Sparse Representation for Face Recognition based on Discriminative Low-Rank Dictionary Learning

Long Ma, Chunheng Wang, Baihua Xiao, Wen Zhou
State Key Laboratory of Management and Control for Complex Systems
Institute of Automation Chinese Academy of Sciences
95 Zhongguancun East Road, 100190, BEIJING, CHINA
{long.ma, chunheng.wang, baihua.xiao, wen.zhou}@ia.ac.cn

Abstract

In this paper, we propose a discriminative low-rank dictionary learning algorithm for sparse representation. Sparse representation seeks the sparsest coefficients to represent the test signal as linear combination of the bases in an over-complete dictionary. Motivated by low-rank matrix recovery and completion, assume that the data from the same pattern are linearly correlated, if we stack these data points as column vectors of a dictionary, then the dictionary should be approximately low-rank. An objective function with sparse coefficients, class discrimination and rank minimization is proposed and optimized during dictionary learning. We have applied the algorithm for face recognition. Numerous experiments with improved performances over previous dictionary learning methods validate the effectiveness of the proposed algorithm.

1. Introduction

Sparse representation(or sparse coding) is a powerful tool for statistical signals processing, and it has been widely researched recently[9–11, 15, 26, 28]. Given an over-complete dictionary whose columns are prototype signal-atoms, signals are sparsest linearly represented by these atoms. Sparse representation has been supported by studies of human vision. Recent researches in human perception have revealled that human neurons are selective for a mount of stimuli in both low-level and mid-level human vision. An over-complete dictionary of base signal elements can be formed by these neurons, and the response of the neurons to an input image is highly sparse.

Wright[28] used sparse representation for face recognition and the performance is impressive. The algorithm used by Wright can be explained by solving the following optimization problem:

$$\min_x \|x\|_1 \qquad s.t. \qquad y = Dx \tag{1}$$

where D is an over-complete dictionary, x is the sparse coefficient vector, and y is the test signal. To improve the performance of sparse representation, Yang[30] proposed robust sparse coding to model the sparse coding as a sparsity-constrained robust regression problem; Liu[21] constrained the sparse coefficients to be nonnegative; Huang [14] exploited the clustering tends in nonzero coefficients. These algorithms used the off-the-shelf bases as the dictionary. Learning the dictionary has been proved to improve the signal reconstruction dramatically [8]. Several algorithms have been proposed to optimize the atoms. Aharon[1] generalized the k-means clustering process and proposed K-SVD algorithm, the algorithm iteratively updated the sparse coding of the samples based on the current dictionary and then optimized the dictionary atoms to better fit the data. Mairal [23] proposed an energy formulation with both sparse reconstruction and class discriminative components. An online dictionary learning[22] algorithm based on stochastic approximations is developed to handle large datasets with millions of training samples. Yang[29] proposed a fisher discrimination dictionary learning (FDDL) scheme for sparse representation. Furthermore, nonnegative matrix factorization(NMF)[18] algorithm has been proposed for sparse matrix factorization, and there exist also discriminant variants of the NMF algorithm[17].

One problem arises with these algorithms, the above algorithms aim at learning a dictionary to reconstruct the signals based on the training samples, and this strategy works well for the clean signals or signals that corrupted by small noises. Assume there are sparse large noises in the signals, the dictionary atoms have to be optimized to reconstruct the signals including the noises, and this will introduce corruptions into the atoms. Although an identity matrix I can be introduced as a dictionary to code the corrupted pixels, it

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

will influence the sparsity of the coefficients. It's beneficial for classification or reconstruction if the corrupted pixels can be separated or suppressed. Figure 1 shows some face images with self-shadowing and specularities from Extended Yale B face database[12, 19].

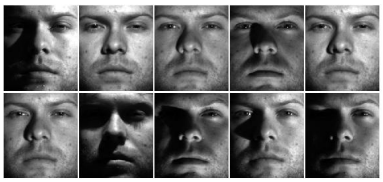


Figure 1. Face images with self-shadowing and specularities from Extended Yale B face database.

Given a matrix A whose columns come from the same pattern, these columns are linearly correlated in many situations and the matrix A should be approximately low-rank. If the entries in A are corrupted by sparse matrix E, we can nearly recover A by separating the noisy matrix E via rank minimization. Based on low-rank matrix recovery and completion[5–7, 16, 27], robust principal analysis(Robust PCA)[5, 27] was proposed to recover the underlying low-rank structure in the data.

Inspired by the above work, we integrate rank minimization into sparse representation for dictionary learning. The proposed algorithm separates the sparse noises from the signals while simultaneously optimizes the dictionary atoms to reconstruct the denoised signals. Compared to the other dictionary learning algorithms, the novelty of the proposed approach is threefold: First, the sparse noises in the training samples are corrected so that the atoms in the dictionary can be optimized to be pure. Second, the dictionary is learned with an explicit discriminative goal, making the proposed model suitable for recognition. Third, compared to Robust PCA, the dictionary is optimized under the framework of sparse representation, the similarity within one category is enhanced and the structure is strengthened to remove the sparse large noises, and the dictionary atoms are optimized to be better fit for sparse representation.

The remainder of the paper is organized as follows. The proposed dictionary learning model is presented in Section 2. Section 3 presents the optimization of the proposed model. Section 4 is devoted to experimental results and analysis, and Section 5 concludes this paper.

2. Discriminative Low-Rank Dictionary Learning for Sparse Representation(DLRD_SR)

To improve the performance of dictionary learning algorithm with noises, we propose a discriminative low-rank dictionary learning scheme. We learn a sub-dictionary D_i for the ith class, then we can get a structured dictionary $D = [D_1, D_2, \cdots D_c]$, where c is the number of classes. Then classification can be applied based on D.

Given a set of training data vectors $Y = [Y_1, Y_2, \cdots Y_c] \in R_{d,N}$, where Y_i is the samples from class i, d is the feature dimension, and N is the number of total training samples. Assume that X is the coding coefficient matrix of Y over D, then we can write $Y = DX + E$, where E is the sparse noises separated by DLRD_SR. We can write X as $X = [X_1, X_2, \cdots, X_c]$, where X_i is the sub-matrix containing the coding coefficients associated with the the training samples Y_i over D. In the proposed dictionary learning model, we require that the sub-dictionary D_i should be pure and compact so that the noises E in training samples Y can be separated, and the structured dictionary D should have powerful discriminative and reconstructive capability of samples Y.

2.1. Dictionary based on Matrix Rank Minimization

Given the samples Y_i from class i, the samples are linearly correlated in many situations. More precisely, the matrix $Y_i = [Y_{i,1}, Y_{i,2}, \cdots Y_{i,K_i}]$ should be approximately low-rank, where K_i is the number of the samples. This assumption holds generally, for example, according to the work of Basri and Jacobs [2], images of convex and Lambertian objects which taken under different illumination lie near an approximately nine-dimensional linear subspace that known as the *harmonic plane*.

Based on sparse representation, we seek a low-rank sub-dictionary D_i in which the bases can be linearly combined to represent the samples from class i. In practice, the low-rank structure can be easily violated with occluded or corrupted samples. So, a matrix E should be added to approximate the sparse error. Take sub-dictionary D_i as an example, the following model is proposed for dictionary learning:

$$\begin{aligned} &\min_{D_i,X,E_i} \|X_i\|_0 + \alpha rank(D_i) + \beta \|E_i\|_0 \qquad (2) \\ &s.t. \qquad Y_i = DX_i + E_i \end{aligned}$$

where X_i is the coding coefficients of Y_i over D, E_i is the error matrix corresponding to Y_i, α and β are positive weighting parameters that trade off the rank of the sub-dictionary and the additive error.

2.2. Dictionary based on Discriminative Learning

The coding coefficients of Y_i over D can be written as $X_i = [X_{i,1}; X_{i,2}; \cdots; X_{i,C}]$, where $X_{i,j}$ is the coding coefficient of Y_i corresponding to D_j. The discriminative power of D_i comes from following: First, the sub-dictionary D_i should be able to well represent Y_i, and there is $Y_i = D_i X_{i,i} + E_i$. Second, for the samples from class $j(j \neq i)$, $X_{j,i}$ should have nearly zero coefficients such that $r(D_i) = \sum_{j=1,j\neq i}^{C} \|D_i X_{j,i}\|_F^2$ is small, which means that

the correlation between D_i and $Y_j(j \neq i)$ should be updated to be small . Then problem (2) can be developed to the following problem:

$$\begin{aligned} \min_{D_i,E_i,X_i} \|X_i\|_0 + \alpha rank(D_i) + \beta \|E_i\|_0 + \lambda r(D_i) \\ s.t. \quad Y_i = DX_i + E_i, \;\; Y_i = D_i X_{i,i} + E_i \end{aligned} \tag{3}$$

Solving problem (3) is difficulty with the rank function and ℓ_0-norm, recent researches in low-rank completion and sparse representation suggest that we can replace the rank function by the convex surrogate, that's $\|D_i\|_*$ for $rank(D_i)$, and replace $\|X_i\|_1$ for $\|X_i\|_0$, $\|E_i\|_1$ for $\|E_i\|_0$ under certain condition. Thus, problem (3) yields the following optimization problem:

$$\begin{aligned} \min_{D_i,E_i,X_i} \|X_i\|_1 + \alpha \|D_i\|_* + \beta \|E_i\|_1 + \lambda r(D_i) \\ s.t. \quad Y_i = DX_i + E_i, \;\; Y_i = D_i X_i + E_i \end{aligned} \tag{4}$$

where $\|.\|_*$ denotes nuclear norm of a matrix(i.e., the sum of singular values of the matrix), $\|.\|_1$ denotes the sum of absolute values of matrix entries.

3. Formulation of DLRD_SR

3.1. Introduction of Augmented Lagrange Multiplier Method (ALM)

In this section, we will review the ALM algorithm [3, 4], and the algorithm will be used to solve problem (4) and its variations. Generally, the Augmented Lagrange Multipliers algorithm is developed to solve the following optimization problem:

$$\min_X f(X) \qquad s.t. \qquad h(X) = 0 \tag{5}$$

where f is a convex function, h is a linear function. We can define the following augmented Lagrangian function:

$$Ł(X, T, \mu) = f(X) + \langle T, h(X)\rangle + \frac{\mu}{2}\|h(X)\|_2^2 \tag{6}$$

where μ is a positive scalar, T is a Lagrange multiplier vector, and $\langle A, B\rangle = trace(A^t B)$. Bertsekas [3] established that with appropriate selection of T and a large enough value of μ, the solution to the problem (5) equals that of problem (6). With such convertion, the constrained optimization problem (5) is converted into the unconstrained problem (6) which has the same solution. The process of ALM algorithm is outlined as Algorithm 1. Algorithm 1 works efficiently if it's easy to seek X which can minimize $L_{u_k}(X, T_k)$. In this paper, the following key property of the matrix nuclear norm and ℓ_1-norm will be used in calling ALM algorithm:

$$\begin{aligned} S_\mu(W_1 + W_2) = \arg\min_X \mu\|X\|_1 - \langle X, W_1\rangle \\ + \frac{1}{2}\|X - W_2\|_F^2 \end{aligned} \tag{7}$$

Algorithm 1: General Method of Augmented Lagrange Multipliers

$\rho \geq 1$ **while** *not converged* **do**
  Solve $X_{k+1} = \arg\min_X L_{u_k}(X, T_k)$;
  Update T: $T_{k+1} = T_k + \mu_k h(X_{k+1})$;
  Update μ: $\mu_{k+1} = \mu_k + \rho\mu_k$;
end
Output: X_k

$$\begin{aligned} US_\mu[\textstyle\sum]V^* = \arg\min_X \mu\|X\|_* - \langle X, W_1\rangle \\ + \frac{1}{2}\|X - W_2\|_F^2 \end{aligned} \tag{8}$$

where μ is any non-negative real constant, $U\sum V^*$ is the singular value decomposition of the matrix $(W_1 + W_2)$. and $S_\mu[\sum]$ is the soft-thresholding or shrinkage operator:

$$S_\mu[x] = sign(x)(|x| - \mu) \tag{9}$$

3.2. Solution of DLRD_SR

The objective function in problem (4) can be divided into two sub-problems: First sub-problem: updating $X_i, (i = 1, 2...C)$ one by one by fixing D_i and $X_j(j \neq i)$, then we can get all coefficient matrix $X = [X_1; X_2; \cdots; X_c]$; Second sub-problem: updating D_i by fixing $X_j(j \neq i)$. One problem arises in the second sub-problem, if D_i is updated, the corresponding coefficients $X_{i,i}$ in coding Y_i should be updated to meet the condition $Y_i = D_i X_{i,i} + E_i$. So, $X_{i,i}$ is also updated in the second sub-problem. This two steps are iteratively operated to get the discriminative sub-dictionary D_i, the coefficients X_i, and the sparse error E_i. E_i is updated in each sub-problem, and the parameters β in this two sub-problems are set differently to adjust the weighting of the error E_i.

In the first sub-problem, suppose that the structured dictionary D is given, the coefficients matrix $X_i(i = 1, 2....C)$ is update one by one, then the problem is converted to the following sparse coding problem:

$$\min_{E_i,X_i} \|X_i\|_1 + \beta_1 \|E_i\|_1 \;\; s.t. \; Y_i = DX_i + E_i \tag{10}$$

We first convert Problem (10) to the following equivalent problem:

$$\begin{aligned} \min_{E_i,X_i} \|H\|_1 + \beta_1 \|E_i\|_1 \\ s.t. \quad Y_i = DX_i + E_i, \quad X_i = H \end{aligned} \tag{11}$$

Problem (11) can be solved by solving the following Augmented Lagrange Multiplier problem:

$$\begin{aligned} \min_{E_i,X_i} \|H\|_1 + \beta_1 \|E_i\|_1 + tr[T_1^t(X_i - H)] \\ + tr[T_2^t(Y_i - DX_i - E_i)] \\ + \frac{\mu}{2}(\|X_i - H\|_F^2 + \|Y_i - DX_i - E_i\|_F^2) \end{aligned} \tag{12}$$

where T_1 and T_2 are Lagrange multipliers and μ is a positive penalty parameter. The optimization process of problem (12) in given in Algorithm 2.

Algorithm 2: Solving Problem (12) via Inexact ALM

Input: Matrix Y_i, Initial Dictionary D_i, Parameters α, β_1, λ

Output: D_i, E_i, X_i

Initialize : $H = 0, E_i = 0, T_1 = 0, T_2 = 0$, $\mu = 10^{-6}$, $max_\mu = 10^{30}$, $\rho = 1.1$, $\varepsilon = 10^{-8}$;

while *not converged* **do**

fix others and update H by:

$$H = \arg\min_H \alpha \|H\|_1 - \langle T_1, H\rangle + \frac{\mu}{2}\|X_i - H\|_F^2$$

fix thers and update X_i by:

$$X_i = (I + D^tD)^{-1}(H + D^tY_i - D^tE_i + (D^tT_2 - T_1)/\mu)$$

fix the others and update E_i by:

$$E_i = \arg\min_E \beta_1 \|E\|_1 - \langle T_2, E\rangle + \frac{\mu}{2}\|Y_i - DX_i - E\|_F^2$$

update the multipliers by:

$$T_1 = T_1 + \mu(X_i - H)$$
$$T_2 = T_2 + \mu(Y_i - DX_i - E_i)$$

update the parameter μ by:

$$\mu = min(\rho\mu, max_\mu)$$

check the convergence conditions:

$$\|X_i - H\|_\infty < \varepsilon \ \ and \ \ \|Y_i - DX_i - E_i\|_\infty < \varepsilon$$

end

When $X_j(j = 1, 2...C)$ is fixed, sub-dictionary $D_i(i = 1, 2...C)$ is updated one by one, the second sub-problem is converted to the following problem:

$$\min_{D_i,E_i,X_{i,i}} \|X_{i,i}\|_1 + \alpha\|D_i\|_* + \beta_2\|E_i\|_1 + \lambda r(D_i) \quad s.t. \quad Y_i = D_iX_{i,i} + E_i, \tag{13}$$

We first convert problem (13) to the following equivalent problem:

$$\min_{D_i,E_i,X_{i,i}} \|Z\|_1 + \alpha\|J\|_* + \beta_2\|E_i\|_1 + \lambda r(D_i) \quad s.t. \quad Y_i = D_iX_{i,i} + E_i, \quad D_i = J, \quad X_{i,i} = Z \tag{14}$$

Problem (14) can be solved by solving the following Augmented Lagrange Multiplier problem:

$$\begin{aligned}\min_{D_i,E_i,X_{i,i}} \ & \|Z\|_1 + \alpha\|J\|_* + \beta_2\|E_i\|_1 + \lambda r(D_i) \\ & + tr[T_1^t(D_i - J)] + tr[T_2^t(Y_i - D_iX_{i,i} - E_i)] \\ & + tr[T_3^t(X_{i,i} - Z)] + \frac{\mu}{2}(\|D_i - J\|_F^2 \\ & + \|Y_i - D_iX_{i,i} - E_i\|_F^2 + \|X_{i,i} - Z\|_F^2)\end{aligned} \tag{15}$$

where T_1, T_2 and T_3 are Lagrange multipliers and μ is a positive penalty parameter.

The optimization process of Problem (15) is presented in Algorithm 3, and we require that each column of the dictionary is a unit vector.

The algorithm of DLRD_SR is summarized in Algorithm 4. Problem (12) and (15) can be solved by either *exact* or *inexact* ALM. The convergence property of Algorithm 2 could be proved similarity as Lin[20] did. We can not guarantee the convergence of Algorithm 3, so a maximum iteration number is required, but Algorithm 3 converges with a smaller iteration number than max_k in the experiments.

3.3. Classification based on DLRD_SR

In the dictionary learning process, there is an error matrix E which is used to approximate the noises or occlusions in the training process. In practical, noises exists both in training and testing samples, so a component that measure the sparsity of noises can be added in the classification process. Specially, given the dictionary $D = [D_1, D_2...D_N]$, where D_i is the learned dictionary for the ith pattern. Assume that x_j is an test sample from some class, and we can get the sparse coefficient vector by solving:

$$\min_{x,e} \|x\|_1 + \beta_1\|e\|_1 \qquad s.t. \qquad y = Dx + e \tag{16}$$

Similarly, problem (14) can be solved by ALM algorithm. For the ith class, we can approximate y_j by selecting the coefficients associated with the ith class:

$$\overline{y_j} = y_j - D\delta_i(x) - e \tag{17}$$

where $\delta_i(x)$ is a characteristic function which select the coefficients associated with the ith class. y_j is classified to the class with the minimum residual:

$$identity(y_j) = \arg\min_i \|y_j - D\delta_i(x) - e\|_F^2 \tag{18}$$

4. Experiments

In this section, firstly, we will apply the DLRD_SR algorithm for face recognition. We will test the robustness of DLRD_SR to illumination changes, pixel corruptions, uniform noises and block occlusions. Experimental results will be presented and be analysed in this section.

Algorithm 3: Solving Problem (15) via Inexact ALM

Input: Matrix Y_i, Initial Dictionary D_i, Parameters $\alpha, \beta_2, \lambda, max_k$

Output: $D_i, E_i, X_{i,i}$

Initialize : $J_0 = 0, E_{i,0} = 0, T_1 = 0, T_2 = 0, T_3 = 0$, $\mu = 10^{-6}$, $max_\mu = 10^{30}$, $\rho = 1.1$, $\varepsilon = 10^{-8}$, *k=0*;

while *not converged and* $k \leq \max_k$ **do**

fix others and update Z by:

$$Z = \arg\min_{\mathrm{Z}} \alpha \|Z\|_1 - \langle T_3, Z\rangle + \frac{\mu}{2}\|X_{i,i} - Z\|_F^2$$

fix others and update $X_{i,i}$ by:

$$X_{ii} = (D_i^t D_i + I)^{-1}(D_i^t Y_i - D_i^t E_i + Z + (D_i^t T_2 - T_3)/\mu)$$

fix others and update J by:

$$J = \arg\min_{\mathrm{J}} \alpha \|J\|_* - \langle T_1, J\rangle + \frac{\mu}{2}\|D_i - J\|_F^2$$

normalize the columns in J to unit vector;

fix thers and update D_i by:

$$D_i = (J + Y_i X_{i,i}^T - E_i X_{i,i}^T + (T_2 X_{i,i}^T - T_1)/\mu)(I + X_{i,i} X_{i,i}^T + V)^{-1}$$

normalize the columns in D_i to unit vector;

fix the others and update E_i by:

$$E_i = \arg\min_{\mathrm{E}} \beta_2 \|E\|_1 - \langle T_2, E\rangle + \frac{\mu}{2}\|Y_i - D_i X_{i,i} - E\|_F^2$$

update the multipliers by:

$$T_1 = T_1 + \mu(D_i - J)$$
$$T_2 = T_2 + \mu(Y_i - D_i X_{i,i} - E_i)$$
$$T_3 = T_3 + \mu(X_{i,i} - Z)$$

update the parameter μ by:

$$\mu = min(\rho\mu, max_\mu)$$

check the convergence conditions: $\|D_i - J\|_\infty < \varepsilon$ *and* $\|Y_i - D_i X_{i,i} - E_i\|_\infty < \varepsilon$ *and* $\|X_{i,i} - Z\|_\infty < \varepsilon$

end

$$V = \frac{2\lambda}{\mu} \sum_{j=1, j\neq i}^{C} X_{j,i} X_{j,i}^T$$

Algorithm 4: Algorithm of DLRD_SR

Input: Train Samples Y, Initial Dictionary D, Parameters α, β, λ

Output: D, E, X

Fix D and update X_i (i=1,2...C) one by one by solving problem (12);

Fix X and update each sub-dictionary D_i (i=1,2...C) one by one by solving problem (15);

Return to Step 1 until the sub-dictionary converge or the maximum number of iterations is reached;

4.1. Parameter selection

Although there are four parameters in DLRD_SR, we found that β_1 and β_2 play a more important roles in recognition, the other parameters are set as $\alpha = 1$,$\lambda = 1$ for all experiments in this paper, β_1 and β_2 are set by 5-fold cross validation. All parameters that used in competing methods are evaluated by 5-fold cross validation. Generally, the size of the dictionary should be set in advance. The original SR algorithm use all the training samples as dictionary, so the number of columns in the dictionary in DLRD_SR is same as the number of the training samples.

4.2. Face Recognition

In this section, we apply DLRD_SR for face recognition and compare the results with SR, SVM (Gaussian kernel), and the other two dictionary learning algorithm: Robust PCA[27], and FDDL[29]. We apply the algorithm on four datasets: the Extended Yale B database[12, 19], the UMIST database[13], the AR database[24] and the ORL[25] database. Note that unlike the other sparse coding algorithm, corruptions and occlusions exist both in the training and the testing images in our experiments. We will test the robustness of DLRD_SR to illumination changes, pixel corruptions, uniform distributed noises and block occlusions.

Extended Yale B Face Database. The Extended Yale B database consists of 2414 frontal-face images of 38 individuals. The images were captured under various laboratory-controlled lighting conditions. For each class, there are about 64 images in each class, half images are randomly selected as training images, and the rest are testing images. The columns of each image is concatenated as features. We apply the algorithm on the down-sampled images with ratios of 1/32, 1/24, 1/16, and 1/8. We repeat each experiment for 10 times and the accuracy is averaged. Figure 2 presents an example of low-rank dictionary learning process on Extended Yale B with a ratio of 1/8. Figure 2 (a) and (b) shows the discriminative power of the sub-dictionary $D_i (i = 1)$ and the structured dictionary D respectively, that's $r(D_i)$ and $R(D) = \sum_{i=1}^{C} r(D_i)$. Figure 2 (c) shows the maximum iteration numbers in Algorithm 2 and Algorithm 3 across each iteration in Algorithm 4. Figure 2 (d) shows the rank of the sub-dictionary D_i. Table 1 shows the recognition rates with different feature dimension. After dictionary learning, DLRD_SR outperformed SR with 2 % on average. When the feature is reduced to 30-dimension, SR performs better than DLRD_SR. Figure 3 shows an example of learned low-rank dictionary from Extended Yale B database.

UMIST Face Database. The UMIST face database

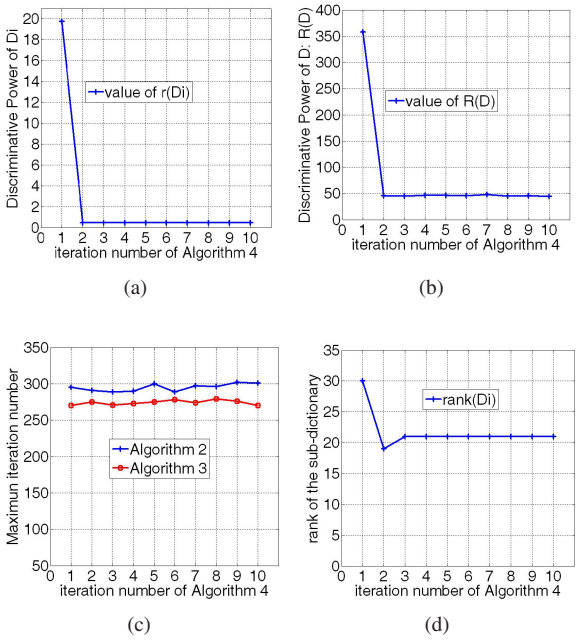


Figure 2. An example of low-rank dictionary learning process on Extended Yale B. (a)the discriminative power of the sub-dictionaryD_i: $r(D_i)$. (b)the discriminative power of the structured dictionary$D$: $R(D) = \sum_{i=1}^{C} r(D_i)$, (c)the maximum iteration number in Algorithm 2 and Algorithm 3 across each iteration of Algorithm 4. (d) the rank of the sub-dictionary D_i across iteration of Algorithm 4

Dimension	30	56	120	504
DLRD_SR (%)	75.0	**89.7**	**95.8**	**98.2**
SR (%)	75.3	86.5	93.6	97.0
NS (%)	80.8	88.2	91.1	93.4
SVM(%)	48.9	69.5	79.0	91.6
NN (%)	51.7	62.6	71.6	78.0

Table 1. Recognition accuracy on Extended Yale B database. The results of NS, SVM, and NN comes from Wright's work[28], ($\beta_1 = 1, \beta_2 = 0.2$).

Figure 3. An example of learned low-rank dictionary from Extended Yale B database

consists of 564 images of 20 individuals. Each individual is shown in a range of poses from profile to frontal views, which is challenging in computer vision because the variations between the images of the same face in viewing direction are almost always larger than image variations in face identity. In the experiment, we use the first 18 images for each individual, and half images are randomly selected to learn the dictionary and the left as testing images. All these images database are cropped and resized into 28* 23 images. For each training image and testing image, a certain percentage of pixels are randomly selected and replaced with value 255. We repeat the experiments for 10 times and average the recognition rates.

Figure 4 shows an example of face recognition from UMIST database with pixel corruptions. The recognition results are presented in Figure 5. With a little pixel corruption, the proposed DLRD_SR reduce the diversity within one category and the performance is not as good as FDDL. With more missing, the advantage of DLRD_SR over Robust PCA and FDDL is clear. The proposed DLRD_SR outperforms Robust PCA by 3% improvement on average. The recognition rates of FDDL decreases fast with increasing corruptions.

Figure 4. An example of DLRD_SR algorithm from UMIST database with 20% pixel corruptions. First row: original training images. Second row: corrupted training images. Third row: learned dictionary by DLRD_SR. Fourth row: separated noises from the second row by DLRD_SR. Fifth row: recovered images of the second row by DLRD_SR.

AR Face Database. In this experiment, we test the robustness of DLRD_SR to illumination and expression changes on AR database. The AR database consists of over 4000 frontal images from 126 individuals. For each individual, 26 pictures were taken in tow separate sessions. The images are cropped with dimension 27*20 and converted to grayscale. A subset that contains 50 males and 50 females was chosen from AR dataset.

In the experiment, the 7 images with illumination and expression changes from Session 1 are used for training, and the other 7 images with the same changes from Session 2 are used as testing images. We replace a percentage of randomly chosen pixels from each of training and testing images with iid samples from a uniform distribution

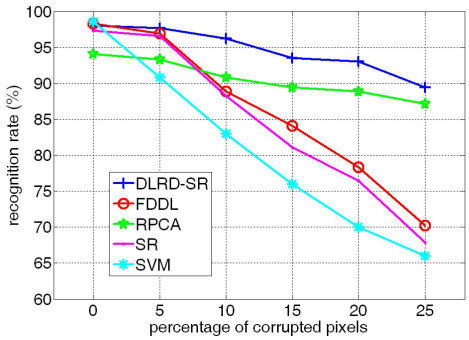
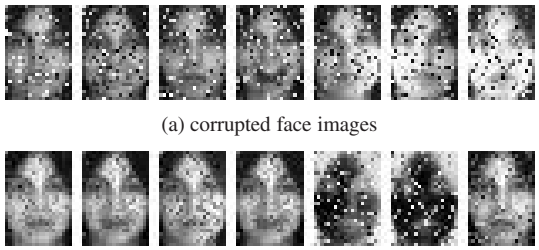


Figure 5. Recognition accuracy on UMIST database with pixel corruptions ($\beta_1 = 10, \beta_2 = 0.1$).

as Wright[27] did, that's samples uniform over $[0, V_{\max}]$, where $V_{\max}$ is the largest possible pixel value in the image. Figure 6 shows an example of dictionary learning algorithm of DLRD_SR. The recognition rates under different levels of noises are given in Figure 7. Apparently, all recognition rates decrease with increasing noises. With more noises, DLRD_SR performs better than FDDL by 4% improvement on average.



(a) corrupted face images

(b) dictionary learned by DLRD_SR

Figure 6. An example of dictionary learning algorithm by DLRD_SR from AR database with 20% uniform noises.

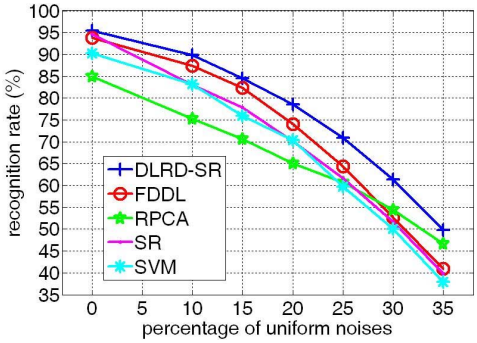


Figure 7. Recognition accuracy on AR database with pixel corruption ($\beta_1 = 5, \beta_2 = 0.1$).

ORL Face Database. The ORL database contains 400 images of 40 individuals. The images were taken at different times, varying the lighting, facial expressions and facial details. We use the cropped and normalized 28*23 face images. For each class, half images are randomly selected as training images, and the rest are testing images. We replace a randomly located block of each image with an unrelated random image. We repeat each experiment for 10 times and the accuracy is averaged. Figure 8 shows an example of training and testing face images with random block occlusions from ORL database. Table 2 lists the recognition rate with different levels of block occlusions. Compared to DLRD_SR and FDDL, Robust PCA doesn't handle the block occlusion well. DLRD_SR performs better than FDDL with a small advantage, and there is still 1.5% improvement with 50% occlusion.

Occlusions	0	10	20	30	40	50
DLRD_SR	95.9	**94.4**	**91.1**	**86.0**	**76.7**	**69.9**
FDDL	**96.7**	94.0	89.8	85.1	76.1	68.3
RPCA	89.3	88	83	76.6	72.0	66.2
SR	95.2	91.7	86.0	75.8	61.8	54
SVM	94.6	88.5	80.6	71.6	57.3	42

Table 2. Recognition accuracy (%) on ORL database with different level of occlusions (%) ($\beta_1 = 1, \beta_2 = 0.1$).

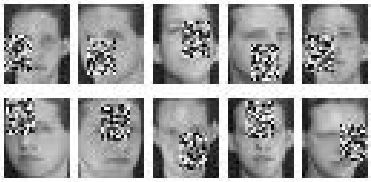


Figure 8. Training images (top) and testing (bottom) images from ORL database with 20% random block occlusions.

4.3. Results Analysis

In all the experiments, we can find that DLRD_SR performs better with noises, and sometimes the superiority is clear. For example, with uniform distribution noises, there is a 4% improvements over FDDL. When the rank of the sub-dictionary is minimized, the diversity of the bases in the sub-dictionary is reduced which may affect the sparse coding algorithm. The performances of Robust PCA in the experiments has verified this: with little noises, Robust PCA is inferior to SR algorithm, the superiority of rank minimization can be demonstrated with more noises. In DLRD_SR, the rank of the sub-dictionary is optimized under the frame of sparse representation, and the discrimination between the sub-dictionaries is exploited. Compared to DLRD_SR, the robustness of FDDL decrease fast with increasing noises.

5. Conclusion

In this paper, we proposed a discriminative low-rank dictionary learning for sparse representation. The learned dic-

tionary has two characteristics: First, the sub-dictionary is optimized to be low-rank, Second, the dictionary is updated to be more discriminative. Minimizing the rank of the sub-dictionary separates the noises in the training images and makes the sub-dictionary compact. The discriminative power of the dictionary comes from minimizing the correlation between the sub-dictionary and the samples that belong to other classes. The proposed DLRD_SR can be extended to separate noises, recover details and enhance global structures in the images. It can also be used for background modeling, video image restoration and so on. But DLRD_SR may reduce the diversity within the class which will affecting the sparse coding algorithm, and experiments on large dataset should be tested. Researches of these problems is on our future research agenda.

Acknowledgment

This work is supported by NSFC 60835001, 60802055, and 60933010.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. rmk-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[2] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 383–390. IEEE, 2001.

[3] D. Bertsekas. *Constrained optimization and Lagrange multiplier methods*, volume 410. Academic Press New York, 1982.

[4] D. Bertsekas. *Nonlinear programming*. Athena Scientific Belmont, MA, 1999.

[5] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? 2009.

[6] E. Candes and Y. Plan. Matrix completion with noise. *Arxiv preprint arXiv:0903.3131*, 2009.

[7] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[8] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

[9] Y. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *Information Theory, IEEE Transactions on*, 55(11):5302–5316, 2009.

[10] E. Elhamifar and R. Vidal. Sparse subspace clustering. 2009.

[11] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering.

[12] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[13] D. Graham and N. Allinson. Characterising virtual eigensignatures for general purpose face recognition. *NATO ASI series. Series F: computer and system sciences*, pages 446–456, 1998.

[14] J. Huang, X. Huang, and D. Metaxas. Learning with dynamic group sparsity. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 64–71. IEEE, 2009.

[15] K. Huang and S. Aviyente. Sparse representation for signal classification. *Advances in neural information processing systems*, 19:609, 2007.

[16] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 99:2057–2078, 2010.

[17] I. Kotsia, S. Zafeiriou, and I. Pitas. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *Information Forensics and Security, IEEE Transactions on*, 2(3):588–595, 2007.

[18] D. Lee, H. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[19] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[20] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Arxiv preprint arXiv:1009.5055*, 2010.

[21] Y. Liu, F. Wu, Z. Zhang, Y. Zhuang, and S. Yan. Sparse representation using nonnegative curds and whey. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3578–3585. IEEE.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

[23] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. 2008.

[24] A. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[25] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142. IEEE, 1994.

[26] E. Vidal. Robust classification using structured sparse representation.

[27] J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *submitted to Journal of the ACM*, 2009.

[28] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 210–227, 2008.

[29] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 543–550. IEEE, 2011.

[30] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition.