AU R-CNN: Encoding Expert Prior Knowledge into R-CNN for Action Unit Detection

Chen Ma[a], Li Chen[a,*], Junhai Yong[a]

[a]*School of Software Department, Tsinghua University, Beijing, 100084, PR China*

Abstract

Modeling action units (AUs) on human faces is challenging because various AUs cause subtle facial appearance changes over various regions at different scales. Current works have attempted to recognize AUs by emphasizing important regions. However, the incorporation of prior knowledge into region definition remains under-exploited, and current AU detection systems do not use regional convolutional neural networks (R-CNN) with expert prior knowledge to directly focus on AU-related regions adaptively. By incorporating expert prior knowledge, we propose a novel R-CNN based model named AU R-CNN. The proposed solution offers two main contributions: (1) AU R-CNN directly observes different facial regions, where various AUs are located. Expert prior knowledge is encoded in the region and the RoI-level label definition. This design produces considerably better detection performance than do existing approaches. (2) We also integrate various dynamic models (including convolutional long short-term memory, two stream network, conditional random field, and temporal action localization network) into AU R-CNN and then investigate and analyze the reason behind the performance of dynamic models. Experiment results demonstrate that *only* static RGB image information and no optical flow-based AU R-CNN surpasses the one fused with dynamic models. AU R-CNN is also superior to traditional CNNs that use the same backbone on varying image resolutions. State-of-the-art recognition performance of AU detection is achieved. The complete network is end-to-end trainable. Experiments on BP4D and DISFA datasets show the effectiveness of our approach. Code will be made available.

Keywords: Action unit detection, Expert prior knowledge, R-CNN, Facial Action Coding System

1. Introduction

Facial expressions reveal people's emotions and intentions. Facial Action Coding System (FACS)[1] has defined 44 action units (AUs) related to the movement of specific facial muscles; these units can anatomically represent all possible facial expressions, considering the crucial importance of facial expression analysis. Since then, AU detection has been studied for decades, and many interesting approaches have been proposed. Automatic detection of AUs has a wide range of applications, such as human-machine interfaces, affective computing, and car-driving monitoring.

The goal of AU detection is to recognize and predict AU labels on each frame of the facial expression video. Since the human face may present complex facial expression, and AUs appear in the form of subtle appearance changes on the local regions of face, that current classifiers cannot easily recognize. This problem is the main obstacle of current AU detection systems. Various approaches focus on fusion with extra information in convolutional neural networks (CNNs), *e.g.* , the optical flow information[2] or landmark information[3, 4], to help AU detection systems capture such subtle facial expressions. However, these approaches have high detection error rates, due to the lack of using prior knowledge. Human can easily recognize micro facial expression by their long accumulated experience. Hence, Integrating the expert prior knowledge of FACS[1] to AU detection system is promising. With fusing of this prior knowledge, our proposed approach addresses the AU detection problem by partitioning the face to easily recognizable AU-related subregions, then the prediction of each subregion is merged to obtain the image-level prediction. Fig. 1 shows our approach's framework, we design an "AU partition rule" to encode the expert prior knowledge. This AU partition rule decomposes the image into a bunch of AU-related bounding boxes. Then AU R-CNN head module focuses on recognizing each bounding box. This design can well address the three problems of existing approaches.

First, existing approaches[3–13] have been proposed to extract features near landmarks (namely, "AU center"), which is trivially defined and leading to emphasize on inaccurate places. AUs occur in regions around specific facial muscles that may be inaccurately located on a landmark or an AU center due to the limitation of the facial muscle's activity place. Thus, most AUs limit their activities in specific irregular regions of a face, and we call this limitation the "space constraint". Our approach reviews the FACS and designs the "AU partition rule" to represent this space constraint accurately. This well-designed "AU partition rule" is called the "expert prior knowledge" in our approach which is built on the basis of the space-constrained regional recognition, so it reduces the detection error rate caused by inaccurate landmark positioning (See experiment Section 4.3.1).

*Corresponding author at: School of Software Department, Tsinghua University, Beijing, 100084, PR China. Declarations of interest: none

Email addresses: sharpstill@163.com (Chen Ma), chenlee@tsinghua.edu.cn (Li Chen), yongjh@tsinghua.edu.cn (Junhai Yong)

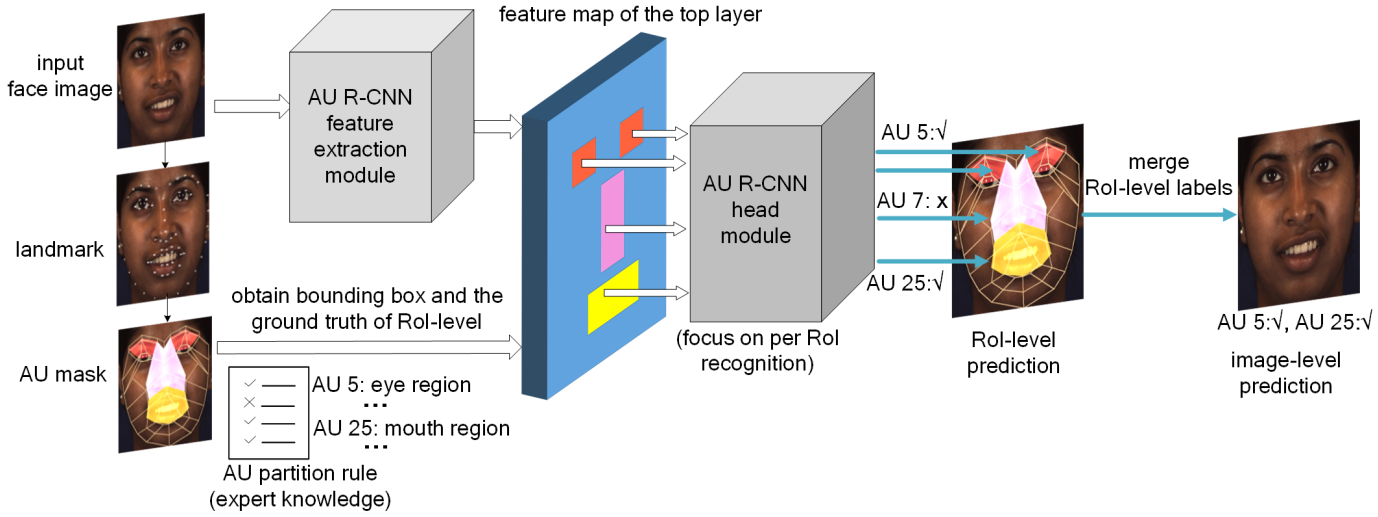


Fig. 1. AU R-CNN framework recognizes each RoI-level's label based on the AU partition rule, which uses the landmark point location information to encode the expert prior knowledge. This rule indicates where the AUs may occur. AU R-CNN head module focuses on recognizing each bounding box to improve performance.

Second, existing approaches still use CNNs to recognize a full face image[3, 4, 14, 15] and do not learn to recognize individual region's labels, which may not use the correct image context to detect. For example, a CNN may use an unreliable context, such as mouth area features, to recognize eye-area-related AUs (*e.g.* AU 1, AU 2). Recent success in the object detection model of Fast/Faster R-CNN[16, 17] has inspired us to utilize the power of R-CNN based models to learn robust regional features of AUs under space constraints. We propose the use of AU R-CNN to detect AUs only from AU-related regions by limiting its vision inside space-constrained areas. In this process, unrelated areas can be excluded to avoid interference, which is key to improving detection accuracy.

Third, the multi-label learning problem in AU detection can be addressed at a fine-grained level under AU-related RoI space constraint. Previous approaches[4, 14, 18] adopt the sigmoid cross-entropy cost function to learn the image-level multi-label and emphasize the important regions, but such a solution is not sufficiently fine-grained. The multi-label relationship can be captured accurately in the RoI-level supervised information constraint. Most facial muscles can show diverse expressions that lead to RoI-level multi-label learning. For example, AU 12 (lip corner puller) is often present in a smile, which may also occur together with AU 10 (upper lip raiser), and deepen the nasolabial fold, as shown in Fig. 2. Therefore, in the definition of the AU partition rule, AUs are grouped by the definition of FACS and related facial muscles. Each AU group shares the same region, and such AU groups can be represented by a binary vector that denotes 1 if the corresponding AU occurs in the ground truth and 0 otherwise. The sigmoid cross-entropy cost function is adopted in the RoI-level learning. In our experiments, we determine that using RoI-level labels to train and predict and then merging the RoI-level prediction result to that of the image level surpass the previous approaches.

Furthermore, the effects and analyses of fusing temporal features on dynamic models are determined. We conduct complete comparison experiments to investigate the effects of integrating dynamic models, including convolutional long short-term memory (ConvLSTM)[19], two-stream network[20], general graph conditional random field (CRF) model, and TAL-Net[21], into AU R-CNN. We analyze the reason behind such effects and the cases under which the dynamic models are effective. Our AU R-CNN with *only* static RGB images and no optical flow achieves 63% average F1 score on BP4D, and outperforms all dynamic models. The main contributions of our study are as follows.

1) AU R-CNN is proposed to learn regional features adaptively under RoI-level multi-label supervised information. Specifically, we encode the expert prior knowledge by defining the AU partition rule, including the AU groups and related regions, according to FACS[1].

2) We investigate the effects of various dynamic models, including two-stream network, ConvLSTM, CRF model and TAL-Net, on the BP4D[22] and DISFA[23] databases. The reasons behind such experiment effects and the effective cases are analyzed. The experiment results show that our static RGB image-based AU R-CNN achieves the best average F1 score in BP4D and is close to the performance of the best dynamic model in DISFA. Our approach achieves state-of-the-art performance in AU detection.

2. Related Work

Extensive works on AU detection have been proposed to extract effective facial features. The facial features in AU detection can be grouped into appearance and geometric features. Appearance features portray the local or global changes

in facial components. Most popular approaches in this category adopt Haar feature[24], local binary pattern[25], Garbor wavelets[26, 27], and canonical appearance feature[28]. Geometric features represent the salient facial point or skin changing direction or distance. Geometric changes can be measured by optical flows[29] or displacement of facial landmark points[28, 30]. Landmark plays an important role in geometry approaches, and many methods have been proposed to extract features near landmark points[5–12, 31]. Fabian *et al*. [32] proposed a method that combines geometric changes and local texture information. Wu and Ji[33] investigated the combination of facial AU recognition and facial landmark detection. Zhao *et al*. [13] proposed joint patch and multi-label learning (JPML) for AU detection with a scale-invariant feature transform descriptor near landmarks. These traditional approaches focus on extracting handcraft features near landmark points. With the recent success of deep learning, CNN[34] have been widely adopted to extract AU features[**?**]. Zhao *et al*. [14] proposed a deep region and multi-label learning (DRML) network to divide the face images into 8×8 blocks and used individual convolutional kernels to convolve each block. Although this approach treats each face as a group of individual parts, it divides blocks uniformly and does not consider the FACS knowledge, thereby leading to poor performance. Wei Li *et al*. [4] proposed Enhancing and Cropping Net (EAC-Net), which intends to give significant attention to individual AU centers; however, this approach defines the AU center trivially and it uses image-level context to learn. Its CNN backbone may use incorrect context to classify and the lack of RoI-level supervised information can only give coarse guidance. Song *et al*. [35] studied the sparsity and co-occurrence of AUs. Han *et al*. [15] proposed an Optimized Filter Size CNN (OFS-CNN) to simultaneously learn the filter sizes and weights of all conv layer. Other related problems, including the effects of dataset size[36], the action detection in videos[37], the pose-based feature of action recognition[38], and generalized multimodal factorized high-order pooling for visual question answering[39] have also been studied. Previous works have mainly focused on landmark-based subregions or learning multiple subregions with convolutional kernels separately. Detection with expert prior knowledge and utilizing RoI-level labels are important but have been undervalued in previous methods.

Researchers have utilized temporal dependencies in video sequences over the last few years. Romero *et al*. [2] advocated a two-stream CNN model that combines optical flow and RGB information, and their result was promising. However, they used one binary classification model for each AU, which caused their approach to be time consuming to train and yield numerous model parameters. The CNN and LSTM hybrid network architectures are studied in Chu *et al*. [40], Wei Li *et al*. [3] and He *et al*. [41], which feed the CNN-produced features to LSTM to improve performance by capturing the temporal relationship across frames. However, their solutions are inefficient because they are not an end-to-end networks. In our experiments, we also investigate the effects of using temporal feature relationships in the time axis of videos. We use various dynamic models (including two-stream network, ConvLSTM etc.) that are incorporated into AU R-CNN. Such temporal dependency cannot always improve performance in all cases (Section 4.5).

Unlike existing approaches, AU R-CNN is a unified end-to-end learning model that encodes expert prior knowledge and outperforms state-of-the-art approaches. Thus, it is a simple and practical model.

3. Proposed Method

3.1. Overview

AU detection can be considered a multi-label classification problem. The most popular image classification approach is the CNN, and the basic assumption for a standard CNN is the shared convolutional kernels for an entire image. For a highly structural image, such as a human face, a standard CNN will fail to capture subtle appearance changes. To address this issue, we propose AU R-CNN, in which expert prior knowledge is encoded. We review FACS[1] and define a rule ("AU partition rule") for partitioning a face on the basis of FACS knowledge using landmarks. With this rule, we can treat each face image as a group of separate regions and AU R-CNN is proposed to recognize each region. The overall procedure is composed of two steps. First, the face image's landmark key points is obtained, and then the face is partitioned into regions on the basis of the AU partition rule and the landmark coordinates. An "AU mask" is generated in this step, and the expert prior knowledge is encoded into the AU mask. Second, the face images is input into the AU R-CNN's backbone, and the produced feature map and the minimum bounding box of the AU mask are then fed into AU R-CNN's RoI pooling layer together. The final fully-connected(fc) layer's output can be treated as classification probabilities. The image-level ground truth label is also partitioned to RoI-level in the learning. After AU R-CNN is trained over, the prediction is performed on the RoI-level. Then, we use a "bit-wise OR" operator to merge RoI-level prediction labels to image-level ones. In this section, we introduce the AU partition rule and then AU R-CNN. We also introduce a dynamic model extension of AU R-CNN in the end of this section.

3.2. AU partition rule

AUs appear in specific regions of a face but are not limited to facial landmark points; previous AU feature extraction approaches directly use facial landmarks or offsets of the landmarks as AU centers[3, 4, 13], but actual places where activities occur may be missed, and sensitivity may increase. Instead of identifying the AU center, we adopt the domain-related expertise to guide the partition of AU-related RoIs. The first step is to utilize the dlib[42] toolkit to obtain 68 landmark points. The landmark points provide rich information about the face, and the landmark points help us focus on areas where AUs may occur. Fig. 2 shows the region partition of a face, and several extra points are calculated using 68 landmarks. A typical example is shown in Fig. 2 right. The face image is partitioned into 43 basic RoIs using landmarks. Then, on the basis of FACS

Table 1. FACS definition of AUs and related anatomy muscles[1]

AU number	AU name	Muscle Basis
1	Inner brow raiser	Frontalis
2	Outer brow raiser	Frontalis
4	Brow lowerer	Corrugator supercilii
6	Cheek raiser	Orbicularis oculi
7	Lid tightener	Orbicularis oculi
10	Upper lip raiser	Levator labii superioris
12	Lip corner puller	Zygomaticus major
14	Dimpler	Buccinator
15	Lip corner depressor	Depressor anguli oris
17	Chin raiser	Mentalis
23	Lip tightener	Orbicularis oris
24	Lip pressor	Orbicularis oris
25	Lips part	Depressor labii inferioris
26	Jaw drop	Masseter

definition[1] (Table 1) and the anatomy of facial muscle structure[2], the AU partition rule and the AU mask can be defined for representing the expert prior knowledge. For this purpose, we classify AUs into four cases.

1) The RoIs defined in Fig. 2 are the basic building blocks. One AU contains multiple basic RoIs; hence, multiple RoIs are selected to be grouped and assigned to AUs by RoI numbers (Table 2). The principle of such RoI assignment is the FACS muscle definition (Table 1). The grouped RoIs are called the "AU mask".

2) Most muscles can present multiple AUs—in other words, some AUs can co-occur in the same place. For example, AU 12 (lip corner puller) and AU 10 (upper lip raiser) are often present together in a smile, which requires lifting of the muscle and may also deepen the nasolabial fold, as shown in Fig. 4(e). Therefore, we group AUs into 8 "AU groups" on the basis of AU-related muscles defined in FACS (Table 1) and the AU co-occurrence statistics of the database. Each AU group has its own mask, whose region is shared by the AUs. One AU group contains multiple basic RoIs, which are defined in Fig. 2, to form an AU mask (Fig. 4).

3) Some AU groups are defined in a hierarchical structure, that is, these AU group masks have a broad area, which contains other AU group's small areas. For example, AU group # 6 contains AU group # 7 (Fig. 4). The reason behind such a design is that AU group # 7 (AU 17) is caused by the movement of the mentalis (Table 1), which is in the chin. The bone structure of the chin makes it a relatively stable area, which limits the possible occurrence of AU 17. Therefore, we can define a detailed area in AU group # 7 (Fig.4(g)). However, AU group # 6 consists of AU 16, AU 20, AU 25 and AU 26, and it is located in the mouth area. The mouth area contains several possible movement locations (mouth open, mouth close, smell, laugh, etc.), and the chin area follows mouth opening and closing. Therefore we define AU group # 6 to contain the area of AU group # 7 (Fig. 4(f)). The partition of the face image naturally leads to RoI-level label assignment. In this case, the AU group # 6 must contain RoI-level labels of AU group # 7. We define operator "label fetch" #7∈#6 to enable AU group # 6 to fetch labels from AU group #7 (Table 2).

4) Some AU groups have overlapping areas with other AU groups' areas. For example, AU group # 3's mask, which is across the nose area (Fig. 7(c)), will also contain labels of AU group # 4 (Fig. 7(e)); thus, we also define operator "label fetch" #4∈#3 to fetch labels from AU group # 4 in this case. (Table 2).

In summary, Table 2 and Fig. 4 show the AU partition rule and the AU mask. The AU group definition is related not only to the RoI partition of the face, but also to the RoI-level label assignment.

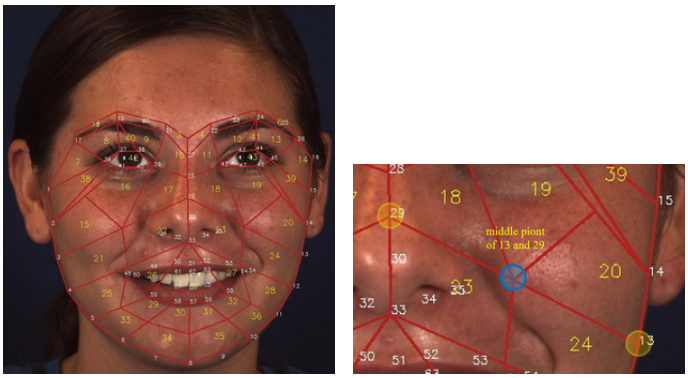


Fig. 2. Landmark and region partition of face. Yellow and white numbers indicate the RoI number and landmark number respectively. **Left:** Partition of 43 RoIs. **Right:** Position of blue point is the average position of landmark 13 and 29.

3.3. AU R-CNN

AU R-CNN is composed of two modules, namely, feature extraction and head modules. This model can use ResNet[43] or VGG[44] as its backbone. Here, we use ResNet-101 to illustrate. The feature extraction module comprises conv layers that produce the feature maps (ResNet-101's conv1, bn1, res2, res3, res4 layers), and the head module includes an RoI pooling layer and the subsequent top layers (res5, avg-pool, and fc layers). After AU masks are obtained, unrelated areas can be excluded. However, each AU mask is an irregular polygon area, which means it cannot be directly fed into the fc layer. Therefore, we introduce the RoI pooling layer originally from Fast R-CNN[16]. The RoI pooling layer is designed to convert the features inside any rectangle RoI (or bounding box) into a small feature map with a fixed spatial extent of $H \times W$. To utilize the RoI pooling layer, each AU mask is converted into a minimum bounding box (named "AU bounding box") around the mask to input [3](Fig. 5). The RoI pooling layer needs a parameter named "RoI size", indicates the RoI's height and width after pooling.

[1] https://www.cs.cmu.edu/~face/facs.htm

[2] https://en.wikipedia.org/wiki/Facial_muscles

[3] AU group #1 contains two separate symmetrical regions, thus it contains two bounding boxes, which results in total 9 AU bounding boxes, one more than AU group number.

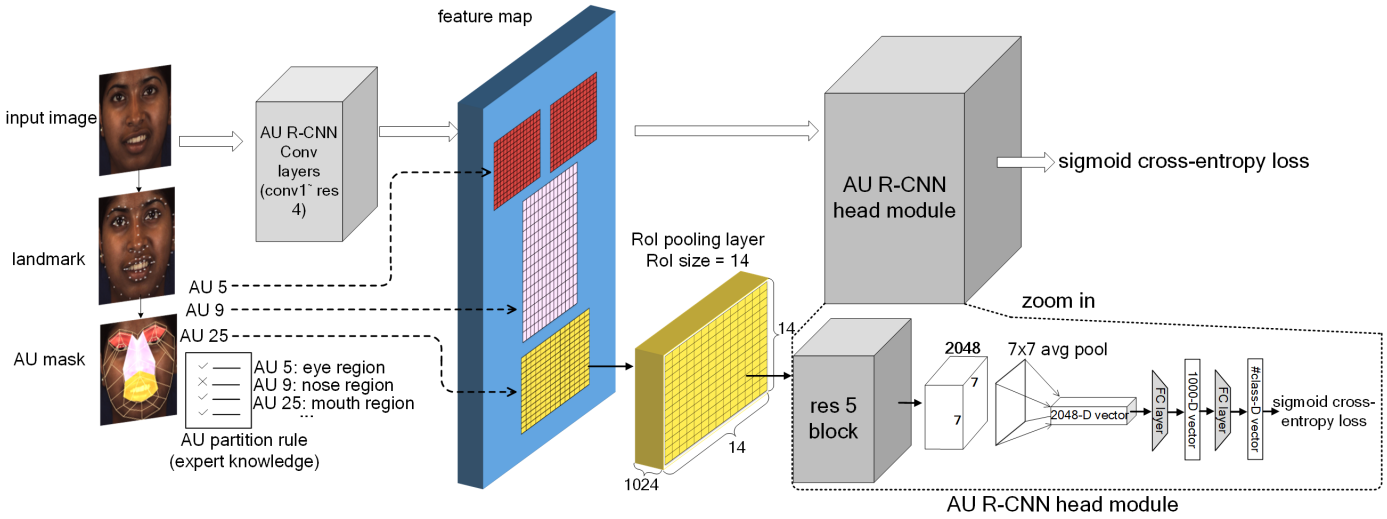


Fig. 3. AU R-CNN using ResNet-101 backbone architecture, where #class denotes the AU category number we wish to discriminate.

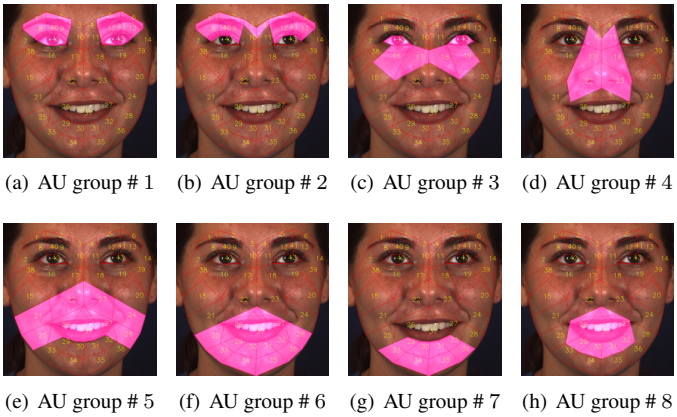
Table 2. AU partition rule

AU group	AU NO	RoI NO
# 1* (∈# 2)	AU 1 , AU 2 , AU 5 , AU 7	1, 2, 5, 6, 8, 9, 12, 13, 40, 41, 42, 43
# 2	AU 4	1, 2, 3, 4, 5, 6, 8, 9, 12, 13, 40, 41
# 3	AU 6	16, 17, 18, 19, 42, 43
# 4 (∈ # 3)	AU 9	10, 11, 17, 18, 22, 23
# 5 (∈ # 6)	AU 10 , AU 11 , AU 12 , AU 13 , AU 14 , AU 15	21, 22, 23, 24, 25, 26, 27, 28, 37
# 6 (∈ # 5)	AU 16 , AU 20 , AU 25 , AU26 , AU 27	25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37
# 7 (∈ # 6)	AU 17	29, 30, 31, 32, 33, 34, 35, 36
# 8 (∈ # 5, # 6)	AU 18 , AU 22 , AU 23 , AU 24 , AU 28	26, 27, 29, 30, 31, 32, 37

Note: Symbol ∗ means the corresponding AU group have symmetrical regions. Symbol ∈ indicates the "label fetch".

In our experiment, we set RoI size to 14×14 in ResNet101 backbone and 7×7 in VGG-16 and VGG-19 backbone.

Object detection networks, such as Fast R-CNN, aim to identify and localize the object. Benefiting from the design of the AU mask, we have strong confidence in where the AUs should occur; thus, we can concentrate on what the AUs are. Fig. 3 depicts the AU R-CNN's forward process. In the RoI pooling layer, we input the AU bounding box and feature map (The bounding box coordinate and feature map are usually $16\times$ smaller than the input image resolution). We treat the last fully connected layer's output vector as predicted label probabilities.



(a) AU group # 1 (b) AU group # 2 (c) AU group # 3 (d) AU group # 4

(e) AU group # 5 (f) AU group # 6 (g) AU group # 7 (h) AU group # 8

Fig. 4. Action Unit masks for AU group #1 $\sim$ #8 (see Table 2).

The total AU category number we wish to discriminate is set as L[4]; the number of bounding boxes in each image is R[5]; the ground truth $\mathbf{y} \in \{0,1\}^{R\times L}$, $\mathbf{y}_{i,j}$ indicates the (i,j)-th element of $\mathbf{y}$, where $\mathbf{y}_{i,j} = 0$ denotes AU j is inactive in bounding box i, and AU j is active if $\mathbf{y}_{i,j} = 1$. The ground truth $\mathbf{y}$ must satisfy the AU partition rule's constraint: $\mathbf{y}_{i,j} = 0$ if AU j does not belong to bounding box i's corresponding AU group (Fig. 5 and Table 2). The RoI-level prediction probability is $\hat{\mathbf{y}} \in \mathbb{R}^{R\times L}$. Given multiple labels inside each RoI (*e.g.* AU 10 and AU 12 often occur together in the mouth area), we adopt the multi-label sigmoid cross-entropy loss function, namely,

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{C}\{\mathbf{y}_{rc}\log(\hat{\mathbf{y}}_{rc})\} \quad (1)$$

Unlike ROI-Nets[3] and EAC-Net[4], AU R-CNN has considerably fewer parameters due to the sharing of conv layer in the

[4] $L = 22$ in BP4D database and $L = 12$ in DISFA database.

[5] $R = 9$ in BP4D database (Fig. 5) and $R = 7$ in DISFA database (since DISFA doesn't contain AU group# 7 and AU group # 8).

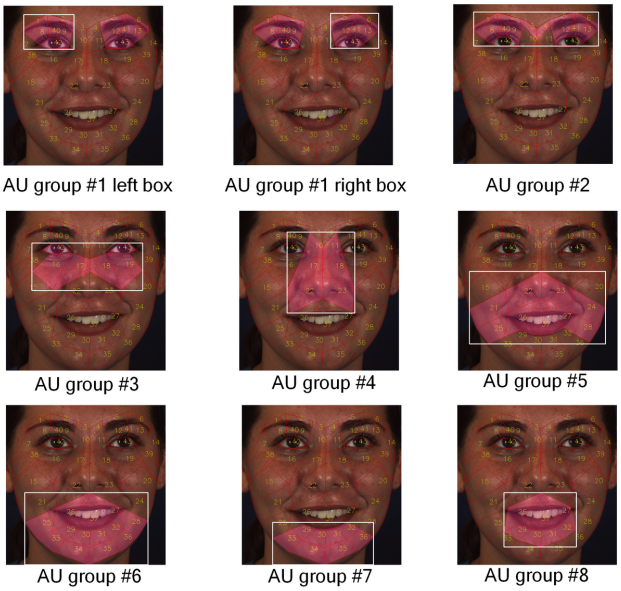


Fig. 5. AU bounding boxes, which are defined as the minimum bounding box around each AU mask. Since AU group #1 has two symmetrical regions, the bounding box number is 9.

feature extraction module, which leads to space and time saving. The RoI pooling layer and RoI-level label also help improve classifier performance through the space constraint and supervised information of the RoIs.

In the inference stage, the last fc layer's output is converted to a binary integer prediction vector using the threshold of zero (the elements that greater than 0 set to 1, others set to 0). Multiple RoIs' prediction results are merged via a "bit-wise OR" operator to obtain the image-level label. We report this merged image-level prediction result in Section 4.

3.4. Dynamic model extension of AU R-CNN

AU R-CNN can use only static RGB images to learn. A natural extension is to use the RoI feature map extracted from AU R-CNN to model the temporal dependency of RoIs across frames. In this extension, we can adopt various dynamic models to observe RoI-level appearance changes (Experiments are shown in Section 4.5). In this section, we introduce one extension that integrates ConvLSTM into the AU R-CNN architecture.

Fig. 6 shows the AU R-CNN integrated with ConvLSTM architecture. In each image, we first extract nine AU group RoI features ($7 \times 7 \times 2048$) correspond to nine AU bounding boxes of Fig. 5 from the last conv layer. To represent the evolvement of facial local regions, we construct an RoI parallel line stream with nine timelines. The timeline is constructed by skipping four frames per time-step in the video to eliminate the similar frames. In total, we set 10 time-steps for each iteration. In each timeline, we connect the RoI at the current frame to the corresponding RoI at the adjacent frames, *e.g.* the mouth area has only temporal correlation to the next/previous frame's mouth area. Therefore, each timeline corresponds to an AU bounding box's evolution across time. Nine ConvLSTM kernels are used to process on the nine timelines. The output of each ConvLSTM kernel are fed into two fc layers to produce the prediction probability. More specifically, Let's denote the mini-batch size as N. the time-steps as T, the channel, height and width of RoI feature as C, H and W respectively. The concatenation of ConvLSTM's all time-step's output is a five-dimensional tensor of shape $[N, T, C, H, W]$. We reshape this tensor to a two-dimensional tensor of shape $[N \times T, C \times H \times W]$, the first dimension is treated as the mini-batch of shape $[N \times T]$. This tensor is input to two fc layers to get a prediction probability vector $[N \times T, Class]$, where $Class$ denotes AU category number. We adopt sigmoid cross-entropy loss function to minimize difference between the prediction probability vector and ground truth, which is the same as Eq. 1. In the inference stage, we use the last frame's prediction result of the 10-frame video clip to evaluate. This model, named "AR$_{ConvLSTM}$", is trained together with AU R-CNN in an end-to-end form.

The introduction of dynamic model extension brings new issues, as shown in our experiments (Section 4.5), the dynamic model cannot always improve overall performance as expected. We use database statistics and a data visualization technique to identify the effective cases. Various statistics of BP4D and DISFA databases are collected, including the AU duration of each database and the AU group bounding box areas. Li Wei *et al.* [4] found that the occurrence of AUs in the database have the influence of static-image-based AU detection classifiers. However, in the ConvLSTM extension model, the average AU activity duration of videos and AR$_{ConvLSTM}$ classification performance are correlated. Fig. 9 provides an intuitive figure of such correlation, when the AU duration increases at high peak, the performance of AR$_{ConvLSTM}$ can be always improved. Therefore, in situations such as long-duration activities, AR$_{ConvLSTM}$ can be adopted to improve performance. Other dynamic models can also be integrated into AU R-CNN, including the two-stream network, TAL-Net, and the general graph CRF model. In Section 4.5, we collect the experiment results and analyze various dynamic models in detail.

4. Experiments and Results

4.1. Settings

4.1.1. Dataset description

We evaluate our method on two datasets, namely, BP4D dataset[22] and DISFA dataset[23]. For both datasets, we adopt a 3-fold partition to ensure that the subjects are mutually exclusive in the train/test split sets. AUs that present more than 5% base rate are included for evaluation. In total, we select 12 AUs on BP4D and 8 AUs on DISFA to report the experiment results.

(1) BP4D[22] contains 41 young adults of different races and genders (23 females and 18 males). We use 328 videos (41 participants $\times 8$ videos) captured in total, which result in $\sim$ 140,000 valid face images. We select positive samples as those with AU intensities equal to or higher than A-level, and the rest are negative samples. We use 3-fold splits exactly the same as [4, 18] partition to ensure that the training and testing subjects are mutually exclusive. The average AU activity duration of all videos in BP4D and the total activity segment count

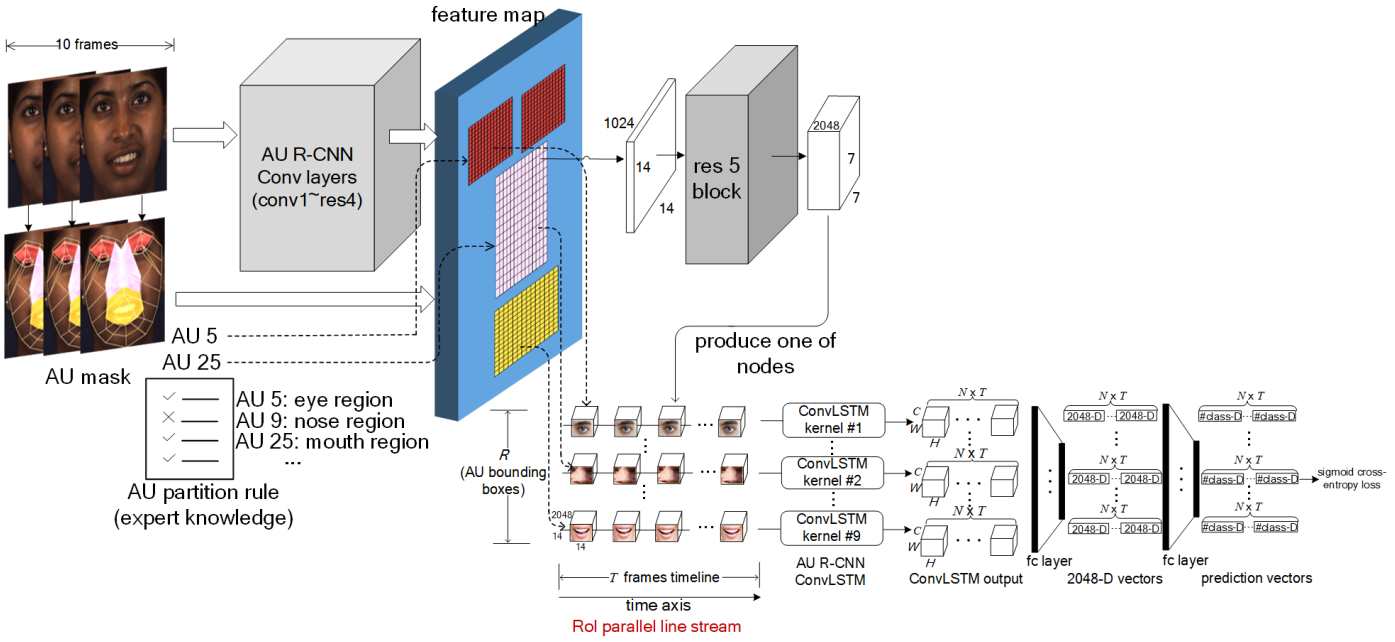


Fig. 6. AU R-CNN integrated with ConvLSTM architecture, where N denotes mini-batch size; T denotes the frames to process in each iteration; R denotes AU bounding box number; C, H, and W denotes the ConvLSTM's output channel number, height and width respectively. #class denotes the AU category number we wish to discriminate.

are shown in Table 14. The average AU mask bounding box area is provided in Table 9.

(2) DISFA[23] contains 27 subjects. We determine $\sim$ 260,000 valid face images and 54 videos (27 videos captured by left camera and 27 videos captured by right camera). We also use the 3-fold split partition protocol in the DISFA experiment. The average AU activity duration of all videos in DISFA and the total activity segment count are shown in Table 15. The average AU mask bounding box area is given in Table 10.

4.1.2. Evaluation metric

Our task is to detect whether the AUs are active, which is a multi-label binary classification problem. Since our approach focuses on RoI prediction for each bounding box (Fig. 5), the RoI-level prediction is a binary vector with L elements, where L denotes the total AU category number we wish to discriminative. We use the image-level prediction to evaluate, which is obtained by using a "bit-wise OR" operator for merging an image's RoI-level predictions. After obtaining the image-level prediction, we directly use the database provided image-level ground truth labels to evaluate, which are also a binary vectors with elements equal 1 for active AUs and equal 0 for inactive AUs. The F1 score can be used as an indicator of the performances of the algorithms on each AU and is widely employed in AU detection. In our evaluation, we compute frame-based F1 score[9] for 12 AUs in BP4D and 8 AUs in DISFA on image-level prediction. The overall performance of the algorithm is described by the average F1 score(denoted as *Avg.*).

Table 3. Compared models details

Model	E2E	ML	RGB	LANDMARK	CONVERGE	VIDEO
CNN_{res}	✓	✓	✓	×	✓	×
AR_{vgg16}	✓	✓	✓	✓	✓	×
AR_{vgg19}	✓	✓	✓	✓	✓	×
AR_{res}	✓	✓	✓	✓	✓	×
AR_{mean_box}	✓	✓	✓	×	✓	×
AR_{FPN}	✓	✓	✓	✓	✓	×
$\text{AR}_{ConvLSTM}$	✓	✓	✓	✓	✓	✓
$\text{AR}_{2stream}$	✓	✓	×	✓	✓	✓
AR_{CRF}	×	×	✓	✓	✓	✓
AR_{TAL}	×	✓	×	✓	×	✓

* **E2E**: end-to-end trainable, **ML**: multi-label learning, **RGB**: only use RGB information, not incorporate optical flow, **LANDMARK**: use landmark point, **CONVERGE**: the model converged in training, **VIDEO**: need video context.

4.1.3. Compared methods

We collect the F1 scores of the most popular state-of-the-art approaches that used the same 3-fold protocol in Table 4 and Table 7 to compare our approaches with other methods. These techniques include a linear support vector machine (LSVM), active patch learning (APL[45]), JPML[13], a confidence-preserving machine (CPM[10]), a block-based region learning CNN (DRML[14]), an enhancing and cropping nets (EAC-net[4]), an ROI adaption net (ROI-Nets[3]), and LSTM fused with a simple CNN (CNN+LSTM[40]), an optimized filter size CNN (OFS-CNN[15]). We also conduct complete control experiments of AU R-CNN in Table 5 and Table 8, including traditional ResNet-101 based CNN that classifies the entire face images (CNN_{res}), ResNet-101 based AU R-CNN (AR_{res}), VGG-16 based AU R-CNN (AR_{vgg16}), VGG-19 based AU R-CNN (AR_{vgg19}), mean bounding boxes

(a) AU 1, 4, 7, 10 (b) AU 6, 7, 10, 12 (c) AU 2, 14, 17 (d) AU 6, 7, 12 (e) AU 7, 14, 17

Fig. 7. Example figures of detection result.

version AU R-CNN (AR_{mean_box}), AU R-CNN incorporate with Feature Pyramid Network[46](AR_{FPN}), AU R-CNN integrated with ConvLSTM[19] ($AR_{ConvLSTM}$), AU R-CNN with optical flow and RGB feature fusion two-stream network architecture[20]($AR_{2stream}$), general graph CRF with features extracted by AU R-CNN(AR_{CRF}), and AU R-CNN with a temporal action localization in video network, TAL-Net[21](AR_{TAL}). We use ResNet-101 based CNN(CNN_{res}) as our baseline model. The details of the compared models are summarized in Table 3.

4.1.4. Implementation details

We resize the face images to 512×512 after cropping the face areas. Each image and bounding box are horizontally mirrored randomly before being sent to AU R-CNN for data augmentation. We also subtract the pixel value of all the images in the dataset before sending to AU R-CNN. We use dlib[42] to landmark faces, and the landmark operator is consequently time consuming. We cache the mask in the memcached database to accelerate speed in later epochs. The VGG and ResNet-101 backbones of AU-RCNN use pre-trained ImageNet ILSVRC dataset[47] weights to initialize. AU R-CNN is initialized with a learning rate of 0.001 and further reduced by a factor of 0.1 after every 10 epochs. In all experiments, we select momentum stochastic gradient descent to train AU R-CNN for 25 epochs and set momentum to 0.9 and weight decay to 0.0005. The mini-batch size is set to 5.

4.2. Conventional CNN versus AU R-CNN

AU R-CNN is proposed for adaptive regional learning in Section 3.3. Thus, our first experiment aims to determine whether it can perform better than the baseline conventional CNN, which uses entire face images to learn. We suppose that by learning the adaptive RoIs separately, recognition capability can be improved. We train CNN_{res} and AR_{res} on the BP4D and DISFA datasets using the same ResNet-101 backbone for comparison. Twelve AUs in BP4D and eight AUs in DISFA are used together; therefore, AR_{res} and CNN_{res} use the sigmoid cross-entropy loss function, as shown in Eq. 1. Both models are based on static images. During each iteration, we randomly select five images to comprise one mini-batch to train and initialize the learning rate to 0.001.

Fig. 7 demonstrates the example detection results of our approach. Table 5 and Table 8 show the BP4D and DISFA results, in which the margin is larger in DISFA (3.69%) than in BP4D (2.1%). These results can be attributed to the relatively lower resolution images in DISFA, which cause AR_{res} to benefit more. We also show that AU R-CNN performs efficiently with varying image resolutions. Experiments have been conducted to compare the proposed AU R-CNN and baseline CNN using the same ResNet-101 backbone on the BP4D database with different resolutions of the input image. Table 6 shows the result, and the resolutions of images are set to 256×256, 416×416, 512×512, and 608×608. Most AU results prefer AU R-CNN model by observing subtle cues of facial appearance changes. In 256×256, although the resolution is nearly half of that in 512×512, the performance is close to that in 512×512. This similarity leads to efficient detection when using 256×256. But in the highest resolution 608×608, the F1 score is lower than that of 512×512, we believe this performance drop can be attribute to two possible reasons. (1) As pointed out by [15], when the image resolution increases to 608×608, the receptive field covers a smaller actual area of the entire face when using the same convolution filter size. The smaller receptive field deteriorates the vision. (2) Larger images produce larger feature maps before RoI pooling layer in AR_{res}, or larger feature maps before the final avg pooling layer in CNN_{res}. The increasing of feature map size also increases each pooling grid cell's covered size dramatically in RoI pooling/avg pooling layer, which has negative impact on high level features. Regardless of the overall improvement of AU R-CNN across the 12 AUs. In AU 10 and AU 12, CNN and AU R-CNN obtain similar results. One explanation is that AU 10 and AU 12 have relatively sufficient training samples compared with other AUs.

In the DISFA dataset (Table 8, Table 7), AR_{res} outperforms CNN_{res} in six out of eight AUs. The two remaining AUs are AU 12 and AU 25. As shown in Table 10, AU 12 and AU 25 have the largest area proportions (29.8 % and 26.6 %) on the face images. In BP4D and DISFA, AU 1 (inner brow raiser) has a significant improvement in AR_{res} because of the relatively small area on the face.

4.3. ROI-Nets versus AU R-CNN

Our proposed AU R-CNN in Section 3.3 is designed to recognize local regional AUs in static images under AU mask. Previous state-of-the-art static image AU detection approach ROI-Nets[3] also focuses on regional learning; they attempt to learn regional features by using individual conv layer over subregions centered on AU center (Fig. 8). The two models are based on static images, whereas our AU R-CNN uses the shared conv layer in feature extraction module and RoI-level supervised information. This choice saves space and time, and provides accurate guidance. Instead of using the concept of the AU center area, we introduce the AU mask. We believe that AU mask can preserve more context information than cropping from AU center. ROI-Nets adopts VGG-19 as backbone. For fair comparison, we adopt VGG-19 based AU R-CNN (denoted as AR_{vgg19}) to compare. AR_{vgg19} outperforms ROI-Nets in 8 out of 12 AUs in BP4D (Table 4).

The interesting part lies in AU 23 (lip pressor) and AU 24 (lip tighter), in which AR_{vgg19} significantly outperforms ROI-Nets by 7.8% and 10.9%, respectively. This superiority is because the lip area is a relatively small area on face; AU R-CNN uses AU mask and RoI-level label so that it can concentrate on this

Table 4. F1 score result comparison with state-of-the-art methods on **BP4D** dataset. Bracketed bold numbers indicate the best score; bold numbers indicate the second best.

AU	LSVM	JPML[13]	DRML[14]	CPM[10]	CNN+LSTM[40]	EAC-Net[4]	OFS-CNN[15]	ROI-Nets[3]	FERA[48]	AR_{vgg16}	AR_{vgg19}	AR_{res}
1	23.2	32.6	36.4	43.4	31.4	39	41.6	36.2	28	**47.5**	44.8	[**50.2**]
2	22.8	25.6	41.8	40.7	31.1	35.2	30.5	31.6	28	40.5	**43.5**	[**43.7**]
4	23.1	37.4	43	43.4	[**71.4**]	48.6	39.1	43.4	34	55.1	52.2	**57**
6	27.2	42.3	55	59.2	63.3	76.1	74.5	**77.1**	70	73.8	75.7	[**78.5**]
7	47.1	50.5	67	61.3	77.1	72.9	62.8	73.7	**78**	76.6	75.2	[**78.5**]
10	77.2	72.2	66.3	62.1	45	81.9	74.3	[**85**]	81	82	**82.7**	82.6
12	63.7	74.1	65.8	68.5	82.6	86.2	81.2	[**87**]	78	85.2	85.9	[**87**]
14	64.3	65.7	54.1	52.5	**72.9**	58.8	55.5	62.6	[**75**]	64.9	63.4	67.7
15	18.4	38.1	36.7	34	34	37.5	32.6	45.7	20	**48.8**	45.3	[**49.1**]
17	33	40	48	54.3	53.9	59.1	56.8	58	36	**60.6**	60	[**62.4**]
23	19.4	30.4	31.7	39.5	38.6	35.9	41.3	38.3	41	43.9	**46.1**	[**50.4**]
24	20.7	42.3	30	37.8	37	35.8	-	37.4	-	[**49.3**]	48.3	[**49.3**]
Avg	35.3	45.9	48.3	50	53.2	55.9	53.7	56.4	51.7	**60.7**	60.3	[**63**]

Table 5. **Control experiments for BP4D**. Results are reported using F1 score on 3-fold protocol.

AU	CNN_{res}	AR_{vgg16}	AR_{vgg19}	AR_{res}	AR_{mean_box}	AR_{FPN}	$AR_{ConvLSTM}$	$AR_{2stream}$	AR_{CRF}	AR_{TAL}
1	45.8	47.5	44.8	[**50.2**]	45.8	46.4	48	46.6	**50.1**	41.3
2	43.2	40.5	**43.5**	[**43.7**]	41.1	40.7	43.2	42.1	35	37.4
4	54.3	55.1	52.2	[**57**]	[**57**]	47.5	53.1	52.4	45.2	44.5
6	**77.4**	73.8	75.7	[**78.5**]	75.1	76.4	76.9	75.4	71.4	64.4
7	77.9	76.6	75.2	[**78.5**]	77.7	76.9	**78.4**	77.3	77.7	73.6
10	81.8	82	**82.7**	82.6	82.2	81.3	[**82.8**]	82.1	82.1	76.2
12	85.8	85.2	85.9	87	86.5	85.4	[**87.9**]	**87.1**	86.9	80
14	60.8	64.9	63.4	[**67.7**]	62	63.5	[**67.7**]	62.7	67.2	64.9
15	[**50**]	48.8	45.3	49.1	48	44.9	45.6	**49.6**	47.6	45.7
17	58.3	60.6	60	62.4	61.5	57.9	[**63.4**]	**63.2**	58.7	53.3
23	47.6	43.9	46.1	[**50.4**]	48.7	42.3	47.9	**49.9**	36.8	39.1
24	48.4	49.3	48.3	49.3	53.2	46.6	**56.4**	[**57.6**]	51.6	49.5
Avg	60.9	60.7	60.3	[**63**]	61.6	59.2	**62.6**	62.2	59.2	55.8

Table 6. F1 score of varying resolutions comparison result on **BP4D** dataset. The **bold** highlights the best performance in each resolution experiment.

resolution	256 × 256		416 × 416		512 × 512		608 × 608	
AU	CNN_{res}	AR_{res}	CNN_{res}	AR_{res}	CNN_{res}	AR_{res}	CNN_{res}	AR_{res}
1	45.6	**50.1**	47.4	**49.3**	45.8	**50.2**	44.3	**47.5**
2	43.6	**46.5**	38.3	**42.1**	43.2	**43.7**	**40.1**	39.2
4	52.2	**54.6**	**53.3**	50.0	54.3	**57.0**	49.5	**53.5**
6	74.9	**77.7**	**75.7**	75.2	77.4	**78.5**	76.3	**76.9**
7	76.3	**78.3**	75.7	**78.7**	77.9	**78.5**	76.4	**78.6**
10	**82.5**	81.7	**82.4**	82.3	81.8	**82.6**	81.5	**82.7**
12	86.5	**87.5**	**87.2**	86.5	85.8	**87.0**	**87.5**	85.5
14	55.4	**62.1**	59.5	**61.9**	60.8	**67.7**	59.5	**62.0**
15	48.0	**51.2**	44.1	**49.2**	**50.0**	49.1	44.9	**49.6**
17	59.9	**61.8**	57.5	**61.4**	58.3	**62.4**	57.4	**61.3**
23	44.7	**46.2**	41.2	**44.9**	47.6	**50.4**	**45.6**	45.1
24	46.9	**52.3**	44.5	**47.7**	48.4	**49.3**	48.2	**51.1**
Avg	59.7	**62.5**	58.9	**60.8**	60.9	**63.0**	59.3	**61.1**

Table 7. F1 score result comparison with state-of-the-art methods on **DISFA** dataset. Bracketed bold numbers indicate the best score; bold numbers indicate the second best.

AU	LSVM	APL[45]	DRML[14]	ROI-Nets[3]	CNN_{res}	AR_{vgg16}	AR_{vgg19}	AR_{res}
1	10.8	11.4	17.3	[**41.5**]	26.3	24.9	26.9	**32.1**
2	10	12	17.7	[**26.4**]	23.4	23.5	21	**25.9**
4	21.8	30.1	37.4	[**66.4**]	51.2	55.5	59.6	**59.8**
6	15.7	12.4	29	50.7	48.1	51	[**56.5**]	**55.3**
9	11.5	10.1	10.7	8.5	29.9	**41.8**	[**46**]	39.8
12	**70.4**	65.9	37.7	[**89.3**]	69.4	68	67.7	67.7
25	12	21.4	38.5	[**88.9**]	**80.1**	74.9	79.8	77.4
26	22.1	26.9	20.1	15.6	**52.4**	49.4	47.6	[**52.6**]
Avg	21.8	23.8	26.7	48.5	47.6	48.6	**50.7**	[**51.3**]

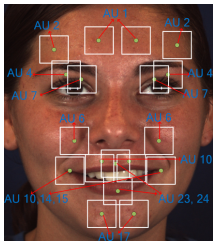


Fig. 8. The AU centers of ROI-Nets, each AU center location is an offset of landmark point, and the 3×3 bounding boxes centered at AU centers from top layer's feature map are cropped.

area. This fact can be verified from Table 9 that AU 23 and AU 24 bounding box only occupy 14.7% area of the face image. Other typical cases are AU 1, AU 2, and AU 4, which are located in the areas around eyebrows and eyes; AR_{vgg19} outperforms ROI-Nets by 8.5%, 11.9%, and 8.8%, respectively. In AU 6 (cheek raiser, see Fig. 7(c)) and AU 10, AU 12, AU 14, and AU 15 results, ROI-Nets and AU R-CNN achieve close results. These areas occupy relatively large proportions in the image (Table 9), and ROI-Nets focus on the central large area of the image. The experiment in DISFA dataset (Table 7) demonstrates the similar result. The above comparisons prove that, AU R-CNN better expresses the classification information of local regions than ROI-Nets. We also found that ResNet-based AU R-CNN (AR_{res}) outperforms AR_{vgg19} in the BP4D and DISFA datasets, and achieves the best performance over all static-image-based approaches. For better representation of AU features, we conduct our remaining experiments on the basis of AR_{res} features.

Table 8. **Control experiments for DISFA**. Results are reported using F1 score on 3-fold protocol.

AU	CNN_{res}	AR_{vgg16}	AR_{vgg19}	AR_{res}	AR_{mean_box}	AR_{FPN}	$\text{AR}_{ConvLSTM}$	$\text{AR}_{2stream}$	AR_{CRF}
1	26.3	24.9	26.9	32.1	31.3	[**39.9**]	26.9	**34.3**	24.1
2	23.4	23.5	21	25.9	**28.3**	[**33.3**]	24.4	27.4	26.5
4	51.2	55.5	**59.6**	[**59.8**]	59.3	59.3	58.6	59.4	51.7
6	48.1	51	56.5	55.3	55.4	49.3	49.7	[**59.8**]	**57.8**
9	29.9	41.8	[**46**]	39.8	38.4	32.5	34.2	**42.1**	33
12	**69.4**	68	67.7	67.7	67.7	65.5	[**71.3**]	65	65.5
25	**80.1**	74.9	79.8	77.4	77.2	72.6	[**83.4**]	77.4	71
26	52.4	49.4	47.6	52.6	**52.8**	47.9	51.4	50.1	[**53.5**]
Avg	47.6	48.6	50.7	**51.3**	**51.3**	50	50	[**51.9**]	47.9

Table 9. Average bounding box area in **BP4D**

AU group	# 1	# 2	# 3	# 5	# 7	# 8
AU index	1,2,7	4	6	10,12, 14,15	17	23,24
Avg box area	17785	46101	54832	103875	42388	38470
Area proportion	6.8%	17.6%	20.9%	39.6 %	16.2 %	14.7 %

Table 10. Average bounding box area in **DISFA**

AU group	# 1	# 2	# 3	# 4	# 5	# 6
AU index	1,2	4	6	9	12	25,26
Avg box area	17545	45046	46317	48393	78131	69624
Area proportion	6.7%	17%	17.7%	18.5 %	29.8 %	26.6 %

We further evaluate the inference time of our approach, LCN (CNN with locally connected layer[49]) and ROI-Nets on a Nvidia Geforce GTX 1080Ti GPU. We run each network for 20 trails over 1000 iterations with mini-batch size sets to 1; then we evaluate the running time for each iteration, and finally computed the mean and standard deviation over the 20 trials. The inference time is showed in Table 11, we can see our approach benefits from the RoI pooling layer's parallel computing over multiple bounding boxes, its inference time is lower than LCN and ROI-Nets. The RoI-Nets adopt 20 individual conv layers for 20 bounding boxes, thus it results worst performance.

4.3.1. AU R-CNN + Mean Box

The computation of each image's precise landmark point location is time consuming. We believe it is enough to use a "mean" AU bounding box coordinate to represent all images' bounding box. In this section, we collect the average coordinates of the overall images of nine AU group bounding boxes in each database to form a unified "mean box" across all images (Table 12 and Table 13). We use this "mean box" coordinates to replace the real bounding box coordinates calculated from the landmark in each image to evaluate. The experiment results are shown in Table 5 and Table 8, denoted as AR_{mean_box}. Although most images of BP4D and DISFA dataset are the frontal face, the deviation of mean bounding box coordinates from real box location exists. However, the F1 score is remarkably close to AR_{res}, because the RoI pooling layer in AU R-CNN performs a coarse spatial quantization. This performance similarity demonstrates that AU R-CNN is robust to small landmark location error, and the computation consumption of each image's landmark can be saved via using "mean box".

Table 11. Inference time(ms) of VGG-19 backbone on 512×512 images

Ours	ROI-Nets[3]	LCN[49]
27.4 ± 0.0005	67.7 ± 0.0004	34.7 ± 0.008

Table 12. Mean box coordinates of 512×512 resolution images in **BP4D**

AU group	AU index	mean boxes coordinates (y_{min}, x_{min}, y_{max}, x_{max} format)
# 1	1,2,7	(30.4, 58.1, 140.3, 222.5), (30.1, 297.2, 140.9, 456.5)
# 2	4	(23.9, 57.8, 139, 455.9)
# 3	6	(109.4, 79.8, 264.5, 431.8)
# 5	10,12,14,15	(198.9, 35.2, 437.0, 472.6)
# 7	17	(378.7, 94.5, 510.9, 416.6)
# 8	23,24	(282.7,145.5,455.0,368.3)

4.4. AU R-CNN + Feature Pyramid Network

In the previous sections, we use the single scale ($16\times$ smaller scale) RoI feature to detect. Feature Pyramid Network (FPN)[46] is a popular architecture for leveraging a CNN's pyramidal features in the object detection field, which have semantics from low to high levels. In this experiment, FPN is integrated into AU R-CNN's backbone as feature extractor that extracts RoI features from the feature pyramid. The assignment of an RoI of width w and height h to the level k of FPN is as follows [46]:

$$k = \lceil k_0 + \log_2(\sqrt{wh}/224) \tag{2}$$

The experiment results (denoted as AR_{FPN}) are shown in Table 5 and Table 8. The AR_{FPN} performs worse than the single-scale RoI feature counterpart AR_{res}. This is because AU R-CNN needs high-level RoI features to classify AUs well and does not need to perform box coordinate regression. Furthermore, the bounding boxes in AU R-CNN are not too small to detect compared with those in the object detection scenario. Therefore, pyramidal features are not needed in detection.

4.5. Static versus Dynamic

Can the previous state of facial expression action always improve AU detection? In this section, we conduct a series of experiments using the most popular dynamic models that are integrated into AU R-CNN, including $\text{AR}_{ConvLSTM}$, as described in Section 3.4, to determine the answer.

Table 13. Mean box coordinates of 512×512 resolution images in **DISFA**

AU group	AU index	mean box coordinates (y_{min}, x_{min}, y_{max}, x_{max} format)
# 1	1,2	(55.5, 71.3, 168.6, 220.0), (53.5, 277.6, 167.6, 431.4)
# 2	4	(48.5, 58.7, 165.1, 444.0)
# 3	6	(141.4, 86.7, 281.5, 418.9)
# 4	9	(107.8, 152.2, 348.8, 352.8)
# 5	12	(236.9, 53.5, 433.3, 454.4)
# 6	25,26	(316.4, 73.8, 511.0, 433.4)

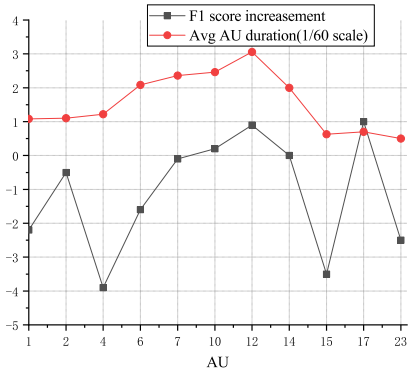


Fig. 9. Correlation between F1 score improvement of that in $\text{AR}_{ConvLSTM}$ over that in AR_{res} and AU activity duration, AU activity duration is rescaled presenting clarity.

4.5.1. AU R-CNN + ConvLSTM

Table 14. AU average activity duration & segments count in **BP4D**

AU	1	2	4	6	7	10	12	14	15	17	23	24
Avg Duration	65	66	73	125	142	148	184	120	**38**	42	**30**	49
Seg Count	474	380	408	540	569	591	448	571	647	1203	806	458

In this section, we conduct experiments on $\text{AR}_{ConvLSTM}$, whose architecture is described in Section 3.4. Table 5 and Table 8 present that $\text{AR}_{ConvLSTM}$ has a slightly lower average F1 score than AR_{res}. The main reason of the overall performance drop is that the action duration varies drastically in different AUs (Table 14 and Table 15); if the temporal length of AU duration is short, ConvLSTM model does not have sufficient capability to observe such actions. The switch of action is so rapid that ConvLSTM cannot infer such label change when processing in the video. We draw a plot of F1 score improvement of $\text{AR}_{ConvLSTM}$ over AR_{res} and average AU duration (rescale to 1/60 scale) in Fig. 9 to justify our hypothesis. Other factors also influence the performance of ConvLSTM, we can see the red line and the black line have strong correlation in most AUs except AU 1, 2, 4 and AU 15, 17, 23. The reason of high F1 score improvement in AU 17 is that AU 17 have much more segment count (1203) than AU 15 and AU 23 (Table 14), which results in sufficient training samples of AU 17. The AU 4 have lower F1 score improvement than that of AU 1, 2, because AU 4's bounding box (corresponding AU group #2) is twice times the size of AU 1 and AU 2 (Fig. 5), the larger bounding box leads to weaker recognition capability of capturing the subtle skin change between eyebrows. Most AUs do not have long activity duration; hence, AR_{res} surpasses $\text{AR}_{ConvLSTM}$ in average F1 score.

Table 15. AU average activity duration & segments count in **DISFA**

AU	1	2	4	6	9	12	25	26
Avg Duration	**55**	68	112	115	96	133	154	**82**
Seg Count	320	218	438	340	148	464	600	606

4.5.2. AU R-CNN + Two-Stream Network

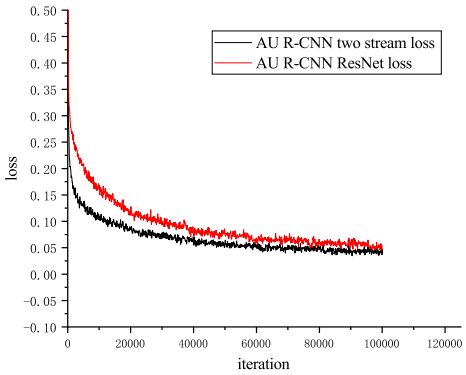


Fig. 10. AR_{res} vs. $\text{AR}_{2stream}$ train loss curve

Convolutional two-stream network[20] achieves impressive results in video action recognition. In this experiment, we experiment a two-stream network integrated into the AU R-CNN architecture for comparison, denoted as "$\text{AR}_{2stream}$". We use a 10-frame optical flow and a single corresponding RGB image[6], which are fed into two AU R-CNNs. Both AU R-CNN branches use the same bounding boxes, which is the corresponding bounding boxes of RGB image branch, for classification. Two produced $7 \times 7 \times 2048$ RoI features are concatenated along the channel dimension. The channel size of 4096 feature map is yielded, which will be reduced to 2048 channels using one kernel size of 1 convolutional layer. The features are sent to two fc layers to obtain the classification scores. The ground truth label involved in calculating the loss function adopts the single RGB image's labels.

The performance of the two-stream network $\text{AR}_{2stream}$ is remarkably close to that of RGB-image-based AR_{res}, which is slightly worse in the BP4D database (Table 5) and is better in the DISFA database (Table 8). In BP4D, the score significantly increases in AU 17 and AU 24 in AR_{res}. All these AUs are in the lip area. We attribute this result to the relative small area in the lip area causes the optical flow to be an obvious signal to classify. If we check the result in DISFA dataset in Table 8, this reason can be verified — AU 1, AU 6, and AU 9 in

[6]This corresponding RGB image is in the corresponding location that centers on 10 flow frames

Table 16. The features and applications of dynamic models extension

Model	Application	Training Speed	Feature
AR_{res}	most cases, no need for the video context	fast	high accuracy and universal application
AR_{FPN}	inappropriate	medium	low accuracy and has more layers than AR_{res}
$\text{AR}_{ConvLSTM}$	suitable for long AU activity duration case	slow	accuracy can be improved in long duration activities
$\text{AR}_{2stream}$	suitable for AUs in small sub-regions especially eye or mouth area	fast but need pre-compute optical flow	need pre-compute optical flow first
AR_{CRF}	application in the case of CPU only	medium and need pre-compute features	small model parameter size and no need to use GPU
AR_{TAL}	inappropriate	fast	the training cannot fully converged

the DISFA dataset have the smallest AU group areas (See Table 10.), and the F1 scores of these AUs increase. However, the performance of $\text{AR}_{ConvLSTM}$ in these AUs cannot be improved compared with $\text{AR}_{2stream}$. This justifies that AU group bounding box area is not the reason of the performance improvement in $\text{AR}_{ConvLSTM}$ but is the reason of performance improvement in $\text{AR}_{2stream}$. Although the average F1 score of $\text{AR}_{2stream}$ is worse than that of AR_{res} in the BP4D database, an interesting property exists in $\text{AR}_{2stream}$— the training convergence speed is faster than that in AR_{res} (See loss curve comparison in Fig. 10.).

4.5.3. AU R-CNN + TAL-Net

TAL-Net[21] follows the Faster R-CNN detection paradigm for temporal action location, and its goal is to detect 1D temporal segments in the time axis of videos. In this experiment, we regard video sequence as separate segments, and each segment has one label with it. We use the same RoI parallel line stream with the $\text{AR}_{ConvLSTM}$ because we want to detect each region's activity temporal segments. We reformulate the labels of segments in the AU video sequence as a label inside a start and end time interval. In TAL-Net, we use pre-computed $\text{AR}_{2stream}$ features. We stack 10 1-D 3×3 kernel conv layer on the 1-D feature map in the segment proposal network module to generate segment proposals, and we directly feed the pre-computed 1-D feature map into the SoI pooling layer and subsequent fc layers. This network is denoted as "AR_{TAL}".

In our experiment, we determine that AR_{TAL} cannot converge easily, and the loss can only decrease to approximately 1.3 at most (starting from approximately 2.7), which causes AR_{TAL} to perform worse than do $\text{AR}_{ConvLSTM}$ (Table 5). We can attribute this result to two reasons. First, facial expression is more subtle than the obvious human body action, and the temporal action localization mechanism cannot work efficiently. Second, training 1-D conv layer and fc layers requires millions of data samples, which cannot be satisfied when converting an entire video to a 1-D feature map. Therefore, this model has the worst performance among all dynamic models.

4.5.4. AU R-CNN + General Graph CRF

CRF model is a classical model for graph inference. We experiment with an interesting idea that involves connecting all separate parts of faces in a video to construct a spatio-temporal graph and then using the general graph CRF to learn from such a graph. This model is denoted as "AR_{CRF}". We not only connect RoIs with the same AU group number in the adjacent frames of the time axis but also fully connect different RoIs inside each frame, thereby yielding a spatio-temporal graph. In this method, the entire facial expression video is converted to a spatio-temporal graph using pre-computed 2048-D features extracted by AR_{res} (average pooling layer's output). This graph encodes not only the temporal dependencies of RoIs but also the spatial dependencies of each frame's RoIs. Table 5 and Table 8 present that AR_{CRF} has a lower score than does AR_{res} in BP4D and DISFA. We attribute this score decrease to the number of weight parameters. In AR_{CRF}, we have only $|\mathcal{F}| \times |\mathcal{Y}| + |\mathcal{E}| \times |\mathcal{Y}|^2$ weight parameters in total (in BP4D, it is $45{,}540$), where $|\mathcal{F}|$ denotes the feature dimension, $|\mathcal{Y}|$ denotes the class number, and $|\mathcal{E}|$ denotes the number of edge type. We extract 2048-D features from the average pooling layer. Two other fc layers exist on top of the average pooling layer in AR_{res}, and their weight matrices are 2048×1000 and 1000×22, which result in $2{,}070{,}000$ parameters that are much more than $45{,}540$ in AR_{CRF}. Therefore, classification performance is influenced not only by correlation but also by model capacity (including the number of learned parameters).

4.5.5. Dynamic models summary

After above discussion, the features and application of dynamic models extension can be summarized in Table 16.

5. Conclusion

In this paper, we present AU R-CNN for AU detection. It focuses on adaptive regional learning using expert prior knowledge, whose introduction provides accurate supervised information and fine-grained guidance for the model. Complete comparison experiments are conducted, and the results show that the presented model outperforms state-of-the-art approaches and the conventional CNN baseline model which uses the same backbone, which prove the benefit of introducing the expert prior knowledge. We also investigate various dynamic architectures that are integrated into AU R-CNN, which demonstrate that the static-image-based AU R-CNN outperforms all the dynamic models. Experiments conducted on the BP4D and DISFA databases manifest the effectiveness of our approach.

Acknowledgement

This research is partially supported by the National Key R&D Program of China (Grant No. 2017YFB1304301) and National Natural Science Foundation of China (Grant Nos. 61572274, 61672307, 61272225).

References

[1] P. Ekman, W. V. Friesen, Facial action coding system.

[2] A. Romero, J. León, P. Arbeláez, Multi-view dynamic facial action unit detection, arXiv preprint arXiv:1704.07863.

[3] W. Li, F. Abitahi, Z. Zhu, Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing, arXiv preprint arXiv:1704.03067.

[4] W. Li, F. Abtahi, Z. Zhu, L. Yin, Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 103–110.

[5] S. Eleftheriadis, O. Rudovic, M. Pantic, Multi-conditional latent variable model for joint facial action unit detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3792–3800.

[6] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE transactions on pattern analysis and machine intelligence 32 (11) (2010) 1940–1954.

[7] Z. Wang, Y. Li, S. Wang, Q. Ji, Capturing global semantic relationships for facial action unit recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3304–3311.

[8] W.-S. Chu, F. De la Torre, J. F. Cohn, Selective transfer machine for personalized facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3515–3522.

[9] X. Ding, W.-S. Chu, F. De la Torre, J. F. Cohn, Q. Wang, Facial action unit event detection by cascade of tasks, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2400–2407.

[10] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, Z. Xiong, Confidence preserving machine for facial action unit detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3622–3630.

[11] M. Liu, S. Li, S. Shan, X. Chen, Au-aware deep networks for facial expression recognition, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–6.

[12] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, J. F. Cohn, Fera 2015-second facial expression recognition and analysis challenge, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, Vol. 6, IEEE, 2015, pp. 1–8.

[13] K. Zhao, W.-s. Chu, S. Member, F. D. Torre, J. F. Cohn, H. Zhang, S. Member, Joint Patch and Multi-label Learning for Facial Action Unit and Holistic Expression Recognition 25 (8) (2016) 3931–3946. doi:10.1109/TIP.2016.2570550.

[14] K. Zhao, W.-S. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3391–3399.

[15] S. Han, Z. Meng, J. O'Reilly, J. Cai, X. Wang, Y. Tong, Optimizing filter size in convolutional neural networks for facial action unit recognition, arXiv preprint arXiv:1707.08630.

[16] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[17] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networksarXiv:1506.01497v3.

[18] W. Li, F. Abtahi, Z. Zhu, Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 00, 2017, pp. 6766–6775. doi:10.1109/CVPR.2017.716.
URL doi.ieeecomputersociety.org/10.1109/CVPR.2017.716

[19] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, 2015, pp. 802–810.

[20] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.

[21] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, R. Sukthankar, Rethinking the faster r-cnn architecture for temporal action localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1130–1139.

[22] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J. M. Girard, Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database, Image and Vision Computing 32 (10) (2014) 692–706.

[23] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, J. F. Cohn, Disfa: A spontaneous facial action intensity database, IEEE Transactions on Affective Computing 4 (2) (2013) 151–160.

[24] J. Whitehill, C. W. Omlin, Haar features for facs au recognition, in: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE, 2006, pp. 5–pp.

[25] B. Jiang, M. F. Valstar, M. Pantic, Action unit detection using sparse appearance descriptors in space-time video volumes, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 314–321.

[26] J. J. Bazzo, M. V. Lamar, Recognizing facial actions using gabor wavelets with neutral face average difference, in: Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, IEEE, 2004, pp. 505–510.

[27] M. Valstar, M. Pantic, Fully automatic facial action unit detection and temporal analysis, in: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, IEEE, 2006, pp. 149–149.

[28] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 94–101.

[29] J. J.-J. Lien, T. Kanade, J. F. Cohn, C.-C. Li, Detection, tracking, and classification of action units in facial expression, Robotics and Autonomous Systems 31 (3) (2000) 131–146.

[30] M. F. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42 (1) (2012) 28–43.

[31] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, D. Tao, Rethinking diversified and discriminative proposal generation for visual grounding, International Joint Conference on Artificial Intelligence (IJCAI).

[32] C. Fabian Benitez-Quiroz, R. Srinivasan, A. M. Martinez, Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5562–5570.

[33] Y. Wu, Q. Ji, Constrained Joint Cascade Regression Framework for Simultaneous Facial Action Unit Recognition and Facial Landmark Detection, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3400–3408doi:10.1109/CVPR.2016.370.

[34] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[35] Y. Song, D. McDuff, D. Vasisht, A. Kapoor, Exploiting sparsity and co-occurrence structure for action unit recognition, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, Vol. 1, IEEE, 2015, pp. 1–8.

[36] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, F. De la Torre, How much training data for facial action unit detection?, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, Vol. 1, IEEE, 2015, pp. 1–8.

[37] R. Hou, C. Chen, M. Shah, Tube convolutional neural network (t-cnn) for action detection in videos, in: IEEE international conference on computer vision, 2017.

[38] G. Chéron, I. Laptev, C. Schmid, P-cnn: Pose-based cnn features for action recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 3218–3226.

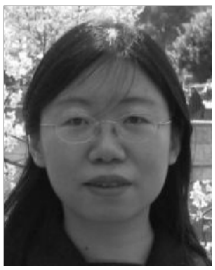
[39] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE Transactions on Neural Networks and Learning Systems (99) (2018) 1–13.

[40] W. S. Chu, F. D. L. Torre, J. F. Cohn, Learning spatial and temporal cues for multi-label facial action unit detection, in: IEEE International Conference on Automatic Face & Gesture Recognition, 2017, pp. 25–32.

[41] J. He, D. Li, B. Yang, S. Cao, B. Sun, L. Yu, Multi view facial action unit detection based on cnn and blstm-rnn, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 848–853.

[42] D. E. King, Dlib-ml: A machine learning toolkit, Journal of Machine Learning Research 10 (Jul) (2009) 1755–1758.
[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[44] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognitionarXiv:1409.1556v6.
[45] L. Zhong, Q. Liu, P. Yang, J. Huang, D. N. Metaxas, Learning multi-scale active facial patches for expression analysis, IEEE transactions on cybernetics 45 (8) (2015) 1499–1510.
[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: CVPR, Vol. 1, 2017, p. 4.
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.
[48] S. Jaiswal, M. Valstar, Deep learning the dynamic appearance and shape of facial action units, in: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE, 2016, pp. 1–8.
[49] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.

Chen Ma is currently pursuing the Ph.D. degree with the School of Software Department, Tsinghua University, Beijing, China. He received the Master degree from Beijing University of Posts and Telecommunications, in 2012. His research interests include facial expression an'alysis, action unit detection, and deep learning interpretability.



Li Chen received the Ph.D. degree in visualization from Zhejiang University, Hangzhou, China, in 1996. She is currently an Associate Professor with the Institute of Computer Graphics and Computer Aided Design, School of Software, Tsinghua University, Beijing, China. Her research interests include data visualization, mesh generation, and parallel algorithm.

Jun-Hai Yong is currently a Professor with the School of Software, Tsinghua University, Beijing, China, where he received the B.S. and Ph.D. degrees in computer science, in 1996 and 2001, respectively. He held a visiting researcher position with the Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, in 2000. He was a Post-Doctoral Fellow with the Department of Computer Science, University of Kentucky, Lexington, KY, USA, from 2000 to 2002. He received several awards, such as the National Excellent Doctoral Dissertation Award, the National Science Fund for Distinguished Young Scholars, the Best Paper Award of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, the Outstanding Service Award as an Associate Editor of the Computers and Graphics (Elsevier) journal, and several National Excellent Textbook Awards. His main research interests include computer-aided design and computer graphics.