Discriminative Dictionary Learning with Low-Rank Regularization for Face Recognition

Liangyue Li, Sheng Li, and Yun Fu
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115, USA
{li.liangy, li.shengl}@husky.neu.edu, yunfu@ece.neu.edu

Abstract—We consider learning a discriminative dictionary in sparse representation and specifically focus on face recognition application to improve its performance. This paper presents an algorithm to learn a discriminative dictionary with low-rank regularization on the dictionary. To make the dictionary more discriminative, we apply the Fisher discriminant function to the coding coefficients with the goal that they have a small ratio of the within-class scatter to between-class scatter. However, noise in the training samples will undermine the discrimination power of the dictionary. To handle this problem, we base on low-rank matrix recovery theory and apply a low-rank regularization on the dictionary. The proposed discriminative dictionary learning with low-rank regularization ($D^2 L^2 R^2$) algorithm is evaluated on several face image datasets in comparison with existing representative dictionary learning and classification algorithms. The experimental results demonstrate its superiority.

I. INTRODUCTION

Face recognition has been studied extensively due to its real world challenges and great value in numerous applications [1][2][3][4][5]. Wright et al.[6] addressed face recognition problem from sparse signal representation theory and achieved promising results. Given a test signal and an over-complete dictionary with prototype signals as atoms, sparse representation seeks a sparsest representation of the test signal among all the linear combinations of the dictionary atoms. This sparsity has been supported by research in human visual system which found that nerve cells in the connecting pathway only react to a certain amount of stimuli[7].

Given an over-complete dictionary $D$ and a query sample $y$, the essence of sparse representation is to solve the following minimization problem:

$$\min_{\alpha} ||\alpha||_1 \quad s.t. \quad y = D\alpha$$

where $\alpha$ is the coding coefficient whose non-zero elements are those corresponding to the category $y$ belongs to. Here the dictionary $D$ can either be pre-specified or gradually adapted to fit the training samples given. Wright et al. [6] pre-specified the dictionary as the original training samples. A problem with this strategy is that the original images in the training set may not faithfully represent the test samples due to the noise and uncertainty in it. Besides, the distinctive message resided in the training set might be ignored in this way. As a result, we need to adapt the learned dictionary based on the specific training set.

In order to have a well adapted dictionary for discriminative representation of test samples, research progress on dictionary learning has been achieved. Generalizing K-means clustering process, Aharon et al.[8] presented K-SVD algorithm to learn an over-complete dictionary by updating dictionary atoms and sparse representations iteratively. Recently, a discriminative K-SVD method that considers classification error when learning the dictionary was proposed [9]. Jiang et al.[10] associated label information with each dictionary atom to enforce discriminability. To reduce computational complexity, [11] [12] emphasized specific discriminative criteria to learn an over-complete dictionary. In [13], the authors introduced the Fisher criterion to learn a structured dictionary. Studer et al.[14] investigated dictionary learning from spuriously corrupted signals. However, the methods above can only work well on clean training samples or with small noise and sparse corruption. Imagine the training samples are corrupted with large noise, then in order for representing the training samples, the dictionary atoms will also get corrupted.

Given a matrix $M$ of low rank, matrix completion aims at recovering it from noisy observations of a random small portion of its elements. It has been proved that under certain assumptions, the problem can be exactly solved and several methods have been proposed[15][16][17]. In our case of dictionary learning for face recognition, training samples in the same class are linearly correlated and live in a low dimensional manifold. Therefore, a sub-dictionary for representing...
samples from one class should reasonably of low rank. Ma et al.[5] integrated rank minimization into sparse representation and achieved impressive results especially when corruption existed.

Inspired by the previous work, we aim at learning a discriminative dictionary for face recognition which can handle training samples that are corrupted with large noise. We propose a discriminative dictionary learning with low-rank regularization (D²L²R²) algorithm and illustrate it in Fig. 1. In the figure, the training dataset can approximately be recovered by the multiplication of the dictionary and the coding coefficient matrix. Each sub-dictionary is of low rank (can be seen as multiplication of two matrices of smaller size) to reduce the negative effect of noise contained in training samples. The coding coefficients conform to Fisher discrimination criterion. Benefiting from the above design, our algorithm has the following advantages: First, the Fisher discriminant function can help us achieve a small ratio of the within-class scatter to between-class scatter on the coefficients, making the dictionary learned has strong discriminative power. Second, low-rank regularization will return us a compact and pure dictionary that can reconstruct the denoised images even when the training samples are contaminated.

Different from FDDL proposed in [13], our D²L²R² algorithm can well cope with training samples with large noise and can still achieve impressive performance due to the low-rank regularization on the sub-dictionaries. Though DLRD SR proposed in [5] was claimed to be able to handle noisy samples as well, it may suffer from certain information loss because of the low-rank regularization, our D²L²R² algorithm compensates this by enforcing the Fisher criterion on the coding coefficients of the training sets.

The rest of this paper is organized as follows. Section II reviews sparse representation based classification method. Section III introduces our Discriminative Dictionary Learning with Low-Rank Regularization (D²L²R²). Section IV presents the optimization solution for our model. Section V describes the classification scheme. Section VI presents experimental results. We draw conclusion in section VII.

II. SPARSE REPRESENTATION BASED CLASSIFICATION (SRC) REVISITED

We briefly review robust face recognition using sparse representation proposed in [6]. Define matrix $Y$ as the entire training set which consists of $n$ training samples from all $c$ different classes: $Y = [Y_1, Y_2, \ldots, Y_c]$ where $Y_i \in R^{d \times n_i}$ is all the training samples from $i$-th class, $d$ is the dimensionality of each sample vector, and $n_i$ is $i$-th class’ sample size.

From $Y$, we want to learn a discriminative dictionary for future classification task. Rather than learning the dictionary as a whole from all the training samples, we learn a sub-dictionary $D_i$ for the $i$-th class separately. With all the sub-dictionaries learned, we will get the whole dictionary as $D = [D_1, D_2, \ldots, D_c]$, where $c$ is the number of classes, $D_i$ is the sub-dictionary for the $i$-th class, each $D_i$ is of size $d \times m_i$, $d$ is the dimension of each dictionary atom which is the same with the feature dimension of each training sample, and $m_i$ is the number of atoms in the $i$-th sub-dictionary.

We represent the entire training set $Y$ using the whole dictionary $D$ and denote by $X$ the sparse coefficient matrix we obtain. We should have $Y \approx DX$, and $X$ could be written as $X = [X_1, X_2, \ldots, X_c]$, where $X_i$ is the sub-matrix that is the coefficients for representing $Y_i$ using $D$. In this paper, we propose the following D²L²R² model:

$$ J_{D,X} = \arg \min_{D,X} \{ R(D,X) + \lambda_1 \|X\|_1 + \lambda_2 F(X) + \alpha \sum_i \|D_i\|_* \} $$

where $R(D,X)$ is reconstruction error term for expressing the discrimination power of $D$, $\|X\|_1$ is the $l_1$ regularization on coding coefficient matrix, $F(X)$ is the Fisher discriminant function of the coefficients $X$, and $\|D_i\|_*$ is the nuclear norm of each sub-dictionary $D_i$, which is the convex envelope of its matrix rank. We will break down the model in the following subsections.

B. Discriminative reconstruction error

Sub-dictionary $D_i$ should have the capacity to well represent samples from $i$-th class. To illustrate this mathematically, we rewrite $X_i$, the coding coefficient matrix of $Y_i$ over $D$, as $X_i = [X^1_i; X^2_i; \ldots; X^r_i]$, where $X^j_i \in R^{m_j \times n_i}$ is the coding coefficient matrix of $Y_i$ over $D_j$. We will have to minimize $\|Y_i - D_jX^j_i\|_F^2$. On the other hand, $D_j$ should not be able to represent samples from other classes, that is $\sum_{j=1,j\neq i}^r \|D_jX^j_i\|_F^2$ should be as small as possible, where each $X^j_i$ has nearly zero elements. Lastly, it is obvious that the whole dictionary $D$ can well represent samples from any class $Y_i$, so we require the minimization of $\|Y_i - DX_i\|_F^2$.

the smallest residual:

$$ \min_i r_i(y) = \|y - Y\delta_i(\alpha)\|_2 \quad (2) $$

where $\delta_i(\alpha)$ is a function that picks the coefficients corresponding to $i$-th class.

III. DISCRIMINATIVE DICTIONARY LEARNING WITH LOW-RANK REGULARIZATION (D²L²R²)

To learn a discriminative dictionary even when large noise exists in the training samples, we propose a discriminative dictionary learning algorithm with low-rank regularization.

A. Notations and model overview

Define matrix $Y$ as the entire training set which consists of $n$ training samples from all $c$ different classes: $Y = [Y_1, Y_2, \ldots, Y_c]$ where $Y_i \in R^{d \times n_i}$ is all the training samples from $i$-th class, $d$ is the dimensionality of each sample vector, and $n_i$ is $i$-th class’ sample size.

We represent the entire training set $Y$ using the whole dictionary $D$ and denote by $X$ the sparse coefficient matrix we obtain. We should have $Y \approx DX$, and $X$ could be written as $X = [X_1, X_2, \ldots, X_c]$, where $X_i$ is the sub-matrix that is the coefficients for representing $Y_i$ using $D$. In this paper, we propose the following D²L²R² model:

$$ J_{D,X} = \arg \min_{D,X} \{ R(D,X) + \lambda_1 \|X\|_1 + \lambda_2 F(X) + \alpha \sum_i \|D_i\|_* \} $$

where $R(D,X)$ is reconstruction error term for expressing the discrimination power of $D$, $\|X\|_1$ is the $l_1$ regularization on coding coefficient matrix, $F(X)$ is the Fisher discriminant function of the coefficients $X$, and $\|D_i\|_*$ is the nuclear norm of each sub-dictionary $D_i$, which is the convex envelope of its matrix rank. We will break down the model in the following subsections.

B. Discriminative reconstruction error

Sub-dictionary $D_i$ should have the capacity to well represent samples from $i$-th class. To illustrate this mathematically, we rewrite $X_i$, the coding coefficient matrix of $Y_i$ over $D$, as $X_i = [X^1_i; X^2_i; \ldots; X^r_i]$, where $X^j_i \in R^{m_j \times n_i}$ is the coding coefficient matrix of $Y_i$ over $D_j$. We will have to minimize $\|Y_i - D_jX^j_i\|_F^2$. On the other hand, $D_j$ should not be able to represent samples from other classes, that is $\sum_{j=1,j\neq i}^r \|D_jX^j_i\|_F^2$ should be as small as possible, where each $X^j_i$ has nearly zero elements. Lastly, it is obvious that the whole dictionary $D$ can well represent samples from any class $Y_i$, so we require the minimization of $\|Y_i - DX_i\|_F^2$.

the smallest residual:

$$ \min_i r_i(y) = \|y - Y\delta_i(\alpha)\|_2 \quad (2) $$

where $\delta_i(\alpha)$ is a function that picks the coefficients corresponding to $i$-th class.
Denote $R(D_i, X_i) = \|Y_i - D_i X_i\|_F^2 + \sum_{j=1,j\neq i}^{c} \|D_j X_j\|_F^2 + \|Y_i - D X_i\|_F^2$ as the discriminative reconstruction error term for sub-dictionary $D_i$, we want to minimize the value of $R(D_i, X_i)$.

C. Fisher discriminant for sparse codings

In addition to the discriminative reconstruction term, we want to make the coding coefficient matrix $X$ discriminative as well. In this way, $D$ will have discriminative power for training samples $Y$. We apply Fisher discrimination criteria[18] on the coding coefficient matrix $X$ so that the ratio of within-class scatter to between-class scatter will be minimized and samples from different classes can be separated. Denote by $S_W(X)$ the within-class scatter of $X$ and $S_B(X)$ the between-class scatter of $X$, they are defined as

$$S_W(X) = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \bar{x}_i)(x_k - \bar{x}_i)^T$$

$$S_B(X) = \sum_{i=1}^{c} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

where $\bar{x}_i$ is the mean sample of $X_i$, $\bar{x}$ is the mean sample of $X$, and $n_i$ is the number of samples in $i$-th class.

As we have mentioned, we will introduce Fisher criterion to $X$ and denote this term as $F(X)$ which is defined as

$$F(X) = tr(S_W(X)) - tr(S_B(X)) + \eta \|X\|_F^2$$ (3)

Note here that minimizing $tr(S_W(X)) - tr(S_B(X))$ is equivalent to minimizing the ratio of within-class scatter to between-class scatter. The last term $\eta \|X\|_F^2$ is to make the function convex and stable in which $\eta$ is set as $\eta = 1$ [13].

D. Low-Rank regularization on the dictionary

In face recognition, training samples in the same class are linearly correlated and reside in a low dimensional subspace. Therefore, a sub-dictionary for representing samples from one class should be reasonably of low rank. Besides, requiring that the sub-dictionaries are of low rank can separate the noisy information and make the dictionary more pure and compact. Of all the possible sub-dictionaries $D_i$ that can represent samples from $i$-th class, we want to find the one with the most compact bases, that is to minimize $\|D_i\|_*$. 

E. The $D^2L^2R^2$ model

Considering the discriminative reconstruction error, Fisher discrimination criterion on the coding coefficients and the low-rank regularization on the dictionary all together, we have the following $D^2L^2R^2$ model:

$$J_{(D,X)} = \arg \min_{D,X} \left( \sum_{i=1}^{c} R(D_i, X_i) + \lambda_1 \|X\|_1 + \lambda_2 F(X) + \alpha \sum_{i=1}^{c} \|D_i\|_* \right)$$ (4)

In the next section, we solve our model by alternatively optimizing $D$ and $X$. 

Algorithm 1 Inexact ALM algorithm for Eq. (7)

**Input:** Initial Dictionary $D$, Matrix $Y$, parameters $\alpha, \beta, \lambda$

**Output:** $D_i, E_i, X_i^*$

**Initialize:** $J = 0, E_i = 0, T_1 = 0, T_2 = 0, T_3 = 0, \mu = 10^{-6}, \max x_\mu = 10^{30}, \varepsilon = 10^{-8}, \rho = 1.1$

while not converged do

1. Fix other variables and update $Z$ by:

   $$Z = \arg \min_{Z} \left( \frac{1}{\mu} \|Z\|_1 + \frac{1}{2} \|Z - (X_i + \frac{T_3}{\mu})\|_F^2 \right)$$

2. Fix other variables and update $X_i^*$ by:

   $$X_i^* = (D_i^T D_i + I)^{-1} (D_i^T Y_i - E_i) + Z + \frac{D_i^T (T_1 - T_3)}{\mu}$$

3. Fix other variables and update $J$ by:

   $$J = \arg \min_{J} \left( \alpha \|J\|_* + \frac{1}{2} \|J - (D_i + \frac{T_2}{\mu})\|_F^2 \right)$$

length normalization for each column in $J$;

4. Fix other variables and update $D_i$ by:

   $$D_i = \frac{[2\mu (Y_i X_i^T + (\sum_{j=1,j\neq i}^{c} D_j X_j^T) X_i^T)]^T}{(\mu + 1)X_i^T + I}$$

length normalization for each atom in $D_i$;

5. Fix other variables and update $E_i$ by:

   $$E_i = \arg \min_{E_i} \left( \frac{\alpha}{\mu} \|E_i\|_{2,1} + \frac{\gamma}{2} \|E_i - (Y_i - D_i X_i^T + \frac{T_1}{\mu})\|_F^2 \right)$$

6. Update $T_1, T_2$ and $T_3$ by:

   $$T_1 = T_1 + \mu (Y_i - D_i X_i^T - E_i)$$

   $$T_2 = T_2 + \mu (D_i - J)$$

   $$T_3 = T_3 + \mu (X_i^T - Z)$$

7. Update $\mu$ by:

   $$\mu = \min(\rho \mu, \max x_\mu)$$

8. Determine when to stop the iterations:

   $\|D_i - J\|_\infty < \varepsilon$ and $\|Y_i - D_i X_i^T - E_i\|_\infty < \varepsilon$ and $\|X_i^T - Z\|_\infty < \varepsilon$

end while

IV. OPTIMIZATION OF $D^2L^2R^2$

We consider solving the proposed objective function in Eq.(4) by dividing it into two sub-problems: First, updating $X_i (i = 1, 2, . . . , c)$ by fixing dictionary $D$ and all $X_j (j \neq i)$ and we can get coding coefficient matrix $X$ by putting all the $X_i (i = 1, 2, . . . , c)$ together; Second, updating $D_i$ by fixing $X$ and $D_j (j \neq i)$. However, when $D_i$ is updated, the coding coefficient of $Y_i$ over $D_i$ should also be updated to reflect this change, that is, $X_i^*$ is also updated. Once we have initialized $D$ we can proceed by iteratively repeating the above process until a stopping criterion is reached.
A. Updating coding coefficient matrix X

First, we assume D is fixed, the original objective function Eq.(4) reduces to a sparse coding problem. We update each X_i one by one and make all X_j (j ≠ i) fixed. This can be done by solving the following:

$$J(X_i) = \arg \min_{X_i} \left( \frac{1}{2} \| Y_i - D_i X_i \|_F^2 + \| Y_i - D_j X_j \|_F^2 + \sum_{j=1, j \neq i}^{c} \| D_j X_j - \bar{D} \|_1^2 + \lambda_i \| X_i \|_1 + \lambda_2 F_i(X_i) \right)$$  \hspace{1cm} (5)

where $F_i(X_i) = \| X_i - \bar{X}_i \|_F^2 - \sum_{k=1}^{n} \| X_k - \bar{X} \|_F^2 + \eta \| X_i \|_F^2$ and $\bar{X}_k$ and $\bar{X}$ are matrices composed of the mean of vectors of k-th class and all classes. This reduced objective function can be solved by using Iterative Projection Method in [19] by rewriting it as

$$J(X_i) = \arg \min_{X_i} \left\{ Q(X_i) + 2\tau \| X_i \|_1 \right\}$$

More details can be referred to [13].

B. Updating sub-dictionaries D_i

Next, if X is fixed, we update D_i by fixing all the other D_j (j ≠ i). Notice that the coding coefficient of Y_i over D_i should also be updated at the same time. The objective function then reduces to

$$J(D_i) = \arg \min_{D_i, X_i} \left( \frac{1}{2} \| Y_i - D_i X_i \|_F^2 + \sum_{j=1, j \neq i}^{c} \| D_j X_j - \bar{D} \|_1^2 + \| Y_i - D_i X_i \|_F^2 \right) + \alpha \| D_i \|_n$$

Denote $r(D_i) = \| Y_i - D_i X_i \|_F^2 - \sum_{j=1, j \neq i}^{c} \| D_j X_j - \bar{D} \|_1^2 + \sum_{j=1, j \neq i}^{c} \| D_j X_j \|_F^2$, the above objective function can be converted to the following:

$$\min_{D_i, E_i, X_i} \| X_i \|_1 + \alpha \| D_i \|_n + \beta \| E_i \|_2,1 + \lambda r(D_i)$$

s.t. $Y_i = D_i X_i + E_i$

which can be solved as the following essentially the same problem:

$$\min_{D_i, E_i, X_i} \| Z \|_1 + \alpha \| J \|_n + \beta \| E_i \|_2,1 + \lambda r(D_i)$$

s.t. $Y_i = D_i X_i + E_i, D_i = J, X_i = Z$

The above problem can then be solved by the following Augmented Lagrange Multiplier [20] method:

$$\min_{D_i, E_i, X_i} \| Z \|_1 + \alpha \| J \|_n + \beta \| E_i \|_2,1 + \lambda r(D_i) + \mu \left( \| Y_i - D_i X_i - E_i \|_F^2 + \| Y_i - D_i - J \|_F^2 + \| X_i - Z \|_F^2 \right)$$

where $T_1, T_2, T_3$ are Lagrange multipliers and $\mu (\mu > 0)$ is a balance parameter. The details of solving the problem can be referred to Algorithm 1. Each atom of the dictionary is normalized to a unit vector. A similar proof to demonstrate the convergence property of Algorithm 1 can be found in Lin et. al’s work [21].

We summarize our algorithm for $D^2 L^2 R^2$ in Table I.

<table>
<thead>
<tr>
<th>Table I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Algorithm for $D^2 L^2 R^2$ model</strong></td>
</tr>
<tr>
<td>1. Initialize dictionary D</td>
</tr>
<tr>
<td>2. Update the coefficient matrix X</td>
</tr>
<tr>
<td>3. Update dictionary D</td>
</tr>
<tr>
<td>4. Output</td>
</tr>
<tr>
<td>Check the values of $J_{D,D_i}$ in two consecutive iterations, if their difference is small enough or the maximum loops are done, output X and D, otherwise continue with step 2.</td>
</tr>
</tbody>
</table>

V. Classification based on our method

We code a query sample y against the dictionary D learned and obtain the coding coefficient by solving

$$x = \arg \min_{x} \left\{ \| x - D x \|_2^2 + \gamma \| x \|_1 \right\}$$  \hspace{1cm} (8)

Denote by $x = [x_1; x_2; \ldots; x_c]$, where $x_i$ is the coefficient vector over sub-dictionary $D_i$. We can calculate the residual associated with i-th class as

$$e_i = \| y - D_i x_i \|_2^2 + \| x - \bar{x}_i \|_2^2$$  \hspace{1cm} (9)

where $\bar{x}_i$ is the learned mean coefficient of class i, and $\omega$ is a preset weight parameter. The identity of testing sample $y$ is determined according to

$$\text{identity}(y) = \arg \min_{i} \{ e_i \}$$  \hspace{1cm} (10)

VI. Experimental results

We apply our algorithm to face recognition on ORL[22], Extend Yale B [23] and CMU PIE[24] databases to verify its performance. The robustness of our $D^2 L^2 R^2$ algorithm to illumination changes, pixel corruptions, block corruptions, and noise will be tested. Experimental results will be presented along with some analysis.

A. Parameter selection

We set the number of dictionary columns of each class as training size. There are 5 parameters in our algorithm: $\lambda_1$ and $\lambda_2$ in Eq.(5) and $\alpha$ and $\lambda$ in Eq.(6). In the experiment, we found that changing $\alpha$ and $\lambda$ wouldn’t affect the result that much, and we set them both as 1. Parameters of the comparison algorithms are chosen by cross validation in 5-fold fashion. For ORL, $\lambda_1 = 0.005, \lambda_2 = 0.05, \beta = 0.1$; for Extended Yale B, $\lambda_1 = 0.005, \lambda_2 = 0.005, \beta = 0.01$; for PIE, $\lambda_1 = 0.005, \lambda_2 = 0.05, \beta = 0.1$. $\lambda_1$ and $\lambda_2$ in Eq.(5) and $\alpha$ and $\lambda$ in Eq.(6).

B. Face recognition

We compare $D^2 L^2 R^2$ with DLRD_SR[5], FDDL[13], LDA[1], SRC[6] and LRC[25]. DLRD_SR applies low-rank regularization on the dictionary but without Fisher criterion on the coefficients, FDDL introduces Fisher criterion but has no low-rank requirement for the dictionary, LDA has Fisher criterion alone but no discriminative reconstruction error minimization. LRC is a recently proposed classifier which falls in the category of nearest subspace classification.
The ORL Face Database. The ORL database contains 400 images in total, ten different images for each of 40 different subjects. The background of the images is uniform and dark while the subjects are in frontal, upright posture. The images were shot under different lighting condition and with various facial expression and details[22]. For each class, we select half of the images randomly as training samples and rest as testing and repeat the experiment on five random splits. The images are normalized to 32 × 32. The images are manually corrupted by an unrelated block image at a random location. Fig. 2 shows an example of images with 20% block corruptions. We list the recognition accuracies under different levels of occlusions in Table II. From the table, we can see that our algorithm constantly performs the best under different levels of corruptions (> 0%). FDDL achieves the best result when there is no corruption, however, when the percentage of occlusions increases, performance of FDDL along with that of LRC, SRC and LDA drops rapidly, but \(D^2L^2R^2\) and DLRD_SR can still obtain much better recognition rates. This demonstrates the effectiveness of low-rank regularization when noise exists. Comparing \(D^2L^2R^2\) with DLRD_SR, \(D^2L^2R^2\) performs better due to the Fisher criterion on the coefficients.

The CMU PIE database consists of 41368 images of 68 subjects, each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions[24]. We use the first 15 subjects and select those images with frontal position and various expression and illumination. Each subject has 50 images, and we randomly select 10 images as training and the rest as testing and repeat our experiment on five random splits. The images are normalized to size 32 × 32. We replace a certain percentage of randomly selected pixels of each image with pixel value 255. Fig. 3 exemplifies random pixel corruption on both training and testing face samples. Fig. 4 shows the recognition accuracy under various noise percentage. \(D^2L^2R^2\) performs the best most of the time, but when there is no corruption or the percentage of corruption is very small, \(D^2L^2R^2\) cannot beat FDDL. We see that the low-rank regularization doesn’t help much in this case, but can on the hand degrade the performance. However, compared with DLRD_SR, \(D^2L^2R^2\) still obtain better accuracy due to the benefit from Fisher criterion. This can also be validated from LDA’s good performance.

Extended Yale B. The Extended Yale B dataset contains 2414 frontal-face images of 38 subjects captured under various laboratory-controlled lighting conditions [23]. We choose the first 15 subjects and each subject has around 60 images. We randomly take half as training samples, and the rest as testing samples and repeat the experiment five times. The images are normalized to size 32 × 32. We replace a certain percentage of randomly selected pixels from the images with pixel value 255. We list the recognition accuracies in Table III. Both \(D^2L^2R^2\) and DLRD_SR perform well when noise exists whereas recognition rates of LRC, SRC and LDA decrease fast with increasing

### Table II

<table>
<thead>
<tr>
<th>Occlusions</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D^2L^2R^2)[ours]</td>
<td>94.3</td>
<td>91.1</td>
<td>82.6</td>
<td>77.2</td>
<td>68.9</td>
<td>59.8</td>
</tr>
<tr>
<td>DLRD_SR[5]</td>
<td>93.2</td>
<td>90.4</td>
<td>81.3</td>
<td>76.5</td>
<td>67.9</td>
<td>57.3</td>
</tr>
<tr>
<td>FDDL[13]</td>
<td>96.8</td>
<td>86.8</td>
<td>74.5</td>
<td>61.9</td>
<td>49.2</td>
<td>36.9</td>
</tr>
<tr>
<td>LRC[25]</td>
<td>91.2</td>
<td>83.6</td>
<td>70.3</td>
<td>62.6</td>
<td>47.2</td>
<td>40.1</td>
</tr>
<tr>
<td>SRC[6]</td>
<td>90.3</td>
<td>79.8</td>
<td>63.0</td>
<td>53.9</td>
<td>38.2</td>
<td>26.7</td>
</tr>
<tr>
<td>LDA[1]</td>
<td>91.7</td>
<td>72.2</td>
<td>54.8</td>
<td>39.4</td>
<td>26.1</td>
<td>19.7</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** Example of ORL images with 20% block corruptions. Top: original images. Bottom: corresponding corrupted images with 20% block corruptions.

**Fig. 3.** Example of 30% random pixel corruptions on both training (1st row) and testing (2nd row) samples from PIE database.

**Fig. 4.** Average recognition accuracy on PIE database with different percentage of pixel corruptions (averaging over five random splits).
noise, which demonstrates the superiority of low-rank regularization in terms of handling noise.

C. Results discussion

Both $D^2 L^2 R^2$ and DLRD_SR perform better under noise condition from the results of the three datasets, which no doubt demonstrates low-rank regularization’s advantage in dealing with noise. Comparing $D^2 L^2 R^2$ with DLRD_SR, the former one can achieve better results almost all the time. This is due to the Fisher discriminant function on the coefficient matrix, which can make the dictionary learned more discriminative. However, this function is defined on the coefficients of training set, so when the training set is not sufficient compared to the testing set, $D^2 L^2 R^2$ and FDDL might not perform that well.

VII. CONCLUSION

This paper presents an algorithm to learn a discriminative dictionary with low-rank regularization ($D^2 L^2 R^2$) for face recognition. The $D^2 L^2 R^2$ can learn a structured dictionary where each sub-dictionary is of low rank, which can separate the noise information in the training samples. The discrimination power comes as two fold: First, the correlation between each sub-dictionary and samples from other classes is minimized; Second, Fisher discrimination criterion is applied to the coding coefficient matrix. We verify our algorithm in face recognition, and the experimental results clearly demonstrate its superiority to many other state-of-the-art methods, especially when there is noise or corruption present.

ACKNOWLEDGMENTS

This research is supported in part by the NSF CNS 1135660 and 1314484, Office of Naval Research award N00014-12-1-0125, Air Force Office of Scientific Research award FA9550-12-1-0201, and IC Postdoctoral Research Fellowship award 2011-11071400006.

REFERENCES


