J. Meas. Sci. Instrum., 2025, 16(1): 85-95. http://jmsi.publish.founderss.cn/
DOI: 10.62756/jmsi.1674-8042.2025009 jmsi@nuc.edu.cn

Face recognition algorithm using collaborative sparse representation based on CNN features

ZHAO Shilin[1,2*], XU Chengjun[1], LIU Changrong[1]

1. School of Digital Media, Lanzhou University of Arts and Science, Lanzhou 730010, China;
2. School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

***Corresponding author:** ZHAO Shilin (1000706@luas.edu.cn)

Received: Feburary 19, 2024 **Revised:** April 16, 2024 **Accepted:** May 2, 2024

Abstract: Considering that the algorithm accuracy of the traditional sparse representation models is not high under the influence of multiple complex environmental factors, this study focuses on the improvement of feature extraction and model construction. Firstly, the convolutional neural network (CNN) features of the face are extracted by the trained deep learning network. Next, the steady-state and dynamic classifiers for face recognition are constructed based on the CNN features and Haar features respectively, with two-stage sparse representation introduced in the process of constructing the steady-state classifier and the feature templates with high reliability are dynamically selected as alternative templates from the sparse representation template dictionary constructed using the CNN features. Finally, the results of face recognition are given based on the classification results of the steady-state classifier and the dynamic classifier together. Based on this, the feature weights of the steady-state classifier template are adjusted in real time and the dictionary set is dynamically updated to reduce the probability of irrelevant features entering the dictionary set. The average recognition accuracy of this method is 94.45% on the CMU PIE face database and 96.58% on the AR face database, which is significantly improved compared with that of the traditional face recognition methods.

Key words: sparse representation; deep learning; face recognition; dictionary update; feature extraction

0 Introduction

Face recognition is a technology that utilizes the features reflected in the face image to make identity judgments or authentication, and it originated earlier and entered into wide application later. With the maturity of the technology, face recognition technology has achieved wide applications in identity authentication, automated management, self-service, security criminal investigation, and many other aspects. The earliest face recognition technology to enter the public's life is the face ID equipped on Apple's iPhone X cell phone in 2017. Face ID technology is powerful but extremely difficult to realize and in the absence of not open source. Especially for close-range face recognition, it has attracted large number of researchers.

Face recognition system as a non-contact means of authentication has achieved a wide range of applications and high accuracy, but the recognition results of the algorithms are highly dependent on training set and test samples, and recognition effect is greatly reduced in some complex scenes with serious occlusion, dramatic change in lighting and accompanied changes in posture and expression. Researchers have made many breakthroughs in face feature extraction and recognition algorithm optimization, but these problems are still hot issues in the field. Initial face recognition using classical classification algorithms[1] has achieved good classification results in conventional recognition environments. As the problem deepens further, more cutting-edge approaches have been proposed. Sparse coding[2] as a classical signal decomposition and reconstruction algorithm in the field of signal processing has also been gradually discovered and applied in the field of image processing. Based on the theory of sparse representation, Wright et al. [3] proposed a sparse representation-based image classification method, which achieves good classification results in complex environments such as faces subject to partial occlusion, lighting changes, and noise interference. Subsequently, many sparse representation methods have been continuously proposed, such as Fisher discriminant classifier[4,5], dictionary update learning[6,7], deep

learning[8-11], and other feature extraction methods[12,13]. Although these methods have achieved better recognition results, they are still not able to achieve reliable recognition in specific scenarios. Therefore, some improvements are proposed. For example, Yang et al.[14] utilized global sparse representation and the weighted integrals of component sparse representation to construct L1, and then L1 paradigm after weighting approximates original L0 paradigm, which exhibits better performance than other works in low signal-to-noise ratio scenarios. Nakachi et al.[15] improved the encryption algorithm and used legacy encryption algorithm to solve L0 paradigm based on an orthogonal matching tracking (OMP) algorithm, which well protects the privacy of observed signal. These improvements in face recognition methods based on paradigm solving, dictionary learning, and fast optimization bring the rapid development of sparse representation-based face recognition methods. However, Due to a variety of complex environments with small samples, the sparse representation may update some background information similar to the features of target samples when updating the dictionary of the samples, coupled with the fact that it is difficult to represent face models completely in the complex background using a single traditional feature extraction algorithm. To completely represent the face model in the complex background, some neural network models have been proposed. Blauch et al. [16] trained the convolutional neural network (CNN) with a large-scale dataset as a prerequisite and used it for face recognition, and the results are amazing. Bhasha et al.[17] used images with pose and illumination variations to train a deep convolutional inverse map network and then generated a new image, which shows a powerful feature extraction capability. Song et al. [18] proposed an improved method based on a multi-task cascaded CNN model, which can classify faces based on task relevance. Sun et al. [19] earlier proposed a face recognition method based on sparse representation of deeply learned features, exposing the advantages of deep learning in face feature learning. Based on the method[19], Guo et al.[20-25] improved the network in terms of network lightweight, small sample identification, personalized convolution, and noise interference, respectively, which achieved good results. The above research results expose that the combination of deep learning network-based features and sparse representation classification algorithms has great potential in the field of face recognition.

Currently, facial occlusion in face recognition is a universal phenomenon. Therefore, how to improve the recognition rate of algorithms in multiple complex backgrounds has become a crucial issue. Inspired by previous work, we propose a method that reconstructs the target's appearance model using two types of features: one is the commonly used Haar feature, which can effectively deal with the occlusion of target; the other is the deep feature extracted from the trained CNN, which has better separability and migration[8]. Based on this, a collaborative face recognition using a two-stage sparse representation model is established. Experimental results show that the use of two types of features to jointly model the target can improve the accuracy of face recognition algorithms in complex environments and has certain practical application value.

1 Basic theory

The function of sparse representation is to analyze the characteristics of the signal and construct an overcomplete dictionary of the signal, eliminate the spatial correlation between signal elements, and use a small number of as sparse as possible signals and an overcomplete dictionary to represent the original signal. The process of signal sparse representation is also the process of solving sparse coefficients, which can be further transformed into the solution of paradigms according to its mathematical representation.

1.1 Face recognition based on sparse representation

The process of sparse representation is shown in Fig.1, where $\boldsymbol{y}$ is the image to be represented, $\boldsymbol{D}$ is the overcomplete dictionary of the image, $\boldsymbol{x}$ is the coefficient matrix of the sparse representation, and the blanks of $\boldsymbol{x}$ denote the atoms with the value of 0. The more the atoms in $\boldsymbol{x}$ with the value of 0, the sparser the representation of the signal.

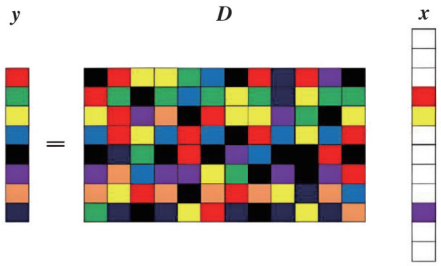

Fig. 1 Signal sparse representation

The process of sparse representation can be transformed into a sparse approximation solving problem as

$$\arg\min_{x}\|\boldsymbol{e}\|_2 \quad subject\ to\ \boldsymbol{y}=\boldsymbol{D}\boldsymbol{x}+\boldsymbol{e} \text{ and } \|\boldsymbol{x}\|_1\leqslant m, \tag{1}$$

where $\boldsymbol{e}$ represents the reconstruction error in sparse representation; $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}$, $\boldsymbol{x}_i$ is the coefficient of the representation corresponding to each

class; $D=\{D_1,D_2,\cdots,D_i,\cdots,D_n\}$ is the training sample, D_i is the dictionary of the ith class; $\boldsymbol{y}$ is the test sample; and m is a constant representing the upper bound of L1 norm during optimization.

The process of sparse representation face recognition method[3] (SRC) is summarized as constructing training samples and test samples based on the face library, the training samples are the overcomplete dictionaries of each category of faces, and the test samples are the face images to be categorized. The category of target images is judged from the reconstruction error of each category of training samples (overcomplete dictionaries) on the face image under test, which is the basic principle of the sparse representation-based face recognition method and is expressed as

$$\hat{\boldsymbol{x}}=\arg\min_{\boldsymbol{x}}\|\boldsymbol{y}-\boldsymbol{D}\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{x}\|_1, \tag{2}$$

where λ is a regularization parameter to control the level of sparsity.

The categorization process is represented as

$$D_{\text{identy}}=\arg\min_{i}\|\boldsymbol{y}-\boldsymbol{D}_i\hat{\boldsymbol{x}}_i\|_2^2. \tag{3}$$

The sparse representation method for classifying target images needs each category of training samples up to required quantity, i.e., the dictionary for each category of samples must be complete. This condition often results in misclassification when there are few training samples and target images are affected by occlusion, pose change, illumination change, expression change, etc. Additionally, since the structural design of sparse representation itself requires the target images to be split for representation, the spatial properties among image atoms are ignored. To solve this problem, Akahavan et al.[26-29] introduced the image compensation dictionary into the sparse representation-based face recognition algorithm, but could not eliminate external irrelevant features into the target feature dictionary set.

1.2 Deep CNN

As early as 1962, Hubel and Wiesel[30], neuroscientists from Canada, by studying the visual cortex system of the cat brain, proposed a kind of cell named "sensory field" that could respond to localized visual images, and gave the mechanism of co-processing the information in the visual cortex pathway. The mechanism of information co-processing in the visual cortex pathway has been further studied. Waibel et al.[31] from the University of Toronto revealed that the neural network containing multiple hidden layers has a strong feature learning ability, and through the training of the model, it can express the original data more abstractly and essentially; the mechanism of layer-by-layer initialization can realize the hierarchical expression of information, reduce the difficulty of model training, and put forward the concept of deep learning for the first time. In the subsequent decade, deep CNNs have been widely used in image classification tasks. Based on the deep DNN model, scholars further improved its accuracy by 20%—30%, and this method has become the mainstream method in the field of image recognition at present.

The input of the CNN can be the original image or other images that have been transformed, and the expected output can be obtained directly from the input without any human intervention in the process, which is an end-to-end model. The data from input to output need to go through the role of the intermediate layers to finally get a specific output, the output can be expressed as an objective function, and the training process of CNN can be understood as the optimization process of the objective function, which can be refined as forward propagation to get the network error and back propagation to adjust the parameters of each layer of the network according to the error function. CNN usually takes the value of the last fully connected layer as the value of the original data feature extraction, ensuring the reliability of the features through network training. Fig.2 shows the CNN processing.

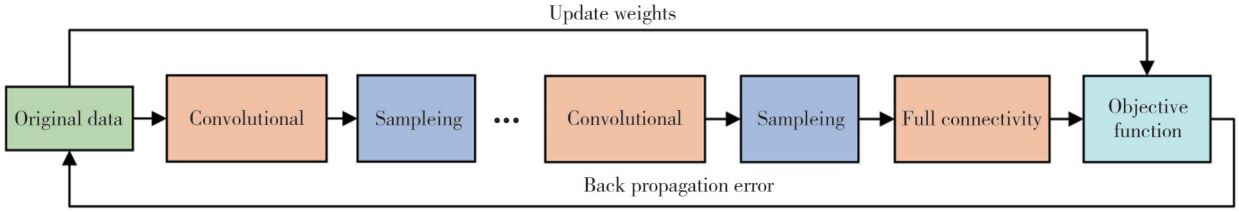

Fig. 2 CNN processing

Deep learning features have been proven to have better differentiation, portability, and subspace properties in many research experiments. To better solve the problem of low accuracy of face recognition algorithms due to few training samples[8], we adopt the deep learning network for the extraction of target features and the collaborative sparse representation classification method proposed in the following section

for the effective recognition of face images.

2 Collaborative sparse representation for face recognition based on deep features

Based on Ref. [32], the CNN structure is further improved and optimized by training the network via the existing dataset and extracting the depth features of the face with the help of the trained network model. Then, the advantages and disadvantages of sparse representation models are thoroughly analyzed, and collaborative sparse representation for face recognition is proposed. Based on the deep learning feature extraction algorithm and collaborative sparse representation for face recognition, a collaborative sparse representation via deep learning features-based classification (CSRDFC) is proposed. The experimental validation shows that the CSRDFC method has good robustness in dealing with the face recognition problems affected by occlusion, illumination changes, and multiple complex factors.

2.1 Principle of CSRDFC algorithm

The CSRDFC based on deep learning contains four processes: feature extraction, collaborative sparse representation, face recognition classification, and dictionary updates. The specific process is shown in Fig.3.

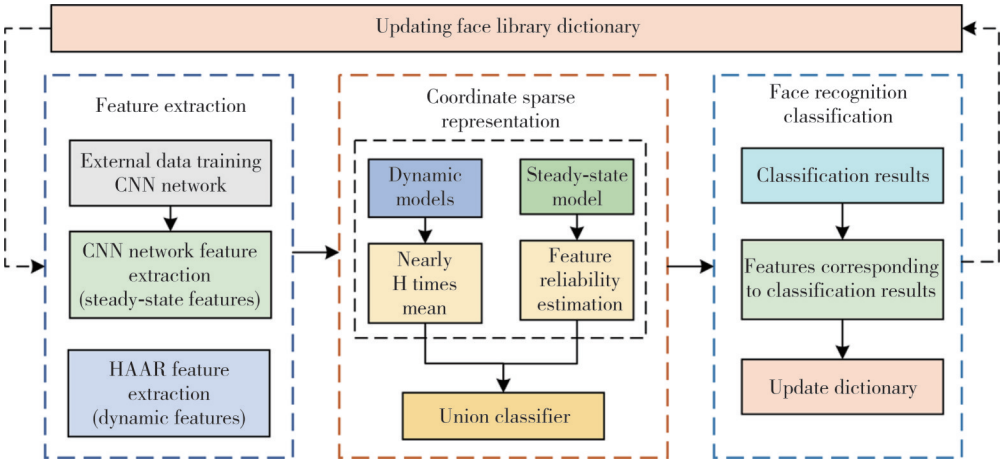

Fig. 3 CSRDFC methodology process

As shown in Fig. 3, feature extraction contains three stages: the first stage is to train the CNN based on the external face data of the system and determine the network parameters, and the specific training process can be found in Section 2.2; the second stage uses the trained network model to extract the depth features of the face data in the face recognition system as the steady-state face features; and the third stage extracts the face Haar features as dynamic face features. Collaborative sparse representation is the core of this algorithm, and its main task is to use the two-stage collaborative sparse representation model to construct a sparse representation classifier based on the dynamic and steady-state features of the face for reliability estimation of the classified features (Section 2.3), and dynamically update the dictionary set. The face recognition process can use the features with higher reliability in the previous process to reconstruct the face to be tested and update the complete dictionary in real time according to the classification results to ensure the robustness of face recognition in dynamic environments.

2.2 Feature extraction based on CNN

The feature extraction process of the CSRDFC method is done by the deep CNN, the training of the network mainly uses local connection, weight sharing, and subsampling operations, which makes both the complexity of the model and the number of weights decrease. The input of the network is a gray face image, where different colors represent different network layers. Considering that network nodes are more, only the overall network structure is given in Fig.4. In the training process of the network, the activation function is chosen to be ReLU, and the dropout probability is set to 0.5. The 400 000 images of nearly seven thousand types of face images from the CASIA dataset which contains the effects of various factors such as occlusion, pose change, expression, and illumination are chosen for the training of the network, and all the images are applied with the DLIB open-source library for face alignment.

The training process uses supervised learning, the input of the network is the original image, the output is the labeled values of the image, and the output of the fully connected layer is used as the extracted face features. The trained network is tested on the LFW dataset (the distance function uses the Euclidean distance), and the accuracy of the model is between

0.951 2 and 0.973 4. The results show that the features extracted by the CNN network proposed in this study have good distinguishability and portability, and the specific parameters of each layer of the network are shown in Table 1.

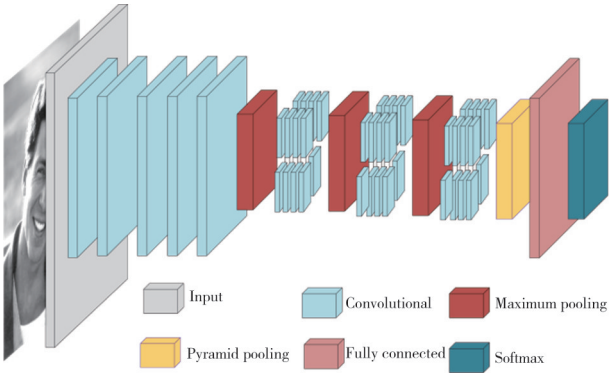

Fig. 4 CSRDFC feature extraction network structure

Table 1 Parameters of feature extraction network

Network layer	Parameter
Input	64×64
Convolutional 1	64×9×9, step=2
Maximum pooling 1	5×5, step=1
Convolutional 2	128×2×2, step=2
Convolutional 3,4,5	128×2×2, step=1
Maximum pooling 2	2×2, step=1
Convolutional 6	256×2×2, step=2
Convolutional 7,8,9	256×2×2, step=1
Maximum pooling 3	2×2, step=1
Convolutional 10	512×2×2, step=2
Convolutional 11,12,13	512×2×2, step=2
Space pyramid pooling	{6, 3, 2, 1}
Full connectivity	512
Softmax	7 000

2.3 Collaborative sparse representation

2.3.1 Dynamic classifier for face recognition

In the collaborative sparse representation algorithm, firstly, the stability and dynamics of face appearance changes are represented by steady-state features (CNN features) and dynamic features (Haar features), respectively; then, dynamic features are used to construct a dynamic appearance model of the face, and steady-state features and sparse representation algorithm are used to represent the steady-state appearance model of the face; finally, a face recognition classifier is constructed based on the appearance models of the above two processes, and the classification feature weights are adjusted dynamically according to the recognition results so as to dynamically adjust the classification feature weights.

$$\bar{\boldsymbol{f}}_{i,t}^{k}=\frac{1}{H}\sum_{h=1}^{H}\boldsymbol{f}_{i,t-h}^{k}, \tag{4}$$

where $\bar{\boldsymbol{f}}_{i,t}^{k}$ is denoted as the dynamic appearance model of the face at moment t. The dynamic appearance model of the face at moment t is the average of the last H recognized face appearance models.

2.3.2 Steady-state classifier for face recognition

Defining $\boldsymbol{f}_{s,t}^{k}$ as the steady-state appearance model of the face at moment t and using the face steady-state features and the sparse representation algorithm to jointly represent the face steady-state appearance model as $\boldsymbol{y}_t^{i,k}$, the steady-state appearance model of the face is represented as

$$\boldsymbol{y}_t^{i,k}\approx\boldsymbol{f}_{s,t}^{k}+\boldsymbol{\varepsilon}^{i,k}=\boldsymbol{f}_{1,t}^{i,k}\boldsymbol{\alpha}_{1,t}^{i,k}+\boldsymbol{f}_{2,t}^{i,k}\boldsymbol{\alpha}_{2,t}^{i,k}+\cdots+\boldsymbol{f}_{s,t}^{i,k}\boldsymbol{\alpha}_{s,t}^{i,k}. \tag{5}$$

In the process of recognizing a target face, the face can usually be reconstructed by multiple dictionaries, and face recognition methods select the dictionary with the smallest reconstruction error via sparse representation. However, the environment surrounding the target face is often complex and variable, which makes only some of the face features and the background environment distinguishable. In this study, a face recognition method based on collaborative sparse representation is proposed to obtain the minimum reconstruction error in reconstructing the target face. The L2 regular least squares method is used to solve the sparse coefficients in the collaborative sparse representation method, so as to reduce the computational complexity.

Defining $\boldsymbol{f}_1=\left[\boldsymbol{f}_{s,t}^{k},\boldsymbol{I}^{k}\right]$, $\boldsymbol{\alpha}_1=\begin{bmatrix}\boldsymbol{\alpha}_t^{i,k}\\ \boldsymbol{\beta}_t^{i,k}\end{bmatrix}$, $\boldsymbol{\alpha}_t^{i,k}=\left[\boldsymbol{\alpha}_{1,t}^{i,k},\cdots,\boldsymbol{\alpha}_{r,t}^{i,k}\right]^{\mathrm{T}}\in\mathbf{R}^{r}$ as sparse coefficient vectors, where noise coefficient vector $\boldsymbol{\beta}_t^{i,k}=\left[\boldsymbol{\beta}_{1,t}^{i,k},\cdots,\boldsymbol{\beta}_{d^k,t}^{i,k}\right]^{\mathrm{T}}\in\mathbf{R}^{d^k}$, and $\boldsymbol{I}^{k}\in\mathbf{R}^{d^k\times d^k}$ is a unit matrix, thus the candidate sample $\boldsymbol{y}_t^{i,k}$ can be sparsely represented as

$$\boldsymbol{y}_t^{i,k}=\boldsymbol{f}_{s,t}^{k}\boldsymbol{\alpha}_t^{i,k}+\boldsymbol{\varepsilon}^{i,k}=\left[\boldsymbol{f}_{s,t}^{k}\,\boldsymbol{I}\right]\begin{bmatrix}\boldsymbol{\alpha}_t^{i,k}\\ \boldsymbol{\beta}_t^{i,k}\end{bmatrix}. \tag{6}$$

When the face is recognized at moment t, two recognition results $\{\hat{\boldsymbol{x}}_t^{i}|i=1,2\}$ will be obtained. In the ith classifier, $\boldsymbol{y}_t^{i,k}$ denotes the candidate image block represented by the kth feature $\boldsymbol{f}_{s,t}^{k}$ and $\boldsymbol{\alpha}_t^{i,k}$. Then, a two-stage cooperative sparse coding method is used to solve the sparse coefficient vector $\boldsymbol{\alpha}_t^{i,k}$ and the noise coefficient vector $\boldsymbol{\beta}_t^{i,k}$, that is

$$\boldsymbol{\alpha}_1=\underset{\boldsymbol{\alpha}_t^{i,k},\boldsymbol{\beta}_t^{i,k}}{\operatorname{argmin}}\left\|\boldsymbol{f}_1\boldsymbol{\alpha}_1-\boldsymbol{y}_t^{i,k}\right\|_2,$$

$$\text{s.t.}\left\|\boldsymbol{\alpha}_t^{i,k}\right\|_2\leqslant K_1,\ \left\|\boldsymbol{\beta}_t^{i,k}\right\|_2\leqslant K_2, \tag{7}$$

where K_1 and K_2 are non-zero components. Since the dimensionality of the target face is usually very high, efficiently processing high-dimensional feature data is the key to the algorithm. In our work, a diagonal matrix $\boldsymbol{W}$ is used to reduce the dimensionality of the feature space. For sample $\boldsymbol{X}=\{\boldsymbol{\chi}^{i}\in\mathbf{R}^{p}|i=1,2\}$, the joint

sparse approximation method is expressed as

$$(\boldsymbol{\alpha}_1, \boldsymbol{W}) = \arg\min \lambda \left\| \boldsymbol{W}\boldsymbol{f}_1\boldsymbol{\alpha}_1 - \boldsymbol{W}\boldsymbol{y}_t^{i,k} \right\|_2^2 + \gamma G(\boldsymbol{W}, \boldsymbol{X}) + \tau_1 \left\| \boldsymbol{\alpha}_1 \right\|_2^2 + \tau_2 \left\| \operatorname{diag}(\boldsymbol{W}) \right\|_2^2, \quad (8)$$

where $G(\boldsymbol{W}, \boldsymbol{X})$ is the loss function, τ_1 and τ_2 are sparse parameters. If $\boldsymbol{W}_{ii} \neq 0$ denotes that the ith feature is used in the classification, the loss function can be defined as

$$G(\boldsymbol{W}, \boldsymbol{X}) = \exp\left(-\sum_{i=1}^{2} (\boldsymbol{x}_t^i \boldsymbol{w}_t^i) \right), \quad (9)$$

where $\{ \boldsymbol{w}_t^i \in \mathbf{R}^p | i = 1, 2 \}$ is a sparse vector, and the ith feature is selected if $\boldsymbol{W}_i \neq 0$. Then, the value of the loss function is reduced by solving the sparse solution problem as

$$(\boldsymbol{w}_t^i)^* = \arg\min \left\| \boldsymbol{X}\boldsymbol{w}_t^i \right\|_2, \quad s.t. \left\| \boldsymbol{w}_t^i \right\|_2 \leqslant K_0, \quad (10)$$

where K_0 denotes the maximum number of features that can be selected. In order to solve the problem that the traditional sparse representation cannot represent the spatial information of the image block, $\boldsymbol{N}_{\boldsymbol{w}_t^i}(i, j)$ is defined as the jth neighboring feature of the ith feature. Then, the vector set of the ith feature can be expressed as

$$\boldsymbol{y}_t^i = (\boldsymbol{w}_t^i) + \sum_{j=1}^{\tau} \theta_j^2 \boldsymbol{N}_{\boldsymbol{w}_t^i}^2(i, j), \quad i = 1, 2, \cdots, p, \quad (11)$$

where θ denotes the weight of the jth feature. The diagonal matrix $\boldsymbol{W}$ can be expressed as

$$(\boldsymbol{W}_t^i)_{j,j} = \begin{cases} 1, & (\boldsymbol{w}_t^i)_j^* \neq 0, \\ 0, & \text{otherwise}. \end{cases} \quad (12)$$

The spatial relationship of neighboring features can be obtained at the end of the first stage of the sparse representation. To improve the robustness of the face recognition algorithm, the dictionary of the target face needs to be updated periodically. In the process of updating, some feature information that is not related to the target face is usually updated, thus the neighboring features of the features are used to differentiate the face features from the background features, so as to reduce the possibility of the irrelevant features entering into the dictionary and the complexity of the sparse representation process. By using the first stage of sparse coding, a more accurate and complete dictionary can be obtained, and the time spent on face feature selection during face reconstruction can be shortened to some extent. By using the second stage of sparse representation, the values of $\boldsymbol{\alpha}_t^{i,k}$ and $\boldsymbol{\beta}_t^{i,k}$ can be obtained as

$$(\boldsymbol{\alpha}_t^{i,k}, \boldsymbol{\beta}_t^{i,k}) = \arg\min_{\boldsymbol{\alpha}_t^{i,k}, \boldsymbol{\beta}_t^{i,k}} \left\| \boldsymbol{W}\boldsymbol{f}_1\boldsymbol{\alpha}_1 - \boldsymbol{W}\boldsymbol{y}_t^{i,k} \right\|_2, \quad s.t. \left\| \boldsymbol{\alpha}_t^{i,k} \right\|_2 \leqslant K_1, \quad \left\| \boldsymbol{\beta}_t^{i,k} \right\|_2 \leqslant K_2. \quad (13)$$

Defining the non-zero rows of matrix $\boldsymbol{W}$ as $\boldsymbol{W}' \in \mathbf{R}^{K_0 p}$, when $\boldsymbol{f}_1' = \boldsymbol{W}\boldsymbol{f}_1$, $\boldsymbol{y}_t' = \boldsymbol{W}'\boldsymbol{y}_t$, $\boldsymbol{\beta}' = \boldsymbol{W}'\boldsymbol{\beta}$, there is

$$((\boldsymbol{\alpha}_t^{i,k})^*, (\boldsymbol{\beta}_t^{i,k})^*) = \arg\min \left\| \left[\boldsymbol{f}_1', \boldsymbol{W}' \right] \begin{bmatrix} \boldsymbol{\alpha}_t^{i,k} \\ \boldsymbol{\beta}_t^{i,k} \end{bmatrix} - \boldsymbol{y}_t' \right\|, \quad s.t. \left\| \boldsymbol{\alpha} \right\|_2 \leqslant K_1, \left\| \boldsymbol{\beta} \right\|_2 \leqslant K_2, \quad (14)$$

where K_1 and K_2 denote the sparse parameters of sparse coefficient and noise coefficient representation of the target face image when it is subjected to sparse representation, respectively. The appearance model of $\boldsymbol{y}_t^{i,k}$ reconstructed candidate face is defined as

$$\bar{\boldsymbol{f}}_{s,t}^{i,k} = \boldsymbol{f}_{s,t}^{i,k} \boldsymbol{\alpha}_t^{i,k}. \quad (15)$$

After the sparse reconstruction based on the above, the dimensions of the features range from $p \times L$ to $K_0 \times L$, where L denotes the number of templates in the face template dictionary. The reliability of the ith feature at the moment t can be expressed as

$$\hat{\boldsymbol{x}}_t^i = \arg\max p(\boldsymbol{x}_t | \boldsymbol{y}_{1:t}, k_t^i). \quad (16)$$

2.3.3 Face recognition with co-sparse representation

The feature i with high reliability is chosen to represent the steady-state appearance model of face as $\bar{\boldsymbol{f}}_{s,t}^{i,k}$, the dynamic appearance model of face is denoted as $\boldsymbol{f}_{i,t}^k$, and the final classifier for face recognition can be denoted as

$$D_{\text{identy}} = \underset{i}{\operatorname{argmin}} \left\| \boldsymbol{f}_{\text{d}} - \boldsymbol{f}_{i,t}^k \right\|_2^2 + \underset{i}{\operatorname{argmin}} \left\| \boldsymbol{f}_{\text{s}} - \dot{\boldsymbol{f}}_{s,t}^{i,k} \right\|_2^2, \quad (17)$$

where $\boldsymbol{f}_{\text{d}}$ is the face appearance model to be tested for dynamic feature representation, and $\boldsymbol{f}_{\text{s}}$ is the face appearance model to be tested for sparse representation and steady-state feature representation.

2.4 Algorithm description

The basic process of the face recognition algorithm based on collaborative sparse representation of CNN features can be divided into five steps: data input, network learning, feature extraction, face recognition, and output. The implementation of the algorithm can be represented as follows.

1) Input: Training sample $\boldsymbol{X}$ and test sample $\boldsymbol{y}$.

2) Training network $\boldsymbol{F}(\boldsymbol{X})$

2-1. Train the network using CASIA dataset;

2-2. Test the network using LFW dataset;

2-3. Repeat 1-2 until the accuracy is greater than B and stop the network iteration (B=0.97);

2-4. Input the training set $\boldsymbol{X}$ into network $\boldsymbol{F}(\boldsymbol{X})$, get the mapping $\boldsymbol{D}_i = F(\boldsymbol{X}_i)$;

2-5. Further obtain the training dictionary set $\boldsymbol{D}$.

3) Feature extraction

3-1. Extract the Haar features to construct face

dynamic model f_d;.

3-2. Extract the depth feature to construct face steady-state model f_s;

4) Face recognition

4-1. Based on the Haar features nearly H times the mean value of classification results is obtained as $f_{i,t}^k$;

4-2. Reconstruction of f_s using dictionary D, represented by two-stage sparsity as $\bar{f}_{s,t}^{i,k} = f_{s,t}^{i,k} \alpha_t^{i,k}$;

4-3. Obtain final face classification based on Eq. 17;

4-4. Update the dictionary set D based on the categorization result;

5) Output: Face classification results

3 Experimental results and analysis

Experiments were conducted on two commonly used datasets, the CMU PIE face database and the AR face database, corresponding to face recognition in different intraclass variation environments, and the images were uniformly cropped to a size of 64×64 pixels. To verify the effectiveness of the proposed algorithm, Bayesian model recognition method based on gray features (Pixel+Bayesian), sparse representation recognition method based on gray features (Pixel+SRC), Bayesian recognition method based on depth features (DL+Bayesian), sparse representation recognition method based on depth features (DL+SRC), and the collaborative sparse representation recognition method based on depth features (DL+CSRC, CSRDFC) proposed in this study were compared. These methods were implemented using Matlab R2021b programming on a laptop computer with Intel Core i7-8750H @2.20GHz hexa-core, 32G of RAM, and Windows 10 (64-bit).

3.1 Face recognition experiment on CMU PIE face database

The CMU PIE face database covers more than 4 000 face images of 68 people, and the experiments were conducted using the images with intraclass interference containing different poses, lighting, and expressions, totaling 2 792 images. some representative images from the CMU PIE dataset are shown in Fig.5. All images were face-aligned using the open-source library DLIB. 1—25 face images from each class sample are randomly selected as the training set dictionary and the rest of the data as the test set, and 15 experiments were conducted and the results were compared.

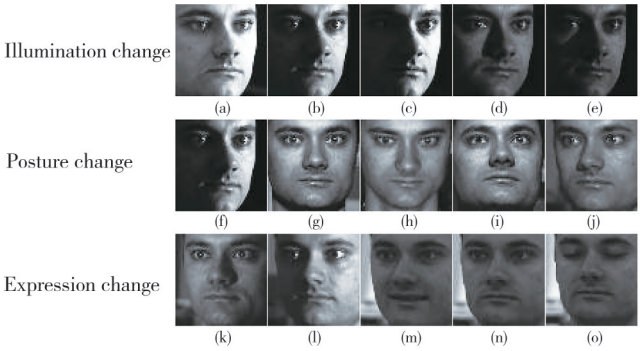

Fig. 5 Examples from CMU PIE face database. (a)-(o), face images with illumination changes; (f)-(j), face images with posture changes; and (k)-(o), face images with expression changes

In the CMU PIE face database, face images are affected by intraclass factors in multiple poses and multiple expressions as well as external factors in multiple illuminations. From the data in Table 2, it can be seen that all the above methods can recognize the target face in the CMU PIE face database to some extent, while our method achieves the highest average recognition accuracy when the dictionary number is taken to be any value from 1 to 8, with nearly 20%—50% improvement in the average recognition accuracy compared to the traditional face recognition methods using grayscale features, and nearly 1%—10% improvement compared to the face recognition methods using deep learning features.

Table 2 Average recognition accuracy on CMU PIE

Method	Number of sample dictionaries per class/%							
	1	2	3	4	5	6	7	8
Pixel+Bayesian	15.37	29.91	35.76	42.51	47.26	50.75	54.08	55.31
Pixel+SRC	24.91	38.71	47.68	54.84	58.95	62.62	65.85	67.15
DL+Bayesian	53.16	68.92	74.15	77.87	81.35	83.73	85.79	85.93
DL+SRC	59.01	77.71	83.47	85.58	88.41	90.69	91.87	92.12
DL+CSRC	**59.12**	**80.11**	**87.23**	**90.56**	**91.13**	**94.87**	**92.88**	**93.01**

Fig.6 shows that the average recognition accuracy of the method using deep learning features is overall higher than that of the method using traditional grayscale features because the deep learning features have better robustness to the interference of intraclass variations in face images. The recognition effect of our method is better than that of the methods using both deep features and sparse representation by first adjusting the structure of sparse representation to collaborative sparse representation, then dividing the process of sparse

representation into two phases to better represents the spatial characteristics of the image, and finally adopting the dictionary updating strategy to further improve the robustness of the classifier.

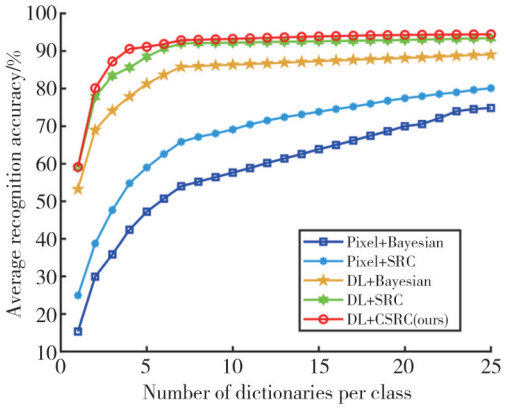

Fig. 6 Average recognition accuracy on CMU PIE face database

3.2 Face recognition experiment on AR database

The AR face database used in this experiment contains 2 600 images from 100 people (50 males, 50 females), and 26 images of each person were collect in different scenarios, including lighting changes, expression changes, and occlusion scenarios, where occlusion includes two types of images with sunglasses and scarves. Some representative images of the AR database are shown in Fig.7. All images are face aligned using the open source library DLIB, with the top 1 — 25 face images of each class of samples are selected as the training set dictionary, and the rest of the data as the test set, and the results of 15 experiments are averaged and compared.

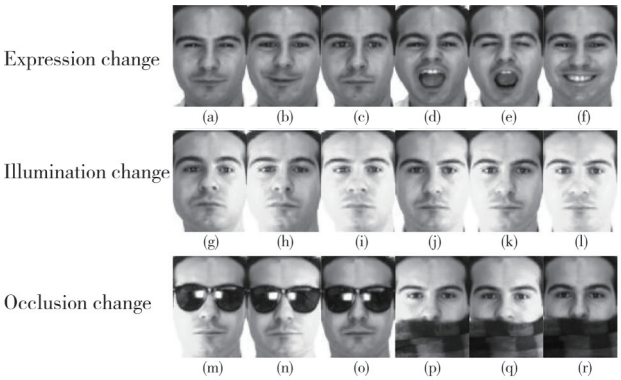

Fig. 7 Examples of AR face database. (a)–(f), face images with expression changes; (j)–(l), face images with illumination changes; and (m)–(r), face images with occlusion changes.

In the AR face database, face images are affected by external factors such as lighting and occlusion, and also by intraclass factors such as expression changes. From the data in Table 3, it can be seen that all the above methods can recognize the target face in the AR face database to some extent, however, our method achieves the highest recognition accuracy when the dictionary number is taken as any value from 1 to 8, with an average recognition accuracy improvement by 25% — 60% compared to the traditional face recognition method using grayscale features, and an improvement by 1% — 12% compared to the face recognition method using deep learning features.

Table 3 Average recognition accuracy on AR face database

Method	Number of sample dictionaries per class/%							
	1	2	3	4	5	6	7	8
Pixel+Bayesian	15.66	23.13	28.61	33.07	36.93	40.86	44.19	48.23
Pixel+SRC	16.37	29.49	39.91	48.17	54.57	59.66	64.08	69.07
DL+Bayesian	34.54	48.29	56.61	65.09	70.82	74.67	77.89	81.29
DL+SRC	41.19	58.83	67.87	75.81	81.07	84.41	86.59	87.97
DL+CSRC	**41.43**	**60.21**	**70.12**	**77.35**	**84.11**	**86.67**	**87.09**	**88.62**

Fig. 8 shows that the features extracted by the deep learning algorithm have better distinguishability and better scene robustness than the grayscale features under the influence of factors such as occlusion, illumination, and expression changes. The recognition effect of our method is better than that of the traditional sparse representation based on deep features because it uses two stages of sparse representation to construct dynamic and steady-state classifiers respectively, in which the dynamic face classifier based on the Haar feature has better characteristics for the target's partial occlusion. Based on the above analysis, it can be verified that our method is effective.

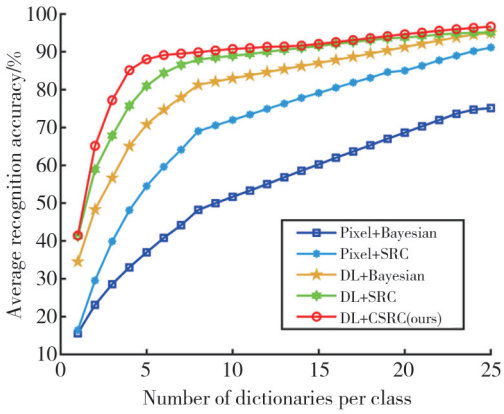

Fig. 8 Average recognition accuracy on AR face database

3.3 Analysis of model complexity

For the complexity of the algorithms in this study, a statistical analysis was conducted, all experimental models were implemented in Matlab R2021b, and the GPU was NVIDIA GeForce GTX 1050 Ti. The running results and specific metrics of different methods were used to verify the time complexity of the methods in this study. Here, four typical methods were selected: Pixel+Bayesian, Pixel+

SRC, DL+Bayesian, and DL+SRC. The running time of the experiments were compared on CMU PIE and AR face databases, where the running time includes the time for dictionary learning, model construction, and face detection, etc. The experimental results are shown in Table 4.

Table 4 Model complexity on CMU PIE and AR databaes

Method	CMU PIE /s		AR /s	
	Dict(8)	Dict(10)	Dict(8)	Dict(10)
Pixel+Bayesian	53.63	54.65	56.32	57.31
Pixel+SRC	173.34	190.23	76.52	90.13
DL+Bayesian	63.21	66.33	53.45	58.81
DL+SRC	192.33	208.42	132.43	151.16
DL+CSRC	161.70	176.20	71.29	89.11

The overall sample size of the CMU PIE database is higher than that of the AR database, so the overall overhead of the model on the AR database is relatively small. Methods with sparse representations require the computation of L1 paradigms, and the complexity of the Pixel+Bayesian and DL+Bayesian methods performs better than that of Pixel+SRC, DL+SRC, and DL+CSRC (our methods). The methods using deep learning features need to train the network in advance, and the complexity of DL+SRC and DL+CSRC methods is higher than that of the traditional sparse representation method Pixel+SRC. Our method adopts a two-stage collaborative sparse representation model to dynamically select the classifiers, and the algorithm running time overhead is lower than that of the Pixel+SRC and DL+SRC methods, and our method performs optimally in terms of overall performance.

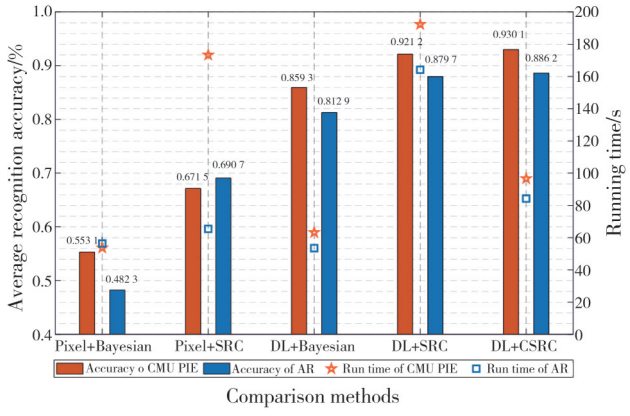

Fig. 9 Average recognition rate and running time on CUM PIE and AR databases(the number of dictionaries is 8)

As shown in Fig. 9, compared with the other four methods, our method can achieve the best recognition results in case that the number of training dictionaries is 8. In terms of the running time, because our method adopts the collaborative sparse representation model to dynamically select the classifier, its running time falls in between that of the ordinary classification method and that of the sparse representation method. Since the best recognition results can be obtained in a shorter running time, our method has a certain degree of practicality.

4 Conclusions

In the study, Haar features were extracted, which are robust to face occlusion, along with deep learning-based features during the feature extraction stage to construct a face model. At the classifier construction stage, the limitations of traditional sparse representation, which failed to fully capture the spatial characteristics of images, were addressed. A collaborative sparse model was proposed that utilized both dynamic and steady-state features to effectively model the target. These two types of features were used to build separate classifiers. As the face recognition environment changed, the weights of these features were dynamically adjusted during the recognition process. Additionally, the face dictionary was continually updated using an incremental method based on current classification results.

The experimental results demonstrated that the method achieved better recognition accuracy and robustness than traditional face recognition methods, particularly in complex environments with challenges like occlusion, lighting changes, expression variations, and more, even with a small number of samples. However, the structure of the proposed method was complex. Future research will focus on simplifying the structure of the method and addressing face recognition issues in noisy environments.

Acknowledgement

We are grateful for the financial support from Natural Science Foundation of Gansu Province (Nos. 22JR5RA217, 22JR5RA216), Lanzhou Science and Technology Program (No. 2022-2-111), Lanzhou University of Arts and Sciences School Innovation Fund Project (No. XJ2022000103), and Lanzhou College of Arts and Sciences 2023 Talent Cultivation Quality Improvement Project (No. 2023-ZL-jxzz-03). We thank all the editors and reviewers who provided valuable comments on the article.

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] CHAUDHRY A, ELGAZZAR H. Design and implementation of a hybrid face recognition technique//2019 IEEE 9th Annual Computing and Communication Workshop and Conference, January 7-9, 2019, Las Vegas, NV, USA. New York: IEEE, 2019: 384-391.

[2] DONOHO D L. Compressed sensing. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[3] WRIGHT J, YANG A Y, GANESH A, et al. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227.

[4] SUN Y B, LIU Q S, TANG J H, et al. Learning discriminative dictionary for group sparse representation. IEEE Transactions on Image Processing, 2014, 23(9): 3816-3828.

[5] ZAMANI F, JAMZAD M, RABIEE H R. Atom specific multiple kernel dictionary based sparse representation classifier for medium scale image classification. Journal of Visual Communication and Image Representation, 2021, 79: 103228.

[6] YAN C M, ZHANG Y Y, ZHANG D. Face recognition based on improved label consistent KSVD dictionary learning. Computer Engineering & Science, 2022, 44(2): 291-297.

[7] DI L, JIE H W, LIANG J Z. Sparse integrated dictionary learning for small-sample face recognition. Journal of Intelligent Systems, 2021, 16(2): 218-227.

[8] MA X, ZHANG P D, FENG J F. Sparse representation via deep learning features based face recognition method. CAAI Transactions on Intelligent Systems, 2016, 11(3): 279-286.

[9] HUANG X Y, LIU F, BAO Q Y, et al. A face correction recognition method for generative adversarial networks based on multi-task learning and identity constraints. Journal of Electronics, 2023(10): 2936-2949.

[10] HU L R, LI X, TAN K. Mask face recognition algorithm based on improved attention mechanism. Computer Simulation, 2023, 40(7): 180-183.

[11] LI X S, DING W J, FANG Y, et al. Occluded face recognition based on image depth prior and robust markov random field. Computer Science, 2024, 51(7): 244-256.

[12] WANG J J, SONG S H, DU Y M. Simulation of local blur face image recognition considering Gabor wavelet features. Computer Simulation, 2023, 40(9): 237-241.

[13] LIU X M, ZHAO H P. Research on face recognition under illumination variation based on Retinex algorithm. Laser Journal, 2023, 44(7): 218-222.

[14] ZUO M, XIE S G, ZHANG X, et al. DOA estimation based on weighted l_1-norm sparse representation for low SNR scenarios. Sensors, 2021, 21(13): 4614.

[15] NAKACHI T, KIYA H. L0 norm optimization in scrambled sparse representation domain and its application to EtC system. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, 2020, 103(12): 1589-1598.

[16] BLAUCH N M, BEHRMANN M, PLAUT D C. Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. Cognition, 2021, 208: 104341.

[17] BHASHA A V, VENKATRAMANA REDDY B D. Automated image super resolution with the aid of activation function optimized deep CNN and adaptive wavelet lifting approach. International Journal of Image and Graphics, 2022, 22(5): 2250046.

[18] SONG W F, LI S, LIU J, et al. Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. IEEE Journal of Biomedical and Health Informatics, 2019, 23(3): 1215-1224.

[19] SUN Z, CHIONG R, HU Z P, et al. Deep subspace learning for expression recognition driven by a two-phase representation classifier. Signal, Image and Video Processing, 2020, 14(3): 437-444.

[20] GUO J M, YANG J S, SESHATHIRI S, et al. A light-weight CNN for object detection with sparse model and knowledge distillation. Electronics, 2022, 11(4): 575.

[21] DU H, SHI H L, ZENG D, et al. The elements of end-to-end deep face recognition: a survey of recent advances. ACM Computing Surveys, 2022, 54(10s): 1-42.

[22] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[23] HAN C R, SHAN S G, KAN M N, et al. Personalized convolution for face recognition. International Journal of Computer Vision, 2022, 130(2): 344-362.

[24] HAMMOUCHE R, ATTIA A, AKHROUF S, et al. Gabor filter bank with deep autoencoder based face recognition system. Expert Systems with Applications, 2022, 197: 116743.

[25] NGUYEN T S, LUONG M, KAANICHE M, et al. A novel multi-branch wavelet neural network for sparse representation based object classification. Pattern Recognition, 2023, 135: 109155.

[26] AKHAVAN S, BAGHESTANI F, KAZEMI P, et al. Dictionary learning for sparse representation of signals with hidden Markov model dependency. Digital Signal Processing, 2022, 123: 103420.

[27] LIAO M M, GU X D. Face recognition based on dictionary learning and subspace learning. Digital Signal Processing, 2019, 90: 110-124.

[28] YANG S C, WEN Y, HE L H, et al. Sparse common feature representation for undersampled face recognition. IEEE Internet of Things Journal, 2021, 8(7): 5607-5618.

[29] KONG Y, WANG T Y, FENG Z P, et al. Discriminative dictionary learning based sparse representation classification for intelligent fault identification of planet bearings in wind turbine. Renewable Energy, 2020, 152: 754-769.

[30] HUBEL D H, WIESEL T N. Receptive fields, binocular interaction and functional architecture in the cat′s visual cortex. The Journal of Physiology, 1962, 160(1): 106-154.
[31] WAIBEL A, HANAZAWA T, HINTON G, et al. Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37(3): 328-339.
[32] YANG X L, LI M, ZHAO S L, et al. Facial expression recognition algorithm based on CNN and LBP feature fusion//3rd International Conference on Robotics and Artificial Intelligence, December 29-31, 2017, Shanghai, China. New York: ACM, 2017: 33-38.

基于CNN特征的协同稀疏表示人脸识别算法

赵世林[1,2*], 徐成俊[1], 刘昌荣[1]

1. 兰州文理学院 数字媒体学院, 甘肃 兰州 730010;
2. 兰州交通大学 自动化与电气工程学院, 甘肃 兰州 730070

摘　要： 针对传统稀疏表示模型在多种复杂环境因素影响下算法精度不高的问题，从特征提取和模型构建两方面加以改进。首先，利用训练好的深度学习网络提取人脸的卷积神经网络（Convolutional neural network，CNN）特征。其次，分别基于CNN特征和哈尔（Haar）特征构建人脸识别稳态和动态分类器：稳态分类器构建过程中引入两阶段稀疏表示，然后从CNN特征构建的稀疏表示模板字典集中动态选取可靠性高的特征模板作为备选模板。最后，基于稳态分类器和动态分类器的分类结果共同给出人脸识别结果，并根据识别结果实时调整稳态分类器模板特征权重，动态更新字典集，降低无关特征进入字典集的概率。实验结果表明，该方法在CMU PIE人脸库上的平均识别准确率为94.45%，在AR人脸库上的平均识别准确率为96.58%。相比传统人脸识别方法，其识别准确率显著提高。

关键词： 稀疏表示；深度学习；人脸识别；字典更新；特征提取

引用格式： ZHAO Shilin, XU Chengjun, LIU Changrong. Face recognition algorithm using collaborative sparse representation based on CNN features. Journal of Measurement Science and Instrumentation, 2025, 16(1): 85-95. DOI: 10.62756/jmsi.1674-8042.2025009