An Action Unit co-occurrence constraint 3DCNN based Action Unit recognition approach

Xibin Jia[1], Weiting Li[1], Yuechen Wang[1], SungChan Hong[2] and Xing Su[1]*

[1] Faculty of Information Technology, Department of Computer Science, Beijing University of Technology, Beijing
Beijing, China.R 100124 - China
[e-mail: xingsu@bjut.edu.cn]
Department of information science and telecom, Hanshin University
Seoul, South Korea
[e-mail: schong@hs.ac.kr]
*Corresponding author: Xing Su

Received September 18, 2019; revised December 1, 2019; accepted December 25, 2020; published March 31, 2020

Abstract

The facial expression is diverse and various among persons due to the impact of the psychology factor. Whilst the facial action is comparatively steady because of the fixedness of the anatomic structure. Therefore, to improve performance of the action unit recognition will facilitate the facial expression recognition and provide profound basis for the mental state analysis, etc. However, it still a challenge job and recognition accuracy rate is limited, because the muscle movements around the face are tiny and the facial actions are not obvious accordingly. Taking account of the moving of muscles impact each other when person express their emotion, we propose to make full use of co-occurrence relationship among action units (AUs) in this paper. Considering the dynamic characteristic of AUs as well, we adopt the 3D Convolutional Neural Network(3DCNN) as base framework and proposed to recognize multiple action units around brows, nose and mouth specially contributing in the emotion expression with putting their co-occurrence relationships as constrain. The experiments have been conducted on a typical public dataset CASME and its variant CASME2 dataset. The experiment results show that our proposed AU co-occurrence constraint 3DCNN based AU recognition approach outperforms current approaches and demonstrate the effectiveness of taking use of AUs relationship in AU recognition.

Keywords: Action Unit Recognition; 3DCNN; Correlation; FACS

A preliminary version of this paper appeared in APIC-IST 2019, Beijing, China. This version of the work includes a new AU recognition algorithm that incorporates AUs' Co-occurrence relationship. This research is supported by the Beijing Natural Science Foundation under Grant (No. 4202004) and the International Research Cooperation Seed Fund of Beijing University of Technology（No. 2018A02）

http://doi.org/10.3837/tiis.2020.03.001 **ISSN : 1976-7277**

1. Introduction

In the 1970s, Swedish anatomist Carl-Herman Hjortsj designed a Facial Action Coding System (FACS), which decomposes human's facial activities through facial expressions[1]. The basic components of FACS are Action Units (AUs), which describe the states of facial muscle movements and provide micro-information for the facial expression. Therefore, AU recognition plays an important role in the fields of facial expression recognition, mental state analysis, etc.

With the increasing demands for the intelligent life, high-quality human-computer interaction experience is the goal of many research fields. By analyzing the facial expressions, we can better capture the mental state of users and the robot will be more user-friendly. From the perspective of anatomy, the diversity and complexity of facial muscle movements lead to a variety of facial expressions. However, the current approaches on the facial expression recognition mainly focus on the six basic facial expressions, which are happiness, sadness, surprise, anger, fear and disgust. Although the trained classifier for each basic facial expression can achieve accurate basic facial expression recognition[2], they cannot perform well for a variety of complex facial expressions. Since AUs can describe facial muscle movements, the states of them are highly related to the facial expressions. Therefore, we can get effective information for the complex facial expression according to the AU recognition.

AUs also play an important role in the research of the behavioral science. They can accurately describe the minor differences among facial expressions. According to AUs, we can accurately judge the mental states of humans, such as fatigue, irritability, confusion, etc. And then improve the effects for safe driving, Internet education, medical diagnosis and other fields[3、4、5、6、7]. However, the cost of manual AU recognition is very expensive. It takes 100 hours to train an AU recognition worker and 2 hours for a skilled AU recognition worker to annotate a one-minute video[3]. In addition, there is a lot of useful information about the dynamic change of AUs, but it is hard to analyze them manually. The AU recognition is significant and valuable as many challenges need to be solved.

After reviewing and analyzing the researches on the AU recognition in recent years, this paper proposes a 3D Convolutional Neural Network (CNN) AU recognition framework, which can achieve accurate AU recognition by considering AU co-occurrence relationships. In general, the contributions of the proposed approach can be summarized as follows.

1. According to the association between AUs and expressions of FACS, the proposed framework selects the AUs with the expression degree of 70% or more, and performs AU image segmentation on the original face image to preprocess facial expression images;
2. We proposes an AU recognition framework based on the co-occurrence relationships of AUs. Specifically, the proposed framework extracts the spatiotemporal features of the AU image sequence by using the 3D CNN; and synchronously trains the 13 AU recognition models by using the same objective function and the AU co-occurrence relationship constraints.
3. The comparison experiments between 2D and 3D CNN based AU recognition models as well as between AU recognition models with and without considering AU co-occurrence relationships are conducted. The experimental results indicate that our 3D CNN and AU co-occurrence relationship based AU recognition framework can improve the accuracy of the AU recognition significantly;

The rest of the paper is organized as follows. In Section 2 the works related to the AU recognition are given. In Section 3, the AU and their co-occurrence relationships are demonstrated. In Section 4, our AU recognition framework is introduced in details. In Section 5, the experiments are illustrated and their results are analyzed. This paper is concluded in Section 6.

2. Related Work

In computer vision researches, the two-dimensional convolution is a general algorithm. However, when the learning tasks are video data, dynamic information between multiple consecutive video frames is required. In order to extract dynamic features, Ji [58] et al. used three-dimensional convolution in the convolution phase of CNN to achieve the feature extraction from both spatial and temporal dimensions. **Fig. 1** shows the structure of the three-dimensional convolutional neural network proposed by Ji et al.

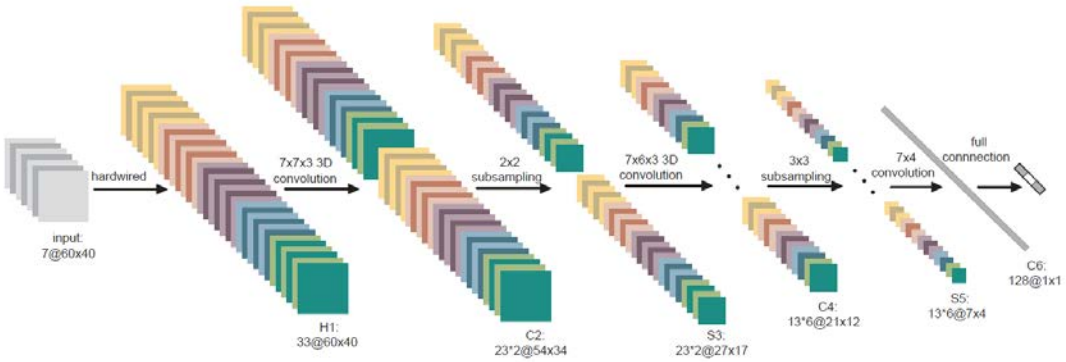


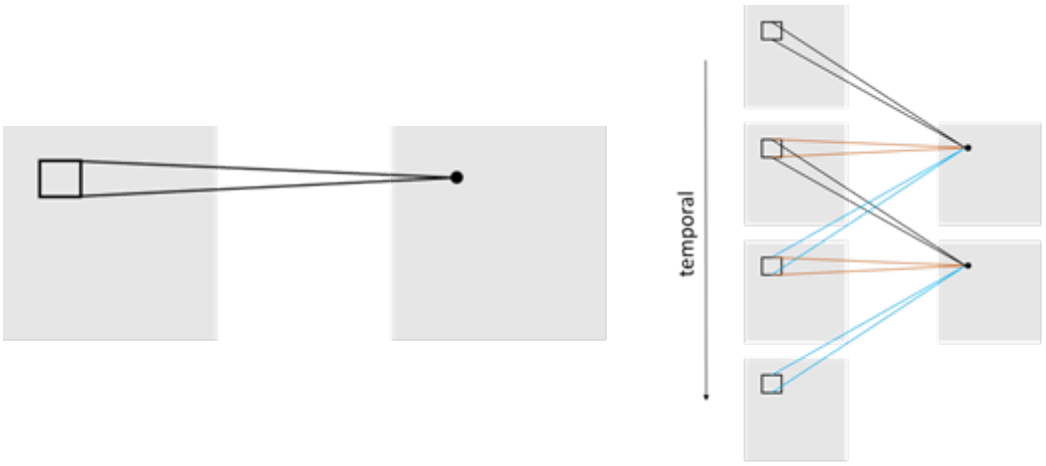
Fig. 1. The structure of the three-dimensional convolutional neural network

The three-dimensional convolution uses a three-dimensional convolution kernel to extract spatiotemporal features from a cube which is stacked by the consecutive video frames. It allows the feature map in the convolutional layer to be connected to a plurality of consecutive video frames in the previous layer, so as to capture the motion information in the video.

Fig. 2 shows the comparison between **(a)** a two-dimensional convolution and **(b)** a three-dimensional convolution. In the three-dimensional convolution, there are three convolution kernels for the temporal dimension. Since the weights of three-dimensional convolution kernels are shared in the whole video data cube, the same three-dimensional convolution kernel can only extract one feature from the cube (**Fig. 2**). The same color of the lines indicates that the convolution kernels have the same weight values. Therefore, if multiple features need to be extracted, multiple convolution kernels can be used for the feature extraction.

American psychologists Paul Ekman and Wallace V. Friesen summarized and further improved Carl-Herman Hjortsjs's work and proposed the FACS 2002[8].

The AU recognition has always been a very challenging subject in the computer vision field. With the development of the psychology and human-computer interaction in recent years, more and more researchers begin to pay attention to the AU recognition. Although there are a lot of related research works on the AU recognition, it is very difficult to achieve the high precision AU recognition, because of the complexity and diversity of the AU occurrence.



(a) **(b)**

Fig. 2. The comparison between two-dimensional convolution and three-dimensional convolution

Bartlett [9] et al. proposed a combination of support vector machine (SVM) and hidden Markov machine (HMM) based AU recognition method for the AU recognition of AU images and AU image sequences. The proposed method first trains the SVM by using image data with strong AU features, then extracts the AU features from each frame image in the AU image sequence by using the trained SVM, and finally uses the features extracted by the SVM as the inputs of the HMM to train the HMM for the AU Identification. The experiments have proved that their method has better recognition effect than the AU recognition approaches that only use SVM. Pantic[10、11] et al. proposed an approach to analyze the dynamics of the AU by extracting the geometric features from the front and side of faces. Their approach identifies 27 AUs based on the dynamics and rules of the AU. Some researchers extracted Gabor features from AU images and used SVM, HMM, Adaboost, DBN and neural networks as the classifiers for the AU recognition and verified the effectiveness of these classifiers on the AU datasets, such as CK and MMI [12、13、14、15、16].

3. AU and AUs' Co-occurrence

3.1 Construction

In FACS 2002, many detailed analyses have been performed. Such as the changes of facial muscle movements and the observable expressions caused by these facial muscle movements. According to these analyses, facial muscle movements are divided into some basic AUs. Parts AUs of FACS 2002 are shown in **Table 1**.

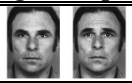
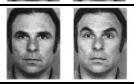
Table 1. Parts AUs of FACS 2002

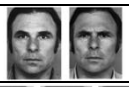
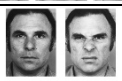
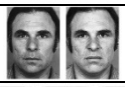
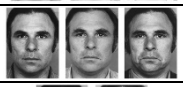
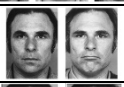
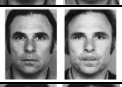
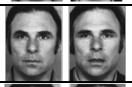
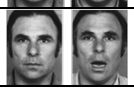
AU	name	feature
AU1	Inner brow raiser	
AU2	Outer brow raiser	
AU4	Brow lowerer	
AU5	Upper lid raiser	
AU6	Cheek raiser	

AU7	Lid tightener	
AU9	Nose wrinkle	
AU10	Upper lip raiser	
AU11	Nasolabial deepener	
AU12	Lip corner puller	
AU15	Lip corner depressor	
AU16	Lower lip depressor	
AU17	Chin raiser	
AU18	Lip pucker	
AU20	Lip stretcher	
AU22	Lip funneler	
AU23	Lip tightener	
AU24	Lip pressor	
AU25	Lips part	
SU26	Jaw drop	

According to the correlation between AUs and six basic facial expressions in FACS, 13 AUs with a correlation of more than 70% with six basic facial expressions were selected in this study, which are
shown in **Table 2**. The 13 AUs are mainly distributed around eye-brow, nose and lip regions. Among them, AU1, AU2, AU4 and AU7 describe the muscle movements around eyes' and brows' regions. AU9 and AU17 describe the muscle movements around nose regions. AU10, AU12, AU15, AU20, AU24, AU25 and AU26 describe the muscle movements around lip regions.

Table 2. The descriptions of 13 AUs

AU	Region display	feature
AU1		Inner brow raiser
AU2		Outer brow raiser

AU4		Brow lower
AU7		Lid tightener
AU9		Nose wrinkle
AU10		Upper lip raiser
AU12		Lip corner puller
AU15		Lip corner depressor
AU17		Chin raiser
AU20		Lip stretcher
AU24		Lip pressor
AU25		Lips part
SU26		Jaw drop

3.2 AUs' Co-occurrence

Through the detailed introduction of FACS, it can be seen that the states of AUs can help us to understand the human facial expressions and the movement of each AU is controlled by one or a group of facial muscles. Therefore, some AUs may have co-occurrence relationships. Combined with the results of facial anatomy, it can be found that some AUs often appear at the same time, while some AUs rarely appear at the same time. Since AUs are controlled by facial muscles, their co-occurrence should rely on the facial anatomy. The muscle movement caused AUs co-occurrence is called the expression-independent dependency between AUs, which consists of two parts: positive co-occurrence and negative co-occurrence.

Positive co-occurrence refers to the simultaneous occurrence of certain AUs, because they are controlled by the same or adjacent muscles. For example, AU1, which is the inner brow raiser, and AU2, which is the outer brow raiser, are controlled by occipitofrontal muscle. The tail contraction of occipitofrontal muscle causes AU2, and the middle contraction of occipitofrontal muscle causes AU1. Since the contraction of the tail part of the occipitofrontal muscle usually occurs simultaneously with that of the middle part of the occipitofrontal muscle, the occurrence of AU2 increases the chance of AU1 occurrence. Therefore, AU1 and AU2 are positive co-occurrence.

On the other hand, negative co-occurrence refers to the fact that some AUs rarely or never appear at the same time. **Table 3** summarizes the positive co-occurrence and negative co-occurrence of AUs in FACS. For example, AU12, which is the lip corner puller, is controlled by the zygomaticus Major muscle group, and AU15, which is the lip corner depressor, is controlled by the anatomical angular muscle group. Obviously, it is impossible

for the two AUs to appear simultaneously, since the contraction of zygomatic major muscle constricts the movement of descending oral horn muscle. The occurrence of AU12 reduces the occurrence of AU15. Therefore, AU12 and AU15 are negative co-occurrence. **Table 3** summarizes the positive co-occurrence and negative co-occurrence of AUs in FACS.

Table 3. The AU dependencies

AU co-occurrence	AU pair
Positive	(AU1,AU2),(AU4,AU7), (AU4, AU9),(AU7, AU9), (AU6,AU12),(AU9,AU17), (AU12, AU17),(AU15,AU24), (AU17, AU24), (AU23, AU24)
Negative	(AU2, AU6), (AU2, AU7), (AU12, AU15), (AU12, AU17)

4. AU Recognition Framework

The overall architecture of the proposed AU recognition framework is shown in **Fig. 3**. In our framework, first, the facial image is segmented to obtain AU regional image data. In order to ensure the accuracy of the AU feature extraction, three kinds of AU recognition models are set up according to recognition regions. Then, AU regional image data are the inputs of three AU recognition models. Finally, three types of AU recognition models are constrained by the co-occurrence relationships of AUs in the training process of the AU recognition models. Thus, AU features can be supplemented and modified by the AUs co-occurrence in the AU recognition process.

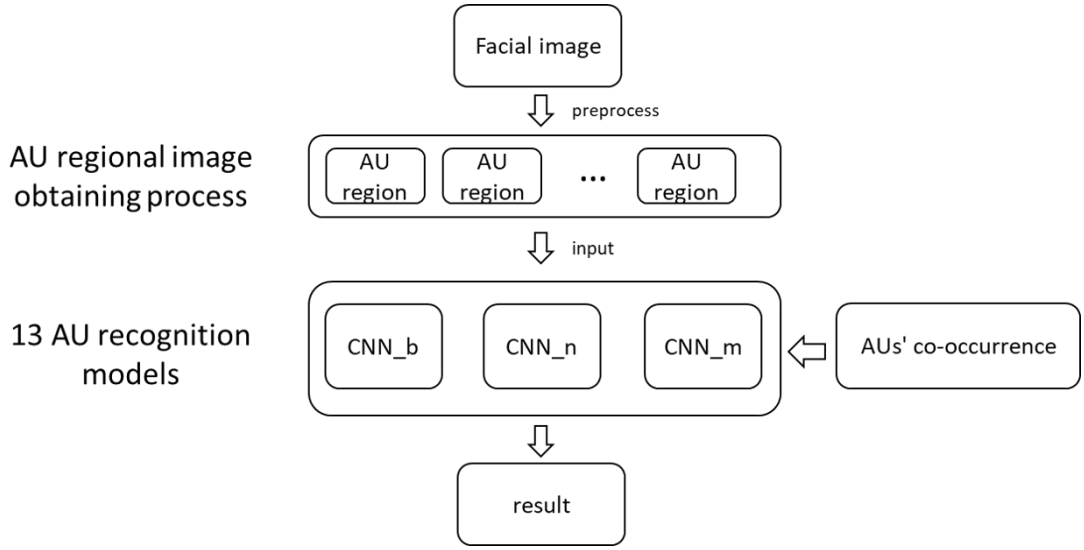


Fig. 3. The comparison between two-dimensional convolution and three-dimensional convolution

As shown in **Fig. 3**, the overall architecture of our framework includes two parts: an AU regional image obtaining process and 13 AU recognition models. The AU regional image obtaining process obtains the regional AU images from facial images. The 13 AU recognition models are trained to recognize the AUs in the corresponding regions by considering their co-occurrence relationships.

4.1 The preprocess module

In this module, the facial image is segmented according to the existing 68 feature points. The 68 feature points of a human face are shown in **Fig. 4**.

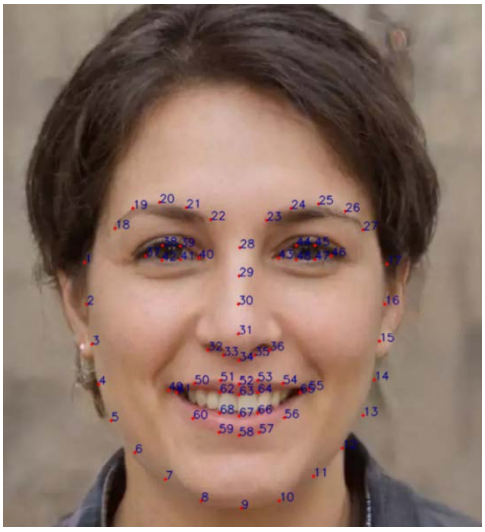


Fig. 4. 68 feature points of a face

Assuming that the AU recognition region is represented by $region\_AU$, which is a rectangular region. The location of the region is determined by the $left\_top(x, y)$ and $right\_bottom(x, y)$ which are the vertices of the upper left corner and the lower right corner of the rectangle. According to the location distribution of 68 feature points, we selects 18, 20, 27 and 29 feature points (shown in **Fig. 4**) as the AU region around eyes and brows, 18, 29, 27 and 34 feature points (shown in **Fig. 4**) as the AU region around the nose and 20, 34, 25 and 9 feature points (shown in **Fig. 4**) as the AU region around the lip. $feature\_p1(x, y)$, $feature\_p2(x, y)$, $feature\_p3(x, y)$ and $feature\_p4(x, y)$ are used to represent the four feature points selected for each facial region. The $left\_top(x, y)$ and $right\_bottom(x, y)$ are the vertices of the upper left corner and the lower right corner of $region\_AU$, which are calculated by Equations 1 to 4.

$$left\_top.x = feature\_p1.x2 \tag{1}$$

$$left\_top.y = feature\_p2.y2 \tag{2}$$

$$right\_bottom.x = feature\_p3.x + 2 \tag{3}$$

$$right\_bottom.y = feature\_p4.y + 2 \tag{4}$$

where $left\_top.x$ and $left\_top.y$ represent the x coordinate of the upper left corner vertex and the y coordinate of the upper left vertex of the $region\_AU$. $right\_bottom.x$ and $right\_bottom.y$ represent the x coordinate of the lower right corner vertex and the y coordinate of the lower right corner vertex of the $region\_AU$. $feature\_p1.x$, $feature\_p2.y$, $feature\_p3.x$ and $feature\_p4.y$ represent the x coordinates of $feature\_p1$ and $feature\_p3$ and the y coordinates of $feature\_p2$ and $feature\_p4$ are selected for the facial region, respectively.

AU regional image obtaining process mainly includes three parts: Locating 68 feature points from a facial image by Dlib facial feature points; Selecting feature points to locate $region\_AU$ according to detailed description of AU in FACS. Segmenting the facial image according to the selected feature points. The locations of 68 feature points in a facial image are shown in **Fig. 5**.

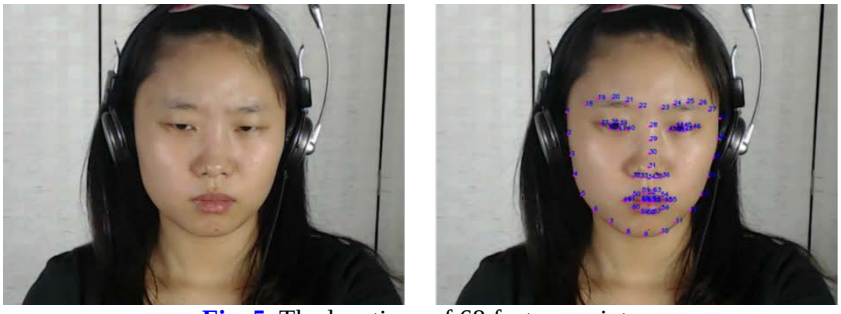


Fig. 5. The locations of 68 feature points

According to the detailed AU description of FACS, the face feature points of $region\_AU$ are selected and located, the process of which is shown in **Fig. 6**.

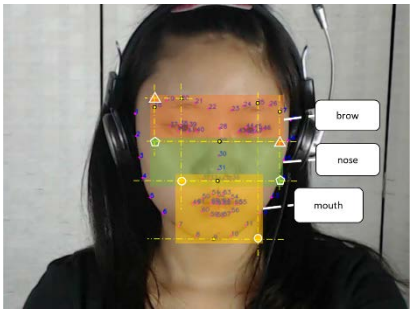


Fig. 6. Location of AU region

In **Fig. 6**, the three colored areas (yellow, green and orange) are the facial feature point region (i.e., $region\_AU$). The orange area with triangles as the vertices represents the eye-brow region. The green area with pentagons as the vertices represents the nose region. The yellow area with circles as the vertices represents the lip region.

The AU regional image obtaining of three-dimensional AU region image sequences is shown in **Fig. 7**.

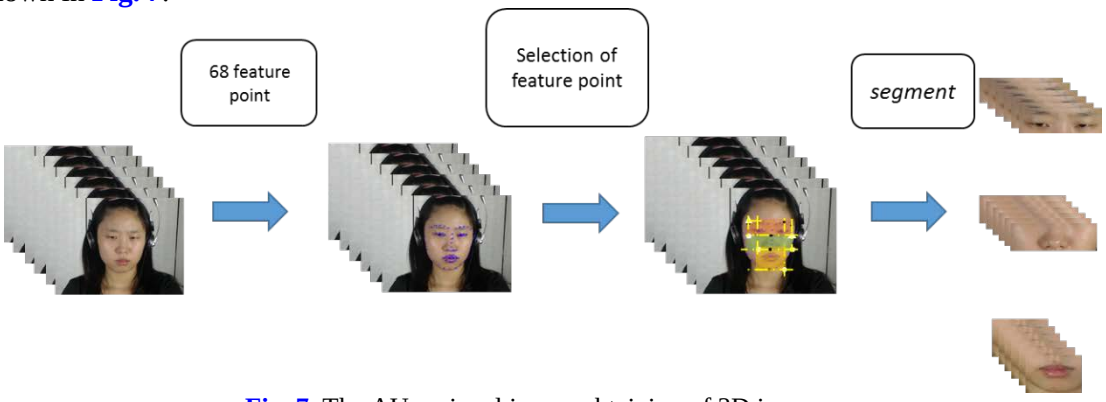


Fig. 7. The AU regional image obtaining of 3D images

4.2 The AU Recognition Models

Based on the 2D AU regional image obtaining in Section 4.2, a 2DCNN_AU recognition model for the 2D AU regional image recognition is designed. The structure of the model is shown in **Fig. 8**. The model consists of two modules: the convolution feature extraction module and the AU classification module. The convolution feature extraction module includes four convolution layers and four pooling layers, which is used to extract the convolution features of the input 2D AU regional images. The AU classification module includes a feature stretching layer and a full connection layer. And a sigmoid classification layer is connected at the end of the network.

In addition, the co-occurrence relationship constraints of AUs are added to the AU recognition model to provide auxiliary information for the AU recognition and improve the recognition effect of AU. Due to the biological structural interconnection among facial muscles, there exists evitable co-occurrence among AU pairs with the movement of facial muscles. In order to make full use of this prior knowledge of AU co-occurrence relationship in the data-driven based AU recognition, we propose in this paper adding an additional AU co-occurrence relationship fusion layer after thirteen independent 3DCNN-based AU classifier layers. As shown in the below formula, the prediction rate of ith AU is calculated with counting of the occurrence rate of relative AUs. By weighting with co-occurrence rate of relative AUs, the certainty of AU determination will be increased with taking account of the fact that strongly related AU happens at same time with high possibility. Therefore, with our proposed co-occurrence relationship constraint method into the CNN network, the 13 AU recognition CNN networks are integrated as a whole with the more robust AU recognition results.

$$\overline{p}(AU_i) = \sum_{j \in D_i} w_{ij} \hat{p}(AU_j) \tag{5}$$

Where, $\hat{p}(AU_j)$ represents the predicted probability of the j-th AU obtained by the j-th AU classifier, and w_{ij} represents the co-occurrence relationship between the i-th AU and the j-th AU. D_i is a set, whose elements are all AUs that have a co-occurrence relationship with AU_i, and $\overline{p}(AU_i)$ is the predicted probability of the final fusion of AU symbiosis.

Based on the above itegrated fusion network, the co-occurrence relationship constraints of AUs are hereby proposed to be added to the loss function of the AU recognition model, which are used in jointly training of the 13 AU classifiers as the auxillary information for the AU recognition model.

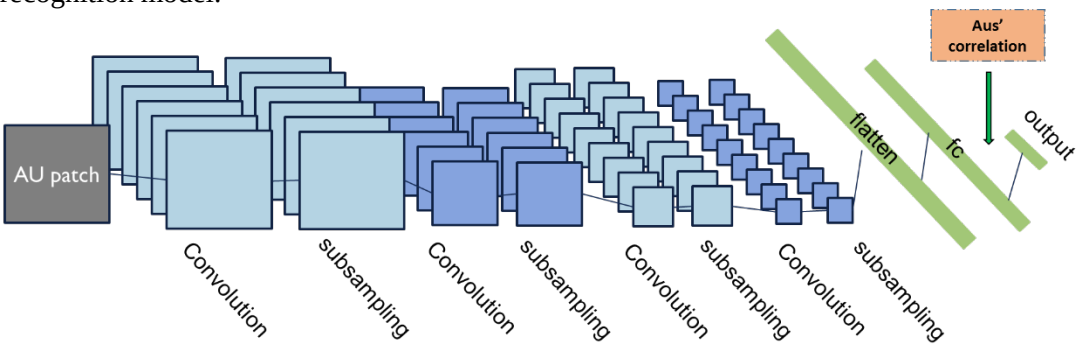


Fig. 8. The structure of 2DCNN model for the AU recognition

For the loss function of the model, we consider the error between results and labels and the co-occurrence relationships among 13AUs. The loss function of AU recognition model is shown in Equation 5:

$$Loss\_au_i = L\left(y_i \cdot \hat{y}_i\right) - \sum_{j=1}^{13} y_j \cdot d_{ij} \tag{6}$$

where $L(\bullet)$ denotes the cross-entropy loss function in the form of Equation 6:

$$L\left(y_i \cdot \hat{y}_i\right) = \left[y \cdot \log \hat{y} + (1-y) \cdot \log(1-\hat{y})\right] \tag{7}$$

In Equations 6 and 7, i represents the predicted value of the i^{th} AU. y_i represents the predicted value of the i^{th} AU classifier. $\hat{y}_i$ represents the real value of the AU image to be predicted. j represents the predicted value of the j^{th} AU classifier. y_j represents the predicted value of the j^{th} AU classifier. And d_{ij} represents the co-occurrence relationship between the i^{th} AU and the j^{th} AU.

Based on the 3D AU regional image sequence obtain in section 4.2, we designed a 3DCNN_AU recognition model for the 3D AU regional image sequence recognition. The structure of the model is shown in **Fig. 9**. The overall architecture of the 3DCNN_AU recognition model is consistent with the 2DCNN_AU recognition model. The model includes convolution feature extraction module and an AU classification module. However, since the 3DCNN_AU recognition model needs the AU recognition for 3D AU regional image sequence, it adopts 3DCNN structure. which enables the model to extract the 2D structure information from AU regional image sequence as well as the third dimension of AU image sequence-time dimension feature information. Compared with 2DCNN_AU recognition model, 3DCNN_AU recognition model increases the extraction of time-varying features of AU image sequence when extracting features. In theory, the recognition effect of 3DCNN_AU recognition model should be better than that of 2DCNN_AU recognition model.

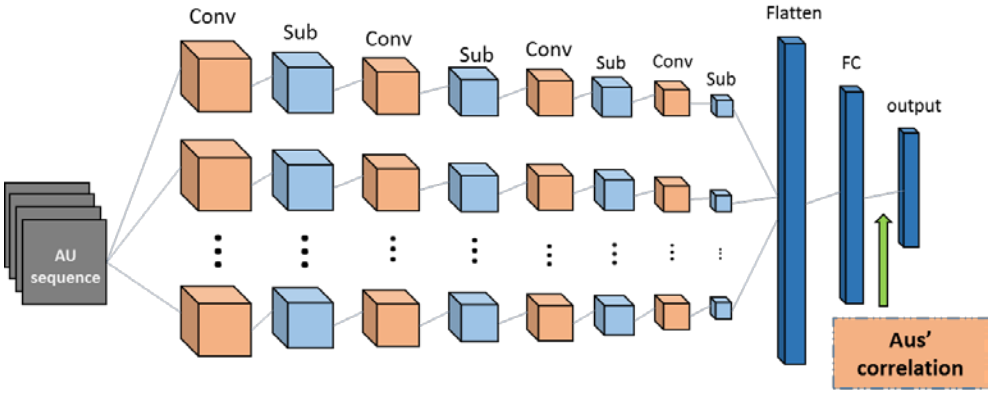


Fig. 9. The architecture of 3DCNN model for the AU recognition

The architecture of the training process of the AU recognition model is shown in **Fig. 10**.

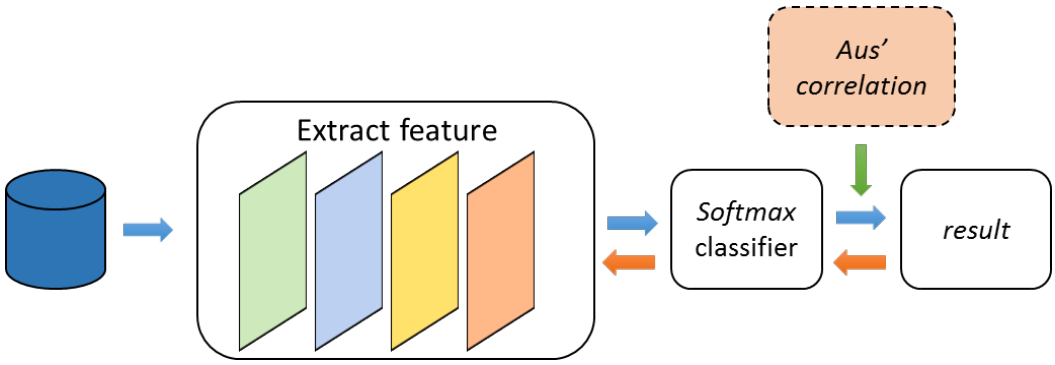


Fig. 10. The training process of our AU recognition model

As shown in **Fig. 10**, the training process of our AU recognition model mainly includes two processes: 1) Extracting convolution features of AU block image and recognizing AU according to the extracted features and 2) Comparing the recognition results with the real AU label to get the loss value of the AU recognition model and adjusting the model through the loss reverse. The AU recognition model is a multi-classification model. The model is composed of 13 binary AU classifiers, which share a joint objective function for the synchronous training. The output of the 13 AUs classifiers is a vector, which is shown in Equation 7.

$$output_{aus} = (au_1, au_2, \cdots, au_i, \cdots, au_{13}) \tag{8}$$

where au_i represents the output of the first AU classifier and $output_{aus}$ represents the comprehensive output of 13 AUs classifiers. The real AU labels of an expression image is described by the vector form, which is shown in Equation 8.

$$Y_{aus} = (y_{au}^1, y_{au}^2, \cdots, y_{au}^i, \cdots, y_{au}^{13}) \tag{9}$$

where y_{au}^i denotes the real label of the first AU with a value of 0 (AU does not appear) or 1 (AU appear). The simultaneous training of 13 AUs classifiers uses union objective functions, which is shown in Equation 9.

$$Loss = ave(\sum_{i=1}^{13} Loss\_au_i) \tag{10}$$

where, $Loss\_au_i$ is the loss function of a single AU classifier described in Equation 5. In $Loss\_au_i$, $y_i \in output_{aus}$ and $\hat{y}_i \in Y_{aus} \cdot Loss$ is a joint objective function of 13 AUs classifiers, which combines the recognition results and losses of each AU and integrates the co-occurrence relationships among AUs into the objective function. By doing so, the objective function makes the co-occurrence relationships among AUs to play an effective role in the training of the AU recognition model and effectively improve some of the phenomena of poor recognition effect of AU.

5. Experiment

5.1 Database and Preprocess

The data set used in this experiment is CASME database designed by Fu Xiaolan team of

Institute of Psychology, Chinese Academy of Sciences in 2013 and CAS (ME) 2 database designed by the team in 2014. The CASME database contains 195 micro-expression videos of 35 subjects (13 women and 22 men). The temporal resolution of the database is 60 frames per second, the image resolution is 640*480, and the resolution of the facial region is 150*190. CAS (ME) 2 database contains 57 microfacial videos from 22 subjects. The temporal resolution of the database is 30 frames per second and the image resolution is 640*480. Both databases contain expression tags and AU tags. Examples of expression images in CAEME series databases are shown in **Fig. 11**.

Fig. 11. Samples of CASME

Since the proposed AU recognition model is divided into 2D AU recognition model and 3D AU recognition model, it is necessary to process the original video data into two formats of AU regional image data sets when using CASME series databases. For the 2D AU recognition model, the expression peak frame in the original video data is selected as the image data to be used. For the 3D model, seven consecutive image sequences near the expression peak frame in the original video data are selected as the image sequence data to be used. The AU samples obtained after pretreatment of CASME series data sets are shown in **Table 4**.

Table 4. Sample size of AU images

AU	samples	AU	samples
AU1	49	AU15	23
AU2	51	AU17	30
AU4	248	AU20	3
AU7	68	AU24	26
AU9	24	AU25	10
AU10	20	AU26	1
AU12	162		

5.2 Experiment Design and Results

In order to verify the effectiveness of the proposed AU recognition method based on the constraints of AUs' correlation, two sets of comparative experiments are designed and implemented.

1. The comparative experiment between AU recognition model based on 2DCNN and AU recognition model based on 3D CNN.
2. The comparative experiment on the recog- nition effect of the AU recognition model based on the constraints of the symbiotic relationship between AUs and the AU recognition model without considering the AUs' correlation.

Two evaluation indexes, Accuracy and AUC, are selected to analyze the effects of 2DCNN_AU recognition model and 3D CNN_AU recognition model on 13 AUs.

Tables 5 and 6 show the classification results of 13 AU data in CASME library using 2D CNN_AU recognition model and 3D CNN_AU recognition model respectively. Through the analysis of the recognition results of the two AU recognition models, we can find that the recognition accuracy of the other 11 AUs is over 83%, and the recognition accuracy of eight AUs is over 92%, except for AU4 and AU12.

Table 5. Statics of results of 2DCNN_AU recognition model

	AU1	AU2	AU4	AU7	AU9	AU10
Accuracy	88.36	87.88	40.62	83.61	**94.29**	**95.25**
AUC	0.863	0.877	0.388	0.833	0.911	0.903

	AU12	AU15	AU17	AU20	AU24	AU25	AU26
Accuracy	61.06	**94.54**	**92.87**	**99.29**	**93.82**	**97.62**	**99.76**
AUC	0.607	0.935	0.903	0.99	0.931	0.959	0.993

Table 6. Statics of results of 3DCNN_AU recognition mode

	AU1	AU2	AU4	AU7	AU9	AU10
Accuracy	89.55	89.13	47.12	85.5	**94.88**	**95.74**
AUC	0.888	0.887	0.465	0.833	0.943	0.955
	AU12	**AU15**	**AU17**	**AU20**	**AU24**	**AU25**	**AU26**
Accuracy	65.46	**95.1**	**93.6**	**99.36**	**94.46**	**97.87**	**99.79**
AUC	0.638	0.947	0.922	0.988	0.936	0.968	0.996

AU4 is defined in FACS as lowering eyebrows. In the macroscopic expression, the characteristics of this muscle movement change more obviously (as shown in **Fig. 12a**), which includes the prominent features of muscle texture caused by the muscle contraction in the middle of two eyebrows and the obvious decrease of eyebrow position. But in micro-expressions, these characteristics of muscle movement are not evident (as shown in **Fig. 12b**). Similarly, the pull-up of the corners of the mouth and the prominent features of the statutory lines in AU12 are not evident in micro-expressions.

(a) **(b)**

Fig. 12. The changed feature figure of AU4 in macro-expression and micro-expression

Through the statistics of the recognition results of two AU recognition models in **Tables 5 and 6**, the recognition accuracy of the two models is compared with that of AUC. As shown in **Fig. 13**.

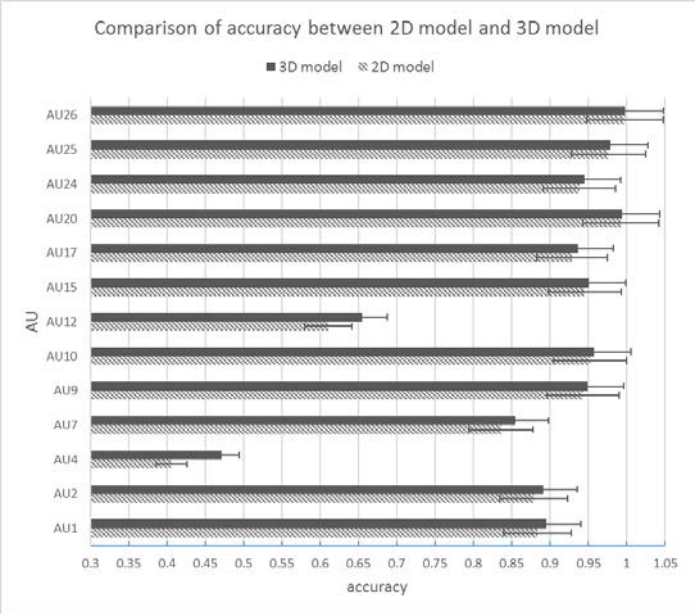


Fig. 13. Comparison of accuracy between 2D model and 3D model

Through the display of **Fig. 13**, it can be clearly seen that the recognition accuracy of the 3D AU recognition model for 13 kinds of AUs is improved as a whole compared with that of the 2D AU recognition model. Especially in AU4 and AU12, which have low recognition effect, the recognition accuracy of 3D model is 16% and 7% higher than that of 2D model. Through comparative analysis, it can be seen that the temporal features extracted by 3D convolution neural network play an effective role in the AU recognition of microfacial datasets compared with 2D convolution neural network, and can help improve the recognition effect of AU.

In order to better analyze the recognition effect of the two AU recognition models, the AUC of the recognition results of the two AU recognition models on 13 AUs is counted. The statistical results are shown in **Fig. 14**.

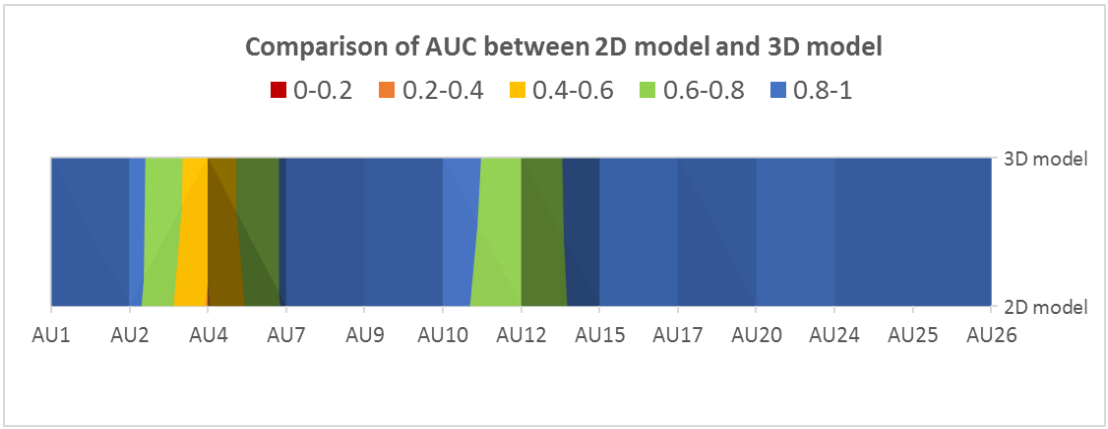


Fig. 14. Comparison of AUC between 2D model and 3D model

Among them, the blue part indicates that AUC is between 0.8 and 1.0. The green part indicates that AUC is between 0.6 and 0.8. And the yellow part indicates between 0.4 and 0.2. According to the introduction of AUC evaluation index in the front part of this section, the bigger the AUC of the model, the better the recognition effect. Combining the model recognition results shown in **Tables 5 and 6**, it can be seen that the 2D AU recognition model and the 3D AU recognition model perform well on these 13 AUs. This further proves that the convolution feature of AU image and the temporal feature of AU image sequence can be effectively applied to AU recognition tasks.

The experimental results of the 2D and 3D AU recognition model show that the convolution features of AU images and the temporal features of AU image sequences contain the key feature information of AU. So the model can be effectively described by using these features. However, we also find that only relying on these features for AU recognition will not do well on a variety of AU data through the above experimental results. In order to handle this problem, we introduces the idea of correlation among AUs into AU recognition model. And trains 13 AU recognition models synchronously by using the AUs' correlation. The statistical analysis results between with and without considering co-occurenced relationship with 3D and 2D respecitvely are shown in **Fig. 15**.

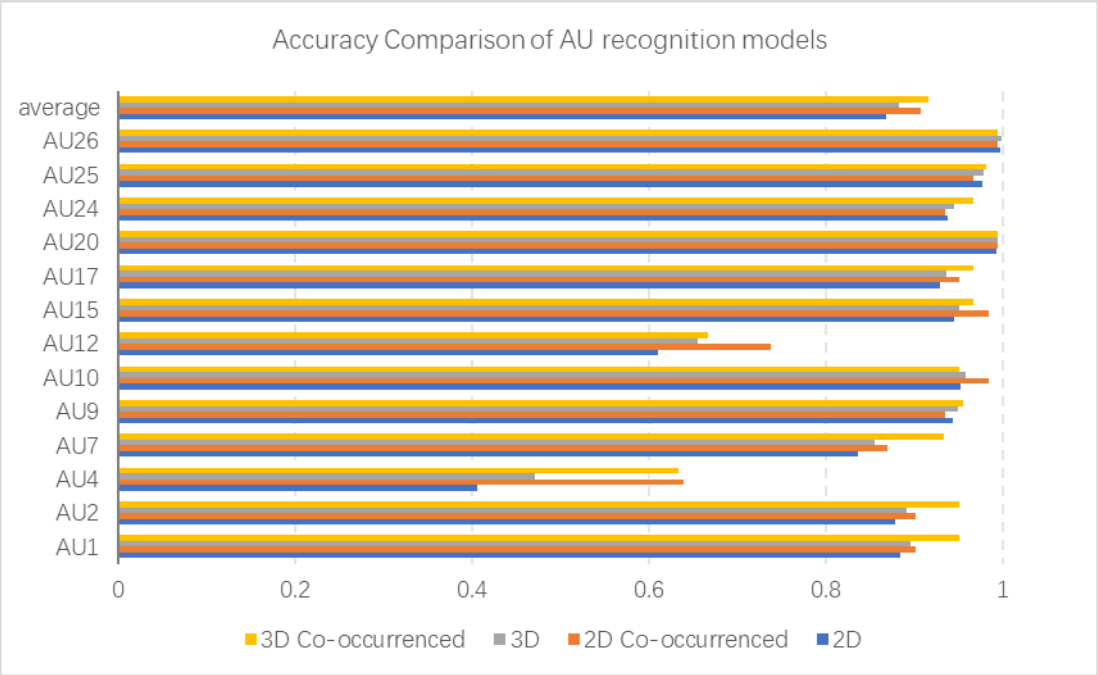


Fig. 15. Accuracy Comparison of AU recognition models

From the comparison results shown in **Fig. 15**, it could be found that adding the additional AU co-occurrence relationship fusion layer, most thirteen AUs recognition accuracy rate and the average recognition rate with both 2D and 3D models are higher than that without counting the co-occurrence realationship. Among them, the recognition accuracy rate of AU4 and AU12 especially increased by 16% and 12% respectively. Analysis the reason, there exits strong positive symbiotic relationship between AU4 and AU7, and between AU4 and AU9 in the AU co-occurrenc relationship among AU pairs.Therefore the recognition accuracy of

undiscriminable action unit AU4 is greatly increased by taking account of the positive symbiotic relationship with discrimative AU7 and AU9, which have the good recognition results independently. Similar contribution in improving the recogntion rate even when the relationship is negative. Like we find that there is a strong negative AU co-occurrence relationship between AU12 and AU15, AU12 and AU17. Therefore the recognition accuracy of AU12 is also greatly improved with impacting by the high recognition results of AU15 and AU17.

Above all, the experiments results demostrates that the additional AU co-occurrence relationship fusion layer plays an effective role in AU recognition on the micro-expression dataset. With counting the corraltive relationship among AUs, the recognition performance of AUs improves significantly, especially of those small unobtrusive actions correlative with some other easily detective AUs.

6. Conclusion

In this paper, we present an interactive method to manipulate a 3DCNN model and AUs' co-occurrence for AU recognition. 3DCNN is used to extract temporal and spatial features of image sequences, and AUs' co-occurrence are used to restrict the training of multiple AU classifiers, which enriches the feature information of AU. Finally, experiments show that the AU recognition model proposed in this paper have good performance in recognition accuracy. The main research work of this paper is as follows:

1. We selected 13 AUs (AU1, AU2, AU4, AU7, AU9, AU10, AU12, AU15, AU17, AU20, AU24, AU25 and AU26) with more than 70% of the correlation degree between AU and facial expression based on Facial Action Coding System (FACS).

We proposed an AU recognition algorithm based on AUs' **c**o-occurrence. Referring to the state description and feature analysis of AU in FACS, this paper uses Dlib's face key point detection algorithm to locate facial feature points and determine the AU segmentation area. The spatial and temporal features of AU image sequences are extracted by using 3D CNN, and AUs' **c**o-occurrence are added to the training process of AU classification network. Experiments on CASME databases show that the proposed algorithm based on AUs' **c**o-occurrence has good performance in recognition accuracy and recall rate. The comparison of recognition effect between 2DCNN AU recognition model and 3D CNN AU recognition model proves that temporal feature plays an important role in improving the recognition effect of AU. The AUs' **c**o-occurrence plays an important role in improving the accuracy of AU with inferior recognition accuracy.

References

[1] Hjorztsjö, *Man's face and mimic language*, Studentlitteratur, Lund, 1970. Article (CrossRef Link)

[2] Fasel B, Luettin J, "Automatic facial expression analysis: a survey," *Pattern Recognition*, Vol 36, No.1, pp. 259-275, 2003. Article (CrossRef Link).

[3] Bartlett M S, Littlewort G C, Frank M G, et al, “Automatic recognition official actions in spontaneous expressions,” *Journal of Multimedia*, Vol 1, No.6, pp.22-35, 2006.

[4] Ekman P, Davidson R J, Friesen W V, “The Duchenne smile: emotional expression and brain physiology Ⅱ,” *Journal of Personality and Social Psychology*, Vol 58, No.2, pp. 342-353, 1990. Article (CrossRef Link)

[5] Ekman P, Rosenberg E L, *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)*, Oxford University Press, New York, 1997. Article (CrossRef Link)
[6] Frank M G, Ekman P, "Appearing truthful generalizes across different deception situations," *Journal of Personality & Social Psychology*, Vol 86, No.3, pp. 486-495, 2004. Article (CrossRef Link)
[7] Bartlett M S, *Face image analysis by unsupervised learning and redundancy reduction*, University of California, San Diego, 1998. Article (CrossRef Link)
[8] P. Ekman and W. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," *Facial Action Coding System (FACS)*, 1978. Article (CrossRef Link)
[9] BARTLETT M S, MOVELLAN J R, LITTLEWORT G, et al, "Towards automatic recognition of spontaneous facial actions," *Oxford University Press, New York*, 393-412, 2005. Article (CrossRef Link)
[10] PANTIC M, ROTHKRANTZ L J M, " Facial Action Recognition for Facial Expression Analysis From Static Face Images," *IEEE TRANSACTIONS ON CYBERNETICS*, 34(3), 1449-1461. 2004. Article (CrossRef Link).
[11] BARTLETT M S, LITTLEWON G C, FRANK M G, et al, " Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, 1(6), 22-35, 2006.
[12] COHN J F, KANADE T, MORIYAMA T, et al, "A comparative study of alternative FACS coding algorithms," *CMU Final Report: FACS Coding Algorithms*, 200l. Article (CrossRef Link)
[13] BARTLETT M S, BRAATHEN B, GWEN L F, et al, *Automatic analysis of spontaneous facial behavior: a final project report*, u11iversity of California, San Diego, 2001.
[14] SMITH E, BARTLETT M S, MOVELLAN J, "Computer recognition of facial actions: a study of co-articulation effects," in *Proc. of the 8th Symposium on Neural Computation, La Jolla*, 1-6, 2006. Article (CrossRef Link)
[15] TONG Y, LIAO W, JI Q, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10), 1683-1699, 2007. Article (CrossRef Link).
[16] CAMPOS C P, TONG Y, JI Q, "Constrained Maximum Likelihood Learning of Bayesian Networks for Facial Action Recognition," in *Proc. of European Conference on Computer Vision, Springer-Verlag*, 168-181, 2008. Article (CrossRef Link)

Xibin Jia, received Ph.D. degree in computer science and technology from Beijing University of Technology in 2007, M.S. degree in intelligent instrument from North China Institute of Technology in 1996 and B.S. degree in wireless technology from Chongqing University in 1991. She is a Professor in the Faculty of Information at the Beijing University of Technology in Beijing, China. Her areas of interest include visual information cognition and computing, affective computing and intelligent medical image analysis and diagnosis.

Weiting Li received his BSc in computer science and technology from Beijing University of Technology, China in 2018. From 2018, Waiting Li studys as a postgraduate in College of Computer Science and Technology in Beijing University of Technology, China. His research interests include multi-agent systems, disaster management, computer vision and affective computing.

Yuechen Wang received her BSc in computer science and technology from Hebei Normal University, China in 2016. Yuechen received her MSc in College of Computer Science and Technology in Beijing University of Technology, China in 2019. Her research interests include multi-agent systems, disaster management, computer vision and affective computing.

Sung-Chan Hong received the B.S degree in Department of Statistics from Korea University, Seoul, Korea in 1983. He received M.S and Ph.D. degree from Keio University, Tokyo, Japan, in 1990 and 1994, respectively. From 1993 to 1995, he joined LG CNS, Seoul, Korea, as a Senior Research Engineer. He was president of the Korean Society for Internet Information, KSII, from 2011 to 2012. Currently, he is a Professor at Division of Information and Telecommunications, Hanshin University. His main research interests are in the areas of big data, artificial intelligence and information systems.

Xing Su received his BSc in software engineering from Beijing University of Technology, China in 2007. He received his MSc and PhD in computer science from University of Wollongong, Australia in 2012 and 2015, respectively. From 2016, he works a lecturer in the Faculty of Information at the Beijing University of Technology in Beijing, China. His research interests include distributed artificial intelligence, multi-agent systems, disaster management and service-oriented computing