training. First, he summarizes a large amount of domain knowledge about AU relationships and sample pseudo AU data based on the summarized domain knowledge. After that, he proposes a novel adversarial network for AU recognition, with the goal of making the distribution of AU classifiers' output converge to the distribution of the pseudo AU data generated from domain knowledge. Specifically, the proposed AU recognition adversary network consists of two models: a recognition model R, which learns AU classifiers, and a discriminative model D, which estimates the probability that AU labels generated from domain knowledge rather than the recognized AU labels from R. These two models are trained simultaneously through an adversarial process. The training procedure for R is to maximize the probability of D making a mistake, while the training procedure for D clearly distinguishes the pseudo AU data generated with domain knowledge from the predicted AU labels of the recognition model. By leveraging this adversarial mechanism, the distribution of recognized AUs is closed to AU prior distribution from domain knowledge after training. Furthermore, he extends the proposed weakly supervised AU recognition to semi supervised learning scenarios when partially AU-annotated images are available by adding a cross-entropy term for the AU-annotated images. My current work also adopt adversarial learning to do multi task. We are leveraging multi-task adversarial learning to reduce the each subject's ID, and adapt from a source data distribution to a target data distribution(one modality to another).

Domain adaptation is another solution for learning a

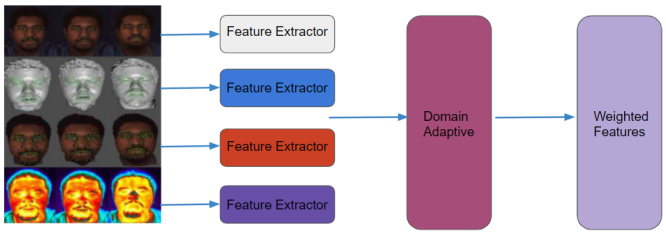
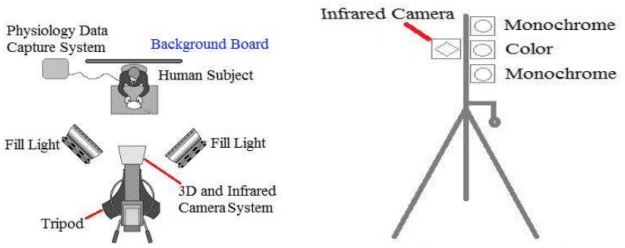


Fig. 3: Schematic diagram of the domain adaptation module



(a) System aerial view (b) Relative position of cameras

Fig. 4: Recording system

task from multi-modalities. Currently, Our team is doing some research about domain adaptation for the action unit recognition task. Unlike the domain-adversarial neural network, which can generate domain-invariant data features that are discriminative for the classification task whereas indiscriminate for the shift between the source and target domains. We propose an universal domain adaptation module which doesn't require any prior knowledge of domain labels. This model aims to automatically recognize and weigh the different contribution degrees of signals from different domains or modalities. We hope the weighted deep features can improve the performance of our current models. For the schematic diagram of the model, see Fig.3

Furthermore, data imbalance is a general issue in recognition task, especially in multi-label training. The most popular method is to use adaptive loss function to control the training process. But I'm trying to use generative adversarial networks (GAN) [2] train a model which can take the continuous face attributes vectors and face image as the input to output the corresponding face image. By generating designated data, we can synthesize an enhanced extension data to solve imbalanced annotation issue. We can even create more data with generated head position and face ID which can enhance the lab data. Fig. 5 shows the synthetic faces with specific action unit using samples from BP4D. In order to reduce the negative impact of the generated data on the original data, I added a domain attention [7] module to adapt the generated data in my deep learning model. So far, my method has made some performance improvement in automatic facial action recognition task.

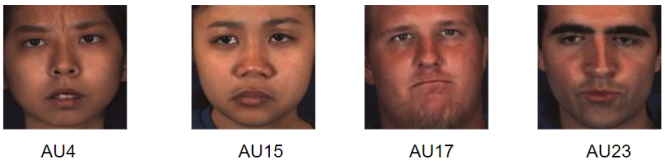


Fig. 5: Synthetic data with single AU label

II. Working Plan

Our future work will expand the size to a larger-scale for both our EEG based multi-modal emotion dataset, and will conduct EEG based AU detection and EEG-expression based fusion for AU detection. We also would like to discover how to build a model by using constructed AU relationship knowledge-graph as the extra guidance information to make our model robust. Other than that, we also pay attention to dynamic analysis of facial action, considering spatial-temporal information can take performance improvement to our current work. We are considering both hand-crafted feature and long short-term memory(LSTM) to address the issue.

III. Future Plans and Challenges

Although the main focus in machine analysis of facial action has shifted to the analysis of spontaneous expressions, state-of-the-art methods can not be used in fully unconstrained environmental conditions effectively. Challenges preventing this include handling occlusions, non-frontal head poses,rigid and non-rigid facial motions, varying illumination conditions and lack of data. The good news is that with the fast development of deep learning, some of the challenges are no longer an insurmountable mountain. In the next few years, we will avoid areas which are too crowded and put more efforts on the spaces that few researchers stare at. We expect to achieve automatic generating more realistic multi-modal data with limited laboratory data; we expect to achieve automatic annotation for large scale database of human behavior analysis; we expect to interpret or translate affection information from one domain to another for making machine intelligent to recognize human emotion; we expect to explore more novel modalities which can be used for improving emotion recognition. All of these areas are with huge potential to be excavated.

References

[1] P. Ekman and W. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 1971.
[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al. Generative adversarial networks, 2014.
[3] X. Li, X. Zhang, H. Yang, W. Duan, W. Dai, and L. Yin. An eeg-based multi-modal emotion database with both posed andauthentic facial actions for emotion analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020.
[4] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 10(3):325–347, jul 2019.
[5] G. Peng and S. Wang. Weakly supervised facial action unit recognition through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[6] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses, 2017.

[7] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos. Towards universal object detection by domain attention. In *IEEEConference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[8] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[9] X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. Girard. BP4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[10] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] Z. Zhang, S. Zhai, and L. Yin. Identity-based adversarial training of deep cnns for facial action unit recognition. In *BMVC*, 2018.