Research Article

Yilihamu Yaermaimaiti, Tusongjiang Kari*, and Guohang Zhuang

Research on facial expression recognition based on an improved fusion algorithm

https://doi.org/10.1515/nleng-2022-0015
received October 19, 2021; accepted March 17, 2022

Abstract: This article puts forward a facial expression recognition (FER) algorithm based on multi-feature fusion and convolutional neural network (CNN) to solve the problem that FER is susceptible to interference factors such as non-uniform illumination, thereby reducing the recognition rate of facial expressions. It starts by extracting the multi-layer representation information (asymmetric region local binary pattern [AR-LBP]) of facial expression images and cascading them to minimize the loss of facial expression texture information. In addition, an improved algorithm called divided local directional pattern (DLDP) is used to extract the original facial expression image features, which not only retains the original texture information but also reduces the time consumption. With a weighted fusion of the features extracted from the above two facial expressions, new AR-LBP-DLDP facial local features are then obtained. Later, CNN is used to extract global features of facial expressions, and the local features of AR-LBP-DLDP obtained by weighted fusion are cascaded and fused with the global features extracted by the CNN, thereby producing the final facial expression features. Ultimately, the final facial expression features are input into Softmax for training and classification. The results show that the proposed algorithm, with good robustness and real-time performance, effectively improves the recognition rate of facial expressions.

Keywords: non-uniform illumination, multi-feature fusion, facial expression, convolutional neural network, robustness

1 Introduction

The human face, with the most prominent and easy-to-use features, has attracted significant attention from researchers. Facial expressions, as a way to reveal the inner world, contain rich emotional information and play a vital role in our social interactions. Facial expression recognition (FER) technology is a basic pattern recognition technology and drives the development of artificial intelligence and human–computer interaction. This technology is widely used in many fields (such as machine image vision, artificial intelligence, and pattern recognition) and is gradually becoming a research hotspot [1].

FER mainly includes facial expression pre-processing, facial expression feature extraction, and facial expression classification. Among these, facial expression feature extraction is the most critical part because its effectiveness plays a decisive role in the accuracy of subsequent FER and classification performance. Therefore, the bottleneck problem is how to effectively extract high-quality facial expression features [2–4]. The main function of feature extraction is to convert the image from pixel-level information to high-level representation, such as histogram of oriented gradients (HOG) of appearance features [5], Gabor wavelet [6], active appearance model (AAM) [7], and deep learning features. The aforementioned algorithms ignore the spatial arrangement information among local features and extract information in different scales and directions through wavelet extraction, so they have better distinguishability for different expressions. However, with the high dimensions and large calculation amounts of their extracted features, these algorithms are difficult to be applied in real-time systems. The AAM can obtain face parameters effectively, but it has some shortcomings, such as difficulty in obtaining initial parameters and time-consuming calculations.

Local binary pattern (LBP) stands out due to its clear principles, low computational complexity, and strong feature descriptive powers. It has been widely used in facial expression feature extraction. Although LBP is simple in calculation and efficient and quick in feature

*** Corresponding author: Tusongjiang Kari,** School of Electrical Engineering, Xinjiang University, Urumqi, Xinjiang 830047, China, e-mail: karlsjtu@163.com, aonangnang188@163.com
Yilihamu Yaermaimaiti, Guohang Zhuang: School of Electrical Engineering, Xinjiang University, Urumqi, Xinjiang 830047, China

extraction, it cannot solve well the problems of poor scalability and continuity of facial expression feature histograms and low effectiveness of discrimination. Therefore, Naika *et al.* [8] proposed an asymmetric region local binary pattern (AR-LBP) algorithm based on the traditional LBP algorithm. The AR-LBP algorithm reduced the loss of facial expression texture information features and strengthened the recognition ability of facial expression features. Jabid *et al.* [9] put forward a local directional mode (LDP), which better solved the influence of illumination changes and random noise on the target image by introducing the Kirsch operator. FER methods based on traditional algorithms have weak real-time feature extraction and poor pertinence. So, Luo *et al.* [10] proposed an improved feature extraction algorithm called divided local directional pattern (DLDP). The DLDP algorithm further shortened the time of FER by reducing the feature size of facial expressions and improved the real-time performance of FER. He and Chen [11] proposed an expression recognition algorithm based on improved LBP and HOSVD using k-nearest neighbors for classification and the similarity between images for secondary classification. Fekri-Ershad [12] proposed that an improved rotation-invariant method of LBPs is used for gender classification, first extracting facial feature vectors using the ILBP algorithm and then classifying gender using the Kullback–Leibler divergence classifier. This method shows good results when classifying gender, but in the case of multiple labels of expressions, it is still necessary to further extract more facial details.

Convolutional neural network (CNN) has been well developed and applied in artificial intelligence, pattern recognition, and other related fields, especially facial recognition [13]. As the number of network layers increases, CNN can better learn to recognize non-specific features of things [14]. However, the texture information of the face cannot be obtained by using CNN deficiency. Therefore, for those image features obtained through human operations, the facial features automatically obtained by CNN can better reflect the changing trend of facial expressions, thereby improving the recognition rate of facial expressions [15–17]. In image fusion, Zhu *et al.* [18] proposed a novel image fusion scheme based on image cartoon-texture decomposition and sparse representation in order to fuse complementary information from multimodal images into images. Meanwhile, in order to reduce the interference of external factors on the image, Zhu *et al.* [19] proposed an image fusion-based algorithm to enhance the performance and robustness of image defogging. Ravi *et al.* [20] proposed a method combining CNN and LBP for face recognition – first using LBP and CNN for feature extraction and then using support vector machine (SVM) to classify the extracted features. Still, there is a traditional LBP that cannot process faces. The scalability of the expression feature histogram, the length of the FER, and the complex structure of the CNN network can easily lead to problems such as slow recognition speed.

Given the aforementioned factors, this article proposes a new FER algorithm based on multi-feature fusion and CNN. AR-LBP can well solve the problems of traditional LBP, while DLDP is found to reduce the computing time well when extracting features, so this article uses AR-LBP- and DLDP-extracted features for weighted fusion. Because CNN networks are found to achieve good results in extracting facial features, this article uses a 5×5 convolutional kernel to extract large regions of face expression information, which can effectively extract the global information of facial expressions, and finally, the features extracted by CNN and the features fused by AR-LBP and DLDP are used for Softmax classification. The contributions made by each section are as follows:

(1) AR-LBP: AR-LBP can deal well with the scalability problem of facial expression histogram and can effectively reduce the loss of expression texture information.
(2) DLDP: DLDP can reduce the influence of illumination and expression changes, further shorten the time of FER, and further improve the real-time performance of FER.
(3) CNN: CNN network is also a powerful feature extractor that can effectively extract global information.

It can be specifically divided into six steps. First, AR-LBP is used to extract the multi-layer feature information domain of facial expressions. Second, the improved algorithm DLDP is used to extract facial expression features. Third, through the weighted fusion of AR-LBP and DLDP, the local features of facial expressions are obtained. Fourth, CNN is used to extract global features of facial expressions. Fifth, a cascade is conducted on local and global features of facial expressions. Finally, the final features of the cascaded facial expressions are inputted to Softmax for training and classification.

2 AR-LBP feature description

The traditional local feature extraction method is LBP. The definition of LBP is in a 3×3 window. The center of the window is used as the threshold, and the pixel value of the center point is compared with the adjacent eight pixel values. If the size of the adjacent pixel value is larger than the center point pixel value, the point is marked as 1; otherwise, it is 0. LBP cannot deal with the scalability of the facial expression feature histogram and the length of FER. Naika *et al.* [8] proposed an asymmetric local binary model AR-LBP based on traditional LBP. The feature extraction method between LBP and AR-LBP is shown in Figure 1.

AR-LBP can retain the basic texture information features of facial expressions when extracting facial expression features without losing a large amount of texture information, thereby enhancing the ability to discriminate facial expression features. The AR-LBP operator selects a larger-scale neighborhood with a detailed size of $(2m + 1) \times (2n + 1)$. In $1 \le m \le \left\lfloor \frac{h-1}{2} \right\rfloor$, $1 \le n \le \left\lfloor \frac{\omega-1}{2} \right\rfloor$, h represents the height, ω is the width, and $\lfloor\ \rfloor$ means rounding down; the neighborhood can be divided into nine sub-neighborhoods, each denoted as R_i; and the average value of pixels in the sub-neighborhoods is selected as the pixel value of the sub-neighborhood, which is denoted as X_i, $i = 1, 2, \ldots, 8$. The calculation formula is given in the following equation:

$$X_i = \frac{1}{n_i}\sum_{j=1}^{n_i} p_{ij}. \tag{1}$$

In formula (1), n_i represents the number of pixels in sub-neighborhood no. i, and p_{ij} represents the jth pixel value in R_i.

The formula for calculating the AR-LBP value is shown in formula (2). The center pixel is defined as the threshold value, and the facial expression gray values of the adjacent eight pixels are compared with it.

$$\text{AR-LBP}(x, y) = \sum_{i=0}^{N-1} s(X_i - X_c)\cdot 2^i. \tag{2}$$

If the center pixel value of the surrounding sub-neighborhood is less than its pixel value, the pixel is marked as 1; otherwise, it is 0. The formula is shown as follows:

$$s(x) = \begin{cases} 1, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{3}$$

In Eq. (2), AR-LBP(x, y) stands for the facial expression feature values of (x, y); X_c represents the pixel value of the center point (x, y), and X_i is that of the sub-neighborhood R_i. Finally, the AR-LBP histogram of the face image is counted. In order to retain most of the local information features of facial expressions, the facial expression image is first divided into many independent sub-blocks of the same size. Then, the AR-LBP histogram of each sub-block is counted according to formulas (4) and (5). Finally, the feature histograms of all AR-LBP sub-blocks are cascaded as the feature value of the final facial expression image. The tth calculation formula of the AR-LBP histogram of the first sub-block is as follows:

$$H_t(i) = \sum_{x,y} I\{\text{AR-LBP}(x, y) = i\}, \quad i = 0, 1, \ldots, 2^N - 1, \tag{4}$$

$$I\{A\} = \begin{cases} 1, & A = \text{true}, \\ 0, & A = \text{false}. \end{cases} \tag{5}$$

Then we cascade the histograms of all sub-blocks to generate an H histogram of the entire facial expression image, the formula of which is shown as follows:

$$H = \{H_1, H_2, \cdots, H_t, \cdots, H_k\}. \tag{6}$$

where H_t represents the histogram of the tth sub-block, and k represents the number of sub-blocks.

AR-LBP is able to extract local information of faces and can achieve a better one than the traditional LBP operator in processing local information; therefore, AR-LBP

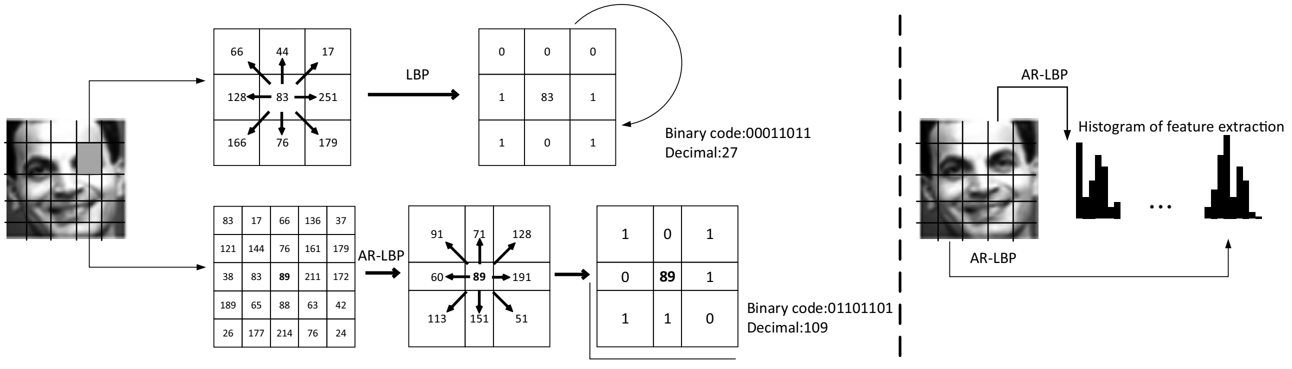

Figure 1: Feature extraction for LBP and AR-LBP.

is introduced in this article to improve the local information representation of facial expressions.

3 DLDP algorithm overview

According to the traditional LDP, it is necessary to calculate the edge response values in eight directions, take the absolute value, and then perform the sorting process. This method greatly increases the time consumption of facial expression feature extraction. In response to this problem, Luo *et al.* proposed an improved DLDP feature extraction algorithm, which could retain the facial expression texture feature information of the original LDP algorithm, reduce the running time, and improve the real-time performance.

The DLDP operator divides the eight direction templates of the Kirsch operator into two independent sub-direction templates according to four corners and the four directions of up, down, left, and right. The kirsch mask shown in Figure 2 performs a convolution operation on the 3 × 3 neighborhood image block where each pixel in the image is located and calculates the response value of each pixel in eight directions. In addition, it uses the random pixel point X of the facial expression image as the center point. At the same time, two independent sub-templates are used to calculate the edge response values of the four corners and the four directions of up, down, left, and right to obtain two four-direction edge response values, namely, m_{10}, m_{11}, m_{12}, m_{13} and m_{20}, m_{21}, m_{22}, m_{23}. Then, the absolute value is taken from these eight edge response values, and a sorting operation is performed, followed by setting the top three edge response value to 1 and the others to 0. Finally, coding DLDP1 and DLDP2 are obtained.

According to the symbols of the facial expression image in the four directions, the symbols are processed by the binary coding mode to construct the partial descriptor information of the facial expression image. The coding definition formulas are given as:

$$\text{DLDP1} = \sum_{i=0}^{3} \phi(m_{1i} - m_k)\cdot 2^i, \tag{7}$$

$$\text{DLDP2} = \sum_{i=0}^{3} \phi(m_{2i} - m_k)\cdot 2^i, \tag{8}$$

$$\phi(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases} \tag{9}$$

In the above formulas, m_{1i} and m_{2i} are edge response values, and m_k is the kth largest among all edge response values. The DLDP codes of the pixels are arranged according to the coordinates of the original map to form a corresponding DLDP code map. After calculating the DLDP code of each pixel (x, y), the $H_{\text{DLDP}i}$ histogram as a DLDP descriptor is defined as follows (10):

$$H_{\text{DLDP}i} = \sum_{x,y} f(\text{DLDP}i(x, y), R). \tag{10}$$

In the above formula

$$f(a, x) = \begin{cases} 1, & a = x, \\ 0, & \text{other}, \end{cases}$$

where R represents the DLDPi code value in the sub-direction, $i = 1, 2$. The DLDP histogram (H_{DLDP}) can be obtained by cascading the DLDP1 histogram and the DLDP2 histogram, as shown in formula (11).

-3	-3	5
-3	0	5
-3	-3	5

m_{10}

-3	5	5
-3	0	5
-3	-3	-3

m_{11}

5	5	5
-3	0	-3
-3	-3	-3

m_{12}

5	5	5
5	0	-3
-3	-3	-3

m_{13}

5	-3	-3
5	0	-3
5	-3	-3

m_{20}

-3	-3	-3
5	0	-3
5	5	-3

m_{21}

-3	-3	-3
-3	0	-3
5	5	5

m_{22}

-3	-3	-3
-3	0	5
-3	5	5

m_{23}

Figure 2: Kirsch mask.

$$H_{\mathrm{DLDP}} = \sum_{i=1}^{2} H_{\mathrm{DLDP}i}. \tag{11}$$

The DLDP descriptor provides detailed information about facial expression texture features. The histograms of DLDP1 and DLDP2 can be obtained by formula (10), and then, the histograms of DLDP1 and DLDP2 are connected in formula (11) to obtain the DLDP histogram (H_{DLDP}) as the final facial expression image feature information.

The DLDP algorithm calculates the eight directional masks of the Kirsch operator separately. The article performs local feature weighting fusion by extracting features for faces with DLDP and AR-LBP, *i.e.*, it preserves the texture information and gradient information of the extracted facial features and also improves the overall computational speed.

4 Global feature extraction from CNNs

As one of the deep learning algorithms, CNN is a feed-forward neural network whose convolution operation is completed by the convolution kernel and has a deep structure. It is widely used in various fields, especially in the field of image recognition.

CNNs generally include convolutional layer, pooling layer, and full connection layer. The convolution layer mainly extracts the local features of the image. The pooling layer is mainly to compress the relevant image features extracted from the previous layer, namely, the convolution layer, to obtain new and smaller image features, so that the main image features can be retained. The full connection layer is an effective method to extract global features by cascading and fusing all local features extracted from the convolution layer and the pooling layer to obtain global features. Finally, softmax is used to achieve the classification effect. The CNN network frame diagram is shown in Figure 3.

Since the convolutional neural network is superimposed too deeply, it is easy to cause the gradient to disappear, resulting in a decline in performance. The CNN model in this article is composed of three parts, each part is composed of a convolutional level and a pooling layer and finally connected with a full connection layer and a Softmax layer. First of all, 48 × 48 grayscale facial expression images are input and convolution operation of the first convolution layer is performed with ten convolution checks. Then, the output of the first pooling layer is convolved with 20 convolution checks. Second, the output of the second pooling layer is convolved with 40 convolution checks. Finally, expand it into a fully connected structure. Among them, the size of convolution kernel is 5 × 5, the pooling layer adopts 2 × 2 maximum pooling, the number of neurons is 100, and the expression classification is carried out in the softmax layer.

Generally, the calculation formula of the convolution layer is as follows Eq. (12):

$$C_j = f\left(\sum_{i=1}^{N} M_i * L_{i,j} + p_j\right). \tag{12}$$

where M_i is the input matrix, $L_{i,j}$ is the value of the convolution kernel, p_j represents the offset term, C_j is the output matrix, $f(\cdot)$ is the activation function, RELU is used as the activation function in this article, and its definition is as follows Eq. (13):

$$f(x) = \max(0, x). \tag{13}$$

In this article, we design a CNN network that can be trained end-to-end. In order to be able to extract global information, a convolutional kernel of size 5 × 5 is used. Compared to CNNs with large parameters, we design a network structure that is easy to extract global information, while the training time is short and more suitable for the data of the experiments in this article.

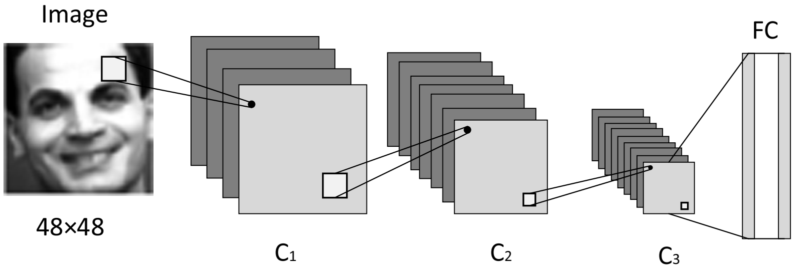

Figure 3: CNN network.

5 Overview of the fused FER algorithm

The overall process of the fusion FER algorithm in this article is shown in Figure 4.

This article starts by using the AR-LBP algorithm to extract the information domain and the improved algorithm called DLDP to extract facial expression features and then performs a weighted fusion operation on the facial expression features extracted by the algorithms AR-LBP and DLDP. The facial expression features obtained by this weighted fusion are taken as the local features of the facial expressions. Later, the global facial expressions are extracted through the CNN, and the proposed local features are cascaded with the global features. Ultimately, the final features of the cascaded facial expressions are input into the CNN for training and classification. The algorithm process is as follows:

(1) obtain a 48 × 48 grayscale image set of the original facial expression image through pre-processing operation;
(2) extract and cascade multi-layer AR-LBP facial expression features of facial expression images to reduce the loss of facial expression and texture information and enhance the capability of discriminating facial expression features;
(3) extract the original facial expression image features using the DLDP algorithm, which not only retains the texture information of the original LDP but also reduces time consumption through algorithm improvements; and
(4) perform the weighted fusion of the AR-LBP facial expression features obtained in steps (2) and (3) with the DLDP facial expression image features to get the local feature expression of the facial expression image. The formula is as follows (14):

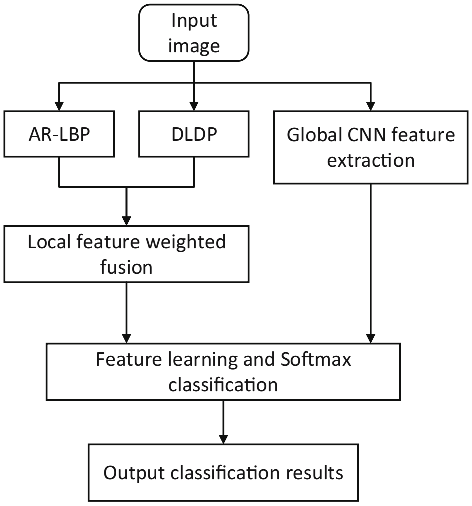
Input image
AR-LBP
DLDP
Global CNN feature extraction
Local feature weighted fusion
Feature learning and Softmax classification
Output classification results

Figure 4: Flow chart of FER algorithm.

$$f_w = \alpha \cdot f_x + (1 - \alpha)\cdot f_y, \tag{14}$$

where f_w represents the local feature expression of the facial expression image after weighted fusion, and f_x is labeled as the facial expression feature. α represents the weighting coefficient of the AR-LBP feature, and f_y is the DLDP facial expression image feature. Also, $(1 - \alpha)$ refers to the weighting coefficient of the DLDP feature. In order to make AR-LBP and DLDP obtain the same contribution in fusion, set α to 0.5:

(1) Perform a normalization operation on the local features f_w obtained by weighted fusion and the output global features of the CNN fully connected layer f_z and then perform a series of cascade operations on the two sets of features to obtain the final facial expression feature vector f. The formula is as follows (15):

$$f = (f_w, f_z). \tag{15}$$

(2) Input the final fusion feature f into softmax for training and classification.

Previous algorithms often cannot extract face information well and capture texture and gradient information of faces. In this article, in order to improve feature extraction of face information, AR-LBP and DLDP algorithms are introduced to extract features and weighted fusion, while large convolutional kernels are used to capture global information of faces through CNNs, and finally, the obtained information is spliced in the fully connected layer of neural networks for classification. The performance of the dataset can reflect the effectiveness and real-time performance of the proposed algorithm.

6 Simulation experiment

6.1 Facial expression database

In this article, the extended Cohn–Kanade dataset (CK+) and Facial Expression Recognition 2013 Dataset (FER2013) are used to carry out FER simulation experiments. The two datasets are shown in Figure 5.

The CK+ database comes from 123 individuals with a total of 593 personal facial expression image sequences, among which 327 personal facial expression image data

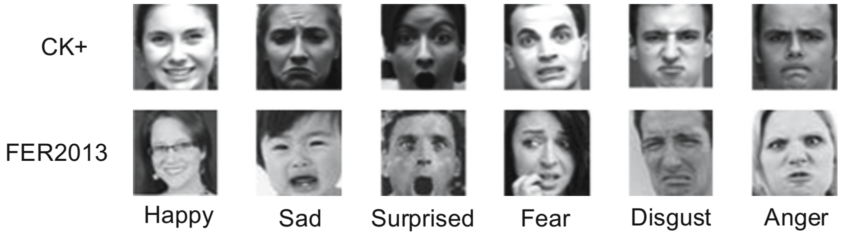

Figure 5: Example of the CK+ dataset and FER2013 dataset.

bear marks. This article uses 1,356 facial expression images, including 6 different expressions of anger, disgust, fear, happiness, sadness, and surprise. The training set uses 1,256 peak facial expression images, whereas the test set uses 100 facial expression images.

The FER2013 database stems from the facial expression image data set in the 2013 Kaggle competition. It contains 35,000 facial expression images, and each picture is grayscale standardized to 48 × 48 pixels. FER2013 contains seven different facial expressions: neutral expression, anger, disgust, fear, happiness, grievance, and surprise. In order to verify the effectiveness of the algorithm, we eliminated the neutral expressions in the FER2013 dataset and tested the remaining six facial expressions. Among them, 28,000 frames are used as the training set and 3,000 frames are used as the test set. However, the FER2013 dataset is not balanced. The images refer to various races, ages, lighting, occlusion, side faces, postures, *etc*. Also, the images are grayscale ones with a size of 48 × 48 pixels. Besides, there are problems such as missing labels and label errors. All these make it a challenge to conduct FER in the FER2013 dataset.

6.2 Training method

The simulated experimental environment is an Inteli7 1.80 GHz CPU, 1660ti graphics card, 8 GB memory, Win10, and MatlabR2016a software. This article uses the FER rate as an index to evaluate the effect of FER, which is defined as follows:

$$\mathrm{RR} = \frac{\text{Number of correctly identified samples}}{\text{Number of test samples}} \times 100\%.$$

During our training process, the input image size is 48 × 48, the step size is 1, and the size 5 × 5 convolution kernel is used to convolve the facial expression image. Compared with other larger convolution kernels, the 5 × 5 convolution kernel not only reduces the standard parameters of the convolution layer but also saves the size and space of facial expression images or expression feature maps, thereby promoting the classification accuracy of the network. Simultaneously, when using the dropout method to propagate forward, we randomly stop some hidden neurons according to a given probability p, that is, set the input and output of these neurons to 0. These hidden neurons do not participate in the forward propagation and backpropagation of errors so that the entire network will not rely too much on some local features. In training the network, the dropout was set to 0.5. The Adam algorithm was used to optimize the network. The initial learning rate is 0.0001. At the same time, the number of iterations of datasets was set to 100.

6.3 Analysis of simulation experiment results

We experimented to compare the proposed algorithm with the four algorithms, including CNN, SIFT + CNN [21], LDP + DWT + Sobel + CNN [22], and LBP + DCT + CNN [23]. The results are shown in Tables 1 and 2.

It can be seen from the conclusions of Tables 1 and 2 that the algorithm proposed in this article exerts the highest recognition rate of facial expressions. In order to compare more intuitively the recognition rates of the comparison algorithm and the proposed algorithm in each expression category, the results are visualized in this article, as shown in Figures 6 and 7. It can be seen from the figure that the proposed algorithm is still relatively accurate in recognizing six types of expressions, especially the happy expressions, which are able to reach 100% accuracy in both datasets. In the CK+ facial expression database and the FER2013 facial expression database, the average recognition rates of facial expressions obtained by this algorithm are 94.7 and 91%, respectively. Also, the algorithm can adaptively weigh the difference in the amount of information of facial expression images. The weighted fusion of AR-LBP facial expression features and DLDP facial expression image features not

Table 1: Result comparison of different algorithms in the CK+ dataset, where Ci means correct identification, ACC means recognition rate

Expression	CNN		SIFT + CNN		LBP + DCT + CNN		LDP + DWT + Sobel + CNN		Proposed	
	Ci (times)	ACC (%)	Ci (times)	ACC (%)	Ci (times)	ACC (%)	Ci (times)	ACC (%)	Ci (times)	ACC (%)
Happy	45	92	47	94	48	96	49	98	50	100
Sad	43	87	44	90	45	92	46	98	47	95
Surprised	44	90	46	92	47	95	48	96	49	97
Fear	41	82	45	92	45	92	46	92	47	92
Disgust	39	79	47	91	41	86	43	87	41	88
Anger	44	90	46	93	47	95	48	96	48	96

Table 2: Result comparison of different algorithms in the FER2013 dataset, where Ci means correct identification and ACC means recognition rate

Expression	CNN		SIFT + CNN		LBP + DCT + CNN		LDP + DWT + Sobel + CNN		Proposed	
	Ci (times)	ACC (%)	Ci (times)	ACC (%)	Ci (times)	ACC (%)	Ci (times)	ACC (%)	Ci (times)	ACC (%)
Happy	44	89	45	92	46	90	47	93	50	100
Sad	41	85	41	85	42	87	44	89	46	90
Surprised	41	85	44	89	45	91	46	92	47	93
Fear	39	80	43	88	43	88	44	89	46	90
Disgust	36	76	45	91	39	80	41	85	40	80
Anger	42	87	44	89	44	89	46	92	47	93

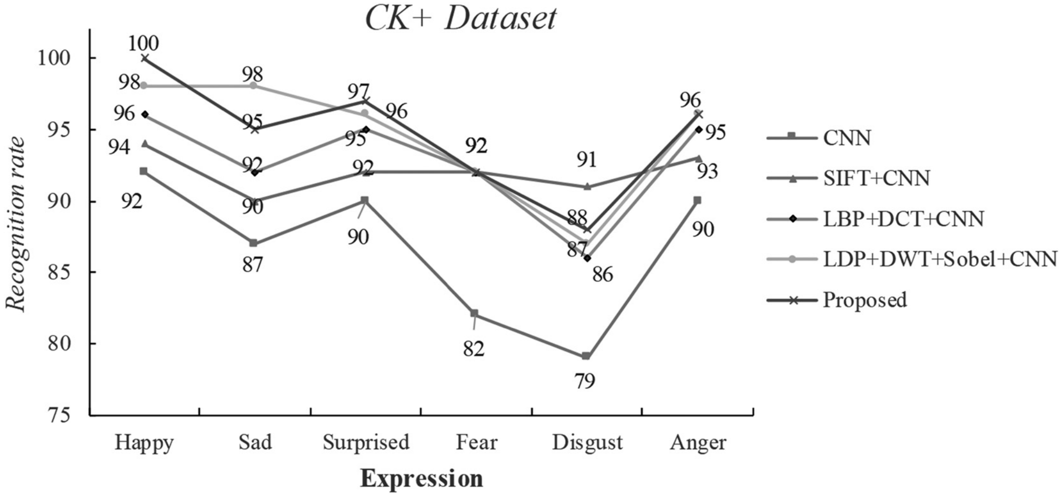

Figure 6: Result comparison of different algorithms in the CK+ dataset.

only highlights areas with rich facial expression information but also improves the effectiveness of local features. Besides, the algorithm integrates the global features of the CNN of facial expression images, which complements the facial expression features from another angle. After cascading the obtained local facial expression features and the global facial expression features obtained by convolution, the final feature vector of the facial expression is obtained.

To deal with the low recognition rate of individual facial expressions and to show the classification results more intuitively, we made the expression distribution obtained by multiple experiments into a confusion matrix, as shown in Tables 3 and 4.

According to Tables 3 and 4, the two expressions with the highest recognition rate of facial expressions are happiness and surprise. Because compared with other expressions, these two are easier to recognize and classify for

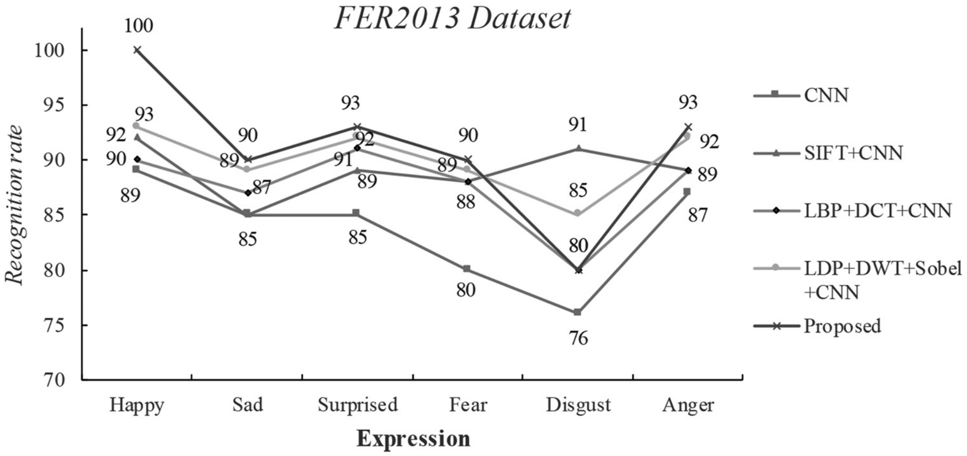

Figure 7: Result comparison of different algorithms in the FER2013 dataset.

Table 3: Confusion matrix of the CK+ dataset

Expression	Happy	Sad	Surprised	Fear	Disgust	Anger
Happy	97.97	0.19	0.71	0.45	0.29	0.39
Sad	0.12	95.98	0.33	0.24	0.16	3.17
Surprised	0.25	0.27	98.08	1.11	0.29	0
Fear	0	0.31	0.12	94.88	3.07	1.62
Disgust	0.16	0.85	0.55	1.81	94.74	1.89
Anger	0	0.49	0.08	0.85	0.47	98.11

Table 4: Confusion matrix of the FER2013 dataset

Expression	Happy	Sad	Surprised	Fear	Disgust	Anger
Happy	92.35	1.71	2.85	1.11	0	1.98
Sad	4.60	78.45	3.41	8.05	2.51	2.98
Surprised	5.85	1.40	87.31	3.39	0.58	1.47
Fear	3.92	12.58	10.94	60.15	2.38	10.03
Disgust	3.22	0	2.71	4.52	77.48	12.07
Anger	2.54	8.96	2.24	5.82	1.92	78.52

their larger motion range, while the other four facial expressions are more difficult to recognize. The expressions of fear and sadness are similar because they both are featured by the wrinkled forehead and opened lips. Anger and disgust are similar in facial features, such as frowning and mouth corners, which can easily cause confusion in recognition. It can be seen from Table 3 that the proposed algorithm performs well in the CK+ dataset, and it exerts the highest recognition rate among the three facial expressions: happiness, anger, and surprise. However, it is not perfect in identifying expressions of disgust, fear, and sadness, which is mainly due to the similarity shared by the expressions of disgust, fear, and sadness. This similarity, to some extent, adds a challenge to the feature extraction process of these three different expressions, thereby leading to a decrease in the recognition rate during detection.

Meanwhile, we found through experimental simulations that the recognition effect of the same type on the CK+ dataset is better than that of the FER2013 dataset. This is due to the fact that the facial expression images in the FER2013 dataset are closer to the real-life situation, thereby producing more interference factors. The CK+ dataset is composed of standard laboratory pictures, which have relatively few interferences and influencing factors. In addition, the sample quality of the CK+ facial expression database is superior to that of the FER2013 facial expression database, so the recognition rate of the CK+ facial expression database is much higher than that of FER2013.

This article also conducts an experiment on the accuracy curve of iteration times for training and testing two datasets of CK+ and FER2013, as shown in Figures 8 and 9.

Figure 8 shows that on the CK+ dataset when the model is iterated to the 50th time, its accuracy curves

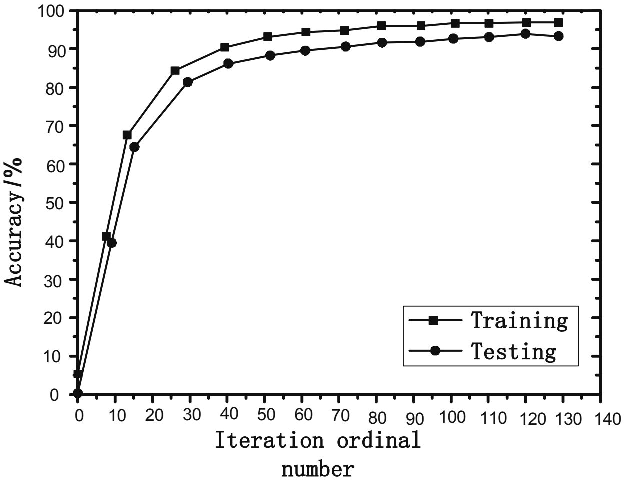

Figure 8: Accuracy curve of training and testing on the CK+ dataset.

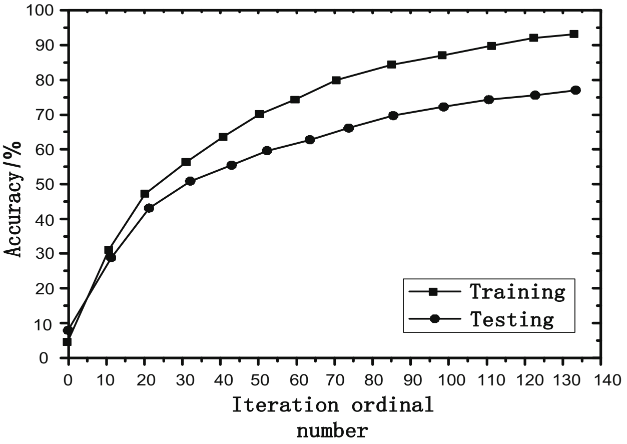

Figure 9: Accuracy curve of training and testing on the FER2013 dataset.

of training and testing converge and tend to be stable. On the FER2013 dataset in Figure 9, the model gradually stabilizes after the 90th iteration. This is mainly because the facial expression images on the CK+ dataset are taken under laboratory conditions and bear fewer external interference factors, so they are more stable than the FER2013 dataset. As a result, the CK+ dataset shows a higher recognition rate with its accuracy curves of training and testing converging faster. At the same time, it can be seen from Figure 9 that the difference between the accuracy of training on the FER2013 dataset and the recognition accuracy in the test is greater than the difference on the CK+ dataset. The recognition accuracy rate of the FER2013 dataset is relatively low. The main reason for this result is that the images in the FER2013 dataset are closer to the facial expression images in real life, with more influencing factors, lower resolution, and problems such as missing tags and tag errors. The sample quality of the FER2013 dataset is not as good as that of the CK+ dataset, which reflects that the facial expression database in the experiment has a significant influence on the accuracy of the model.

Finally, a time test is carried out in order to verify the superiority of the algorithm proposed in this article in terms of real-time performance. The test uses CK+ facial expression database as the test face database and selects ten facial expression images of each person as training samples and other facial expression images as test samples. The time test results are shown in Table 5.

Table 5 reflects that the algorithm proposed in this article shows the optimal performance regardless of the average training time or the average recognition time. Although the proposed algorithm is slightly inferior to LDP + DWT + Sobel + CNN in terms of time running, it exerts the highest average recognition rate. In summary, the proposed algorithm performs well in terms of both recognition rate and real-time capability.

7 Conclusion

This article puts forward an FER algorithm based on multi-feature fusion and CNN to solve the problem that FER is susceptible to interference factors such as non-uniform illumination. It first extracts the multi-layer representation information (AR-LBP) of facial expression images

Table 5: Time test results of different algorithms

Algorithm	Average recognition rate (%)	Average training time (ms)	Average recognition time (ms)
CNN	86.67	2,856	2,456
SIFT + CNN	92.00	2,655	2,355
LBP + DCT + CNN	92.67	2,597	2,297
LDP + DWT + Sobel + CNN	94.50	2,387	2,087
Proposed	94.67	2,445	2,145

and concatenates them, which can effectively extract the texture information of the original picture, and uses an improved algorithm called DLDP to extract the original image. The facial expression image feature method can effectively reduce the influence of illumination and expression changes. With the weighted fusion of the features extracted from the above two facial expressions, new AR-LBP-DLDP facial local features are then obtained. Later, CNN is used to extract global features of facial expressions, and the local features of AR-LBP-DLDP obtained by weighted fusion are cascaded and fused with the global features extracted by the CNN, thereby producing the final facial expression features. Ultimately, the final facial expression features are input into Softmax for training and classification. Improve the robustness of features to changes in lighting, pose, expression, occlusion, *etc*. The results on datasets CK+ and FER2013 show that the algorithm has good robustness and real-time performance.

Acknowledgements: The authors acknowledge the financial support from the National Natural Science Foundation of China (Project Nos. 61866037 and 61462082).

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors in this article declare that they have no conflict of interests.

References

[1] Liao CT, Chuang HJ, Duan CH, Lai SH. Learning spatial weighting for facial expression analysis via constrained quadratic programming. Pattern Recognit. 2013;46(11):3103–16.

[2] Zavaschi THH, Britto AS, Oliveira LES, Koerich AL. Fusion of feature sets and classifiers for facial expression recognition. Expert Syst Appl. 2013;40(2):646–55.

[3] Hu B, Wang J. 3D facial expression recognition method based on bimodal and semantic knowledge. Yi Qi Yi Biao Xue Bao/Chinese J Sci Instrum. 2013;34(4):873–80.

[4] Sandbach G, Zafeiriou S, Pantic M, Yin L. Static and dynamic 3D facial expression recognition: A comprehensive survey. Image Vis Comput. 2012;30(10):683–97.

[5] Chen JH, Takiguchi T, Ariki Y. Rotation-reversal invariant HOG cascade for facial expression recognition. Signal Image Video Process. 2017:11(1–3):1–8.

[6] Hegde G, Seetha M, Hegde N. Facial expression recognition using entire gabor filter matching score level fusion approach based on subspace methods. Microelectron Reliab. 2015;52(3):497–502.

[7] Zhou H, Lam KM, He X. Shape-appearance correlated active appearance model. Pattern Recognit. 2016;56(C):88–99.

[8] Naika CLS, Das PK, Nair SB. Asymmetric region local binary pattern operator for person-dependent facial expression recognition. IEEE International Conference on Computing, Communication and Applications; 2012 Feb 22–24; Dindigul, India. IEEE; 2012. p. 1–5.

[9] Jabid T, Kabir MH, Chae O. Robust facial expression recognition based on local directional pattern. ETRI J. 2010;32(5):784–94.

[10] Luo Y, Yu CJ, Zhang Y, Liu L. Facial expression recognition algorithm based on improved local direction pattern. J Chongqing Univ. 2019;42(3):85–91.

[11] He Y, Chen S. Person-independent facial expression recognition based on improved local binary pattern and higher-order singular value decomposition. IEEE Access. 2020;10(8):190184–93.

[12] Fekri-Ershad S. Gender classification in human face images for smart phone applications based on local texture information and evaluated Kullback-Leibler divergence. Traitement du Signal. 2019;36(6):507–14.

[13] Li H, Lin Z, Shen X, Bradt J, Hua G. A convolutional neural network cascade for face detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, USA. IEEE; 2015. p. 5325–34.

[14] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.

[15] Xu L, Zhao HT, Sun SY. Monocular infrared image depth estimation based on deep convolutional neural networks. Acta Optica Sin. 2016;36(7):196–205.

[16] Liu YZ, Jiang ZQ, Ma F, Zhang CH. Hyperspectral image classification based on hypergraph and convolutional neural network. Laser Optoelectron Prog. 2019;56(11):162–9.

[17] Ou P, Zhang Z, Lu K, Liu ZY. Object detection of remote sensing images based on convolutional neural networks. Laser Optoelectron Prog. 2019;56(5):74–80.

[18] Zhu Z, Yin H, Chai Y, Li Y., Qi G. A novel multi-modality image fusion method based on image decomposition and sparse representation. Inf Sci. 2018;432(3):516–29.

[19] Zhu Y, Zhu B, Liu HHT, Qin K. A novel fast single image dehazing algorithm based on artificial multiexposure image fusion. IEEE Trans Instrum Meas. 2020;70(9):1–23.

[20] Ravi R, Yadhukrishna SV, Prithviraj R. A face expression recognition using CNN & LBP. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC); 2020 Mar 11–13; Erode, India. IEEE; 2020. p. 684–9.

[21] Zhang YQ, He N, Wei RC. Face expression recognition based on convolutional neural network fusing SIFT features. Computer Appl Softw. 2019;36(11):161–7.

[22] Yu M, An MT, Liu Y. Facial expression recognition based on multiple features and convolutional neural networks. Sci Technol Eng. 2018;18(13):104–10.

[23] Wang JX, Lei ZC. A convolutional neural network based on feature fusion for face recognition. Laser Optoelectron Prog. 2020;57(10):339–45.