Automated Facial Action Unit Recognition using Local Relationships and Multimodal Data

Nagashri N Lakshminarayana
Department of Computer Science and Engineering,
University at Buffalo,
New York , USA

I. INTRODUCTION

The study of non verbal cues is an important step towards human behavioural understanding. Facial expressions are key indicators of the internal state of mind. Computers that can understand and interpret human emotions effectively can engage in a more meaningful human-machine dialogue. In my thesis, I work on recognizing expressions at the granularity of facial muscle activations called Action Units (AU). A common way of quantifying expressions is through the basic universal emotions like happy, sad, surprise etc. The Facial action coding system (FACS) further deconstructs the macro emotions into visually discernible movements on face sans the subjective bias linked with emotion labels.

Although FACS are used extensively by human annotators to make higher level decisions, the process of automating it is impeded by several challenges. The key challenges for automatic AU recognition are listed below:

- The AUs occur due to the contractions and expansions of specific facial muscle groups. While some of the muscle activities cause geometric distortions, others cause skin texture changes. This makes AU activation very subtle and local.
- The AUs act as building blocks for macro expressions. Thus several AUs co-occur forming an expression. The correlations/co-occurrences of AUs are complex and hard to define making AU recognition a challenging multi-label, multi-class classification problem.
- Some of the AU combinations can be additive. The manifestations of some AUs both in terms of appearance and the location can be altered by the other co-occurring AUs.
- The appearance based changes can be directly identified in the visible domain. But recognizing AUs using images is severely limited by illumination changes, head pose variations, occlusions and other external factors.

In my thesis, I aim to develop a unified framework that can address the above mentioned challenges in a principled manner. Attention learning is an interesting field of machine learning that draws inspiration from natural vision. Rather than processing the image as a whole, the spatial attentions selectively emphasize the regions of image that are salient to the given task. In fields like object recognition, activity recognition, spatio-temporal attentions have improved the performance of pre-existing deep learning models. Not all regions of face contain equal amount of information for recognizing AUs. Certain regions around the facial landmarks like eyebrows, lips etc contain more predominant cues than the others. However the attentions of the individual localized regions are not independent owing to the complex relationships between the various AUs that effect each other nature. Therefore In my thesis the attention and the AU relations are modelled jointly. While the AU relationships are learnt heuristically, attention learning can be used to model the prevalent latent relationships at the level of regions of interest automatically. An end-to-end framework combining attention, AU relation learning and classification can lead to more robust representations of AUs and thus, a better understanding of human facial expressions. Although Facial Action Units Recognition is well explored in the visible light domain (VLD), the RGB images suffer from illumination changes and can only capture the visual changes that occur as an effect of the AUs. There are however some physiological changes that the face undergoes during the occurrence of the Action Units, such as skin temperature changes, variation in the heart rate , blood pressure, and respiration rate that can't be captured using the visible images. Extending the unimodal recognition to include multimodal data requires careful modelling of the relationships between the modalities such that the final representation is a meaningful aggregation of the individual complementary sources

II. RELATED WORK

The general architecture of AU recognition consists of image pre-processing, feature extraction and classification stages. The input is a cropped facial image frontalised to register face to a common template. Following the pre-processing stage is the feature extraction and classification stages that are jointly modelled using deep learning frameworks. The traditional convolutional nueral networks (CNNs) parse the entire image/feature space using same filters. Based on the locational importance of regions of face, several methods [4], [5] divided face into uniform regions and learnt separate filters for each region. In recent methods [1] [2], rather than crudely dividing facial regions into grids, the AU centers were defined based on the expert knowledge around facial keypoints. One of the key aspects of the action units are the inter-relationships between the different AUs. The above methods do not explicitly model the AU relations. In [9] the relationships were learnt by

learning individual patch prediction models and modeling AU correlations after the feature extraction using Conditional Random Fields (CRF). In [6] shao et al. modelled pixel level relationships using CRFs. In the aforementioned methods attention and relations are learnt jointly as the posteriors of the feature extraction stage. Due to the practical limitations of compiling spontaneous emotion databases and annotating them, only a handful of datasets for AU precognition exist. We perform experiments on BP4D, DISFA and MMSE dataset. Both BP4D and DISFA are spontaneous datasets with – and 27 subjects respectively. Since the action unit occurrences can't be controlled in a spontaneous setting, the AUs used for classification is different each of the datasets. MMSE is the only dataset with multimodal data for AU recognition. It has 140 subjects. In addition to subject videos, their physiological response like Blood pressure, pulse rate and respiration rate is also provided by the dataset.

III. AU Guided Attention

Modelling AU relations is a complex task. Capturing the AU correlations at the stage of AU predictions have lead to significant improvement in the performances as demonstrated by the previous methods. However most of the methods disregard the additive nature of AUs. For example when AU4 (brow lowerer) occurs independently, brows are drawn together and lowered. When occurring with AU 1 (inner brow rise), brows are drawn together but raised. In order to account for such AU interactions, we need to imbibe the AU co-dependency into the feature construction itself.

In our most recent work, we introduce the notion of AU relationships early on into the network using AU guided attention maps. We designed an attention map that modelled the AU correlations from the ground truth data. Through our attentions we modelled both the AU region saliency and relationships while still preserving the feature sharing property of the convolutional kernels. The later layers construct features from the AU guided features therefore instilling a notion of AUs throughout the feature extraction process.

We jointly model the feature extraction and classification using the densenet-121 architecture. The backbone network has four blocks of feature extraction. At the second stage we introduce attention maps that are superimposed with the network activations to form a more richer and concentrated representation. Rather than hard coding the attention maps we learn the maps from scratch. Ideally the attention maps give more importance to the regions that correspond to the AUs that are present. The centers corresponding to each of the AUs are defined by [1] and we follow the same convention. We select the AU centers corresponding to those AUs that are present in the current image. The desired AU attention map with respect to the ground truth is shown in figure 1.

To learn such attention maps implicitly in an end-to-end fashion requires substantial data. Since most of the expression datasets are limited owing to the practical considerations, we learn the AU guided attention mask using the groundtruth masks as shown in the Figure 1. We construct a mask network that uses the activations of the densenet block-2 layers as the input. Since the block-2 features are superimposed with the learned attentions and further propagated into the proceeding layers, the features benefit from the dual tasks. Firstly the gradients from the mask learning network induce the AU saliency knowledge into the features, secondly the gradients from the classification network help steer the features towards robust discriminative representation.

TABLE I
Comparison of Guided Attention model with related State of the art methods.

Method	Params	BP4D	DISFA	MMSE
EAC	260M	55.9	48.5	-
DSIN	20M	61.7	53.6	-
GA-net	9M	60.0	54.33	59.39

We perform experiments on three standard datasets BP4D, MMSE and DISFA that have 41, 27 and 140 subjects respectively. Each frame of the video is annotated with FACS. Currently my thesis is focused on detecting AUs at the frame level. Using the proposed method, we obtain competitive performance on all three datasets as shown in Figure III. Unlike other methods we don't change the topology of the existing network which causes a substantial increase in parameters to the original backbone network. In DSIN, a separate CNN is learnt for every patch thereby causing a significant parameterization. In contrast, our method only adds few parameters to the current backbone making it efficient both in terms of performance and size. Also, our attention network is modular and can be plugged into any deep learning based backbone.

IV. Multimodal Learning

The use of physiological signals in emotion recognition has been widely studied in the affective research community. The survey paper Shu et al [3] gives an extensive compilation of various such methods that have adopted physiological signals for emotion recognition. While the visual inputs efficiently reveal muscle movements on the face, bio-signals have shown to be useful in predicting the valence and arousal of expressions. We could say that physiological signals have an auxiliary view to the appearance based changes. The multi-channel information could help constructing more robust and holistic representations for AU recognition. In our prior work [10] we explored the combination of thermal images with visible images to encode complementary aspects of AU recognition. The individual modalities were enhanced in their own subspace such that they are better aligned to be combined with each other. The feature enhancement for a subspace was performed taking into consideration the context of the counterpart.

More recently in [8], we also conducted experiments with physiological signals as metadata to test the hypothesis that bio-signals can help improve the performance of unimodal representations for AU recognition. The physiological signals

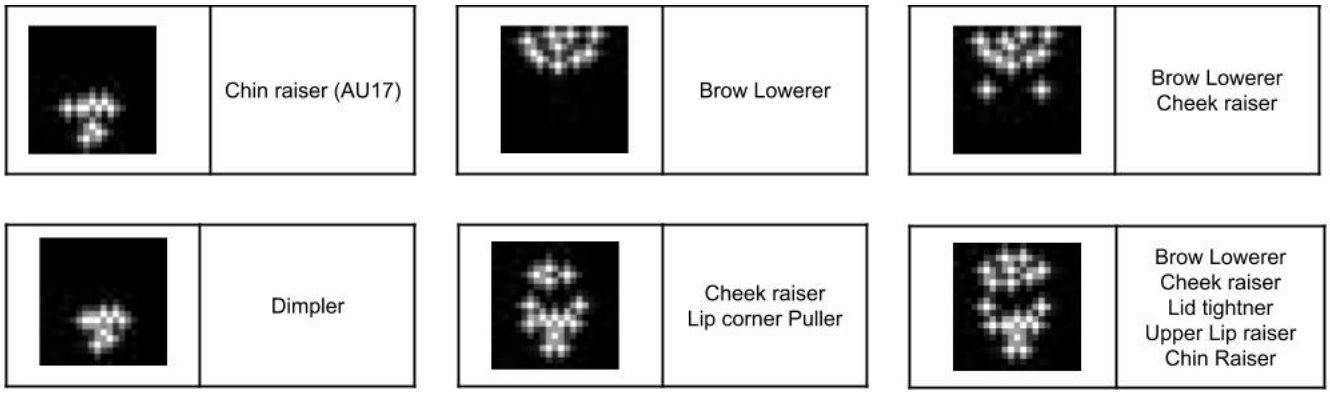

Fig. 1. The ground-truth masks generated for mask learning. The attention of different regions of face is conditioned on the presence of the AUs in that region

alone may not carry any explicit information regarding the occurrence of the action units, but may give certain important insights into the data. We use two networks, the visible feature extractor and the metadata network to extract the corresponding features and perform a feature level concatenation. The metadata aggregation network takes as input the concatenated features and learns a combined subspace for action unit recognition. We conducted experiments on MMSE dataset that has signals like GSR (galvanic skin response), BP (blood pressure) and respiration rate. The multimodal learning improved the performance of the unimodal network from 56.8% to 58.1% confirming our hypothesis.

V. Future work

AU relations form the core of expression recognition. Yet, the current state of the art methods model these relationships superficially using the AU correlations that are inferred from the dataset. Since existing expression datasets are skewed and capture only a subset of AU relations, modelling AU relationships based on the co-occurrences observed in the dataset is not comprehensive. Besides the AU co-occurrences are just one of many kinds of relationships between the muscle activations on the face. Since the facial muscular morphology control the facial expressions, it is more meaningful to model the relationships at the level of regions of muscle activations rather than AU predictions. Therefore, the subsequent direction of my thesis is to model the underlying structure of AU relationships at the level of the muscle regions or ROIs of expression. A comprehensive relationship formulation that takes into the account the muscular basis of expressions could lead to a more contextual representation for facial action unit recognition. My next steps would be to explore graph based models to automatically infer the latent optimal structure to model both global and local ROI relationships in a unified framework. While the previous methods require defining explicit connections between the regions, my approach would be to learn those relationships using spatial attentions. A framework for joint attention and relation learning thereby eliminates the efforts to define specific dataset based AU relations and can be generalized to other dataset with minimal human supervision.

The multimodal data that have disparate views of the data could eliminate the shortcomings of the other modalities by augmenting the representation with multiple views. In preliminary experiments, we explored the use of multimodal data in facial action unit recognition. We used image as a primary modality and all the bio-signals as the secondary modalities. Since each of the bio-signals encode different aspects of emotions and have diverse signal profiles, the role of the signals with respect to capturing the residuals of the images needs to be accounted for independently. In my thesis I aim to factorize the individual relationships of the bio-signals to the AU occurrence when used with image features. Further, a multimodal attention mechanism can be defined to selectively choose the most important multimodal interactions. I plan to extend the current multimodal framework to include contextual attention. In this framework the context refers to the information encoded by the other modalities.

References

[1] Li, Wei and Abtahi, Farnaz and Zhu, Zhigang and Yin, Lijun, Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection, *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*,2017, pp 103–110.

[2] Shao, Zhiwen and Liu, Zhilei and Cai, Jianfei and Ma, Lizhuang, Lijun, Deep adaptive attention for joint facial action unit detection and face alignment, *2Proceedings of the European Conference on Computer Vision (ECCV)*,2018, pp 705–720.

[3] Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X. (2018). A Review of Emotion Recognition Using Physiological Signals. Sensors (Basel, Switzerland), 18(7), 2074. https://doi.org/10.3390/s18072074

[4] Zhong, Lin and Liu, Qingshan and Yang, Peng and Huang, Junzhou and Metaxas, Dimitris N, earning multiscale active facial patches for expression analysis, *IEEE transactions on cybernetics.*, vol. 45, 2015, pp 1499-1510.

[5] Liu, Ping and Zhou, Joey Tianyi and Tsang, Ivor Wai-Hung and Meng, Zibo and Han, Shizhong and Tong, Yan, Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis, *European Conference on Computer Vision*,2014, pp 151-166.

[6] Z. Shao and Z. Liu and J. Cai and Y. Wu and L. Ma, Facial Action Unit Detection Using Attention and Relation Learning, Transactions on Affective Computing,2019.

[7] N. N. Lakshminarayana and N. Sankaran and S. Setlur and V. Govindaraju, Multimodal Deep Feature Aggregation for Facial Action Unit Recognition using Visible Images and Physiological Signals, *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*,2019, pp 1-4.

[8] N. N. Lakshminarayana and N. Sankaran and S. Setlur and V. Govindaraju, Multimodal Deep Feature Aggregation for Facial Action Unit Recognition using Visible Images and Physiological Signals, *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*,2019, pp 1-4.

[9] Corneanu, Ciprian and Madadi, Meysam and Escalera, Sergio, Deep Structure Inference Network for Facial Action Unit Recognition, *"Computer Vision – ECCV 2018*, pp 309-324.

[10] Lakshminarayana, Nagashri N and Mohan, Deen Dayal and Sankaran, Nishant and Setlur, Srirangaraj and Govindaraju, Venu, Multi-modal Conditional Feature Enhancement for Facial Action Unit Recognition, *Domain Adaptation for Visual Understanding*, 2020 pp 95–109.