Facial Action Unit Detection by Exploring the Weak Relationships Between AU Labels

Mengke Tian[1,2], Hengliang Zhu[3(✉)], Yong Wang[2], Yimao Cai[1], Feng Liu[4], Pengrong Lin[2], Yingzhuo Huang[2], and Xiaochen Xie[2]

[1] School of Integrated Circuits, Peking University, Beijing 100871, China
mtianaa@connect.ust.hk
[2] Beijing Microelectronics Technology Institute, Beijing 100076, China
[3] Fujian University of Technology, Fuzhou 350118, China
hengliang_zhu@fjut.edu.cn
[4] Ericsson Communications Co. Ltd., Beijing 100102, China

Abstract. In recent years, facial action unit (AU) detection attracts more and more attentions and great progress has been made. However, few approaches solve AU detection problem by applying the emotion information, and the specific influence of emotion categories to AU detection is not investigated. In this paper, we firstly explore the relationship between emotion categories and AU labels, and study the influence of emotion for AU detection. With emotion weak labels, we propose a simple yet efficient deep network that uses limited emotion labels to constraint the AU detection. The proposed network contains two architectures: a main net and an assistant net. The main net can learn semantic relation between AUs, especially the AUs related to emotions. Moreover, we design a dual pooling module embedded into the main net to further promote the results. Extensive experiments on two datasets show that the AU detection can obtain benefits with the weak labels of AUs. The proposed method has a significant improvement on baseline and achieves state-of-the-art performance compared with other methods. Furthermore, because only the main net is used for testing, our model is very fast and achieves over 278 fps.

Keywords: Action Unit(AU) detection · Emotion · Semantic relation

1 Introduction

Facial action unit detection plays an important role in various facial related tasks, such as expression analysis, interactive games, affective computing and behavioral science. Facial action unit is coded by the Facial Action Coding System (FACS) [1] and this system captures the slight different instant changes on facial appearance. Each AU depicts the movement of individual facial muscle, and some specific AUs can show high-level semantic expression when they are

H. Gao et al. (Eds.): CollaborateCom 2022, LNICST 461, pp. 478–495, 2022.
https://doi.org/10.1007/978-3-031-24386-8_26

co-occurrence. Action unit detection is a multi-label classification problem and lots of excellent works [2–4] solve this task by learning multiple AUs together. However, we observe that the different AU has different occurrence rate. It is difficult to train multi-label images by using the data with unbalanced distribution. On the other hand, the relation between all AUs is weak, but only specific AUs contain strong correlations, such as AU6 and AU12 define the happiness emotion. So both emotion categories and AUs can be used to represent facial behaviors. Though some researchers [3,5] applied the strategies of region enhancement for AU detection, the relations between emotion and AUs are ignored.

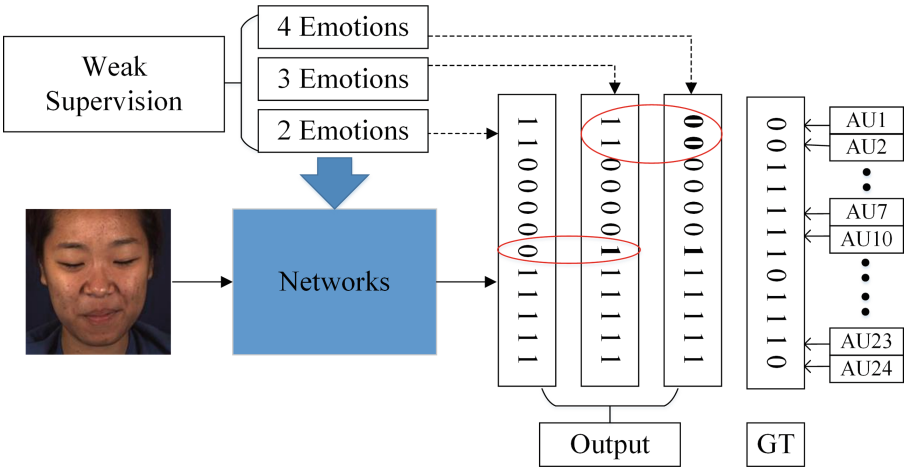

Fig. 1. Illustration of using different weak relationships. With more fine-grained emotion categories, more AUs are precisely detected. For example, AUs in red circles regions are corrected. (Color figure online)

In order to solve the above problems, the relations between emotion and AU labels are considered in this paper. Human beings have seven basic emotions (i.e., happiness, sadness, surprise, fear, anger, disgust and contempt) and each emotion is consisted of different AUs. From the Table 1, some specific AUs come from the same emotion and have a strong relationship between them. These co-occurrence AUs describe facial semantic information. Therefore, we explore the relations between emotion category and specific AUs to improve the performance of AU detection. However, the public AU databases have limited AU labels, for example, BP4D database provides only 12 AU labels, but there have 28 main facial AU codes. In other words, some emotions can not be identified due to the lacking of the AU labels. To alleviate this issue, we treat emotion categories as weak labels and propose a simple partition of emotion categories for AU dataset. We design three groups, including 2, 3 and 4 emotion categories. Take 4 emotion categories as example, three emotions can be ensured by using the provided AU

labels and other uncertain emotions are classified as one category. With emotion-level supervision, we propose a light-weighted architecture for AU detection. As shown in Fig. 1, with the weak relationships of limited emotion categories, the AU detection can obtain benefits. We can see that more fine-grained emotion categories, the results are better.

As mentioned above, the emotion categories can provide rich context among AUs. So how to efficiently use emotion categories to supervise facial AU multi-label learning is challenging. Deep learning based methods achieve superior performance and extract more discriminative features, especially in recognition tasks. In this paper, we proposed a efficient AU detection method to solve this problem. Because AU detection is sensitive to facial details and the pooling layer may lead to facial features loss, a new dual pooling module is presented in our framework. The module can maintain more useful information of AU regions. Furthermore, to further improve the accuracy of AU detection, we design an assistant net that inspired by the success of excellent works [6,7]. The assistant net reuses the features of shallow layer and serves as a supervision to the main net. This auxiliary constraint is beneficial to make the AU detection robust. Experimental results show that our proposed method performs well.

In the paper, by leveraging the limited facial emotion categories as weak labels, we propose a deep learning model to detect the facial AUs. The contributions of the paper are summarized as follows:

- We explored the relationship between the emotion categories and specific AUs, and depicted the benefit of AU detection with different number of emotion categories. To our knowledge, this is the first work that investigates the influence of the limited emotion labels for AU detection.
- We present a novel end-to-end deep learning network with high speed and accuracy, which consists of a main net and an assistant net. Moreover, a new dual pooling module is embedded in the network to preserve the more useful facial details.
- Experimental analysis shows the effectiveness of the proposed components in the whole network and demonstrates superiority over the state-of-the-art approaches.

Table 1. Emotion-related facial action units.

Emotion	Action Units
Happiness	6 + 12
Sadness	1 + 4 + 15
Surprise	1 + 2 + 5 + 26
Fear	1 + 2 + 4 + 5 + 20 + 26
Anger	4 + 5 + 7 + 23
Disgust	9 + 15 + 16
Contempt	12 + 14

2 Related Work

Automated AU detection has a noticeable improvement in recent years. Researchers extract different features from a face image to represent the desired semantic information. According to the difference of features, these methods are classified into conventional methods and deep learning methods.

2.1 Conventional Methods

In conventional approaches, various handcraft features are applied to represent a face and classical classifiers are used to learn the models. Some researchers extract features from the whole face image and some researchers extract features from regions around facial landmarks. Valstar et al. [8] applied Gabor wavelet features extracted around facial landmarks and learned the representation by Adaboost framework, and the final labels are classified by using SVM. Some works [9–11] also applied handcraft features extracted around facial landmarks. Jiang et al. [2] applied histograms of Local Phase Quantization (LPQ) to extract the discriminative features for AU detection. Zhao et al. [12] proposed a joint learning method that detecting the patch and AUs label simultaneously. By fusing the geometry information and multiple orientation Gabor features, Fabian et al. [13] achieved fast and accurate AUs detection performance. The patch is around the facial landmarks and SIFT features are extracted. Song et al. [14] addressed this topic by analyzing co-occurrence and the sparsity of action units. Wu et al. [4] proposed a new constraint to jointly learning the AU labels and facial landmarks localization. Girard et al. [15] exploited a regression framework to estimate the intensity of action unit regions and addressed this problem by employing linear partial least squares. In summary, these conventional methods made efforts by applying discriminative features and robust classifiers.

2.2 Deep Learning Methods

Due to the strong ability of learning discriminative features, deep learning is one of the most hot topics in the last few years, and CNNs have been used in almost all the computer version tasks. Inspired by the locally connected layer [16], Zhao et al. [5] proposed a region-based CNN method to capture structural information in different facial regions. In this method, a new region layer is proposed to divide the image into small patches and each patch is learned individually. In the last convolutional layer, these patches are combined into one image. EAC [3] proposed two novel nets to make the neural network to pay more attention to AUs interest regions to improve the accuracy. Corneanu et al. [17] learned the AU detection in two stages and these two stages are patch learning and structure learning. Han et al. [18] proposed an Optimized Filter Size (OFS) for AU detection. In this model, the filter size is alterable. That is to say when the model is learning AU labels, filters weights and sizes could be learned at the same time. Hao et al. [19] proposed a three-layer hybrid Bayesian network and expression information was used to assist the AU recognition. Shao et al. [20,21] use facial landmarks to extract the meaningful local

features for AU detection, but there network is very complex. Zhang et al. [22] proposed a method that leveraging prior probabilities of expression-independent and expression-dependent AU to learn the multiple AU classifiers. Hao et al. [19] captured the global relationships among AUs and expressions For BP4D, they used 4612 apex frames as samples in the experiment. Zhang et al. [22] proposed prior probabilities on AUs, including expression-independent and expression-dependent AU probabilities. They utilized 391 apex frames and 8 AUs on BP4D, while we use 12 AUs and total 146577 frames.

Different with the above two methods, we design the coarse emotion labels based on the AU labels and use emotion to supervise AU detection. Shao et al. [23] used the attention mechanism to capture the AU-related local features and pixel-level relations for AUs. Li et al. [24] utilized the self-supervised representation learning method to to encode the movements of AUs and head motions. Due to the lack of accurate annotations, Shao et al. [25] proposed an end-to-end unconstrained facial AU detection to deal with the situations with the unconstrained variability in facial appearances.

3 The Proposed Method

In Sect. 3.1, we introduce the generation of emotion categories and the weak relationships of AUs for AU detection. Then, the structure of the whole network is described in Sect. 3.2, as shown in Fig. 2. In Sect. 3.3, we design a dual pooling module and demonstrate its effectiveness.

3.1 Weak Relationships of AUs

AU detection is affected with various factors, such as pose, illumination, facial appearance and pedestrians identity. So how to improve the performance of facial

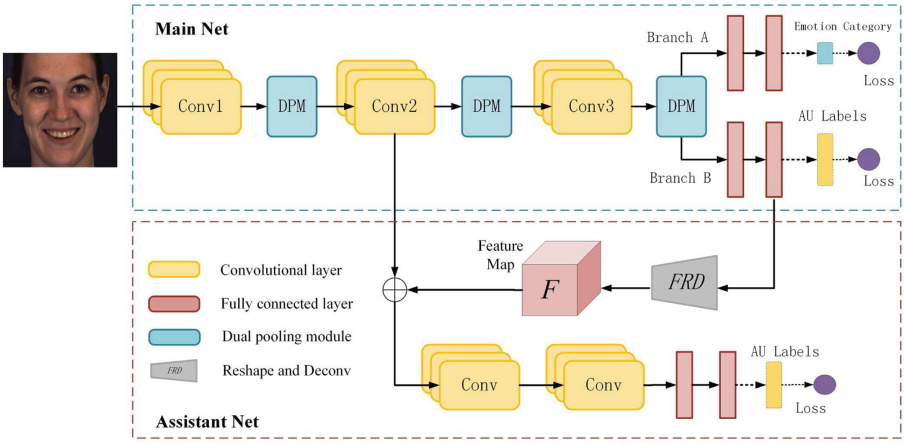

Fig. 2. The architecture of our AU detection model (best viewed in color). (Color figure online)

AU detection is a challenging problem. In order to improve the performance of AU detection, some researches focus on the facial region by applying the facial landmarks information. For example, [3,20,26] used facial landmarks to help the facial AU detection. However, these methods learn features that are sensitive to local regions without considering the global information of the face. We observe that facial emotions are produced by the change of muscle group, and can represent the global facial semantic features. Different group of AUs occurrence can generate different emotions. Their are some works [19,22] used expression information to assist AU detection. These works ignored the situation that how to use the image without emotion labels. In contrast, we regard the face without emotion label as hybrid emotion category.

Table 2. The 4 coarse emotion categories in the BP4D and DISFA datasets.

BP4D	Happiness	Sadness	Contempt	Hybrid
AUs	6 + 12	1 + 4 + 15	12 + 14	others
DISFA	Happiness	Sadness	Surprise	Hybrid
AUs	6 + 12	1 + 4 + 15	1 + 2 + 5 + 6	others

Without giving the ground truth labels on emotions, it is really hard to infer a definite label by only looking at the combination of AUs. We found that the facial emotion has close relations with AUs [27]. Emotion categories are easily defined as a unique set of AUs (see Table 1). For example, sadness contains AU1, AU4 and AU15. When these specific AUs co-occurrence, the face appears corresponding emotion and the semantic information of the face is involved. So we leverage the emotion categories to learn facial semantic features to provide a constraint for the AUs recognition. Because the labels of emotion are not provided in the BP4D and DISFA databases, the emotion categories are defined by using the combination of specific AUs. Based on the FACS, we design three groups of emotion category, including 2, 3 and 4 emotion categories. The detail partitions of 2 and 3 coarse emotion categories are described in Sect. 4.4. Here, we talk about the 4 coarse emotion categories, as shown in Table 2. It includes 3 emotions that can be identified and one hybrid emotion that can not be identified. These 4 coarse labels of emotion are used as weak relationships for the AU detection. Thus, our model can deal with the face images that emotions are uncertain.

Given the AU detection training dataset $S = (I^{(n)}, Y^{(n)}), n = 1, 2, ..., N$ with N training images, where I denotes an face image with ground-truth labels Y. Then we defined a vector $Y = [y_1, y_2, ..., y_C]$, y_i is a binary variable of each AU. We set $y_i = 1$ if the $i - th$ AU is occurrence in an image I, and $y_i = 0$ otherwise. C is the number of AU labels in the dataset S. Let the symbol $\mathbf{W}$ as the parameters of the network, $\widehat{Y} = [\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_C]$ denotes the detected results of AU labels.

For emotion supervision, we use a softmax layer to predict the probability of emotion. The loss of emotion recognition (branch A) is obtained by

$$\mathcal{L}_e(\mathbf{W}, \theta_e) = \varphi_{softmax}(I; \mathbf{W}, \theta_e) \tag{1}$$

where θ_e denotes the classifier parameter for the emotion recognition. The AU detection can be regarded as a multi-label binary classification problem, and we use the multi-label sigmoid cross-entropy loss for AU detection [5]. The loss of AU detection (branch B) is defined as follows.

$$\begin{aligned}\mathcal{L}_m(\mathbf{W}, \theta_m) = &- \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C}\xi(y_i) log P(\widehat{y_i}^{(n)}|I; \mathbf{W}, \theta_m) \\ &+ (1 - \xi(y_i)) log P(\widehat{y_i}^{(n)}|I; \mathbf{W}, \theta_m).\end{aligned} \tag{2}$$

where $\xi(\cdot)$ is a sign function, and it returns 1 when the $i-th$ AU is occurrence, otherwise returns 0. θ_m is the classifier parameter for the AU detection in the main net. $P(\widehat{y_i}|I; \mathbf{W}, \theta_m)$ denotes the confidence score of the $i-th$ AU detection. Note that, our method does not use external data and domain adaption technologies. We believe that this weak relationships method can also be used in other visual tasks with limited labels.

3.2 Overview of Framework

As shown in Fig. 2, the architecture of our AU detection model consists of three components: the main net, the dual pooling module and the assistant net. The main net is a task for AU detection and the assistant net applies weak relationships of AUs to provide a constraint to the AU branch.

Main Net for AU Detection. Our main net is very simple and contains three convolutional layers. Because the surroundings of AU are regional and subtle, the texture changes of these facial regions could influence the accuracy of different AU detection. We observe that when the multiple continual convolution layers are used, the details of facial information will be lost in the final output (the feature map). In order to hold sufficient facial information from its previous layer, we only use three convolutional layers in the main net, as illustrated in Fig. 2. The parameters of three convolutional layers are shared between the emotion classification and AU detection. The features collected from the shared convolutional layers across different semantic levels. At the end of the main net, we use two fully connected layers (FCL) to model the spatial correlations of the entire image. This designing is simple and efficient for features extraction.

The input image is a 200×200 RGB image, then it is sent to a convolutional layer (Conv1) with 96 filters, kernel size 11×11 and stride 4. The non-linear transformation (ReLU) is also used after each convolutional layer, and the outputs of Conv1 are size of 48×48 feature maps. For simplicity, the parameters of Conv1 are denoted as $96 \times 48 \times 48, 11 \times 11, 4$. The parameters of the main

Table 3. The parameter settings of the main net.

No	Layer	Parameters
0	Input	$200 \times 200 \times 3$
1	Conv1	$96 \times 48 \times 48, 11 \times 11, 4$
2	DPM	$96 \times 24 \times 24$
3	Conv2	$256 \times 24 \times 24, 5 \times 5, 2$
4	DPM	$256 \times 12 \times 12$
5	Conv3	$384 \times 12 \times 12, 3 \times 3, 1$
6	DPM	$384 \times 6 \times 6$

net can be found in Table 3. The Conv1 layer generates 96 feature maps, which are passed into a dual pooling layer (DPM, the blue box in Fig. 2). The DPM outputs 96 feature maps with size 24×24, and more details will be provided in Sect. 3.3. In the main net, we use three convolutional layers and three DPMs to avoid losing too much facial features. Finally, two branches behind the last DPM with each two fully connected layers to capture the global spatial information across the input image. Branch A is used for emotion classification, and branch B is used for AU detection. For each branch, the output of fully connected layers is a 4096 feature vector, which can extract the discriminative features.

Assistant Net with Feature Fusing. Previous works [28–30] indicate that the high-level semantic features help the category recognition of image, and the low-level visual features contribute to preserve detailed structures.

Motivated by the skip connections [31], we combine the deep layer and shallow layer to improve AU detection. We propose an assistant net to improve the ability of feature extraction in our model. The assistant net structure mainly contains two convolutional layers with kernel size 3 and stride 1, and two FCLs with size 4096. The FRD block in the assistant net contains a feature reshaper and deconvolution structure (the gray trapezoid in Fig. 2). The FRD reshapes the feature vector of FCL into $64 \times 8 \times 8$ feature maps. Then, the feature maps are up-sampled to the same size with the outputs of Conv2 layer. The Conv2 layer and feature map are combined to feed into the following convoltuional layer. The integrated feature maps $\boldsymbol{F}$ can be defined as

$$Comb = Concat(conv2, \boldsymbol{F}), \tag{3}$$

where $Concat$ is the cross-channel concatenation operator. By this way, the whole network can learn rich structure information to improve the feature representations around the AU regions. Experimental results show that the assistant net can promote the accuracy of AU detection.

The Total Loss. Our proposed deep model has three loss layers, as shown in Fig. 2 (the purple circle). The loss $L_a(\mathbf{W}, \theta_a)$ in the assistant net is similar to

the main net, where θ_a is the classifier parameter for the AU detection in the assistant net.

The total loss for AU detection can be written by

$$\begin{aligned}\mathcal{L}_{final}(\mathbf{W},\theta) =& \alpha_1\mathcal{L}_m(\mathbf{W},\theta_m) + \alpha_2\mathcal{L}_a(\mathbf{W},\theta_a) \\ &+ \alpha_3\mathcal{L}_e(\mathbf{W},\theta_e).\end{aligned} \tag{4}$$

where α_i is the loss weight to balance the loss of each task. In our experiment, we set $\alpha_i = 1, i = 1, 2, 3$. To solve above the loss function, we utilize the stochastic gradient descent (SGD) algorithm to get the optimal values.

3.3 Dual Pooling Module

Facial action units detection is sensitive to details of the face, in order to extract more useful features, we design a dual pooling module that can retain more detail features. The size of output feature map after convolutional layer is much smaller than the input, for example, the size of the image changed from 200×200 to 48×48 via the Conv1 layer. Therefore, lots of facial features lost in this process.

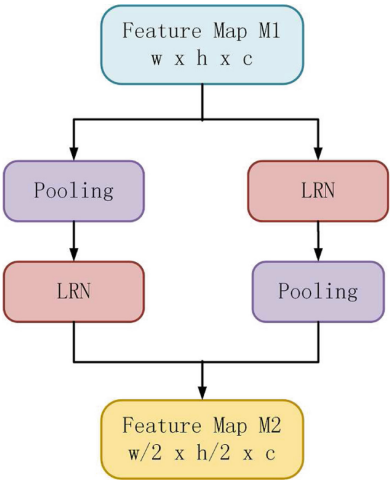

Fig. 3. The architecture of the dual pooling module. The DPM contains two components, and each one takes feature maps with $w \times h \times c$ resolutions as input. Then pooling and response normalization operators down-sample the feature maps to the same spatial size. Finally, the concatenation and ReLU non-linearity layer are used to output the integrated feature maps.

In order to transfer more facial details to the following convolutional layer, we embed the DPM into the main net to improve the performance of AU detection, the structure of DPM is displayed in Fig. 3. The module consists of two complementary and symmetric components. Each component contains one 3×3

pooling layer with stride 2 and one Local Response Normalization (LRN) layer with local size 5. Given the input feature maps $\mathbf{F(I)}$ with size $t = [w \times h \times c]$, the size of output feature maps is $r = [w/2 \times h/2 \times c]$. Thus, the integrated feature maps are generated by

$$Comb = Concat(P_l(\mathbf{F(I)}; \Omega_l), P_r(\mathbf{F(I)}; \Omega_r)) \tag{5}$$

where $P_i(\mathbf{F(I)}; \Omega_i)$ denotes left or right components operator with parameters Ω_i, $i \in \{l, r\}$. The function $P_i(\cdot; \cdot)$ helps to down-sample and normalize the input high-resolution feature maps.

By applying the DPM, more facial features can be retained to boost the performance in AU detection. As illustrated in Fig. 4, without the DPM, the feature maps lost much information (the middle column). When using the DPM, more facial information are preserved (the third column). Experimental results show that the proposed DPM conspicuously improves the accuracy of AU detection by 0.8% on BP4D dataset, as list in Table 8. We think that the DPM module can be applied into other detail-sensitive visual tasks.

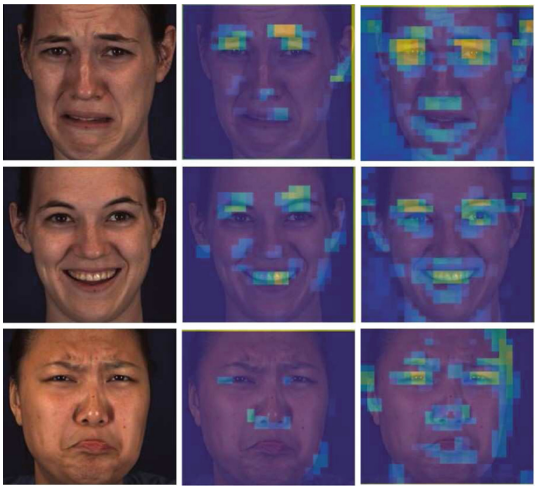

Fig. 4. Visualization of feature maps with or without DPM, the middle column without DPM and the third column is using DPM.

4 Experiments and Results

4.1 Settings

Datasets: The public databases used in this paper are BP4D [32] and DISFA [33]. The AU databases are difficult to obtain due to huge manual labeling work is needed. Here we give a brief review of these two AU databases.

BP4D: There are 41 participants and each participant is involved in 8 sessions that capture both 2D and 3D videos. More than 140000 frames can be obtained from these 328 videos, and 49 facial landmarks are provided to crop the face from the original images. For fair comparison, 12 AUs are evaluated and 3-fold cross validation are conducted like DRML [5] and EAC [3].

DISFA: This dataset contains 27 participants including 12 women and 15 men, each of participant has 4845 frames. Facial landmarks and AU intensities are provided. The AU intensities are from 0 to 5 and we use the images with intensities equal or over 2 like DRML [5]. There are more than 100000 images and we use 58000+ images. Similar to EAC [3], we use the pre-trained model from BP4D and fine-tuning on this database. The 3-fold cross validation is conducted and 8 AUs are evaluated.

Evaluation Metrics: Two metrics: the F_β-score and the average accuracy (%) are used to measure the performance of AU detection. We compute the performance with accuracy (%) following the previous work [3]. F_β-score metric is widely used in AU detection [9,34], and it consists of two items: precision and recall. The F_β-score of each AU label is given by

$$F_\beta = \frac{(1+\beta^2) \times precision_i \times recall_i}{\beta^2 \times precision_i + recall_i} \tag{6}$$

Following the previous works [3,5], we set $\beta = 1$ in our experiment where recall and precision are treated as equally relevant. Then we get F1-score evaluation for all AUs. In addition, we compute the average results over all AU labels. For simplicity, we omit % in all the results in our experiments.

Table 4. F1-score results on BP4D database. The best results are shown in **bold**.

AU	LSVM	JPML	DRML	CPM	EAC	ROI	DSIN	OFS	Ours
1	23.2	32.6	36.4	43.4	39.0	36.2	**51.7**	41.6	45.6
2	22.8	25.6	**41.8**	40.7	35.2	31.6	40.4	30.5	**41.8**
4	23.1	37.4	43.0	43.3	48.6	43.4	**56.0**	39.1	54.6
6	27.2	42.3	55.0	59.2	76.1	77.1	76.1	74.5	**78.5**
7	47.1	50.5	67.0	61.3	72.9	**73.7**	73.5	62.8	73.4
10	77.2	72.2	66.3	62.1	81.9	**85.0**	79.9	74.3	82.0
12	63.7	74.1	65.8	68.5	86.2	87.0	85.4	81.2	**87.7**
14	64.3	**65.7**	54.1	52.5	58.8	62.6	62.7	55.5	62.2
15	18.4	38.1	33.2	36.7	37.5	**45.7**	37.3	32.6	38.9
17	33.0	40.0	48.0	54.3	59.1	58.0	**62.9**	56.8	61.7
23	19.4	30.4	31.7	39.5	35.9	38.3	38.8	41.3	**43.6**
24	20.7	42.3	31.0	37.8	35.8	37.4	41.6	–	**47.3**
AVG	35.3	45.9	48.3	50.0	55.9	56.4	58.9	53.7	**59.8**

Implementation Details: To train a deep learning model, we need larger numbers of face images. Similar to DRML's experiment settings [5], we choose BP4D to train our model. We first split the dataset into 3 folds, and each time two folds are used for training and the third fold for testing. A post-processing is used in our method. We use 3 emotions which can be ensured to judge the corresponding AU labels. This process can give 0.1% improvement on final results. In our network, all the weights of loss function are set to 1 without optimization. If the weights are optimized, the results can be further improved. We found that the distribution of AUs in the database was unbalance, and some AUs were much more than others. However, we did not do any data balancing operation for the training, but directly trained on the original data. Our network could still achieve good results in the case of unbalance distribution of data samples.

For each face image, we crop and scale the original image into a $200 \times 200 \times 3$ image. In order to enhance the diversity of training data, horizontally flipping is used. We train our model with an open source deep learning framework Caffe [35], and directly feed the input images into the network. The proposed network is trained on an Intel Core computer with an i7-6850K CPU and a single GeForce GTX 1080Ti GPU. In our experiments, the base learning rate is initialized with 0.0001, which is reduced after every 10000 iterations. We set the total number of iterations to 40000. The momentum is 0.9 and weight decay of 0.0005 is used.

Running Time: For training stage, it takes us about 2 h to train the deep model. Because the output of branch B is the final result of AUs detection, the structures of network that irrelevant to this output can be pruned in testing stage, such as branch A and the assistant net. Thus, the speed of AU detection is very fast. In testing, our network takes 0.0036s (278 FPS) to process an image (average 200×200).

4.2 Comparison with State-of-the-Art Methods

We compare our method with state-of-the-art methods in this Section, the compared approaches including LSVM [36], JPML [12], APL [37], DRML [5], CPM [38], EAC-Net [3], ROI [39], DSIN [17] and OFS [18]. LSVM [36], JPML [12] and APL [37] are conventional methods and other methods are deep learning-based methods.

Tables 4 and 5 show the F1-score and accuracy results of 12 AUs on BP4D database. We can see that our algorithm outperforms all other methods on this challenging database in term of average results. For some AUs, the performance is significantly improved. DSIN [17] and OFS [18] are the most recent works which utilized the CNN models and our method also give better results compared with them. OFS [18] only reports the average F1-score result of 11 AUs and average accuracy is 72.2%. DSIN [17] does not provide the accuracy results. Therefore, we do not show the AU accuracy result of DSIN [17]. In our method, some AUs are not as good as others, the reason may be is the unbalance of AUs distribution.

Table 5. Accuracy results on BP4D database. The best results are shown in **bold**.

AU	LSVM	JPML	DRML	EAC	Ours
1	20.7	40.7	55.7	68.9	**72.8**
2	17.7	42.1	54.5	73.9	**76.3**
4	22.9	46.2	58.8	78.1	**78.8**
6	20.3	40.0	56.6	78.5	**79.3**
7	44.8	50.0	61.0	69.0	**70.1**
10	73.4	75.2	53.6	77.6	**77.9**
12	55.3	60.5	60.8	84.6	**85.9**
14	46.8	53.6	57.0	60.6	**63.5**
15	18.3	50.1	56.2	78.1	**79.2**
17	36.4	42.5	50.0	70.6	**72.7**
23	19.2	51.9	53.9	81.0	**81.4**
24	11.7	53.2	53.9	82.4	**84.0**
AVG	32.2	50.5	56.0	75.2	**76.8**

Table 6. F1-score results on DISFA database.

AU	APL	DRML	EAC	DSIN	Ours
1	11.4	17.3	41.5	**42.4**	39.1
2	12.0	17.7	26.4	39.0	**65.2**
4	30.1	37.4	66.4	**68.4**	67.9
6	12.4	29.0	**50.7**	28.6	40.8
9	10.1	10.7	**80.5**	46.8	46.4
12	65.9	37.7	**89.3**	70.8	73.2
25	21.4	38.5	88.9	**90.4**	89.8
26	26.9	20.1	15.6	**42.2**	37.2
AVG	23.8	26.7	48.5	53.6	**57.4**

Tables 6 and 7 show the F1-score and accuracy results of 8 AUs on DISFA database. On this database, our method is the best in term of average results. For the average F1-score, the result of our method is increased by 7% compared with DSIN. For the average accuracy, our result is 85%, better than EAC-Net [3] (80%). In addition, OFS [18] uses 10 AUs to train and conducts 9 folds cross-validation on DISFA database, the F1-score is 55.3% and accuracy is 85.0%. From the Table 7, we also see that the results of some AUs (i.e., AU1, AU2, AU6, AU25 and AU26) are greatly improved when applying our method.

Table 7. Accuracy results on DISFA database.

AU	APL	DRML	EAC	Ours
1	32.7	53.3	85.6	**90.5**
2	27.8	53.2	84.9	**94.5**
4	37.9	60.0	**79.1**	76.8
6	13.6	54.9	69.1	**79.4**
9	64.4	51.5	88.1	**91.1**
12	94.2	54.6	**90.0**	84.1
25	50.4	45.6	80.5	**88.1**
26	47.1	45.3	64.8	**75.5**
AVG	46.0	52.3	80.6	**85.0**

4.3 Ablation Study

In this paper, we propose three components to improve the performance of AU detection and each component shows a benefit to the whole process. In order to show the affect of each component for AU detection, we conduct our experiments on BP4D database. Notably, we follow the three subsets partition of DRML [5]. The subset {1,2} are used for training and the subset {3} is used for testing. For simple notation, we define as follows: the main net without DPM and branch A is used as baseline (BL), the main net without DPM as MT-net, the main net as M-net, the whole network as Final-net. Table 8 shows the results of our experiments. From the table, we can see that each component gives an improvement to the results. For example, when using emotion categories, the result of MT-net increases 1.6 point compared with BL. When both the emotion categories and DPM are used, the accuracy of M-net is 58.4, achieving 0.8 point improvement over the MT-net. The assistant net rises about 1.3 point based on M-net, reaching 59.7% of accuracy. Therefore, both emotion supervision, DPM module and assistant net have significant contributions to the whole framework.

4.4 Analysis and Future Work

In this paper, we propose a post-processing that uses results of emotion to ensure the labels of related AUs. Though this step gives little improvement, it is useful and we will integrate it into the network to make the whole procedure full automation. The accuracy of emotion classification are 84.6% and 77.7% for DISFA and BP4D databases respectively. The accuracy result of BP4D is not good, due to the unbalance of emotion category. Over half of the images on BP4D database are hybrid emotion category, and images of other three emotion categories are less. That is to say, only limited part of data are benefited from branch A in the main net. We will use other methods to further divide the images belonging to the hybrid emotion category.

Table 8. F1-score results with different components in our model.

AU	BL	MT-net	M-net	Final-net
1	41.8	49.7	45.8	50.3
2	33.9	31.1	40.9	42.7
4	55.8	55.4	56.5	51.3
6	79.6	78.8	80.3	79.9
7	66.4	68.1	69.5	66.1
10	86.9	86.4	85.4	85.9
12	89.1	87.7	88.8	87.9
14	59.9	62.3	58.5	63.9
15	32.0	37.5	37.2	32.7
17	56.8	59.7	60.0	63.0
23	32.6	34.6	37.4	47.5
24	36.6	40.0	40.0	45.3
AVG	56.0	57.6	58.4	59.7

5 Conclusion

In this paper, we defined the label of emotions by using the combination of specific AUs. Then, a light-weighted deep network was proposed to apply the weak relationships to constraint AU detection. Specifically, we observed that with more emotion categories, the performance of AU detection was better. We also designed an assistant net and proposed a dual pooling module to be embedded in the main net. The DPM can protect the detailed facial structure and the assistant net can further improve the main net to capture semantic information. Experimental results show that our method is effectiveness for AU detection. Moreover, the proposed network is very simple and runs in real time, so our model can be applied for facial related tasks in mobile applications.

References

1. Ekman, P., Rosenberg, E.L.: What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (facs), 2nd ed. (2005)
2. Jiang, B., Martínez, B., Valstar, M.F., Pantic, M.: Decision level fusion of domain specific regions for facial action recognition. In: 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, 24–28 Aug 2014, pp. 1776–1781 (2014)
3. Li, W., Abtahi, F., Zhu, Z., Yin, L.: EAC-net: Deep nets with enhancing and cropping for facial action unit detection. IEEE Trans. Pattern Anal. Mach. Intell. **40**(11), 2583–2596 (2018)
4. Wu, Y., Ji, Q.: Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 3400–3408 (2016)

5. Zhao, K., Chu, W.-S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 3391–3399 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778 (2016)
7. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 2261–2269 (2017)
8. Valstar, M.F., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2006, New York, NY, USA, 17–22 June 2006, p. 149 (2006)
9. Eleftheriadis, S., Rudovic, O., Pantic, M.: Multi-conditional latent variable model for joint facial action unit detection. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 Dec 2015, pp. 3792–3800 (2015)
10. Koelstra, S., Pantic, M., Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11), 1940–1954 (2010)
11. Wang, Z., Li, Y., Wang, S., Ji, Q.: Capturing global semantic relationships for facial action unit recognition. In: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, 1–8 Dec 2013, pp. 3304–3311 (2013)
12. Zhao, K., Chu, W.-S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 2207–2216 (2015)
13. Benitez-Quiroz, C.F., Srinivasan, R., Martínez, A.M.: Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 5562–5570 (2016)
14. Song, Y., McDuff, D., Vasisht, D., Kapoor, A.: Exploiting sparsity and co-occurrence structure for action unit recognition. In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, 4–8 May 2015, pp. 1–8 (2015)
15. Gehrig, T., Al-Halah, Z., Ekenel, H.K., Stiefelhagen, R.: Action unit intensity estimation using hierarchical partial least squares. In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, 4–8 May 2015, pp. 1–6 (2015)
16. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014, pp. 1701–1708 (2014)
17. Corneanu, C., Madadi, M., Escalera, S.: Deep Structure Inference Network for Facial Action Unit Recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 309–324. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_19
18. Han, S., Meng, Z., O'Reilly, J., Cai, J., Wang, X., Tong, Y.: Optimizing filter size in convolutional neural networks for facial action unit recognition. CoRR, vol. abs/1707.08630 (2017)

19. Hao, L., Wang, S., Peng, G., Ji, Q.: Facial action unit recognition augmented by their dependencies. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, 15–19 May 2018, pp. 187–194 (2018)
20. Shao, Z., Liu, Z., Cai, J., Ma, L.: Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 725–740. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_43
21. Shao, Z., Liu, Z., Cai, J., Ma, L.: Jaâ-net: joint facial action unit detection and face alignment via adaptive attention. Int. J. Comput. Vis. **129**, 1–20 (2021)
22. Zhang, Y., Dong, W., Hu, B.-G., Ji, Q.: Classifier learning with prior probabilities for facial action unit recognition. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 5108–5116 (2018)
23. Shao, Z., Liu, Z., Cai, J., Wu, Y., Ma, L.: Facial action unit detection using attention and relation learning. In: IEEE Transactions on Affective Computing, vol. PP, p. 1 (2019)
24. Li, Y., Zeng, J., Shan, S., Chen, X.: Self-supervised representation learning from videos for facial action unit detection. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10924–10933 (2019)
25. Shao, Z., Cai, J., Cham, T.-J., Lu, X., Ma, L.: Unconstrained facial action unit detection via latent feature domain. In: IEEE Transactions on Affective Computing, vol. PP, p. 1 (2021)
26. Devries, T., Biswaranjan, K., Taylor, G.W.: Multi-task learning of facial landmarks and expression. In: Canadian Conference on Computer and Robot Vision, CRV 2014, Montreal, QC, Canada, 6–9 May 2014, pp. 98–103 (2014)
27. Martinez, A.M.: Computational models of face perception. Curr. Dir. Psychol. Sci. **26**(3), 263 (2017)
28. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 3319–3327 (2017)
29. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision–ECCV 2014. ECCV 2014. LNCS, vol. 8689, pp. 818–833 Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
30. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 5188–5196 (2015)
31. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
32. Zhang, X., et al.: A high-resolution spontaneous 3D dynamic facial expression database. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–6 (2013)
33. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: DISFA: a spontaneous facial action intensity database. In: IEEE Transactions on Affective Computing, vol. 4, no. 2, pp. 151–160 (2013)
34. Valstar, M.F., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2006, New York, NY, USA, 17–22 June 2006, p. 149 (2006)
35. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: ACM MM, pp. 675–678 (2014)

36. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
37. Zhong, L., Liu, Q., Yang, P., Huang, J., Metaxas, D.N.: Learning multiscale active facial patches for expression analysis. IEEE Trans. Cybernetics **45**(8), 1499–1510 (2015)
38. Zeng, J., Chu, W.-S., De la Torre, F., Cohn, J.F., Xiong, Z.: Confidence preserving machine for facial action unit detection. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 Dec 2015, pp. 3622–3630 (2015)
39. Li, W., Abtahi, F., Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 6766–6775 (2017)