Current Science
CS 5 (5), 4508-4522 (2025)
Received 20 June 2025 | **Revised 18 July** 2025 | **Accepted 16 August** 2025 | **Online Available 21 September** 2025
https://doi.org0/10.52845/CS/2025-5-5-16

Current Science

Current Science

OPEN ACCESS

ISSN(O)2795-8639

Original Article

Super-Resolution Image Facial Expression Recognition using Deep Learning: A Complete Improved Approach

Chenxing Xue[1], Xiangye Ji[2]

[1]China Ship Research and Development Academy, Beijing, 100097, China

[2]China Ship Research and Development Academy, Beijing, 100097, China

[*]Corresponding Author: Chenxing Xue

Abstract:
In the last few years, there have been significant advances studying how computers can better understand human emotions. The technique of facial expression recognition (FER) is the focal point of this effort. In this paper, we propose a face detection and facial expression recognition method based on an improved Faster Regions with convolutional neural networks (Faster R-CNN). In the complete improved method, the model is improved from the two points: the structure of network and data optimizing. We introduce a new multi-scale fusion strategy to the Faster R-CNN, in terms of the structure, which can get both shallow and deep information to increase the gap between similar facial expressions. Data optimizing has two enhancement modules which are beneficial for FER. By reconstructing the super-resolution image, the one is to improve the quality of the image. The another is combination of channels, which can help focus on facial texture information. Experimental results illustrate the advancement of the FER method in the Japanese female facial expression (JAFFE), the Extended Cohn-Kanade (CK+) database and SFEW. It is a complete real-time facial expression recognition method.

Index Terms: Facial Expression Recognition, Faster R-CNN, Super-resolution image, Multi-Scale Fusion

1. Introduction

Facial expression is one of the most universal, direct and useful tools for human beings to express their emotional states and intentions [1,2]. The Facial Expression Recognition (FER) plays a critical role in human-computer interaction system. In the field of computer vision and machine learning, FER has been an active research topic for many years [3]. To build a highly accurate FER system, a wide variety of methods have been proposed.

Generally, FER methods can be classified into two major categories: traditional methods and methods based on deep learning. Traditional methods usually sequentially perform two individual stages process: feature extraction and simple classifier construction. In [4], Mase et al. realized automatic FER on computer for the first time. The region-based optical flow model was used to extract the movement features of facial muscles, then the k-nearest neighbor rule was used to classify the expressions. Subsequently, based on this mind, a series of methods [5-15] were proposed. Zhang et al. [6] adopted the Gabor filter to extract facial expression features and classified them with a double-layer perceptron, which achieved good results. In [13], Asthana et al. proposed a feature extraction method based on geometric features: Active Appearance Mode (AAM), and used a multi-class SVM to classify facial expressions. Chen et al. [7] used the Clustering based Discriminant Analysis (CDA) method to extract facial expressions with discriminant power, and selected a one-to-many classification scheme for facial expression classification. The CDA method provided an

efficient feature reduction and extraction schemes which are useful for FER. In [10], Zhao et al. proposed a simpler Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) histograms computed from three orthogonal planes, made it easy to extract facial expression related features. A nearest-neighbor classifier is adopted to classify facial expressions. Their approach is computationally simple and robust in terms of grayscale and rotation variations, making it very promising for real application problems [10]. Chu et al [15]. used the methods of Gradient Directional Pattern (GDP) and Local Binary Pattern (LBP) to extract features, and the Sparse Representation Classification (SRC) is used to classify facial expressions. This method improved the recognition rate.

On the basis of predecessors, variety of improved methods on FER have been published in recent years. To overcome high dimension and characteristic redundancy of Completed Local Binary Pattern (CLBP) features, Zhou et al. [18] proposed a method for facial expression recognition based on discriminative CLBP. The method could select different facial expressions corresponding to different facial features and classify them. In [19], Li et al. used LBP and Histogram of Optical Flow (HOOF) to extract features, and a Linear Support Vector Machine (LSVM) is used as classifier for Micro FER. This method has many potential applications such as in lie detection, law enforcement and psychotherapy [19]. In [20], Liu et al. mainly focus on the stage of extracting features. They computed optical flow to extract the Main Directional Mean Optical-flow (MDMO) features, and introduce sparsity into the original MDMO features. Meanwhile, they used Support Vector Machine (SVM) to classify facial expressions. MDMO is an effective sparse representation method. To achieve better recognition effect, extensive empirical studies are needed to search for an optimal combination of feature extraction and classifier [20]. Therefore, it would be highly expected to find an intelligent approach, which can automatic FER in a simple and efficient way.

With the vigorous development of artificial intelligence, deep learning methods have seen an explosion of interest and are being successfully applied across an extraordinary range of problem domains in the last few years. It had been also introduced into the field of FER [20-24]. In [20], Liu et al constructed a FER method by dividing face images into 80 image blocks, then establishing a deep confidence network for each block, and constituted a boosted 80 confidence networks. Although this network structure achieves better detection results, the detection speed is slow due to the complexity of the network. Li et al [23] improved the LeNet5 network through multi-scale fusion strategy, and achieved certain results in facial expression detection. However, because of the limitation of LeNet5 network model, it has only three convolution layers, which makes it difficult to extract deep network features better. In [21], Hamester et al. proposed a multi-Channel Convolutional Neutral Network (MCCNN). Two hard-coded feature extractors are replaced by a single channel which is partially trained in an unsupervised fashion as a Convolutional AutoEncoder (CAE). One additional channel that contains a standard CNN is left unchanged. Information from both channels converges in a fully connected layer and is then used for classification [21]. However, two hard-coded feature extractors are complex, and this structure has not proved entirely useful. Because of the network can automatically extract features without artificial interference, it is a promising method to solve FER with deep learning.

In our opinions, the deep learning based FER can be improved from two points. In the first one, efforts are concentrated on distinguishing slightly differences of similar expressions. We draw on the experience of Chen et al [31] who introduce multiscale fusion strategy to obtain more comprehensive feature information in the foggy environment to get both shallow and deep facial expression features. In the second point of view, efforts are put into enhancing the expression correlation of facial images. For example, Zhou et al [18] thought facial expression features are mainly concentrated near the mouth, nose and eyes. Therefore, they put forward a method of fusion of expression sub region and whole image features to improve the expression correlation of facial images.

In this paper, we propose a novel and complete FER method based on deep learning. Through the structure of channel combination for images,

which is effective to improve the expression correlation of facial images. The significant edge image, the global edge information image and the original data are combined through channels to form the channel combined image. Specifically, we introduce a new multi-scale fusion strategy, which accounts for recognition performance of both the Shallow features and deep features, to distinguish similar expressions. We select Faster Regions with Convolutional Neural Networks (Faster R-CNN) [25] as the basic structure, which is proposed by Ross Girshick. Due to the higher efficiency of the model in target detection, it was quickly recognized by the industry and consequently widely used in various research fields [26,27,28]. Such as, Song et al [26] converted the Faster R-CNN multi-classification into a two-class problem and applied it to vehicle detection in complex scenes. At present, Faster R-CNN is one of the mainstream algorithms for target detection.

The rest of this paper is organized as follows. In section 2, we introduce our work from three aspects: the structure of network, reconstructing the super-resolution image and combination of channels. Section 3 is mainly about the experiment content. We summarize our achievements, explaining the limitations of the article and looking forward to the next work in Section 4.

I. Our Proposed Method Based on Faster R-Cnn

A. Brief Review Faster R-Cnn

The Faster R-CNN model is shown as Fig. 1. The first step of this method is to send the data to the network, and then scales the picture according to equation 2 and 3. The second step feeds the scaled data into the VGG16 feature extraction module. The module references the transfer learning strategy [29] with VGG16 as a pre-trained model. The third step sends the feature maps to the Region Proposal Networks (RPN) for classification of foreground and background and regression of bounding boxes. In the fourth step, the data processed by the RPN and the convolved feature map of conv5_3 were sent to the final foreground classification layer and the bounding box regression layer. It's loss function for an image is defined as equation 1:

$$L(\{P_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_{i=1}^{n} L_{cls}(P_i, P_i^*) + \lambda \frac{1}{N_{reg}}\sum_{i=1}^{n} P_i^* L_{reg}(t_i, t_i^*) \qquad (1)$$

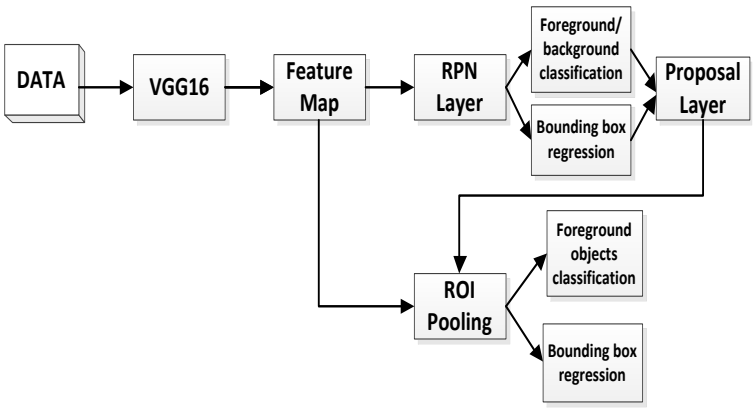

Figure 1 The model of Faster R-CNN.

Where, i is index of the box, P_i is the predicted of box i being an object, t_i is the parameterized coordinates of box, P_i^* is used to determine whether it is the ground-truth label, t_i^* is the box when P_i^* is 1. Usually, λ is set to 10 to keep the N_{cls} and N_{reg} terms roughly equal, N_{cls} represents the mini-batch size, N_{reg} is the number of anchor.

For equation 3, the input pictures have a long side of M and a short side of N. The model limits the minimum side (L_{min}) to 600 and the maximum side (L_{max}) to 1000, where the two scaling ratios

are δ_1 and δ_2 If equation 2 is satisfied, δ_2 is selected as the scaling ratio, otherwise it is δ_1.

$$\delta_1 * M > L_{\max} \tag{2}$$

$$\delta_1 = \frac{L_{\min}}{N} \qquad \delta_2 = \frac{L_{\max}}{N} \tag{3}$$

Due to the efficiency of Faster RCNN in the field of target detection, the model is employed as the basic structure. This paper compares the application status of that model, replaces the foreground and background classification with face detection, and replaces the classification of foreground objects with facial expression classification. More importantly, we introduced a multi-scale fusion strategy to improve the main feature extraction module (VGG16) of Faster RCNN.

B. Our Proposed Impproved Multiscale Convolutional Neural Network

As the depth increases, the extracted image features become more abstract in the network, and makes the shallow features gradually negligible [30]. The original VGG16 feature extraction module is implemented by stacking the 3*3 convolution kernels and the 2*2 pooling layers. To make the shallow features of the model available, the convolutional layer in the feature extraction module is more closely connected, that is, a new multi-scale feature is employed.

To further improve the information flow between layers, we propose a different connectivity pattern: the structure of Net-Block. Fig. 2 illustrates the layout of the structure schematically. We draw on the method of multi-scale fusion of channels in the literature [31,32,42], but it does not achieve better detection results. Surprisingly, the method of multi-scale fusion of height proposed in our paper has greatly improved the detection accuracy. The method merges the features of different layers into the same channel of the feature map to promote similar expressions for classification. The shallow features and the deep features were merged into a channel, which is very effective for the detection of multiple expressions under a single target. It is a new advanced feature.

The flowchart of the detecting process is shown in Fig. 3 and Fig. 4. The difference from the original network is described as follows. Firstly, the model takes the features extracted by the Conv3_3 and the Conv4_3 layer and superimposes them to generate a Net1 layer. Then the features extracted from Net1 layer and Conv5_3 layer are combined by concat to generate Net2, which are then sent into RPN network and ROI Pooling respectively for subsequent classification and regression tasks. The initial parameters of VGG16 are trained based on the ImageNet dataset. In addition, the parameters of the first six layers are set to freeze, and the last seven layers of parameters are updated with the model training.

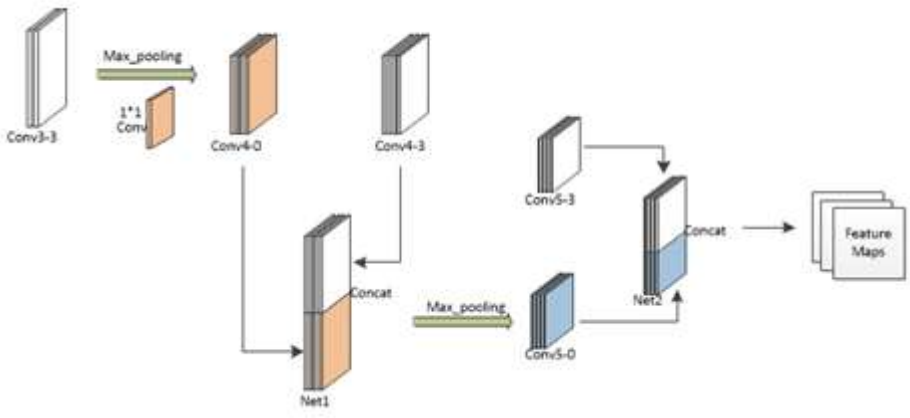

Figure 2: The structure of Net-Block

C. Our Proposed Data Preprocessing

To solve the problem that the old data set with grayscale image cannot be widely used, we propose novel approach of multi-channel combination of super-resolution images. The method enhances the expression correlation of facial images while improving image quality.

In the deep learning, the quality of the data determines the effectiveness of the model. Therefore, we mainly strengthen data through two aspects: reconstructing super-resolution images and Image fusion. In addition, our method is different from the algorithms [5-15] in the data preprocessing stage. The classification methods of those references are only based on the classification after face detection, not a complete facial emotion recognition. This paper achieves face detection and facial expression classification with a single small-scale background, and does not need to cut the original data.

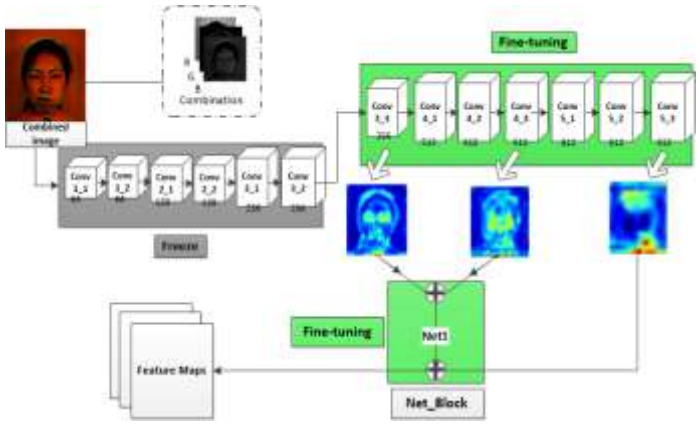

Figure 3 Improved feature extraction module

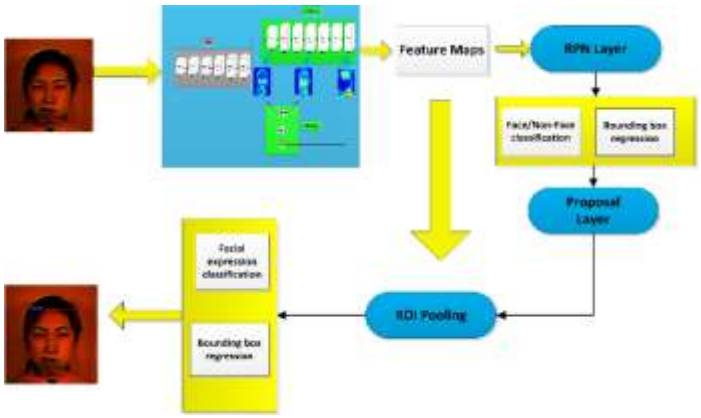

Figure 4 Facial expression recognition system

Reconstructing Super-Resolution Images

By studying the structure of Faster R-CNN, we found that the data should be resized before being fed into the network. The long side cannot exceed 1000 while the short side cannot exceed 600. However, the JAFFE database data has a uniform size of 256*256. The size is set to 600 * 600 by equation 2 and 3. Image quality is seriously affected by the scaling method. Thus, we refer to the Generative Adversarial Net to enhance data. The result of the comparison is shown as Fig 5.

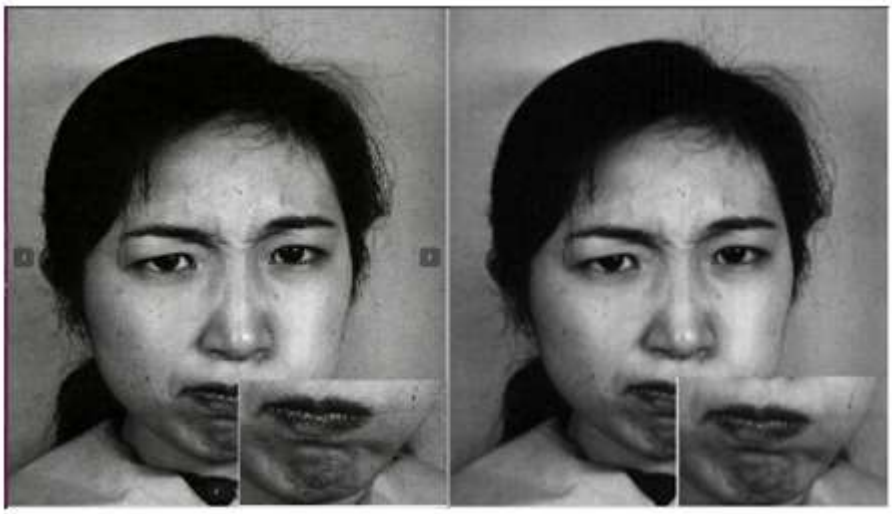

Figure 5 The result of the comparison, the left side is the super resolution image, and the right side is the scaled original image.

To improve the image quality, this paper employs the Super-resolution Generative Adversarial network (SRGAN) [34] to reconstruct super-resolution images. The SRGAN network generates a high-resolution image from a low-resolution image, and its discriminant network determines whether the generated image is a forged image or an original image. If the generated image is created in the confrontation network and the discriminant network cannot detect it, a super-resolution image can be obtained, and the resolution of the image is improved by using the antagonistic loss (4) and the content loss (5), (6). The equations are:

$$L_{Gen}^{SR} = \sum_{n=1}^{n} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{4}$$

$$L_{MSE}^{SR} = \frac{1}{r^2 WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{5}$$

$$L_{VGG|i,j}^{SR} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR})_{x,y}))^2 \tag{6}$$

The variable G_{θ_G} represents the probability that the generated image is judged to be a true image, and $G_{\theta_G}(I^{LR})$ represents the reconstructed super-resolution image. The content loss function includes two parts: the first is the pixel space error MSEloss and the second is VGGloss.

The Combined Image of Three Channels

Every expression is communicated through the movement of the facial muscles. It is not difficult to find that when the muscles of the face move, the corresponding parts will stretch or contract in FACS [35]. For instance, in AU10, the upper lip is pulled up. These actions will deepen the facial edges of the corresponding movements so that others can better understand the emotion. To help the neural network understand these facial expressions, we propose a multi-channel combined method, which makes the data more suitable for Faster R-CNN, while deepening the edge information. The original data is firstly processed by the Sobel operator and Adam operator to separately generate significant edge image and global edge information image.

The type of input data for the model requires three-channel data, while the JAFFE public library image is a set of single-channel grayscale image. In order to make the grayscale image be widely used, we generate a new kind of composite images by MATLAB software, as shown in Fig. 6. The image is composed of the original image, the edge image of Sobel operator and the edge image of Adam operator by channel combination method

We propose the Adam operator which can get the global edge information. By using the method of full convolution network for reference, we construct a simple convolution network [40]. The operator performs a sliding window operation from the original image to generate a corresponding map, as shown in Fig. 7. It provides the overall information and also enhances the edge information. It also contains the complete and continuous edge information.

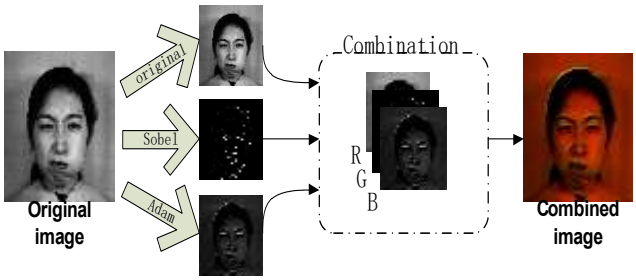

Figure 6 Processing flow chart of the three-channel combined image.

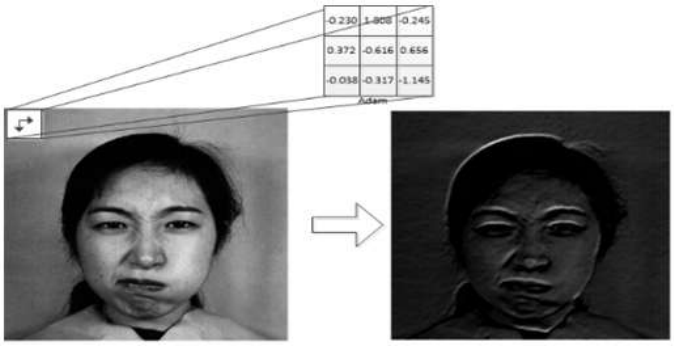

Figure 7 dam operator processing flow.

II. Experiment and Discussion

In this experiment, Faster R-CNN, based on TensorFlow framework using the Linux platform, is used to analyze the proposed method. The data preprocessing platform uses MATLAB software. The experimental platform is configured with 16 GB memory, and a NVIDIA 1660-6G graphic card.

The Faster R-CNN initial learning rate was 0.001, the learning rate of 50,000 steps per iteration is reduced by an order of magnitude. RPN positive sample threshold is 0.7. Threshold for non-maximum suppression is 0.7. Number of top scoring boxes to keep before apply NMS to RPN proposals is 12000. Number of top scoring boxes to keep after applying NMS to RPN proposals is 2000. We did not change the mechanism of anchors in the original model.

The optimizer is an SGD optimizer with momentum, and the equation is shown in Equation 7,8, where V_t represents the acceleration at t, α represents the power size, and is generally set to 0.9, X represents the sample, Y represents the label corresponding to X, and W_t represents the Model parameters at t.

$$V_t = \alpha V_{t-1} + \eta_t \nabla J\left(W_t, X^{(is)}, Y^{(is)}\right) \quad (7)$$

$$W_{t+1} = W_t - V_t \quad (8)$$

In this paper, we feed experimental data to the improved network for training iterations 100,000 times. To demonstrate the effectiveness of the proposed method, we have performed extensive experiments on JAFFE database and extended Cohn-Kanade database. Firstly, we have prepared four sets of experimental data and a more

comprehensive evaluation function. Secondly, according to the characteristics of data, different data partition methods are adopted through comparing experiments. In JAFFE, we take human-based strategies as a way to test and train experimental data. In Cohn-Kanade, we take the 5-fold cross-validation technique. Finally, we verified our improved model, and compared it with its experimental results. At the same time, we tested our model on our own data.

A. Experimental Evaluation Function

we compared two network performance evaluation parameters. Expression (9) shows the formula for accuracy (Accuracy, Acc). Expression (10-11) shows Average-Precision (AP). Acc is often employed in facial expression recognition papers [4-24,37-39,41,42].

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

$$AP = \int_0^1 P(r)d(r) \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \tag{11}$$

In these equations, TP indicates that a positive sample is correctly identified as a positive sample, TN indicates that a negative sample is correctly identified as a negative sample, FP indicates that a negative sample is misidentified as a positive sample, and FN indicates that a positive sample is incorrectly identified as a negative sample. Besides, p is the precision, and r is the recall. Equation (11) shows the definitions of those two variables. These expressions show that the expression Acc (Accuracy) can evaluate the global accuracy, but it cannot fully evaluate the performance of model. AP (average precision) can analyze model capability from multiple angles, so that it can more comprehensively assess the results. In conclusion, we hire AP to evaluate the model.

B. Datasets

The Japanese female facial expression (JAFFE) [33] database was proposed in the 1990s. It is a mage dataset captured in laboratory that contains 213 samples of posed expressions from 10 Japanese females. Each person has 3~4 images with every basic facial expression and one image with a neutral expression. All images have an equal resolution of 256*256 pixels and are in grayscale. The database is challenging because it contains few examples per subject/expression [1].

Fig. 8 shows an image of a normal expression and the six basic expressions.

Figure 8 Public facial expression data of JAFFE

The Extended Cohn-Kanade (CK+) database has 593 sequences from 123 subjects. Image sequences were processed into either 640x490 or 640x480 pixel arrays with 8- bit gray-scale or 24-bit color values. It has seven popular facial expressions, i.e. angry, disgust, contempt, fear, sadness, surprise and happy. The contempt expression is not adopted. Some samples of remaining six expressions and a normal expression on the CK+ database are described in Fig. 9.

Figure 9 Public facial expression data of CK+

SFEW[43] is a static frame edited from movies , which contains seven facial expressions. it is a very challenging set of data. The data contains a large number of facial expression images that are difficult to distinguish, such as facial occlusion, blurred lights, complex backgrounds, etc. SFEW is divided into three groups: Train (958 samples), Val (436 samples) and Test (372 samples). Among them, the training set and test set are publicly available, and the label of the test set is reserved by the government. Therefore, this article will use the validation set as the test set. The data sample is as follows.

Figure 10 Public facial expression data of SFEW

We hired JAFFE as an example to explain the experimental methods of data enhancement and testing. More importantly, we analyzed and studied the distribution of data and selected the most rigorous method to evaluate the model.

C. Experimental Data of the Jaffe

The basic data comes from JAFFE public database. In further verify our proposed methods, four sets of images were prepared by the combined method of 2.3.2. These are the original grayscale image M-1, the reconstructed super-resolution grayscale image M-2, the original combined image M-3 and reconstructed super combined image M-4, as shown in Table I. Fig. 10 shows the combined images. The size of M-2 and M-4 is 1024*1024, and the size of M-1 and M-3 is 256*256. To better display and compare the images in this figure, the images are uniformly scaled to 600*600 pixels. M-1 and M-3 are both channel combination images. The three channels of images are all gray images of the same type. In addition, the amount of data was doubled by using a flipping strategy.

Table I Composition of image data. Original is original grayscale image of jaffe database.

	R	G	B
M-1	original	original	original
M-2	super-resolution	super-resolution	super-resolution
M-3	original	sobel	Adam
M-4	super-resolution	sobel	Adam

Super-resolution is Image super-resolution based on JAFFE database. Sobel and Adam are both edge detection operators. They are all based on the R channel image for edge detection.

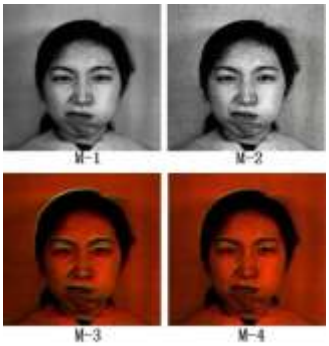

Figure 11 Experimental data

D. The Instructions of Experimental Data

The algorithm evaluation based on the JAFFE is divided into two methods. The first randomly divides into several groups according to the expressions, and the model is trained and verified, as described in the literature [23,36,37]. The other method is the human-based strategy mentioned in the literature [11,15,37]. It uses all the expressions of one or two people as a verification set, and the expression of the remaining person as a training set in the library. In this database, we have found that the same facial expression similarity of a single person is very high. For instance, in Fig. 11, the expression of the same happy person is only slightly changed in the three images. We believe that if the first type of evaluation algorithm is adopted, the similarity of the same expression of the single person will greatly affect the rigor of the experiment, and it is very easy to over-fit.

Figure 12 Comparison of the same facial expression from the same person.

In order to evaluate our method better, a group of comparative experiments are carried out. The model uses the original model of Faster R-CNN, the verification data set (Val) is the whole expression of two people, and the remaining expression data is trained and tested according to the first type of evaluation method. The test set is one-third of the remaining data, and the training set is the rest. The results are shown in Table II. The table shows that the first evaluation method has high test accuracy, but the generalization is very low, and the model has an obvious over-fitting phenomenon. Therefore, we adopt a human-based strategy in the paper.

In addition, the CK+ dataset is utilized to evaluate the model by the above methods. we report the results using the 5-fold cross-validation technique and then averaging the recognition rates over five folds.

E. Result Analysis

In Table III, a comparative analysis is performed using M-1, M-2, M-3, and M-4 as experimental data with the Faster R-CNN model. The test data is all of the expressions of the first two people, and the training data is the remaining eight people. The experimental results show that reconstructing the super-resolution image can effectively improve the classification accuracy. However, the proposed combined channel image M-3 does not improve the accuracy, and the classification accuracy is lower than that of the three-channel image M-1. By comparing M-2 with M-4, one can infer that a multi-channel combined image containing reconstructed super-resolution images facilitates facial expression classification. Given the above results, we consider that expanding the size of image makes the image obscured, which will be negatively affected in the edge detection stage. It is the reason that M-3 is the worst in four groups of experimental data. Therefore, super-resolution strategy is used in M-4, and its final effect is better than other types of data.

Table IV shows the comparison experiment between Faster R-CNN (FRCNN) and improved multi-scale Faster R-CNN (M-FRCNN). The evaluation method of Table II is used, and the data is M-4. From this table, it can be found that M-FRCNN combines high-level abstract features and low-level detailed information through cross-linking, so that the network can better understand that expression features, such as the sadness are significantly improved.

Table V shows the different expression classification results of the algorithm for the JAFFE database. To strictly verify the algorithm, this paper divides the database into five groups of data where each group is the whole expression of two people, and five corresponding experiments have been done. The overall trend of the results is that the expressions “happy” and “angry” achieve the highest recognition rate, and the recognition rate of “sad” and “disgusting” are the lowest. In addition, the expressions of “surprise” and “neutral” are improved. Fig. 12 shows the test results on the data set. We also tested our approach in terms of gender and background in

Fig. 13. In the case of similar background of experimental data, the male's expression recognition rate has also reached a high level. In the complex background, we choose the data with a pair of glasses to test our method in Fig. 14. Although our method has low accuracy, it works.

Figure 13 experimental test results on the JAFFE

Figure 14 experimental test results on our data

Table II. Comparative experiment of evaluation methods AP(%)

	Happiness	Anger	Fear	Neutral	Sadness	Surprise	Disgust	MAP
Test	100	100	97.90	100	100	100	100	99.88
Val	97.73	65.98	92.42	74.30	79.98	97.40	80.47	84.04

TABLE III. FACIAL EXPRESSION CLASSIFICATION OF DIFFERENT SAMPLES AP (%)

	Happiness	Anger	Fear	Neutral	Sadness	Surprise	Disgust	MAP
M-1	100	88.74	87.22	100	66.67	100	100	91.80
M-2	100	87.88	90.31	100	81.82	100	100	94.29
M-3	97.73	85.15	100	84.92	84.85	89.09	90.91	90.38
M-4	100	95.45	97.73	95.45	79.22	100	100	95.41

TABLE IV FACIAL EXPRESSION CLASSIFICATION OF DIFFERENT MODELS AP(%). M-4 IS SELECTED AS EXPERIMENTAL DATA.

	Happiness	Anger	Fear	Neutral	Sadness	Surprise	Disgust	MAP
FRCNN	100	95.45	97.73	95.45	79.22	100	100	95.41
M-FRCNN	100	92.42	95.96	100	88.24	100	100	96.66

TABLE V. CLASSIFICATION OF DIFFERENT EXPRESSIONS OF JAFFE DATABASE IN THIS PAPER AP(%)

	Happiness	Anger	Fear	Neutral	Sadness	Surprise	Disgust	MAP
Test 1	100	92.42	95.96	100	88.24	100	100	96.66
Test 2	100	100	100	100	92.42	100	78.44	95.84
Test 3	100	100	100	100	88.24	100	72.73	94.42
Test 4	91.74	100	67.38	95.45	73.86	100	82.95	87.34
Test 5	100	97.40	93.18	86.15	100	90.91	74.55	91.74
MAP	98.34	97.96	91.30	96.32	88.55	98.18	81.73	93.26

TABLE VI CLASSIFICATION OF DIFFERENT EXPRESSIONS OF CK+ DATABASE IN THIS PAPER AP(%)

	Happiness	Anger	Fear	Neutral	Sadness	Surprise	Disgust	MAP
Test 1	100	93.39	100	78.94	81.82	100	100	96.66
Test 2	100	100	100	100	93.39	100	100	99.06

Test 3	100	100	100	100	94.66	100	97.73	98.91
Test 4	97.73	100	97.73	100	100	100	97.73	99.03
Test 5	100	74.60	100	100	97.98	100	100	96.08
MAP	99.55	93.60	99.55	95.78	93.57	100	99.09	97.94

TABLE VIII COMPARISON OF DEEP LEARNING ALGORITHMS ON THE JAFFE DATASET AND THE CK+ DATASET.

Source	Method	JAFFE(%)	CK+(%)
Liu et al. [20]	DBN	91.80(ACC)	-
Yang et al. [22]	Weighted mixture double channel CNN	92.20(ACC)	97.00(ACC)
Salmam et al. [41]	Fiducial Point detection and NN	93.8(ACC)	99.00(ACC)
Hamester et al. [21]	MCCNN	95.8(ACC)	-
LI et al [23]	Cross-connected LeNet-5	94.37(ACC)	83.74(ACC)
Ours	M-Faster R-CNN	93.26(MAP)	97.94(MAP)

Table VI shows the different expression classification results of the algorithm for the CK+ database. We used 5-fold cross-validation strategy as a method to test our model. The experimental results show that our method performs well on the CK + dataset.

Table VII compares the recognition accuracy of different traditional methods in the JAFFE. It is clear from the results that the deep learning algorithm, M-Faster R-CNN, is a significant improvement in classification accuracy over the previous methods.

TABLE VII COMPARISON OF TRADITIONAL ALGORITHMS ON THE JAFFE DATASET

Source	Method	Precision (%)
Ying et al. [11]	LBP+LE	70.48(ACC)
Huang et al. [38]	ASM	86.96(ACC)
Gu et al. [39]	Gabor	89.67(ACC)
Cheng et al. [37]	Gaussian process	55.24(ACC)
Cheng et al. [37]	Naive Bayes	40.10(ACC)
Chu et al. [15]	LBP+GDP+SRC	69.52(ACC)
Ours	M-Faster R-CNN	93.26(MAP)

As can be seen from the table VIII, although our proposed method is not state-of the-art, it is rigorous and effective. Firstly, taking JAFFE as an example, we discussed the distribution of the data set. Through the experiments in Table II, a rigorous data usage method for JAFFE is proposed. In the table VIII, the method proposed by Salmam et al [41]. is more accurate than our method on the JAFFE. However, their databases were divided into 60% for training, 10% for validation, and 30% for test. Table II has shown that continuous data of the same expression may over fit the model. The method proposed by LI et al. [23] has the same problem, and on the ck+ data set, it is far lower than our method. Secondly, Hamester et al. [21] adopted the rigorous strategy in Table 2, and its experimental accuracy was higher than our method. But MCCNN is composed of a two-channel network, the model is complex and some principles cannot be explained. Our method is simple and fast.

TABLE IX. COMPARISON OF ALGORITHMS ON THE SFEW DATASET.

Source	Method	Precision (%)
Meng et al. [44]	Identity-aware CNN	50.98(ACC)
Wadhawan et al. [45]	EL	44.50(ACC)
Gu et al. [39]	ppf-svm	38.38(ACC)
Ours	M-Faster R-CNN	43.10(MAP)

In the table IX, the performance of the network model is tested with the verification set of SFEW. Meng et al. [44] and Wadhawan et al. [45] hired CNN model to recognize facial expressions. But the experimental data they use are all cropped images after face recognition. Our method needs to search for faces i in a complex background for expression classification. Gu et al. [39] proposed to classify expressions based on traditional methods. It can be seen from the results that our method is far superior to theirs.

On the three databases, the method in this article can detect a picture in 0.15s. More importantly, our method is different from the above methods. It can locate and classify facial expressions directly from the data sets without clipping. It is a complete facial expression recognition system.

III. Conclusion and Perspective

The paper proposes a complete facial expression recognition method based on improved Faster R-CNN. Using the JAFFE public database, the CK+ database and the SFEW database, two evaluation methods were compared and tested, and the performance of the model was tested with the most rigorous evaluation methods. Firstly, this paper proposes that when the data size is too small, the super-resolution strategy can be used to improve accuracy. Secondly, the three-channel edge information combination method and multi-scale fusion strategy can effectively improve the accuracy of FER. What's more, according to the analysis in Section 3.1, we chose a more comprehensive method than most of the facial expression evaluation methods. Finally, based on the Faster RCNN network, this article can efficiently detect target expressions in a single background, and the detection speed can reach 5fps.

Therefore, our complete FER method based on improved Faster RCNN is an advanced and effective technology.

Although our method has achieved good overall results, it was found during the experiment that there are many occluded facial expressions in the SFEW that have not been accurately identified. It is necessary to add other strategies in subsequent experiments to solve occlusion and other problems. Future areas of further development include constructing the facial expression database in complex scenes and developing a real-time facial expression detection system. This work is currently underway. With the development of AI, we would expect that an intelligent FER system can be used to predict the facial expression with all kinds of complex environments, which is quite a step forward in pattern recognition.

References

1. S. Li, W. Deng, “Deep facial expression recognition: a survey,” 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018.
2. X. Wu, C. Feng, M. Xu, T. F. Zheng and A. Hamdulla, "DialoguePCN: Perception and Cognition Network for Emotion Recognition in Conversations," in IEEE Access IEEE, vol. 11, 2023: pp. 141251-141260.
3. Ö. Ezerceli and M. T. Eskil, "Convolutional Neural Network (CNN) Algorithm Based Facial Emotion Recognition (FER) System for FER-2013 Dataset," 2022 International Conference on Electrical, Computer, Commun -ications and Mechatronics Engineering (ICECCME), Maldives, Maldives,2022,pp.1-6.
4. Kenji Mase and Alex Pentland, “Recognition of Facial Expression from optimal flow” IEICE TRANSACTIONS on Information and Systems, 1991, pp.1-43.
5. Sharifnejad M, Shahbahrami A , Akoushideh A , et al. Facial expression recognition using a combination of enhanced local binary pattern and pyramid histogram of oriented gradients features extraction[J]. Image Processing, IET, 2020, 15(4).
6. M. Pantic, “Automatic Analysis of Facial Expressions: The State of the Art [J] ” , IEEE Trans. PAMI, 2000, pp.1424-1445.
7. Bellamkonda S, Gopalan NP. An Enhanced Facial Expression Recognition Model Using Local Feature Fusion of Gabor Wavelets and Local Directionality Patterns[J]. International Journal of Ambient Computing and Intelligence (IJACI), 2020, 11.
8. S. Banoth, V. Kukreja, N. Thapliyal, M. Aeri and R. Sharma, "Beyond Basic Emotions: High-Accuracy Facial Expression Classification Using a CNN-SVM Hybrid Model," 2024 International Conference on Emerging Technologies in Computer Science

for Interdisciplinary Applications (ICETCS), Bengaluru, India, 2024, pp. 1-4
9. Gharbi, A. Gattal and I. Bendib, "Basic Emotion Recognition in Facial Expressions using Deep CNN," 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), EL OUED, Algeria, 2024, pp. 1-6
10. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE transactions on pattern analysis and machine intelligence, 2007, pp.915–928.
11. Z. Ying, L. Cai, J. Gan, et al. "Facial Expression Recognition with Local Binary Pattern and Laplacian Eigenmaps[C]"// International Conference on Emerging Intelligent Computing Technology & Applications. Springer-Verlag, 2009.
12. Asthana, J. Saragih, M. Wagner, et al. "Evaluating aam fitting methods for facial expression recognition[C]" // International Conference on Affective Computing and Intelligent Interaction and Workshops. IEEE, 2019, pp. 1-8.
13. R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2011, pp.38–52.
14. Pradeep, Madhushree, B. S. Sumukha, G. R. Richards and S. P. Prashant, "Facial Emotion Detection using CNN and OpenCV," 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS), Bengaluru, India, 2024, pp. 1-6
15. Chu Wenjin. "Facial Expression Recognition Based on Local Binary Pattern and Gradient Directional Pattern."2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing IEEE, 2013.
16. Chakrabarti, D. Dutta. "Facial expression recognition using PCA and various distance classifiers[M]" // Emerging Trends in Computing and Communication, India: Springer, 2014, pp.79-85.
17. LI Yang, Guo Haiqiao. "Facial Expression Recognition Based on LBP and SVM Decision Tree." Modern Computer: professional edition 2014, pp.40-44.
18. ZHOU Yuxuan, WU Qin, LIANG Jiuzhen, et al. "Facial expression recognition based on discriminative CLBP." Computer Engineering and Applications, 2017, pp.163-169.
19. X. Li, X. Hong, A. Moilanen, et al. "Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods[J]." IEEE Transactions on Affective Computing, 2017, pp.1-1.
20. P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1805–1812.
21. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in Neural Networks (IJCNN), 2015 International Joint Conference on. IEEE, 2015.
22. B. Yang, J. Cao, et al. "Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images[J]," IEEE Access, 2018.
23. L. I. Yong, L. Xiao-Zhu, J. Meng-Ying., "Expression Recognition with Cross-connect Le Net-5 Network," Acta Automatica Sinica, 2018, pp. 176−182
24. K. Thakur, S. K. Bisoy, P. Mishra, R. Pratap Patra and R. Ranjan, "Speech Emotion Recognition Using Machine Learning and Deep Learning," 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU), Bhubaneswar, India, 2024, pp. 1-6
25. S. Ren, K. He, R. Girshick, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]." IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, pp.1137-1149.
26. S. Huansheng, Z. Xiangqing, Z. Baofeng, et al. "Vehicle detection based on deep learning in complex scene[J]." Application Research of Computers, 2018.
27. L. I. Weishan, WEI Chen, WANG Lin, "Improved Faster R-CNN approach for pedestrian detection in underground coal mine," Computer Engineering and Applications, 2019, pp.200-207.

28. Zhigang Liu, Yang Lyu, Liyou Wang, et al. “Detection Approach Based on an Improved Faster RCNN for Brace Sleeve Screws in High-Speed Railways[J].” IEEE Transactions on Instrumentation and Measurement, 2020, 69(7):4395-4403.
29. Pan S. J., Q. Yang, “A Survey on Transfer Learning [J],” IEEE Transactions on Knowledge & Data Engineering, 2010,22(10): 1345-1359.
30. Z. Duan, F. Wang, B. Wang, G. Luo and Z. Jiang, "An Adapted ResNet-50 Architecture for Predicting Flow Fields of an Underwater Vehicle," in IEEE Access IEEE, vol. 12, 2024, pp. 66398-66407,
31. Chen Yong, Guo Hongguang, Ai Yapeng, “Single Image Dehazing Method Based on Multi-scale Convolution Neural Network [J].” Acta Optica Sinica, 2019,39(10).
32. W. Chen, H. Zhao and Z. Wang, "Defect Detection Model of Printed Circuit Board Components Based on the Fusion of Multi-Scale Features and Efficient Channel Attention Mechanism," in IEEE Access IEEE, vol. 12, 2024:pp. 62964-62974.
33. M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with Gabor wavelets,” in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998, pp. 200–205.
34. C. Ledig, L. Theis, F. Huszar, et al. “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network[C]” // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
35. P. Ekman and W. V. Friesen, “Facial action coding system (FACS): a technique for the measurement of facial actions,” Rivista Di Psichiatria, 1978,47(2), pp.126-38.
36. S. Arafin and M. G. R. Alam, "Enhancing the ability to recognize facial expressions in young children: A method using few-shot learning and cross-dataset validation," 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, 2024, pp. 640-64
37. F.Y. Shih, C. F. Chuang, P. S. P. Wang, “Performance comparisons of facial expression recognition in JAFFE database[J],” International Journal of Pattern Recognition and Artificial Intelligence, 2008, 22(03): 445-459.
38. K.C. Huang, Y.H. Kuo, M.F. Horng, et al. “Emotion recognition by a novel triangular facial feature extraction method[J],” International Journal of Innovative Computing Information and Control, 2012, 8(11): 7729-7746.
39. W.F. Gu, C. Xiang, Y.V. Venkatesh, D. Huang, H Lin, “Facial expression recognition using radial encoding of local Gabor features and classifier synthesis,” Pattern Recogn. 2012, 45(1), pp.80–91
40. Chenxing Xue, Jun Zhang, et al. “Research on Edge Detection Operator of a Convolutional Neural Network[C]” 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2019.
41. F.Z. Salmam, A. Madani, M. Kissi, “Emotion recognition from facial expression based on fiducial points detection and using neural network,” Int. J. Electr. Comput. Eng, 2018,8(1), 52.
42. C. Zhu, Y. Zheng, K. Luu, et al. “Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection[J].” Deep learning for biometrics, Cham, 2017, pp.57-79.
43. A Dhall, R Goecke, S Lucey, et al. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark[C] IEEE International Conference on Computer Vision Workshops. IEEE, 2011
44. Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-Aware convolutional neural network for facial expression recognition. IEEE,2017, pp.558–565.
45. Wadhawan, Rohan, and T. K. Gandhi. "Landmark-Aware and Part-based Ensemble Transfer Learning Network for Facial Expression Recognition from Static images." 2021
46. Kacem Anis, Daoudi Mohamed, et al. Alvarez-Paiva. "Barycentric Representation and Metric Learning for Facial Expression Recognition." IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2018.