Facial expression recognition based on facial action unit

Jiannan Yang
Computer Science and Technology
Nanjing Tech University
Jiangsu Province, China
201861120038@njtech.edu.cn

Fan Zhang
IBM Watson Group
IBM Massachusette Lab
Littleton, MA
fzhang@us.ibm.com

Bike Chen
Xinktech
Building 22, No. -02, Xuanwu Avenue
Jiangsu Province, China
chenbike@xinktech.com

Samee U. Khan
Electrical and Computer Eng
North Dakota State Univ
Fargo, ND
samee.khan@ndsu.edu

***Abstract*—In the past few years, there has been increasing interest in the perception of human expressions and mental states by machines, and Facial Expression Recognition (FER) has attracted increasing attention. Facial Action Unit (AU) is an early proposed method to describe facial muscle movements, which can effectively reflect the changes in people's facial expressions. In this paper, we propose a high-performance facial expression recognition method based on facial action unit, which can run on low-configuration computer and realize video and real-time camera FER. Our method is mainly divided into two parts. In the first part, 68 facial landmarks and image Histograms of Oriented Gradients (HOG) are obtained, and the feature values of action units are calculated accordingly. The second part uses three classification methods to realize the mapping from AUs to FER. We have conducted many experiments on the popular human FER benchmark datasets (CK+ and Oulu_CASIA) to demonstrate the effectiveness of our method.**

***Keywords*-facial expression recognition; facial action unit; facial landmark;**

I. Introduction

Intuitively, facial expression recognition recognizes basic human expressions [1] (e.g., surprise, sadness, happiness, disgust, anger, etc.) by processing and analyzing face image features. Moreover, the potential applications of facial expression recognition are very extensive, such as service industry, criminal investigation and interrogation, medical help [2] and so on. Most methods of facial expression recognition are mainly divided into two steps: feature extraction and classification. Feature extraction mainly analyzes the face image and obtains the potential features of the image. Since different expressions have different facial expression characteristics, it is possible to obtain effective potential features for better classification.

Facial action unit is by studying the movement of facial muscles [3], and a method of describing facial movement changes. In 1978, American psychologist Ekman Paul and Friesen [4] developed Facial Action Coding System (FACS), the system for almost all the muscles of the facial expression behavior carried on the detailed classification, is the facial expression enjoys wide application in measurement technology, one of the most representative methods. Although action units can accurately express facial expressions, they are rarely used in facial expression recognition due to the difficulty in accurate positioning. Currently, more and more attention has been paid to the study of facial action units. In recent years, there have been many studies on the positioning and detection of AUs. Ding X [5] et al. proposed a method based on Cascade of Tasks to detect facial action units.Baltrusaitis [6] et al. presented a real-time facial action unit intensity estimation and occurrence system based on appearance features (Histograms of Oriented Gradients) and geometric features (shape parameters and landmark locations). Based on the detection method of AUs proposed by Baltrusaitis et al., this paper puts forward a method of facial expression recognition using face action units, which can meet the requirements of low-configuration equipment. The main contributions of this paper are as follows:

We proposed a facial expression recognition based on facial action unit and used multiple classification methods to realize the mapping of AUs to facial expression.

The proposed method has relatively low requirements for computer configuration and does not rely on GPU devices. It can run on a notebook computer with i5-8300 processor and 8G of memory, and the recognition speed of each image is maintained at about 30ms. So our program can realize video and real-time camera facial expression recognition.

II. Related Work

With the continuous improvement of computer performance, many more accurate facial expression recognition methods have been developed. In the past, traditional recognition methods have shown excellent performance in facial expression recognition. He [7] et al. proposed a facial expression recognition method based on classical LBP. Wang

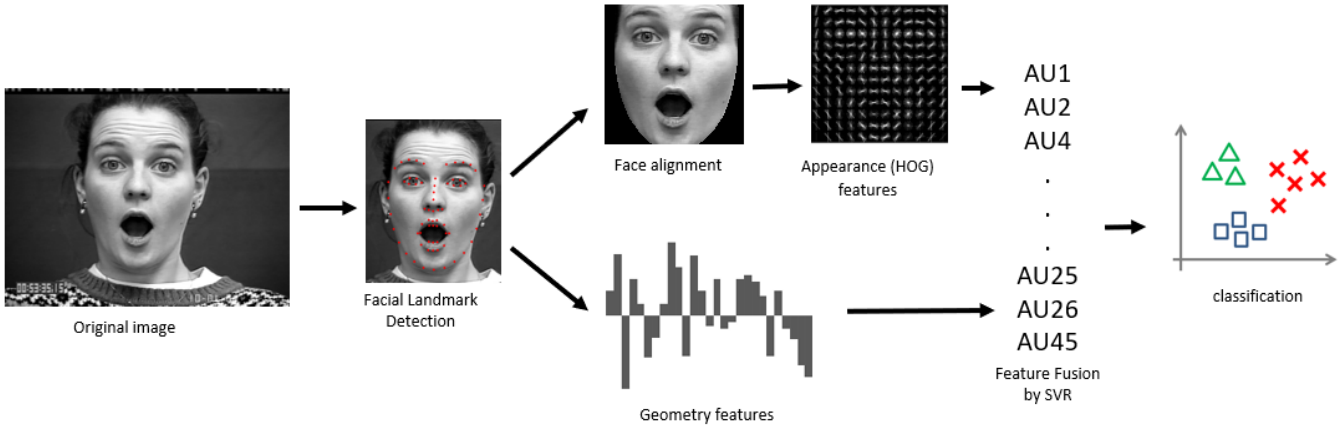


Figure 1. Overview of facial expression recognition method based on facial action unit.

[8] et al. proposed a method of HOG and Weber Local Descriptor (WLD) feature fusion to realize facial expression recognition, in order to solve the problem of lack of contour and shape information. Berretti [9] and others put forward a set of facial feature points calculation of depth image SIFT descriptors, and then select the most relevant feature subset of the original method.

With the advent of deep learning, facial expression recognition has been further developed. Liu [10] et al. proposed a generalized adaptive (N+M)-tuplet clusters loss function together with the identity-aware hard-negative mining and online positive mining scheme for facial expression recognition, which combined the deep metric loss and softmax loss in a unified two fully connected layer branches framework via joint optimization. Li [11] et al.proposed a novel multi-scale CNN integrated with an attention-based learning layer (AMSCNN) for robust facial expression recognition. The attention-based learning layer is designed to automatically learn the importance of different receptive fields in the face during training. Meng [12] et al. developed an identity-aware convolutional neural network which utilizes identity information during the procedure of training model to alleviate variations introduced by personal attributes. In addition, contrastive loss and softmax loss are adopted jointly as supervision signals to optimize the neural network. Jung [13] et al. proposed a two branches deep neural network to do the task of FER, one of which extracts temporal appearance features from an image sequence, the other of which extracts temporal geometry features from temporal facial landmarks. Wu [14] et al. proposed a novel facial expression recognition method for different pose faces based on special landmark detection (FERMPI-SFL) to solve the problem that self-occlusion of facial posture will seriously affect the accuracy of facial expression recognition.

III. Proposed Method

In this section, the detection methods of facial action units are first introduced, and then describe the three classification methods to realize the mapping of AUs to eight facial expressions (neutral, anger, contempt, happiness, disgust, sadness, fear and surprise). In the experiment, we extracted 16 AUs with prominent facial features. Figure 1 shows the detailed process of the method.

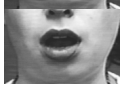
A. Facial Action Unit

We calculated 16 AUs with prominent facial features in the experiment. The description of AUs is shown in Table 1.

AU features extraction, using two main types of features: appearance features and geometric features. For appearance features, we get them by extracting direction gradient histogram. For geometric features, we rely on the coordinate results of the detection and tracking of facial landmark for the calculation of facial alignment. Here are the details.

For facial geometric features, Convolutional Experts Constrained Local Model (CE-CLM) proposed by Zadeh [15] et al. is used for the detection and analysis of facial landmark to obtain the coordinate positions. CE-CLM is an instance of Constrained Local Model (CLM) [16], It uses a local detector - Convolutional Experts Network (CEN) - that brings together the advantages of neural architectures and mixtures of experts in an end-to-end framework to optimize the results of CLM. We used CE-CLM model to mark 68 landmarks on the face as shown in Figure 2, which can clearly reflect the expression changes of various parts of the face (eyes, eyebrows, mouth, nose, etc.), and then we calculated the geometric features of the face image by comparing with the facial landmarks of the neutral expression. It is a difficult problem to standardize the landmarks of neutral expression because the facial landmarks of each person are not uniform. Therefore, we conduct a comparison based on dynamics, and we capture the geometric features of the face by training and

Table I
LIST OF AUS.

AU	Description	Example image
1	Inner Brow Raiser	
2	Outer Brow Raiser	
4	Brow Lowererr	
5	Upper Lid Raiserr	
6	Cheek Raiserr	
7	Lid Tightenerr	
9	Nose Wrinklerr	
10	Upper Lip Raiser	
12	Lip Corner Puller	
14	Dimpler	
15	Lip Corner Depressor	
17	Chin Raiser	
20	Lip stretcher	
23	Lip Tightener	
25	Lips part	
26	Jaw Drop	

learning the changes in the position of 68 landmarks. For the subsequent calculation of appearance features, we separated the face area and aligned it to fixed image size (112×112).

After face image alignment, we can extract appearance features, as the method proposed by Baltrusaitis [6] et al. When extracting Histograms of Oriented Gradients (HOG), we use Principal Component Analysis (PCA) method to reduce the dimension of HOG feature vector, because we need to reduce the dimension to meet the problem of expression analysis.

Finally, we used support vector regression to estimate the intensity of AUs. Among them, we used a linear kernel to improve the training speed, because we were interested in real-time detection methods.

B. Classification

After obtaining the values of the AUs through the above method, We need to preprocess the AU values, formula (1) shows the method of processing AUs. Min-max Normalization is firstly used to normalize the values, avoid characteristic value is too large or too small. Then we will square the

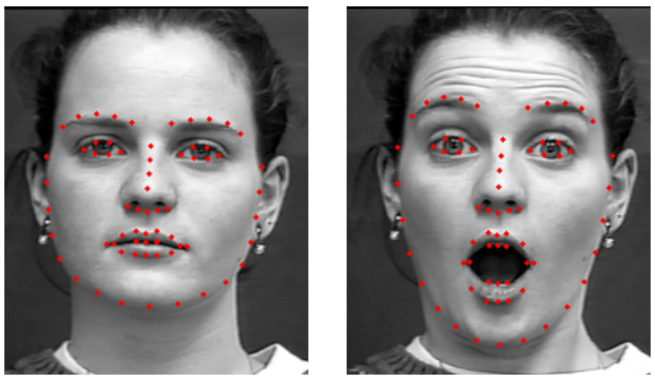


Figure 2. The results of 68 landmarks in CE-CLM model are shown in the figure, neutral expression on the left and surprised expression on the right. The geometric features of the face are obtained by comparing the changes in the position of landmarks.

results, in order to make the features of the corresponding expression more outstanding, reduce the influences of other irrelevant features, and thus more accurate classification.

$$AU_x{}' = \left(\frac{AU_x - min\{ AU_i \}}{max\{ AU_i \} - min\{ AU_i \}} \right)^2 \quad (1)$$

where, x is for every particular AU subscript, and i is for all the AU subscripts.

Then we combined AUs according to the changes of facial muscles corresponding to different expressions. Figure 3- Figure 9 show the changes in AUs corresponding to expressions.

Figure 3. The prominent changes of AUs corresponding to the anger expression.

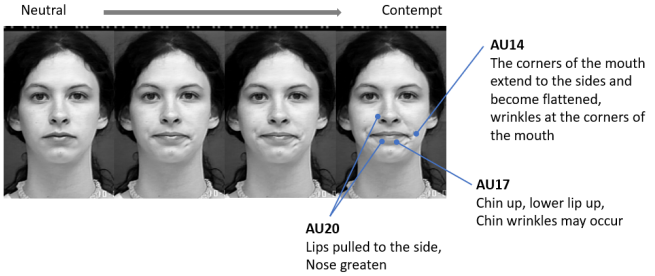


Figure 4. The prominent changes of AUs corresponding to the contempt expression.

Table II
EXPRESSION-RELATED AU COMBINATIONS

Emotion	Combination of AUs
Anger	4+7+17, 4+5+7+23
Contempt	14+17+20
Disgust	4+6+7+9, 7+9+10+17
Fear	1+2+20+25, 4+5+20
Happiness	6+7+12+25
Sadness	1+4+15+17
Surprise	1+2+5+25+26

From the figures, we can see that each expression triggers the movement of facial muscles, and the performance of each expression are basically similar, so we combined AUs for 7 expressions(not including neutral expression) to highlight the characteristics of the expression in the classification process. The results of the combinations are shown in Table 2. We get the characteristic values of the combination through the following formula:

$$AU_{E_k} = \lfloor \frac{2\sum AU_i'}{\sum max_{2x}\{\ AU_j'\ \}} \rfloor \quad (2)$$

where, E_k is the k-th combination of the corresponding expression(E is the abbreviation of emotion), i is the subscript of AU in the combination, x is the number of AUs in each combination, j is for all subscripts of AUs, max_{2x} means taking the maximum 2x values of AUs.

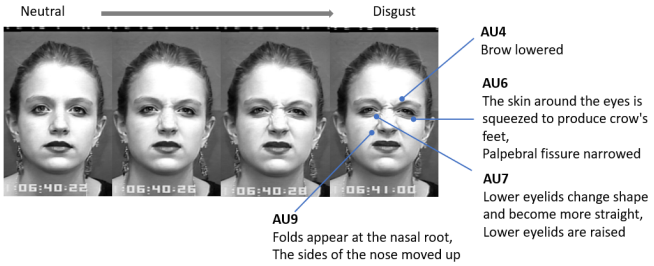


Figure 5. The prominent changes of AUs corresponding to the disgust expression.

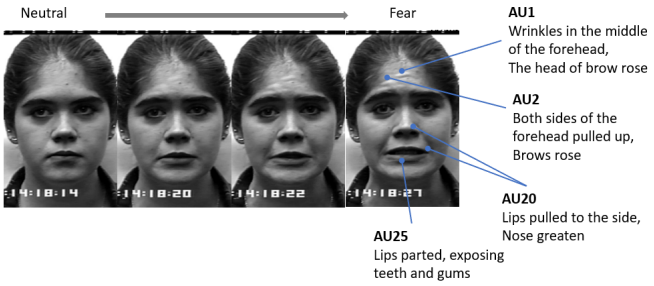


Figure 6. The prominent changes of AUs corresponding to the fear expression.

Finally, we used three classification methods to achieve the mapping of AUs to eight expressions, namely Support

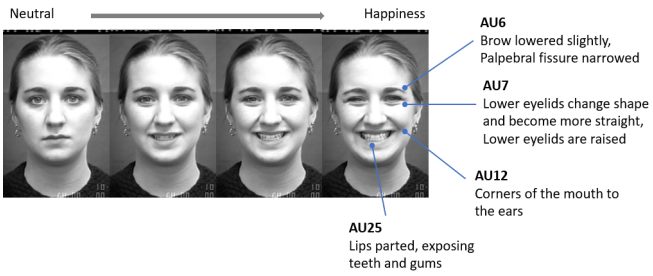


Figure 7. The prominent changes of AUs corresponding to the happy expression.

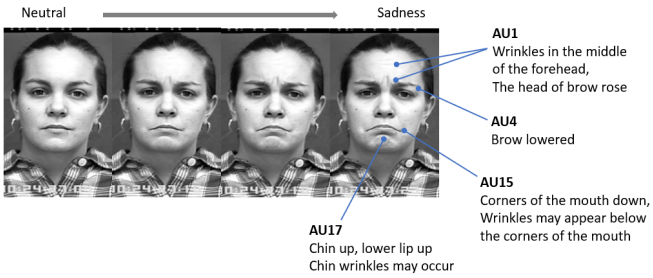


Figure 8. The prominent changes of AUs corresponding to the sad expression.

Vector Machine (SVM), Extreme Gradient Boosting (Xgboost) and Deep Neural Networks (DNN). At the same time, we use 10-fold cross-validation for each method and compare the results of classification to verify the effectiveness of our method. For SVM, we used linear kernel for the kernel function, because it can effectively improve our computational efficiency and has little impact on the classification results. Meanwhile, we verified the penalty item C parameter of the SVM model during the training. We choose the gbtree model as booster parameter input, which uses the tree-based model for booster calculation in Xgboost. We use the deep neural network to build a simple classifier for classification, and design a three-layer neural network, which contains 10, 20, 10 neurons, the input layer are AU features, the output layer are eight basic expressions. Finally, we trained the neural network 30,00 times.

IV. EXPERIMENTS

A. Dataset

CK+ dataset: CK + dataset [17] is composed of image sequences of eight expressions video recorded by 118 subjects. It is widely used to evaluate facial expression recognition methods. Each sequence in the dataset was marked with one of eight expression labels, such as neutral, anger, contempt, happiness, disgust, sadness, fear and surprise. The label is only provided for the last frame (called peak frame) of each image sequence. In order to obtain more experimental data, the dataset we trained contains the last 3-6 images (obvious facial expression features) of each image sequence, which

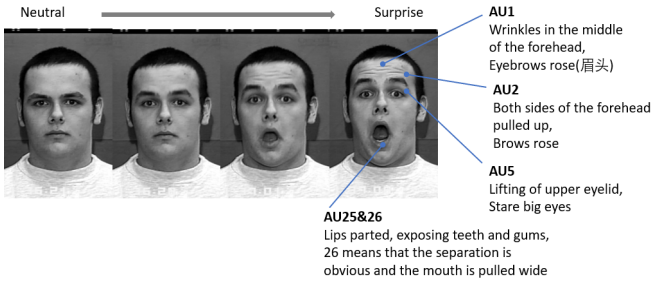


Figure 9. The prominent changes of AUs corresponding to the surprised expression.

is composed of 1492 images. We divide CK + dataset into 10 subsets, each of which is strictly independent of the principal. Data from eight subsets were used for training, and the remaining two were used for testing and validation, with 10-fold cross-validation used during training.

Oulu_CASIA dataset: Oulu_CASIA dataset [18] consists of 80 subjects, each of which contains 6 basic expressions (happiness, sadness, anger, fear, disgust, and surprise). Images of each facial expression are captured with two imaging systems: near-infrared (NIR) and visible light (VIS) under one of three different illumination conditions: normal indoor illumination, weak illumination, and dim illumination. Thence, Oulu_CASIA dataset includes 2880 image sequences. Following the previous approaches evaluated on the Oulu_CASIA dataset, only 480 image sequences taken under normal illumination conditions by VIS system are utilized in our experiments. The last three frames are collected as peak frames of the labeled expression. Thus, a total number of 1440 images in the Oulu_CASIA dataset is used to evaluate our facial expression recognition system. Similar to the process utilized on CK+ dataset, we also split the Oulu_CASIA dataset into 10 subsets and each of them is strictly subjected independent. Data from 8 subsets are employed for training and the remaining 2 subsets are adopted to validation and testing respectively. 10-fold cross-validation is also employed during the training.

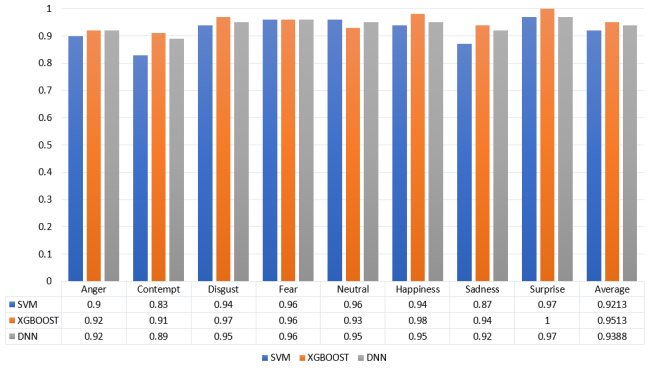


Figure 10. The recognition rate of AU combinations on different categories of facial expression on CK+ dataset.

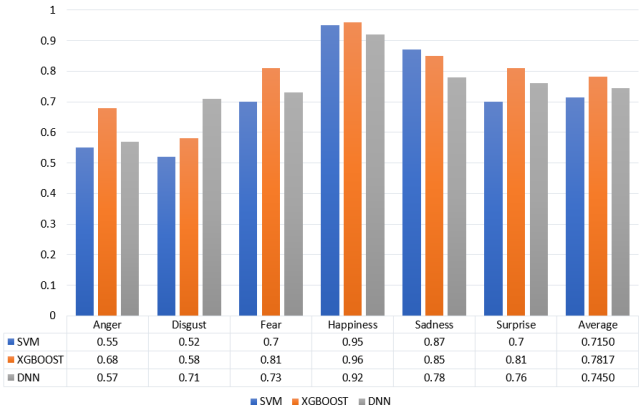


Figure 11. The recognition rate of AU combinations on different categories of facial expression on Oulu_CASIA dataset.

	AN	CO	DI	FE	NE	HA	SA	SU
AN	**0.92**	0	0	0	0.08	0	0	0
CO	0	**0.91**	0	0.05	0	0	0	0.04
DI	0	0	**0.97**	0	0	0.015	0.015	0
FE	0	0	0	**0.96**	0	0.04	0	0
NE	0	0.05	0	0.02	**0.93**	0	0	0
HA	0	0.02	0	0	0	**0.98**	0	0
SA	0.02	0	0	0	0.04	0	**0.94**	0
SU	0	0	0	0	0	0	0	**1**

Figure 12. Confusion matrix of Xgboost classification method evaluated on the CK+ dataset(AN: angry, DI: disgust, HA: happiness, SA: sadness, SU: surprise, FE: fear, CO: contempt).

	AN	DI	FE	HA	SA	SU
AN	**0.68**	0.16	0	0	0.16	0
DI	0.42	**0.58**	0	0	0	0
FE	0	0	**0.81**	0.19	0	0
HA	0.04	0	0	**0.96**	0	0
SA	0.15	0	0	0	**0.85**	0
SU	0	0	0.13	0	0.06	**0.81**

Figure 13. Confusion matrix of Xgboost classification method evaluated on the Oulu_CASIA dataset(AN: angry, DI: disgust, HA: happiness, SA: sadness, SU: surprise, FE: fear, CO: contempt).

B. Result

For the experimental evaluation, we compared the results of three different classification methods on two datasets. For the test set, we used about 40 separate images for each expression (not training), totaling 337 test images on CK+ dataset and totaling 240 test images on Oulu_CASIA

dataset. We calculated and counted the recognition accuracy of each expression, and calculated the average accuracy corresponding to the three methods. Figure 10 and Figure 11 show the details. Meanwhile, we list the confusion matrix of the Xgboost classification method on both datasets as shown in the Figure 12 and Figure 13.

It can be clearly seen that for CK+ dataset our methods achieved better results, especially the Xgboost classifier with higher accuracy. And for Oulu_CASIA dataset, the average accuracy of all three methods is relatively low. For the two datasets, CK+'s images(640 $\times$ 490) are more clear and the expressions of people are more obvious. The position of human faces is relatively fixed, while the resolution of the images in Oulu_CASIA dataset is relatively small, only 320 $\times$ 240, so the expressions are blurred. We realize expression recognition based on AUs mapping to expressions, and the accuracy of AU feature acquisition directly affects the final result. Therefore, our method requires relatively high qualities of pictures to obtain more accurate AU values, and clear pictures can often get more accurate results.

At the same time, we tested the real-time performance of facial expression recognition. We obtained face images through the camera, took each frame as the input image of the program, with size 640x480, and tested it in a noteBook computer with Intel i5-8300 processor and 8G of memory. We calculated the speed of facial expression recognition at about 30ms (33.3fps) through the input time of the image and the output time of the recognition result, and continuously output each frame containing the classification result. This proves that our method can realize video or camera facial expression recognition in real time.

Finally, we verified the effectiveness of our AU combination method by comparing the experimental results before and after the combinations of AUs on CK+ dataset. Figure 14 shows the comparison results. It can be seen that the combinations of AUs have significantly improved the recognition results of each expression.

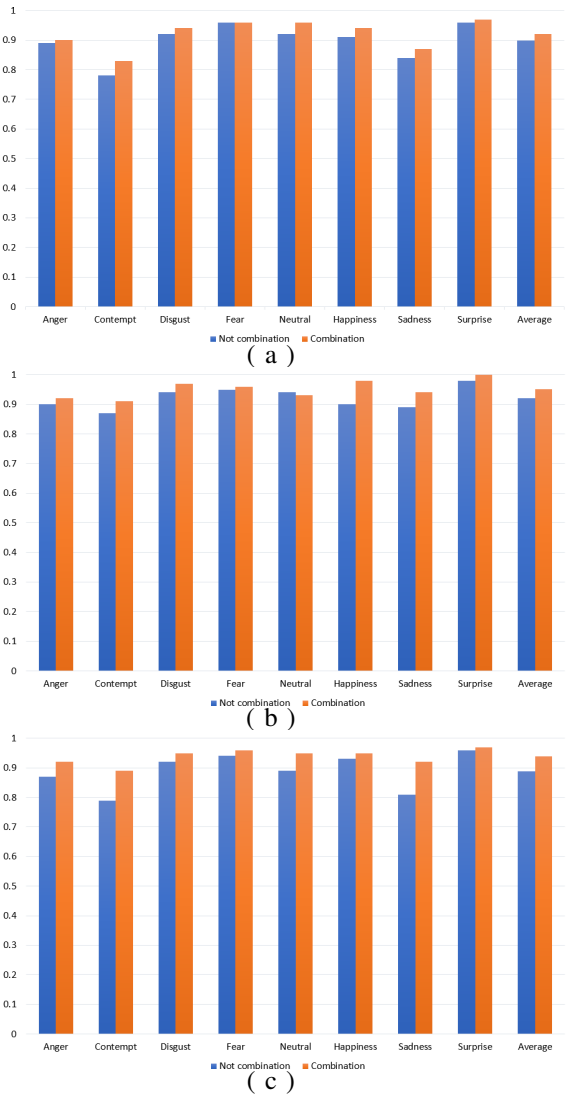


Figure 14. Comparison of the experimental results after AU combinations and the results of original data directly used for classification on CK+ dataset, the three charts are respectively the classification results of (a) SVM, (b) XGBOOST and (c) DNN.

V. Conclusion

In this paper, the mapping from facial action unit to facial expression recognition is realized. Firstly, facial landmark detection and Histograms of Oriented Gradients are used to acquire facial action units according to facial appearance and geometric features, and then different classification methods are used to train and classify the above results, corresponding to eight kinds of expressions. Experiments are carried out on benchmark datasets (CK+ and Oulu_CASIA) to verify the effectiveness of the application of facial action units in the field of facial expression recognition. With the development of technology, we hope to get facial action units more accurately and make it more widely used.

References

[1] R. Peter and E. K. Rana, "Computation of emotions in man and machines," *Philosophical Transactions of the Royal Society of London*, vol. 364, no. 1535, pp. 3441–3447, 2009.

[2] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis." in *IEEE International Conference & Workshops on Automatic Face & Gesture Recognition*, 2013.

[3] J. C. Hager, "A comparison of units for visually measuring facial actions," *Behavior Research Methods Instruments & Computers*, vol. 17, no. 4, pp. 450–468, 1985.

[4] E. Friesen and P. Ekman, "Facial action coding system: a

technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[5] X. Ding, W. S. Chu, l. T. F. De, J. F. Cohn, and Q. Wang, "Facial action unit event detection by cascade of tasks," 2013.

[6] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," 2015.

[7] L. He, C. Zou, L. Zhao, and D. Hu, "An enhanced lbp feature based on facial expression recognition." in *IEEE Engineering in Medicine & Biology Conference*, 2005.

[8] X. Wang, J. Chao, L. Wei, H. Min, L. Xu, and F. Ren, "Feature fusion of hog and wld for facial expression recognition," in *IEEE/SICE International Symposium on System Integration*, 2014.

[9] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *International Conference on Pattern Recognition*, 2010.

[10] X. Liu, B. V. K. V. Kumar, J. You, and J. Ping, "Adaptive deep metric learning for identity-aware facial expression recognition," in *IEEE Conference on Computer Vision & Pattern Recognition Workshops*, 2017.

[11] Z. Li, S. Wu, and G. Xiao, "Facial expression recognition by multi-scale cnn with regularized center loss," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3384–3389.

[12] Z. Meng, L. Ping, C. Jie, S. Han, and T. Yan, "Identity-aware convolutional neural network for facial expression recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.

[13] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *IEEE International Conference on Computer Vision*, 2015.

[14] W. Wu, Y. Yin, Y. Wang, X. Wang, and D. Xu, "Facial expression recognition for different pose faces based on special landmark detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1524–1529.

[15] A. Zadeh, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts network for facial landmark detection," in *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, vol. 3, no. 5, 2017, p. 6.

[16] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc British Machine Vision Conference*, 2006.

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision & Pattern Recognition Workshops*, 2010.

[18] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image & Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.