# Facial Action Unit Detection with Multilayer Fused Multi-Task and Multi-Label Deep Learning Network

**Jun He[1], Dongliang Li[1], Sun Bo[1*] and Lejun Yu[1]**
[1] College of Information Science and Technology, Beijing Normal University
No.19, XinJieKouWai St., HaiDian District,
Beijing 100875, P.R.China
[e-mail: tosunbo@bnu.edu.cn]
*Corresponding author: Sun Bo

*Received October 7, 2018; revised April 30, 2019; accepted June 4, 2019; published November 30, 2019*

## Abstract

Facial action units (AUs) have recently drawn increased attention because they can be used to recognize facial expressions. A variety of methods have been designed for frontal-view AU detection, but few have been able to handle multi-view face images. In this paper we propose a method for multi-view facial AU detection using a fused multilayer, multi-task, and multi-label deep learning network. The network can complete two tasks: AU detection and facial view detection. AU detection is a multi-label problem and facial view detection is a single-label problem. A residual network and multilayer fusion are applied to obtain more representative features. Our method is effective and performs well. The F1 score on FERA 2017 is 13.1% higher than the baseline. The facial view recognition accuracy is 0.991. This shows that our multi-task, multi-label model could achieve good performance on the two tasks.

***Keywords:*** facial action unit, multi-task learning, multi-label learning, multilayer fusion, deep learning

**http://doi.org/10.3837/tiis.2019.11.015** **ISSN : 1976-7277**

## 1. Introduction

In human day-to-day life and society life, interpersonal communication depends largely on facial expressions. Psychologist J.A. Russell believes that 55% of information in daily communication is conveyed through facial expressions [1]. American psychologist Paul Ekman described six basic facial expressions [2]: happiness, disgust, sadness, surprise, fear and anger. In the 1970s, Paul Ekman released the facial action coding system (FACS) [3], which defines more than 40 different facial action units (AUs). Since then, many scholars have explored facial expression recognition by establishing a relationship between AUs and basic expressions. This is challenging work compared with frontal facial AU detection. To handle multiple pose variations, we design a multi-task network capable of working on faces captured from multiple views, and jointly complete AU detection and facial view detection.

The basic process of AU detection includes pre-processing, facial feature extraction and classifier design. In pre-processing phase, the target face is detected by a trained binary classifier. Then some pre-processing is often performed on the face image. Next, facial features are extracted, which is absolutely vital for AU recognition. For a high-dimensional facial image matrix, the dimension of target vector is deduced greatly after feature extraction, so that the computation is greatly reduced. Finally, a classifier can be trained by those feature vectors extracted from training images. Then it can be used to recognize an AU in test images. Fig. 1 illustrates the flowchart of facial AU detection system.

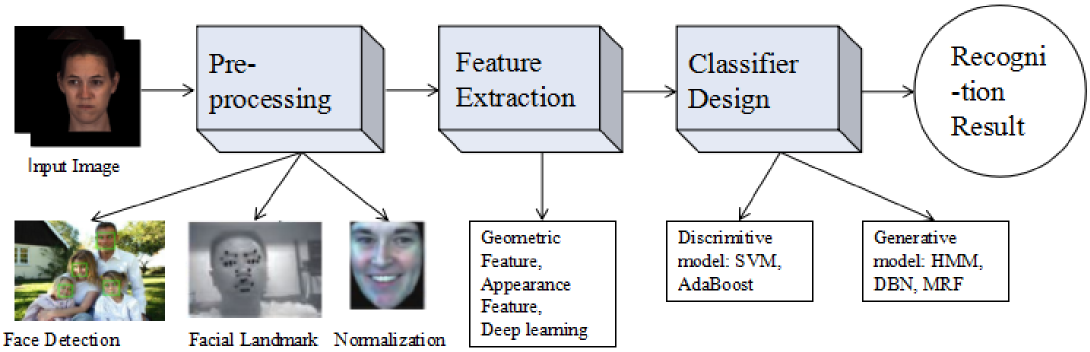


Fig. 1. Flowchart of facial AU detection system

AU detection has been widely used in many fields. In distance education, a teacher could know students’ learning effectiveness throughout facial action units, through which he could react accordingly. Additionally, facial action unit detection is necessary for workers under pressure doing hazardous work, e.g. for drivers’ fatigue detection. If the driver is found dozed off by monitoring his AUs, such as eyes, an alarm should be raised. If the driver is found sleeping, the car should be stopped forcibly and automatically. In a similar way, facial action unit detection could also be used for security or assisted medical care. Therefore, the research and development of human-computer interaction based on facial action unit analysis has attracted more and more attention from many scholars.

In this paper, we propose a multi-view facial AU detection approach using a novel multilayer fusion network based on multi-task and multi-label deep learning. We first designed a multi-task and multi-label network with a residual network [4] as the basic component for obtaining more representative features. Using this network, we trained a different network, by changing its structure and applying multilayer fusion to it. Based on these two networks, we produced a decision-level fusion. When we applied our method to the FERA 2017 database [5], we obtained outstanding results with the fewest networks. The contributions of our paper are as follows:

(1) Applying a deep residual network for facial AU detection. At present, most large-scale convolutional networks for AU detection are based on VGG16. We are the first to use a deep residual network to address this issue. Our experiments showed that a residual network could learn more representative features with fewer parameters.

(2) Using multi-task and multi-label learning in a network. Multi-label learning is applied for detecting different AUs. The multi-task learning is concentrated on simultaneously detecting AUs and facial views. Multi-label learning could jointly learn multiple AUs as one classification problem, while also considering AU correlations.

(3) Applying multilayer fusion to utilize information from different layers. In this method, we can simultaneously utilize strong spatial information regarding the convolutional layer and high-level abstract information regarding the fully-connected layer. This approach could enhance the performance of several AUs.

## 2. Related Work

Recently, the detection of facial AUs has arouse attention of many scholars. Facial Action Coding System (FACS) proposed by Ekman is the most widely used coding system in human behavior. FACS describes all possible facial expressions through a set of facial muscle anatomical movements called AU. FACS could describe and analyze facial expressions in greater detail.

AU is considered to be the smallest meaningful unit of the human face, and the combination of AUs can represent different expressions. Although FACS defines a limited number of human facial action units, but so far more than 7000 AU combinations have been found. Therefore, FACS is widely used in the study of human facial motion. **Fig. 2** illustrates some examples of facial action units.

As mentioned above, feature extraction is the key of the AU detection method, we discuss it in the view of its features. At present, the features used for AU detection can be divided into the following four categories: geometric features, appearance features, combining features, and deep-learning features.

Geometric features mainly describe the location of facial points and the shape of facial components. For example, the baseline method of FERA 2015 [6] used geometric features, which are calculated using 49 facial landmarks. The experimental results show that for most facial AUs the geometric features have better performance than appearance features. Yue Wu et al. used the position of facial points as the features for AU detection. They proposed the Constrained Joint Cascade Regression Framework (CJCRF) [7] to learn the relationship among facial action units and face shapes, and this model can enhance the performances of both facial action unit detection and facial landmark detection simultaneously. Arnaud Dapogny et al. [8] proposed local expression-driven features along with a facial mesh refined using an adaptive strategy based on facial points. Longfei Hao et al. [9] proposed hybrid Bayesian networks with multi layers whose bottom two layers are Bayesian Networks and top two layers consist of a latent regression Bayesian network. Their method uses facial points feature for AU recognition. Shangfei Wang et al. [10] proposed a new weakly supervised AU detection method based on facial landmark features to jointly learn multiple AU classifiers with expression labels without any AU labels by leveraging domain knowledge. Appearance features represent the facial texture including furrows, bulges and wrinkles. Common Appearance features include LGBP [11], HOG [12] etc. Because geometric and appearance features have their own advantages and disadvantages for AU detection, many researchers have combined them as new features. Thibaud Senechal et al. [13] used a multi kernel SVM model. Their method used both LGBP feature and AAM coefficients to detect facial AUs. Zuheng Ming et al. [14] used the fusion of the different geometry and appearance features. They also used a multi kernel SVM model to detect the intensities of facial AUs.

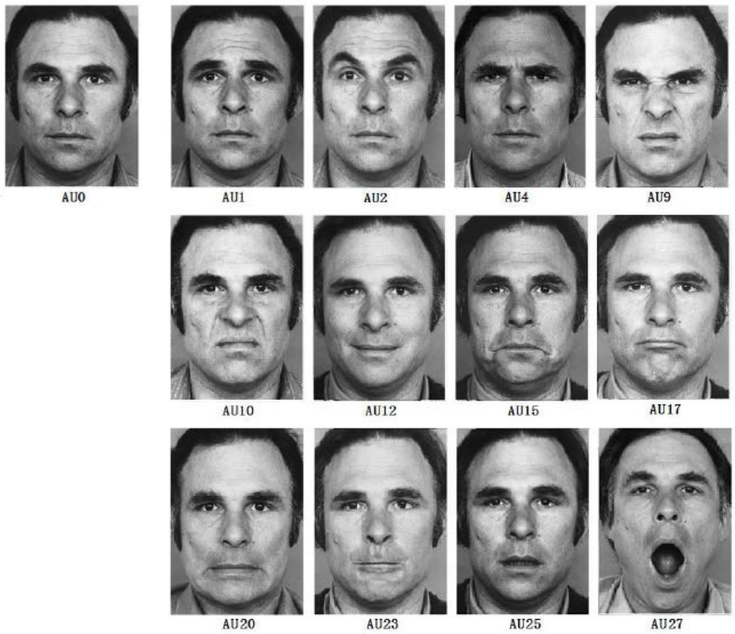


**Fig. 2.** Examples of facial action unit

Deep learning is an important branch of machine learning. Additionally, its excellent analytic and modeling capabilities provide new research directions for image classification. Because of these features, deep learning has also been used to detect AUs. Amogh Gudi et al. [15] proposed a convolutional neural network for facial AU occurrence and intensity detection. Their deep learning model consists of three convolutional layers, one pooling layer and one fully connected layer. It has achieved a better performance without any temporal information. Shashank Jaiswal et al. [16] proposed a CNN-LSTM model, which can learn the geometric, appearance and dynamics information jointly for AU detection. Their work proved that using temporal information throughout LSTM method could improve the performance of facial AU detection. Kaili Zhao et al. [17] presented a Deep Region and Multi-label Learning (DRML) method to detect facial action units. Their method solved two problems simultaneously, namely region learning (RL) and multi-label learning (ML). They considered that AUs are active on sparse facial regions, and RL just aims to identify these regions for a better specificity. In addition, multi-label learning (ML) attempted to jointly learn multiple AUs as one classification problem. It shows possibilities of utilizing AU correlations. DRML address both problems by construction, allowing these two seemingly unrelated issues to interact more directly. Also with the multi-label learning method, Sayan Ghosh et al. [18]

proposed a multi-label CNN network designed to learn a shared representation among multiple AUs directly from the training images. They obtained competitive results on CK+, DISFA and BP4D databases. Robert Walecki et al. [19] proposed a new Copula convolutional neural network method aimed to model multivariate ordinal variables. Their model explained the ordinal structure of output variables and their nonlinear dependencies throughout copula functions. Wei Li [20] proposed a deep learning framework for AU detection with region of interest adaptation and multi-label learning .In addition, their method also use LSTM in order to utilize temporal information.

In addition to these methods based on frontal face, multi view facial action unit detection has received increasingly more attentions. Zoltan Toser et al. [21] presented a method to complete AU detection over a wider range of head pose by using 3D information to augment video data, and used a deep learning method to extract the effective features for AU detection. Michel F. Valstar et al. [5] used tracked facial points locations as geometric feature and used Conditional Random Field (CRF) for temporal dynamics modelling. Xinrui Li et al. [22] utilized LBPTOP and CNN features throughout multiple classifiers including random forest and SVM to achieve multi view AU detection. Their method also encoded the temporal dynamics and static facial appearance respectively. Júlio César Batista et al. [23] proposed a unified CNN network named AUMPNet to perform AU detection and intensity estimation on facial images under multiple head poses. Jun He et al. [24] proposed CNN and BLSTM-RNN for multi view action units' detection. This method could utilize both visual features and temporal information. Chuangao Tang et al. [25] trained a network by fine tuning the VGG Face network on FERA 2017 multi view facial database in order to extract more effective features used for facial action unit detection.

Multi-task learning could solve several recognition tasks simultaneously by utilizing shared information. Multi-task learning has been used many fields such as facial landmark detection, pose estimation, action recognition, face detection and so on. Zhang Z et al. [26] presented a Tasks Constrained Deep Convolutional Network (TCDCN). TCDCN use several tasks consist of appearance attribute detection (such as wearing glasses), expression recognition, demographic detection (such as gender) and head pose detection. These tasks could enhance the performance of facial landmark detection. The result shows that their method could dealing with faces with head pose variation and severe occlusion. Junho Yim et al. [27] used multi task deep learning to rotate human face. The main task could rotate the face to a custom angle and the additional task rotate the face to 30 degrees. Their experiment shows that the additional task could improve the effect of main task. Cha Zhang et al. [28] proposed a multi task deep CNN network aimed to achieve the face/noface detection, face pose estimation and the facial landmark localization detection simultaneously.

## 3. Method

**Fig. 1** shows the architecture of our network, which is composed of two different structures. Network1 contains ResNet (see the green box portion of **Fig. 3**). Network1 is a multi-task network that has two branches as outputs. Network2 contains ResNet (see the orange box portion of **Fig. 3**) using multilayer fusion. Then, we use decision-level fusion for AU detection.

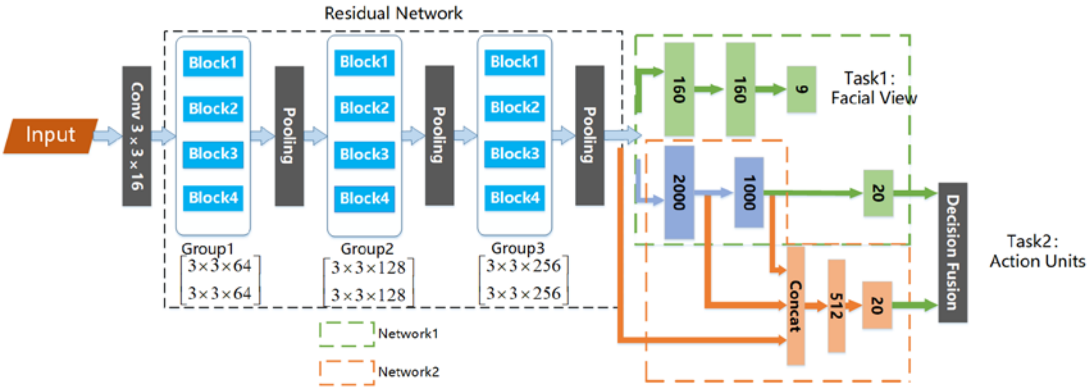


**Fig. 3.** Architecture of our multi-task and multi-label network. The size of the input image is 132 × 132. The activation function for all convolutional layers is ReLU. We use a Parametric ReLU after each fully connected layer, with a parameter was set to 0.3. Additionally, all of the pooling layers are average pooling layers.

### 3.1 Residual Network

A residual network is applied as the basic component. There are four groups of residual networks. In each group, there are four blocks, whose structure is shown in **Fig. 4**.

A residual block with identity mapping can be represented by the following formula:

$$x_{l+1} = x_l + F(x_l, w_l) \tag{1}$$

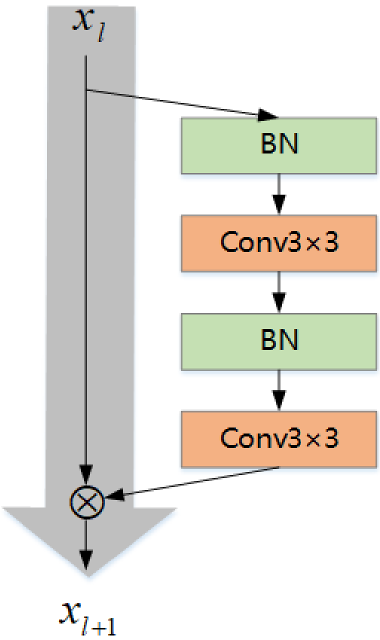


**Fig. 4.** Residual block architecture.

In this formula, $x_l$ and $x_{l+1}$ are input and output of the 1-th unit in the network respectively, F is a residual function and $w_l$ are parameters of the block. F represents the residual mapping to be learned. In our network, F has two consecutive 3×3 convolutions with batch normalization and ReLU activation function.

### 3.2 Multi-task and Multi-label Learning

After the residual network, we design and sequentially train two networks (Network1 and Network2). As is shown in **Fig. 1**, Network1 is a multi-task and multi-label learning network. We use multi-task learning (MTL) to improve the generalization performance of multiple related tasks by learning them jointly. Assuming that there are T tasks, the training data for the t-th task are denoted as $(X_i^t, y_i^t)$, where $t = \{1, \dots, T\}$, $i = \{1, \dots, N\}$, with $X_i^t \in R^d$ and $y_i^t \in R$ being the feature vector and label. The goal of multi-task learning is as follows:

$$\underset{\{w^t\}}{argmin} \sum_{t=1}^{T} \sum_{i=1}^{N} \ell\left(y_i^t, f\left(X_i^t; w^t\right)\right) + \Phi(w^t) \tag{2}$$

where $f\left(X_i^t; w^t\right)$ is a function of $X_i^t$ and is parameterized by as weight vector $w^t$. The loss function is denoted by $\ell(\cdot)$, which is a hinge loss. $\Phi(w^t)$ is the regularization term penalizing weight complexity.

The images can be considered to have two types of labels: facial AUs and facial views. We use a multi-output CNN to achieve multi-task learning. The network has two outputs: one for AU detection and one for facial view detection. Loss in our model is calculated by the following formula:

$$L = \lambda_1 L_1 + \lambda_2 L_2 \tag{3}$$

where $L$ is the main loss in our model, $L_1$ is the loss in facial action unit detection, and $L_2$ is the loss of facial view detection. $\lambda_1$ and $\lambda_2$ are two parameters, set manually, that are used to balance the two sub-tasks. $L_1$ is the multi-label cross-entropy loss (introduced in part 3.2) and $L_2$ is the softmax cross-entropy loss.

In Network1, after the residual network, there is a flattened layer that could reshape high-dimension inputs to a vector. Then, the network has two branches: one for facial view detection and another for AU detection. In the facial view detection branch, the output layer has only one node corresponding to the single view label. In the AU detection branch, the output is a multi-label layer, and the multi-label cross-entropy loss can be calculated using the following formula:

$$L\left(Y, \hat{Y}\right) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} \left\{[Y_{nc} > 0] log \hat{Y}_{nc} + [Y_{nc} < 0] log\left(1 - \hat{Y}_{nc}\right)\right\} \tag{4}$$

Where the number of AUs is C, the number of samples is N, and the ground truth $Y \in \{-1,0,1\}^{N \times C}$. $Y_{ij}$ is the $(i, j)$-th element of Y, and the predictions are $\hat{Y} \in R^{N \times C}$. We use 2 neurons for each label to avoid threshold selection. One represents positive activation, while the other represents negative activation. The larger value of the two neurons is chosen to represent whether or not the AU is activated. A dropout layer [29] is set after FC2 to suppress overfitting. Being trained, this network achieved good performance. In order to use all information for the convolutional layer and fully connected layer, we make some improvements in Network2.

### 3.3 Multilayer Fused and Decision Level Fusion

We found that the residual network outputs contained weak semantic information but strong spatial information. Additionally, the output of the FC layer contained high-level semantic and abstract information [30]. To fully use these two types of information, we trained Network2 based on the pre-trained residual network. In this step, with the pre-trained residual network in Network1, we only trained the last layers of the orange part in **Fig. 3**. Network2 has a multilayer fused structure. We obtained a concatenated layer by combining the last two FC layers (i.e. FC1 and FC2) and the residual network outputs. This branch is only used for AU recognition. Multilayer Fused method could use semantic information and spatial information. The results showed that several AUs could produce better performance in this new network.

After training Network2, we obtain two models. With these models, we perform decision level fusion for AU detection. For facial view detection, only the first model could produce excellent results, so fusion is unnecessary for this task. The decision fusion formula is as follows:

$$Y_k = sgn\left(\sum_{i=1}^{l} w_i y_{2k-1}^{(i)} - \sum_{i=1}^{l} w_i\, y_{2k}^{(i)}\right) \tag{5}$$

In this formula, $Y_k$ is the recognition result of the k-th AU, whose value is 0 or 1. The value of 0 indicates that AU has not occurred, while 1 indicates that AU has occurred. $y_k^{(i)}$ represents the output value of k-th neuron in the last layer in the i-th network. l is the number of fused models (which is 2 in our method). $w_i$ is the weight of i-th model.

## 4. Experiment

### 4.1 Database

In this paper, we tested our method using the FERA 2017 database. FERA 2017 database was used in FERA 2017 challenge. The FERA 2017 database has 3 parts: one training part, one validation part and one test part. The training part for FERA 2017 database is derived from BP4D-Spontaneous database [31], the validation part and test part are derived from a subset of BP4D+ database [32]. The training and validation data are publicly available and the test data is held back by the FERA 2017 organizers. There are 1326123 images for training and 679185 images for validation. FERA 2017 focus on 10 frequent occurrence facial AUs consist of AU1, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17 and AU23. These AUs were selected based on their frequency of occurrence and sufficiently high inter-rater reliability score. Different from the frontal view images in FERA 2015, the images in FERA 2017 database are generated by 9 different views under different camera views, as shown in **Fig. 5**. Therefore, the target of the FERA 2017 is to detect facial action units across different views, this increases AU detection difficulties relative to previous challenges.

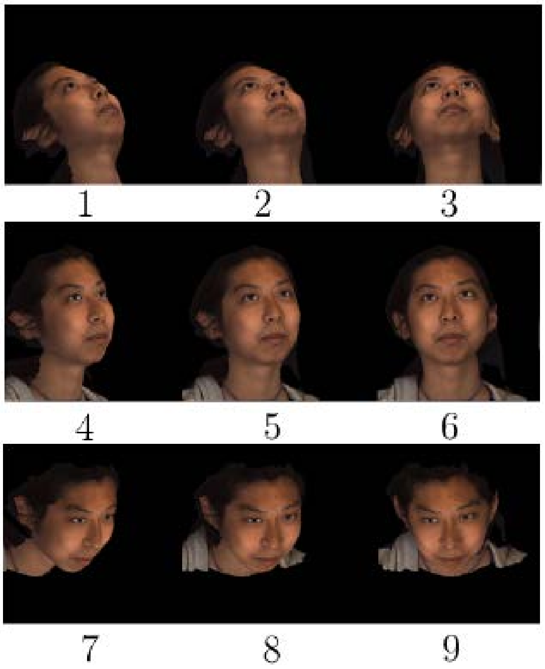


**Fig. 5.** Each of the different views of FERA 2017 database

## 4.2 Pre-processing

The first step of AU recognition is pre-processing. **Fig. 6** illustrates the flowchart of pre-processing. In order to reduce the influence of background, we use human face detection model to obtain face image and remove irrelevant background. After this step, some other pre-processing steps are required including resizing the image, image normalization and mean removal. When we get the face-crop image, we resize it so that the approximate scale of the face is constant. As a result, the face image is resizing to 132×132. In order to minimize the computational complexity of color images with multiple color channels, we change face image to grayscale image. In image normalization step, we normalize images and the pixel values are between zero and one. In mean removal step, we calculate the mean value of training part and remove it from both training part and test part. After pre-processing steps, we could minimize the computational complexity and achieve higher convergent rate.

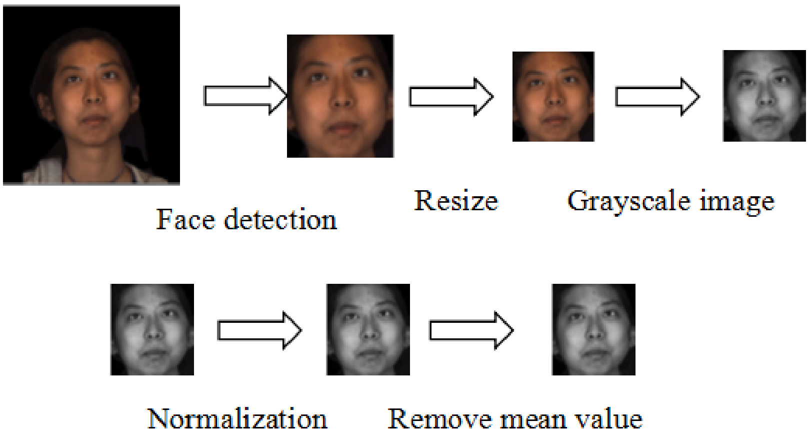


**Fig. 6.** The flowchart of pre-processing

### 4.3 Experiment Results

We evaluated performance using accuracy and an F1-measure, which is the harmonic mean of recall and precision. For an AU with precision P and recall R, F1 score is calculated as:

$$F_1 = \frac{2PR}{P+R} \tag{6}$$

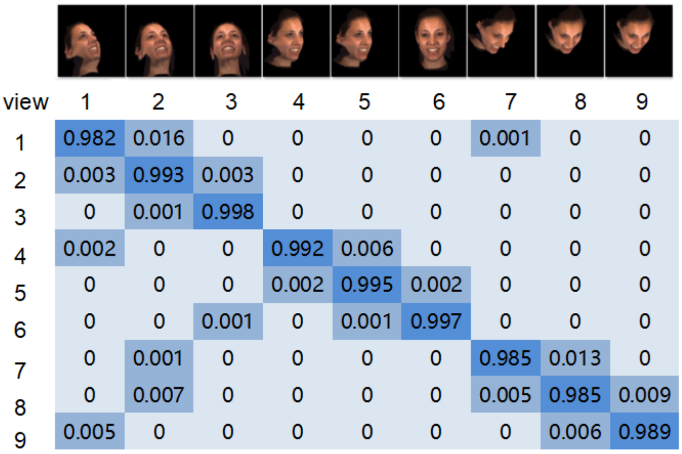
In our experiment, we set the multi-task parameters $\lambda_1$ and $\lambda_2$ in formula (3) to 0.5. Additionally, we set the decision-level fusion weight parameters $w_1$ and $w_2$ in formula (5) to 0.5. We employ the TensorFlow platform implementation for CNN training, and use the NVIDIA GTX1080Ti GPU device to train our models. The loss function is a categorical cross-entropy function and the optimizer used in the experiment was Adam. We set $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$.

The multi-view AU recognition results are shown in **Table 1**. The F1 score of our method (0.547) was 0.131 times higher than the baseline (0.416). In addition, the facial view recognition accuracy was 0.991, which shows that our model could recognize facial views with great accuracy. The confusion is shown in **Fig. 7**.

**Table 1.** F1 Score on FERA 2017 DATABASE

| **Action Unit** | **FERA 2017 Baseline** [5] | **Multi Feature Fusion** [22] | **AUMP Net**[23] | **CNN-LSTM** [24] | **Single-label VGG 16**[25] | **Our method** | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | **Network 1** | **Network 2** | **Decision Fusion** |
| **F1 Score** | | | | | | | | |
| AU1 | 0.154 | 0.288 | 0.345 | 0.369 | 0.304 | 0.306 | 0.340 | 0.357 |
| AU4 | 0.172 | 0.255 | 0.278 | 0.264 | 0.362 | 0.317 | 0.318 | 0.354 |
| AU6 | 0.564 | 0.600 | 0.677 | 0.678 | 0.712 | 0.729 | 0.714 | 0.744 |
| AU7 | 0.727 | 0.749 | 0.794 | 0.763 | 0.779 | 0.782 | 0.781 | 0.796 |
| AU10 | 0.692 | 0.751 | 0.785 | 0.801 | 0.836 | 0.810 | 0.798 | 0.818 |
| AU12 | 0.647 | 0.730 | 0.762 | 0.796 | 0.840 | 0.795 | 0.792 | 0.806 |
| AU14 | 0.622 | 0.606 | 0.692 | 0.664 | 0.697 | 0.609 | 0.649 | 0.636 |
| AU15 | 0.146 | 0.246 | 0.267 | 0.269 | 0.353 | 0.344 | 0.271 | 0.325 |
| AU17 | 0.224 | 0.284 | 0.364 | 0.366 | 0.442 | 0.424 | 0.378 | 0.420 |
| AU23 | 0.207 | 0.248 | 0.250 | 0.248 | 0.475 | 0.260 | 0.211 | 0.216 |
| Mean | 0.416 | 0.473 | 0.521 | 0.522 | 0.580 | 0.538 | 0.525 | 0.547 |

As is shown in **Table 1**, when our method was applied to the FERA 2017 database, it produced better results than other methods. Our method used only two networks to detect all ten AUs and all nine facial views. Network1 achieved a mean F1 score of 0.538. After combining three branches of layers, Network2 successfully improved the results for AU1, AU4 and AU14. As a result, it could achieve an F1 score of 0.547 after decision-level fusion. Although the single-label VGG16 produced better results than ours, their method uses ten VGG16 networks, which have nearly 1380 M, a huge number of parameters to train. Our networks used only 192 M parameters, greatly reducing model complexity and saving storage space relative to single-label VGG16 networks. Our method could achieve better performance with fewer parameters.



| view | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.982 | 0.016 | 0 | 0 | 0 | 0 | 0.001 | 0 | 0 |
| 2 | 0.003 | 0.993 | 0.003 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0.001 | 0.998 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.002 | 0 | 0 | 0.992 | 0.006 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0.002 | 0.995 | 0.002 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0.001 | 0 | 0.001 | 0.997 | 0 | 0 | 0 |
| 7 | 0 | 0.001 | 0 | 0 | 0 | 0 | 0.985 | 0.013 | 0 |
| 8 | 0 | 0.007 | 0 | 0 | 0 | 0 | 0.005 | 0.985 | 0.009 |
| 9 | 0.005 | 0 | 0 | 0 | 0 | 0 | 0 | 0.006 | 0.989 |

**Fig. 7.** Confusion matrix results for recognizing nine facial views

Besides F1 score, we also used accuracy as performance measurement. Because [22] and [23] did not include accuracy results in their paper, we compared our result with [21], [24] and [25]. As is shown in **Table 2**, the accuracy of our method is higher than others. Our decision fusion method achieve highest accuracy (0.817) which is 0.02 higher than Single-label VGG16 [25]. The result shows that our method has good performance.

**Table 2.** Accuracy on FERA 2017 DATABASE

| **Action Unit** | **FERA 2017 Baseline[5]** | **CNN-LSTM [24]** | **Single-label VGG 16[25]** | **Our method** | | |
|---|---|---|---|---|---|---|
| | | | | **Network 1** | **Network 2** | **Decision Fusion** |
| | | | **Accuracy** | | | |
| AU1 | 0.570 | 0.880 | 0.782 | 0.878 | 0.888 | 0.901 |
| AU4 | 0.520 | 0.811 | 0.808 | 0.847 | 0.871 | 0.870 |
| AU6 | 0.676 | 0.765 | 0.799 | 0.798 | 0.781 | 0.882 |
| AU7 | 0.642 | 0.679 | 0.737 | 0.738 | 0.735 | 0.807 |
| AU10 | 0.638 | 0.783 | 0.829 | 0.791 | 0.776 | 0.753 |
| AU12 | 0.660 | 0.812 | 0.860 | 0.804 | 0.801 | 0.798 |
| AU14 | 0.622 | 0.640 | 0.667 | 0.649 | 0.664 | 0.815 |
| AU15 | 0.307 | 0.836 | 0.806 | 0.906 | 0.903 | 0.666 |
| AU17 | 0.485 | 0.733 | 0.822 | 0.758 | 0.731 | 0.914 |
| AU23 | 0.373 | 0.757 | 0.862 | 0.870 | 0.861 | 0.758 |
| Mean | 0.549 | 0.770 | 0.797 | 0.804 | 0.801 | **0.817** |

## 5. Conclusion

In this paper, we present our work on AU detection for multiple facial views. We used multi-task and multi-label networks to detect AUs and facial views, and enabled the relationships to be learned between AUs and multiple tasks. According to our results, our method achieved good performance. Multi-label learning could jointly learn multiple AUs as one classification, and could utilize correlations between AUs to improve performance.

Multi-label learning also reduced the number of required models. In future research, the architectures of CNN models and sizes of temporal windows may be optimized for improving performance.

## References

[1] M. Pantic, L. J.M. Rothkranz, "Expert System for Automatic Analysis of Facial Expression," *Image and Vision Computing*, 18(11), 881-905, 2000. Article (CrossRef Link).

[2] Ekman P., "An argument for basic emotions," *Cognition & emotion*, 6(3-4), 169-200, 1992.

[3] Ekman P, Friesen W, "Facial Action Coding System," *Facial Action Coding System (FACS)*, 1978. Article (CrossRef Link).

[4] He, K., Zhang, X., Ren, S., Sun, J., "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016. Article (CrossRef Link).

[5] Michel F. Valstar, Enrique Sanchez-Lozano, Jeff F. Cohn, Laszlo A. Jeni, Jeff. M. Girard, Lijun Yin, Zheng Zhang, Maja Pantic, "FERA 2017 - Addressing Head Pose in the Third Facial Expression Recognition and Analysis Challenge," in *Proc. of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017. Article (CrossRef Link).

[6] Michel F. Valstar, T. Almaev, Jeff M. Girard, G. McKeown, M. Mehu, Lijun Yin, & Jeff F. Cohn, "FERA 2015 - Second Facial Expression Recognition and Analysis Challenge," in *Proc. of 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015. Article (CrossRef Link).

[7] Y. Wu and Q. Ji, "Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. Article (CrossRef Link).

[8] Dapogny, A., Bailly, K., & Dubuisson, S., "Multi-Output Random Forests for Facial Action Unit Detection," in *Proc. of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 135-140, May 2017. Article (CrossRef Link).

[9] Hao, L., Wang, S., Peng, G., & Ji, Q., "Facial action unit recognition augmented by their dependencies," in *Proc. of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 187-194, May 2018. Article (CrossRef Link).

[10] Wang, S., Peng, G., Chen, S., & Ji, Q., "Weakly Supervised Facial Action Unit Recognition with Domain Knowledge," *IEEE transactions on cybernetics*, 48(11), 3265-3276, 2018. Article (CrossRef Link).

[11] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *Proc. of IEEE International Conference on Computer Vision*, 1, 786– 791, 2005. Article (CrossRef Link).

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Computer Vision and Pattern Recognition*, vol. l, pp. 886-893, 2005. Article (CrossRef Link).

[13] Thibaud Senechal, Vincent Rapp, Hanan Salam, Renaud Seguier, Kevin Bailly, and Lionel Prevost, "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning," *Systems Man and Cybernetics Part B: Cybernetics*, 42(4), 993-1005, 2012. Article (CrossRef Link).

[14] Z. Ming, A. Bugeau, J.-L. Rouas, T Shochi, "Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine Automatic face and gesture recognition (FG)," in *Proc. of 2015 11th IEEE International Conference and Workshops on 6, IEEE*, pp. 1–6, 2015. Article (CrossRef Link).

[15] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis, "Deep learning based facs action unit occurrence and intensity estimation," in *Proc. of Facial Expression Recognition and Analysis Challenge, conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015. Article (CrossRef Link).

[16] Shashank Jaiswal, Michel Valstar, "Deep Learning the Dynamic Appearance and Shape of Facial Action Units," in *Proc. of WACV*, 2016. Article (CrossRef Link).

[17] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391– 3399, 2016. Article (CrossRef Link).

[18] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency, "A multi-label convolutional neural network approach to cross-domain action unit detectionm," *Affective Computing and Intelligent Interaction(ACII)*, 2015. Article (CrossRef Link).

[19] Walecki, R., Pavlovic, V., Schuller, B., & Pantic, M, "Deep structured learning for facial action unit intensity estimation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3405-3414, 2017. Article (CrossRef Link).

[20] S. Li, W., Abtahi, F., & Zhu, Z., "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1841-1850, 2017. Article (CrossRef Link).

[21] Z Tősér, LA Jeni, A Lőrincz, JF Cohn, "Deep Learning for Facial Action Unit Detection Under Large Head Poses," in *Proc. of Computer Vision – ECCV 2016 Workshops*, pp. 359-371, 2016. Article (CrossRef Link).

[22] Li, X., Chen, S., & Jin, Q., "Facial action units detection with multi-features and-aus fusion," in *Proc. of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 860-865, 2017 May. Article (CrossRef Link).

[23] Batista, J. C., Albiero, V., Bellon, O. R., & Silva, L., "Aumpnet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network," in *Proc. of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 866-871, May 2017. Article (CrossRef Link).

[24] He, J., Li, D., Yang, B., Cao, S., Sun, B., & Yu, L, "Multi view facial action unit detection based on cnn and blstm-rnn," in *Proc. of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 848-853, May 2017. Article (CrossRef Link).

[25] Tang, C., Zheng, W., Yan, J., Li, Q., Li, Y., Zhang, T., & Cui, Z, "View-independent facial action unit detection," in *Proc. of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 878-882, May 2017. Article (CrossRef Link).

[26] Zhanpeng Zhang, Ping Luo, Chen Change Loy and Xiaoou Tang, "Facial Landmark Detection by Deep Multi-task Learning," in *Proc. of ECCV*, pp. 94-108, 2014. Article (CrossRef Link).

[27] Junho Yim, Heechul Jung ByungIn Yoo, Changkyu Choi, Dusik Park and Junmo Kim, "Rotating Your Face Using Multi-task Deep Neural Network," in *Proc. of CVPR*, 2015. Article (CrossRef Link).

[28] Cha Zhang and Zhengyou Zhang, "Improving Multiview Face Detection with Multi-Task Deep Convolutuinal Neural Networks," in *Proc. of WACV*, 2014. Article (CrossRef Link).

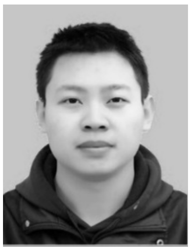
[29] Nitish Srivastava, "Improving neural networks with dropout," *Ph.D. thesis, University of Toronto*, 2013. Article (CrossRef Link).

[30] Zhu, Yi, and S. Newsam, "Efficient Action Detection in Untrimmed Videos via Multi-task Learning," in *Proc. of Applications of Computer Vision IEEE*, 2017. Article (CrossRef Link).

[31] Xing Zhang, Lijun Yin, Jeff Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeff Girard, "BP4D-Spontaneous: A high resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, 32(10), pp. 692–706, 2014. (special issue of the Best of FG13). Article (CrossRef Link).

[32] Zheng Zhang, Jeff Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Micheal Reale, Andy Horowitz, Huiyuan Yang, Jeff Cohn, Qiang Ji, and Lijun Yin, "Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis," in *Proc. of CVPR*, 2016. Article (CrossRef Link).

**Jun He** was born in Beijing City, China in 1976. She received the B.S. degree in Optical Engineering from Beijing Institute of Technology, Beijing, China in 1998 and Ph.D. degree in Physical Electronics from the same university, in 2003 directly through an accelerated Ph.D. program. From 2003 to 2010, she was a Research Assistant with the College of Information Science and Technology of Beijing Normal University, Beijing, China. Since 2010, she has been an Assistant Professor with the same college. Her research interests include image processing application and pattern recognition, deep learning, and face recognition, emotion recognition.



**Dongliang Li** received a BSc in School of Software and Microelectronics from Northwestern Polytechnical University, China. He is currently working toward a M.S. in Signal and Information Processing at Beijing Normal University. His research interests include deep learning and affect recognition.

**Bo Sun** received the B.S. degree in computer science from Beihang University (BUAA), Beijing, China, in 1988, the M.S. and Ph.D. degrees separately in natural language process and computer-aided education from Beijing Normal University (BNU), Beijing, in 1991 and 2003. From 1999 to 2004, he was an Associate Professor with the Computer Science Department at BNU. Since 2004, he has been a Professor with the College of Information Science and Technology, BNU. His research interests include machine learning, deep neural networks, pattern recognition and natural language processing.

**Lejun Yu** received the B.S. degree in computer science from Beijing Normal University, Beijing, China in 1999 and the M.S. degree in electronic engineering from Beijing Normal University, Beijing, China, in 2002. He received the Ph.D. degree in computer technology from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2007. From 2007, he has been working with Beijing Normal University. In 2014, he worked with Vienna University of Technology as a visiting scholar. His research interest includes the video analyzing and artificial intelligence