Paper:

Facial Expression Recognition with Feature Matrix for Low-resolution Images

Jian Wang, Kaoru Hirota, Yaping Dai, Zhiyang Jia*

School of Automation, Beijing Instituet of Technology
E-mail: wang_jian@bit.edu.cn

Abstract. Low-resolution (LR) facial images usually lack enough visual information for feature extraction, which increases the difficulty of expression recognition and reduces the recognition accuracy. In this paper, we propose a new model framework for LR images, which main idea is to extract the emotion feature matrix of the images and add it to the convolutional neural network (CNN). More specifically, the emotion feature matrix determined by analyzing the distribution of salient regions in the dataset can eliminate redundant regions and retain salient regions. The emotion feature matrix is used to construct CNN model to extract limited features from LR images, which can enhance the expression of salient region features. Experimental results on several facial expression datasets, including CK+, JAFFE and FER2013 show the superior performance of the proposed method for LR facial expression recognition, compared with several state-of-the-art methods.

Keywords: Expression recognition; Low-resolution; Convolutional neural network; Emotion feature matrix; Salient region

1. Introduction

In the field of computer vision and image recognition, facial expression recognition (FER) has received extensive attention in recent years. With the extensive application of deep learning, especially CNNs, in the field of image recognition, FER has higher recognition rate, speed of identification and accuracy [1, 2]. More and more high-precision expression recognition methods are applied in human-computer interaction system (HCI) [3], safe driving and other fields. Although FER methods with high precision and strong robustness emerge in endlessly, LR images often have a negative impact on the algorithm performance due to the lack of sufficient visual information [4]. LR images are common in the field of effective computing, such as the commonly used facial images in CK+ and JAFFE datasets, the face images far away from the camera, etc. How to make full use of the feature information of LR images is the key to improve the recognition rate of LR images.

At present, there are various methods to solve LR image expression recognition problem, which can be roughly divided into two strategies: (1) Super-resolution based method [5–7] and (2) facial image representation [8–11].

Super-resolution based methods for face/expression recognition are all related to HR images. On the one hand, reconstruct HR images from LR images. For example, Lim et al. [12] set up the deep super-resolution network (EDSR) by removing unnecessary modules in conventional residual networks, which HR images reconstructed by EDSR get significant performance improvement. On the other hand, learn the relationship between LR and HR images. Zou et al. [13] and Haghighat M et al. [7] respectively utilize different methods to extract LR images and HR images features, analyze their relationship, and achieve good performance in face recognition; Dong et al. [6] employ the CNN to learn a nonlinear mapping function between LR and HR images based on a large scale image dataset. Clearly, other ingenious methods can also obtain the required information by processing LR and HR images. Chu et al. [5] use cluster-based regularized simultaneous discriminant analysis (C-RSDA) method to map the high and low resolution images features to the same feature subspace, which enhance the discriminative power of the feature subspace. Xing et al. [14] project different resolution image sets into a unified feature subspace, and then they use two heterogeneous sample sets to construct generalized bipartite graph, which contains more complete information for classification. The methods introduced above solve the problem of LR images recognition rate by connecting HR images. However, this strategy is computationally inefficiency, and it is difficult to ensure that the reconstructed or connected HR images is the best image for expression recognition, and the effective feature information of the image cannot be fully utilized.

Facial image representation contains a variety of methods, mainly for facial feature processing. Traditional facial feature extraction methods mainly include local binary patterns (LBP), Haar like features [15] and Gabor wavelet features. These methods have been successfully applied to FER problem and achieved excellent performance. With the rapid development of neural network especially CNN, the methods of feature extraction based on CNN emerge in endlessly, which can not only extract surface features, but also show strong advantages in depth feature extraction. For example, Shao et al. [1] design three CNNs based on the study of various neural net-

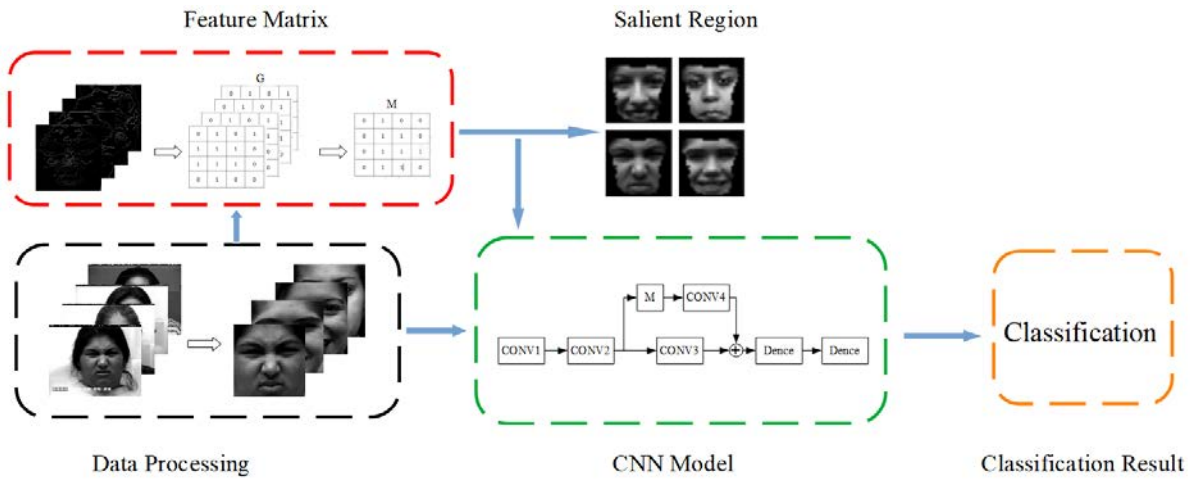


Fig. 1. Model framework

works, and prove their effectiveness on different datasets. In addition, feature dimensionality reduction such as linear discriminant analysis (LDA) and principle component analysis (PCA) can reduce the complexity of features and achieve better classification performance.

All of the above methods take the face as a whole, and do not emphasize the influence of salient regions on the recognition. In fact, the salient regions including eyes, mouth, eyebrows, etc. can cover almost all the critical information that can be used for expression recognition. Too much redundant feature information will have adverse effects on recognition and classification. Therefore, many methods [8, 9, 11, 16] emphasize and attach importance to the role of salient regions. Chen et al. [16] propose an attention model for image segmentation. Xie et al. [9] build a deep attention multi-path CNN (DAM-CNN) model with this model, which can not only automatically evaluate the importance of different facial regions, but also generate a variation robust representation for expression classification.

Motivated by these works, this paper proposes a new model to make full use of the features of salient regions to strengthen the training process, as shown in **Fig. 1**. It is mainly divided into four parts, including dataset preprocessing, emotion feature matrix acquisition, constructing deep CNN with emotion feature matrix for feature extraction and classification. The main contributions of our work are summarized as follows:

(1) A new model framework is proposed to deal with FER for LR images with limited features. Different from the traditional methods, we use the emotion feature matrix to make full use of the salient regions features.

(2) The feature matrix is used to design CNN to increase the proportion of useful feature information when extracting emotion features, which can enhance expression of the salient regions.

(3) Experiments with superior performance on different datasets are carried out and compared with several state-of-the-art method.

2. Proposed method

Our proposed model is mainly divided into four steps. The first step is data preprocessing to align and cut the face images. The second step is to obtain the feature matrix to enhance the expression of salient face regions. The third step is to construct a CNN by using this matrix for feature extraction. The last step is to apply the commonly used softmax regression model to classify. The main contents of the first three steps are described in detail below.

2.1. Data preprocessing

The quality of data directly affects the training results, so before the experiment, it is necessary to process the data according to the requirements. For FER, we pay more attention to the facial regions, while other regions such as hands, neck, even hair and forehead can be considered useless. In the datasets commonly used for expression recognition, such as CK+ and JAFFE datasets, there is a large amount of redundant information to be eliminated for expression recognition tasks, so it is necessary to intercept the face region in the image and scale it to the appropriate size according to the experimental requirements. In addition, in order to obtain the feature matrix with better performance, we will use the feature point detection technology [17] to locate the the five sense organs, and then complete the face alignment operation.

2.2. Emotion feature matrix

When extracting facial image features, not all the features of the facial region are conducive to the recognition of facial expressions, and only the feature of salient regions are critical to the recognition. However, most of the CNN extract the features of facial images without difference, which inevitably leads to the increase of redundant feature information and enhance the difficulty of classification.

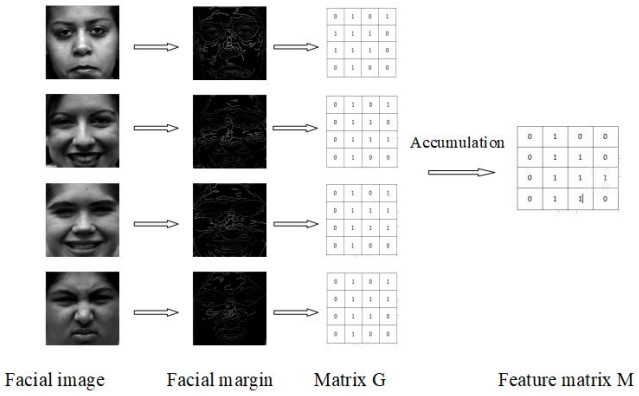


Fig. 2. Emotion feature matrix acquisition

In order to increase the proportion and expression of salient regions in feature extraction process, we use feature matrix to construct CNN. The framework of feature matrix acquisition method is shown in **Fig. 2**. Firstly, we randomly select a certain number of facial images from dataset. The uniform size images are divided into $N \times N$ feature blocks after determining the dimension parameters of emotion feature matrix according to the input image size and CNN model. Next the elements in single image matrix $\boldsymbol{G}$ are filled by analyzing the edge information of the feature blocks at corresponding positions. In the end, every matrix $\boldsymbol{G}$ is accumulated and processed to obtain the emotion feature matrix $\boldsymbol{M}$, which can be formulated as follows:

$$\boldsymbol{M} = f\left(\sum_{i=1}^{l} \boldsymbol{G} + \boldsymbol{B}\right) \tag{1}$$

where l is the number of single image matrix $\boldsymbol{G}$ and $\boldsymbol{B}$ is the correlation bias, which is convenient for human control feature extraction; $f(\alpha)$ is a nonlinear function, which is expressed as follows:

$$f(\alpha) = \begin{cases} 1, \alpha \geq \beta \\ 0, \alpha < \beta \end{cases} \tag{2}$$

where β is the threshold used to control the elements in emotion feature matrix $\boldsymbol{M}$. The elements in the feature matrix $\boldsymbol{M}$ can roughly estimate the importance of each region. When CNN is trained, it can increase the proportion and enhance the expression of the salient region features, and finally have a favorable impact on the final classification task.

2.3. CNN based emotion feature matrix

LR images lack sufficient visual feature information for classification, so a limited number of features must be fully utilized. The salient region contains almost all the features that can distinguish the expression category, which we need to pay more attention to the feature extraction of the salient regions. The feature matrix $\boldsymbol{M}$ is used to construct CNN, which the features of the salient region are extracted twice. After that, the output is processed into a unified dimension through maximum pooling layer, which is convenient for the addition of extracted features. The specific CNN model structure is shown in **Fig. 3**. In the emotion features extracted by this CNN model, the proportion of salient regions features is larger than that of redundant regions, which helps the classification and recognition of Softmax regression model.

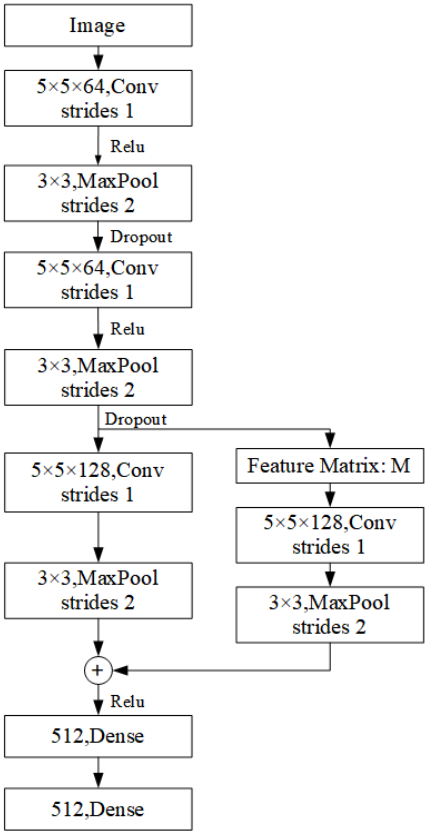


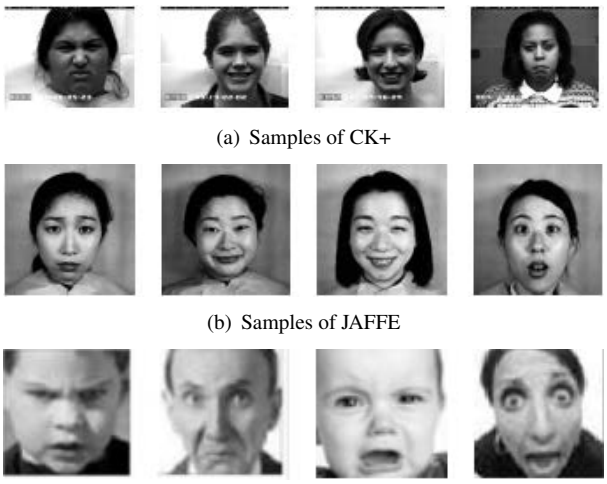
Fig. 3. CNN model

3. Experimental evaluation

In order to verify the effectiveness of our proposed method, experiments are carried out on three commonly used datasets. Some image samples are shown in **Fig. 4**. The experimental details will be described in this section.

3.1. Performance of feature matrix

After preprocessing the CK+ and JAFFE datasets, the facial images were aligned and scaled to a uniform size of 128×128. However, due to the complexity of the content of the FER2013 dataset, we do not preprocess it. According to the CNN model, the dimension of feature matrix is determined. The images randomly selected from the CK+ and JAFFE datasets are evenly divided into 32×32 feature blocks (16×16 for FER2013). Do edge extraction

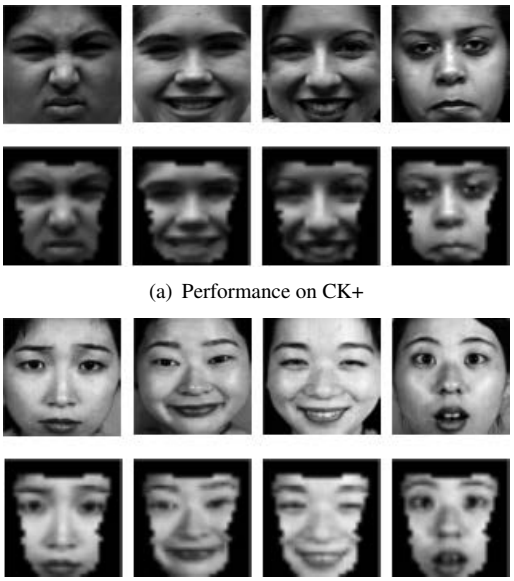


(a) Samples of CK+

(b) Samples of JAFFE

(c) Samples of FER2013

Fig. 4. Samples of three datasets



(a) Performance on CK+

(b) Performance on JAFFE

Fig. 5. Performance of feature matrix

and obtain the edge information of each feature block. For the feature blocks with edge information, the corresponding position coefficient in the single image feature matrix $\boldsymbol{G}$ is set to 1, else without edge information, it is set to 0. According to the above Equation (1), the feature matrix $\boldsymbol{M}$ is obtained by matrix $\boldsymbol{G}$ accumulation and analysis. The parameters in Equation (1) and (2) are set as follows: Number of images $l = 50, \beta = 25$. Matrix $\boldsymbol{B}$ is obtained by experience, in that the shape of facial five senses is simply outlined. The position coefficient of five senses is set to 5, and the rest positions are set to 0. The performance of feature matrix on partial dataset images is as shown in **Fig. 5**. It can be seen that the salient regions are retained and the redundant feature regionss are discarded, which indicates that the feature matrix can ensure the feature extraction of salient regions.

3.2. Experiments on three popular datasets

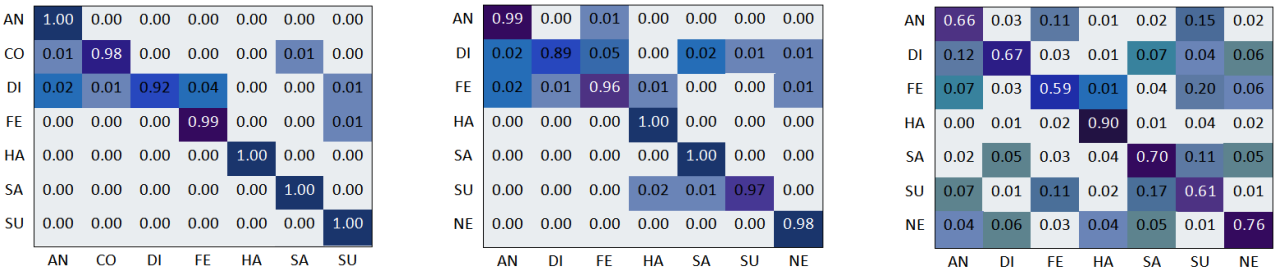
We conduct experiments on the widely used datasets of CK+, JAFFE and FER2013. The CK+ dataset includes expressions of seven labels: anger, contempt, disgust, fear, happy, sadness and surprise. The JAFFE and FER2013 datasets have seven labels: anger, disgust, fear, happy, sad, surprise and neutral. Each dataset is randomly divided into training set, verification set and test set. In the training process, we randomly rotate the face image for image enhancement, and the rotation angle is less than 10° to avoid weakening the effect of the feature matrix.

To evaluate the overall performance, the confusion matrices of our proposed methods on three datasets are illustrated in **Fig. 6**. The method achieves good performance on all the expressions. From the confusion matrix of CK+ dataset, we can see that for anger, happiness, sad and surprise expressions, we have a good recognition accuracy. However, we can also observe that some samples corresponding to the disgust expressions are misclassified as the angry expression, and similar situations are found in other datasets. From the confusion matrix of JAFFE and FER2013 datasets, there are some instances that their true label is angry but the classifier has misclassified it as fear or surprise. This is because some of the features of these expressions are very similar, for example, both the disgust and sad expressions have the wrinkled eyebrows. It's easy to misunderstand people when they express anger and surprise, which is likely to be confused by the trained networks.

Moreover, we plotted the obtained accuracy and loss of CK+ to observe the training process during epochs in **Fig. 7**. Clearly, the training accuracy curve rises steadily. Although the verification accuracy fluctuates, it shows an obvious upward trend and finally tends to be stable. Similarly, the loss curve of the training process decreased steadily, while the verification loss curve fluctuated slightly but tended to decline steadily.

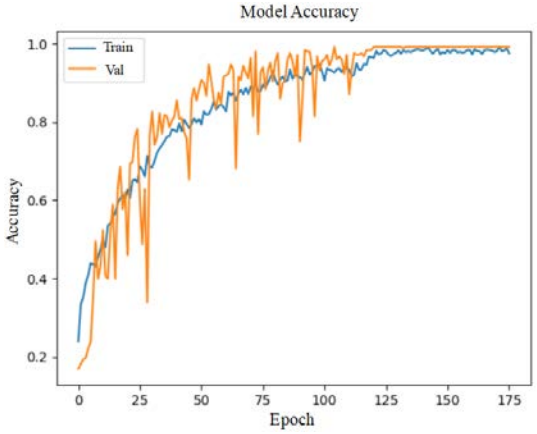
3.3. Comparisons with the state-of-the-art methods

We compare the proposed method with several state-of-the-art methods. These methods include the traditional classification methods (such as PCA [18], multi-class LDA [18], LBP [19]) and deep learning methods (such as Light-CNN [1], DAM-CNN [9] and IFSL [10]) to compare for obtain the comparison results of recognition accuracy on the same datasets. The choice of these competing methods is based on the following reasons: 1) PCA, multi-class LDA and LBP are three widely-used traditional methods for FER and achieve excellent performance. 2) Light CNN and other methods including our proposed method use CNN for feature extraction and classification. 3) DAM-CNN designs a deep multi-path convolutional neural network by taking advantage of salient region attention, which is similar to the idea of feature matrix. 4) IFSL also aims at the expression recognition of LR images. Tables list the accuracy of our proposed and the



(a) CK+ (b) JAFFE (c) FER2013

Fig. 6. Confusion matrices on three expression databases



(a) Accuracy curve

(b) Loss curve

Fig. 7. Training curve of CK+

Table 1. Comparison results obtained on the CK+ dataset.

Methods	Accuracy(%)
PCA(k-NN) [18]	43.8
PCA (SVM) [18]	47.3
multi-class LDA (k-NN) [18]	84.7
multi-class LDA (SVM) [18]	87.1
m-LBP [19]	88.4
DAM-CNN [9]	95.9
Light-CNN [1]	92.9
dual-branch CNN [1]	85.7
pre-trained CNN [1]	95.3
IFSL (SVM) [10]	98.7
IFSL (k-NN) [10]	96.6
Our method	**98.9**

Table 2. Comparison results obtained on the JAFFE dataset.

Methods	Accuracy(%)
PCA(k-NN) [18]	52.4
PCA (SVM) [18]	55.6
multi-class LDA (k-NN) [18]	62.7
multi-class LDA (SVM) [18]	64.4
MSCNN [20]	85.1
DAM-CNN [9]	99.3
IFSL (SVM) [10]	88.2
IFSL (k-NN) [10]	76.4
Our method	**98.6**

state-of-the-art algorithms.

For CK+ dataset, as shown in **Table 1**, the accuracy of our method is superior to most of the other listed advanced methods. The traditional classification methods perform poorly. But the accuracy of our method is as high as 99.8% on the verification set and 98.9% on the test set, which confirms that the feature matrix enhances the expression of features in salient regions. We remove more redundant features and use data enhancement methods, which may be the reason for our high accuracy.

For JAFFE dataset, as shown in **Table 2**, the traditional classification methods are not competent, but CNNs models including our model maintains superior performance. The performance of only few existing methods is evaluated on JAFFE, since it is a small dataset. DAM-CNN uses VGG16 network and attention model to achieve the highest recognition accuracy of JAFFE dataset. The important reason why the accuracy is higher than ours is that the model is more complex and has more parameters.

For FER2013 dataset, as shown in **Table 3**, we still have a satisfactory performance in the most challenging dataset. Since there are many side face images in

Table 3. Comparison results obtained on the FER2013 dataset.

Methods	Accuracy(%)
Shen et al. [21]	61.9
CNN [22]	66.4
Light-CNN [1]	68.0
dual-branch CNN [1]	54.6
pre-trained CNN [1]	71.1
Our method	**66.7**

FER2013 dataset, even many of them are not faces, it cannot be preprocessed, which the effect of feature matrix is greatly weakened. Its performance is only above average. The accuracy is lower than some of CNN models, because our model is relatively simple.

3.4. Discussion

The experimental results show that our CNN model achieve state-of-the-art performance with simple structure and has good performance on the three commonly used datasets. Through the experimental comparison, we also find some problems. High accuracy is obtained in the classification task of non-spontaneous expression. But the performance is average for spontaneous expression, and the recognition effect is even a little terrible for the side face image. The main reason is that there will be errors in the acquisition of salient feature regions, which weakens the role of emotion feature matrix. CNN model is simple and can not learn more useful features, which may also lead to this problem. In addition, whether the existence of the emotion feature matrix will affect the robustness and reduce the generalization ability of the model needs further verification. If it happens, we need to further study how to obtain the best feature matrix to avoid this situation.

4. Conclusion

In this paper, we propose a new expression recognition model for low-resolution images through extracting emotion feature matrix and adding it to the CNN to enhance the expression of salient region features. The model framework is mainly divided into four parts, including data preprocessing, feature matrix acquisition, convolution neural network construction and classification. The purpose of this method is to increase the proportion of salient features and reduce the proportion of redundant features, so as to improve the classification accuracy. After that, we carry out experiments on three commonly used datasets. Compared with several state-of-the-art methods, the proposed method achieves superior results.

Acknowledgements

This work was supported by the Beijing Municipal Natural Science Foundation under Grant No. 3192028

References:

[1] Shao J, Qian Y. Three convolutional neural network models for facial expression recognition in the wild[J]. Neurocomputing, 2019: 82-92.

[2] Alizadeh S, Fazel A. Convolutional Neural Networks for Facial Expression Recognition.[J]. arXiv: Computer Vision and Pattern Recognition, 2017.

[3] C.A. Corneanu, M. Oliu, J.F. Cohn, S. Escalera,Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends,and affect-related applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (12)(2016) 1548-1568.

[4] R.A. Khan, A. Meyer, H. Konik, S. Bouakaz,Framework for reliable, real-time facial expression recognition for low resolution images, Pattern Recognit. Lett.34 (10) (2013) 1159-1168.

[5] Chu Y, Ahmad T, Bebis G, et al. Low-resolution face recognition with single sample per person[J]. Signal Processing, 2017: 144-157.

[6] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2014) 295-307.

[7] Haghighat M, Abdelmottaleb M. Lower Resolution Face Recognition in Surveillance Systems Using Discriminant Correlation Analysis[C]. ieee international conference on automatic face gesture recognition, 2017: 912-917.

[8] S.L. Happy, A. Routray, Automatic facial expression recognition using features of salient facial patches, IEEE Trans. Affect. Comput. 6 (1) (2015) 1-12.

[9] Xie S, Hu H, Wu Y, et al. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition[J]. Pattern Recognition, 2019: 177-191.

[10] Yan Y, Zhang Z, Chen S, et al. Low-resolution Facial Expression Recognition: A Filter Learning Perspective[J]. Signal Processing, 2020.

[11] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, IEEE Trans. Image Process. 28 (5) (2019) 2439-2450.

[12] Lim B, Son S, Kim H, et al. Enhanced Deep Residual Networks for Single Image Super-Resolution[C]. computer vision and pattern recognition, 2017: 1132-1140.

[13] Zou W W, Yuen P C. Very Low Resolution Face Recognition Problem[J]. IEEE Transactions on Image Processing, 2012, 21(1): 327-340.

[14] X. Xing, K. Wang, Couple manifold discriminant analysis with bipartite graph embedding for low-resolution face recognition, Signal Process. 125 (2016)329-335.

[15] Whitehill J, Omlin C W. Haar features for FACS AU recognition[C]. international conference on automatic face and gesture recognition, 2006: 97-101.

[16] Chen L, Yang Y, Wang J, et al. Attention to Scale: Scale-Aware Semantic Image Segmentation[C]. computer vision and pattern recognition, 2016: 3640-3649.

[17] Wang N, Gao X, Tao D, et al. Facial feature point detection: A comprehensive survey[J]. Neurocomputing, 2018: 50-65.

[18] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19(7) (1997) 711-720.

[19] Shan C, Gong S, Mcowan P W, et al. Robust facial expression recognition using local binary patterns[C]. international conference on image processing, 2005: 370-373.

[20] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Trans. Image Process. 26 (9) (2017)4193-4203.

[21] Zeng G, Zhou J, Jia X, et al. Hand-Crafted Feature Guided Deep Learning for Facial Expression Recognition[C]. ieee international conference on automatic face gesture recognition, 2018: 423-430.

[22] Mollahosseini A, Chan D M, Mahoor M H, et al. Going deeper in facial expression recognition using deep neural networks[C]. workshop on applications of computer vision, 2016: 1-10.