Deep Partial Occlusion Facial Expression Recognition via Improved CNN

Yujian Chen and Shiguang Liu(✉)

College of Intelligence and Computing, Tianjin University,
Tianjin 300350, People's Republic of China
lsg@tju.edu.cn

Abstract. Facial expression recognition (FER) can indicate a person's emotion state, that is of great importance in virtual human modelling and communication. However, FER suffers from a partial occlusion problem when applied under an unconstrained environment. In this paper, we propose to use facial expressions with partial occlusion for FER. This differs from the most conventional FER problems which assume that facial images are detected without any occlusion. To this end, by reconstructing the partially occluded facial expression database, we propose a 20-layer "VGG + residual" CNN network based on the improved VGG16 network, and adapt a hybrid feature strategy to parallelize the Gabor filter with the above CNN. We also optimize the components of the model by LMCL and momentum SGD. The results are then combined with a certain weight to get the classification results. The advantages of this method are demonstrated by multiple sets of experiments and cross-database tests.

Keywords: Facial expression recognition · CNN · Gabor · LMCL · SGD

1 Introduction

Expressions can be defined as a facial change that corresponds to a person's internal emotional state, intention or social interaction. The rich and small changes in the face can denote a variety of expressions [24]. Facial expression recognition (FER), i.e., calculation of the changes in muscles, morphology and key features of a person's face by computer, is a very active and challenging area that plays an important role in virtual human modelling and communication [4,11,15]. It has also been widely used in social robots, safe driving, public monitoring, polygraph technology, interactive games, etc., which attracted much attention in recent years [14].

Most recent research [3,10] in this area has focused on unconstrained FER by CNN. The CNN model exhibits excellent performance on this task because it is based on a combination of low-level features to find advanced features that are capable of extracting features which are robust to the changes in the training data (if sufficient changes of samples are included).

Supported by the Natural Science Foundation of China under grant nos. 61672375 and 61170118.

G. Bebis et al. (Eds.): ISVC 2020, LNCS 12509, pp. 451–462, 2020.
https://doi.org/10.1007/978-3-030-64556-4_35

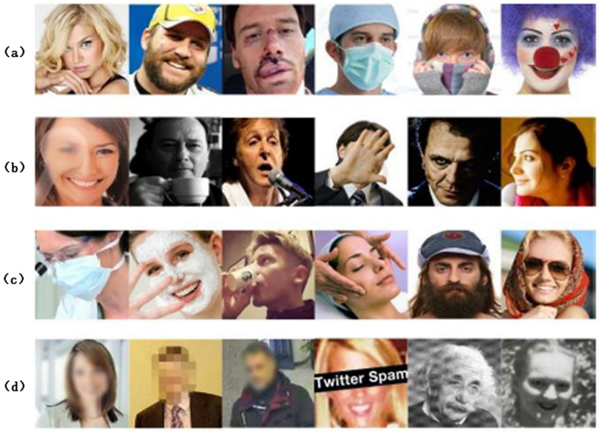

Fig. 1. Systematic occlusion (a), temporary occlusion (b), mixed occlusion (c), and other special occlusions (d).

However, in our unrestricted real life, the scenes are likely to have partial occlusion of the face. As shown in Fig. 1, facial occlusion includes systemic occlusion, temporary occlusion, mixed occlusion, and other special occlusions [19,25]. The presence of partial occlusion of the face can have two effects on facial expression recognition: first, occlusion can significantly change the visual appearance of the face, seriously affecting the performance of the FER system; secondly, occlusion can result in inaccurate feature positions or inaccurate face alignment, which increases the difficulty of extracting discriminant features from occluded faces. Partial occlusion of the face has become one of the main limitations of developing a robust FER system.

Therefore, in this work, we show that some facial occlusions still have problems for unconstrained facial recognition. This is because that most databases used for training do not provide samples of occluded faces to learn how to handle them. One way to solve this problem is to train the CNN model with a data set that contains more occluded faces. However, this task can be challenging because the main source of facial images is usually the internet, where there are fewer marked faces with occlusion.

With this in mind, we manually occluded the mouth and eyes of the KDEF [5] and JAFFE [13] datasets and added them as occluded samples. Based on the improvement of VGG16, a 20-layer deep CN network with two residual blocks is designed, which can solve the problem of degradation of deep networks. In the network, by fedding the image of the occluded eyes and mouth into the network for training, the model can adapt to different occlusion areas, and the strong occlusion feature is learned from the unoccluded area. Excellent results have been achieved in training and testing. In order to solve the problem of cross-database verification caused by insufficient amount and insufficient feature extraction of partially occluded facial expressions data. We combine the Gabor filter and CNN and obtain their respective classification results through parallel processing. We then combine the above two parts according to a certain weight to produce the final classification vector. Secondly, the traditional Softmax Loss in deep CNN can make the calculation of the model more stable, that is beneficial for optimizing the distance between classes. It is weaker, however, for the optimization class distance and the feature distinguishing ability is also insufficient. To solve this problem in our FER system, we applied the latest, validated and well-functioning LMCL (Large Margin Cosine Loss) [21] to our model and optimized the classification results. We also use momentum SGD (Stochastic Gradient Descent) to optimize the traditional SGD, thus speeding up the gradient descent optimization process. Through comparative experiments, we evaluated the relationship of the recognition rates in different conditions.

2 Related Work

The study of facial expressions began in the 19th century. In 1872, Darwin described the connection and difference of facial expressions between humanity and animals [20].

In recent years, the latest research and development of convolutional neural networks [12,22] has greatly improved the effects of various computer vision tasks, which has made DCNN a mainstream machine learning method for computer vision. As one of the most common computer vision tasks, facial expression recognition has been widely studied and become a new hot research direction. The early research was based on traditional methods and shallow models of low-level features, and today's facial expression recognition has made great progress under the impetus of DCNN. However, most of the previous researches on facial expression recognition are based on the premise of unobstructed conditions in the laboratory. In real life, there are often various occlusions on the face, which has become a major problem of FER. The bottlenecks and challenges of the system have also attracted attention of more and more researchers.

2.1 Facial Expression Recognition Without Occlusion

Most of the related researches are based on FER without occlusion under laboratory constraints. Traditionally, Facial Action Coding System (FACS) [3] expresses different emotions by describing subtle changes in facial expressions, which can better describe the details of facial expressions.

In terms of deep learning methods, Liu et al. [10] proposed a new method called elevated deep confidence network (BDBN). Their experiments were performed in Cohn-Kanade and JAFFE, with the accuracy of 96.7% and 68.0%, respectively. However, this method ignores the correlation between different facial areas.

2.2 Facial Expression Recognition with Partial Occlusion

As mentioned above, in the early investigation of facial expression recognition, no research has been reported on overcoming facial occlusion. Nevertheless, due to the recognition of the great influence of facial occlusion on facial expressions, many studies have begun to try face facial expression recognition with partial occlusion. It can be seen from the existing research that FER methods under partial occlusion can be divided into two categories, traditional methods and deep learning methods.

Traditional Methods. Kotsia et al. [9] used Gabor wavelet texture information extraction based on discriminant non-negative matrix factorization and shape-based methods to classify partial occlusion images. Hammal and Arguin [6] proposed the improved transferable belief model (TBM) for FER under partial occlusions. But the recognition of sadness is not high (25%). FER based on traditional methods are often affected by the extraction of more occluded physical features.

Deep Learning Methods. Zoltan et al. [26] used an 8-layer CNN for action unit detection to detect self-occlusion caused by large attitude changes in 3D pose data, but it may not be the best to use CNN alone. Brink and Vadapalli [1] proposed a hybrid model by combining the variable processing model and CNN. This model greatly reduces the error rate of feature extraction, but they do not pay more attention to occlusion expression recognition. Cheng et al. [2] proposed the deep structure of facial expression recognition with partial occlusion, and extracted features from face images using the Gabor filter as input to 3-layer deep Boltzmann machine for sentiment classification. The accuracy reaches up to 82%. The performance of this work still has room to improve, and the amount of database used is only 213, which is not enough for learning the deep structure. Lu et al. [23] proposed a Wasserstein Generative Adversarial Network-based method to perform occluded FER. This method consists of a generator G and two discriminators D1 and D2. FER is completed by introducing the classification loss into D2.

3 Method

The flow chart of the models is shown in Fig. 2. We propose a two part model structure. Given source images, we first use the improved CNN network and the Gabor filter for parallel processing, and then combine the two resulting vectors by weighting to obtain the classification result. The first part is a 20-layer CNN. The input is a 256 × 256 image, which is preprocessed and then fed to the VGG16, followed by a 4-layer residual network at the back of the network. In the second part, we combine the traditional Gabor filter with Adaboost to construct another processing classifier. Finally, the classification results of the two parts are combined by a certain weight (DCNN + Gabor). In addition, we use LMCL to optimize the traditional Softmax Loss, and use the SGD with momentum to optimize the traditional SGD, and the results of our model classification are further optimized.

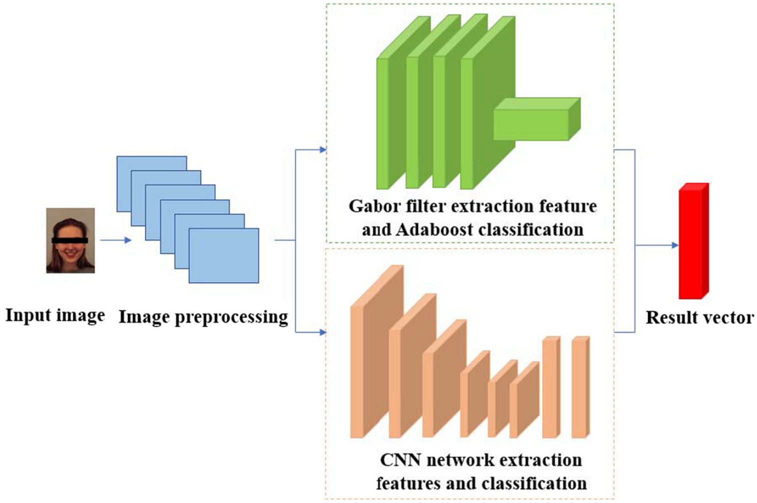
Gabor filter extraction feature
and Adaboost classification
Input image
Image preprocessing
Result vector
CNN network extraction
features and classification

Fig. 2. Overview of our approach.

3.1 Data Set

We first constructed partial occlusion dataset for occlusion of the eyes and mouth using the public no-occlusion data sets JAFFE [13] and KDEF [5], then classify it as 7 expressions (i.e., angry, disgusted, fearful, happy, sad, surprised and neutral) for comparative testing. We denote their abbreviations as AN, DI, FE, HA, SA, SU and NE, respectively. We will process all the images into 256 × 256 and perform a series of data preprocessing, including standardizing the space and augmented data and generating synthetic images by artificially rotating real images [17].

3.2 CNN Network with Residual Block

As shown in Fig. 3, the first part of our models is based on the VGG16 [18] architecture, which consists of 13 convolutional layers and 3 fully connected layers, where the convolutional layer is divided by 5 max-pooling layers. Note that the filter size of all convolutional layers is 3 × 3 and the step size is 1 for padding. The input is a preprocessed image that passes through multiple convolutional layers, pooling layers and two residual blocks. Then, after processing through the fully connected layer and Softmax, a seven-dimensional output vector is obtained. Finally, the resulting expression is judged based on the probability value.

During training facial expressions, when the number of network layers is deepened, it may cause gradient dispersion or gradient explosion problems. This problem may be solved through regularization, however, another one still remain unsolved, i.e., the degradation of the network. Although the number of network layers increases, the accuracy of the training set is saturated or even decreased. To this end, He et al. [7] proposed a residual network and proved to achieve

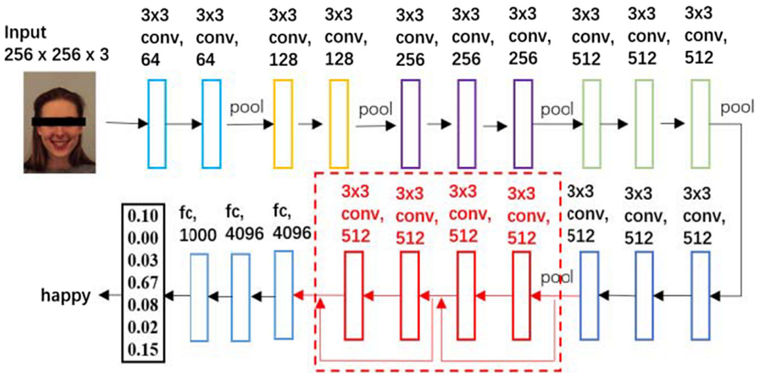

Fig. 3. Illustration of the proposed network.

good results in deep networks. Inspired by their work, we add a 4-layer residual network to VGG16 to improve the network. It can solve the problem of degradation of a deep network, so that the model can make good use of the depth of the network, and the learning of the training set is more sufficient.

The residual block transfers the input x by adding a jump connection between every two layers. The residual map $F(x)$ is constructed at the same time. Finally, $F(x)$ is summed with x to obtain the result map $H(x)$. The optimization of the network can also avoid the instability of the training error when the network is deepened, which can help alleviate the problem of gradient disappearance and gradient explosion, and also ensure the good performance of the network and avoid the network becoming “bloated”.

3.3 Hybrid Model Structure Combined with Gabor Filter

The second part of our model structure is a facial expression recognition system based on the Gabor filter and the Adaboost classifier. The Gabor filter can be regarded as a feature filter that can better imitate the human visual system. It can well describe the feature information of facial expression images, including texture features, edge features and directional features. The process of this model is described below.

Image Preprocessing. We mainly perform two pre-processing steps, i.e., gray normalization and geometric normalization. First, the Sobel operator [8] is used to extract the edge of the facial feature region, and the automatic threshold binarization method [16] is employed to obtain the segmentation graph of the facial features. PCA (Principal Component Analysis) is introduced to select true facial features, combined with the constraints of a priori relationship between facial features and face size. We can obtain important position parts such as nose, eyes, mouth, cheeks, lower bars and foreheads. Then, the feature images corresponding to the feature regions are extracted from the facial expression image.

Feature Extraction. In this paper, Gabor filter banks with different scales and different directions are used. Different parameters are selected according to actual requirements, and they are convoluted with the preprocessed image to obtain the filtered Gabor features. Specifically, the filter bank used here includes a total of 40 filters consisting of 5 different scales and 8 different directional Gabor filters.

Feature Selection and Expression Classification. Due to the characteristics of the Gabor filter, the feature dimension obtained after passing through the Gabor filter bank will be very large, and if it is directly passed as input to the classifier, the amount of computation will be very large. And the high-dimensional feature vector contains more redundant information, which will have a great impact on the final classification accuracy. Therefore, Adaboost is used for feature selection and the classifier. The weak learner is the classification and regression tree (CART). Through each round of adjustment of the sample weights, we learn 10 weak learners in series, and finally get a series of classifier parameters and weak learners. The weight value is then combined with the weight value according to a certain strategy to obtain the final classification result.

As shown in Fig. 2, we obtain a hybrid structure consisting of the CNN method and the Gabor method. By linearly combining the resulting vectors obtained by the above two models, the final classification result can be obtained. By experiments, the weight of the model structure of CNN and that of the Gabor filter model are set as 0.6 and 0.4, respectively.

3.4 Component Optimization for Deep Networks

In the optimization of the model structure, we replace the traditional Softmax Loss with LMCL, so that the loss function has great feature classification ability to optimize the classification accuracy. We also improve the traditional SGD as the momentum SGD in order to greatly accelerate the convergence process of the model.

4 Experiments

In this section, we introduce the experimental environment, data sets, model frameworks, and training processes, etc. We performed the experiments using the model described above.

4.1 Implementation

Our model is based on the Tensorflow framework, which uses the GPU (Graphics Processing Unit) for parallel computing. We run our method under a 6G discrete graphics card. The experimental data set is a mixed data set of KDEF and JAFFE using artificial occlusion. Among them, the 90% of the total data are

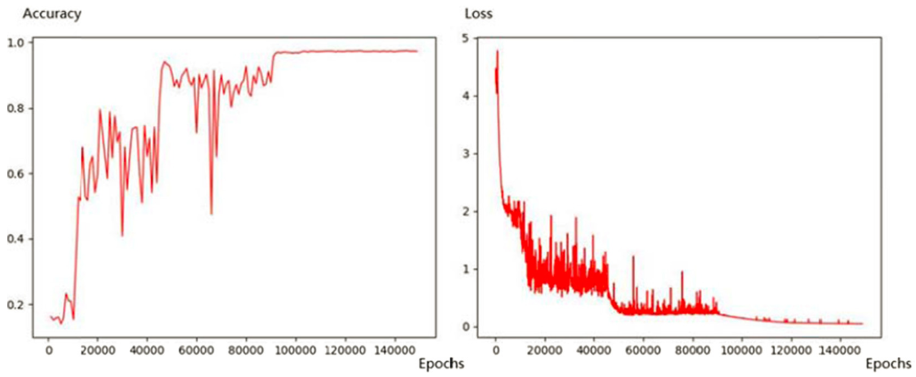

Fig. 4. The testing accuracy (left) and testing loss (right).

served as training set, and the remaining is test set. The size of each image is processed to 256 × 256. After a series of pre-processing, the image is fed into the model for iterative training.

As shown in the left image of Fig. 4, the total iteration of testing reaches more than 100,000 times. In the beginning, the accuracy of the testing was low. After about 1000 rounds of fine-tuning, the accuracy quickly rises to be around 50%. Then the accuracy fluctuated and reaches about 90% after about 5000 iterations. Finally, after constant fluctuations and small adjustments, the accuracy of the testing set can reach higher values. When the testing iteration is approximating 90,000 times, the testing accuracy reaches a peak. Corresponding to this stage, as shown in the right image of Fig. 4, the loss value of the testing is also decreasing until it reaches a minimum at the 90,000th iteration.

Table 1. The confusion matrix of our model on the KDEF database.

%	AN			DI			FE			HA			SA			SU			NE		
	NO	EY	MO	NO	EY	MO	NO	EY	MO	NO	EY	MO	NO	EY	MO	NO	EY	MO	NO	EY	MO
AN	**83.5**	**80.5**	**80.2**	5.9	10.7	9.2	1.1	0	0	0	0	0	6.4	9.1	7.5	1.0	2.5	0	0	0	0
DI	8.9	7.0	6.4	**86.1**	**80.6**	**81.9**	0	0	0	0.1	0	0	2.4	1.6	2.6	0	0.7	3.1	0	0	0
FE	0	1.5	4.7	0	0	0	**91.8**	**89.1**	**86.0**	1.5	4.7	7.6	0	0	0	3.2	2.5	4.4	0	0	0
HA	0	0	0	0	0	0	5.2	5.9	5.1	**98.3**	**95.3**	**88.5**	1.5	2.4	5.6	0	0	0	1.6	0.5	0.9
SA	7.3	11.0	8.7	7.6	7.0	5.9	1.9	5.0	8.9	0	0	3.9	**89.6**	**86.9**	**85.9**	0	0	1.1	1.5	2.6	0.8
SU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	**95.8**	**90.3**	**91.4**	1.3	3.4	2.7
NE	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0.5	0	1.0	0	**95.9**	**93.5**	**95.6**

4.2 Experimental Results and Analysis

KDEF. As shown in Table 1, the accuracy of 91.6%, 88.3 and 87.5% was obtained in the cases of no-occlusion, occlusion of the eyes and occlusion of the mouth, respectively. Note that the average accuracy of the JAFFE dataset is slightly smaller than KDEF, but the overall law is consistent, which is not

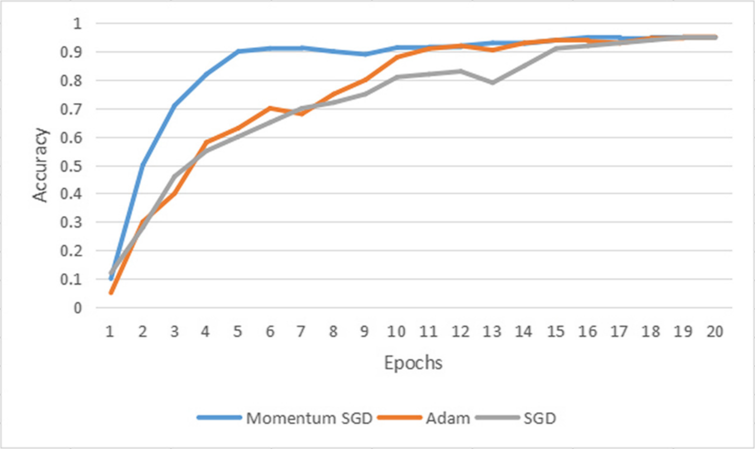

Fig. 5. Convergence time of different optimization algorithms.

listed here. It can be found that, in general, mouth occlusion will have a greater impact than eye occlusion, which also means that the mouth is more important in expression recognition. Mouth occlusion has a greater impact on the recognition accuracy of expressions of angry, fearful, happy and sad, while eye occlusion has a greater impact on the recognition accuracy of disgusted, surprised and neutral expressions.

Specifically, for angry, disgusted and sad, these three types of expressions are more likely to be confused when judged; happy is the expression with the highest accuracy. Relatively speaking, some happy expressions are recognized as fearful. This may be due to that these two expressions have some similarities in the change of facial morphology; neutral is not easily misjudged by other expressions, and their accuracy is relatively high and more stable under various circumstances; surprise is more likely to be misjudged as fearful and angry.

4.3 Evaluation of the Effects of Different Optimization Algorithms

In order to evaluate the performance of SGD, we select SGD, Adam and RMSProp for comparison and perform the above mentioned DCNN with Gabor. As shown in Fig. 5, the results show that there is great improvement in terms of the accuracy of training. Note that the improvement may be restricted by the amount of training data. Moreover, the training speed and gradient reduction process of the whole model have been greatly optimized.

4.4 Comparison with the State-of-the-Arts

As shown in Fig. 6 and Fig. 7, we can find that our experimental results are effective and robust, especially under occlusion compared with LGBPHS, Deep Nonlinear Network Structure, Gabor and DNMF.

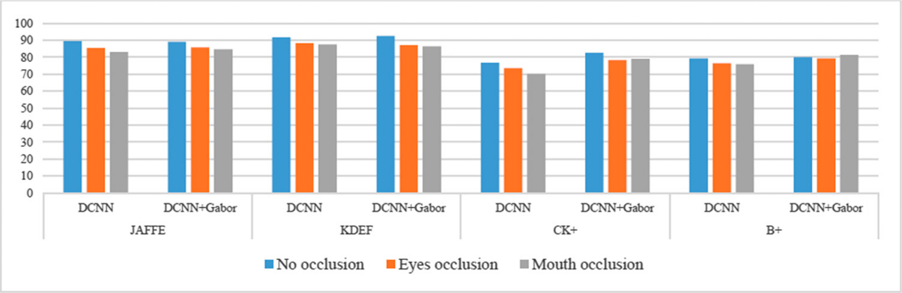

Fig. 6. Comparison among different methods in cross-data validation.

JAFFE Dataset. It can be observed from the results that all seven methods have higher precision under no-occlusion, both exceed 85%. Our approach has slight advantage, up to 89.5%. Under partial occlusion, each method has a different degree of reduction, especially the occlusion of the mouth. The results show that our method has excellent accuracy in eye occlusion and mouth occlusion, which arrive 85.5% and 83.1%, respectively. The accuracy is slightly lower than no occlusion condition, but is superior to all other methods. Among other methods, the Gabor method achieves better results in the case of no occlusion. As for partial occlusion, the unmodified VGG16 and LGBPHS perform slightly lower than other methods.

KDEF Dataset. We found that the accuracy under the KDEF dataset is higher than that under JAFFE, and the improvement of the model brings the improvement of accuracy, which shows that the improvement of our algorithm is effective. This may be owing to more explicit representation of the KDEF dataset and the larger amount of data than JAFFE.

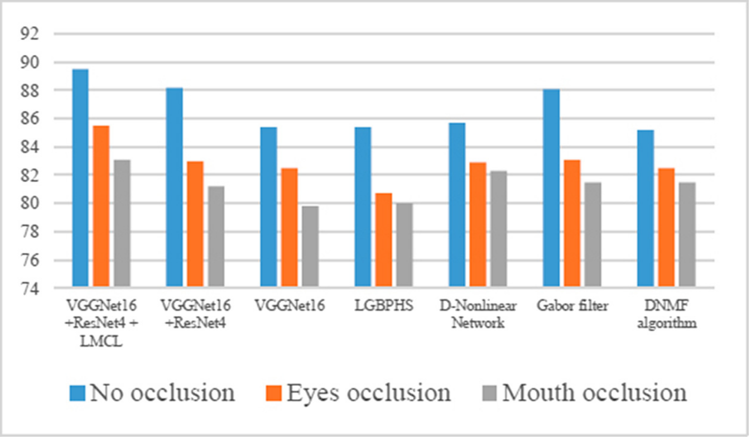

Fig. 7. Comparison between the proposed method and other methods in JAFFE.

5 Conclusion and Future Work

This paper has proposed a novel deep learning method for FER with various partial occlusion. We first reconstructed the facial expression data set under partial occlusion and perform preprocessing. We developed a novel "VGG + residual" CNN network based on the improved VGG16 network, and adapt a hybrid feature strategy to parallelize the Gabor filter with the above CNN. We used LMCL and momentum SGD to improve the traditional methods. Our model can greatly accomplish FER under various partial occlusion and achieve higher recognition accuracy than the state-of-the-arts. We have found out from the statistical results that the mouth occlusion has greater impact on the resulting recognition accuracy than that of eye occlusion.

Our method is not without limitation. The amount of the partially occluded facial expression data is not large enough. We will introduce more challenging partially occluded facial expression data under natural state into our data set. The recognition accuracy and efficiency still have room for improvement. It may work by introducing more advanced optimization strategies into our network. We will also attempt to apply our method in virtual human modelling systems. It would endow a virtual human more ability to "see" one's emotion by accurately recognizing various facial expressions, and thereby benefit the communication in the virtual world.

References

1. Brink, H., Vadapalli, H.B.: Deformable part models with CNN features for facial landmark detection under occlusion. In: ACM Press the South African Institute of Computer Scientists and Information Technologists, pp. 681–685 (2017)
2. Cheng, Y., Jiang, B., Jia, K.: A deep structure for facial expression recognition under partial occlusion. In: IEEE Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 211–214 (2014)
3. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124–129 (1971)
4. García-Rojas, A., et al.: Emotional face expression profiles supported by virtual human ontology: research articles. Comput. Animation Virtual Worlds **17**(3–4), 259–269 (2006)
5. Goeleven, E., De-Raedt, R., Leyman, L., Verschuere, B.: The Karolinska directed emotional faces: a validation study. Cogn. Emot. **22**(6), 1094–1118 (2008)
6. Hammal, Z., Arguin, M.: Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions. J. Vis. **9**(2), 22–28 (2009)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–12 (2015)
8. Kanopoulos, N., Vasanthavada, N., Baker, R.L.: Design of an image edge detection filter using the Sobel operator. IEEE J. Solid State Circuits **23**(2), 358–367 (1988)
9. Kotsia, I., Zafeiriou, S., Pitas, I.: Texture and shape information fusion for facial expression and facial action unit recognition. Pattern Recogn. **41**(3), 833–851 (2008)

10. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2014)
11. Liu, S., Yang, X., Wang, Z., Xiao, Z., Zhang, J.: Real-time facial expression transfer with single video camera. Comput. Animation Virtual Worlds **27**(3–4), 301–310 (2016)
12. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: International Conference on Machine Learning, pp. 507–516 (2016)
13. Lyons, M.J., Akamatsu, S., Kamachi, M.: Coding facial expressions with Gabor wavelets. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998)
14. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1445 (2000)
15. Qiao, F., Yao, N., Jiao, Z., Li, Z.: Emotional facial expression transfer from a single image via generative adversarial nets. Comput. Animation Virtual Worlds **29**(6), e1819 (2018)
16. Shi, J., Ray, N., Zhang, H.: Shape based local thresholding for binarization of document images. Pattern Recogn. Lett. **33**(1), 24–32 (2012)
17. Simard, P., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition, pp. 958–963 (2003)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference of Learning Representation, pp. 1409–1417 (2014)
19. Towner, H., Slater, M.: Reconstruction and recognition of occluded facial expressions using PCA. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 36–47. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74889-2_4
20. Wallace, C.: A note on Darwins work on the expression of the emotions in man and animals. J. Abnorm. Psychol. Soc. Psychol. **16**(5), 356–366 (1921)
21. Wang, H.: Cosface large margin cosine loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1801–1807 (2018)
22. Wen, Y., Zhang, K., Li, Z., Qiao, Yu.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
23. Yang, L., Wang, S., Zhao, W., Zhao, Y.: Wgan-based robust occluded facial expression recognition. IEEE Access **7**, 93594–93610 (2019)
24. Zhang, L., Brijesh, V., Dian, T., Vinod, C.: Facial expression analysis under partial occlusion: a survey. ACM Comput. Surv. **51**(2), 1–49 (2018)
25. Zhuo, J., Chen, Z., Lai, J., Wang, G.: Occluded person re-identification. In: International Conference on Multimedia and Expo, pp. 1–6 (2018)
26. Tősér, Z., Jeni, L.A., Lőrincz, A., Cohn, J.F.: Deep learning for facial action unit detection under large head poses. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 359–371. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_29