Local Relationship Learning with Person-specific Shape Regularization for Facial Action Unit Detection

Xuesong Niu[1,3], Hu Han[1,2], Songfan Yang[5,6], Yan Huang[6], Shiguang Shan[1,2,3,4]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
[5] College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, China
[6] TAL Education Group, Beijing, China
xuesong@vipl.ict.ac.cn, {hanhu, sgshan}@ict.ac.cn, {yangsongfan, galehuang}@100tal.com

Encoding individual facial expressions via action units (AUs) coded by the Facial Action Coding System (FACS) has been found to be an effective approach in resolving the ambiguity issue among different expressions. While a number of methods have been proposed for AU detection, robust AU detection in the wild remains a challenging problem because of the diverse baseline AU intensities across individual subjects, and the weakness of appearance signal of AUs. To resolve these issues, in this work, we propose a novel AU detection method by utilizing local information and the relationship of individual local face regions. Through such a local relationship learning, we expect to utilize rich local information to improve the AU detection robustness against the potential perceptual inconsistency of individual local regions. In addition, considering the diversity in the baseline AU intensities of individual subjects, we further regularize local relationship learning via person-specific face shape information, i.e., reducing the influence of person-specific shape information, and obtaining more AU discriminative features. The proposed approach outperforms the state-of-the-art methods on two widely used AU detection datasets in the public domain (BP4D and DISFA).

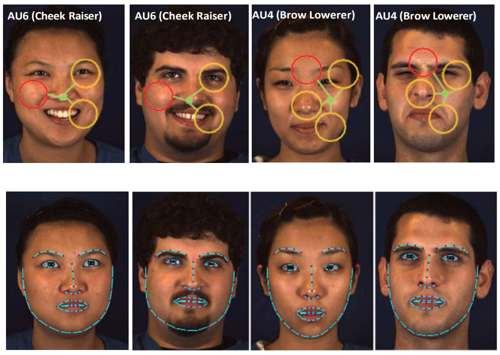

Figure 1: Each single local facial region defined for AUs in FACS (red circles) can be ambiguous because of face variations in pose, illumination, etc.; therefore, taking into account the relationship of multiple related face regions (yellow circles) can provide more robustness than using individual single local regions separately. At the same time, person-specific face shape information also influences the AU detection performance, i.e., detection of AU4 (Brow Lowerer) is highly influenced by the eye-eyebrow distance, which may vary significantly among different subjects. Therefore, we expect to reduce the influence of such person-specific shape information to the AU detection task, i.e., through regularization during feature learning.

1. Introduction

Facial expression is a natural and powerful means for human communications, which is highly associated with human's intention, attitude or mental state. Therefore, facial expression analysis has wide potential applications in diagnosing mental health [32], improving e-learning experiences [30], and detecting deception [12]. However, direct facial expression recognition in the wild can be challenging because of ambiguities between several expressions. One of the effective methods in resolving the ambiguity issue is to represent individual expression using the Facial Action Coding System (FACS) [10], in which each expression is identified as a specific configuration of multiple basic facial AUs. Therefore, a robust facial AU detection system is important for the accurate analysis of facial expressions.

Since different AUs correspond to different muscular activations of the face, the appearance of multiple local regions jointly reflects the presence of individual AUs, and the local information is crucial for AU detection. The early works on facial AUs detection represented different local facial areas using the traditional hand-crafted features, which can be not discriminative enough for capturing the facial morphology [46, 48]. Recently, deep learning has been widely applied for facial representation learning, including using the deep representation for more effective AU detection [5, 8, 24, 35, 47].

However, besides learning more AU discriminative features, the relationship of individual facial regions can be very important for AU detection. As shown in Fig. 1, each single local face region defined in FACS can be ambiguous for AU detection because of face variations in pose, illumination, etc.; therefore, taking into account the relationship of multiple face regions can provide more robustness than using a single local region. For instance, the cheek area and the mouth corner of the face usually active simultaneously in a common facial behavior called Duchenne smile, resulting in high correlations between AU6 (cheek raiser) and AU12 (lip corner puller). Some approaches tried to utilize such local relationship information by using multi-label learning [24, 46, 47], but only holistic feature representations were used. A meticulous modeling method is required for effectively leveraging the relationship of different local facial regions to perform robust AU detection.

Another critical characteristic of AU is that the appearance of the same AU may vary among different subjects due to the different morphological aspects and ways to express the emotions of different subjects (see Fig. 1). This is the reason why designing a person-specific AU detector can improve the AU detection accuracy. However, existing person-specific AU detection methods require either retraining the model for the new subjects [7, 43], or additional data of the new subjects for model generation [1, 33] or normalization [2]. These constraints limit the range of applications of the existing AU detection methods.

In this paper, we propose an end-to-end trainable network for AU detection using Local relationship learning with Person-specific shape regularization (namely LP-Net). The LP-Net consists of a stem network, a local relationship learning module (L-Net) and a person-specific shape regularization module (P-Net). The stem network mainly contains convolutional layers for local region feature extraction. The extracted local features are then fed to the local relationship learning module for relationship learning and predicting the AU occurrence probabilities. At the same time, P-Net aims to learn features that are independent with the features by L-Net, and thus works as a regularization term to reduces the influence of person-specific shape information. As a result, the final features learned by L-Net are more discriminative and generalizable for AU detection.

The contributions of this work are three-fold: (i) we propose a novel end-to-end trainable framework for AU detection, which is able to leverage not only the local information but also the relationship of individual regions to improve the AU detection robustness; (ii) we regularize local relationship learning via person-specific face shape information to obtain more discriminative and generalizable features related to AU detection; (iii) the proposed approach outperforms the state-of-the-art methods on two widely used AU detection datasets BP4D and DISFA.

2. Related Work

Automatic facial action unit detection has been studied for decades, and several works have been proposed. Various features [4, 20, 25, 26] and classifiers [7, 38, 44, 46] have been applied to build a robust facial action unit detection system under realistic situations. Recently, CNNs have shown great power in many computer vision tasks such as face verification [37], objection detection [13], and image recognition [17], and have been successfully applied to automatic facial action unit detection [5, 15]. The reader can refer the recent surveys and challenges [9, 27, 39] for more information. In the following paragraphs, we will review the relative works to ours.

Since facial AUs are defined as patterns of different facial muscular movements, the ways they perform the facial expressions are relatively based on the local facial appearance. Several works are based on this character and use local information for facial AU detection. Zhong *et al.* [48] divided the face area into multiple uniform patches, and use the common and specific patches to describe different expressions. Taheri *et al.* [36] defined fixed regions for different AUs and used sparse coding to recover facial expressions using the composition rules of AUs. Zhao *et al.* [46] performed a patch selection method based on facial landmarks and group sparsity learning. All these methods used traditional features to represent the face local information and these features are not sufficiently expressive.

Besides of traditional features, the great modeling power of CNNs has also been successfully leveraged to facial action unit detection. In [47], Zhao *et al.* proposed a region layer to induce the CNN to focus on important facial regions for better feature learning. In [23], Li *et al.* trained different CNNs using different parts of a face and merged the features from different areas in an early fusion fashion using fully connected layers. In [24], Li *et al.* proposed a local feature learning method based on enhanced and cropped facial area. In [35], Shao *et al.* proposed an end-to-end deep learning framework for joint AU detection and face alignment, which used the alignment feature to compute

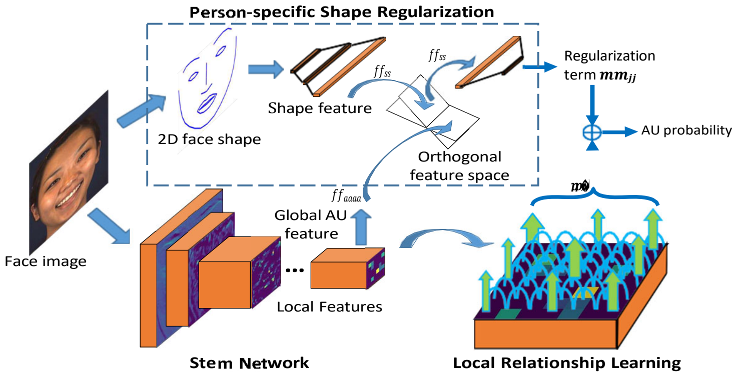

Figure 2: An overview of our approach for AU detection, which consists of a stem network, a local relationship learning module (L-Net) and a person-specific shape regularization module (P-Net). By using P-Net to model the person-specific shape information, and enforcing the person-specific shape features are independent with the features learned by L-Net, we expect the final features for AU detection can be more discriminative and generalizable.

an adaptive attention map for better local feature learning. These methods have drastically improved the performance of facial AU detection with the great modeling power of deep learning. However, all these methods only focused on different regions and failed to consider the relationship of different local areas. At the same time, the appearance of different facial local areas usually changes simultaneously because of the underlying facial anatomy, and this relationship of different local regions will also benefit the detection of AUs.

Besides directly using local features to predict AUs, another way of modeling the relationship of AUs is to use the correlations of different AUs. Walecki *et al.* [40] proposed a method to model the AU relations and feature representations simultaneously by combining conditional random field (CRF) with deep learning. In [41], Wang *et al.* proposed a restricted Boltzmann machine to capture high-order AU interactions. In [8], Corneanu *et al.* applied a graphical model inference approach to passing AU probabilities between different AU labels. All these methods took the probabilistic dependencies between different AUs into consideration and used the correlations to refine the predicted results. However, most of these methods computed the AU probabilities using the feature generated from the entire face area. Local information has been ignored, which can be very important for facial AU detection. At the same time, AU-to-AU relationships are mainly generated using facial anatomy and FACS [10] based on posed expressions, and their generalization ability to spontaneous expressions is not known.

Another key characteristic of AUs is that the appearance of the same AU may vary among different subjects. This is the reason why many person-specific AU analysis models have been proposed, and have been found to be effective for AU detection. Chu *et al.* [7] proposed a selective transfer machine to personalize the AU detector by re-weighting of the source distribution to match that of the target distribution. Zeng *et al.* [43] applied a similar re-weighting strategy and learned a person-specific classifier using synthetic labels provided by confident classifiers. This kind of person-specific AU detectors requires re-training the model for each subject, which can be time-consuming. Besides re-weighting the source distribution, Sangineto *et al.* [33] proposed a transfer process to learn discriminative mappings between the data distribution associated with each source subject and the corresponding parameters. Almaev *et al.* [1] proposed a multi-task learning structure to learn the latent relations among tasks using one single AU and transfer the latent relations to other AUs. In [2], Baltrusˇaitis *et al.* proposed a simple but efficient way for person-specific feature normalization using the median of all the features in a video. Although all these methods do not need to re-train the model for a new subject, they still need additional data to generate a new AU predictor, which limits the application scope in practical scenarios.

In contrast to these existing methods, we employ an end-to-end deep framework LP-Net to predict AUs. We not only consider the local information for facial AUs prediction but also take the relationship of different facial regions into consideration. At the same time, person-specific shape regularization is also utilized to reduce the influence of the diverse baseline AU intensities among different subjects.

3. Proposed Method

Fig. 2 shows the overall framework of our LP-Net for facial AU detection, which consists of a stem network, a local relationship learning module (L-Net) and a person-specific shape regularization module (P-Net). We detail the proposed approach in the following sections.

3.1. Overview of LP-Net

Feature representation is the key component of building a robust AU detection system, in which CNN has shown its great power and achieved great success in many computer vision tasks [13, 17, 37]. Traditional CNNs usually feed the output of convolutional layers to a global pooling layer in order to get a robust global feature. However, such an operation would fail to capture the local information for structured objects like faces and thus ignoring some local but important information related to AU detection.

To overcome these limitations, as shown in Fig. 2, we remove the global pooling layer in CNN and directly use the output feature maps from the convolutional layers as the representation of local features. CNN networks like ResNet [17] have been proved to have a strong ability for local features generation with only convolutional layers. So, here, we choose ResNet-34 [17] as our stem network for local feature learning. The output of the last convolutional layer of ResNet-34, which contains 512 feature maps with a size of 7 × 7, is regarded as the set of local features and utilized for further processing. Thus, we in total obtain 49 local features of 512-dimension from the stem network.

After we get the local features generated from the stem network, a local relationship learning module based on Long Short-Term Memory (LSTM) [18] (L-Net) is introduced to automatically explore the underline relationship of individual local facial regions in the feature space. Our L-Net jointly considers the features of local regions and their relationship and outputs the probabilities of individual AUs. As summarized in Section 1, another challenge is that different subjects may have different baseline AU intensities because of the face shape differences. A person-specific shape regularization module (P-Net) is used to model such person-specific information based on 2D face shape. The features encoded by P-Net are expected to be independent with the features encoded by L-Net, and further used to calculate the regularization term to refine the AU probabilities predicted by L-Net. Thus P-Net works as a regularization module to enforce the L-Net to learn more subject-independent features for AU detection, and the refined AU probabilities by P-Net are used as the final prediction of our LP-Net.

3.2. Local Relationship Learning via L-Net

Fig. 3 gives the detailed structure of our L-Net for local relationship learning. Since the feature maps generated by the stem network are from the last convolutional layer of ResNet-34, each element (1 × 1 × 512) in the feature map highlights the characteristic of a facial region. Therefore, we use each element on the feature maps as a representation of the local face area and use it to perform local relationship learning.

Specifically, we get k local features $f_1, f_2, \cdots, f_k$ from the stem network ($k = 49$ for ResNet-34). Each local feature f_i will be used for AU prediction, and output an AU occurrence probability. The LSTM structure is utilized to learn the relationship and outputs the probabilities of different local features. Since different AUs have different muscular activations, and the contributions of individual local features for predicting the probability should be different. Therefore, we predict the occurrence probability of each AU separately, i.e., using C LSTM structures to predict the probabilities of all the C AUs.

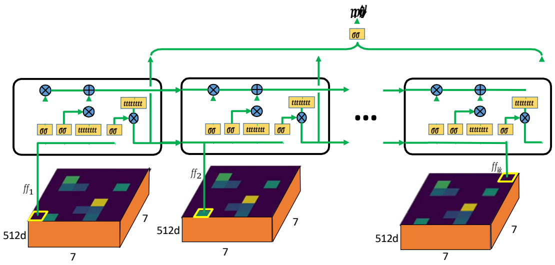

Figure 3: Each element (49 elements in total) of the feature map generated by the stem network is treated as a representation of a local region and used as the input of our L-Net based on LSTM. L-Net explores the underline relationship of individual local regions, and outputs the probabilities.

At the same time, we believe that every local feature can be helpful for detecting individual AUs, and thus all the k local features are fed to each LSTM structure. The final decision for the detection of each AU is obtained by combining all prediction results and the final predicted AU occurrence probabilities by L-Net can be written as

$$\hat{p}_j^l = \sigma\left(\frac{1}{k}\sum_{i=1}^{k} LSTM_j(f_i)\right) \qquad j = 1, 2, \cdots, C \tag{1}$$

where σ is a sigmoid function.

3.3. Person-specific Shape Regularization via P-Net

P-Net aims to reduce the influence of person-specific shape information and obtaining more discriminative and general features for AU detection. As shown in Fig. 4, we use 2D facial landmarks as a representation of the face shape [14, 21]. Specifically, we use a robust facial landmarks detector (Convolutional Experts Constrained Local Model [3, 42]) to detect 68 facial landmarks $P_1, P_2, \cdots, P_{68}$, and then all the face images based on the two eye centers to reduce the influence of head pose. After the face images are aligned, each landmark point is normalized using

$$P_{norm} = \frac{P - P_{center}}{d} \tag{2}$$