A probabilistic collaborative dictionary learning-based approach for face recognition

Shilin Lv\(^1\) | Jiuzhen Liang\(^1\) | Lan Di\(^2\) | Xia Yunfei\(^3\) | ZhenJie Hou\(^1\)

\(^1\) School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu 213164, China
\(^2\) School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu 214122, China
\(^3\) Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, North Carolina, United States

Correspondence
Jiuzhen Liang, School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu 213164, China.
Email: jzliang@cczu.edu.cn

Abstract
Although Sparse Representation based Classifier (SRC), a non-parametric model, can obtain an interesting result for pattern recognition, a reasonable interpretation has been lacked for its classification mechanism. What is more, the training samples are used as off-the-shelf dictionary directly in SRC, which can make the feature hidden in the training samples hard be extracted. At the same time, the complexity of the algorithm is increased because of too many atoms of the dictionary. The authors first explains in detail the classification mechanism of SRC from the view of probabilistic collaborative subspace and offer the process to improve the stability of the algorithm using the joint probability in the case of the multi-subspace. Then, the authors introduce the dictionary learning (DL) and Fisher criterion into the model to further enhance the discrimination of the coding coefficient. In order to ensure the convexity of the discrimination term and further enhance the discrimination, the authors add the $L_{21}$-norm term into the Fisher discrimination term and offer the proof for its convexity. Finally, the experimental result on a series of benchmark databases, such as AR, Extended Yale B, LFW3D-hassner, LFW3D-sdm and LFW3D-Dlib, show that PCDDL outperforms existing classical classification models.

1 INTRODUCTION

Recently sparse representation techniques have led to interesting results in pattern classification, such as face recognition, handwritten digit recognition, etc. The success of sparse representation—based classifier should attribute to the fact that a high-dimensional image can be represented or coded by some representative samples from the same class in a low-dimensional manifold, and the recent progress of $L_0$-norm and $L_1$-norm minimization techniques.

In general, the sparse representation schemes are usually divided into two categories. The first one is the non-parametric methods, such as SRC [10, 16], which can predict the query samples using the training samples directly and do not need to train a parametric model. Another, on the contrary, is the parametric methods, such as dictionary learning (DL) [24], which aim to train a parametric model used for predicting the testing samples.

For the non-parametric model, the classifier based on the distance is used widely in a series of vision recognition tasks (e.g. the nearest subspace classifier [25]). However, non-parametric model based on distance depends heavily on the pre-defined distance or similarity metric criterion. The commonly used metrics (e.g. Euclidean distance, manifold distance and principal angle-based correlation) can describe intuitively the variations among samples, but it cannot reflect accurately the intrinsic similarity among objects. To better characterize the similarity, one appropriate choice is to introduce indeterminacy into the output of the classifier, such as probabilistic Support Vector Machines (SVM) [17–19]. The other choice that can replace the probabilistic SVM is the probabilistic subspace method, such as probabilistic principal component analysis [20, 22] and probabilistic linear discrimination analysis [21], which reformulate the subspace method as a possible variable model and optimize the parameter by maximum likelihood estimation.

Besides, a critical problem is how to characterize the testing samples for the distance-based non-parametric classifier. For SRC, the linear combination of training samples is used to represent the testing samples and $\ell_1$-norm regularization term is employed to impose sparsity constraint. Inspired by the probabilistic subspace methods [1], this paper analyses the intrinsic mechanism of SRC, and proposes probabilistic collaboration subspace-based sparsity representation classifier model.
Figure 1  The basic framework in the paper

Though SRC shows promising performance in face recognition tasks, the dictionary used in it may not be effective enough to characterize the testing samples because of the potential indeterminacy and noise in training samples. What is more, using the original training samples as the dictionary can be hard to extract the hidden discriminative information [39] in the training samples. Likewise, using off-the-shelf bases [14] as dictionary, such as Haar wavelets and Gabor wavelets, may be universal to all kinds of images but will not be appropriate enough for the images of the specific types. Fortunately, all the problems mentioned above can be solved through learning a proper parametric dictionary from the training samples. Hence, Yang et al. [6] propose that one can get more promising performance by learning a class sub-dictionary from each class samples. Such sub-dictionary, however, might not capture the similarity feature among various classes.

Therefore, we introduce the DL [26, 35–37] into the probabilistic collaboration subspace-based SRC (PCSRC) to get Probabilistic Collaboration subspace-based Dictionary Learning (PCDL) model. To improve the effectiveness of DL, Yang et al. [5] and Ramirez et al. [29] propose to enhance the discrimination and the reconstructive ability of the dictionary through learning a class sub-dictionary and adding the irrelevant enhancement items. Yang et al. introduce Fisher discrimination criterion into the DL to get the Fisher Discrimination Dictionary Learning (FDDL) [2, 13, 15] model and obtain promising results. But, due to the intrinsic deficiency in the residual term, FDDL could not further promote the performance of the classifier. Inspired by this, we introduce the Fisher discrimination criterion into the PCDL model to get the Probabilistic Collaboration subspace-based Discrimination Dictionary Learning (PCDDL) model and apply the model to final classification tasks. In brief, the main contributions in our study are as follows:

(1) Our study analyses the classification mechanism of SRC in detail from the perspective of probabilistic collaboration subspace, provides the processing procedure using joint probability to improve the stability of the algorithm in the case of multi-subspace and proposes the PCSRC model.

(2) We first introduce the DL into the PCSRC model to get PCDL model, then add the Fisher discrimination term into the PCDL model to obtain the PCDDL model used for the final classification tasks. At the same time, to ensure the convexity of the Fisher discrimination term, we introduce the term $\|X\|_{2,1}$ into it, and we also offer the proof of the convexity.

The rest paper is organized as follows: Section 2 briefly introduces the concept of probabilistic collaboration subspace in [1], and analyses the classification mechanism of SRC from the perspective of probabilistic collaboration subspace. Section 3 provides the processing procedure to improve the stability of the algorithm in the case of multi-subspace. Section 4 adds the DL and Fisher discrimination term to obtain the PCDDL model. Section 5 conducts experiments and analyses the results. Section 6 concludes the paper. The basic framework in the paper is illustrated in Figure 1.

2  |  RELATED WORK

2.1  |  Probabilistic collaboration subspace

Given a training sample sets $X = \{X_1, ..., X_k\}$, where $X_k, k = 1, 2, ..., K$ is the sample matrix of class $k$, and each column in $X_k$ means a sample vector. Let $\mathcal{L}_X$ be the label sets of all the class in $X$ and denote by $\mathcal{S}$ the linear collaboration subspace spanned by all the training samples. So, every data point $x$ in subspace $\mathcal{S}$ can be represented as a linear combination of the samples in $\mathcal{X}$, i.e. $x = X \alpha$, where $\alpha$ is the coding coefficient vector.

In general, different norms correspond to different distances, and the $L_1$ norm is a Manhattan distance (also known as the chessboard distance). As shown in Figure 2, the shortest path...
2.2 Probabilistic representation of samples outside the collaborative subspace

Because the testing sample $y$ usually lies on the outside of the subspace $S$. In a bid to measure the probability that the label of sample $y$ belongs to $l_k$, i.e. $P(l(y) \in l_k)$, Cai et al. proposed to compute the probability as follows:

$$P(l(y) \in l_k) = P(l(\infty) = l(y)|l(\infty) \in l_k) \cdot P(l(\infty) \in l_k),$$

(2)

where $P(l(\infty) = l(y)|l(\infty) \in l_k)$ can be measured by the similarity between $y$ and $x$. And Cai et al. chose the Gaussian kernel to define the similarity.

According to Equations (1)–(2), the task to maximize the probability $P(l(y) \in l_k)$ will convert to the following object:

$$\max P(l(y) \in l_k) = \min_{\alpha} \|y - X\alpha\|_2^2 + \lambda \|\alpha\|_1,$$

(3)

where $\lambda$ is a constant. We can see that the equation above offer a probabilistic representation that $y$ lies in subspace $S$, and provide a reasonable interpretation from the view of probability.

3 PROBABILISTIC COLLABORATION SUBSPACE-BASED SRC

3.1 Probability to each class-specific subspace

A sample $x$ within the subspace $S$ can be collaboratively represented as $x = X\alpha = \sum_{k=1}^{K} X_k \alpha_k$, $\alpha = [\alpha_1; \alpha_2; \ldots ; \alpha_k]$, where $\alpha_k$ is the coding vector related to $X_k$. Note that $x_k$ is a sample point in the subspace of class $k$, i.e. $x_k = X_k \alpha_k$. Here we, using Gaussian kernel, define the probability that $x$ belongs to the specific class $k$:

$$P(l(\infty) = k | l(\infty) \in l_k) \propto \exp \left(-\delta \|x - X_k \alpha_k\|_2^2\right),$$

(4)

where $\delta$ is a constant.

For the testing sample $y$ outside the subspace, the probability $P(l(y) = k)$ can be calculated as follows:

$$P(l(y) = k) = P(l(y) \in l_k) \cdot P(l(\infty) = k | l(\infty) \in l_k)$$

$$\propto \exp \left(-\left(\|y - X\alpha\|_2^2 + \lambda \|\alpha\|_1 + \gamma \|X\alpha - X_k \alpha_k\|_2^2\right)\right),$$

(5)

where $\gamma$ is a constant.

3.2 PCSRC model

To determine the label of $y$ by checking which subspace the sample $y$ belongs to will make the method become instability...
because of change of the data point \( x \) which is from \( S \) and corresponds to the sample \( y \).

Therefore, it is critical that we need to find a common data point \( x \) from \( S \) and maximize the joint probability \( P(l(y) = 1, \ldots, l(y) = K) \). Then, the class label of \( y \) can be determined by checking the largest probability. By assuming the events \( l(y) = k \) are independent, we have

\[
\max P(l(y) = 1, \ldots, l(y) = K) = \max \prod_{k=1}^{K} P(l(y) = k). \tag{6}
\]

Applying the logarithmic operator to Equation (6) and ignoring the constant term, we have

\[
(\tilde{\alpha}) = \arg \min_{\alpha} \left\{ \|y - X\alpha\|_2^2 + \lambda \|\alpha\|_1 + \gamma \sum_{k=1}^{K} \left( \|X\alpha - X_k\alpha_k\|_2^2 \right) \right\}. \tag{7}
\]

In Equation (7), the first two terms \( \|y - X\alpha\|_2^2 \) and \( \lambda \|\alpha\|_1 \) are collaboration representation term, which aims to find a data point \( x = X\alpha \) from collaboration subspace \( S \) as close to \( y \) as possible. The last term \( \sum_{k=1}^{K} \left( \|X\alpha - X_k\alpha_k\|_2^2 \right) \) is used to find a data point \( X_k\alpha_k \) from every class subspace \( S_k \) as close to common data point \( x \) as possible. The parameter \( \lambda \) and \( \gamma \) balance the role of the three terms, which can be set based on the prior knowledge of the problem.

## 4 Probabilistic collaboration criterion-based PCDL model (PCDDL)

### 4.1 Probabilistic collaboration subspace-based DL (PCDL) model

As mentioned above, in a bid to ensure the stability of the classification results, we need to maximize the joint probability \( P(l(y) = 1, \ldots, l(y) = K) \) in PCSRC model, which results from that the unknown label testing samples, in its classification rule, are represented by the dictionary constructed by the training samples. Instead, PCDL needs to learn a parametric dictionary and the samples used for training dictionary have known labels, so we do not need to use joint probability anymore, but use the single probability \( P(l(y) = k) \). The specific expression of PCDL model is formulated as follows:

\[
J_{l(X)} = \arg \min_{\alpha, \lambda, \gamma} \left\{ \sum_{i=1}^{K} \left( \|A_i - DX_i\|_F^2 \right) + \gamma \|DX_i - D_X\|_2 \right\}, \tag{8}
\]

where \( D \) is the dictionary trained from training samples, \( X_i = [X_i^1, X_i^2, \ldots, X_i^K] \) is the coding coefficient and \( X_i^k \) is the coding coefficient related to \( D_k \).

### 4.2 Probabilistic collaboration subspace–based discrimination dictionary learning (PCDDL)

To improve the effect of the DL, we introduce Fisher discrimination criterion into PCDL model to get PCDDL model and use it for face recognition.

#### 4.2.1 The PCDDL model

As mentioned above, the specific expression of PCDDL model is formulated as follows:

\[
J_{(D,X)} = \arg \min_{\alpha, \lambda, \gamma} \left\{ \sum_{i=1}^{K} \left( \|A_i - DX_i\|_F^2 \right) + \gamma \|DX_i - D_X\|_2 \right\}. \tag{9}
\]

with

\[
f(X) = (1 - w) \text{tr} (S_W^{-1}X - S_B^{-1}X) + w \|X\|_{2,1}, \tag{10}
\]

where

\[
\gamma = t_1 \cdot \min \left\{ \frac{\|A_i - DX_i\|_F^2 \cdot \|DX_i - D_X\|_2^2}{\|DX_i - D_X\|_2^2} \right\} \tag{11}
\]

and

\[
w = t_2 \cdot \frac{\text{tr} (S_W^{-1}X - S_B^{-1}X)}{\|X\|_{2,1}} \tag{12}
\]

are the adaptive hyperparameter. Equation (11) defines the hyperparameter of the residual term in the model in Equation (9), and the values of the two terms in the residual term change dynamically during iteration. Since it is the first term that plays the leading role, in each iteration, we take the ratio of the smaller term to the larger term as the balance factor (hyperparameter), so as to dynamically adjust the balance of the two terms, where \( T1 \) is a constant. Equation (12) is same as Equation (11). In \( f(X) \), the first term is the Fisher discrimination term and the second term is \( \|X\|_{2,1} \). The term \( \|X\|_{2,1} \) cannot only to ensure the convexity of the first term, but to enhance the effect of clustering. And residual term \( \|A_i - DX_i\|_F^2 + \gamma \|DX_i - D_X\|_2^2 \) have more intuitive interpretation in the two-dimensional vector space as follows: As shown in Figure 4, where \( A_i \) is
the training samples $A_j \approx DX_j = D_jX_j^1 + \cdots + D_jX_j^K = P_l + \cdots + P_l + \cdots + P_l$.

Figure 4(a) denotes the first term $\|A_i - DX_i\|^2_F$, which is just to ensure that the samples $A_j$ could be promisingly reconstructed by $D$. And the other in Figure 4 means the second term $\|DX_i - D_jX_j\|^2_F$, which requires $DX_i = \sum_{j=1}^K P_j \approx P_l$, to enhance the discrimination of $D$. The proof of the clustering term convexity is as follows:

Proof. Let $E_i^j = \{I\}_{i \times j}$ be a matrix of size $n_i \times n_j$ with all entries being 1, and let $N_j = I_{n_i \times n_j} - E_i^j \setminus n_j$, $P_i = E_i^j \setminus n_j$, and $C_i^j = E_i^j \setminus n$.

From $f_i(X_i) = \|X_i - M_i\|^2_F - \sum_{k=1}^C ||M_k - M_i\|^2_F + \eta \|X_i\|_{21}$, we can derive that

$$f_i(X_i) = \|X_i - M_i\|^2_F - \sum_{k=1}^C ||M_k - M_i\|^2_F + \eta \|X_i\|_{21},$$

(13)

where $G = \sum_{k=1}^C X_kC_k^T = \sum_{j=1}^C X_j E_j^C \setminus n_j - \sum_{j=1}^C X_jC_j^C \setminus n_j$.

Rewrite $X_i$ as a column vector, $X_i = [r_{i1}, r_{i2}, ..., r_{id}]^T$, where $r_{ij}$ is the $j$th row vector of $X_i$, and $d$ is the total number of row vectors in $X_i$. Then $f_i(X_i)$ equals to

$$f_i(X_i) = \|\text{diag}(N_i) X_i\|^2 - \|\text{diag}(P_i^T) X_i - \text{diag}(G^T)\|^2_2$$

$$- \sum_{k=1,k\neq i}^C \|\text{diag}(C_k^T) X_i - \text{diag}(Z)\|^2_2 + \eta \|X_i\|_{21},$$

(14)

where diag$(T)$ is to construct a block diagonal matrix with each block on the diagonal being matrix $T$.

The convexity of $f_i(X_i)$ depends on whether its Hessian matrix $\nabla^2 f_i(X_i)$ is positive definite [29] or not. $\nabla^2 f_i(X_i)$ will be positive definite if the following matrix $S$ is positive definite:

$$S = N_iN_i^T - \left(P_iP_i^T + \sum_{k=1,k\neq i}^C C_k^T C_k^T\right) + \eta \Sigma,$$

(15)

where $\Sigma = \text{diag}(1/\|r_{ij}\|_2)$, $j = 1, 2, ..., d$. So if $S$ is positive definite, then $\nabla^2 f_i(X_i)$ is also positive definite.

**ALGORITHM 1** Optimization of PCDDL Model

**Input:** A matrix of training samples $A = \{A_1, A_2, ..., A_K\} \in \mathbb{R}^{n \times N}$ for $K$ classes.

1: Initialize the $D = \{D_1, D_2, ..., D_K\}$ using the eigenvector of the training samples $A$.

2: repeat

3: Fix $D$, update $X_i$ class by class (i.e. fix all $X_j$, $j \neq i$, calculate $X_i$) via the following objective

$$\min_{X_i} \left\{\|A_i - DX_i\|^2_F + \eta\|DX_i - D_jX_j\|^2_F + \lambda_1 \|X_i\|_1 + \lambda_2 f_i(X_i)\right\}$$

$$\|X_i - M_i\|^2_F - \sum_{k=1}^C \|M_k - M_i\|^2_F + \eta \|X_i\|_{21}$$

4: Fix $X_i$, update $D_i = (d_1, d_1, ..., d_i)$ class by class via the following objective

$$\min_{D_i} \left\{\|Z_i - D_iX_i\|^2_F + \eta\|Z_i - D_jX_j\|^2_F\right\}Z_i = A_i - \sum_{j=1,j\neq i}^K D_jX_j,$$

for $i = 0$ to $n$ do

5: $Y_i = Z_i - \sum_{j=1,j\neq i}^K d_jX_j^T V_0 = Z_i - \sum_{j=1,j\neq i}^K d_jX_j^T$

6: $\|Y_i - d_iX_i^T V_0\|^2_F + \|V_0 - d_iX_i\|^2_F + tr[(V_0 - d_iX_i)(V_0 - d_iX_i)^T]$

7: $d_i = \operatorname{argmin}_{d_i} \left(\sum_{j=1,j\neq i}^K d_jX_j^T V_0^T (V_0 - d_iX_i) + (V_0 - d_iX_i)^T (V_0 - d_iX_i)^T\right)$

8: $\eta = \frac{\sum_{j=1,j\neq i}^K d_jX_j^T V_0^T (V_0 - d_iX_i)^T}{\|V_0 - d_iX_i\|^2_F}$

9: $d_i = \operatorname{argmax}_{d_i} \left(\sum_{j=1,j\neq i}^K d_jX_j^T V_0^T (V_0 - d_iX_i) + (V_0 - d_iX_i)^T (V_0 - d_iX_i)^T\right)$

10: end for

11: until maximum iterations or convergence.

**Output:** $D, X$

After some derivations, we have

$$\frac{1}{\eta} = \frac{1}{\eta} - \frac{1}{\eta} E_i^j \left(\frac{2}{n} - \frac{2}{n} + \sum_{k=1}^n \frac{n_k}{n}\right) + \Sigma$$

(16)

is positive definite.

Because the maximal eigenvalue of $E_i^j$ is $n_i$, we should ensure

$$\frac{1}{\eta} - \frac{1}{\eta} E_i^j \left(\frac{2}{n} - \frac{2}{n} + \sum_{k=1}^n \frac{n_k}{n}\right) > 0.$$ 

(17)

For $n = n_1 + \cdots + n_s$, we have $1/\eta - (n - n)/\eta > 0$. Since $n_i - n < 0$, So $\eta < 0$, i.e. $f_i(X_i)$ will be convex to $X_i$ as long as $\eta < 0$.

4.2.2 Optimization

Since the objection is convex to $D$ or $X$, respectively, so the optimization process alternatively updating $D$ or $X$ can be designed. Here we employ the method in literature [6, 7, 9, 34]. Detailed algorithm steps are shown in Algorithm 1.

4.2.3 The classification approach of PCDDL

For a given testing sample, we first reconstructed the sample by the dictionary $D$ learned from Equation (9), then we get a
coding coefficient $\hat{\alpha}$, i.e.

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

(18)

where $\lambda$ is a constant. Here we define the following rule:

$$l(y) = \arg \min_i e_i = \arg \min_i \| D\hat{\alpha} - D_i\hat{\alpha}_i \|_2^2.$$  

(19)

5 EXPERIMENTAL RESULTS AND ANALYSIS

In this part, our main task has three aspects. First of all, we discussed whether the $L_1$-norm or the $L_2$-norm is a better constraint on coding coefficients. We also verify the convergence of the terms in the model. Second, we discuss the effect of the number of dictionary atoms on model performance. Finally, we further validated the model's effectiveness on AR, Extended Yale B and LFW databases.

As can be seen from the definition of the $L_2$-norm, its sensitivity comes from the square operation, allowing elements with large values to dominate the final result. This means that the $L_2$-norm is more sensitive to image changes than the $L_1$-norm. In addition, the $L_1$ norm can lead to certain sparsity, which is helpful to improve the discriminability of the model. At the same time, the sparsity caused by the $L_1$-norm of coding coefficient $X$ does not affect its cooperative representation. As shown in Figure 5, by comparing the two sub-graphs, it can be seen that under the $L_1$-norm constraint, the energy of the encoding vector of the sample in the dictionary space is more concentrated in the subspace of its category, while under the $L_2$-norm constraint, there is no such effect. As can be seen from Figure 6, the recognition rate of the model has a smaller error when the coding coefficient is constrained by the $L_1$-norm.

To intuitively observe the convergence of the PCDDL model, we draw the changing curve of objective function and its each term in Figures 8 and 9. We can see that the objective function and its each term gradually converge with iteration. As shown in Figure 7, PCDDL has stronger convergence than classical FDDL. That is, in the same circumstances, PCDDL employs smaller convergence value than FDDL.

In order to test the performance of PCDDL for face recognition, we conduct the experiment in AR, Extended Yale B and LFW, and compare the results with classical methods, such as SRC [10, 23], Collaborative Representation based Classification (CRC) [16], Nearest Neighbour (NN), SVM, Discriminative KSVD (DKSVD), Dictionary Learning with Structure incoherence (DLsI), FDDL, Support Vector Guided Dictionary Learning (SVGDL) [4]. All the experiments are performed on a computer with an Intel Core i5-8500, 8GB RAM and 64-bit Windows 10 operating system.
For parameter selection, if there is no specific note, the hyperparameters, including $\lambda_1, \tau_1, \tau_2, \lambda_2$, are set to $\lambda_1 = 0.005$, $\tau_1 = 0.05, \tau_2 = 0.05, \lambda = 0.005$. And we employ fivefold cross-validation method to refrain from overfitting.

Besides, there is a special and critical parameter in the DL, i.e. the number $a_i (i = 1, 2, ..., K)$ of atoms in the sub-dictionary $D_i$. We use SRC as the baseline method, and analyse the influence of $a_i$ on the performance of PCDDL. We take face recognition tasks on LFW as an example. SRC directly uses the training samples as the dictionary and we randomly select $a_i$ training samples used as the $i$th sub-dictionary. Since the recognition rate will be various due to the diverse sub-dictionary. Therefore, we conduct 10 experiments to get the average recognition rate of SRC. As illustrated in Figure 10, we can see that in all cases PCDDL
and FDDL have about 2% improvement over SRC and SVDL. Especially, PCDDL with $a_i = 3$ can still have higher recognition rate than SRC with $a_i = 8$. Besides, with varying the number of atoms from 3 to 8, PCDDL’s recognition rate fluctuates by around 2%, while that of SRC is 13%, which demonstrates that PCDDL is able to compute a compact and representative dictionary, and has higher recognition rate than SRC.

I. AR: AR database [8] consists of more than 4000 frontal-face images of 126 individuals. For each person, take 26 photos in two different stages. As in [2], we selected a subset of 50 male and 50 female subjects for the experiment. For each subject, seven images from session 1 with changes in lighting and expression were used for training, and another seven images from session 2 with the same conditions were used for testing. An example in AR is shown in Figure 11. The size of the original face image is 60 × 43. A comparison of the competitive methods is shown in Table 1. We can see that PCDDL is at least 0.5% better than other methods.

II. Extended Yale B: The expanded Yale B database includes 2414 frontal facial images from 38 people (about 64 for each subject), taken under a variety of laboratory-controlled lighting conditions. For each subject, we randomly selected 20 images for training and the remaining images for testing (the experimental setup was more difficult than [2]). All the images are normalized 54 × 48. A comparison of the competing methods is shown in Table 2. It can be seen that PCDDL is at least
The recognition rates of various methods on the LFW3D-sdm and LFW3D-hassner, as well as draw the boxplot of the recognition rates with varying number of atoms. As can be seen from Table 14, when the number of atoms is 5, the length of the box is relatively short, which proves that the results have smaller variance than that with other number of atoms. Meanwhile, the height of the red line in the box is relatively high, which proves that the results employ a larger mean than that with other number of atoms. In short, when the number of atoms is 5, the recognition rate has larger average, smaller variance and fewer outliers than with other numbers of atoms. Therefore, the number of atoms is set to 5 for PCDDL. To get a fair comparison, the counterpart in FDDL is also set to 5.

We next choose 50, 75, 100 subjects that include at least 10 samples, respectively, from LFW3D-hassner and LFW3D-sdm, and every subject kept only 10 samples. Randomly selecting eight samples from each individual are used for training, and the rest samples used for testing. With varying the number of samples for each class in the following experiments. As for PCDDL and FDDL, since they are less affected by the number of atoms, and the complexity of coding is positively correlated with it. So, we next begin to find the appropriate number of the sub-dictionary atoms.

In the part, we conduct two experiments. First, as shown in Table 4, on the each feature dimension, we obtained the recognition rate of PCDDL with varying number of atoms, and we selected the number of atoms which can enable PCDDL to obtain the maximum accuracy on most feature dimensions. We can see that when the number of atoms is 5, PCDDL can obtain the maximum recognition rate on the most feature dimensions. In addition, we conduct another experiment to verify the stability of the algorithm. We randomly select eight samples used for training for each subject from the LFW3D-sdm database, and the rest samples used for testing. With varying the number of atoms from 3 to 8, we repeated it 30 times. To intuitively observe the distribution of the recognition rate, as listed in Table 3, we calculated its mean and variance. As illustrated in Figure 14, we as well as draw the boxplot of the recognition rates with varying number of atoms. As can be seen that from Figure 14, when the number of atoms is 5, the length of the box is relatively short, which proves that the results have smaller variance than that with other number of atoms. Meanwhile, the height of the red line in the box is relatively high, which proves that the results employ a larger mean than that with other number of atoms. In short, when the number of atoms is 5, the recognition rate has larger average, smaller variance and fewer outliers than with other numbers of atoms. Therefore, the number of atoms is set to 5 for PCDDL. To get a fair comparison, the counterpart in FDDL is also set to 5.

We next choose 50, 75, 100 subjects that include at least 10 samples, respectively, from LFW3D-hassner and LFW3D-sdm, and every subject kept only 10 samples. Randomly selecting eight samples from each individual are used for training, and the rest used for testing. We repeated the experiment 10 times on each various dimensions and calculated their mean accuracy and variance. The experimental results are given in Tables 5–10, and the bold font data is the highest average recognition rate achieved for each feature dimension. We can see that our model can get the better recognition result on the higher feature dimension, and this should be the characteristic of DL. Not only that, the performance of our model can improve with increasing the feature dimension. And this is a good trait. In addition to the average recognition rate, we also list the highest recognition rates that each method can achieve, and our model almost always achieves the highest recognition rate. And the highest recognition rate of our model is positively correlated with the feature dimension. To intuitively observe the change of the recognition rate, we also drew the average accuracy line as shown in Figures 15 and 16.

Figures 15 and 16 show the recognition rate curves of various methods on the LFW3D-sdm and LFW3D-hassner, respectively. The sub-figures (a), (b) and (c) correspond to the

### TABLE 1 The recognition rates of various methods on the AR database

| Methods   | SRC | NN  | SVM | DKSVD | FDDL | SVGDL | PCDDL |
|-----------|-----|-----|-----|-------|------|-------|-------|
| Accurate  | 88.80% | 71.40% | 87.10% | 85.40% | 91.43% | 93.14% | 94.43% |

1% better than other methods. As shown in Figure 12, we also provide the confusion matrix of PCDDL on the extended Yale B database.

III. Labeled Faces in the Wild (LFW): LFW face database [11], which is collected in the wild environment, is a popular database at present. It included 13,233 face images from 5749 people, with a different number of samples for each subject. In addition, face recognition faces some challenges due to uniform lighting, inconsistent backgrounds, changes in posture and expression and the number of faces in each image. At present, the common method is to pre-process the samples and then use the processed samples for face recognition. The pre-processing techniques [12] include face selection from images, alignment [28] (i.e. marking feature points), normalization [3, 27, 30, 31], clipping and down-sampling, etc. Fortunately, in the past few years, scholars at home and abroad have made great achievements in the field of face image alignment and normalization. Therefore, ready-made normalized database can be obtained, such as lfw3d-hassner [32, 33, 38] made by developer Hassner of the original 3D normalization algorithm, and lfw3d-sdm based on Supervised Descent Method (SDM) feature point detection method. The same samples from both above-mentioned databases are shown in Figure 13.

The specific experimental settings are as follows. We first down-sample the each normalized face image from LFW3D-hassner and LFW3D-sdm to 90x90. As you can see from Figure 10, SRC and SVDL are more sensitive to the number of atoms than FDDL and PCDDL. And the performance of the former are positively correlated with the number of atoms. The more atoms there are in SRC, the higher its recognition rate will be. Hence, the number of atoms in SRC and SVDL is set to the maximum (i.e. the number of training samples for each class) in the following experiments. As for PCDDL and FDDL, since they are less affected by the number of atoms, and the complexity of coding is positively correlated with it. So, we next begin to find the appropriate number of the sub-dictionary atoms.

In the part, we conduct two experiments. First, as shown in Table 4, on the each feature dimension, we obtained the recognition rate of PCDDL with varying number of atoms, and we selected the number of atoms which can enable PCDDL to obtain the maximum accuracy on most feature dimensions. We can see that when the number of atoms is 5, PCDDL can obtain the maximum recognition rate on the most feature dimensions. In addition, we conduct another experiment to verify the stability of the algorithm. We randomly select eight samples used for training for each subject from the LFW3D-sdm database, and the rest samples used for testing. With varying the number of atoms from 3 to 8, we repeated it 30 times. To intuitively observe the distribution of the recognition rate, as listed in Table 3, we calculated its mean and variance. As illustrated in Figure 14, we as well as draw the boxplot of the recognition rates with varying number of atoms. As can be seen from Figure 14, when the number of atoms is 5, the length of the box is relatively short, which proves that the results have smaller variance than that with other number of atoms. Meanwhile, the height of the red line in the box is relatively high, which proves that the results employ a larger mean than that with other number of atoms. In short, when the number of atoms is 5, the recognition rate has larger average, smaller variance and fewer outliers than with other numbers of atoms. Therefore, the number of atoms is set to 5 for PCDDL. To get a fair comparison, the counterpart in FDDL is also set to 5.

We next choose 50, 75, 100 subjects that include at least 10 samples, respectively, from LFW3D-hassner and LFW3D-sdm, and every subject kept only 10 samples. Randomly selecting eight samples from each individual are used for training, and the rest used for testing. We repeated the experiment 10 times on each various dimensions and calculated their mean accuracy and variance. The experimental results are given in Tables 5–10, and the bold font data is the highest average recognition rate achieved for each feature dimension. We can see that our model can get the better recognition result on the higher feature dimension, and this should be the characteristic of DL. Not only that, the performance of our model can improve with increasing the feature dimension. And this is a good trait. In addition to the average recognition rate, we also list the highest recognition rates that each method can achieve, and our model almost always achieves the highest recognition rate. And the highest recognition rate of our model is positively correlated with the feature dimension. To intuitively observe the change of the recognition rate, we also drew the average accuracy line as shown in Figures 15 and 16.

Figures 15 and 16 show the recognition rate curves of various methods on the LFW3D-sdm and LFW3D-hassner, respectively. The sub-figures (a), (b) and (c) correspond to the

### TABLE 2 The recognition rates of various methods on the Extended Yale B database

| Methods   | SRC | SVM | DKSVD | DLSI | FDDL | SVGDL | PCDDL |
|-----------|-----|-----|-------|------|------|-------|-------|
| Accurate  | 90.00% | 88.80% | 75.30% | 89.00% | 91.90% | 96.10% | 97.97% |
**Figure 12** Confusion matrix

**Figure 13** The illustration of the same individual samples from LFW3D-hassner and LFW3D-sdm (a) LFW3D-hassner (b) LFW3D-sdm

**Table 3** The mean and variance ($\times 10^{-4}$) of accurate under various number of sub-dictionary atoms

| Three atoms | Four atoms | Five atoms | Six atoms | Seven atoms | Eight atoms |
|-------------|------------|------------|-----------|-------------|-------------|
| Mean        | 0.9413     | 0.9500     | 0.9523    | 0.9497      | 0.9603      | 0.9557      |
| Var.        | 3.09       | 6.48       | 3.36      | 7.00        | 6.24        | 5.22        |
TABLE 4  The recognition rate of PCDDL with various number of atoms on each feature dimension

| Dims | Three atoms | Four atoms | Five atoms | Six atoms | Seven atoms | Eight atoms |
|------|-------------|------------|------------|-----------|-------------|-------------|
| 100  | 0.63        | 0.80       | 0.78       | 0.74      | 0.66        | 0.57        |
| 150  | 0.81        | 0.78       | 0.83       | 0.83      | 0.81        | 0.79        |
| 200  | 0.89        | 0.87       | 0.82       | 0.82      | 0.85        | 0.84        |
| 250  | 0.90        | 0.89       | 0.90       | 0.86      | 0.85        | 0.86        |
| 300  | 0.89        | 0.90       | 0.92       | 0.90      | 0.86        | 0.86        |
| 350  | 0.91        | 0.91       | 0.93       | 0.94      | 0.90        | 0.86        |
| 400  | 0.92        | 0.91       | 0.95       | 0.93      | 0.93        | 0.95        |

FIGURE 14  The recognition rate boxplot with various number of dictionary atoms

As we can see, the performance of all the methods gradually decreases with the increasing of the number of classes. And the recognition rates of other methods fluctuate with the increasing of feature dimension, while that of PCDDL show the basic upward tendency in the same case. At the same time, to further evaluate the stability of the algorithm, we also carried out the following experiment. We tested the results of every method on the same feature dimensions (400 dims) on LFW3D-sdm, and repeated it 30 times for each method. At every time, we randomly selected eight samples used for training, and the rest used for testing. We next calculated their mean and variance of recognition rates and listed them in Table 11. In order to intuitively observe the distribution of the data, we as well as draw the boxplot of various algorithms illustrated in Figure 17. We can see that the box length of PCDDL is relatively shorter than that of other methods, and the number of the outliers is also fewer than that of others, which proves that the results of PCDDL have less variance than that of others, i.e. stronger stability.

6  | DISCUSSION

According to the experimental results on the AR, Yale and LFW databases, we summarize the following conclusions:

(1) The experimental results show that the proposed method PCDDL outperforms the other methods in terms of recognition performance on the AR, Extended Yale B and LFW databases. According to the results in Tables 5–10, we can see that PCDDL is able to obtain the highest recognition rate in the highest dimensions feature space. From Figures 15 and 16, as the dimension increases, the recognition rate curves of PCDDL, compared with other methods, present a basic upward trend, which is a great property. However, as the number of the classes increases, we find that the recognition rates of all the methods show a downward trend, which proves that the current pattern classification methods are not good at the classification tasks that possess a large number of class. So, this should be the problem that we will solve next.

(2) The proposed PCDDL has stronger stability than other methods. By Figure 17 and Table 11, we can see that PCDDL possesses smaller variance, larger mean and fewer outliers than other methods at the same time, which proves that PCDDL, in most cases, can get better results than other methods.

7  | CONCLUSION

In this study, the classification mechanism of SRC, first, is clearly interpreted from the view of the probabilistic subspace, and the PCSRC model is offered at the same time. To get over the challenges that the feature hidden in the training samples cannot be fully extracted and the complexity of coding is too much high resulting from the overcomplete basis, we introduce the DL into the PCSRC model to get PCDL model. By learning a dictionary with as less number of atoms as it can, the above problem can be
TABLE 5  The mean accuracy and variance with varying feature dimensions of each algorithm on LFW-hassner(50 class) (the number in parentheses is the highest recognition rate (%))

| Methods | 100 dim | 150 dim | 200 dim | 250 dim | 300 dim | 350 dim | 400 dim |
|---------|---------|---------|---------|---------|---------|---------|---------|
| SVDL    | 89.30 ± 0.02(92.00) | 92.60 ± 0.03(95.00) | 93.20 ± 0.05(97.00) | 93.40 ± 0.04(96.00) | 94.80 ± 0.06(99.00) | 93.40 ± 0.07(99.00) | 93.70 ± 0.06(98.00) |
| SVM     | 86.10 ± 0.11(92.00) | 92.10 ± 0.12(97.00) | 89.70 ± 0.05(93.00) | 90.00 ± 0.06(94.00) | 89.90 ± 0.03(93.00) | 87.60 ± 0.04(92.00) | 84.10 ± 0.05(87.00) |
| FDDL    | 80.70 ± 0.08(85.00) | 82.30 ± 0.12(87.00) | 77.20 ± 0.27(86.00) | 81.90 ± 0.09(85.00) | 89.90 ± 0.11(94.00) | 91.80 ± 0.04(95.00) | 94.10 ± 0.03(96.00) |
| CRC     | 89.90 ± 0.04(93.00) | 93.70 ± 0.06(98.00) | 94.70 ± 0.03(97.00) | 94.10 ± 0.03(97.00) | 95.60 ± 0.05(98.00) | 94.60 ± 0.05(97.00) | 91.70 ± 0.03(94.00) |
| ProCRC  | 87.30 ± 0.06(91.00) | 89.80 ± 0.08(94.00) | 92.60 ± 0.04(94.00) | 89.90 ± 0.05(94.00) | 88.40 ± 0.03(91.00) | 89.10 ± 0.09(95.00) | 93.80 ± 0.02(97.00) |
| SRC     | 91.00 ± 0.03(93.00) | 95.00 ± 0.06(98.00) | 94.60 ± 0.04(98.00) | 95.80 ± 0.02(99.00) | 94.90 ± 0.04(98.00) | 92.80 ± 0.06(96.00) | 91.40 ± 0.02(93.00) |
| PCDDL-L1| 82.30 ± 0.14(90.00) | 87.60 ± 0.13(92.00) | 84.10 ± 0.17(94.00) | 90.20 ± 0.06(96.00) | 94.20 ± 0.04(97.00) | 95.20 ± 0.04(98.00) | 94.80 ± 0.05(98.00) |
| PCDDL-L2| 82.30 ± 0.14(86.00) | 87.30 ± 0.13(94.00) | 91.40 ± 0.17(94.00) | 91.10 ± 0.06(95.00) | 92.50 ± 0.04(98.00) | 92.30 ± 0.04(96.00) | 96.30 ± 0.05(100.00) |

TABLE 6  The mean accuracy and variance with varying feature dimensions of each algorithm on LFW-sdm(50 class) (the number in parentheses is the highest recognition rate (%))

| Methods | 100 dim | 150 dim | 200 dim | 250 dim | 300 dim | 350 dim | 400 dim |
|---------|---------|---------|---------|---------|---------|---------|---------|
| SVDL    | 88.90 ± 0.13(94.00) | 92.50 ± 0.08(96.00) | 94.10 ± 0.05(97.00) | 94.00 ± 0.03(97.00) | 94.30 ± 0.16(98.00) | 94.00 ± 0.05(97.00) | 93.90 ± 0.03(97.00) |
| SVM     | 89.50 ± 0.14(96.00) | 91.30 ± 0.12(96.00) | 91.30 ± 0.08(96.00) | 91.30 ± 0.04(94.00) | 90.60 ± 0.04(94.00) | 89.70 ± 0.06(93.00) | 86.80 ± 0.04(91.00) |
| FDDL    | 79.30 ± 0.12(86.00) | 79.60 ± 0.05(84.00) | 81.40 ± 0.14(86.00) | 82.30 ± 0.07(93.00) | 92.20 ± 0.04(96.00) | 93.70 ± 0.03(97.00) | 93.10 ± 0.08(97.00) |
| CRC     | 90.70 ± 0.10(95.00) | 93.80 ± 0.06(98.00) | 94.10 ± 0.11(98.00) | 93.50 ± 0.04(97.00) | 94.00 ± 0.04(97.00) | 93.80 ± 0.08(98.00) | 90.80 ± 0.10(96.00) |
| ProCRC  | 87.40 ± 0.03(89.00) | 89.30 ± 0.13(95.00) | 88.50 ± 0.13(93.00) | 89.90 ± 0.07(93.00) | 89.60 ± 0.09(93.00) | 91.60 ± 0.02(94.00) | 94.50 ± 0.04(97.00) |
| SRC     | 93.00 ± 0.01(95.00) | 94.50 ± 0.07(98.00) | 94.90 ± 0.02(97.00) | 94.30 ± 0.08(99.00) | 93.70 ± 0.06(96.00) | 93.10 ± 0.09(97.00) | 94.00 ± 0.04(98.00) |
| PCDDL-L1| 86.00 ± 0.11(91.00) | 85.50 ± 0.14(89.00) | 85.40 ± 0.09(89.00) | 91.90 ± 0.03(95.00) | 95.00 ± 0.01(96.00) | 93.20 ± 0.11(97.00) | 95.00 ± 0.09(99.00) |
| PCDDL-L2| 82.30 ± 0.11(86.00) | 88.00 ± 0.14(92.00) | 89.40 ± 0.09(93.00) | 90.50 ± 0.03(93.00) | 92.90 ± 0.01(95.00) | 92.90 ± 0.11(96.00) | 96.30 ± 0.09(100.00) |
### TABLE 7  The mean accuracy and variance with varying feature dimensions of each algorithm on LFW-hassner(75 class) (the number in parentheses is the highest recognition rate (%))

| Methods | 100 dim  | 200 dim  | 300 dim  | 400 dim  | 500 dim  | 600 dim  |
|---------|----------|----------|----------|----------|----------|----------|
| SVDL    | 87.73 ± 0.12(93.33) | 87.87 ± 0.07(90.67) | 89.60 ± 0.10(94.67) | 91.07 ± 0.04(94.67) | 90.53 ± 0.04(94.00) | 91.80 ± 0.08(95.33) |
| SVM     | 85.40 ± 0.05(90.67) | 85.73 ± 0.07(91.33) | 83.87 ± 0.02(88.67) | 83.47 ± 0.03(85.33) | 81.27 ± 0.04(84.00) | 78.67 ± 0.02(81.33) |
| FDDL    | 72.27 ± 0.35(78.67) | 83.67 ± 0.18(88.67) | 79.13 ± 0.11(84.00) | 85.07 ± 0.06(88.00) | 90.00 ± 0.05(92.00) | 90.60 ± 0.07(96.00) |
| CRC     | 87.00 ± 0.14(90.67) | 90.27 ± 0.03(92.00) | 92.00 ± 0.04(94.67) | 89.80 ± 0.04(94.00) | 90.20 ± 0.11(93.33) | 85.20 ± 0.06(88.00) |
| ProCRC  | 83.93 ± 0.05(88.00) | 87.60 ± 0.07(92.00) | 83.67 ± 0.04(92.67) | 89.53 ± 0.05(91.33) | 92.20 ± 0.05(96.00) | 91.33 ± 0.07(95.33) |
| SRC     | 82.93 ± 0.13(88.67) | 90.73 ± 0.06(92.67) | 91.33 ± 0.10(96.00) | 91.00 ± 0.04(94.00) | 89.80 ± 0.03(92.67) | 87.87 ± 0.07(92.00) |
| PCDDL-L1| 78.73 ± 0.12(84.67) | 86.20 ± 0.04(89.33) | 83.53 ± 0.17(88.67) | 87.53 ± 0.05(91.33) | 92.20 ± 0.05(96.00) | 91.33 ± 0.07(95.33) |
| PCDDL-L2| 77.87 ± 0.12(84.00) | 86.33 ± 0.04(91.33) | 87.27 ± 0.17(92.00) | 88.33 ± 0.05(92.00) | 91.87 ± 0.05(92.67) | 91.47 ± 0.07(95.33) |

### TABLE 8  The mean accuracy and variance with varying feature dimensions of each algorithm on LFW-sdm(75 class) (the number in parentheses is the highest recognition rate (%))

| Methods | 100 dim  | 200 dim  | 300 dim  | 400 dim  | 500 dim  | 600 dim  |
|---------|----------|----------|----------|----------|----------|----------|
| SVDL    | 84.00 ± 0.16(88.67) | 90.27 ± 0.04(93.33) | 91.33 ± 0.01(94.00) | 92.00 ± 0.07(95.33) | 91.93 ± 0.03(94.00) | 92.53 ± 0.06(95.33) |
| SVM     | 83.93 ± 0.11(89.33) | 87.53 ± 0.02(89.33) | 87.20 ± 0.09(88.67) | 86.00 ± 0.09(92.00) | 84.53 ± 0.09(91.33) | 84.60 ± 0.10(88.67) |
| FDDL    | 77.47 ± 0.11(84.00) | 81.93 ± 0.03(84.67) | 83.93 ± 0.03(86.00) | 89.53 ± 0.09(94.00) | 91.20 ± 0.02(93.33) | 91.60 ± 0.03(94.00) |
| CRC     | 82.93 ± 0.13(88.67) | 90.73 ± 0.05(95.33) | 92.00 ± 0.06(96.67) | 92.27 ± 0.05(95.33) | 91.33 ± 0.08(95.33) | 88.33 ± 0.07(92.67) |
| ProCRC  | 81.33 ± 0.10(86.00) | 88.73 ± 0.03(92.67) | 89.80 ± 0.03(93.33) | 89.40 ± 0.07(94.00) | 89.67 ± 0.07(93.33) | 93.07 ± 0.03(96.00) |
| SRC     | 85.13 ± 0.10(90.00) | 91.73 ± 0.10(96.00) | 91.47 ± 0.02(93.33) | 93.33 ± 0.05(97.33) | 91.33 ± 0.03(94.00) | 89.40 ± 0.04(91.33) |
| PCDDL-L1| 79.73 ± 0.20(84.67) | 89.47 ± 0.05(92.67) | 85.67 ± 0.03(88.00) | 90.13 ± 0.13(93.33) | 92.20 ± 0.07(96.00) | 93.40 ± 0.02(96.67) |
| PCDDL-L2| 76.80 ± 0.20(81.33) | 88.00 ± 0.05(90.67) | 89.87 ± 0.03(92.67) | 90.13 ± 0.13(94.00) | 91.20 ± 0.07(94.00) | 93.73 ± 0.03(98.00) |
TABLE 9  The mean accuracy and variance with varying feature dimensions of each algorithm on LFW-hassner (100 class) (the number in parentheses is the highest recognition rate (%))

| Methods | 200 dim | 300 dim | 400 dim | 500 dim | 600 dim | 700 dim | 800 dim |
|--------|--------|--------|--------|--------|--------|--------|--------|
| SVDL   | 87.35 ± 0.09(92.50) | 89.95 ± 0.05(93.00) | 87.20 ± 0.09(92.50) | 89.20 ± 0.04(92.00) | 89.90 ± 0.04(93.00) | 89.10 ± 0.05(92.00) | 89.05 ± 0.03(91.50) |
| SVM    | 84.70 ± 0.07(89.00) | 82.20 ± 0.04(85.00) | 80.00 ± 0.05(82.50) | 77.80 ± 0.13(82.00) | 78.70 ± 0.10(84.50) | 76.45 ± 0.11(80.50) | 72.00 ± 0.07(76.50) |
| FDDL   | 79.75 ± 0.06(82.50) | 82.80 ± 0.13(87.00) | 80.30 ± 0.08(86.50) | 81.65 ± 0.07(87.00) | 86.45 ± 0.07(91.00) | 87.65 ± 0.05(92.00) | 86.90 ± 0.03(90.00) |
| CRC    | 86.90 ± 0.03(89.00) | 89.20 ± 0.06(94.00) | 91.20 ± 0.10(95.00) | 89.20 ± 0.05(91.50) | 88.40 ± 0.04(91.00) | 84.95 ± 0.06(88.00) | 80.65 ± 0.02(84.00) |
| ProCRC | 86.25 ± 0.02(89.00) | 86.50 ± 0.10(92.00) | 86.85 ± 0.05(91.00) | 84.95 ± 0.03(86.50) | 85.45 ± 0.14(91.00) | 86.05 ± 0.05(90.00) | 88.10 ± 0.05(93.00) |
| SRC    | 87.85 ± 0.08(93.00) | 89.10 ± 0.05(93.00) | 90.15 ± 0.04(94.50) | 88.65 ± 0.09(94.00) | 86.45 ± 0.02(88.50) | 86.05 ± 0.05(89.50) | 81.75 ± 0.01(83.50) |
| PCDDL  | 83.75 ± 0.03(87.50) | 83.25 ± 0.06(88.00) | 83.75 ± 0.09(88.00) | 86.50 ± 0.02(89.00) | 88.10 ± 0.05(91.50) | 90.50 ± 0.05(95.00) | 89.35 ± 0.11(96.00) |

TABLE 10  The mean accuracy and variance with varying feature dimensions of each algorithm on LFW-sdm (100 class) (the number in parentheses is the highest recognition rate (%))

| Methods | 200 dim | 300 dim | 400 dim | 500 dim | 600 dim | 700 dim | 800 dim |
|--------|--------|--------|--------|--------|--------|--------|--------|
| SVDL   | 86.65 ± 0.04(90.00) | 88.65 ± 0.03(91.50) | 87.90 ± 0.03(91.50) | 89.45 ± 0.01(91.50) | 88.05 ± 0.02(90.00) | 88.50 ± 0.02(92.00) | 88.15 ± 0.02(90.00) |
| SVM    | 84.45 ± 0.04(89.50) | 84.50 ± 0.03(88.50) | 83.60 ± 0.04(86.50) | 82.50 ± 0.06(85.00) | 82.20 ± 0.06(85.00) | 80.35 ± 0.04(83.00) | 75.25 ± 0.11(81.00) |
| FDDL   | 79.15 ± 0.02(81.00) | 78.90 ± 0.06(83.00) | 82.20 ± 0.03(85.00) | 85.70 ± 0.07(91.00) | 86.85 ± 0.05(90.00) | 87.65 ± 0.04(91.00) | 88.90 ± 0.05(92.00) |
| CRC    | 87.20 ± 0.03(91.00) | 88.75 ± 0.02(91.00) | 88.30 ± 0.05(92.00) | 88.35 ± 0.03(90.00) | 90.00 ± 0.05(94.50) | 87.80 ± 0.01(89.50) | 81.90 ± 0.03(84.50) |
| ProCRC | 84.40 ± 0.03(87.00) | 86.20 ± 0.06(90.00) | 86.80 ± 0.02(89.00) | 86.15 ± 0.02(88.50) | 86.15 ± 0.10(92.00) | 86.85 ± 0.02(88.50) | 89.00 ± 0.03(92.00) |
| SRC    | 87.35 ± 0.01(90.00) | 89.90 ± 0.02(91.50) | 89.00 ± 0.03(92.00) | 89.30 ± 0.01(91.50) | 88.80 ± 0.02(91.00) | 86.60 ± 0.07(91.00) | 83.60 ± 0.02(86.00) |
| PCDDL  | 84.15 ± 0.03(86.50) | 84.95 ± 0.04(87.50) | 84.05 ± 0.04(86.50) | 87.30 ± 0.06(92.00) | 89.55 ± 0.04(92.50) | 88.65 ± 0.02(90.50) | 89.00 ± 0.03(93.00) |
solved. Meanwhile, in order to further enhance the discrimination of the coding coefficient, the Fisher criterion is introduced into our model PCDDL used for classification tasks, which can guarantee discrimination of both dictionary and the coefficient. Finally, we test the validity of our model in AR, Extended Yale B, LFW3D-hassner and LFW3D-sdm.

**FIGURE 15** The recognition rates of various methods on the LFW3D-sdm database (a) 50 subjects (b) 75 subjects (c) 100 subjects

**FIGURE 16** The recognition rates of various methods on the LFW3D-hassner database (a) 50 subjects (b) 75 subjects (c) 100 subjects
ACKNOWLEDGEMENTS

The authors would like to thank the online face database providers including AR, Yale and LFW.

REFERENCES

1. Cai, S., et al.: A probabilistic collaborative representation based approach for pattern classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2950–2959. IEEE (2016)
2. Yang, M., et al.: Fisher discrimination dictionary learning for sparse representation. In: ICCV 2011—International Conference on Computer Vision, pp. 6–13. IEEE, Barcelona, Spain (2011)
3. Xu, X., et al.: Weighted similarity and distance metric learning for unconstrained face verification with 3D frontalisation. IET Image Proc. 13(2), 399–408 (2019)
4. Cai, S., et al.: Support vector guided dictionary learning. In: European Conference on Computer Vision, pp. 624–639. Springer, International Publishing (2014)
5. Zhang, Q., Li, B.: Discriminative K-SVD for dictionary learning in face recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2691–2698. IEEE (2010)
6. Yang, M., et al.: Metaface learning for sparse representation based face recognition. In: Proc. IEEE Int. Conf. on Image Processing, pp. 1601–1604. ICIP, Hong Kong, China (2010)
7. Rosasco, L., et al. Iterative projection methods for structured sparsity regularization. MIT Technical Reports, MIT-CSAIL-TR-2009-050, CBCL-282 (2009)
8. Martinez, A., Benavente, R.: The AR face database. CVC Tech. Report No. 24 (1998)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Wright, J., et al.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31(2), 210–227 (2009)
11. Huang, G.B., et al.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition. Marseille, France (2008)
12. Georgiades, A.S., et al.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intell. 23(6), 643–660 (2001)
13. Lee, K., et al.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Pattern Anal. Mach. Intell. 27(5), 684–698 (2005)
14. Huang, K., Aviyente, S.: Sparse representation for signal classification. In: Conf. on Neural Information Processing Systems, pp. 609–616 (2006)
15. Lai, Z., et al. Rotational invariant dimensionality reduction algorithms. IEEE Trans. Cybern. 47(11), 3735–3746 (2017)
16. Zhang, D., et al.: Sparse representation or collaborative representation: Which helps face recognition? In: IEEE Int. Conf. on Computer Vision, pp. 471–478. IEEE (2011)
17. Zhang, W., et al.: Probabilistic Outputs for Multiclass Support Vector Machines Based on M-Ary Classification. J. Comput. Theor. Nanosci. 13(10), 6524–6530 (2016)
18. Lin, H., et al.: A note on Platt’s probabilistic outputs for support vector machines. Mach. Learn. 68(3), 267–276 (2007)
19. Franc, V., et al.: Support vector machines as probabilistic models. In: Int. Conf. on Machine Learning, pp. 665–672. ACM (2011)
20. Fearn, T.: Probabilistic principal component analysis. J. R. Stat. Soc. Series B 53, 23–23 (2014)
21. Prince, S.J., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: ICCV Int. Conf. on Computer Vision, pp. 1–8 (2007)
22. Lawrence, N.D.: Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J. Mach. Learn. Res. 6(60), 1783–1816 (2005)
23. Li, W., et al.: A sparse representation based study for face recognition. M.E. Thesis, Jiang Nan University (2016)
24. Mairal, J., et al.: Task-driven dictionary learning. IEEE Trans. Pattern Anal. Mach. Intell. 34(3), 791–804 (2012)
25. Chien, I.-T., Wu, C.-C.: Discriminant wavelet-faces and nearest feature classifiers for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 24(12), 1644–1649 (2002)
26. Tosić, I., Frossard, P.: Dictionary learning. IEEE Signal Process Mag. 28(2), 27–38 (2011)
27. Zhu, X., et al. High-fidelity pose and expression normalization for face recognition in the wild. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 787–796. IEEE (2015)
28. Zhu, X., et al.: Face alignment across large poses: A 3D solution. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 146–155. IEEE (2016)
29. Ramirez, I., et al.: Classification and clustering via dictionary learning with structured incoherence and shared features. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3501–3508 (2010)
30. Ding, C., Tao, D.: Pose-invariant face recognition with homography-based normalization. Pattern Recognit. 66, 144–152 (2017)
31. Tran, L., et al.: Disentangled representation learning GAN for pose-invariant face recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1283–1292. IEEE (2017)
32. Liu, F., et al.: Joint face alignment and 3D face reconstruction with application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 42(3), 664–678 (2020)
33. Tran, L., Liu, X.: On learning 3D face morphable model from In-the-wild images. IEEE Trans. Pattern Anal. Mach. Intell. 43(1), 157–171 (2021)
34. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B 67(2), 301–320 (2005)
35. Luo, X.L., et al.: Multi-resolution dictionary learning for face recognition. Pattern Recognit. 93, 283–292 (2019)
36. Hou, P., et al.: Multi-label manifold learning. In: National Conf. on Artificial Intelligence, pp. 1680–1686 (2016)
37. Bahrampour, S., et al.: Multimodal task-driven dictionary learning for image classification. IEEE Trans. Image Process. 25(1), 24–38 (2015)
38. Bustard, J.D., Nixon, M.S.: 3D morphable model construction for robust ear and face recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2582–2589. IEEE, San Francisco, CA, USA (2010)

39. Chen, D., et al.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3025–3032 (2013)

How to cite this article: Lv S, Liang J, Di L, Xia Y, Hou Z. A probabilistic collaborative dictionary learning-based approach for face recognition. IET Image Process. 2021;15:868–884. https://doi.org/10.1049/ipr2.12068