Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6639598, 17 pages
https://doi.org/10.1155/2021/6639598

Research Article

Facial Expression Recognition Algorithm Based on Fusion of Transformed Multilevel Features and Improved Weighted Voting SVM

Hao Meng,[1,2] **Fei Yuan**,[1,2] **Yue Wu**,[1,2] **and Tianhao Yan**[1,2]

[1]*College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China*
[2]*Key laboratory of Intelligent Technology and Application of Marine Equipment (Harbin Engineering University), Ministry of Education, Harbin 150001, China*

Correspondence should be addressed to Fei Yuan; bohelion@hrbeu.edu.cn

Received 1 November 2020; Revised 24 February 2021; Accepted 12 March 2021; Published 12 April 2021

Academic Editor: Li Haitao

In allusion to the shortcomings of traditional facial expression recognition (FER) that only uses a single feature and the recognition rate is not high, a FER method based on fusion of transformed multilevel features and improved weighted voting SVM (FTMS) is proposed. The algorithm combines the transformed traditional shallow features and convolutional neural network (CNN) deep semantic features and uses an improved weighted voting method to make a comprehensive decision on the results of the four trained SVM classifiers to obtain the final recognition result. The shallow features include local Gabor features, LBP features, and joint geometric features designed in this study, which are composed of distance and deformation characteristics. The deep feature of CNN is the multilayer feature fusion of CNN proposed in this study. This study also proposes to use a better performance SVM classifier with CNN to replace Softmax since the poor distinction between facial expressions. Experiments on the FERPlus database show that the recognition rate of this method is 17.2% higher than that of the traditional CNN, which proves the effectiveness of the fusion of the multilayer convolutional layer features and SVM. FTMS-based facial expression recognition experiments are carried out on the JAFFE and CK+ datasets. Experimental results show that, compared with the single feature, the proposed algorithm has higher recognition rate and robustness and makes full use of the advantages and characteristics of different features.

1. Introduction

FER refers to the use of computers to analyze human facial expressions and judge human psychology and emotions through pattern recognition and machine learning algorithms, thereby achieving intelligent human-computer interaction [1]. Traditional FER methods generally include three steps: face detection, feature extraction, and expression recognition [2, 3]. The most important part is feature extraction, which directly affects the final recognition result.

Texture features commonly used in FER include Gabor and LBP. The Gabor filter has the same characteristics as the receptive field of visual cells and has the ability to analyze subtle changes in images from multiple scales and directions [4].

LBP is a texture operator which can effectively describe the local information of gray scale images [5]. In order to reduce the dimensionality, we often extract histogram features from the LBP feature map instead of directly using the feature map for classification [6]. Geometric feature is to locate key feature points in important feature areas of human face (such as the eyebrows, eyes, nose, and mouth) and then calculate the distance and angle between them [7].

It is determined by a vector sequence formed between key feature points established by some statistical shape models, which can well describe the changes in size, shape, and position caused by changes of facial expressions. In the past few years, many works [8–14] focus on using Gabor, LBP, and geometric features for FER. Gabor and LBP features have a strong description of local texture and more

detailed expression features, but they are not robust. The relationship between geometric features and expression changes is more direct, easier to understand and analyze, and more robust under certain lighting conditions. However, the local description ability of expression information is weak, and the error is large.

The shallow features of traditional hand-designed can no longer adapt well to various interference factors that have nothing to do with expression in the real world. Deep CNN has the ability to mine the deep potential distributed expression characteristics of data, and it is very effective when using deeper layers to learn features with high-level abstractions [15, 16].

In recent years, CNN [17–19] has been widely used in FER. CNN maps the image layer by layer, and the mapping to the end is the result of feature extraction. Traditional CNN usually only uses the last layer of convolutional layer features for image classification. However, the features extracted from the intermediate convolutional layer also contain some information and have certain expressive power in the image [20–22]. Rashid M [23] proposed a sustainable deep learning architecture for accurate object classification, which utilizes the fusion and selection of multilevel deep features. Ren [24] proposed a CNN-based cosaliency detection model, which consists of two key parts including the integration of multilayer convolutional features extracted from a set of images and the interimage saliency propagation. These indicate that the use of the features of the intermediate convolutional layer can improve the feature representation ability of the image, thereby improving the accuracy of the CNN. In addition, CNN usually uses Softmax for classification, but experiments have shown that Softmax is not suitable in the field of FER due to the low distinction between expressions [25, 26]. Currently, many researchers combine the features extracted by CNN with traditional classifiers to have better performance and achieve good results [27–30]. Liu [31] proposed a multilevel structured hybrid forest (MSHF) for joint head detection and pose estimation, which extends the hybrid framework of classification and regression forest. Touil [32] used convolutional features and an online training SVM classifier to detect targets and improve accuracy. The classification accuracy and robustness of the SVM classifier in traditional classifier are better. Pham [33] evaluated the performance of these methods using ROC curves and methods based on statistical indicators by applying five machine learning methods. The experimental results show that the SVM model has the best performance.

Whether the feature is reasonable and effective, it will directly affect the final recognition rate. Single feature often has more or less deficiencies and defects, which cannot meet the conditions of good real-time, high precision, and robustness. In this study, these features are fused, and then, the decision-making level fusion is carried out, learning from each other's strengths, and a FER algorithm based on FTMS is proposed. In the shallow features, in addition to the simple processing of Gabor and LBP, this study proposes a joint geometric feature design method for facial expressions. In the aspect of deep CNN features, this study proposes to use the multilayer features fusion of CNN. Moreover, the Softmax classification of the traditional CNN is abandoned, and the SVM classifier is used to classify facial expressions. Finally, with the weighted voting method proposed in this study, the four classifiers trained based on four features are fused at the decision level to obtain the final recognition result, and the superiority of the new method proposed in this study is verified through experiments.

The rest of the study is organized as follows. Section 2 is about some basic works related to the follow-up. Section 3 summarizes our new algorithm and describes feature fusion and the improved weighted voting method. Section 4 provides the experimental results. Section 5 concludes the study.

2. Related Basic Work

The images in the expression database are the subjects to interference from various aspects, such as light intensity, noise, and size. At the same time, the original expression image also contains certain nonface parts, such as background, hair, and other redundant information. Therefore, it is necessary to reduce these interferences and eliminate redundant information through some preprocessing methods. Face detection is to extract the face parts of the image, remove the nonface parts, and ensure the effectiveness of subsequent feature extraction [34].

In order to facilitate the unified processing, we use the gray scale formula (1) to process the expression dataset images Among them, RGB is the color representing the three channels of red, green, and blue. Graying reduces the image channel, that is, the data dimension, so that the storage space occupied is smaller, and the calculation speed of image data processing is accelerated. We use the Viola–Jones model [35] to detect the face of gray-scaled image and save it, as shown in Figure 1. Finally, the size of the image after face detection is normalized, and bilinear interpolation is used to scale the image to the uniform resolution of each expression dataset, as shown in Figure 2.

$$\text{Gray} = R \times 0.299 + G \times 0.587 + B \times 0.114. \quad (1)$$

After detecting a human face, there are still a few nonface regions, and the redundant information will reduce our final recognition rate. Then, the normalized face expression image is used to label feature points by using ensemble of regression trees [36]. After calibrating the 68 key feature points of the face, three feature regions of the eyes, nose, and mouth are obtained by clipping, as shown in Figure 3. Among them, in order to get the eye part, we find points 17, 19, 24, 26, and 28 near the eyes. Take the abscissa of point 17 as the vertex abscissa x_{17}, the maximum values of y_{19} and y_{24} as the vertex ordinate y_e, (x_e, y_e) as the vertex, $|x_{26} - x_{17}|$ as the width, and $|y_{28} - y_e|$ as the height to draw a rectangle and crop it to get the eye part and $40 * 20$ size for sampling. In the same way, the nose and mouth parts are obtained and sampled with the size of $20 * 10$ and $40 * 20$, respectively, so that three characteristic areas of the eyes, nose, and mouth are obtained.

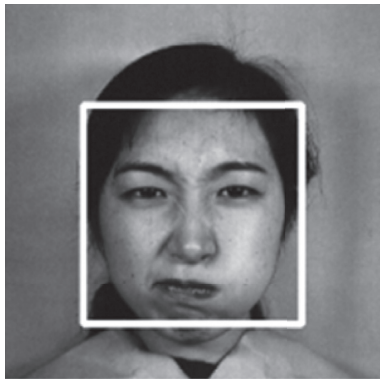

Figure 1: Face detection.

Figure 2: Size normalized.

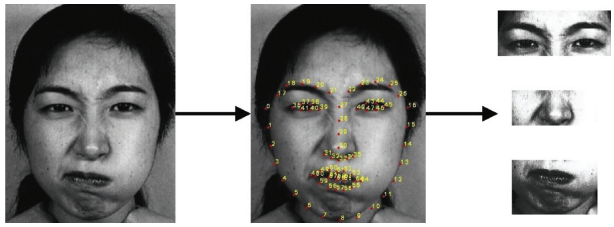

Figure 3: Feature region partitioning of the sample.

3. Approach

3.1. Traditional Shallow Expression Features. Gabor features are obtained from certain feature maps. These feature maps take important facial feature regions (such as the nose and mouth) as input images. By selecting 24 Gabor filters which are closest to the parameters of the receptive field filter of the visual cells, the results are obtained.

LBP histogram features are obtained by connecting 64 small histogram features in sequence. Using the circular LBP operator with a radius of 2, there are 8 points in the field. The LBP feature map is obtained by selecting the uniform LBP mode, and it is evenly divided into 64 small blocks, and each block histogram feature is extracted.

The joint geometric feature proposed in this study is based on the location of feature points. First, the distance feature between feature points is extracted to represent the overall information, the deformation feature is extracted to represent the local information, and the distance feature is connected with the deformation feature finally.

The distance feature represents the overall shape of the face and the distribution information of the eyes, nose, and mouth. We directly calculate the geometric distance between all feature points.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{2}$$

Figure 4 is a calibrated expression image with 68 feature points. Calculate the 67 distance features $d_2^1, d_3^1, \ldots, d_{68}^1$ between the feature point 1 and the other 67 feature points. Calculate the 66 distance features $d_3^2, d_4^2, \ldots, d_{68}^2$ between the feature point 2 and the remaining 66 feature points (no longer seeking distance from the feature point 1 to avoid repetition) and so on. Finally, the relative distance d_{68}^{67} between the 67th feature point and the 68th feature point is calculated. If there are n feature points on the image, the number of all distance features can be calculated as

$$1 + 2 + \cdots + n - 1 = \frac{n(n-1)}{2}. \tag{3}$$

The distance feature vector can be expressed as

$$X = \left(d_2^1, \ldots, d_{n-1}^1\right), \left(d_3^2, \ldots, d_{n-1}^2\right), \ldots, d_{n-1}^{n-1}. \tag{4}$$

The changes in the distance and position of the feature points mostly come from the eyebrows, eyes, mouth, and facial contours, especially when the mouth is open, and the changes in facial contours driven by it will cause significant changes in distance characteristics. Although the distance feature dimension we extracted is not high, there are still some feature redundancies, so we perform principal components analysis (PCA) [37] dimensionality reduction operation. The idea is to map high-dimensional data to low-dimensional space through projection transformation and use the principle of least mean square to obtain the most representative data.

We use indirect deformation features to characterize the deformation information of the local details of the eyes, nose, and mouth regions. Obviously, the local deformation of facial features caused by expressions will cause the changes in the position of feature points in these areas. According to the characteristics of facial muscle movement and facial features deformation, we use a linear combination of distance features between feature points on a part of the facial features area to define nine deformation features. The specific definitions of the nine deformation features are given in Table 1.

After obtaining these nine deformation features, they are correlated with the direct distance features processed by the PCA dimensionality reduction process to obtain joint geometric features. The combined geometric features represent facial expressions from two aspects. At the overall level, the distance features are used to describe the relative positional relationship of important feature points. At the local level, the deformation features are used to describe the facial features caused by changes in expressions.

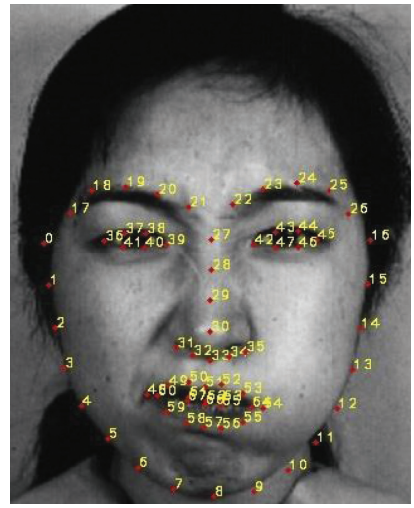

Figure 4: Schematic diagram of distance features.

Table 1: The specific definitions of the nine deformation features.

Location	Calculation method
Curvature of the left eyebrow	$(\mathrm{dis}(18, \mathrm{line}(17,21)) + \mathrm{dis}(19, \mathrm{line}(17,21)) + \mathrm{dis}(20, \mathrm{line}(17,21))/(3 \times \mathrm{dis}(17,21)))$
Curvature of the right eyebrow	$(\mathrm{dis}(23, \mathrm{line}(22,26)) + \mathrm{dis}(24, \mathrm{line}(22,26)) + \mathrm{dis}(25, \mathrm{line}(22,26)))/(3 \times \mathrm{dis}(22,26))$
Left eye closure	$(\mathrm{dis}(37,41) + \mathrm{dis}(38,40))/(2 \times \mathrm{dis}(36,39))$
Right eye closure	$(\mathrm{dis}(43,47) + \mathrm{dis}(44,46))/(2 \times \mathrm{dis}(42,45))$
Nose tip to upper lip distance	$(\mathrm{dis}(33,50) + \mathrm{dis}(33,51) + \mathrm{dis}(33,52))/3$
Inner lip closure	$(\mathrm{dis}(61,67) + \mathrm{dis}(62,66) + \mathrm{dis}(63,65))/(3 \times \mathrm{dis}(61,63))$
Outer lip closure	$(\mathrm{dis}(50,58) + \mathrm{dis}(51,57) + \mathrm{dis}(52,56))/(3 \times \mathrm{dis}(48,54))$
Degree of the mouth to the left	$\lvert y_{33} - y_{48}\rvert / \lvert x_{33} - x_{48}\rvert$
Degree of the mouth to the right	$\lvert y_{33} - y_{54}\rvert / \lvert x_{33} - x_{54}\rvert$

$\mathrm{dis}(m,n)$, the distance from the feature point m to the feature point n; $\mathrm{line}(m,n)$, the straight line determined by the feature points m and n; and (x_n, y_n), the horizontal and vertical coordinate of the feature point n.

3.2. Deep CNN Expression Features. In CNN, different convolution kernels have different sizes and receptive fields. CNN can be regarded as a combination of feature extraction and a classifier. From the perspective of mapping of its various layers, it is similar to the feature extraction process, which extracts features of different levels, through continuous interactive mapping and finally mapped to several tags, with classification function, as shown in Figure 5.

This research uses VGG-16 for feature extraction. After visualizing the convolutional layer through feature map [38], the feature map of each channel can be obtained, and each channel is fused according to 1 : 1 to obtain the fused feature map. Figure 6 shows the convolutional layer map after channel fusion. Through the visualization of the feature map, it can be seen that the shallow features are more inclined to detect the edge of the image and the detected content is more comprehensive. With the deepening of the hierarchy, the feature map becomes more abstract, and the resolution of the image becomes smaller and smaller. In contrast, the deeper the layers, the more representative the extracted features. The traditional VGG model trained on ImageNet only uses the output features of the last convolutional layer, that is, the output vector of the last fully connected layer FC3 before Softmax classification. But the intermediate feature information also has a certain expressive ability for images.

This study proposes to use the features of the subdeep convolutional layer conv5_2 of the CNN to fuse the features of the deepest convolutional layer con5_3. The selection of subdeep features can ensure that deeper features can be obtained when the original features are relatively complete. The deeper the number of layers, the higher the level of semantic information for extracting features and the more sufficient the semantic information. In addition, Softmax is not very suitable for FER because of the low discrimination of facial expression. In this study, a better-performing SVM classifier is selected to improve the accuracy of recognition and the generalization ability of the model. Rely on the powerful learning ability of CNN to learn deep feature representation and then use SVM for expression recognition.

This study established a multilayer CNN structure as shown in Table 2. The output feature vectors of the VGG subdeep convolutional layer conv5_2 and the deepest conv5_3 are fused and sent to the network for training; the feature vector of conv8 is extracted and sent to the SVM classifier for classification training. The multilayer CNN designed in this study does not use the traditional pooling method for downsampling, but the use of convolutional layers for downsampling can strengthen the learning ability of the network. The loss function of SVM is given as formula

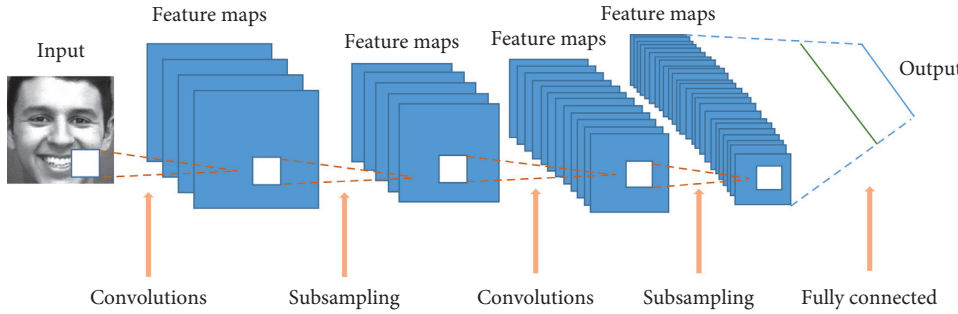

Figure 5: Traditional convolutional neural network structure.

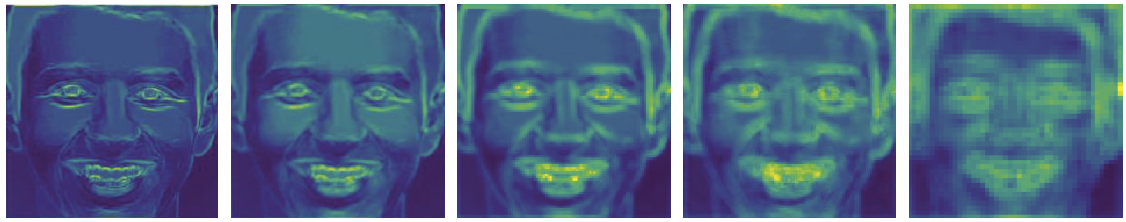

Figure 6: Feature map visualization.

Table 2: Multilayer convolutional network structure.

Layer	Input (W ∗ H ∗ D)	Kernel_num	Kernel_size	Stride	Pad	Out (W ∗ H ∗ D)
Conv1	14 ∗ 14 ∗ 1024	1024	3	1	0	12 ∗ 12 ∗ 1024
Conv2	12 ∗ 12 ∗ 1024	1024	3	1	0	10 ∗ 10 ∗ 1024
Conv3	10 ∗ 10 ∗ 1024	1024	3	1	0	8 ∗ 8 ∗ 1024
Conv4	8 ∗ 8 ∗ 1024	1024	3	1	0	6 ∗ 6 ∗ 1024
Conv5	6 ∗ 6 ∗ 1024	1024	3	1	0	4 ∗ 4 ∗ 1024
Conv5	4 ∗ 4 ∗ 1024	1024	3	1	0	2 ∗ 2 ∗ 1024
Conv6	2 ∗ 2 ∗ 1024	512	2	1	0	1 ∗ 1 ∗ 512
Conv7	1 ∗ 1 ∗ 512	256	1	1	0	1 ∗ 1 ∗ 256
Conv8	1 ∗ 1 ∗ 256	8	1	1	0	1 ∗ 1 ∗ 8

(5). The better the model, the score of the correct category should be higher than the scores of other error categories, as for how high the threshold (Δ) is determined by us. If it is above the threshold, we believe that the correct category is well distinguished from the specific category. We give zero loss to distinguish between these two categories. Conversely, if a wrong category has a higher score than the correct category, it means that the model distinguishes the two categories badly.

$$L_i = \sum_{j \neq y_i} \max\left(\Delta, \omega_j x_i - \omega_{y_i} x_i + 1\right), \quad i \in Z^+, j \in Z. \tag{5}$$

Among them, y_i is the label corresponding to sample x_i , j corresponds to a number of a certain category, $\omega_j x_i$ is the score of misclassification, and $\omega_{y_i} x_i$ is the score of correct classification.

The design network structure flow chart of this study is shown in Figure 7. The facial expression image is input to the VGG network for feature extraction. The subdeep feature vector and the deepest convolutional layer feature vector are extracted and then merged. The fused feature vectors are used as the input of the multilayered CNN established in this study (Table 1). The feature vectors of the conv8 layer are extracted and sent to the SVM classifier.

3.3. Feature Fusion. The so-called feature fusion refers to independently proposing various single expression features, analyzing their advantages and disadvantages and applicable environment, and then making comprehensive decisions to formulate the most reasonable recognition plan. According to the theory of information fusion, information fusion can be realized at four levels: pixel level, feature level, matching level, and decision level [39], which requires an effective fusion strategy.

Comprehensively, consider the extracted Gabor features, LBP features, joint geometric features, and deep CNN features.

(1) From the perspective of feature categories, Gabor and LBP features are used as texture features, joint geometric features as geometric features, and CNN features as deep abstract features; they are relatively independent and have almost no correlation feature categories.

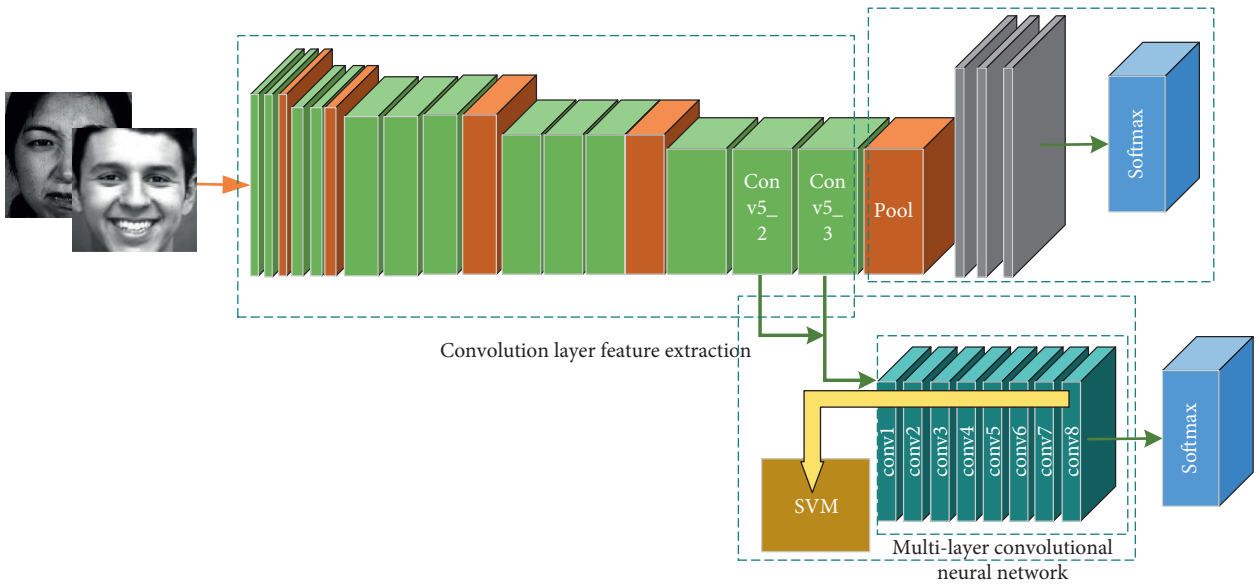

Figure 7: Designed network structure.

(2) From the perspective of feature synthesis, although both Gabor and LBP features are texture features, the calculation methods are quite different. Gabor features exist in the form of directly expanded feature maps, while LBP features extract histogram features from LBP feature maps. There is no strong correlation between them, so their feature-level fusion is easy to lose a lot of information.

(3) From the characteristics of the features, although we added local deformation features to the joint geometric features, its local description ability for expression information is still weak and the error is large, while the Gabor features and LBP features are highly descriptive and highly accurate but not robust, so merging them can achieve the complementary effect.

(4) From the representativeness of features, shallow features are more inclined to detect the edge of the image, and the detected content is comprehensive and key information will also be extracted. As the layers deepen, the feature map becomes more and more abstract, the resolution of the image is getting smaller and smaller, and much information is also ignored. Relatively speaking, the deeper the layer, the more representative the extracted features. The extraction of depth features adds semantic information to the image based on Gabor, LBP, and joint geometric features.

In summary, we choose a higher level of fusion, that is, a decision-level fusion to address our four-feature fusion problem.

3.4. Improved Weighted Voting Classification

3.4.1. Multiclassifier Voting Mechanism. After extracting the features, the facial expressions are classified. This study uses SVM to complete the classification task. Decision-level fusion is actually training the SVM classifier with four features and then multiclassifier combination of the four classifiers. This study proposes to use an improved weighted voting method to make a comprehensive decision on the four SVM classifiers and finally determine the recognition effect.

The voting method is a relatively simple and specific method to realize parallel combination. Its implementation principle is the "one person, one vote" mechanism. But such an overly simple voting rule does not take into account the characteristics of the classifier itself, which will make the classification result worse. From the above analysis, we can see that the feature composition and characteristics we use to train each classifier are different and the recognition capabilities are different; in many cases, we will not use the same classifier, that is, the principles and methods of each classifier are different; even in each classifier, we use different datasets for training. So the recognition ability of each classifier is bound to be different. Obviously, the "one person, one vote" mechanism is not reasonable enough. We adopt the "one person, multiple votes" mechanism, that is, each classifier should be given different weights.

The experimental results show that using the recognition accuracy of a single classifier as the prerequisite and calculation basis for weight setting can further improve the classification effect. The specific process of the expression recognition algorithm proposed in this study using mixed features for weighted voting SVM classification is shown in Figure 8.

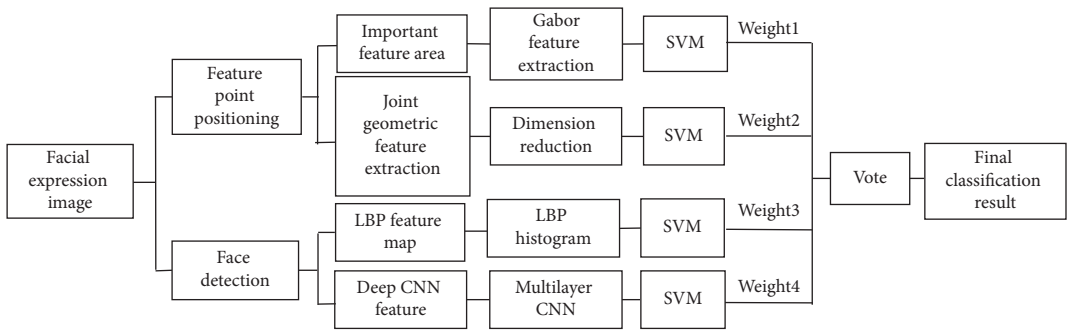

Figure 8: Schematic diagram of weighted voting SVM classification using mixed features.

3.4.2. Weight Calculation. We use the recognition accuracy rate of different features under the same database and a certain expression as the basis of weight setting and take the proportion of the final recognition rate of a certain feature in the sum of the four feature recognition rates as the weight of this feature value. Therefore, features with higher recognition rate can have a greater right to speak and play a greater role in the final decision.

Perform experiment with different features and record the recognition rate of different expressions separately:

Angry: $R_{\text{Gabor}}(a)$, $R_{\text{LBP}}(a)$, $R_{\text{UG}}(a)$, $R_{\text{MC}}(a)$
Disgust: $R_{\text{Gabor}}(d)$, $R_{\text{LBP}}(d)$, $R_{\text{UG}}(d)$, $R_{\text{MC}}(d)$
Fear: $R_{\text{Gabor}}(f)$, $R_{\text{LBP}}(f)$, $R_{\text{UG}}(f)$, $R_{\text{MC}}(f)$
Happy: $R_{\text{Gabor}}(h)$, $R_{\text{LBP}}(h)$, $R_{\text{UG}}(h)$, $R_{\text{MC}}(h)$
Neutral: $R_{\text{Gabor}}(n)$, $R_{\text{LBP}}(n)$, $R_{\text{UG}}(n)$, $R_{\text{MC}}(n)$
Sad: $R_{\text{Gabor}}(\text{sa})$, $R_{\text{LBP}}(\text{sa})$, $R_{\text{UG}}(\text{sa})$, $R_{\text{MC}}(\text{sa})$
Surprise: $R_{\text{Gabor}}(\text{su})$, $R_{\text{LBP}}(\text{su})$, $R_{\text{UG}}(\text{su})$, $R_{\text{MC}}(\text{su})$

For the same expression, calculate the proportion of different feature recognition rates:

Angry:

$$W_m(a) = \frac{R_i(a)}{R_{\text{UG}}(a) + R_{\text{Gabor}}(a) + R_{\text{LBP}}(a) + R_{\text{MC}}(a)}. \tag{6}$$

Disgust:

$$W_m(d) = \frac{R_i(d)}{R_{\text{UG}}(d) + R_{\text{Gabor}}(d) + R_{\text{LBP}}(d) + R_{\text{MC}}(d)}. \tag{7}$$

Fear:

$$W_m(f) = \frac{R_i(f)}{R_{\text{UG}}(f) + R_{\text{Gabor}}(f) + R_{\text{LBP}}(f) + R_{\text{MC}}(f)}. \tag{8}$$

Happy:

$$W_m(h) = \frac{R_i(h)}{R_{\text{UG}}(h) + R_{\text{Gabor}}(h) + R_{\text{LBP}}(h) + R_{\text{MC}}(h)}. \tag{9}$$

Neutral:

$$W_m(n) = \frac{R_i(n)}{R_{\text{UG}}(n) + R_{\text{Gabor}}(n) + R_{\text{LBP}}(n) + R_{\text{MC}}(n)}. \tag{10}$$

Sad:

$$W_m(\text{sa}) = \frac{R_i(\text{sa})}{R_{\text{UG}}(\text{sa}) + R_{\text{Gabor}}(\text{sa}) + R_{\text{LBP}}(\text{sa}) + R_{\text{MC}}(\text{sa})}. \tag{11}$$

Surprise:

$$W_m(\text{su}) = \frac{R_i(\text{su})}{R_{\text{UG}}(\text{su}) + R_{\text{Gabor}}(\text{su}) + R_{\text{LBP}}(\text{su}) + R_{\text{MC}}(\text{su})}, \tag{12}$$

Where m = Gabor, LBP, UG, MC. Take the proportion of the recognition rate of a certain feature in the same expression database and a certain expression in the sum of the four feature recognition rates as the weight of this feature. In the end, the fusion strategy of the improved multiclassifier voting method is

$$n = \arg\max_{i=1,\ldots,N} \left(\sum_{m=1}^{L} W_{mi} \text{vote}_{mi} \right), \tag{13}$$

Where N is the number of expression categories, L is the number of classifiers, and W_{mi} is the weight of the i^{th} expression of the current m^{th} classifier. The value of vote_{mi} is 0 or 1, which indicates whether the recognition result of the current m^{th} classifier is an i-type expression.

4. Experiments

Our experiment results including two parts are shown in this section. Section 4.1, the experiment of the CNN deep features proposed in this study with the FERPlus [40] dataset. After proving the effectiveness of our proposed fusion of multilayer convolutional layer features as CNN deep features, in Section 4.2, the FTMS-based expression recognition experiment was carried out with JAFFE [41] and CK+ [42] databases; the results were compared and analyzed.

The experimental environment are Win10, Python 3, TensorFlow, Visual Studio 2013, and OpenCV 2.4.9, graphics processing unit (GPU) is NVIDIA GeForce RTX 2080 Ti, and video memory is 11 GB.

4.1. Convolutional Neural Network Deep Features

4.1.1. Database. In the FERPlus dataset, there are 10 categories of tags: neutral, happiness, surprise, sadness, anger, disgust, fear, contempt, unknown, and NF. In this study, the unknowns and NF are removed, and there are a total of 8 expression categories. Process the image into a fixed-size data format to facilitate the input of data into the neural network. The size of the face image in the FERPlus dataset is 48 pixels × 48 pixels. In this study, the image is processed into a size of 224 pixels × 224 pixels. The dataset is divided into three parts: training set, validation set, and test set.

4.1.2. Experimental Steps

(1) Send the facial expression dataset directly to the VGG network for training classification and record the final test results

(2) Extract the feature vectors of conv5_2 and conv5_3 from the network model, send them to the multilayer CNN we built (Table 1) after fusion, and then train and record the final test results

(3) Extract the feature vector of the conv8 layer from the multilayer CNN we built, send it to the SVM classifier for classification, and record the final test result

4.1.3. Experimental Results. The FERPlus dataset was directly sent to the VGG network for migration learning and only the last classification layer was changed; the original 1000 categories were changed to 8 categories. Figure 9 is a graph of accuracy and loss function during training. The vectors after the conv5_3 and conv5_3 layer features are fused as the input vectors into the multilayer CNN established in this study, and Softmax is used for classification training. Figure 10 is a graph of accuracy and loss function during training. Among them, Figures 9(a) and 10(a) show a graph of accuracy of training and validation. Figures 9(b) and 10(b) show a graph of loss function of training and validation. The orange represents training and the blue represents validation.

It can be seen from Figure 9 that the test accuracy rate on the final test set is 51.7% and the final average value of the loss function is 1.3. The accuracy rate is relatively low, the loss function value is relatively large, and the curve oscillation is relatively large and unstable. It can be seen from Figure 10 that the test accuracy rate on the final test set is 59.4% and the final average value of the loss function is 1.1. Compared with using only the deepest layer features, the accuracy is improved by 14.9%, the loss function value is reduced by 15.4%, and the oscillation amplitude becomes smaller and the curve is smoother. It proves the effectiveness of our proposed multilayer convolutional layer feature fusion.

Figure 11 shows the confusion matrix directly using VGG for expression classification. The recognition rate of fear, happy, and surprise is relatively high. The recognition rate of angry, disgust, neutral, and sad is all very low, which is already lower than 50%. Among them, the recognition rate of disgust and neutral is lower than 45%, and it is extremely low. According to the classification results, it is found that some facial expression classification results are quite extreme, indicating that the network is not stable enough.

Figure 12 shows the confusion matrix for expression classification using our improved CNN. When a CNN trained with multilayer convolution features is used to classify each type of expression, the average accuracy will be improved by 14.9%. Except that the recognition rate of disgust has just reached 50%, the recognition rate of each type of expression is higher than 55% and the classification result is more stable than the original neural network.

Figure 13 shows the confusion matrix using our improved deep CNN features and SVM classification. When the features processed by the multilayer CNN are fed into the SVM to classify each type of expression, the average accuracy rate will be increased by 2% compared with that in Figure 12 and 17.2% in Figure 11. Except that the recognition rate of disgust is slightly lower, the recognition rate of other expression categories is significantly higher than that of the original CNN and the classification result will be more stable.

The experimental results show that the use of fusion multilayer convolutional layer features can enrich expression features, and our proposed multilayer CNN can reduce the loss of features, thereby improving the accuracy of expression classification. The SVM classifier is more suitable for facial expression classification than the Softmax classifier, which can improve the robustness of the network. It proves the effectiveness of our proposed improved deep CNN feature network.

4.2. Fusion of Transformed Multilevel Features and Improved Weighted Voting SVM

4.2.1. Database. In the JAFFE dataset, each image of each person's expression is selected as the training set with a total of 70 images and the remaining 143 images as the test set to ensure that the number of samples in the test set is sufficient. Select a total of 3 times to obtain the average recognition rate.

In the CK+ dataset, considering that the number of each type of expression in the CK+ database is not balanced, we select 1–4 images with the peak expression (or close to the peak) from each tagged sequence when processing CK+ as the experimental images, a total of 736 images. The selected 736 images included 92 angry expressions, 116 disgusted expressions, 100 fear expressions, 104 happy expressions, 112 sad expressions, 104 surprised expressions, and 108 neutral expressions. Half of each expression image is randomly selected as a total of 368 training images, and the

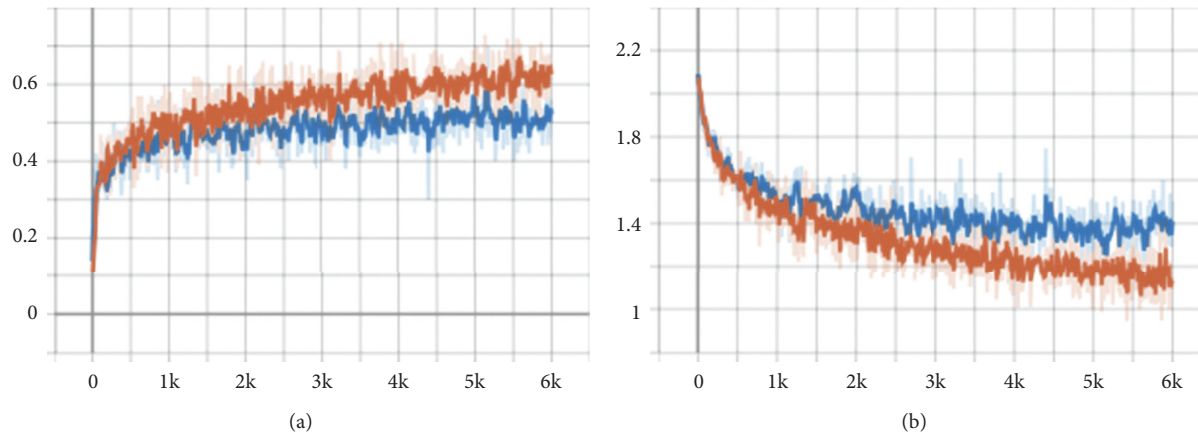

Figure 9: Accuracy curve and loss function of the original VGG network. (a) Accuracy chart. (b) Loss function graph.

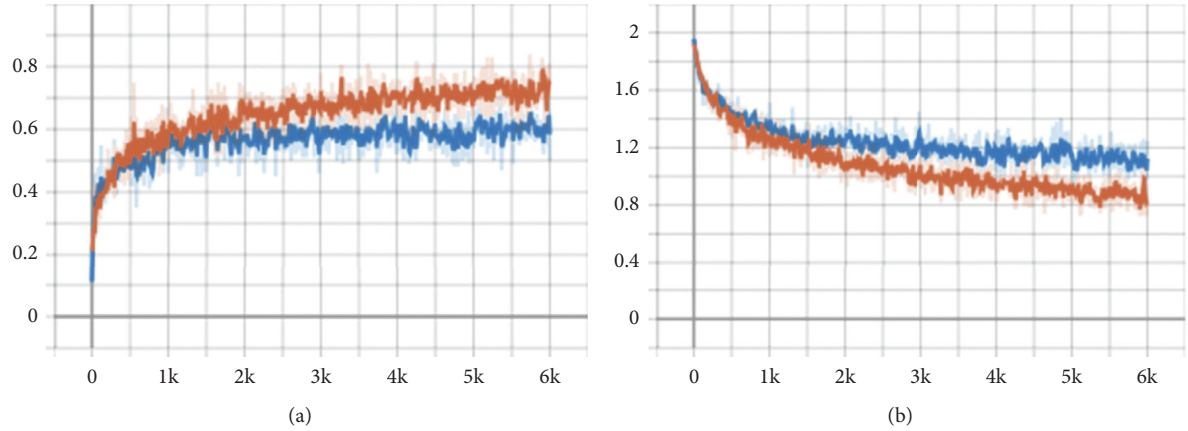

Figure 10: Accuracy curve and loss function of the fusion multilayer convolutional layer feature network. (a) Accuracy chart. (b) Loss function graph.

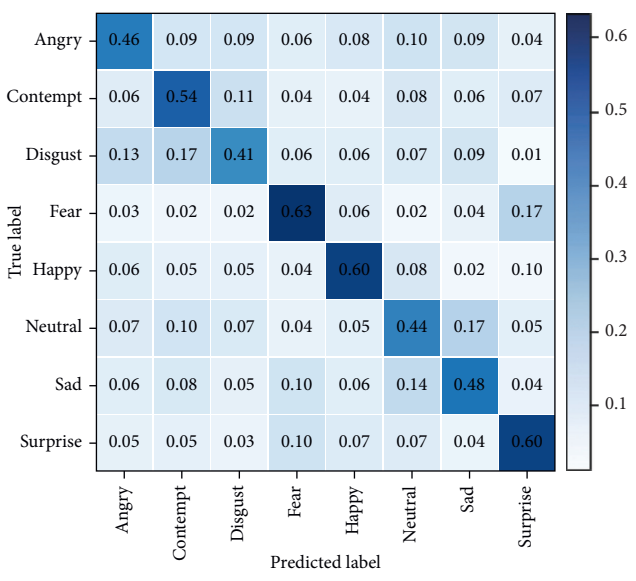

Figure 11: Confusion matrix of facial expression recognition rate based on CNN.

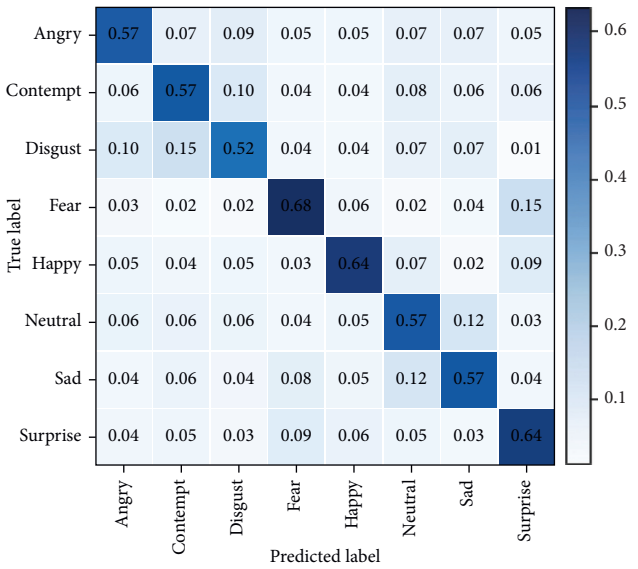

Figure 12: Confusion matrix of facial expression recognition rate based on improved CNN.

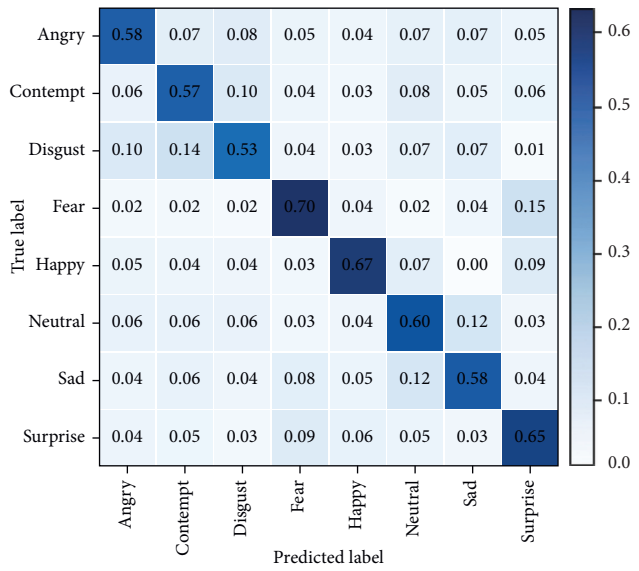

Figure 13: Confusion matrix of facial expression recognition rate based on improved CNN + SVM.

remaining 368 images are used as test images. Repeat the experiment 3 times to get the average value.

4.2.2. Experimental Steps

(1) First, the preprocessing of the facial expression image is needed, including graying, face detection, size normalization, and key feature point positioning of the face

(2) Based on the localization of feature points, three areas of the eyes, nose, and mouth are extracted. Gabor features are extracted from the divided areas, and expression classification is performed using SVM. This study selects the frequency bandwidth $b = 1.4, 1.6, 2.0$ to calculate the wavelength $\lambda = 2.4\sigma, 2.7\sigma, 3.2\sigma$. Select 8 directions and 24 filters in total. The specific experimental process is shown in Figure 14.

(3) Extract the LBP histogram features from the normalized expression images and use SVM to complete the expression classification. The specific experimental process is shown in Figure 15.

(4) Based on the location of feature points, the joint geometric features are extracted and the expression recognition is completed by the SVM. The specific experimental process is shown in Figure 16.

(5) The expression images with normalized size are sent to VGG to extract deep CNN features and then are sent to the SVM for expression recognition. The specific experimental network structure is shown in Figure 7. The SVM classifier has been trained in the experiment as given in Section 4.1.

(6) According to the improved weighted voting method proposed in this study, a decision-level fusion of the four feature training classifiers is carried out to obtain the final classification result.

4.2.3. Experimental Results. A well-trained SVM classifier based on four features was tested on the test set. The final recognition rate results of the four features in the two databases are shown in Figures 17 and 18.

From the results of the recognition rate, we can see

(1) The average recognition rate of the facial expression recognition algorithm based on Gabor features on JAFFE and CK+ reached 88.49% and 92.86%, respectively. The expression recognition algorithm based on LBP features reached 89.27% and 92.35% on JAFFE and CK+, respectively. The expression recognition algorithm based on joint geometric features reached 80.49% and 92.49% on JAFFE and CK+, respectively. The expression recognition algorithm based on deep CNN features has an average recognition rate of 92.7% and 95.61% on JAFFE and CK+, respectively. The feasibility of the recognition algorithm based on four independent feature expressions is verified. The recognition rate of the CK+ is relatively high. The reason is that CK+ has more pictures, a large sample size, and sufficient training, although the CK+ database contains samples of different genders and skin colors.

(2) Compared with Gabor features, LBP features have better performance and more balanced recognition ability in JAFFE; however, the recognition rate of CK+ is lower because CK+ is more complex, including samples with different skin color brightness and poor clarity. It can be concluded that LBP features have higher requirements for image quality and poor noise immunity. Compared with Gabor and LBP features, the recognition rate of joint geometric features and deep CNN features on JAFFE decreases, while the recognition rate on CK+ is still high. This shows that the joint geometric features and the deep CNN features have a poorer recognition effect when the sample size is small; when the sample size is large, it will perform better, especially the deep features of convolutional nerves, which may be overfitting. It is also confirmed that the deep CNN features are robust to a small amount of brightness changes.

According to the method of Section 3.4.2, through calculation, the optimal weight is finally selected as given in Tables 3 and 4.

According to formula (13), decision-level classification is performed. After testing, the final recognition rate results of these two databases are shown in Tables 5-6. Compare it with the results of using these four features separately as shown in Figure 19.

It can be seen from the recognition rate in Tables 5-6 that the average recognition rate of the facial expression recognition algorithm based on FTMS proposed in this study on JAFFE and CK+ databases reached 94.95% and 96.68%, respectively, which verified the feasibility of the algorithm. From the comparison of different databases, CK+ still has the highest recognition rate due to its large sample size and sufficient training.

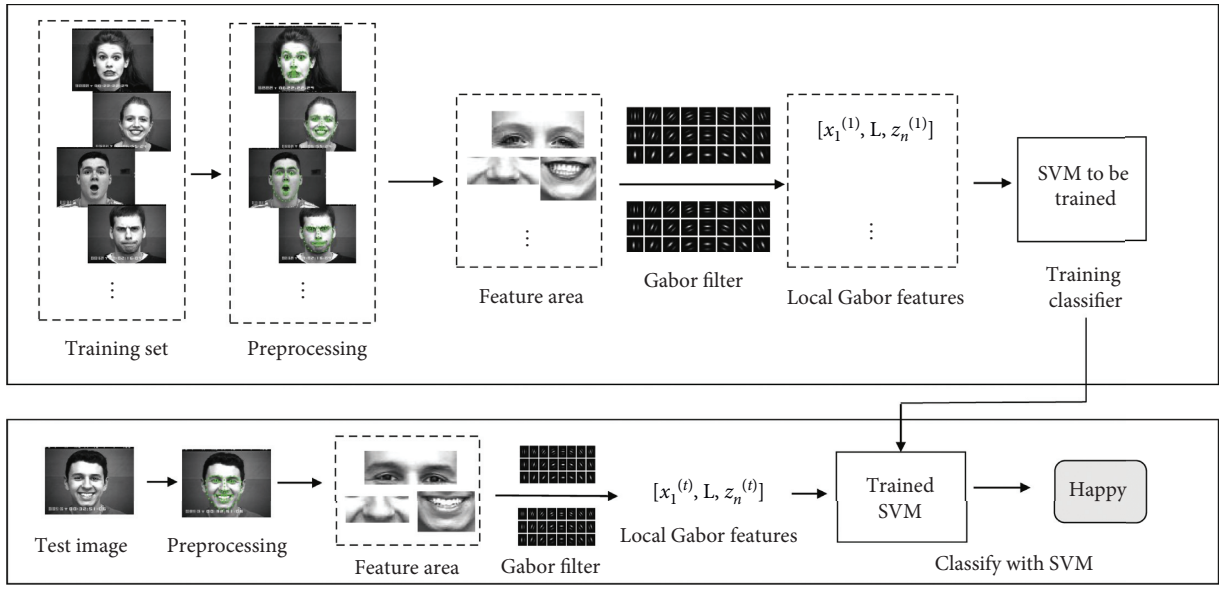

Figure 14: Schematic of the experimental process based on Gabor features.

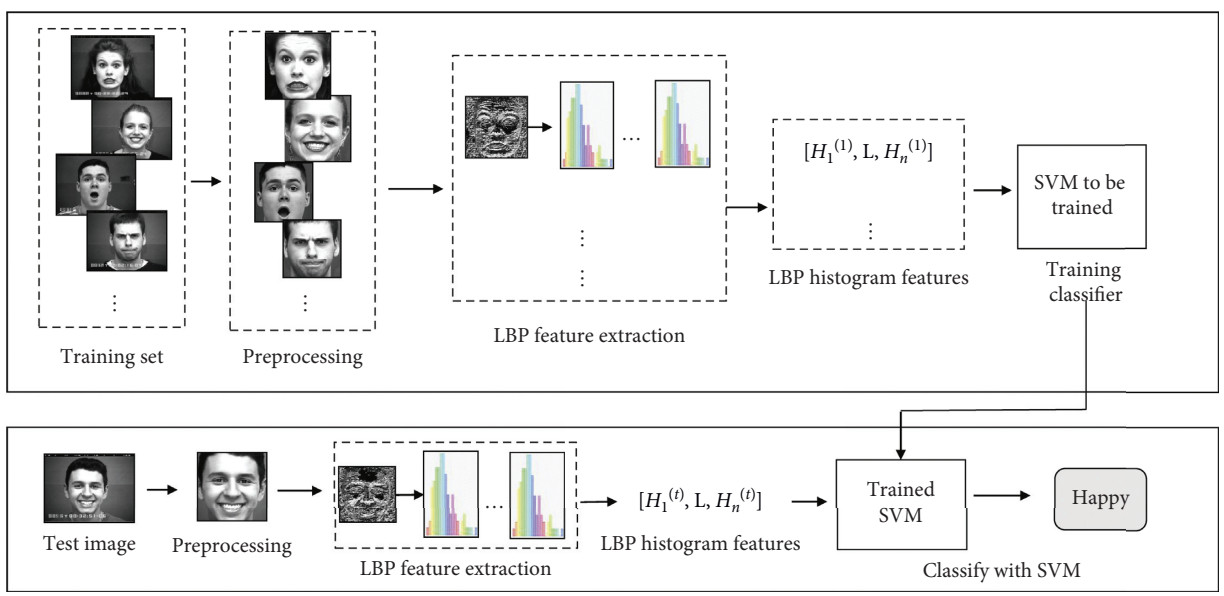

Figure 15: Schematic diagram of LBP feature experiment.

Compared with the experimental results of the four features from Figure 19, on JAFFE, the proposed algorithm increases the recognition rate of Gabor features by 7.3%, the recognition rate of LBP features by 6.4%, the recognition rate of joint geometric features by 18.0%, and the recognition rate of deep CNN features by 2.4% and on CK+, the proposed algorithm increases the recognition rate of Gabor features by 4.1%, the recognition rate of LBP features by 4.7%, the recognition rate of joint geometric features by 4.5%, and the recognition rate of deep CNN features by 1.4%. It can be found that whether it is Gabor, LBP, joint geometric, or deep CNN features, the recognition rate of mixed features in all databases has been significantly improved. And the ability to recognize different expressions is more balanced. This is because the weighted voting system fusion strategy takes advantage of each feature and significantly improves the recognition ability.

Tables 7 and 8 are the confusion matrices after three repeated experiments on the two expression databases. It can be seen that with the use of hybrid features, as the recognition rate increases, the degree of expression confusion decreases. In JAFFE, sadness is easily misjudged as fear and surprise are easily misjudged as happy. In CK+, except for neutral expressions, it is easy to misjudge anger and sadness, and the degree of misjudgment of other expressions is not high. Take JAFFE as an example and print out some misclassified expressions, as shown in Figure 20. The first row below each image represents the predicted expression category and the second row represents the correct label category. It can be seen that some expressions are very

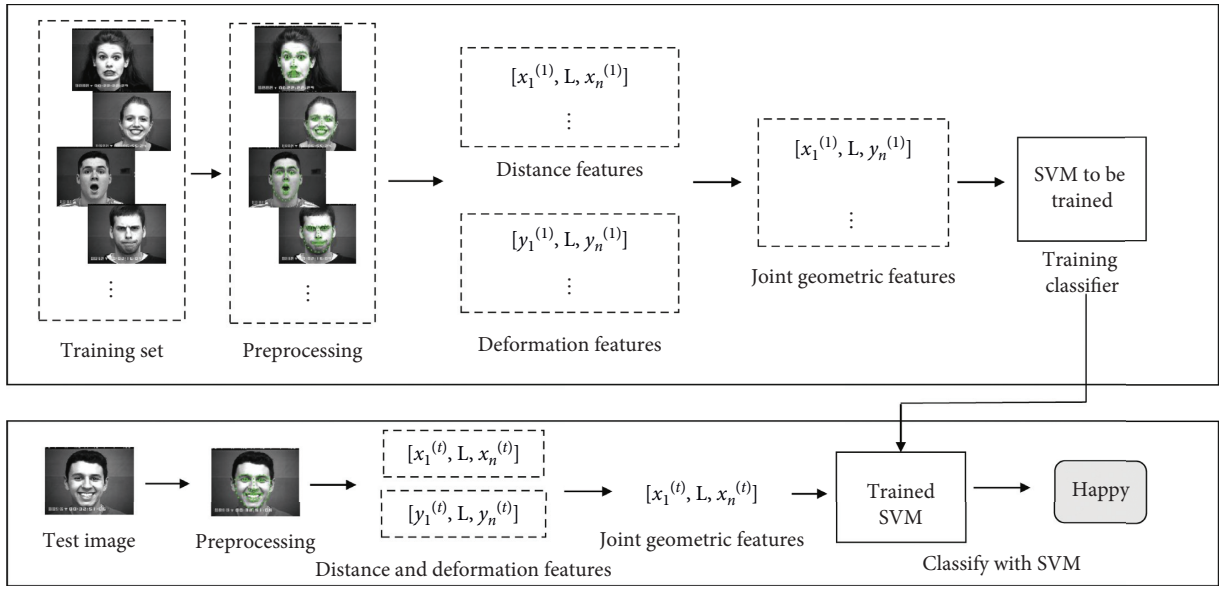

Figure 16: Schematic diagram of the experimental process of joint geometric features.

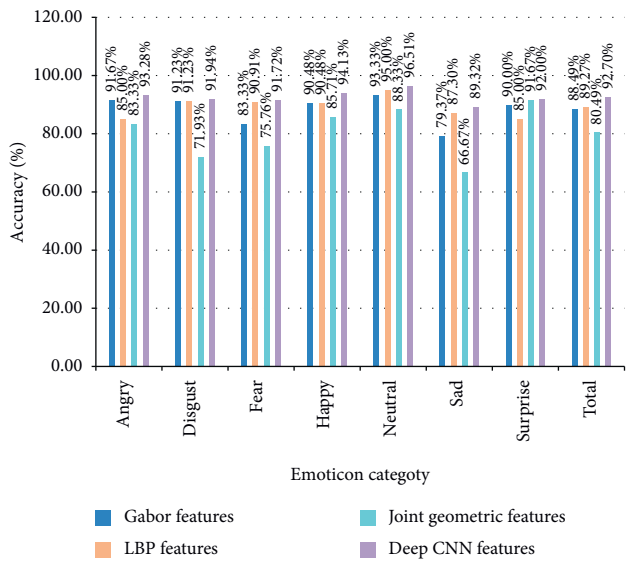

Figure 17: Experimental results of four features on the JAFFE database.

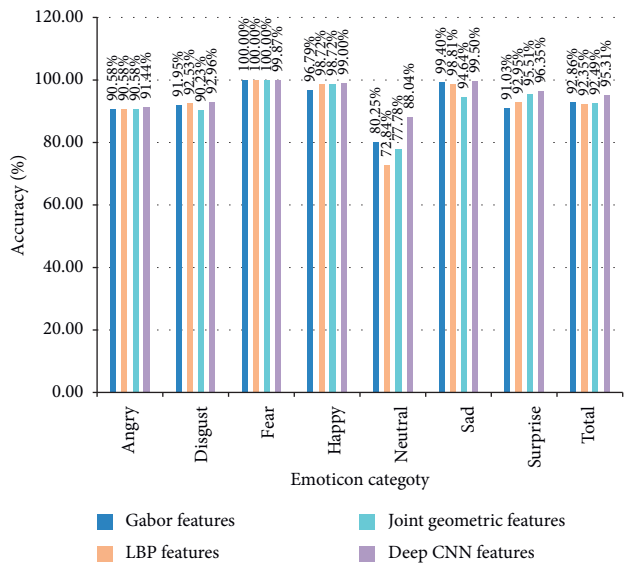

Figure 18: Experimental results of three features on the CK + database.

complicated and difficult to distinguish. For example, people can be angry, happy, sad, disgust, and fear with a blank face. These will be classified as normal expressions. In addition, people cry happily, cry in fear, or cry in anger. These are all classified as sad. Moreover, surprise and fear are often inseparable, surprise and happiness are often inseparable, and exaggerated expression of disgust can easily be classified as sad. Overall, the use of mixed features improves the recognition rate and reduces the degree of misjudgment, which proves the effectiveness of the proposed fusion features in facial expression recognition.

Figure 21 compares our proposed fusion feature with other combinations of fusion features. When the features are fused, the weights are calculated according to the method in Section 3.4.2, so that the respective trained SVM classifiers can be determined comprehensively according to the weights to obtain the result. It can be seen that among these expression recognition methods, the fusion feature performance we proposed is the best. On the JAFFE database, compared with the fusion of Gabor and LBP features, the fusion of Gabor, LBP, and joint geometric features increases the expression recognition rate by 1.5%, while on CK+, it increases by 1.6%, which proves the effectiveness of fusion of joint geometric features. Furthermore, coupled with the CNN deep features, the expression recognition rate is further improved by 1.9% on JAFFE compared to the fusion of Gabor, LBP, and joint geometric features and by 1.3% on CK+, which proves the

Table 3: Optimal weights based on the JAFFE database.

Weight	Gabor	LBP	Joint geometric	Deep CNN
Angry	0.2716	0.2502	0.2469	0.2313
Disgust	0.2543	0.2543	0.2487	0.2427
Fear	0.2500	0.2698	0.2421	0.2378
Happy	0.2563	0.2563	0.2414	0.2460
Neutral	0.2543	0.2634	0.236	0.2463
Sad	0.2568	0.2708	0.2234	0.2490
Surprise	0.2542	0.2354	0.2605	0.2499

Table 4: Optimal weights based on the CK+ database.

Weight	Gabor	LBP	Joint geometric	Deep CNN
Angry	0.2449	0.2449	0.2503	0.2599
Disgust	0.2447	0.2468	0.2452	0.2633
Fear	0.2451	0.2451	0.2451	0.2647
Happy	0.2376	0.2503	0.2534	0.2587
Neutral	0.2536	0.2355	0.2569	0.2540
Sad	0.2520	0.2504	0.2432	0.2544
Surprise	0.2357	0.2516	0.2529	0.2598

Table 5: Experimental results of mixed features on the JAFFE database.

Expression	Number of test samples	Correct number			Average recognition rate (%)
		First	Second	Third	
Angry	20	19	20	19	96.67
Disgust	19	18	18	19	96.49
Fear	22	20	20	19	92.42
Happy	21	20	21	18	93.65
Neutral	20	20	20	20	100.00
Sad	21	20	18	18	92.06
Surprise	20	19	19	18	93.33
Total	143	136	135	132	94. 95

Table 6: Experimental results of mixed features on the CK+ database.

Expression	Number of test samples	Correct number			Average recognition rate (%)
		First	Second	Third	
Angry	46	46	40	46	94.93
Disgust	19	19	18	19	94.25
Fear	50	50	50	50	100.00
Happy	52	52	51	52	99.36
Neutral	54	48	50	46	90.12
Sad	56	56	56	56	100.00
Surprise	52	52	49	52	98.08
Total	368	361	340	357	96.68

effectiveness of the fusion of CNN deep features. In general, the FTMS algorithm we proposed has a certain improvement in the recognition rate of facial expressions and has practical engineering significance.

5. Summary and Discussion

In this study, an expression recognition algorithm based on FTMS is proposed; in this method, the shallow features and deep semantic features extracted are fused effectively. The transformed multilevel features proposed in this study include three shallow features and CNN deep features. The shallow features include local Gabor, LBP, and the joint geometric features designed in this study. For deep features, this study establishes a CNN model that incorporates features of multiple convolutional layers. These four features are used to train four SVM classifiers to obtain the classification results. The improved weighted voting strategy is used to complete the decision-level feature fusion to obtain the final result. The algorithm combines the advantages of

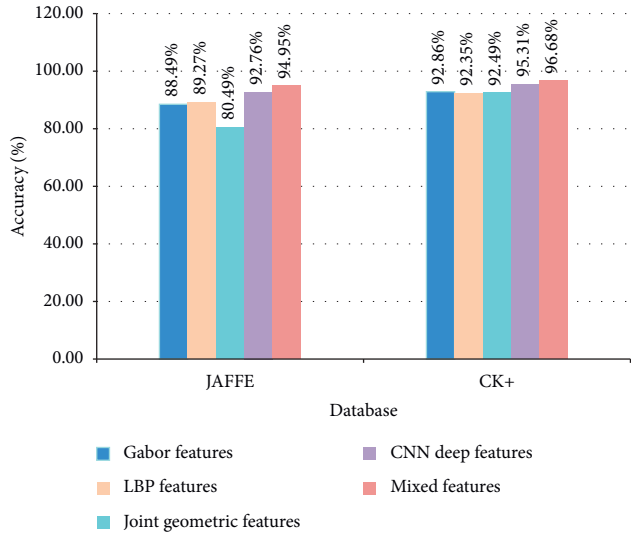

Figure 19: Final result comparison histogram.

Table 7: The total confusion matrix after three experiments on JAFFE using mixed features.

Forecast result / Actual expression	Number of samples	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	60	58	1	0	0	1	0	0
Disgust	57	0	55	0	0	0	2	0
Fear	66	0	1	61	0	1	1	2
Happy	63	0	0	0	59	2	2	0
Neutral	60	0	0	0	0	60	0	0
Sad	63	1	0	3	1	0	58	0
Surprise	60	0	0	1	2	1	0	56

Table 8: The total confusion matrix after three experiments on CK+ using mixed features.

Forecast result / Actual expression	Number of samples	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	138	131	0	0	0	5	2	0
Disgust	174	7	164	0	0	3	0	0
Fear	150	0	0	150	0	0	0	0
Happy	156	0	1	0	155	0	0	0
Neutral	162	6	0	1	0	146	7	2
Sad	168	0	0	0	0	0	168	0
Surprise	156	0	0	0	0	3	0	153

each feature and achieves significantly higher recognition results than the single feature on both expression databases. The average recognition rates of JAFFE and CK+ are 94.75% and 96.86%, respectively. The features are significantly improved, the recognition effect is excellent, and the ability to recognize different expressions is more balanced. The experimental results show that the algorithm has higher recognition rate and robustness than the single feature and fully utilizes the advantages and characteristics of different features.

Although the algorithm proposed in this study has achieved good results in experiments, it still has certain shortcomings. The work that needs further improvement in the future research process includes

(1) The expression recognition algorithm in this study is mainly for static images; but in fact, the change of facial expression is a complex dynamic process. When our recognition object is an image sequence or a dynamic video, we must consider not only the static features but also how to extract effective features from the dynamic sequence, and the algorithm complexity will also increase greatly. Therefore, how to design a dynamic and static expression recognition system is also a problem worthy of exploring.

(2) The facial expression database used in this study is a commonly used database. The expression images are taken in a specific experimental environment and may not get the most real and natural

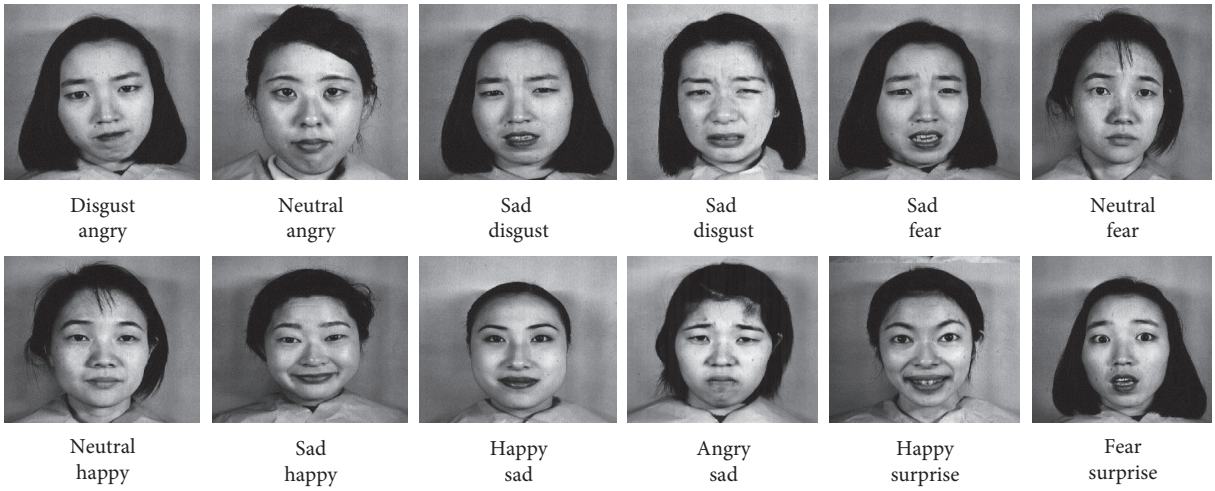

Figure 20: Some misclassified expressions in JAFFE.

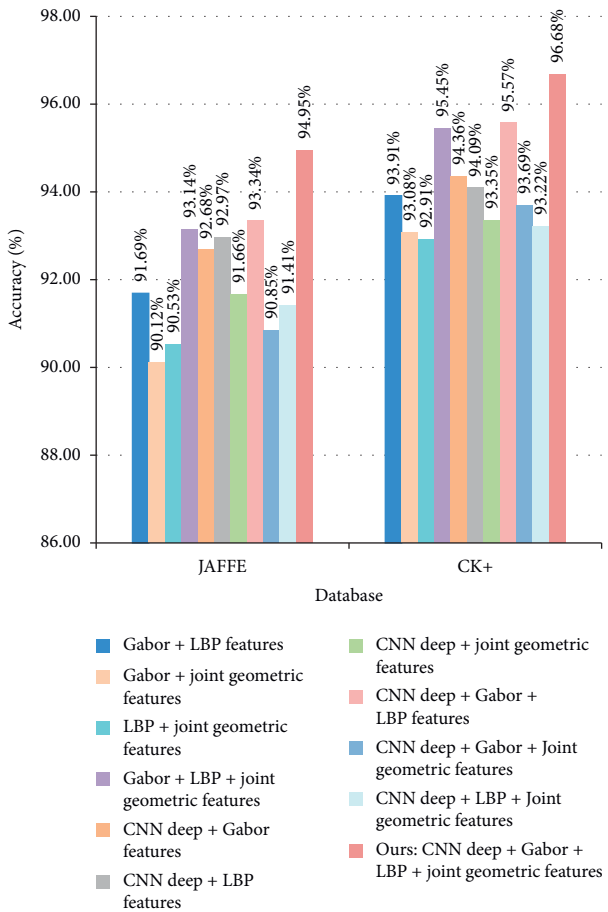

Figure 21: Comparison of our proposed fusion feature and other feature fusion combinations.

expression. In addition, the sample size of the expression database is not sufficient. Even if the two types of expression databases are added, the sample size is still not large. The establishment of a reliable sample balance and an adequate database is an urgent issue.

(3) This study only studies the seven basic expressions of the human face. These seven expressions have obvious characteristics; even then, it is easy to cause confusion between expressions. However, in our real life, we will encounter various expressions and painful and happy mixed expressions and micro-expressions which are not easy to distinguish. The study of these expressions will become a new research direction in the field of expression recognition in the future.

Data Availability

Previously reported [datasets] data were used to support this study and are available at [DOI]. These prior studies (and datasets) are cited at relevant places within the text as references [40–42]. JAFFE (https://zenodo.org/record/3451524#.X54S-egzZPY), CK+ (http://www.pitt.edu/~emotion/ck-spread.htm), and FERPlus (https://www.worldlink.com.cn/osdir/ferplus.html).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the Project of Intelligent Situation Awareness System for Smart Ship (MC-201920-X01).

References

[1] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Information Fusion*, vol. 53, pp. 209–221, 2020.

[2] H.-H. Tsai and Y.-C. Chang, "Facial expression recognition using a combination of multiple facial features and support vector machine," *Soft Computing*, vol. 22, no. 13, pp. 4389–4405, 2018.

[3] A. Lawi and M. S.'R. Machrizzandi, "Facial expression recognition using multiclass ensemble least-square support vector machine," *Journal of Physics Conference Series*, vol. 979, no. 1, Article ID 012032, 2018.

[4] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018.

[5] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[6] L. Ji, Y. Ren, G. Liu et al., "Training-based gradient lbp feature models for multiresolution texture classification," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2683–2696, 2017.

[7] R. Li, J. Tian, and M. C. H. Chua, "Facial expression classification using salient pattern driven integrated geometric and textual features," *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 28971–28983, 2019.

[8] J. Zhou, S. Zhang, H. Mei, and D. Wang, "A method of facial expression recognition based on Gabor and NMF," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 119–124, 2016.

[9] S. Qin, Z. Zhu, Y. Zou, and X. Wang, "Facial expression recognition based on Gabor wavelet transform and 2-channel CNN," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 18, no. 02, p. 2050003, 2019.

[10] F. Kong, "Facial expression recognition method based on deep convolutional neural network combined with improved LBP features," *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 531–539, 2019.

[11] H. Bejaoui, H. Ghazouani, and W. Barhoumi, "Sparse coding-based representation of LBP difference for 3D/4D facial expression recognition," *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 22773–22796, 2019.

[12] S. Shi, H. Si, J. Liu, and Y. Liu, "Facial expression recognition based on Gabor features of salient patches and ACI-LBP," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 4, pp. 2551–2561, 2018.

[13] D. J. Kim, "Facial expression recognition using ASM-based post-processing technique," *Pattern Recognition and Image Analysis*, vol. 26, no. 3, pp. 576–581, 2016.

[14] M. Rahul, N. Kohli, R. Agarwal, and S. Mishra, "Facial expression recognition using geometric features and modified hidden Markov model," *International Journal of Grid and Utility Computing*, vol. 10, no. 5, pp. 488–496, 2019.

[15] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p. 40, 2019.

[16] Y. Quan, Y. Chen, Y. Shao, H. Teng, Y. Xu, and H. Ji, "Image denoising using complex-valued deep CNN," *Pattern Recognition*, vol. 111, p. 107639, 2021.

[17] F. Zhang, T. Zhang, Q. Mao et al., "A unified deep model for joint facial expression recognition, face synthesis, and face alignment," *IEEE Transactions on Image Processing*, vol. 29, pp. 6574–6589, 2020.

[18] X. Sun and M. Lv, "Facial expression recognition based on a hybrid model combining deep and shallow features," *Cognitive Computation*, vol. 11, no. 4, pp. 587–597, 2019.

[19] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018.

[20] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," *IEEE Access*, vol. 7, pp. 18940–18950, 2019.

[21] W. Xie, X. Jia, L. Shen, and M. Yang, "Sparse deep feature learning for facial expression recognition," *Pattern Recognition*, vol. 96, p. 106966, 2019.

[22] H. Li and H. Xu, "Deep reinforcement learning for robust emotional classification in facial expression recognition," *Knowledge-Based Systems*, vol. 204, p. 106172, 2020.

[23] M. Rashid, M. A. Khan, M. Alhaisoni et al., "A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection," *Sustainability*, vol. 12, no. 12, p. 5037, 2020.

[24] J. Ren, Z. Liu, X. Zhou, C. Bai, and G. Sun, "Co-saliency detection via integration of multi-layer convolutional features and inter-image propagation," *Neurocomputing*, vol. 371, no. Jan.2, pp. 137–146, 2020.

[25] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Information Sciences*, vol. 428, pp. 49–61, 2018.

[26] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, 2018.

[27] B. Lin, X. Wei, and Z. Junjie, "Automatic recognition and classification of multi-channel microseismic waveform based on DCNN and SVM," *Computers & Geosciences*, vol. 123, pp. 111–120, 2019.

[28] Y. H. Shao, C. N. Li, L. W. Huang et al., "Joint sample and feature selection via sparse primal and dual LSSVM," *Knowledge-Based Systems*, vol. 185, no. Dec.1, pp. 104915.1–104915.16, 2019.

[29] M. Chaa, Z. Akhtar, and A. Attia, "3D palmprint recognition using unsupervised convolutional deep learning network and SVM classifier," *IET Image Processing*, vol. 13, no. 5, pp. 736–745, 2019.

[30] M. A. Takalkar, M. Xu, and Z. Chaczko, "Manifold feature integration for micro-expression recognition," *Multimedia Systems*, vol. 26, no. 5, pp. 535–551, 2020.

[31] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, and Z. Luo, "Conditional convolution neural network enhanced random forest for facial expression recognition," *Pattern Recognition*, vol. 84, pp. 251–261, 2018.

[32] D. E. Touil, N. Terki, and S. Medouakh, "Hierarchical convolutional features for visual tracking via two combined color spaces with SVM classifier," *Signal, Image and Video Processing*, vol. 13, no. 2, pp. 359–368, 2019.

[33] B. T. Pham, B. Pradhan, D. Tien Bui, I. Prakash, and M. B. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India)," *Environmental Modelling & Software*, vol. 84, pp. 240–250, 2016.

[34] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascade for multi-view face detection with alignment-awareness," *Neurocomputing*, vol. 221, pp. 138–145, 2017.

[35] J. Huang, Y. Shang, and H. Chen, "Improved Viola-Jones face detection algorithm based on HoloLens," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, pp. 1–11, 2019.

[36] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "Face alignment using a 3D deeply-initialized ensemble of regression trees," *Computer Vision and Image Understanding*, vol. 189, Article ID 102846, 2019.

[37] S. Yi, Z. Lai, Z. He, Y.-m. Cheung, and Y. Liu, "Joint sparse principal component analysis," *Pattern Recognition*, vol. 61, pp. 524–536, 2017.

[38] J. Zou, T. Rui, Y. Zhou et al., "Convolutional neural network simplification via feature map pruning," *Computers & Electrical Engineering*, vol. 70, pp. 950–958, Article ID S0045790617326393, 2018.

[39] A. Zitouni, F. Benkouider, F. Chouireb, and M. Belkheiri, "Classification of textured images based on new information fusion methods," *IET Image Processing*, vol. 13, no. 9, pp. 1540–1549, 2019.

[40] V. Tümen, F. Söylemez Ö, and B. Ergen, "Facial emotion recognition on a dataset using convolutional neural network," in *Proceedings of the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5, IEEE, Malatya, Turkey, 2017.

[41] F. Y. Shih, C.-F. Chuang, and P. S. P. Wang, "Performance comparisons of facial expression recognition in jaffe database," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 3, pp. 445–459, 2008.

[42] P. Lucey, J. F. Cohn, T. Kanade et al., "The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the 2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101, IEEE, San Francisco, CA, USA, 2010.