The Visual Computer (2020) 36:391–404
https://doi.org/10.1007/s00371-019-01627-4

ORIGINAL ARTICLE

Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy

Kuan Li[1] · Yi Jin[1] · Muhammad Waqar Akram[1] · Ruize Han[2] · Jiongwei Chen[1]

Published online: 10 January 2019

Abstract
With the recent development and application of human–computer interaction systems, facial expression recognition (FER) has become a popular research area. The recognition of facial expression is a difficult problem for existing machine learning and deep learning models because that the images can vary in brightness, background, pose, etc. Deep learning methods also require the support of big data. It does not perform well when the database is small. Feature extraction is very important for FER, even a simple algorithm can be very effective if the extracted features are sufficient to be separable. However, deep learning methods automatically extract features so that some useless features can interfere with useful features. For these reasons, FER is still a challenging problem in computer vision. In this paper, with the aim of coping with few data and extracting only useful features from image, we propose new face cropping and rotation strategies and simplification of the convolutional neural network (CNN) to make data more abundant and only useful facial features can be extracted. Experiments to evaluate the proposed method were performed on the CK+ and JAFFE databases. High average recognition accuracies of 97.38% and 97.18% were obtained for 7-class experiments on the CK+ and JAFFE databases, respectively. A study of the impact of each proposed data processing method and CNN simplification is also presented. The proposed method is competitive with existing methods in terms of training time, testing time, and recognition accuracy.

Keywords Face cropping · Facial expression recognition · Convolutional neural network · Computer vision

1 Introduction

Facial expressions are one of the most important features to reflect the human emotional state because they convey useful information to the observer [6]. Facial expressions convey 55% of a communicated message, which is more than the part conveyed by the combination of voice and language [22]. Facial expressions can be divided into six basic categories [3], namely anger, disgust, fear, happiness, sadness, and surprise. With the development of human–computer interaction systems, such as social robots, visual-interactive games, and data-driven animation, facial expression recognition (FER) has become a popular field of study in recent years.

Machine learning plays an increasingly significant role in this field. Several methods have been proposed for FER in recent years, particularly using deep learning approaches [13,14,25,32]. Deep learning methods perform well in FER [13,35]. Facial expression recognition methods can be classified into two main categories: those based on an image sequence [23,26,27,34] and those based on static images [5,39]. In the methods based on an image sequence, the sequence changes from a neutral expression to a peak expression, and these two expressions from the same person form a contrast that makes it easier to extract the features of each expression. Static-image-based methods distinguish facial expressions by analysing the peak expression image without temporal information.

Facial expression extraction is an important part of FER. Facial changes caused by different facial expressions are typically extracted using appearance-based methods [2,28,29,38] or geometry-based methods [9,34]. Appearance-based features describe the texture of the face resulting from an

✉ Yi Jin
jinyi08@ustc.edu.cn

[1] Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, 96 Jinzhai road, Baohe District, Hefei 230026, Anhui, People's Republic of China

[2] School of Computer Science and Technology, Tianjin University, 135 Yaguan road, Jinnan District, Tianjin 300350, People's Republic of China

expression, such as wrinkles. In appearance-based FER, facial features are extracted by applying image filters, such as the Gabor wavelets filter [10], local binary patterns (LBP) filter [24], and histogram of oriented gradient (HOG) filter [1], to the whole face or to specific regions. Geometry-based methods extract the shape and components of the face, such as the nose and mouth. The first step in most geometry-based methods is detection and tracking of facial points using an active appearance model (AAM) [19]. The facial shape and other information can be represented by these landmarks, which are designed in different ways.

Geometry-based and appearance-based methods have a common disadvantage, i.e. difficulty in selecting a good feature to replace the facial expression. For geometry-based features, the feature vector is associated with landmarks, which must be selected carefully. For appearance-based features, an experienced designer is required to design a powerful filter. The convolutional neural network (CNN) [18] has been applied to FER to address these limitations. CNN performs better than other deep learning methods [33], such as Deep Belief Network (DBN) which is one of the most widely used networks in FER. For example, CNN can automatically learn the features of data without manual selection, and it can combine different features neatly. Furthermore, CNN has a better effect on feature extraction than DBN, particularly for expressions of contempt, fear, and sadness [33]. CNN randomly initializes a specific number of filters before training and makes these filters better via gradient descent. One of the main advantages of CNN is that the input to the network is an original image rather than a set of hand-coded features. References [20,21,33] use deep convolutional neural network (DCNN) and ensemble convolutional neural network (ECNN) systems, respectively, and achieve good results. However, these systems have a few limitations. DCNN is difficult to train compared to CNN and has high validation error when the network layer is too large [7]. On the other hand, ECNN requires the generation of a large number of convolutional neural networks, which requires substantial computing resources and training time.

To overcome these limitations, we propose a new face cropping and image rotation strategy to improve the accuracy and simplify the CNN structure. The proposed approach was applied to the CK+ [16] and JAFFE [17] databases and compared with other methods. The main contributions of our work are as follows:

(1) Propose a new approach of face cropping to remove the useless regions in an image.
(2) Propose an image rotation strategy to cope with data scarcity.
(3) Build a simplified CNN structure for FER to reduce training/application time and to achieve real-time FER with an ordinary computer.

The remainder of this paper is organized as follows: Sect. 2 presents the most recent related work, while Sect. 3 introduces the proposed method in detail. The experimental results and a relevant discussion are given in Sect. 4. A comparison with other research works is presented in Sect. 5, and the conclusions are given in Sect. 6.

2 Related work

Several deep learning approaches for facial expression recognition were developed in the last decades, particularly the method of CNN. Some recent methods are focused on the construction of advanced networks and the training of model, the fusion of multiple structures and the selection of fusion parameters, and the optimization of classification algorithms.

Mayya et al. [20] proposed an approach to recognize facial expression using DCNN features. They used a DCNN architecture which is used for ImageNet [12] to extract the facial features, and then they obtain a 9216-dimensional vector for validation with support vector machine (SVM) classifier to recognize facial expression. Their experiments were conducted on two databases, CK+ and JAFFE, and achieved an accuracy of 96.02% and 98.12% for 7 classes, respectively. Despite the high accuracy, their validation method is LOSO which is more superior (discussed in Sect. 5). Their approach is not an end-to-end method, it is difficult and time-consuming to train.

Wen et al. [33] presented an ensemble of CNNs with probability-based fusion for FER. In their work, they used random techniques to generate 100 CNNs with rich diversity (different parameters) and then selected 37 CNNs (removed the CNNs which have bad performance) as the base classifier for their final model. Finally, a fusion method, such as majority voting, weighted majority voting, and the probability-based fusion, was employed for ensemble. Their method can reduce the training time by parallel computation, but it requires a large amount of computing resources.

Zhang et al. [37] proposed a novel well-designed CNN which can reduce same-expression variations and enlarge different-expression differences. They used 2-way soft-max function to train their model which requires the researchers have enough experience. However, their method is for smile detection, and they can use 4000 images for one expression which is much bigger than the database which can be used for FER (as explained in Sect. 4.1).

In comparison with the methods above, this work: (1) presents a competitive result in two public databases; (2) employs a simple yet effective CNN structure (not DCNN or ECNN), which is easy and fast to train; (3) does not need to design a tricky algorithm.

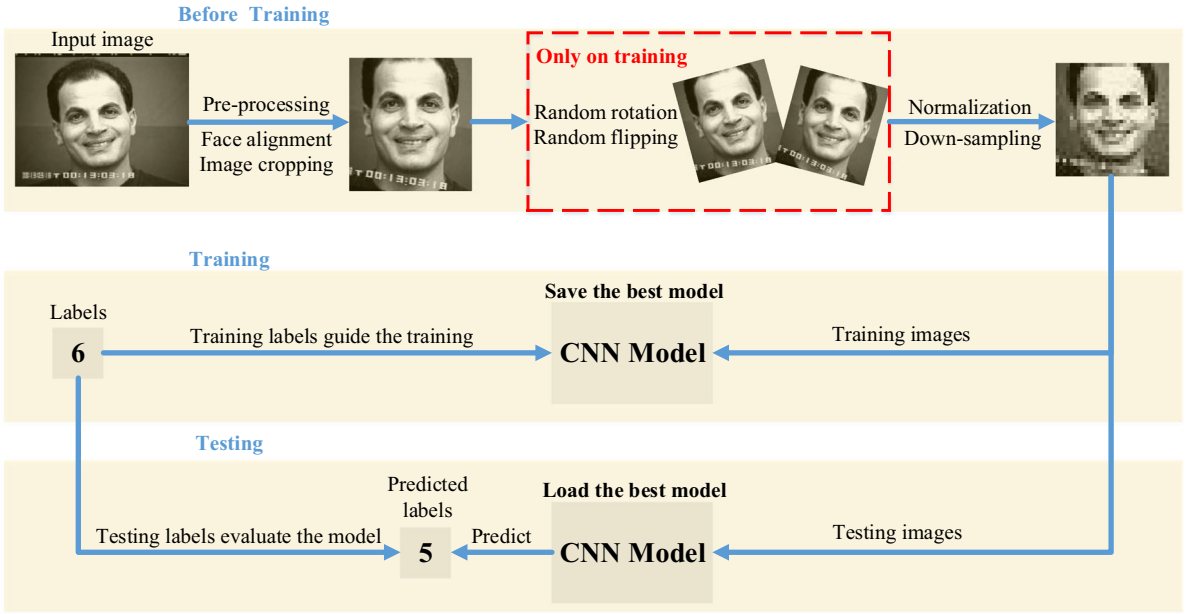

Fig. 1 Outline of the proposed system

3 Proposed method

The proposed FER system is a single classifier based on a CNN. Image pre-processing is required because the images have different colour channels and include people of various races. To cope with the lack of data, we expanded our training images using data augmentation. An outline of the proposed method is shown in Fig. 1. The aligned image was cropped to remove the useless region, and histogram equalization, Z-score normalization, and down-sampling were applied to standardize the image data. During the training phase, random rotation and horizontal flipping were performed to increase the database size. The expanded training data were used to train the CNN, and the best CNN model was saved. During the testing phase, the normalized testing images (without expansion) were sent to the CNN model from the training phase for prediction.

3.1 Face alignment

The images in the databases collected from a laboratory show various postures. These variations affect the system performance. Face alignment was performed to address this problem, as shown in Fig. 2. The algorithm used for face alignment was based on the position of the eyes. The Dlib toolkit [11] was used to obtain the face landmarks. A total of 68 sequential points, each of which could be represented by a coordinate, were identified, but only 12 face landmarks are shown in the figure for clarity. The centres of the left and right eyes were computed based on twelve points (No. 36 to 47). The first six points encircle the left eye centre, and the other six points encircle the right eye centre. The rotation angle is calculated on the basis of these twelve points using Eq. (1).

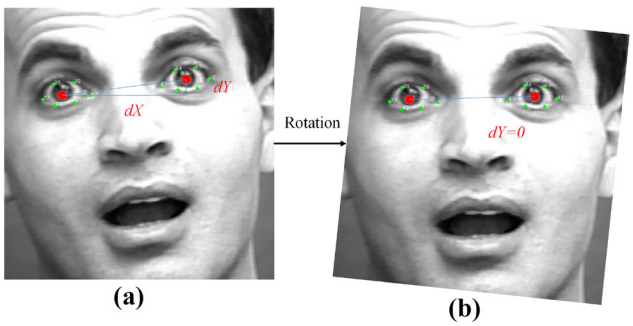

Fig. 2 Face alignment. **a** Before face alignment. **b** After face alignment

$$\text{angle} = \tan^{-1} \frac{\sum_{n=42}^{47} y_n - \sum_{n=36}^{41} y_n}{\sum_{n=42}^{47} x_n - \sum_{n=36}^{41} x_n} \tag{1}$$

where x_n is the x-coordinate of the nth point, and y_n is the y-coordinate of the nth point.

3.2 Image cropping

Image cropping is an important part of present study as we proposed a new method for face cropping. The proposed method was compared with two common methods. Figure 3a shows an image cropped using the OpenCV toolkit used by [14,36]; the cropped image has a little background. Figure 3b shows an image cropped using another common

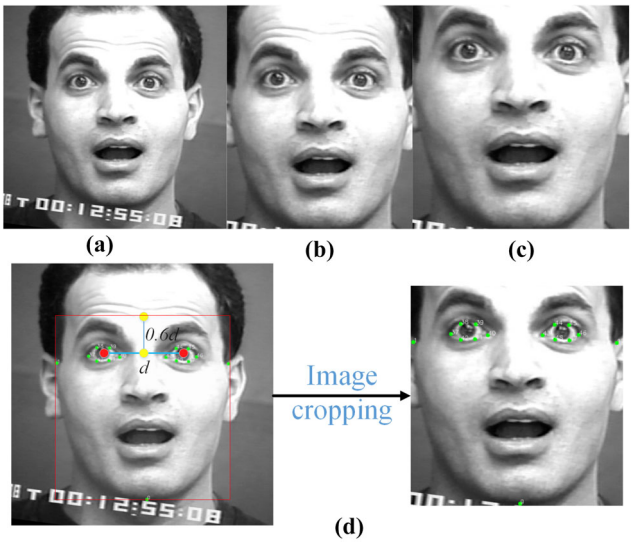

Fig. 3 Image cropping methods. **a** Face with background. **b** Face without background. **c** Face without forehead (proposed). **d** Illustration of the proposed face cropping method

cropping method [2,20] that removes the image background. Figure 3c shows an image cropped by the proposed method. Sixty-eight face landmarks were obtained, but only 15 are shown in the figure for clarity. The horizontal distance d between the eye centres is calculated using Eq. (2).

$$d = \frac{\left(\sum_{n=42}^{47} x_n - \sum_{n=36}^{41} x_n\right)}{6} \tag{2}$$

The forehead region was then removed in such a way that perpendicular distance from the top side of the cropped image to the horizontal line connecting the eye centres is 0.6 d (d is the distance between eye centres), as shown in Fig. 3d. The other three sides of the cropped image are defined by the coordinates of 1st, 9th, and 17th face landmark points.

3.3 Data normalization

Brightness and contrast can differ even between images of the same person with the same expression, as shown in Fig. 4a. Histogram equalization was applied to each image to reduce this variation. Figure 4b shows the images obtained by histogram equalization. The mean values of the normalized images are closer. Z-score normalization was also applied to these images using Eq. (3) to enhance the contrast.

$$x' = \frac{x - \mu}{\sigma} \tag{3}$$

where x' is the value of the new pixel, x is the value of the original pixel, μ is the mean pixel value of all sample images, and σ is the standard deviation of the pixel values of all sample images. Figure 4c shows the images obtained by Z-score normalization, the contrast of the normalized images is enhanced. Finally, the image is down-sampled to 32×32 pixels.

3.4 Data augmentation

After the preceding processing steps, there is slight tilt in the images, as shown in Fig. 5a. To ensure the adaptability and abundancy of our data, we adopted random horizontal flipping and random rotation. In random horizontal flipping, an image can be flipped before training to address the problem of uneven cropping. Random rotation expands the original data by rotating an image by a random angle within an interval. We selected the best rotation angle interval via mesh search method (described in Sect. 4.5). Figure 5b, c shows the random horizontal flipping and random rotation, respectively.

Fig. 4 Data normalization. **a** Original images. **b** After histogram equalization. **c** After Z-score normalization

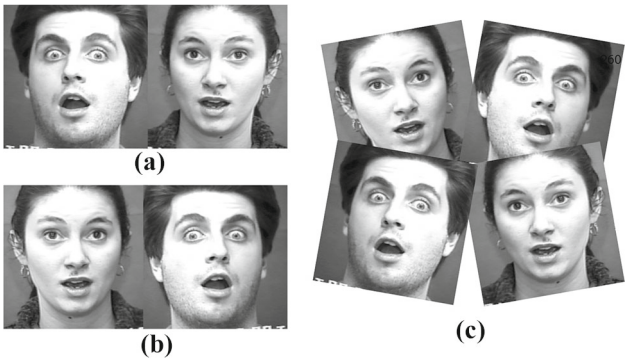

Fig. 5 Data augmentation. **a** Original images. **b** Horizontally flipped images. **c** Randomly rotated images

Fig. 7 Sample images from the CK+ and JAFFE databases

3.5 CNN structure

In the present work, CNN was applied to extract features and categorize expressions. The architecture of the simplified CNN, which has two convolution layers, two sub-sampling layers, and one output layer, is presented in Fig. 6. The first and third layers are convolution layers with 32 and 64 kernels, respectively, which have the size of 5×5. The activation function in the CNN is a rectified linear unit (Relu) function [8]. The second and fourth layers are sub-sampling layers that reduce the image size. We employed max-pooling with a kernel size of 2×2 and step size of 2. We flatted the sub-sampling to a 1600-dimensional vector and directly connected the output layer (with the soft-max activation function). In addition, the simplified CNN uses the momentum optimizer [31], Xavier initializer [4] and cross-entropy loss function.

4 Experiments and discussion

This section introduces the databases and the details of the experiments performed in this study. Experiments were performed to select the best number of neurons, the face cropping method, and the rotation angle. Following these selections, the final experiments were performed on the databases, and the results are discussed in detail.

The proposed method was implemented using OpenCV, Python, and the Neural Network Model library (tflearn-CPU). An Intel Core i5 3.2 GHz CPU was used to conduct all the experiments in an Ubuntu 16.04 environment.

4.1 Databases

Two widely used databases were used in the experiments: the Extended Cohn–Kanade (CK+) database and the Japanese Female Facial Expressions (JAFFE) database. The samples taken from these databases are shown in Fig. 7. The CK+ dataset contains 10,708 images and 327 video sequences collected from 118 participants. The last image of each video sequence is regarded as the peak expression, and they are labelled with seven expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. All sequences begin with a neutral expression, so the dataset contains an additional 327 neutral expressions. Neutral expressions are not discussed in this paper. The numbers of anger, contempt, disgust, fear, happiness, sadness, and surprise expressions are 45, 18, 59, 25, 69, 28, and 83, respectively. JAFFE contains 213 peak expressions collected from ten women, and the expressions are labelled as anger, disgust, fear, happiness, neutral, sadness, and surprise. Each expression is shown in approximately 30 images.

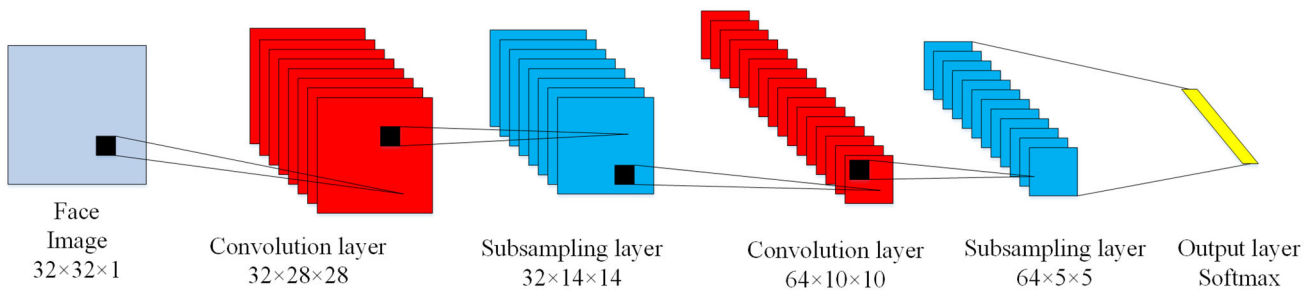

Fig. 6 Structure of the simplified CNN

4.2 Evaluation criteria

In the field of classification, there are many criteria that can be used to evaluate the model. Evaluation criteria are not fixed, we need to choose an appropriate one according to the actual problems we faced. A simple yet effective pair of evaluation criteria are error rate and precision. To evaluate the performance of our model f, we need to compare the predicted results with the real labels y. The error rate and accuracy are calculated using Eqs. 4 and 5, respectively. Precision and recall rate are another pair of criteria, they can provide more reliable information for certain situations (such as data imbalance). Based on the combination of the real label and predicted result, the sample can be divided into: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The precision and recall rate are calculated using Eqs. 6 and 7, respectively. For the multi-classification problem, we can construct a confusion matrix to describe the relation and difference between categories. $F1$ score is also a commonly used criterion, as shown in Eq. 8, it can be seen as a harmonic mean of the precision and recall.

$$\text{err} = \frac{1}{m}\sum_{i=1}^{m} g(f(\mathbf{x}_i) \neq y_i) \tag{4}$$

$$\text{acc} = \frac{1}{m}\sum_{i=1}^{m} g(f(\mathbf{x}_i) = y_i) \tag{5}$$

where g is the indicator function, m is the total number of data, and $\mathbf{x}$ is a given sample.

$$\text{pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$F1 = \frac{2 \times \text{pre} \times \text{rec}}{\text{pre} + \text{rec}} \tag{8}$$

For facial expression recognition, the data are balanced and we want to recognize images as many as possible. Therefore, the accuracy is more appropriate than precision and recall rates. The confusion matrix of multi-classification is used to describe the relation and difference between categories.

4.3 Selection of the neuron number

The effects of the number of neurons of the hidden fully connected layer are considered in this section. A 1600-dimensional vector was extracted after two convolution and sub-sampling layers. Fully connected layer acts as a classifier. We used four different numbers of neurons (0, 256, 512, and 1024) in the fully connected layer to improve the performance. Further, tenfold cross-validation was implemented in each experiment, and four experiments were conducted for each number of neurons to avoid contingency. Each of these experiments used the same optimizer, learning rate, face cropping method (face without background) and dataset (CK+ dataset) order for a fair comparison. The efficiencies of the fully connected layer with different numbers of neurons are shown in Fig. 8. Figure 8a shows the accuracy for the experiment with the hidden layer removed. The average accuracy was 89.40%. Figure 8b–d shows the accuracy for the experiments using neuron number as 256, 512, and 1024. The corresponding average efficiency is 86.34%, 87.99%, and 87.48%, respectively. Figure 8e shows the comparison of accuracy obtained from these four experiments. Compared with the other three networks, our network has a slight improvement in recognition accuracy (at least 1.41% improvement). At the same time, the proposed network has fewer parameters, so it takes less training and testing time competitively. It was observed that the max accuracy for all the experiments is obtained when the epoch is set between 80 and 120, as shown by a segment between vertical lines in Fig. 8e. The recognition accuracy started to decrease when epoch reaches 120 due to lack of data.

4.4 Evaluation on face cropping methods

After demonstrating the effects of the number of neurons, we performed experiments to compare the face cropping methods. Three face cropping methods are discussed in Sect. 3.2, and Fig. 3 shows the images cropped by these methods. Figure 3a is an image of a face with a background, Fig. 3b is an image of a face without a background, and Fig. 3c is an image of a face without a forehead, cropped with our proposed cropping method. Four experiments were conducted for each cropped image, and a CNN without a hidden fully connected layer was adopted. Each experiment used the same optimizer, learning rate, and dataset (CK+ dataset) order for a fair comparison. The average accuracy for the experiment using the image cropped with the proposed method was 92.42%, compared with the average accuracies of 86.90% and 89.40% for the image with the background and the image without the background, respectively. These efficiency curves are shown in Fig. 9.

In order to verify the universality of proposed method, we also conducted comparative experiments on two well-known networks: LeNet5 [18] and AlexNet. LeNet5 and AlexNet have two convolutional layers and five convolutional layers, respectively. The input image sizes are 32×32 and 224×224, respectively. To avoid interference from other factors (such as image random flipping), we only cropped the images in three ways. We conducted many experiments and obtained the average recognition accuracy, as shown in Table 1. The proposed face cropping method is better than

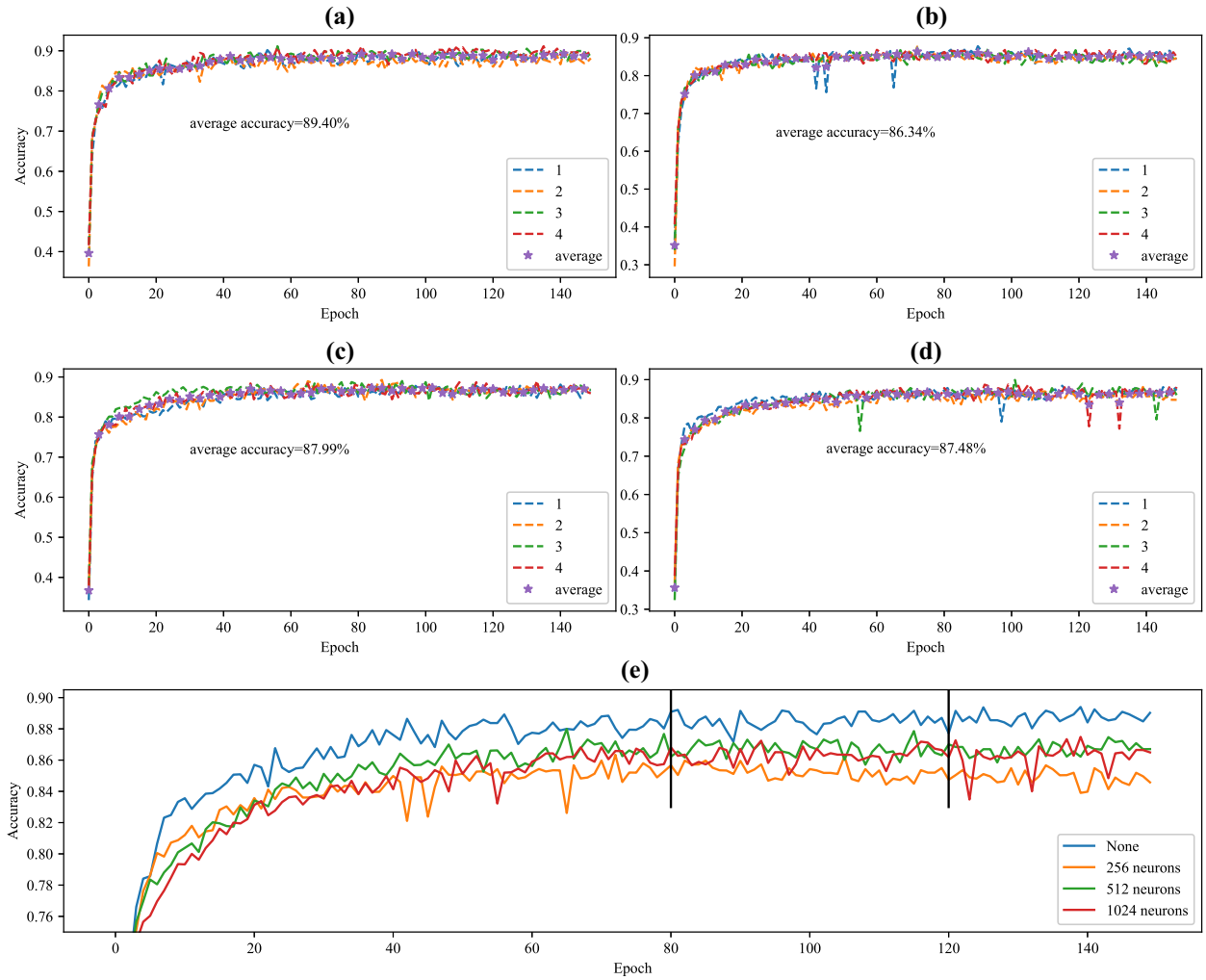

Fig. 8 Recognition accuracy for experiments performed during the selection of the number of neurons. **a** Without the fully connected layer. **b** With 256 neurons. **c** With 512 neurons. **d** With 1024 neurons. **e** Comparison of the accuracy for these four experiments

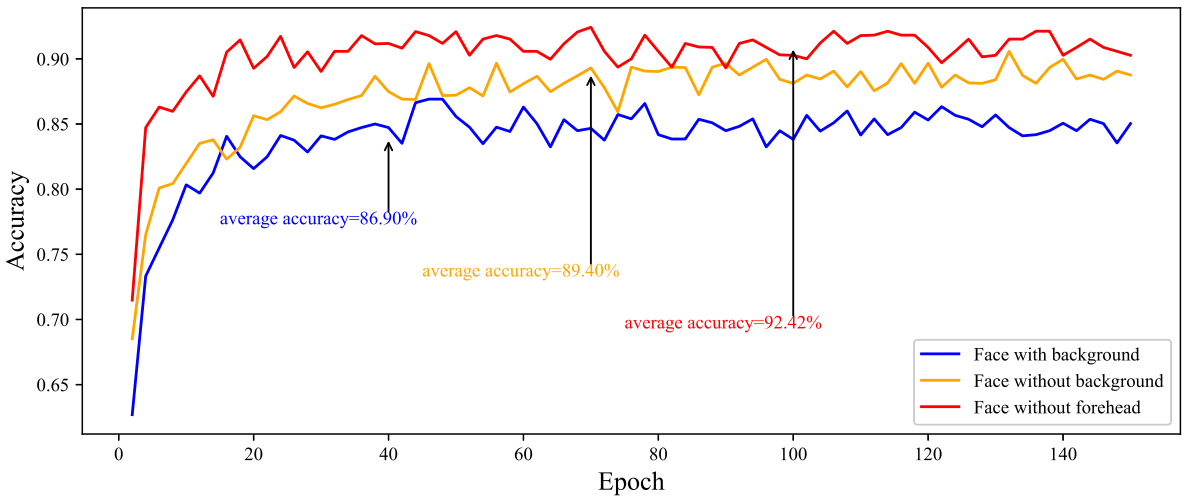

Fig. 9 Recognition accuracy for different image cropping methods

Table 1 Comparative accuracies of three face cropping methods on three networks

Network\method	With background (%)	Without background (%)	Proposed (%)
LeNet5	80.43	84.10	88.07
AlexNet	87.16	88.99	90.52
Proposed	86.90	89.40	92.42

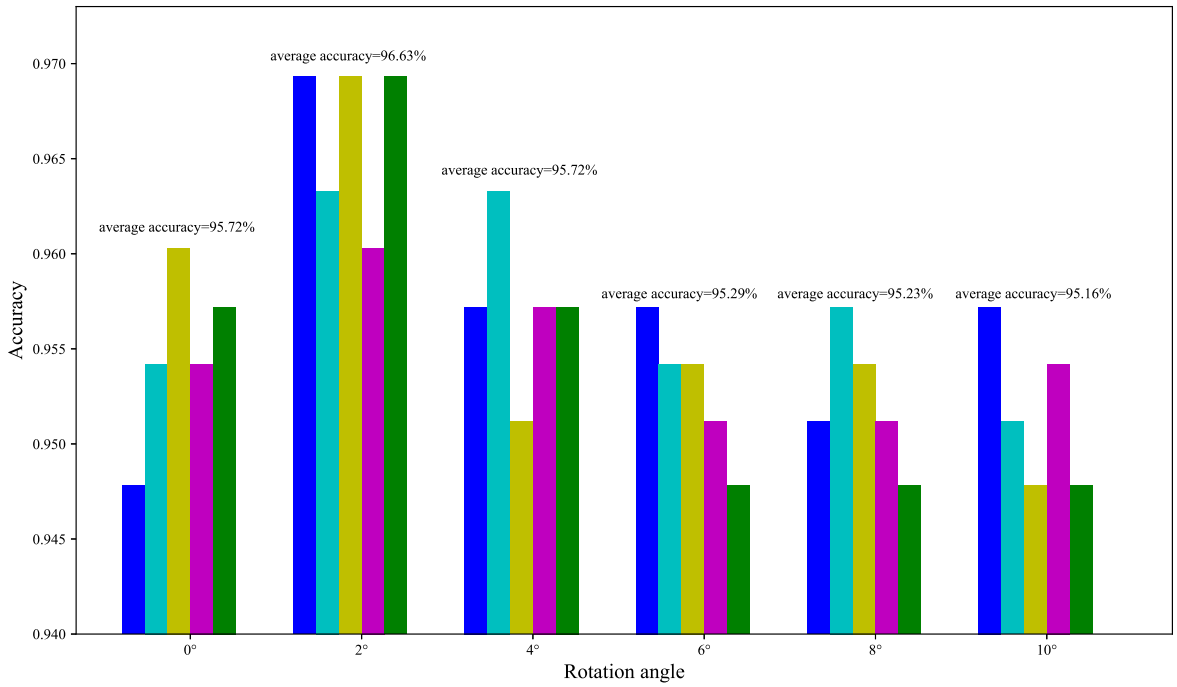

Fig. 10 Recognition accuracy for different rotation angles

Table 2 The effect of pre-processing steps

Step	Accuracy (%)
None	< 89.40
Remove fully connected layer	89.40
Remove fully connected layer + face crop	92.42
Remove fully connected layer + face crop + random flipping	95.72
Remove fully connected layer + face crop + random flipping + random rotation	96.63

the other two methods on all three networks. Moreover, the proposed method is more advantageous in small networks, such as LeNet5 and proposed network.

4.5 Selection of the random rotation angle

After the previous two sets of experiments, a third experiment was conducted to determine the effect of the rotation angle. The recognition accuracy decreased when the angle was too large. Therefore, it is important to select an optimal rotation angle. Four experiments were performed for different rotation angles (0 , 2, 4, 6 , 8, and 10), and the average accuracy was determined. Each experiment used the same optimizer, learning rate, face cropping method (proposed method), and dataset (CK+ dataset) order for a fair comparison. Random horizontal flipping was also implemented in this experiment. The accuracies for the set of experiments with different rotation angles are shown in Fig. 10. The accuracy increases when the image was rotated to 2, but further rotation decreases the accuracy. The maximum average accuracy was obtained for 2 rotation angle. The optimized rotation angle is data-dependent, and it depends on the image collecting condition. The rotation angle needs to be adjusted according to different data.

Table 2 summarizes the effects of these three steps (select the number of neurons, the face cropping method, and the

Table 3 Recognition accuracy for the seven-class and six-class experiments on the CK+ database using the proposed system

Accuracy	1	2	3	4	5
Seven classes (%)	97.27	97.27	97.55	97.55	97.27
Six classes (%)	98.38	98.05	98.38	97.73	98.38
Average of 7 classes: 97.38%		Standard deviation: 0.14%			
Best of 7 classes: 97.55%					
Average of 6 classes: 98.18%		Standard deviation: 0.26%			
Best of 6 classes: 98.38%					

Fig. 11 Misclassified images. The underlined caption is the real expressions, and the other caption is the expression predicted by the CNN

rotation angle). As it can be seen in Table 2, all these three steps we proposed can improve the recognition accuracy.

4.6 Database experiments

The experiments performed to select the best rotation angle, cropping method, and neuron number have a few limitations. For example, the same dataset order was used to train the CNN. Therefore, random experiments were performed using the proposed method. In addition to random rotation and random horizontal flipping, the following random approaches were employed: (1) the training and testing datasets were randomly assigned (they were not assigned evenly); and (2) the training dataset sequence was randomly generated before each epoch. In contrast to [15], the best training sample order was not selected.

CK+ database experiment The second expression in the CK+ database is contempt, which is not present in other databases. Therefore, this database was classified in terms of both seven and six expressions. Eighteen contempt expressions are contained in the database, so there are only 309 images when six expressions are considered. Tenfold cross-validation was applied. When classifying seven expressions, 294 images were used to train the model, and the remaining images were used for testing. Similarly, 278 images were used to train the 6-class CNN. The sample order was randomly generated in each experiment before training. Five experiments were conducted for each classification. Table 3 shows the accuracies achieved by the five experiments on the CK+ database for both classifications. The average and maximum accuracy for the 7-class experiment was 97.38% and 97.55%, respectively, while the average and maximum accuracy for the 6-class experiment was 98.18% and 98.38%, respectively. The results of the five experiments for each classification were similar.

A few images had a high probability of misclassification, and those images are shown in Fig. 11. Anger and sadness were the most likely expressions to be misclassified.

The recognition accuracy of each expression in both experiments is shown in Table 4. In the 6-class experiment, happiness, disgust, and sadness were recognized with 100% accuracy, and fear was recognized with an accuracy of at least 93.6%. The poor classification accuracy of fear was a result of there being only 25 images with an expression of fear, so there may be an uneven division of these images between the training and testing data. In the 7-class experiment, the recognition accuracy for sadness decreased significantly due to the addition of contempt. The proposed CNN model considered sadness and contempt to have similar features. The confusion matrices for the 7 classes and 6 classes in the CK+ database are shown in Tables 5 and 6, respectively. In both cases, more than 6.4% of the fear images were misclassified as happiness, and approximately 4% of the anger images were misclassified into other categories. Moreover, sadness was also wrongly classified as anger in the 7-class experiment.

JAFFE database experiment Similar to the CK+ dataset experiments, five experiments were conducted on the JAFFE dataset. The difference between the CK+ and JAFFE is that the JAFFE database includes neutral expressions instead of the contempt expressions contained in the CK+ database. The JAFFE database has only 213 images, but there is a similar number of images showing each expression. Tenfold cross-validation was applied, and the recognition accuracy for each expression is shown in Table 7. The average and maximum accuracy was 97.18% and 97.65%, respectively. The expressions of anger, fear, and happiness were recognized with 100% accuracy, while the neutral expression was

Table 4 Recognition accuracy obtained for each expression on the CK+ database

Classifier	An	Co	Su	Fe	Ha	Sa	Su
Seven classes (%)	95.56	94.44	98.98	93.60	100	93.57	98.07
Six classes (%)	96.00	–	100	91.20	100	100	98.07

Table 5 Confusion matrix for the seven-expression experiment on the CK+ database

	An	Co	Di	Fe	Ha	Sa	Su
An	215/225	0	0	2/225	5/225	3/225	0
Co	0	85/90	0	0	0	5/90	0
Di	0	0	292/295	0	0	3/295	0
Fe	0	0	0	117/125	8/125	0	0
Ha	0	0	0	0	345/345	0	0
Sa	9/140	0	0	0	0	131/140	0
Su	0	5/415	3/415	0	0	0	407/415

Table 6 Confusion matrix for the 6-expression experiment on the CK+ database

	An	Di	Fe	Ha	Sa	Su
An	216/225	0	0	5/225	4/225	0
Di	0	295/295	0	0	0	0
Fe	0	0	114/125	11/125	0	0
Ha	0	0	0	345/345	0	0
Sa	9/140	0	0	0	140/140	0
Su	0	2/415	6/415	0	0	407/415

Table 7 Recognition accuracy for each expression on the JAFFE database

Classifier	An	Di	Fe	Ha	Ne	Sa	Su
7 classes (%)	100	97.24	100	100	91.33	95.48	94.67
Average of 7 classes: 97.18% standard deviation: 0.30%							
Best of 7 classes: 97.65%							

Table 8 Confusion matrix for the 7-expression experiment on the JAFFE database

	An	Di	Fe	Ha	Ne	Sa	Su
An	150/150	0	0	0	0	0	0
Di	4/145	141/145	0	0	0	0	0
Fe	0	0	160/160	0	0	0	0
Ha	0	0	0	155/155	0	0	0
Ne	9/150	0	0	0	139/150	2/150	0
Sa	0	2/155	0	0	5/155	148/155	0
Su	0	0	0	0	8/155	0	142/150

recognized with an accuracy of at least 91.33%. The confusion matrix for the seven classes of JAFFE database is shown in Table 8. Neutral expressions were responsible for 80% of the misclassified images.

Cross-database experiment In these experiments, the network was trained on one database and tested on the other database. The CK+ database does not contain neutral expressions, and the JAFFE database does not include contempt expressions. Therefore, these expressions were neglected.

Table 9 Cross-database experiment on the CK+ and JAFFE databases

Train	Test	Average (standard deviation)	Best (%)
CK+	JAFFE	39.01% (1.12%)	40.98
JAFFE	CK+	62.78% (1.52%)	64.40

Five experiments were conducted for each case. The recognition accuracy for these experiments is shown in Table 9. The

Table 10 Training parameters used in our experiments

Parameters	Value	Remark
Random rotation	−2° to 2°	
Dropout	0.5	The second sub-sampling layer
Optimizers	Momentum	Learning rate = 0.001
Weights initializer	Xavier	
Batch size	16	
Loss function	Cross-entropy	
Epochs	120	Shuffle the training data

Table 11 Comparison of the proposed algorithm and other studies

Method	Validation	Database	Iteration Time	Recognition accuracy (%)		
				Binary	Six classes	Seven classes
DCNN + SVM [20]	LOSO	CK+	1.91 s		97.08	96.02
		JAFFE				98.12
CNN [15]	Eightfold	CK+	92.68 ms		96.76	95.75
Zeng et al. [36]	Tenfold	CK+				95.79
LBP + SVM [29]	Tenfold	CK+			95.10	91.40
Liu et al. [14]	Eightfold	CK+		96.70		
HOG + SVM [2]	LOSO	CK+			96.40	
Liu et al. [13]	Tenfold	CK+			95.78	
Pu et al. [26]	Tenfold	CK+				96.38
Proposed	Tenfold	CK+	69.16 ms		98.18	97.38
		JAFFE				97.18

average accuracy was 39.01% when the CK+ database was used for training and the JAFFE database was used for testing. In the opposite case, the average accuracy was 62.78%.

The proposed system is a real-time system. The time consumed for image recognition is divided into two parts. The first part is the time taken before the image is sent to the CNN, which includes the time consumed for face alignment, face cropping, histogram equalization, Z-score normalization, and image down-sampling. The other part is the time taken during CNN prediction. The time consumed by landmark generation is not considered because the corresponding files are provided in the CK+ database. A total of 1000 images were predicted by the proposed system, and the time consumed was recorded: 3.93 s and 1.58 s were consumed before and during the CNN process, respectively, i.e. the total processing time was 5.51 s. The proposed approach is summarized in Table 10 in terms of the parameters used in the experiments. A dropout [30] rate of 0.5 was applied to the second sub-sampling layer (1600-dimensional vector), a default learning rate of 0.001 was used, and the batch size was 16. During training, the CNN trained 120 epochs (each epoch was trained on the complete processed training data), and the training data order for each epoch was randomly shuffled.

5 Comparisons

Several novel methods for facial expression recognition have been proposed in recent years. In this section, the experimental results of the proposed approach on the CK+ and JAFFE databases are compared with those of other methods. The comparison is shown in Table 11. Tenfold cross-validation was not used by all researchers; therefore, we compare our method with existing similar or tenfold cross-validation methods. The authors in [20] achieved 98.12% accuracy on the JAFFE database for 7 classes by combining DCNN and support vector machines (SVM), which is 0.94% higher than the accuracy achieved with our method. However, they used leave-one-subject-out (LOSO) validation, which enabled 212 images to be used for training on the JAFFE database, whereas we used only 192 images. We conducted multiple experiments on JAFFE dataset using LOSO validation method, and the obtained average recognition accuracy reaches up to 98.59%, which is 0.47% higher than that obtained in [20]. The validation method in [2] was also LOSO. The researchers in [14,15] used eightfold cross-validation, [15] used the best sample order, and [14] trained seven binary classifiers for each expression. By contrast, we randomly divided the training set for our experiments and trained a seven-class classifier.

Training a neural network is a time-consuming task, especially for DCNN and ECNN. We implemented DCNN [20] with a batch size of 16 and computed the time required for comparison with our work. A total of 1.91 s was required to update the weights of DCNN [20], which is more than the time consumed by our method. We also computed the iteration time consumed by [15], whose network is similar to ours. The time taken by [15] was 92.86 ms, which is reduced to 69.16 ms in our case due to the removal of the hidden fully connected layer. DCNN required more than 12 h to train a model using tenfold cross-validation, while our proposed method required only 26 min.

In addition to the results of the CK+ and JAFFE database experiments, our results for the cross-database experiments were also competitive. The average recognition accuracy when using the CK+ database as the training data and the JAFFE database as the testing data was 39.01%, which is 0.21% higher than that achieved by [15].

6 Conclusion

In this paper, we present an efficient FER approach to simplify the CNN and propose new face cropping and image rotation methods. The impact of the CNN simplification and the proposed data processing methods was studied, and high recognition accuracies were achieved with each technique. The CNN without a hidden fully connected layer has a simpler structure and achieves improved recognition accuracy. The proposed face cropping method retains useful face information and removes useless regions, and the proposed rotation method greatly increases the amount of data. After separate evaluation of each proposed technique, final experiments were conducted on the CK+ and JAFFE databases. The results show that the proposed FER approach achieves competitive results in terms of training time, testing time, and recognition accuracy. Furthermore, the proposed method can be implemented on an ordinary computer without GPU acceleration.

Funding This research was sponsored by the National Natural Science Foundation of China (Grant No. 51605464), National Basic Research Program of China (973 Program) (2014CB049500) and Research on the Major Scientific Instrument of National Natural Science Foundation of China (61727809).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 886–893. https://doi.org/10.1109/CVPR.2005.177 (2005)
2. De la Torre, F., Chu, W.S., Xiong, X., Vicente, F., Ding, X., Cohn, J.F.: Intraface. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp 1–8. https://doi.org/10.1109/FG.2015.7163082 (2015)
3. Ekman, P., Friesen, W.V.: Facial action coding system: a technique for the measurement of facial movement. In: Consulting Psychologists, Palo Alto (1978)
4. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. J. Mach. Learn. Res. **9**, 249–256 (2010)
5. Gogić, I., Manhart, M., Pandžić, I.S., Ahlberg, J.: Fast facial expression recognition using local binary features and shallow neural networks. Vis. Comput. 1–16 (2018). https://doi.org/10.1007/s00371-018-1585-8
6. Goh, K.M., Ng, C.H., Lim, L.L., Sheikh, U.: Micro-expression recognition: an updated review of current trends, challenges and solutions. Vis. Comput. 1–24 (2018). https://doi.org/10.1007/s00371-018-1607-6
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778 (2016)
8. Jarrett, K., Kavukcuoglu, K., Ranzato, M., Lecun, Y.: What is the best multi-stage architecture for object recognition? In: IEEE International Conference on Computer Vision, vol 30, pp 2146–2153 (2009)
9. Jin, H., Wang, X., Lian, Y., Hua, J.: Emotion information visualization through learning of 3d morphable face model. Vis. Comput. 1–14 (2018). https://doi.org/10.1007/s00371-018-1482-1
10. Jones, J.P., Palmer, L.A.: An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. **58**(6), 1233–1258 (1987). https://doi.org/10.1152/jn.1987.58.6.1233
11. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp 1097–1105. Curran Associates, Inc., Lake Tahoe, Nevada, USA (2012)
13. Liu, M., Li, S., Shan, S., Chen, X.: Au-inspired deep networks for facial expression feature learning. Neurocomputing **159**(C), 126–136 (2015). https://doi.org/10.1016/j.neucom.2015.02.011
14. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1805–1812 (2014)
15. Lopes, A.T., Aguiar, E.D., Souza, A.F.D., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit. **61**, 610–628 (2016). https://doi.org/10.1016/j.patcog.2016.07.026
16. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: IEEE conference on computer vision and pattern recognition workshops, pp 94–101. https://doi.org/10.1109/CVPRW.2010.5543262 (2010)
17. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE Trans. Pattern Anal. Mach. Intell. **21**(12), 1357–1362 (1999). https://doi.org/10.1109/34.817413

18. Lcun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
19. Matthews, I., Baker, S.: Active appearance models revisited. Int. J. Comput. Vis. **60**, 135–164 (2004)
20. Mayya, V., Pai, R.M., Pai, M.M.M.: Automatic facial expression recognition using dcnn. Proc. Comput. Sci. **93**, 453–461 (2016a). https://doi.org/10.1016/j.procs.2016.07.233
21. Mayya, V., Pai, R.M., Pai, M.M.M.: Combining temporal interpolation and dcnn for faster recognition of micro-expressions in video sequences. In: International Conference on Advances in Computing, Communications and Informatics, pp 699–703. https://doi.org/10.1109/ICACCI.2016.7732128 (2016)
22. Mehrabian, A.: Communication without words. Commun. Theory, 193–200 (2008)
23. Mohammadi, M.R., Fatemizadeh, E., Mahoor, M.H.: Pca-based dictionary building for accurate facial expression recognition via sparse representation. J. Vis. Commun. Image Represent. **25**(5), 1082–1092 (2014). https://doi.org/10.1016/j.jvcir.2014.03.006
24. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. **29**(1), 51–59 (1996). https://doi.org/10.1016/0031-3203(95)00067-4
25. Owusu, E., Zhan, Y., Mao, Q.R.: An svm-adaboost facial expression recognition system. Appl. Intell. **40**(3), 536–545 (2014)
26. Pu, X., Fan, K., Chen, X., Ji, L., Zhou, Z.: Facial expression recognition from image sequences using twofold random forest classifier. Neurocomputing **168**(C), 1173–1180 (2015). https://doi.org/10.1016/j.neucom.2015.05.005
27. Rashid, M., Abu-Bakar, S., Mokji, M.: Human emotion recognition from videos using spatio-temporal and audio features. Vis. Comput. **29**(12), 1269–1275 (2013)
28. Rivera, A.R., Castillo, J.R., Chae, O.: Local directional number pattern for face analysis: face and expression recognition. IEEE Trans. Image Process. **22**(5), 1740–1752 (2013). https://doi.org/10.1109/TIP.2012.2235848
29. Shan, C., Gong, S., Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009). https://doi.org/10.1016/j.imavis.2008.08.005
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
31. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International Conference on Machine Learning, pp 1139–1147 (2013)
32. Uddin, M.Z., Hassan, M.M., Almogren, A., Zuair, M., Fortino, G., Torresen, J.: A facial expression recognition system using robust face features from depth videos and deep learning. Comput. Electr. Eng. **63**, 114–125 (2017). https://doi.org/10.1016/j.compeleceng.2017.04.019
33. Wen, G., Hou, Z., Li, H., Li, D., Jiang, L., Xun, E.: Ensemble of deep neural networks with probability-based fusion for facial expression recognition. Cogn. Comput. **9**(5), 597–610 (2017). https://doi.org/10.1007/s12559-017-9472-6
34. Yang, P., Liu, Q., Metaxas, D.N.: Boosting coded dynamic features for facial action units and facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1–6. https://doi.org/10.1109/CVPR.2007.383059 (2007)
35. Yu, Z., Liu, Q., Liu, G.: Deeper cascaded peak-piloted network for weak expression recognition. Vis. Comput. **34**(12), 1691–1699 (2018). https://doi.org/10.1007/s00371-017-1443-0
36. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643–649 (2017). https://doi.org/10.1016/j.neucom.2017.08.043
37. Zhang, K., Huang, Y., Wu, H., Wang, L.: Facial smile detection based on deep learning features. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, pp 534–538 (2015)
38. Zhao, G., Pietikinen, M., Member, S.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2008). https://doi.org/10.1109/TPAMI.2007.1110
39. Zhao, J., Mao, X., Zhang, J.: Learning deep facial expression features from image and optical flow sequences using 3D CNN. Vis. Comput. **34**(10), 1461–1475 (2018). https://doi.org/10.1007/s00371-018-1477-y

Kuan Li received his B.S. degree from China University of Mining and Technology, China, in 2016. He is currently pursuing the masters degree from University of Science and Technology of China, Hefei, China. His research interest includes image processing, pattern recognition, and deep learning.

Yi Jin received his Ph.D. degree from University of Science and Technology of China, China, in 2013. He is currently an Associate Professor at University of Science and Technology of China. His current research interest includes human–computer interaction, pattern recognition, and image processing.

Muhammad Waqar Akram received his masters degree from University of Agriculture Faisalabad, Pakistan, in 2015. He is currently pursuing PhD degree in precision machinery and instrumentation at University of Science and Technology of China, Hefei, PR China. His research interest includes solar energy, farm machinery, and computer vision.

Ruize Han received his B.S. degree from Hebei University of Technology, China, in 2016. He is currently pursuing masters degree in Tianjin University, Tianjin, China. His research interest includes image processing and computer vision.

Jiongwei Chen received his B.S. degree from China University of Mining and Technology, China, in 2016. He is currently pursuing masters degree in University of Science and Technology of China, Hefei, China. His research interest includes image processing and pattern recognition.