Robust Face Recognition With Structurally Incoherent Low-Rank Matrix Decomposition

Chia-Po Wei, Chih-Fan Chen, and Yu-Chiang Frank Wang

Abstract—**For the task of robust face recognition, we particularly focus on the scenario in which training and test image data are corrupted due to occlusion or disguise. Prior standard face recognition methods like Eigenfaces or state-of-the-art approaches such as sparse representation-based classification did not consider possible contamination of data during training, and thus their recognition performance on corrupted test data would be degraded. In this paper, we propose a novel face recognition algorithm based on low-rank matrix decomposition to address the aforementioned problem. Besides the capability of decomposing raw training data into a set of representative bases for better modeling the face images, we introduce a constraint of structural incoherence into the proposed algorithm, which enforces the bases learned for different classes to be as independent as possible. As a result, additional discriminating ability is added to the derived base matrices for improved recognition performance. Experimental results on different face databases with a variety of variations verify the effectiveness and robustness of our proposed method.**

Index Terms—**Face recognition, low-rank matrix decomposition, structural incoherence.**

I. Introduction

AMONG biometric approaches for identity recognition, the use of face images can be considered as the most popular one due to its low intrusiveness and high uniqueness [1]. Other physiological or behavioral biometrics (e.g., fingerprint or gait recognition) often require cooperative subjects, which might not always be feasible for real-world applications. Generally, face images can be acquired actively by the user, or they can be captured passively by surveillance cameras. With the increasing needs for security-related applications such as computational forensics and anti-terrorism, face recognition has been an active topic for researchers in the areas of computer vision and image processing.

To address the problem of face recognition, one typically focuses on the extraction of facial features from training image data, and the learning of associated classification models. Unseen test data from the same subjects of interest will be used to evaluate the recognition performance. It is worth noting that, most prior works on face recognition assume that both training and test image data are under pose, illumination, or expression variations. To further assess the *robustness* of the designed face recognition algorithm, only test images are considered to be *corrupted* due to occlusion or disguise in recent literatures [2] and [3]. In other words, while the test data might be corrupted, most prior works consider the training face images to be taken under a well controlled setting (i.e., under reasonable illumination, pose, etc. variations *without* occlusion or disguise). To apply these prior approaches for practical face recognition scenarios, one will need to discard corrupted training images and thus inevitably encounter small sample size and over-fitting problems. Moreover, the disregard of corrupted training face images might give up some valuable information for recognition. For example, in forensic identification, any available information extracted from face images could be the key to identification for forensic investigators [4].

Generally, Eigenfaces [5], Fisherfaces [6], or Laplacianfaces [7] are common face recognition techniques which aim at extracting proper features from face images for recognition using nearest neighbor (NN) or support vector machines (SVM). Although Fisherfaces can extract discriminating features for face recognition, limited number of training data would cause problems when calculating the inverse of the data matrices. To tackle this problem, Jiang *et al.* [8] decompose the derived eigenspace and utilize an eigenspectrum model for improved recognition. Nevertheless, the above approaches are not designed to deal with corrupted training data, and thus their recognition results will be sensitive to the presence of sparse/extreme noise such as occlusion and disguise in face images. We note that, recent methods based on robust PCA have been proposed to deal with data in which sparse noise is presented [9]–[11]. Among them, low-rank matrix recovery can be solved in polynomial-time and has been shown to provide promising results [11]. Although such methods have been shown to be capable of identifying a set of representative bases from corrupted data, there is no guarantee that such a basis set would serve for classification purposes.

Recently, sparse representation-based classification (SRC) [2] has shown very promising results on face recognition, which considers each test image as a sparse linear combination of the training instances. SRC solves an ℓ_1-minimization problem for a test input by deriving the sparse coefficients for the training data, and recognition is achieved based on the minimum class-wise reconstruction error.

Manuscript received October 16, 2013; revised April 7, 2014; accepted May 28, 2014. Date of publication June 6, 2014; date of current version June 24, 2014. This work was supported by the National Science Council of Taiwan under Grant NSC-102-2221-E-001-005-MY2. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikolaos V. Boulgouris.

C.-P. Wei and Y.-C. F. Wang are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan (e-mail: cpwei@citi.sinica.edu.tw; ycwang@citi.sinica.edu.tw).

C.-F. Chen is with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA (e-mail: chihfanc@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2329451

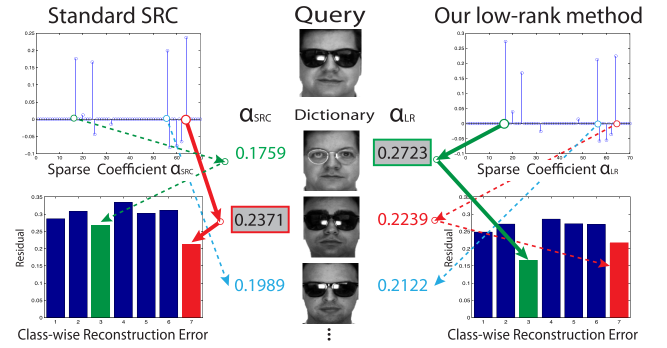


Fig. 1. Comparison between the standard SRC and our method. The standard SRC classifies the test input as the class with most similar training images even if they are occluded (e.g. due to sunglasses), while our approach alleviates this problem and is robust to such occlusions presented in both training and test data.

It has been shown in [2] that if the test image is corrupted due to face occlusion, SRC is able to exhibit excellent robustness and produces promising performance. However, besides requiring the training images to be well aligned for reconstruction purposes, SRC does not allow corrupted data for training (otherwise the performance will be degraded as we verify in our experiments). Inspired by SRC, Wagner *et al.* [3] propose a sequential ℓ_1-minimization algorithm to deal with face misalignment problems, and design a projector-based illumination system to tackle illumination variations. To better handle occlusion, Zhou *et al.* [12] integrate a Markov random field for contiguous occlusion into SRC. Yang *et al.* [13], [14] also modify the SRC framework for handling outliers such as occlusions in face images. Unfortunately, the above SRC based methods might not generalize well if *both* training and test images are corrupted, since none of them consider the possible corruption of training face images.

In this paper, we address the problem of robust face recognition, in which *both* training and test image data are corrupted. We do *not* have the prior knowledge on the type of corruptions (e.g., due to sunglasses, scarf, etc.). We will show that the direct use of dimension reduction techniques such as Eigenfaces for training and testing would degenerate the performance with the presence of corrupted data (see the left half of Fig. 1 for example). To address this problem, we propose a novel low-rank matrix decomposition algorithm with structural incoherence, which allows us to convert raw face image data into a set of representative bases with a corresponding sparse error matrix. We further regularize the derived basis matrix with a structural incoherence constraint. The introduction of such incoherence between the basis extracted from different classes would provide additional discriminating ability to our framework. It is worth noting that we are among the first applying low-rank techniques for face recognition problems. More importantly, our proposed method particularly serves for recognition purposes (not just for reconstruction), as illustrated in the right half of Fig. 1. Our experiments will verify the effectiveness and robustness of our method, and we will show that our method outperforms existing SRC-based approaches when both training and test image data are corrupted by a variety of noise/variations.

The remaining of this paper is organized as follows. Section II reviews related works on low-rank matrix recovery, and discusses the use of SRC for face recognition. In Section III, we present our proposed algorithm based on low-rank matrix decomposition and structural incoherence, including the optimization details. Experimental results on four face image databases are presented in Section IV. Finally, Section V concludes this paper.

II. Related Work

A. Robust PCA and Low-Rank Matrix Recovery

Principal component analysis (PCA) is a popular dimension reduction technique for data analysis applications such as reconstruction and classification. In spite of its effectiveness, PCA is known to be sensitive to sparse errors with large magnitudes [15]. A number of approaches have been proposed in literatures to address this problem, including the introduction of influence functions [9], alternating minimization techniques [10], and low-rank matrix recovery [11] (noted as LR in the remaining for this paper for conciseness). Among these methods (known as robust PCA), LR has been observed to be solved in polynomial time with performance guarantees [11]. Since our work in this paper is inspired by low-rank matrix decomposition, we briefly review its formulation for the sake of completeness.

Low-rank matrix recovery aims at decomposing a data matrix $\mathbf{D}$ into $\mathbf{A}+\mathbf{E}$, in which $\mathbf{A}$ is a low-rank matrix and $\mathbf{E}$ is the associated sparse error. More precisely, to derive the low-rank approximation of the input data matrix $\mathbf{D}$, LR minimizes the rank of matrix $\mathbf{A}$ while reducing the ℓ_0-norm of $\mathbf{E}$. As a result, one will need to solve the following minimization problem:

$$\min_{\mathbf{A},\mathbf{E}} \operatorname{rank}(\mathbf{A}) + \lambda\|\mathbf{E}\|_0 \quad \text{s.t.} \quad \mathbf{D}=\mathbf{A}+\mathbf{E}. \tag{1}$$

From the above formulation, we note that $\|\mathbf{E}\|_0$ calculates the number of non-zero elements in $\mathbf{E}$. Since solving (1) involves the low-rank matrix completion and the ℓ_0-norm minimization problems, it is NP-hard and thus is not easy to solve. To convert (1) into a more tractable optimization problem, Candès *et al.* [11] relax (1) by replacing rank($\mathbf{A}$) with its nuclear norm $\|\mathbf{A}\|_*$ (i.e., the sum of the singular values of $\mathbf{A}$). Instead of solving the minimization of ℓ_0-norm $\|\mathbf{E}\|_0$, that of ℓ_1-norm $\|\mathbf{E}\|_1$ is now considered (i.e., the sum of the absolute values of each entry in $\mathbf{E}$). Consequently, the convex relaxation of (1) has the following form:

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 \quad \text{s.t.} \quad \mathbf{D}=\mathbf{A}+\mathbf{E}. \tag{2}$$

It is shown in [11] that solving this convex relaxation version is equivalent to solving the original low-rank matrix approximation problem, as long as the rank of $\mathbf{A}$ to be recovered is not too large, and the number of non-zero elements in $\mathbf{E}$ is reasonably small (i.e., to be sufficiently sparse). To solve the optimization problem of (2), the technique of augmented Lagrange multipliers (ALM) [16] has been applied due to its computational efficiency. While many image processing applications can be casted as the low-rank matrix recovery problems (e.g., image alignment [17], subspace segmentation [18], collaborative filtering [11], and image tag transduction [19]), we are among the first to apply LR-based techniques for addressing the problem of robust face recognition.

B. Sparse Representation-Based Classification

Wright *et al.* [2] recently proposed a sparse representation-based classification (SRC) algorithm for face recognition. SRC considers each test image as a sparse linear combination of training image data by solving an ℓ_1-minimization problem. Very promising results were reported in [2], even if test image data are corrupted due to occlusion or noise. Several works have been proposed to further extend SRC for improved performance. For example, Yuan and Yan [20] utilized an $\ell_{1,2}$ mixed-norm regularization for computing the joint sparse representation of different features for visual signals. Jenatton *et al.* [21] considered a tree-structured sparse regularization for hierarchical sparse coding. Chao *et al.* [22] integrated the $\ell_{1,2}$ norm with a data locality constraint for improved face recognition.

Since we apply the SRC as our classification rule, we now review this algorithm. Suppose that there exist m training images from N object classes, and each class j has m_j images. Let $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N] \in \mathbb{R}^{d \times m}$ be the training set, where $\mathbf{D}_j \in \mathbb{R}^{d \times m_j}$ contains training images of the jth class as its columns, and d is the dimension of each image. Given a test image $\mathbf{y} \in \mathbb{R}^{d \times 1}$, the SRC algorithm calculates the sparse representation $\boldsymbol{\alpha}$ of $\mathbf{y}$, which is computed via the ℓ_1 minimization process over the entire training image set. More precisely, SRC solves the following optimization problem for deriving the sparse representation $\boldsymbol{\alpha}$:

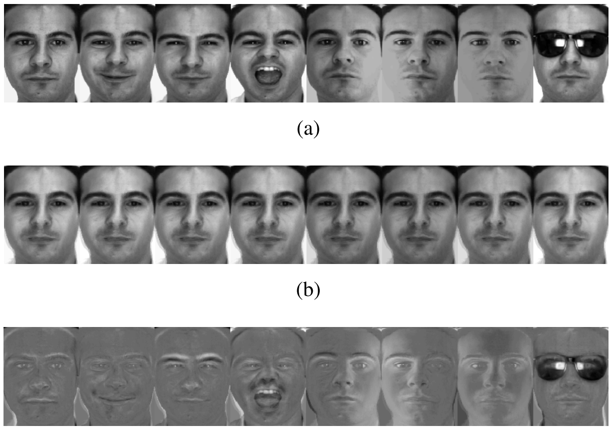
$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1. \tag{3}$$

Let $\delta_i(\boldsymbol{\alpha})$ be a vector in $\mathbb{R}^{m \times 1}$ with nonzero entries as those in $\boldsymbol{\alpha}$ that are associated with class i. Once (3) is solved, the test input $\mathbf{y}$ will be recognized as class j if it satisfies

$$j = \arg\min_i \|\mathbf{y} - \mathbf{D}\,\delta_i(\boldsymbol{\alpha})\|_2^2. \tag{4}$$

In other words, the test image $\mathbf{y}$ will be assigned to the class based on a class-wise minimum reconstruction error. The motivation behind this classification strategy is that the test image $\mathbf{y}$ should lie in the space spanned by the columns $\mathbf{D}_j$ of class j. As a result, most non-zero elements of $\boldsymbol{\alpha}$ will mainly be presented in the non-zero elements of $\delta_j(\boldsymbol{\alpha})$, which results in the minimum reconstruction error. The framework of SRC is depicted by the red arrows in Fig. 3.

Although impressive face recognition results were reported by SRC [2], SRC still requires *clean* (i.e., unoccluded) face images for training. In other words, it might not be preferable for real-world scenarios when corrupted face images are collected during training. As later verified by our experiments, this practical training scenario would result in degraded recognition performance for SRC due to the tendency of recognizing test images as the training ones with the same type of corruption presented. In the following section, we will introduce our proposed algorithm for robust face recognition, in which both training and test image data can be corrupted.



(c)

Fig. 2. Example results of low-rank matrix recovery. (a) Original images **D**. (b) Low-rank and approximated images **A** of (a). (c) Sparse error images **E** of (a).

III. Low-Rank Matrix Recovery With Structural Incoherence for Face Recognition

A. Face Recognition With Low-Rank Matrix Recovery

For face recognition in real-world scenarios, we cannot expect the training image data to be always collected under a well-controlled setting. In addition to illumination, pose, or expression variations, it is possible that one can be taking a scarf, gauze mask, or sunglasses, when his/her face image is taken by the camera. Using such images for training would make the learned face recognition algorithm overfit the extreme noise of occlusion, instead of modeling the face of the subject. As a result, the resulting recognition performance will be degraded.

As discussed earlier in Section II-A, we note that low-rank matrix recovery (LR) can be applied to alleviate the aforementioned problem. Recall that LR decomposes the collected data matrix into two different parts, one is a representative basis matrix with a minimum rank and the other is the corresponding sparse error matrix. It is worth noting that, in order to apply LR for face recognition, the face image data needs to be registered prior to the procedure of low-rank matrix decomposition. In our work, we only consider face images of frontal views (i.e., no pose variations), so that the extracted low-rank matrix would preserve the structure of the face images.

When applying LR for face recognition with N subjects of interest, one can collect training data $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N]$, where $\mathbf{D}_i$ is the training data matrix (with the presence of occlusion or disguise) for subject i, as shown in Fig. 2(a). By performing low-rank matrix recovery, the data matrix $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N]$ will be decomposed into a low-rank matrix $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N]$ and the sparse error matrix $\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_N]$. As shown in Fig. 2(b), the representative images in $\mathbf{A}$ can be considered as preprocessed data with sparse noise removed (see the corresponding images in Fig. 2(c)). Comparing Figs. 2(a) and 2(b), we can see that the low-rank matrix $\mathbf{A}$ has a better representative ability than the original data $\mathbf{D}$ does in describing the face images of the subject of interest.

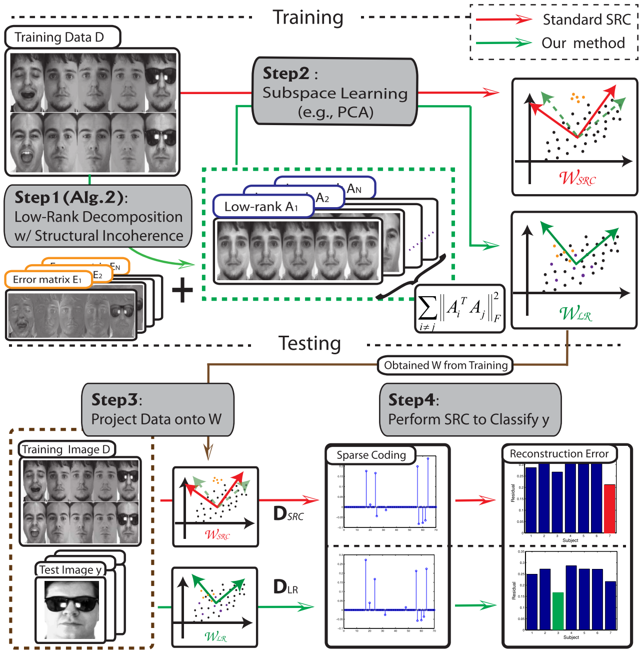


Fig. 3. Illustration of our proposed method. Note that we promote the structural incoherence between low-rank matrices for better modeling and recognizing face images.

Since the face images are typically with high dimensionality, standard dimension reduction techniques such as PCA are typically applied to the face image data before training and testing. Instead of using the Eigenfaces calculated by from the original data matrix $\mathbf{D}$ as most prior works did, one can apply PCA on the low-rank matrix $\mathbf{A}$ (as shown in *Step 2* of Fig. 3), and the resulting subspace can be applied as the dictionary for training and testing purposes (see *Step 3* in Fig. 3). Finally, one can apply SRC and the derived dictionary to classify test inputs, which performs classification based on class-wise minimum reconstruction error (as depicted by *Step 4* in Fig. 3). Later in Section IV, in contrast to the direct use of raw data $\mathbf{D}$ we will verify that LR better handles the problem in which the input training data is under severe illumination variations or is corrupted by occlusion or disguise. Algorithm 1 and Fig. 3 summarize the procedure of integrating low-rank matrix recovery and SRC for face recognition.

B. Low-Rank Matrix Decomposition With Structural Incoherence

1) Proposed Formulation: Although we show that LR is able to process the raw data matrix $\mathbf{D}$ and to produce a low-rank matrix $\mathbf{A}$ for better representation ability, the face images of different subjects might share common (correlated) features (e.g., the locations of eyes, nose, etc.) and thus the derived matrix $\mathbf{A}$ does not contain sufficient discriminating information. Inspired by [23], we propose to promote the incoherence between the derived low-rank matrices of different classes for classification purposes. The introduction of such incoherence would prefer the resulting low-rank matrices to be as independent as possible. Therefore, commonly shared features across different classes will be suppressed while the independent/discriminating ones will be preserved.

Algorithm 1 LR for Face Recognition

```
Input: Training data $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N]$ from $N$ subjects and
    the test input $\mathbf{y}$
    Step 0: Normalize $\mathbf{y}$ and the columns of $\mathbf{D}$ to have unit $\ell_2$-norm
    Step 1: Perform LR on $\mathbf{D}$
    for $i = 1 : N$ do
        $\min_{\mathbf{A}_i,\mathbf{E}_i} \|\mathbf{A}_i\|_* + \lambda\|\mathbf{E}_i\|_1$  s.t.  $\mathbf{D}_i = \mathbf{A}_i + \mathbf{E}_i$
    end for
    Step 2: Calculate principal components $\mathbf{W}$ of $\mathbf{A}$
    $\mathbf{W} \leftarrow \mathcal{PCA}(\mathbf{A})$
    Step 3: Project $\mathbf{D}$ and $\mathbf{y}$ onto $\mathbf{W}$
    $\mathbf{D}_p = \mathbf{W}^T(\mathbf{D} - \boldsymbol{\mu}\mathbf{1}^T)$ and $\mathbf{y}_p = \mathbf{W}^T(\mathbf{y} - \boldsymbol{\mu})$,
    where $\boldsymbol{\mu}$ is the mean of the column vectors of $\mathbf{A}$
    Step 4: Use SRC to classify $\mathbf{y}_p$
    $\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y}_p - \mathbf{D}_p\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$.
    for $i = 1 : N$ do
        $e(i) = \|\mathbf{y}_p - \mathbf{D}_p\,\delta_i(\boldsymbol{\alpha}^*)\|_2^2$
    end for
Output: identity$(\mathbf{y}) \leftarrow \arg\min_i e(i)$
```

As illustrated in *Step 1* of Fig. 3, our method aims at providing additional discriminating ability to the original LR models by promoting their structural incoherence, and the recognition performance is expected to be improved.

Based on the LR formulation in (2), we add a regularization term to the objective function and enforce the incoherence between different low-rank matrices. We now solve the following optimization problem:

$$\min_{\mathbf{A},\mathbf{E}} \sum_{i=1}^{N}\{\|\mathbf{A}_i\|_* + \lambda\|\mathbf{E}_i\|_1\} + \eta\sum_{j\neq i}\|\mathbf{A}_j^T\mathbf{A}_i\|_F^2$$
$$\text{s.t.}\quad \mathbf{D}_i = \mathbf{A}_i + \mathbf{E}_i \quad \text{for} \quad i = 1, 2, \ldots, N. \tag{5}$$

We note that the first term of the objective function in (5) performs the standard low-rank decomposition of the data matrix $\mathbf{D}$. The second term promotes the structural incoherence by summing up the Frobenius norms between different pairs of low-rank matrices $\mathbf{A}_i$ and $\mathbf{A}_j$, which is penalized by the parameter η balancing the low-rank matrix approximation and structural incoherence. We refer to (5) as our proposed low-rank matrix recovery with *structural incoherence*, which will be utilized to provide improved discrimination ability to the original LR model. Since the error matrix $\mathbf{E}$ in (5) is sparse (the same as (2)) and represents extreme noise such as occlusion and disguise presented in face images, we do not enforce extra regularization on $\mathbf{E}$.

While the minimization problem in (5) is nonconvex due to the product term $\mathbf{A}_j^T\mathbf{A}_i$, we do not solve all low-rank matrices $\mathbf{A}_i$ at once and choose to solve class-wise optimization problems across different classes. To be more specific, we iteratively solve the following minimization problem across different classes:

$$\min_{\mathbf{A}_i,\mathbf{E}_i} \|\mathbf{A}_i\|_* + \lambda\|\mathbf{E}_i\|_1 + \eta\sum_{j\neq i}\|\mathbf{A}_j^T\mathbf{A}_i\|_F^2$$
$$\text{s.t.}\quad \mathbf{D}_i = \mathbf{A}_i + \mathbf{E}_i. \tag{6}$$

For each iteration, we aim at solving the low-rank matrices for each class. That is, for class i, we fix $\mathbf{A}_j$ if $j \neq i$, and the variables to be minimized are $\mathbf{A}_i$ and $\mathbf{E}_i$. As a result, (6) turns into a convex optimization problem, and the solution of (6) is guaranteed to be a global minimizer. From (6), we see that the objective function includes the Frobenius norms of product terms of different matrix pairs. To make the optimization problem more tractable, our prior work in [24] applied the Cauchy-Schwarz inequality and replaced the term $\eta \sum_{j \neq i} \|\mathbf{A}_j^T \mathbf{A}_i\|_F^2$ with $\eta' \|\mathbf{A}_i\|_F^2$, in which the influence of low-rank matrices $\mathbf{A}_j$ is absorbed into the parameter η'. However, this relaxation only implicitly addresses the formulation of structural incoherence, and does not guarantee the resulting incoherence between $\mathbf{A}_j$ and $\mathbf{A}_i$. In this paper, we propose to solve the optimization problem of (6) without any relaxation or approximation. More specifically, we introduce auxiliary variables $\mathbf{B}_i$ to (6) to tackle the term of Frobenius norms of different matrix pairs, which leads to

$$\min_{\mathbf{A}_i,\mathbf{B}_i,\mathbf{E}_i} \|\mathbf{A}_i\|_* + \lambda \|\mathbf{E}_i\|_1 + \eta \sum_{j \neq i} \|\mathbf{A}_j^T \mathbf{B}_i\|_F^2$$
$$\text{s.t. } \mathbf{D}_i = \mathbf{A}_i + \mathbf{E}_i \text{ and } \mathbf{B}_i = \mathbf{A}_i. \quad (7)$$

From the above formulation, it is clear that the optimal solutions of (6) and (7) are the effectively same, and hence introducing auxiliary variables does not change the optimization problem that we aim to solve. The strategy of introducing auxiliary variables has been used in [18] for solving the low-rank representation problem for subspace segmentation.

2) Structural Incoherence for Improved Recognition: In our work, we add an additional regularization term $\eta \sum_{j \neq i} \|\mathbf{A}_j^T \mathbf{A}_i\|_F^2$ into the standard formulation of low-rank matrix decomposition (LR). Thus, our proposed algorithm aims at deriving low-rank representations for different classes while minimizing the structural incoherence (SI) between them. We note that, the proposed algorithm balances the low-rank matrix decomposition and the associated structural incoherence. While the former allows us to automatically disregard undesirable noisy patterns from face images, the latter introduces additional data separation between different classes. As a result, the resulting low-rank matrices are not only considered as features for describing face images, they are also utilized for recognizing faces of different subjects due to improved discriminating capabilities.

When addressing pattern recognition problems, it is always desirable to extract features which can be applied to solve the associated recognition task. While we advocate the structural incoherence between the derived low-rank matrices by minimizing their similarities (i.e., correlation), our algorithm effectively searches for data representations of different classes as distinct as possible. The introduced structural incoherence term is regularized by η, which balances between the representation and discrimination capabilities of the derived low-rank matrices for each class. As verified by our experiments, setting $\eta = 0$ would turn the proposed algorithm into the standard LR formulation, and it cannot achieve satisfactory recognition results as ours does.

It is worth noting that, the idea of introducing a regularization term on structural incoherence also appears in recent works on dictionary learning algorithms for image classification (e.g., digit [23], scene [25], or action recognition [26] problems). While aiming at observing dictionaries for solving the associated classification tasks, these approaches also enforce the structure incoherence between the dictionary atoms of different classes in the learning process. In other words, the structural incoherence between the derived dictionaries would imply and thus produce coefficients of different classes as different as possible. When applying the encoded coefficients as features for performing recognition, improved recognition performance have been reported in [23], [25], and [26].

Generally, challenges of face recognition lie in the need to handle image variants due to illumination and expression changes, plus the possible presence of corruptions. Therefore, our proposed low-rank based algorithm with introduced structural incoherence term would produce preferable image features for solving the recognition task.

C. Probabilistic Point of View

We now provide theoretical analysis for supporting our LRSI over the standard LR from probabilistic point of view. We have the training image data as $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N]$, where N is the number of classes. For each class i, we decompose $\mathbf{D}_i$ into $\mathbf{A}_i + \mathbf{E}_i$, where $\mathbf{A}_i$ and $\mathbf{E}_i$ represent the low-rank structure and the corresponding sparse errors of $\mathbf{D}_i$, respectively. Using the Bayes' rule, we have

$$\begin{aligned}\log P(\mathbf{A}, \mathbf{E} \mid \mathbf{D}) &+ \log P(\mathbf{D}) \\ &= \log P(\mathbf{D} \mid \mathbf{A}, \mathbf{E}) + \log P(\mathbf{A}, \mathbf{E}),\end{aligned} \quad (8)$$

in which $\mathbf{A}$ and $\mathbf{E}$ denote the collections of all $\mathbf{A}_i$ and $\mathbf{E}_i$, respectively. In (8), $P(\mathbf{A}, \mathbf{E} \mid \mathbf{D})$ is the posterior probability given the input training data, and $P(\mathbf{D} \mid \mathbf{A}, \mathbf{E})$ is the likelihood function. We consider $P(\mathbf{D})$ as the evidence of $\mathbf{D}$, and $P(\mathbf{A}, \mathbf{E})$ reflects the prior of $(\mathbf{A}, \mathbf{E})$. Based on the maximum a-posteriori (MAP) estimates (see [27, Sec. 1.2.3]), we aim at solving the following optimization problem:

$$\begin{aligned}(\mathbf{A}_{MAP}, \mathbf{E}_{MAP}) &= \arg\max_{\mathbf{A},\mathbf{E}} \log P(\mathbf{D} \mid \mathbf{A}, \mathbf{E}) + \log P(\mathbf{A}, \mathbf{E}) \\ &= \arg\min_{\mathbf{A},\mathbf{E}} - \log P(\mathbf{D} \mid \mathbf{A}, \mathbf{E}) - \log P(\mathbf{A}, \mathbf{E}).\end{aligned} \quad (9)$$

Note that $\log P(\mathbf{D})$ is disregarded in (9) since it is independent of $(\mathbf{A}, \mathbf{E})$. The posterior probability $\log P(\mathbf{A}, \mathbf{E} \mid \mathbf{D})$ is related to the augmented Lagrange function in (17), and is defined as the summation of the terms with $\mathbf{D}_i$ in (17) over $i = 1, 2, \ldots, N$. Since $\mathbf{A}$ and $\mathbf{E}$ are meant to describe distinct characteristics of $\mathbf{D}$, $\mathbf{A}$ and $\mathbf{E}$ will be independent to each other. In other words, we have

$$\log P(\mathbf{A}, \mathbf{E}) = \log P(\mathbf{A}) P(\mathbf{E}) = \log P(\mathbf{A}) + \log P(\mathbf{E}). \quad (10)$$

Since the sparse error matrices of each class have random distributions, we further derive $\log P(\mathbf{E})$ as follows:

$$\begin{aligned}\log P(\mathbf{E}) &= \log P(\mathbf{E}_1) P(\mathbf{E}_2) \cdots P(\mathbf{E}_N) \\ &= \sum_{i=1}^{N} \log P(\mathbf{E}_i) := -\lambda \sum_{i=1}^{N} \|\mathbf{E}_i\|_1.\end{aligned} \quad (11)$$

Note that the smaller the value of $\|\mathbf{E}_i\|_1$, the larger the probability of $\mathbf{E}_i$. Hence, solving the above optimization problem would result in a minimized $\mathbf{E}_{MAP}$, which represents the sparse error components of $\mathbf{D}$.

It is worth noting that, the difference between LR and our LRSI lies in the statistical assumption on the observed low-rank matrices. More specifically, LR assumes that $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N$ are independent, and thus

$$\begin{aligned}\log P(\mathbf{A}) &= \log P(\mathbf{A}_1)P(\mathbf{A}_2)\cdots P(\mathbf{A}_N)\\ &= \sum_{i=1}^{N} \log P(\mathbf{A}_i) := -\sum_{i=1}^{N} \|\mathbf{A}_i\|_*. \end{aligned} \quad (12)$$

Note that smaller $\|\mathbf{A}_i\|_*$, implying $\mathbf{A}_i$ with a lower rank, would correspond to larger $P(\mathbf{A}_i)$. In contrast to LR, our LRSI relaxes the above assumption and allows $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N$ to be dependent. This is practical for face recognition, since in addition to the low-rank constraint, our goal is to observe the low-rank representations of different classes which are as distinct (but not necessarily independent) to each other as possible. Therefore, we rewrite $\log P(\mathbf{A})$ as:

$$\log P(\mathbf{A}) = \log P(\{\mathbf{A}_j, j \neq i\} \mid \mathbf{A}_i) + \log P(\mathbf{A}_i), \quad (13)$$

which holds for $i = 1, 2, \ldots, N$. We note that, equation (13) would reduce to the first equality in (12) if $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N$ are independent. In view of (13), we can further rewrite $\log P(\mathbf{A})$ as:

$$\begin{aligned}\log P(\mathbf{A}) &= \frac{N \log P(\mathbf{A})}{N}\\ &= \frac{1}{N}\sum_{i=1}^{N} \log P(\mathbf{A}_i) + \log P(\{\mathbf{A}_j, j \neq i\} \mid \mathbf{A}_i)\\ &:= \sum_{i=1}^{N}\left[-\|\mathbf{A}_i\|_* - \eta \sum_{j \neq i} \|\mathbf{A}_j^T \mathbf{A}_i\|_F^2\right], \end{aligned} \quad (14)$$

in which the term $-\|\mathbf{A}_i\|_*$ corresponds to $\frac{1}{N} \log P(\mathbf{A}_i)$, and the term $-\eta \sum_{j \neq i} \|\mathbf{A}_j^T \mathbf{A}_i\|_F^2$ corresponds to $\frac{1}{N}$ log $P(\{\mathbf{A}_j, j \neq i\} \mid \mathbf{A}_i)$. The conditional probability log $P(\{\mathbf{A}_j, j \neq i\} \mid \mathbf{A}_i)$ determines the degree of the incoherence between low-rank matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N$, and the parameter η is the weight (or penalty) for the conditional probability (see Section IV-C.3). When setting the value of η to zero, the conditional probability $\log P(\{\mathbf{A}_j, j \neq i\} \mid \mathbf{A}_i)$ vanishes, and our definition of $\log P(\mathbf{A})$ reduces to the case of the standard LR.

Since the choice of the prior belief on $\mathbf{A}$ (i.e., $\log P(\mathbf{A})$) affects the MAP solution, the design of $\log P(\mathbf{A})$ is the key to achieving satisfactory recognition results. To be more precise, better recognition performance can be expected, if $\log P(\mathbf{A})$ is properly designed for the task of face recognition. Because the standard LR was not proposed/designed to address pattern recognition problems, it does not take the dependency between the low-rank matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N$ of different classes into consideration (i.e., LR simply assumes that such low-rank matrices are independent). To improve LR, we consider the relationship between the observed low-rank matrices by introducing the structural incoherence regularization term, which not only corresponds to the conditional probability term in (14) but also addresses the recognition task. With this regularization term, our algorithm is able to obtain better MAP estimates than the standard LR does on recognition problems, and this has been successfully verified by our experiments.

Algorithm 2 Solving LR With Structural Incoherence

```
Input: Data matrix $\mathbf{D}$ and parameters $\eta$ and $\rho$ ($\rho > 1$)
  Use Step1 in Alg. 1 to initialize $\mathbf{A}^0$, $\mathbf{B}^0$, $\mathbf{E}^0$, $\mathbf{Y}^0$, $\mathbf{Z}^0$, $\mu^0$
  while not converged do
    for $i = 1 : N$ do
      while not converged do
        $\mathbf{A}_i^{k+1} = \arg\min_{\mathbf{A}_i} L(\mathbf{A}_i, \mathbf{B}_i^k, \mathbf{E}_i^k, \mathbf{Y}_i^k, \mathbf{Z}_i^k, \mu^k)$
        $\mathbf{E}_i^{k+1} = \arg\min_{\mathbf{E}_i} L(\mathbf{A}_i^{k+1}, \mathbf{B}_i^k, \mathbf{E}_i, \mathbf{Y}_i^k, \mathbf{Z}_i^k, \mu^k)$
        $\mathbf{B}_i^{k+1} = \arg\min_{\mathbf{B}_i} L(\mathbf{A}_i^{k+1}, \mathbf{B}_i, \mathbf{E}_i^{k+1}, \mathbf{Y}_i^k, \mathbf{Z}_i^k, \mu^k)$
        $\mathbf{Y}_i^{k+1} = \mathbf{Y}_i^k + \mu^k(\mathbf{D}_i - \mathbf{A}_i^{k+1} - \mathbf{E}_i^{k+1})$
        $\mathbf{Z}_i^{k+1} = \mathbf{Z}_i^k + \mu^k(\mathbf{B}_i^{k+1} - \mathbf{A}_i^{k+1})$
        $\mu^{k+1} = \rho\mu^k$
      end while
    end for
  end while
Output: $\mathbf{A}$ and $\mathbf{E}$
```

D. Optimization via ALM

Augmented Lagrange multipliers (ALM) have been applied to solve the standard LR problem [11], [16]. In this subsection, we will detail how we extend ALM to solve our proposed LR formulation with regularization on structural incoherence.

Denote the objective function in (7) and the equality constraints in (7) as

$$\begin{aligned} f(\mathbf{X}) &= \|\mathbf{A}_i\|_* + \lambda\|\mathbf{E}_i\|_1 + \eta \sum_{j \neq i} \|\mathbf{A}_j^T \mathbf{B}_i\|_F^2,\\ h_1(\mathbf{X}) &= \mathbf{D}_i - \mathbf{A}_i - \mathbf{E}_i,\ \ h_2(\mathbf{X}) = \mathbf{B}_i - \mathbf{A}_i,\\ h(\mathbf{X}) &= [h_1(\mathbf{X});\ h_2(\mathbf{X})], \end{aligned} \quad (15)$$

and let $\mathbf{X} = (\mathbf{A}_i, \mathbf{B}_i, \mathbf{E}_i)$. For an optimization problem in which $f(\mathbf{X})$ is to be minimized with the constraint $h(\mathbf{X}) = \mathbf{0}$, its ALM function is formulated as follows:

$$L(\mathbf{X}, \mathbf{Y}, \mu) = f(\mathbf{X}) + \langle \boldsymbol{\Phi}, h(\mathbf{X})\rangle + \frac{\mu}{2}\|h(\mathbf{X})\|_F^2, \quad (16)$$

where $\boldsymbol{\Phi} = (\mathbf{Y}_i, \mathbf{Z}_i)$ is a Lagrange multiplier, and μ is a penalty parameter. After substituting (15) into (16), the augmented Lagrangian function for (7) has the form

$$\begin{aligned} &L(\mathbf{A}_i, \mathbf{B}_i, \mathbf{E}_i, \mathbf{Y}_i, \mathbf{Z}_i, \mu)\\ &\quad = \|\mathbf{A}_i\|_* + \lambda\|\mathbf{E}_i\|_1 + \eta \sum_{j \neq i} \|\mathbf{A}_j^T \mathbf{B}_i\|_F^2\\ &\qquad + \langle \mathbf{Z}_i, \mathbf{B}_i - \mathbf{A}_i\rangle + \frac{\mu}{2}\|\mathbf{B}_i - \mathbf{A}_i\|_F^2\\ &\qquad + \langle \mathbf{Y}_i, \mathbf{D}_i - \mathbf{A}_i - \mathbf{E}_i\rangle + \frac{\mu}{2}\|\mathbf{D}_i - \mathbf{A}_i - \mathbf{E}_i\|_F^2. \end{aligned} \quad (17)$$

We apply the alternating direction algorithm [28] to find the minimizer of (17). The pseudo code of our proposed algorithm is shown in Algorithm 2. We now discuss how we update/solve the above variables in each iteration.

1) Updating $\mathbf{A}_i$*:* To update $\mathbf{A}_i^{k+1}$ for class i at the $(k+1)th$ iteration in Algorithm 2, we have fixed variables other than $\mathbf{A}_i$ and solve the following problem accordingly:

$$\begin{aligned}\mathbf{A}_i^{k+1} &= \arg\min_{\mathbf{A}_i} L(\mathbf{A}_i, \mathbf{B}_i^k, \mathbf{E}_i^k, \mathbf{Y}_i^k, \mathbf{Z}_i^k, \mu^k)\\ &= \arg\min_{\mathbf{A}_i} \|\mathbf{A}_i\|_* + \langle \mathbf{Z}_i^k, \mathbf{B}_i^k - \mathbf{A}_i\rangle + \frac{\mu^k}{2}\|\mathbf{B}_i^k - \mathbf{A}_i\|_F^2\\ &\quad + \langle \mathbf{Y}_i^k, \mathbf{D}_i - \mathbf{A}_i - \mathbf{E}_i^k\rangle + \frac{\mu^k}{2}\|\mathbf{D}_i - \mathbf{A}_i - \mathbf{E}_i^k\|_F^2\\ &= \arg\min_{\mathbf{A}_i} \epsilon\|\mathbf{A}_i\|_* + \frac{1}{2}\|\mathbf{X}_a - \mathbf{A}_i\|_F^2,\end{aligned}$$

where $\epsilon = (2\mu^k)^{-1}$ and $\mathbf{X}_a = 0.5(\mathbf{D}_i - \mathbf{E}_i^k + (1/\mu^k)\mathbf{Y}_i^k + \mathbf{B}_i^k + (1/\mu^k)\mathbf{Z}_i^k)$. As shown in [16, Sec. 2.1], the closed-form solution to the above problem is given by

$$\mathbf{A}_i^{k+1} = \mathbf{U}\, T_\epsilon[\mathbf{S}]\mathbf{V}^T, \quad (18)$$

where $\mathbf{USV}^T$ is the singular value decomposition of $\mathbf{X}_a$, and the operator $T_\epsilon[\mathbf{S}]$ in (18) is defined by element-wise ϵ thresholding of $\mathbf{S}$, i.e., $T_\epsilon[\mathbf{S}](i,j) = t_\epsilon[\mathbf{S}(i,j)]$, where $t_\epsilon[s]$ is defined as

$$t_\epsilon[s] = \begin{cases} s - \epsilon, & \text{if } \ s > \epsilon,\\ s + \epsilon, & \text{if } \ s < -\epsilon,\\ 0, & \text{otherwise.}\end{cases} \quad (19)$$

2) Updating $\mathbf{E}_i$*:* To update the error matrix $\mathbf{E}_i$ for class i, we minimize (17) and fix variables other than $\mathbf{E}_i$, which leads to

$$\begin{aligned}\mathbf{E}_i^{k+1} &= \arg\min_{\mathbf{E}_i} L(\mathbf{A}_i^{k+1}, \mathbf{B}_i^k, \mathbf{E}_i, \mathbf{Y}_i^k, \mathbf{Z}_i^k, \mu^k)\\ &= \arg\min_{\mathbf{E}_i} \lambda\|\mathbf{E}_i\|_1 + \langle \mathbf{Y}_i^k, -\mathbf{A}_i^{k+1} - \mathbf{E}_i + \mathbf{D}_i\rangle\\ &\quad + \frac{\mu^k}{2}\| - \mathbf{A}_i^{k+1} - \mathbf{E}_i + \mathbf{D}_i\|_F^2\\ &= \arg\min_{\mathbf{E}_i} \epsilon'\|\mathbf{E}_i\|_1 + \frac{1}{2}\|\mathbf{X}_e - \mathbf{E}_i\|_F^2,\end{aligned}$$

where $\epsilon' = (\lambda/\mu^k)$ and $\mathbf{X}_e = \mathbf{D}_i - \mathbf{A}_i^{k+1} + (1/\mu^k)\mathbf{Y}_i^k$. As shown in [16, Sec. 2.1], the closed-form solution of the above optimization problem is given by $\mathbf{E}_i^{k+1} = T_{\epsilon'}[\mathbf{X}_e]$.

3) Updating $\mathbf{B}_i$*:* To update the auxiliary variable $\mathbf{B}_i$, consider minimizing (17) with variables other than $\mathbf{B}_i$ fixed:

$$\begin{aligned}\mathbf{B}_i^{k+1} &= \arg\min_{\mathbf{B}_i} L(\mathbf{A}_i^{k+1}, \mathbf{B}_i, \mathbf{E}_i^{k+1}, \mathbf{Y}_i^k, \mathbf{Z}_i^k, \mu^k)\\ &= \arg\min_{\mathbf{B}_i} \eta \sum_{j\neq i} \|(\mathbf{A}_j^{k+1})^T\mathbf{B}_i\|_F^2\\ &\quad + \langle \mathbf{Z}_i^k, \mathbf{B}_i - \mathbf{A}_i^{k+1}\rangle + \frac{\mu^k}{2}\|\mathbf{B}_i - \mathbf{A}_i^{k+1}\|_F^2.\end{aligned}$$

Setting the partial derivative of L with respect to $\mathbf{B}_i$ equal to zero gives

$$2\eta\sum_{j\neq i}\mathbf{A}_j^{k+1}(\mathbf{A}_j^{k+1})^T\mathbf{B}_i + \mathbf{Z}_i^k + \mu^k(\mathbf{B}_i - \mathbf{A}_i^{k+1}) = 0,$$

and we obtain

$$\mathbf{B}_i^{k+1} = (2\eta\sum_{j\neq i}\mathbf{A}_j^{k+1}(\mathbf{A}_j^{k+1})^T + \mu^k\mathbf{I})^{-1}(\mu^k\mathbf{A}_i^{k+1} - \mathbf{Z}_i^k).$$

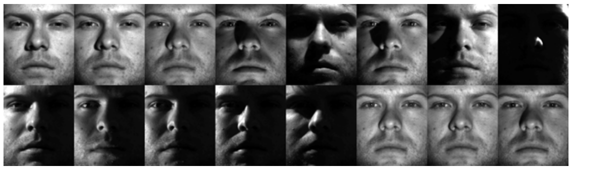


Fig. 4. Example training images randomly selected from the Extended Yale B database.

Once $\mathbf{A}_i$, $\mathbf{E}_i$, and $\mathbf{B}_i$ are obtained, the Lagrange multipliers $\mathbf{Y}_i$ and $\mathbf{Z}_i$ can be simply updated by the corresponding equations in Algorithm 2. The convergence of the variables indicates the termination of the optimization process for our proposed LR algorithm.

E. Convergence Analysis

It can be seen that, the minimization of (5) is non-convex and non-smooth due to the presence of the product term $\mathbf{A}_j^T\mathbf{A}_i$ and the ℓ_1-norm of $\mathbf{E}_i$, respectively. To derive the solution for (5), we iteratively solve (6) across different classes i. During each iteration of minimizing (6), the variables to be minimized are $\mathbf{A}_i$ and $\mathbf{E}_i$, while the remaining variables $\{\mathbf{A}_j : j \neq i\}$ are fixed. This strategy is known as the *block coordinate descent* method ([29, p. 267]). As a result, the objective function in (5) would satisfy the block multiconvex property defined in [30], i.e., the objective function is a convex function of $(\mathbf{A}_i, \mathbf{E}_i)$ with all the other block variables $\{(\mathbf{A}_j, \mathbf{E}_j) : j \neq i\}$ remained fixed. We note that, the convergence and global optimization for the block multiconvex function like (5) has been established and verified in [30], which advances a sophisticated update rule for the block variables under the Kurdyka-Łojasiewicz condition. For the ease of presentation, we choose to update the block variables $(\mathbf{A}_i, \mathbf{E}_i)$ via (6) without introducing additional regularization terms.

We now discuss the convergence rate when solving (6). Note that (6) is now a convex optimization problem with only $\mathbf{A}_i$ and $\mathbf{E}_i$ as variables, and thus a global minimizer can be expected. There exist several approaches which can be utilized to solve (6), including iterative thresholding, accelerated proximal gradient (APG), and augmented Lagrange multiplier (ALM) methods. We adopt the ALM method because of its excellent convergence property (as suggested in [16]). Following the same arguments as in the proof of Theorem 1 in [16], one can prove that the convergence rate of the ALM method is at least $O(\mu_k^{-1})$, where μ_k is the penalty parameter in Algorithm 2. This implies that if μ_k grows geometrically, the ALM algorithm will converge Q-linearly.

IV. Experiments

A. Extended Yale B Database

We first conduct experiments on the Extended Yale B database [31], which consists of 2,414 frontal-face images of 38 subjects (around 59–64 images for each person). The face images are taken under various laboratory-controlled lighting conditions (see Fig. 4 for example) [32]. All images are down-sampled to 64×56 = 3,584 pixels and are converted to gray

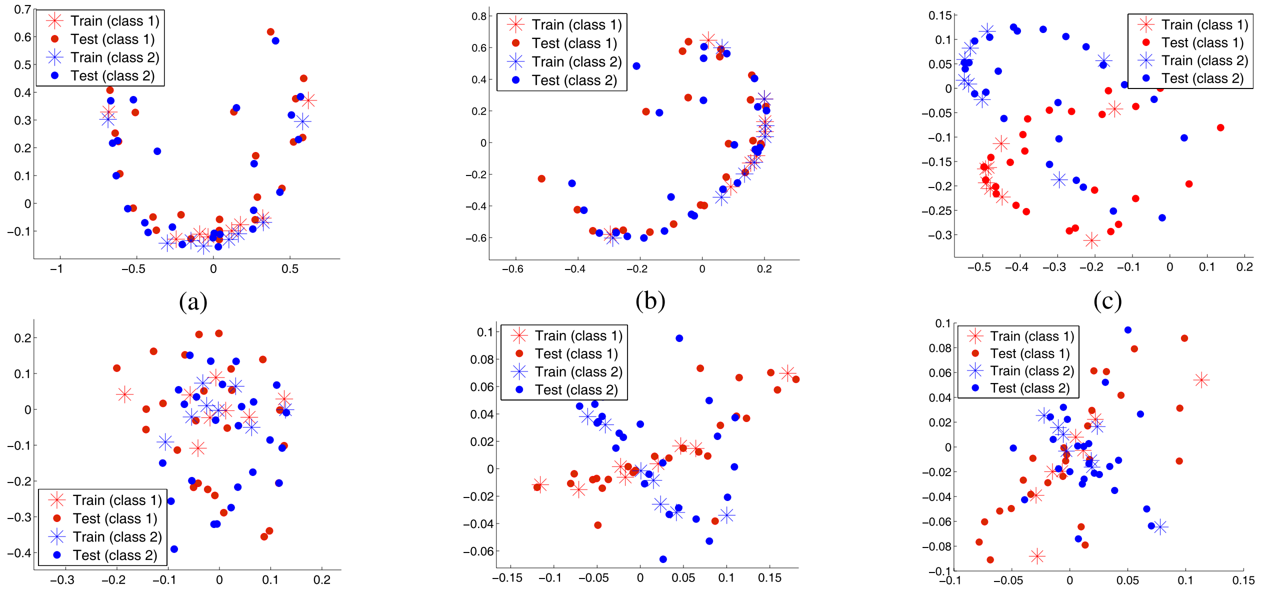


Fig. 5. Data distributions for 2 classes (in blue and red colors ones). The 2D subspace is spanned by the first two eigenvectors of the covariance matrices of (a) the original data matrix **D**, (b) the LR matrix **A** without structural incoherence, and (c) the LR matrix **A** with structural incoherence. The corresponding plots for (a), (b), and (c) when the 2D subspace is spanned by the fifth and sixth eigenvectors are shown in (d), (e), and (f), respectively. Note that training and test instances are denoted as ($*$) and ($\bullet$), respectively.

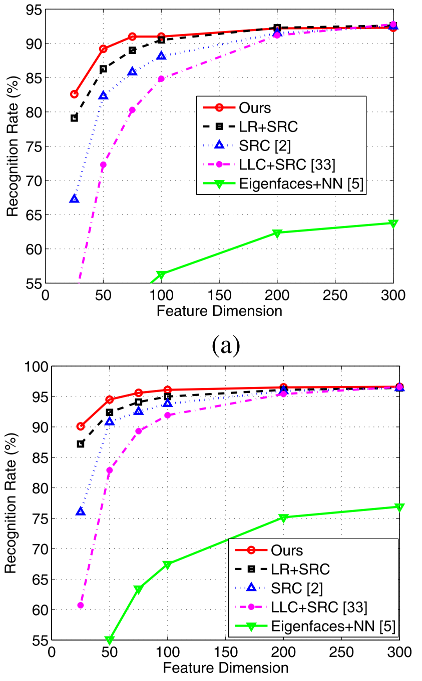
scale images prior to our experiments. Besides the standard LR (without structural incoherence) and our proposed method, we also consider Eigenfaces [5], SRC [2], and LLC [33] for comparisons. Note that LLC is a coding scheme extended from SRC, and it exploits data locality for improved sparse coding. For LLC we use the same classification rule (4) as in the SRC algorithm. In our experiments, we apply the Homotopy method [34] to solve the ℓ_1 minimization problem (3) with $\lambda = 0.001$, which is observed to be accurate and efficient among various ℓ_1 minimization techniques as reported in [34]. For all experiments in this paper, we considered $\eta \in [10^{-2}, 10^2]$ and selected the one with the best recognition performance.

To evaluate our recognition performance using data with different dimensions, we project the data onto the eigenspace derived PCA using our LR models (as shown in Fig. 3). For the standard LR approach, the eigenspace spanned by LR matrices without structural incoherence is considered, while those of other SRC-based methods are derived by the data matrix **D** directly. We vary the dimension of the eigenspace and compare the results in this section.

1) Visualization of The Discrimination Ability: To visualize the effectiveness of our proposed method in recognizing images from different classes, we show the distributions of training and test data from two classes in Fig. 5(a), in which the data are projected onto the first two eigenvectors of the covariance matrix of data matrix **D** (as Eigenfaces and SRC-based approaches do). Moreover, we project the same data onto the subspace derived by the low-rank matrices **A** with and without structural incoherence, and the results are shown in Figs. 5(b) and 5(c), respectively. Compared to Figs. 5(c) and 5(a) (or 5(b)), it is clear that the separation between the two classes (in red and blue colors) is significantly improved, and thus a better recognition rate can be expected using our approach.

Compared to Figs. 5(a), 5(b), and 5(c), we also plot their corresponding 2D subspaces spanned by the fifth and sixth eigenvectors in Figs. 5(d), 5(e), and 5(f), respectively. It can be seen that the same data projected onto the original data matrix **D** (i.e., Fig. 5(d)) still does not exhibit sufficient discrimination property, while the separation between the data projected onto the LR matrices **A** with and without structural incoherence (Figs. 5(e) and 5(f)) are observed to be improved (especially for Fig. 5(e) vs. Fig. 5(b)). However, it is worth noting that our LR matrix **A** with structural incoherence is able to provide better data discrimination at more dominant eigenvectors (see Fig. 5(c)), and thus the use of our derived LR matrix will be expected to achieve better recognition results. The following experiments will confirm this observation.

2) Performance Comparison: To evaluate the recognition performance, we first randomly select 16 images from each class for training and the remaining for test. Therefore, different subjects have training images subject to different lighting conditions, which are close to the cases in practical applications. We vary the dimension of the eigenspace as 25, 50, 75, 100, 200, and 300 to compare the recognition performance between different methods, which are shown in Fig. 6(a). It is clear that while the two LR methods consistently produced higher recognition rates than other Eigenfaces and SRC-based approaches did, our proposed LR method was the best among all. For example, at feature dimension 50, our method achieved a high recognition rate at 89.2%, and those for LR, SRC, LLC, and Eigenfaces were 86.3%, 82.3%, 72.3%, and 45.5%, respectively (see Fig. 6(a)). We repeat the above experiments



(a)

(b)

Fig. 6. Performance comparisons on the Extended Yale B database with different numbers $\bar{m}$ of training images per person. (a) $\bar{m} = 16$. (b) $\bar{m} = 32$.

using 32 training images per person (as shown in Fig. 6(b)), and we observe the same advantages using our proposed method. From these empirical results, we confirm that the use of our LR method alleviates the problem of severe *illumination variations* even when such noise is presented in both training and test data. More importantly, due to the enforcement of structural incoherence between the derived LR matrices, our method exhibits additional classification capability and thus outperforms the standard LR approach.

B. CMU Multi-PIE Database

The CMU Multi-PIE database [35] contains face images of 337 subjects recorded in four different sessions. In each session, every subject has images of two or three facial expressions with 20 different illuminations. In our experiments, we consider the training set of all 249 subjects in Session 1. For each of the 249 subjects, we select face images with the frontal pose, and have illuminations {1, 2, 8, 14} of the neutral expression, and illuminations {15, 17, 19} of the smile expression as training images. Thus, the training set has a total of $7 \times 249 = 1,743$ images. The test set includes the subjects from Sessions 2, 3, and 4 that present in Session 1. For every subject in each session during testing, we use all 20 illuminations of two facial expressions as test images, and thus each session contains about 6,400–7,000 test images. All face images are manually cropped and downsampled into $40 \times 32 = 1,280$ pixels. Example test images from the CMU Multi-PIE database are shown in Fig. 7.

We compare our method with Eigenfaces [5], Fisherfaces [6], standard low-rank matrix recovery (LR), SRC [2], and LRSI-approx [24]. Table I lists and compares the recognition results. From Table I, it can be seen that while SRC-based approaches obtained improved results than baseline methods did (e.g., Eigenfaces and Fisherfaces), our proposed method achieved the highest recognition rates and outperformed all other approaches in all sessions. From our experimental results on the CMU Mult-PIE database, the effectiveness of our proposed algorithm can be verified. In the following subsections, we will consider more challenging datasets with occluded face images for training and testing.

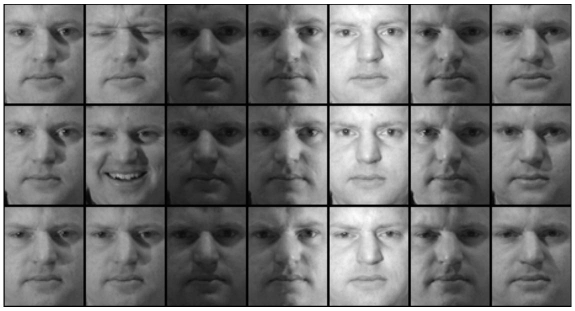


Fig. 7. Example test images from the CMU Multi-PIE database, where the first, second, and third rows are selected from Sessions 2, 3, and 4, respectively.

C. AR Database

The AR database [36] contains over 4,000 frontal images for 126 individuals. For each subject, twenty-six face images are taken under different variations in two separate sessions. There are thirteen images for each session, in which three images with sunglasses, another three with scarfs, and the remaining seven are with illumination and expression variations and thus are considered as clean/neutral images (see Fig. 8 for example). All images are downsampled to $55 \times 40 = 2{,}200$ pixels and converted to gray scale. In our experiments, we choose a subset of the AR database consisting of 50 men and 50 women (as [2] did). It is worth noting that, most prior works using this database only considered the use of neutral images for training. To show the effectiveness of our results, we conduct experiments where the training images are corrupted due to occlusion or random pixel noise.

1) Training Images With Disguise: In this part of the experiments, we consider the scenario in which the training set has *both* neutral and occluded images taken at Session 1 (of a portion of it). There are three cases to be evaluated:

Sunglasses: We first consider occluded training images due to the presence of sunglasses, which occlude about 20% of the face image. We have a total n_c neutral images (randomly chosen) plus n_o image(s) with sunglasses at Session 1 for training (we fix $n_c + n_o = 7$), and 7 neutral images plus 3 images with sunglasses at Session 2 for testing. To assess the influence of the ratio $n_o/(n_c + n_o) = n_o/7$ for robust face recognition, we vary the number of n_o from 0 up to 3.

Scarf: We consider occluded training images occluded by disguise due to the presence of scarfs, which occlude about 40% of the face image. The choice of training and test set data is similar to that for the above (**Sunglasses**) case.

TABLE I

PERFORMANCE COMPARISONS ON THE CMU MULTI-PIE DATABASE. THE FEATURE DIMENSION IS SET AS 300 FOR ALL METHODS EXCEPT FOR FISHERFACES (WHOSE FEATURE DIMENSION IS 248)

Methods	Session 2		Session 3		Session 4		Average
	Neutral	Squint	Neutral	Smile	Neutral 1	Neutral 2	
Eigenfaces+NN [5]	70.24	61.14	69.44	58.72	69.31	69.54	66.40
Fisherfaces+NN [6]	78.13	51.93	81.94	45.31	80.51	79.17	69.50
SRC [2]	92.47	81.63	91.28	76.25	91.74	90.31	87.28
LR+SRC	93.40	84.13	93.03	77.53	93.37	92.29	88.96
LRSI-approx [24]	94.13	**85.18**	**94.28**	79.53	95.37	94.37	90.48
Ours	**94.22**	**85.18**	**94.28**	**79.81**	**95.40**	**94.43**	**90.55**

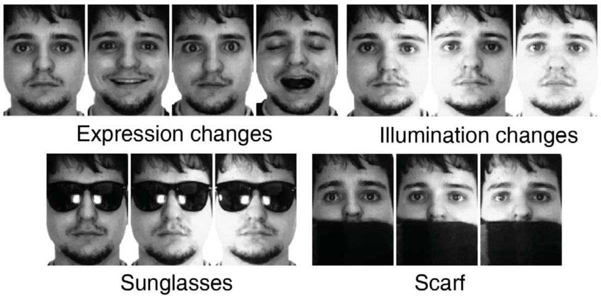


Fig. 8. Example images from Session 1 of the AR database.

Sunglasses+Scarf: In this most challenging case, the training images are occluded due to sunglasses or scarfs. From Session 1, we choose 7 neutral images, n_{sg} images with sunglasses, and n_{sc} images with scarfs for training. The numbers of n_{sg} and n_{sc} are set to be the same, and they range from 0 to 3. The test set consists of 7 neutral images, 3 images with sunglasses, and 3 images with scarfs (all from Session 2). Note that the setting of this scenario is different from those in **Sunglasses** and **Scarf**. The number of training images in the previous two cases is fixed at 7, while the number of training images in this scenario varies with n_{sg} and n_{sc}.

We compare our method with the approaches of Eigenfaces [5], Fisherfaces [6], standard low-rank matrix recovery (LR), SRC [2], and LRSI-approx [24]. Tables II and III show the recognition results of the above three scenarios using different approaches.[1] From these two tables, we see that our method generally outperforms all other approaches across different settings. In Table II, we observe that the recognition rates of SRC are sensitive to the type of occlusions. For example, the difference in recognition rates is 9.6% when the percentage of occluded training images is 14%. Compared to SRC, our method has much smaller performance gap between the two scenarios, and thus our method is much less sensitive to the type of occluded images in the training set.

In Table II, when only neutral images were considered as training images $\mathbf{D}$, the recognition rates were inferior to those using a number of occluded training images. The reason for this is due to the way SRC performs recognition. Recall that in Section II, SRC solves the ℓ_1-minimization problem (3) and determines the identity of the test image $\mathbf{y}$ based on the class-wise reconstruction error (4). In other words, SRC assumes that the test input $\mathbf{y}$ can be well approximated by $\mathbf{D}\boldsymbol{\alpha}$. Given an occluded image $\mathbf{y}$, the reconstruction error $\|\mathbf{y}-\mathbf{D}\boldsymbol{\alpha}\|_2^2$ in (3) will not be negligible if $\mathbf{D}$ contains only neutral training images (i.e., the unoccluded ones), and large reconstruction errors often lead to inferior recognition performance.

It can be seen from Table III that, although LR outperforms SRC for all tests, the difference between their recognition rates becomes smaller when the number of occluded training images increases. This is because that the low-rank matrix $\mathbf{A}$ extracted by standard LR does not contain sufficient discriminating information (as discussed in Section III-B). Unlike LR, our method does not suffer from this due to the enforcement of structural incoherence.

Besides the above experiments, we also vary the feature dimension (via PCA) from 25 up to 500 under (a) **Sunglasses** with $n_o/7 = 14\%$, (b) **Scarf** with $n_o/7 = 14\%$, and (c) **Sunglasses+Scarf** with $(n_{sg}+n_{sc})/(7+n_{sg}+n_{sc}) = 22\%$. We compare the recognition performance of different methods in Fig. 9. From this figure, we see that our approach outperformed all other methods for the three cases. At dimension 100, our approach achieved the best recognition rate 85.1% for **Sunglasses**, 82.4% for **Scarf**, and 80.7% for **Sunglasses+Scarf**. We observe that, since our formulation (5) aims at minimizing the nuclear norm of $\mathbf{A}_i$ and thus reducing the rank of $\mathbf{A}_i$, the first few eigenvalues of the covariance matrix of $[\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N]$ will be the most dominant ones. Since the introduced structural incoherence term $\sum_{j\neq i} \|\mathbf{A}_j^T \mathbf{A}_i\|_F^2$ in (5) encourages the incoherence between different $\mathbf{A}_i$ and $\mathbf{A}_j$, this further suppresses the dominant eigenvalues and makes them even sparser. In the above case, we observe that the rank of the derived matrices $\mathbf{A}_i$ is about 100. This explains why the proposed method favors lower dimensionality while achieving the best recognition performance. From the above experimental results and discussions, we confirm that our method outperformed other state-of-the-art algorithms over a variety of scenarios.

2) Training Images With Random Pixel Corruption: In the second part of the experiments, we consider the training images which are corrupted due to the presence of random noise. We first choose 7 neutral images (without occlusion) from Session 1 for training and 7 neutral images from Session 2 for testing. Next, we randomly choose pixels (and vary the percentages) of training and test images, and those

[1] The feature dimension is set as 300 for all methods except for Fisherfaces. Since the maximal number of valid Fisherfaces is $N - 1$, where N is the number of subjects, the feature dimension of Fisherfaces is fixed at $N - 1$.

TABLE II

COMPARISONS OF RECOGNITION RATES WITH DIFFERENT PERCENTAGES OF OCCLUDED IMAGES ($n_o/7$) PRESENTED IN THE TRAINING SET. THE FEATURE DIMENSION IS SET AS 300 FOR ALL METHODS EXCEPT FOR FISHERFACES

Methods	0% = 0/7		14% = 1/7		29% = 2/7		43% = 3/7	
	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf
Eigenfaces+NN [5]	60.2	52.1	61.7	56.7	60.5	53.9	60.9	50.3
Fisherfaces+NN [6]	67.3	**73.3**	79.0	80.1	77.8	79.3	79.9	78.3
SRC [2]	70.2	64.7	81.0	71.4	79.4	72.8	79.6	68.7
LR+SRC	72.7	66.5	82.6	73.4	81.0	75.5	81.6	73.9
LRSI-approx [24]	70.9	67.1	83.4	80.1	83.2	78.2	83.4	77.7
Ours	**73.0**	72.8	**84.2**	**82.6**	**83.7**	**80.5**	**83.7**	**79.6**

TABLE III

COMPARISONS OF RECOGNITION RATES WITH DIFFERENT PERCENTAGE OF OCCLUDED IMAGES PRESENTED IN THE TRAINING SET. THE FEATURE DIMENSION IS SET AS 300 FOR ALL METHODS EXCEPT FOR FISHERFACES

	Sunglasses+Scarf			
	0/(7+0)	2/(7+2)	4/(7+4)	6/(7+6)
Methods	= 0%	= 22%	= 36%	= 46%
Eigenfaces+NN [5]	48.1	54.8	57.4	60.1
Fisherfaces+NN [6]	62.2	75.4	74.9	76.0
SRC [2]	57.7	73.5	76.5	76.4
LR+SRC	60.3	75.1	77.5	76.9
LRSI-approx [24]	59.8	77.9	80.8	82.2
Ours	**62.8**	**80.8**	**81.8**	**82.8**

pixels are replaced by 0 or 255. The percentage of corrupted pixels ranges from 0 to 40%, as shown in Fig. 10.

Table IV lists the recognition rates with feature dimension set as 300 for all methods except for Fisherfaces. From this table we see that our method again outperformed state-of-the-art algorithms on most cases. Among different methods, we observe that Eigenfaces, Fisherfaces, and SRC degraded significantly as the percentage of corrupted pixels increased. As can be expected, the performance drop for methods utilizing low-rank decomposition (i.e., LR, LRSI-approx, and ours) is less than those using standard subspace learning techniques (i.e., Eigenfaces, Fisherfaces, and SRC), since LR-based approaches exhibit better ability in removing sparse noise.

It is worth noting that, although Fisherfaces [6] also promotes the separation between classes during its learning process, it does not achieve comparable performance as we do. With the percentage of corruption increases, it can be seen that the recognition rate of Fisherfaces was severely degraded. For example, the recognition rate of Fisherfaces decreased from 84.4% to 45.7% when the percentages of random pixel corruption increased from 0% to 10%. This is because of its direct use of corrupted training image data for data separation. As a result, the performance of Fisherfaces will be remarkably degraded due to overfitting the noise presented in training data.

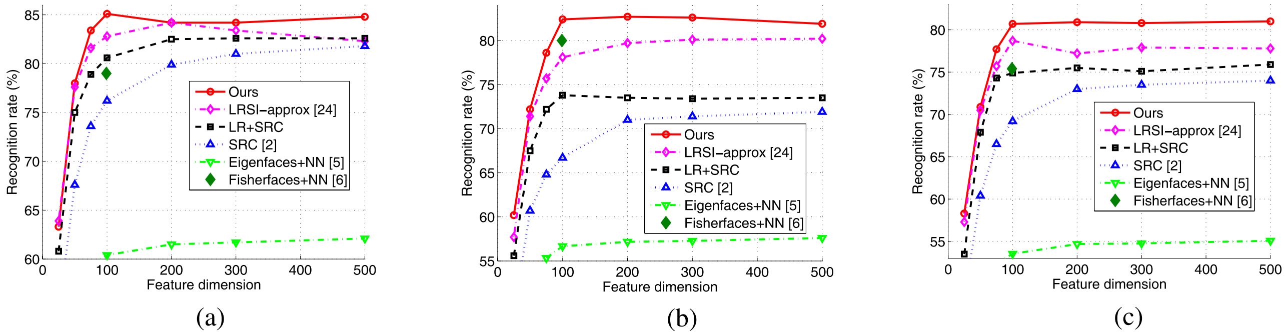
3) Selection of Parameters η and ρ: We now discuss how we determine the parameter η. Similar to SVM or other regularized optimization problems, the introduced regularizer typically solves a particular task, while its weight/penalty balances between the regularizer itself and the original objective function. As can be seen in (5), the parameter η regularizes the incoherence between the low-rank representations of different classes $\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_N$, and such incoherence brings additional discriminating capabilities into the derived solutions as discussed in Section III-B.2.

If the value of η is too small, solving the problem of (5) will focus on minimizing the first term (i.e., the standard low-rank matrix decomposition), and thus there is no guarantee for sufficient incoherence/separation between different classes. However, if the value of η is too large, one would overemphasize the discrimination between difference classes, even such information comes from undesirable patterns or corrupted image regions (e.g., sunglasses or scarves). Since our paper addresses robust face recognition problems, our goal is to select a proper η which ensures sufficient incoherence being introduced for the derived low-rank representations of different subjects. As a result, improved recognition can be achieved.

In our experiments, we considered a range of possible values for η, and we selected the one with the best recognition performance. Take the AR database for example, we plot the recognition rates with $\eta \in [10^{-2}, 10^{2}]$ in Fig. 11, in which the settings were the same as those of the fourth and fifth columns in Table II. For comparison purposes, we also plot the recognition rates using the standard low-rank matrix recovery (LR), which only solved the first term of (5) and directly applied the resulting representations for recognition.

From Fig. 11, it can be seen that our method with proper η choices would consistently outperformed LR. As expected, its performance decreased and was comparable to that of LR when η became small. On the other hand, the recognition performance also degraded if we overemphasized the structured incoherence with a much larger η, which resulted in the lack of the capability of disregarding undesirable noisy patterns for face images. We note that, since there were only few corrupted images available for the databases considered, they were either treated as training or test data for verifying the effectiveness of our proposed algorithm. In other words, we did not select η by performing cross-validation on such a small amount of corrupted data.

As for the parameter ρ in Algorithm 2, it controls the increasing/convergence rate of the augmented Lagrange multiplier μ. In general, the inner while loop in Algorithm 2 converges faster if μ is updated with a larger increasing rate. However, it is also more likely to encounter the



(a) (b) (c)

Fig. 9. The recognition rate across different feature dimensions for various algorithms under (a) **Sunglasses** with $n_o/7 = 14\%$, (b) **Scarf** with $n_o/7 = 14\%$, and (c) **Sunglasses+Scarf** with $(n_{sg} + n_{sc})/(7 + n_{sg} + n_{sc}) = 22\%$.

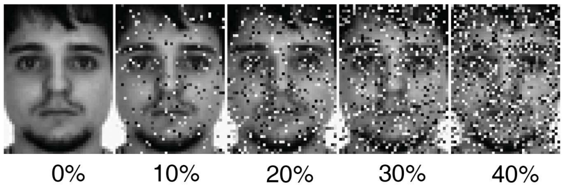


Fig. 10. Example training images under different percentages of random pixel corruption.

TABLE IV

COMPARISONS OF RECOGNITION RATES WITH DIFFERENT PERCENTAGES OF RANDOM PIXEL CORRUPTION

Methods	0%	10%	20%	30%	40%
Eigenfaces+NN [5]	71.1	50.7	34.0	20.7	10.6
Fisherfaces+NN [6]	84.4	45.7	23.3	13.7	4.5
SRC [2]	86.0	73.9	57.2	41.2	27.3
LR+SRC	86.4	81.6	70.1	58.0	42.6
LRSI-approx [24]	86.9	80.9	71.5	**59.7**	45.3
Ours	**89.9**	**82.7**	**72.4**	**59.7**	**46.5**

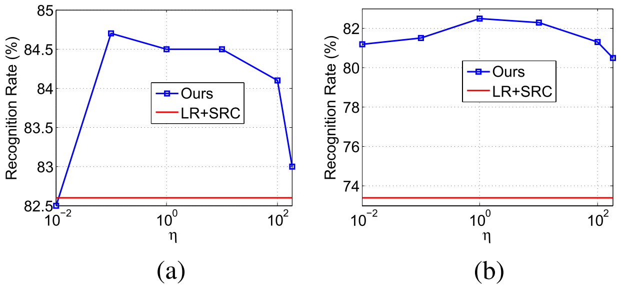


Fig. 11. Recognition rates with different η for the AR database. (a) Sunglasses. (b) Scarf.

ill-condition problem, which prevents one from reaching the optimal solution for the associated Lagrange function. Therefore, there is a tradeoff between fast convergence and the optimum of the solution when solving the optimization problem. We set $\rho = 1.5$ for all experiments in this paper, and this choice always allowed our algorithm to converge within 30 iterations. More discussions on the increasing rate ρ can be found in [29, Sec. 4.2].

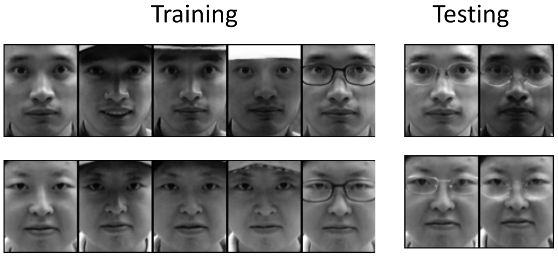


Fig. 12. Example images of the CAS-PEAL database.

TABLE V

PERFORMANCE COMPARISONS ON THE CAS-PEAL DATABASE

Methods	Accuracy
Eigenfaces+NN [5]	73.04
Fisherfaces+NN [6]	88.59
SRC [2]	86.98
LR+SRC	90.78
LRSI-approx [24]	91.36
Ours	**92.51**

D. CAS-PEAL Database

The CAS-PEAL database [37], to the best of our knowledge, is the currently largest public face database with corrupted face images available. To conduct the experiments, we select all 434 subjects from the normal and the accessory categories of CAS-PEAL for training and testing (recall that AR only has face images of 100 subjects). Each subject in CAS-PEAL has 1 neutral image, 3 images with hats, and 3 images with glasses/sunglasses. We select one image with glasses and one image with hats as test images, and the rest for training (including those with sunglasses). Since we only consider recognition of frontal faces in this work, we manually crop out and downsample face images into $40 \times 32 = 1,280$ pixels. Example training and test images are shown in Fig. 12.

Similar to our prior experiments, we compare our method with Eigenfaces [5], Fisherfaces [6], standard low-rank matrix recovery (LR), SRC [2], and LRSI-approx [24]. Table V lists the recognition results. From Table V, we see that our method achieved the highest recognition rate among all methods. Similar to our experiments on the prior two datasets, LR-based approaches (LR, LRSI-approx, and ours) outperformed baseline methods due to the ability of disregarding noisy patterns. It is worth repeating that, our method

TABLE VI
COMPUTATIONAL TIME OF THE TRAINING STAGE OF LOW-RANK BASED ALGORITHMS

Dataset	LR	LRSI-approx [24]	Ours
Extended Yale B	8.77 sec	48.53 sec	1 hr 55 min
Multi-PIE	27.35 sec	161.09 sec	4 hr 35 min
AR	15.30 sec	53.47 sec	2 hr 55 min
CAS-PEAL	45.16 sec	230.18 sec	12 hr 50 min

outperformed standard LR and LRSI-approx because of the advance of structural incoherence, which confirms the use of the proposed algorithm for solving (5) (and thus (7)) in this paper.

E. Runtime Complexity

We now analyze the runtime complexity of our proposed method (i.e., Algorithm 2). The dominant cost of our Algorithm 2 is the inner while loop which updates variables $\mathbf{A}_i$ and $\mathbf{B}_i$ at each iteration. Recall that matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ both are of size $d \times m_i$ with $d \gg m_i$. For updating $\mathbf{A}_i$, the SVD operation in Section III-D.1 has the complexity of $O(dm_i^2)$. To update $\mathbf{B}_i$, we solve the linear equation in Section III-D.3, which requires $\frac{1}{3}d^3$ flops (floating-point operations) for the Cholesky factorization, and $2d^2m_i$ flops for forward and backward substitutions. As a result, the complexity for updating $\mathbf{B}_i$ is $O(d^3)$. Since $d \gg m_i$, updating $\mathbf{B}_i$ dominates the computation complexity, and thus the complexity of the inner while loop of Algorithm 2 is $O(d^3)$. Given the above observations, we conclude that the runtime complexity of Algorithm 2 is $O(d^3Npq)$, where N is the number of classes, and p and q are the numbers of iterations for inner and outer while loops of Algorithm 2, respectively.

In view of the fact that the dominant cost of performing LR is the SVD operation for each of the N classes, the runtime complexity of LR is $O(dm^2Np)$, where $m = \max_i m_i$. Table VI compares the computational time of the *training* stage of LR, LRSI-approx [24], and our method (i.e., Algorithm 2). Note that the runtime estimates are performed on a PC with Intel Core 2 Quad CPU 2.33 GHz and 4G RAM under the MATLAB environment. We note that, LRSI-approx [24] is the prior version of our current approach, which solves a relaxed version of the optimization problem of (5). Although LR requires the least amount of training time, both our Algorithm 2 and LRSI-approx achieved improved recognition performance than LR as shown in our experiments. It is also noting that, the training stages of all low-rank based algorithms can be done offline. As for the testing time, since all low-rank based algorithms utilize the same classification technique of SRC, the computation time of all the above approaches are comparable (e.g., one generally required only 0.5 seconds for classifying an input face image using SRC in our experiments).

F. Limitations and Applications

The same as SRC and most of dictionary learning or reconstruction-based approaches for face recognition, we need registered face images for training and testing. In other words, such approaches cannot directly applied to recognized face images with pose variations (to be more specific, they are not able to recognize face images with out-of-plane rotations). As a result, this type of recognition methods are particularly favorable for applications of access control, automatic teller machine, or other security facilities. In such scenarios, one typically is able to collect controlled (registered) training images in advance, and the test image will be captured under the same (or very similar) environments. Nevertheless, if registered face images are not available for either training or testing (but only shift and in-plane rotation variations are presented), one can apply existing image registration techniques like RASL [17] or IntraFace [38], which would alleviate the above limitations for SRC, etc. approaches.

V. CONCLUSION

We presented a low-rank matrix approximation algorithm with structural incoherence for robust face recognition. The introduction of structural incoherence between low-rank matrices promotes the discrimination between different classes, and thus the associated models exhibit excellent discriminating ability. We provided detailed derivations and showed that the proposed optimization problem can be solved by advancing augmented Lagrange multipliers. Our experiments on four face databases confirmed that our proposed methods is robust to severe illumination variations, occlusion, and random pixel noise corruptions, while our method has been shown to outperform state-of-the-art face recognition algorithms.

REFERENCES

[1] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[3] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[4] A. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications," *IEEE MultiMedia*, vol. 19, no. 1, pp. 20–28, Jan. 2012.

[5] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Comput. Soc. Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 1991, pp. 586–591.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[8] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 383–394, Mar. 2008.

[9] F. De la Torre and M. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, no. 1, pp. 117–142, 2003.

[10] Q. Ke and T. Kanade, "Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming," Ph.D. dissertation, Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2005.

[11] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, article no. 11, 2011.

[12] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using Markov random fields," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 1050–1057.

[13] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," Ph.D. dissertation, Dept. Comput., Hong Kong Polytechnic Univ., Hong Kong, China, 2010.

[14] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 2011, pp. 625–632.

[15] F. De la Torre and M. Black, "Robust principal component analysis for computer vision," in *Proc. IEEE 8th Int. Conf. Comput. Vis. ICCV*, Jul. 2001, pp. 362–369.

[16] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Dept. Elect. Comput. Eng., Univ. Illinois Urbana-Champaign, Champaign, IL, USA, Tech. Rep. UILU-ENG-09-2215, 2009.

[17] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.

[18] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 1–8.

[19] Y. Mu, J. Dong, X. Yuan, and S. Yan, "Accelerated low-rank visual recovery by random projection," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 2011, pp. 2609–2616.

[20] X. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 2010, pp. 3493–3500.

[21] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for hierarchical sparse coding," *J. Mach. Learn. Res.*, vol. 12, pp. 2297–2334, Oct. 2011.

[22] Y.-W. Chao, Y.-R. Yeh, Y.-W. Chen, Y.-J. Lee, and Y.-C. F. Wang, "Locality-constrained group sparse representation for robust face recognition," in *Proc. 8th IEEE Int. Conf. Image Process., ICIP*, Sep. 2011, pp. 761–764.

[23] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3501–3508.

[24] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 2012, pp. 2618–2625.

[25] S. Kong and D. Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," Ph.D. dissertation, Dept. Comput. Sci. Technol., Zhejiang Univ., Hangzhou, China, 2012.

[26] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognit.*, vol. 45, pp. 3902–3911, Oct. 2012.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[28] J. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[29] D. Bertsekas, *Nonlinear Programming*. Nashua, NH, USA: Athena Scientific, 1999.

[30] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[31] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jul. 2001.

[32] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 2010, pp. 3360–3367.

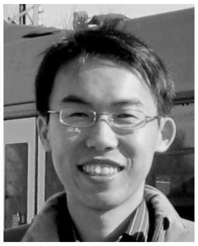
[34] A. Yang, S. Sastry, A. Ganesh, and Y. Ma, "Fast ℓ_1-minimization algorithms and an application in robust face recognition: A review," Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2010-13, 2010.

[35] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[36] A. Martinez and R. Benavente, "The AR face database," CVC, New York, NY, USA, Tech. Rep., 1998.

[37] W. Gao *et al.*, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[38] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 2013, pp. 532–539.



Chia-Po Wei received the B.S. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2004 and 2011, respectively. He is currently a Post-Doctoral Researcher with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include face recognition, dictionary learning, and computer vision.

Chih-Fan Chen received the B.S. and M.S. degrees in mechanical engineering from National Taiwan University, Taipei, Taiwan, in 2007 and 2009, respectively. From 2010 to 2012, he was a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Southern California, Los Angeles, SC, USA. His research interests include computer vision, image processing, and machine learning.

Yu-Chiang Frank Wang (M'04) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2004 and 2009, respectively.

Dr. Wang joined the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, in 2009. He is currently a Tenure-Track Associate Research Fellow and leads the Multimedia and Machine Learning Laboratory at CITI. His research interests span the fields of computer vision, pattern recognition, and machine learning. In 2011, he and his team received the First Place Award at Taiwan Tech Trek by the National Science Council (NSC) of Taiwan. In 2013, he was selected among the Outstanding Young Researchers by NSC.