Nuclear Norm Based Matrix Regression with Applications to Face Recognition with Occlusion and Illumination Changes

Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu

Abstract—Recently, regression analysis has become a popular tool for face recognition. Most existing regression methods use the one-dimensional, pixel-based error model, which characterizes the representation error individually, pixel by pixel, and thus neglects the two-dimensional structure of the error image. We observe that occlusion and illumination changes generally lead, approximately, to a low-rank error image. In order to make use of this low-rank structural information, this paper presents a two-dimensional image-matrix-based error model, namely, nuclear norm based matrix regression (NMR), for face representation and classification. NMR uses the minimal nuclear norm of representation error image as a criterion, and the alternating direction method of multipliers (ADMM) to calculate the regression coefficients. We further develop a fast ADMM algorithm to solve the approximate NMR model and show it has a quadratic rate of convergence. We experiment using five popular face image databases: the Extended Yale B, AR, EURECOM, Multi-PIE and FRGC. Experimental results demonstrate the performance advantage of NMR over the state-of-the-art regression-based methods for face recognition in the presence of occlusion and illumination variations.

Index Terms—Nuclear norm, robust regression, sparse representation, alternating direction method of multipliers (ADMM), face recognition

1 Introduction

Face recognition has aroused broad interest in pattern recognition and computer vision areas during the past 20 years. Meanwhile, numerous face-representation and classification methods have been developed. Recently, linear regression (LR) analysis based methods have become a hot topic in the face recognition community. Naseem et al. presented a linear regression classifier (LRC) for face classification [1]. In fact, several previous works, such as the nearest feature line [2], the nearest feature plane, and the nearest feature space methods [3], are all variants of LR based methods.

To avoid over-fitting, a regularization term is generally imposed upon the LR model. There are two widely-used regularizers: the L_2-norm based regularizer and the L_1-norm based one. LR with the L_2-norm regularizer is generally called Ridge regression, while LR with the L_1-norm regularizer is called Lasso, which is a popular model for sparse representation. Wright et al. [4] presented a sparse representation based classification (SRC) method. To obtain more robustness, they further assumed noise is sparse and built the extended SRC model. The model shows the robust ability to deal with sparse random pixel corruption and block occlusion. Wagner et al. [10] further extended the SRC model and unified face alignment and recognition into a single framework.

Some recent work, on the other hand, began to investigate the role of sparsity in face recognition [13], [14], [15], [16]. Yang et al. [15] gave an insight into SRC and provided some theoretical support for its effectiveness. They argued that it is L_1 constraint rather than L_0 (the inherent sparse constraint) that makes SRC effective. Zhang et al. [16] analyzed the working principle of SRC and believed that the collaborative representation strategy plays a more important role than the L_1-norm based sparsity constraint. They presented a collaborative representation classifier (CRC) based on Ridge regression. CRC, however, does not provide a mechanism for noise removal, so it is not a robust method for face recognition.

In the LRC, CRC and SRC models, the representation residual is measured by the L_2-norm or L_1-norm of the error vector. Such models inherently assume that the representation error follows a Gaussian or Laplacian distribution. However, in real-world face recognition cases, the distribution of representation error is more complicated [6], [7]. So, in theory, the above mentioned models are not sufficiently robust for expected noise. Towards this end, Yang et al. borrowed the idea of robust regression [5] and proposed a regularized robust coding (RSC) method [6], [7]. He et al. made use of the correntropy induced robust error metric and presented the correntropy based sparse representation (CESR) algorithm [8], [9]. It is interesting that, although CESR and RSC are developed from different motivations, in light of the fact that correntropy can be viewed as an M-estimator with varying kernel sizes, they are both in spirit of a sparse robust regression model. Recently, He et al. [34] built a half-quadratic framework which unifies the two kinds of existing sparse robust regression models: the additive model represented by

- *J. Yang, L. Luo, J. Qian, Y. Tai, and F. Zhang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: {csjyang, csjqian}@njust.edu.cn, zzdxpyy3001@163.com, tyshiwo@gmail.com, csfzhang@126.com.*
- *Y. Xu is with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. E-mail: laterfall@hitsz.edu.cn.*

Manuscript received 17 May 2015; revised 6 Feb. 2016; accepted 22 Feb. 2016. Date of publication 25 Feb. 2016; date of current version 12 Dec. 2016.
Recommended for acceptance by D. Maltoni.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2535218

SRC and the multiplicative model represented by CESR and RSC. In addition, Naseem et al. further extended their LRC to the robust linear regression classification (RLRC) using the Huber estimator to deal with severe random pixel noise and illumination changes [17]. All of these robust-regression-related methods have been applied to real-world face recognition problems and yielded promising results.

The existing robust regression methods mentioned above all use the one-dimensional pixel-based error model, in which the error on each pixel is characterized one by one, individually. This model has two problems. First, it assumes that pixel-wise errors are independent and identically distributed. This assumption is reasonable for random pixel corruption, where noise is added independently on each pixel. However, in the cases of many practical face variations, such as occlusion, disguise, or shadow caused by illumination change, this assumption does not hold. For instance, in an occlusion caused by a black scarf, pixel values are zeros. So, the ideal representation errors in the occluded part are correlated, because pixels in a local area of a real-world image are generally highly correlated. (A related example is shown in Fig. S-1 in supplemental materials, which can be found on the Computer Society Digital Library at http://doi.10.1109/TPAMI. 2016.2535218). Therefore, using the one-dimensional pixel-based error model (such as SRC [4], RSC [6], [7], Robust LRC [17] etc.) to address image classification with occlusions is theoretically questionable.

Second, characterizing the representation error individually, pixel by pixel, neglects the whole structure of the error image, since all pixel errors form an error image which may contain meaningful structural information (e.g., the rank of error image). In regression analysis based face recognition methods, we use training images to represent a test image. Ideally, the error image is a zero matrix, and thus it is naturally low-rank. In more general cases, there exist illumination variations and occlusions in test images. Illumination and occlusion are two critical factors that affect the performance of face recognition. In practice, illumination changes, particularly partial illumination variations such as shadows, generally lead to a low-rank (or approximately low-rank) error image, in contrast to the full-rank original image. Occlusion, such as sunglasses and a scarf, also yields a low-rank error image. The aforementioned existing regression methods, characterizing the pixel-error individually, fail to utilize the kind of structural information.

To make full use of the low-rank structural information of error image, this paper presents a *two-dimensional image matrix* based error model, matrix regression, in order to make the image representation and classification straightforward. In contrast, previous methods—ridge regression, Lasso or robust regression—are all vector-based regression models. That is, for dealing with 2D image in the form of matrices, we have to convert images into vectors in advance, before using such regression models. In the converting step, some structural information (e.g., the rank of error image) might be lost. Our matrix regression model does not need the matrix-to-vector converting step. It uses the structural information of images by minimizing the rank of representation residual image to determine the regression coefficients. The rank minimization problem is generally converted into the nuclear norm minimization problem for optimization [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [39]. In this spirit, we use the minimal nuclear norm of the representation residual image as a criterion in our matrix regression model. Thus, our method is named nuclear norm based matrix regression (NMR).

The remainder of the paper is organized as follows: Section 2 outlines the related work. Section 3 introduces the nuclear norm based matrix regression model and presents the alternating direction method of multipliers (ADMM) for the model, and Section 4 further gives an accelerated version of ADMM. In Section 5, we suggest the NMR based classifier for robust classification. In Section 6, we conduct experiments and comparisons with the state-of-the-art methods. Finally, Section 7 concludes the paper.

2 Related Work

With respect to the use of the structural information of errors, in recent literature on face recognition, there are two papers of note: Zhou et al. incorporated the Markov Random Field model into the sparse representation framework for spatial continuity of the occlusion [11]. Li et al. explored the intrinsic structure of continuous occlusion and proposed the structured sparse error coding (SSEC) model [12]. The two methods share a two-step iteration strategy: (1) Detecting error via sparse representation, and (2) Estimating error support (i.e., determining the real occluded part) using graph cuts. The difference is that SSEC uses more elaborate techniques such as the iteratively reweighted sparse coding in the error detection step and a morphological graph model in the error support step for achieving better performance. However, SSEC does not numerically converge to the desired solution; it needs an additional quality assessment model to choose the desired solution from the iteration sequence. Moreover, SSEC contains many parameters to which outcomes are sensitive, and which have a significant effect on performance. Our NMR provides a unified framework to integrate error detection and error support into one simple model. It just has one parameter, which is easily tuned and relatively insensitive to variations of databases.

Recently, structured sparsity was also applied to characterize spatially contiguous occlusions [53], [56]. Structured sparsity is a natural extension of the standard sparsity concept in statistical learning and compressive sensing [54]. The L_1-norms focus on the independent sparsity and do not take into account the potential structural relationships among variables. To induce the structured sparsity patterns, a structured sparsity-inducing norm, built on overlapping groups of variables, was presented [55] by Jenatton et al. Based on structured spare regularization, a structured sparse principal component analysis (PCA) was further developed [56]. The structured sparse PCA encodes not only sparsity but also higher-order priori structural constraints about the data. Mairal et al. [57] also considered a structured sparsity-inducing regularization and used it to learn dictionaries embedded in a particular structure. Actually, the topographic sparse coding [58] and topographic map [59] can also be formulated as a dictionary learning problem with a structured sparsity norm.

Although the structured sparse PCA which has been applied to deal with the occluded faces shows robustness to occlusions compared to non-negative matrix factorization

[56], it was used as a feature extraction method rather than a classifier. Motivated by the idea of structured sparsity, Jia et al. proposed a structured sparse representation classifier (SSRC) by introducing a class of structured sparsity-inducing norms into the SRC framework [53]. They use a hierarchical tree-structured sparsity-inducing norm on the error image of a test face, where overlapping groups of pixels are from local patches of varying size and where each group corresponds to a node of the tree. SSRC turns out to be more effective than SRC for dealing with occlusions, since the structured sparsity-inducing norm is more suitable than the L_1-norm for characterizing the spatially correlated local part.

The structured sparsity-inducing norm, however, is more complicated than the nuclear norm because it needs to predefine group structures, i.e., the set of overlapping groups [55]. How to define a set of optimal groups for real-world, complicated, nonzero patterns caused by occlusions and illumination is a difficult problem. So, NMR is much simpler and more practical than SSRC since it characterizes the holistically low-rank structure of the error image by virtue of the nuclear norm directly. More importantly, the structured sparsity-inducing norm generally uses an L_2 or L_∞ norm as a measure within each group, neither of which can alleviate the correlations between variables. Thus, although the structured sparsity-inducing norm can induce nonzero patterns, it is incapable of eliminating the correlations of variables within the patterns. We know there are local correlations among errors particularly in the case of occlusions. The nuclear norm can alleviate these correlations via the involved singular value decomposition (SVD),[1] while the structured sparsity-inducing norm fails to do this. From this viewpoint, NMR should be more effective than SSRC in handling the occlusions or illumination changes.

In addition, there are a number of illumination-invariant face-recognition methods that have been developed. For instance, Savvides et al. [60] presented the ‘Corefaces’ which links the best of principal component analysis and advanced correlation filters for extracting illumination tolerance features. Chen et al. [61] suggested an illumination normalization approach in which a discrete cosine transform (DCT) is employed to compensate for illumination variations in the logarithm domain. These approaches were demonstrated effective for face recognition with illumination variations, but they might not be powerful for handling arbitrary occlusions. It should be mentioned that in this paper, our focus is not on illumination invariant techniques for face recognition, but on developing a general regression model which is more robust than the existing regression models to occlusion and illumination changes.

It should be mentioned that this paper is an extension of our arXiv paper [63]. Our ICPR paper [64] is a sparsified version of [63], i.e., the L_2 norm based regularization term was replaced by L_1 norm based one in the model. Differing from our previous work, this paper presents a fast ADMM algorithm to solve the NMR model and gives detailed analysis on convergence rate of the algorithm. An extended NMR model is developed for dealing with the face alignment and classification problems simultaneously. In addition, more experiments are conducted to compare our algorithms with state-of-the-art methods.

3 Nuclear Norm Based Matrix Regression Model and Its ADMM Algorithm

This section first presents the nuclear norm based matrix regression model, and then uses the alternating direction method of multipliers to solve the model. Finally, we give a convergence analysis of the proposed algorithm.

3.1 Nuclear Norm Based Matrix Regression Model

Given a set of n image matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n \in R^{p\times q}$ and an image matrix $B \in R^{p\times q}$, let us represent $\mathbf{B}$ linearly using $\mathbf{A}_1, \ldots, \mathbf{A}_n$, i.e.,

$$\mathbf{B} = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 +, \ldots, + x_n\mathbf{A}_n + \mathbf{E}, \tag{1}$$

where $x_1, x_2, \ldots, x_n$ is a set of representation coefficients, and $\mathbf{E}$ is the representation residual.

Let us define the following linear mapping from R^n to $R^{p\times q}$:

$$\mathrm{A}(\mathbf{x}) = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 +, \ldots, + x_n\mathbf{A}_n. \tag{2}$$

Then, the formula (1) becomes

$$\mathbf{B} = \mathrm{A}(\mathbf{x}) + \mathbf{E}. \tag{3}$$

The formula (3) or (1) gives a general form of a linear *matrix regression* model, in contrast with the classical linear *vector regression* model.

Motivated by observations or requirements that the residual image $\mathrm{A}(\mathbf{x}) - \mathbf{B}$ at the optimal solution is typically low rank (or approximately low rank) in many applications, we would like to evaluate the regression coefficients via solving the following nuclear norm approximation problem [27]

$$\min_{\mathbf{x}} \|\mathrm{A}(\mathbf{x}) - \mathbf{B}\|_*. \tag{4}$$

Moreover, borrowing the idea of the Ridge regression, we would like to add a similar regularization term to Eq. (4) and obtain the regularized *matrix regression* model

$$\min_{\mathbf{x}} \|\mathrm{A}(\mathbf{x}) - \mathbf{B}\|_* + \tfrac{1}{2}\lambda\|\mathbf{x}\|_2^2. \tag{5}$$

We will discuss how to solve this model in the following section.

3.2 ADMM Algorithm for NMR

The alternating direction method of multipliers or the augmented Lagrange multipliers (ALM) method has been applied to the nuclear norm optimization problems [28], [29]. For more details of ADMM, see [30]. Motivated by Hansson’s work [29], we here provide details

[1] Ref. [62] proves that the horizontal 2DPCA transform can eliminate the correlation between column vectors of image matrices, and the vertical 2DPCA transform can eliminate the correlation between row vectors of image matrices. If one uses one single error image as the input of 2DPCA model, the horizontal and vertical 2DPCA transform matrices are exactly the orthonormal matrices of SVD. So, SVD can eliminate the correlation between rows and columns of the error image.

of using ADMM to solve the regularized matrix regression problem.

The model in (5) can be rewritten as

$$\min||\mathbf{Y}||_* + \tfrac{1}{2}\lambda||\mathbf{x}||_2^2 \;\; \textit{subject to} \;\; \mathrm{A}(\mathbf{x}) - \mathbf{B} = \mathbf{Y}. \qquad (6)$$

The augmented Lagrangian function L_μ is defined by

$$L_\mu(\mathbf{Y},\mathbf{x},\mathbf{Z}) = ||\mathbf{Y}||_* + \tfrac{1}{2}\lambda||\mathbf{x}||_2^2 + \mathrm{Tr}\big(\mathbf{Z}^T(\mathbf{A}(\mathbf{x}) - \mathbf{Y} - \mathbf{B})\big) + \tfrac{\mu}{2}||\mathbf{A}(\mathbf{x}) - \mathbf{Y} - \mathbf{B}||_F^2, \qquad (7)$$

where $\mu > 0$ is a penalty parameter, $\mathbf{Z}$ is the array of Lagrange multipliers, and $\mathrm{Tr}(\cdot)$ is the trace operator. Note that if $\mu = 0$, Eq. (7) becomes the standard Lagrangian function.

ADMM consists of the following iterations

(i) Given$\mathbf{Y} = \mathbf{Y}^k$ and $\mathbf{Z} = \mathbf{Z}^k$, updating $\mathbf{x}$ by

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} L_\mu(\mathbf{Y},\mathbf{x},\mathbf{Z}). \qquad (8)$$

(ii) Given $\mathbf{x} = \mathbf{x}^{k+1}$ and $\mathbf{Z} = \mathbf{Z}^k$, updating $\mathbf{Y}$ by

$$\mathbf{Y}^{k+1} = \arg\min_{\mathbf{Y}} L_\mu(\mathbf{Y},\mathbf{x},\mathbf{Z}). \qquad (9)$$

(iii) Given $\mathbf{x} = \mathbf{x}^{k+1}$ and $\mathbf{Y} = \mathbf{Y}^{k+1}$, Updating $\mathbf{Z}$ by

$$\mathbf{Z}^{k+1} = \mathbf{Z}^k + \mu(\mathbf{A}(\mathbf{x}) - \mathbf{Y} - \mathbf{B}). \qquad (10)$$

The key steps are to solve the optimization problems in Eqs. (8) and (9).

After some derivations, we can rewrite the augmented Lagrangian function as

$$L_\mu(\mathbf{Y},\mathbf{x},\mathbf{Z}) = ||\mathbf{Y}||_* + \tfrac{1}{2}\lambda||\mathbf{x}||_2^2 + \tfrac{\mu}{2}||\mathbf{A}(\mathbf{x}) - (\mathbf{B} + \mathbf{Y} - \tfrac{1}{\mu}\mathbf{Z})||_F^2 - \tfrac{1}{2\mu}||\mathbf{Z}||_F^2. \qquad (11)$$

Based on the augmented Lagrangian function in Eq. (11), Eq. (8) can be expressed as

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}}\Big(\tfrac{\mu}{2}||\mathbf{A}(\mathbf{x}) - (\mathbf{B} + \mathbf{Y} - \tfrac{1}{\mu}\mathbf{Z})||_F^2 + \tfrac{1}{2}\lambda||\mathbf{x}||_2^2\Big) \qquad (12)$$

Letting $\mathbf{H} = [\mathrm{Vec}(\mathbf{A}_1),\ldots,\mathrm{Vec}(\mathbf{A}_n)]$, we can rewrite $\mathbf{A}(\mathbf{x}) = \sum_{j=1}^n x_j\mathbf{A}_j$ into the matrix form $\mathbf{Hx}$. Denote $\mathbf{g} = \mathrm{Vec}(\mathbf{B} + \mathbf{Y} - \frac{1}{\mu}\mathbf{Z})$. Thus, Eq. (12) is equivalent to

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}}\Big(||\mathbf{Hx} - \mathbf{g}||_2^2 + \tfrac{\lambda}{\mu}||\mathbf{x}||_2^2\Big). \qquad (13)$$

Since Eq. (13) is a standard Ridge regression model, we can get its closed-form solution

$$\mathbf{x}^{k+1} = (\mathbf{H}^T\mathbf{H} + \tfrac{\lambda}{\mu}\mathbf{I})^{-1}\mathbf{H}^T\mathbf{g}. \qquad (14)$$

Based on the augmented Lagrangian function in Eq. (11), Eq. (9) can be expressed as

$$\mathbf{Y}^{k+1} == \arg\min_{\mathbf{Y}}\Big(\tfrac{1}{\mu}||\mathbf{Y}||_* + \tfrac{1}{2}||\mathbf{Y} - (\mathbf{A}(\mathbf{x}) - \mathbf{B} + \tfrac{1}{\mu}\mathbf{Z})||_F^2\Big). \qquad (15)$$

The optimal solution can be computed via the singular value thresholding algorithm [31]. Specifically, given a matrix $\mathbf{Q} \in R^{p\times q}$ of rank r, the singular value decomposition of X is

$$\mathbf{Q} = \mathbf{U}_{p\times r}\Sigma\mathbf{V}_{q\times r}^T, \Sigma = diag(\sigma_1,\ldots,\sigma_r), \qquad (16)$$

where $\sigma_1,\ldots,\sigma_r$ are positive singular values, and $\mathbf{U}_{p\times r}$ and $\mathbf{V}_{q\times r}$ are corresponding matrices with orthogonal columns.

For a given $\tau > 0$, the singular value shrinkage operator is defined as follows

$$D_\varsigma(\mathbf{Q}) = \mathbf{U}_{p\times r}\mathrm{diag}\Big(\{\max(0, \sigma_j - \varsigma)\}_{1\le j\le r}\Big)\mathbf{V}_{q\times r}^T. \qquad (17)$$

Theorem 1. *[31] For each $\mathbf{Q} \in R^{p\times q}$ and $\varsigma > 0$, the singular value shrinkage operator in (17) obeys*

$$D_\varsigma(\mathbf{Q}) = \arg\min_{\mathbf{Y}}\Big(\varsigma||\mathbf{Y}||_* + \tfrac{1}{2}||\mathbf{Y} - \mathbf{Q}||_F^2\Big). \qquad (18)$$

From Theorem 1, the optimal solution of (15) is

$$\mathbf{Y} = D_{\frac{1}{\mu}}\Big(\mathbf{A}(\mathbf{x}) - \mathbf{B} + \tfrac{1}{\mu}\mathbf{Z}\Big). \qquad (19)$$

In summary, the core of ADMM algorithm for the nuclear norm based matrix regression problem involves two sub-problems: the ridge regression and the singular value thresholding.

Boyd et al. [30] give the optimality conditions and stopping criteria of the ADMM algorithm. Based on the results in [29], [30], we use the following termination criterion: the primal and dual residuals must be small, i.e.,

$$||\mathbf{r}_{pri}^k||_2 \le \varepsilon_{pri} \text{ and } ||\mathbf{s}_{dual}^k||_2 \le \varepsilon_{dual} \qquad (20)$$

where $\mathbf{r}_{pri}^k = \mathbf{A}(\mathbf{x}^k) - \mathbf{Y}^k - \mathbf{B}$, $\varepsilon_{pri} = \sqrt{pq}\varepsilon_{abs} + \varepsilon_{rel}\max\{||\mathrm{A}(\mathbf{x})||_F, ||\mathbf{Y}||_F, ||\mathbf{B}||_F\}$, $\mathbf{s}_{dual}^k = \mu\,\mathbf{H}^T\,\mathrm{Vec}(\mathbf{Y}^k - \mathbf{Y}^{k-1})$, and $\varepsilon_{dual} = \sqrt{n}\varepsilon_{abs} + \varepsilon_{rel}||\mathbf{H}^T\,\mathrm{Vec}(\mathbf{Z})||_2$.

Finally, we would like to further elaborate upon our algorithm. If we fix penalty parameter μ in the augmented Lagrangian function, the updating $\mathbf{x}$ step via solving the ridge regression problem can be computed more efficiently. Looking back at Eq. (14), let $\mathbf{M} = (\mathbf{H}^T\mathbf{H} + \frac{\lambda}{\mu}\mathbf{I})^{-1}\mathbf{H}^T$, which is fixed in each iteration, so it can be calculated and stored in advance. Then, in each iteration for updating $\mathbf{x}$, we only need to update $\mathbf{g} = \mathrm{Vec}(\mathbf{B} + \mathbf{Y} - \frac{1}{\mu}\mathbf{Z})$ and carry out the matrix multiplication $\mathbf{Mg}$ once. On the other hand, in the iteration for updating $\mathbf{Y}$, the main computation is consumed for performing the singular value decomposition of the matrix $\mathbf{Q} = \mathbf{A}(\mathbf{x}) - \mathbf{B} + \frac{1}{\mu}\mathbf{Z}$. Since $\mathbf{Q}$ has the same size as the image matrix, the computational complexity of this step only depends on the size of images.

The detailed ADMM algorithm for NMR is summarized in Algorithm 1.

Algorithm 1 can be interpreted in the two-step iteration strategy for robust face recognition as adopted in [11], [12]. Step 3, updating x, is actually an error-detection step for determining the representation coefficients and representation errors; and Step 4, updating $\mathbf{Y}$, is actually an error-support step for determining the real corrupted part. So, we can say that NMR provides a unified framework to integrate error detection and error support into one simple model.

Algorithm 1. ADMM Algorithm for NMR

```
Input: A set of image matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and an image matrix $\mathbf{B} \in R^{p\times q}$, the model parameters $\lambda$ and $\mu$, the termination condition parameters $\varepsilon_{abs}$and $\varepsilon_{rel}$.
1: Let $\mathbf{H} = [\mathrm{Vec}(\mathbf{A}_1), \ldots, \mathrm{Vec}(\mathbf{A}_n)]$. Compute
   $\mathbf{M} = (\mathbf{H}^T\mathbf{H} + \frac{\lambda}{\mu}\mathbf{I})^{-1}\mathbf{H}^T$;
2: $\mathbf{Y}^0 = -\mathbf{B}$, $\mathbf{Z}^k = \mathbf{0}$, $\mathrm{k} = 0$;
3: Updating $\mathbf{x}$: Let $\mathbf{g} = \mathrm{Vec}\Big(\mathbf{B} + \mathbf{Y}^k - \frac{1}{\mu}\mathbf{Z}^k\Big)$. $\mathbf{x}^{k+1} = \mathbf{Mg}$;
4: Updating $\mathbf{Y}$: $\mathbf{Y}^{k+1} = D_{\frac{1}{\mu}}\Big(\mathbf{A}(\mathbf{x}^{k+1}) - \mathbf{B} + \frac{1}{\mu}\mathbf{Z}^k\Big)$;
5: Updating $\mathbf{Z}$: $\mathbf{Z}^{k+1} = \mathbf{Z}^k + \mu\big(\mathbf{A}(\mathbf{x}^{k+1}) - \mathbf{Y}^{k+1} - \mathbf{B}\big)$;
6: If Eq. (20) is not satisfied go to Step 3.
Output: Optimal regression coefficient vector $\mathbf{x}^{k+1}$
```

3.2.1 The Computational Complexity of the NMR Algorithm

Given the training sample size n and the image size $p \times q$, let $m = p \times q$. The computational complexity of Step 3 is $O(mn)$, which is determined by the matrix multiplication $\mathbf{Mg}$. The computational complexity of Step 4 is $O(\min(p^2q, pq^2))$, which is determined by the singular value decomposition of a $p \times q$ matrix $\mathbf{Q} = \mathbf{A}(\mathbf{x}) - \mathbf{B} + \frac{1}{\mu}\mathbf{Z}$. In the case that $p = q$, the computational complexity becomes $O(m^{1.5})$. So, the computational complexity of the NMR Algorithm is $O(k(m^{1.5} + mn))$, where k is the number of iterations.

3.3 Convergence Analysis

In this section, we will give a convergence analysis of ADMM. Algorithm 1 is a special case of a more general class of augmented Lagrange multiplier algorithms known as the alternating directions methods [40]. The convergence of these algorithms has been studied extensively (see, [41], [42] and the many references therein, as well as discussions in [28], [40]). In recent years, the existence of the saddle points is widely assumed for the convergence of algorithms. For instance, Boyd et al. investigated convergence of ADMM by virtue of the properties of the saddle points, and give three important results: Residual convergence, Objective convergence and Dual variable convergence [33]. However, the objective convergence cannot deduce the optimal point of the iterative process. If the optimal point of the iterative process could be identified, the iterative trend would be clearer. Thus, we here mainly study the accumulation points of the iterative variables for Algorithm 1.

Let $(\mathbf{Y}^\star, \mathbf{x}^\star, \mathbf{Z}^\star)$ be a saddle point of the following Lagrangian function $L(\mathbf{Y},\ \mathbf{x},\ \mathbf{Z}) = \|\mathbf{Y}\|_* + \frac{1}{2}\lambda\,||\mathbf{x}||_2^2 + \mathrm{Tr}\big(\mathbf{Z}^T(\mathbf{A}(\mathbf{x}) - \mathbf{Y} - \mathbf{B})\big)$, and $q^k = \|\mathbf{Y}^k\|_* + \frac{\lambda}{2}\|\mathbf{x}^k\|_2^2$, $q^\star = \|\mathbf{Y}^\star\|_* + \frac{\lambda}{2}\|\mathbf{x}^\star\|_2^2$, $\mathbf{R}^k = \mathbf{A}\big(\mathbf{x}^k\big) - \mathbf{Y}^k - \mathbf{B}$, $\mathbf{r}^k = \mathrm{Vec}(\mathbf{R}^k)$. According to the analysis in [33], [43], finding the optimal solutions of original and dual problems is equivalent to finding a saddle point of the function L. Thus, $\mathbf{Z}^\star$ is dual optimal. In addition, we know that $\mathbf{Z}^k \to \mathbf{Z}^\star$, as $k \to \infty$ from [44].

Theorem 2. *If $\mu > 0$, then the sequence $\big\{\big(\mathbf{Y}^k, \mathbf{x}^k, \mathbf{Z}^k\big)\big\}$ generated by Algorithm 1 converges to a saddle point $(\mathbf{Y}^\star, \mathbf{x}^\star, \mathbf{Z}^\star)$ of the Lagrangian function L.*

The proof is given in supplemental materials, available online.

Theorem 2 implies the convergence trend of the sequence $\big\{\big(\mathbf{Y}^k, \mathbf{x}^k, \mathbf{Z}^k\big)\big\}$ generated by Algorithm 1. We know that the convergence rate is another important concept, which reflects the convergence speed of an iterative algorithm. The authors of [45], [46] has showed that ADMM can achieve $O(1/k)$ global convergence, where k is the number of iterations, under a strong convexity assumption. Without this strong convexity assumption, He and Yuan [47] presented the most general result to date of $O(1/k)$ convergence rate for ADMM. Their results only require that both objective-function terms are convex (not necessarily smooth). Since here $||\mathbf{Y}||_*$ and $\frac{1}{2}\lambda||\mathbf{x}||_2^2$ are both convex, Algorithm 1 can achieve $O(1/k)$ convergence.

4 Fast ADMM Algorithm

From Goldstein et al.'s work [46], we know that the ADMM algorithm can be accelerated to achieve the optimal convergence rate of $O(1/k^2)$ under the condition that both additive terms in the objective function are strongly convex. However, the objective function of the NMR model does not satisfy this condition because the first term $\|\mathrm{A}(x) - \mathbf{B}\|_*$ is not strongly convex. To address this problem, we construct an approximate NMR model in which the objective-function terms are both strongly convex. Fortunately, we can prove that the optimal solution of the approximate NMR approaches to that of NMR when the multiplier $\gamma \to 0$.

4.1 Approximate NMR Model

The approximate NMR model is constructed as follows

$$\begin{aligned}&\min||\mathbf{Y}||_* + \gamma\Big(||\mathbf{Y}||_F^2 + \tfrac{1}{2}\lambda||\mathbf{x}||_2^2\Big) + \tfrac{1}{2}\lambda||\mathbf{x}||_2^2\\ &\mathit{subject\ to}\ \mathrm{A}(\mathbf{x}) - \mathbf{B} = \mathbf{Y}\end{aligned} \tag{21}$$

Denoting $\theta = \lambda(1 + \gamma)$, the above model becomes

$$\min||\mathbf{Y}||_* + \gamma\|\mathbf{Y}\|_F^2 + \tfrac{1}{2}\theta||\mathbf{x}||_2^2\ \mathit{subject\ to}\ \mathrm{A}(\mathbf{x}) - \mathbf{B} = \mathbf{Y} \tag{22}$$

In the following, we will show that minimizing the approximate objective function $f(\mathbf{Y}, \mathbf{x}) = ||\mathbf{Y}||_* + \gamma||\mathbf{Y}||_{\mathrm{F}}^2 + \frac{1}{2}\theta||\mathbf{x}||_2^2$ is the same as minimizing the objective for problem (6) in the limit of small γ's.

Theorem 3. *Let $(\mathbf{Y}_\gamma^\star, \mathbf{x}_\gamma^\star)$ be the solution to (22) and $(\mathbf{Y}^\star, \mathbf{x}^\star)$ be the solution to problem (6), then*

$$\min_{\gamma\to 0}\Big\|\mathbf{Y}_\gamma^\star - \mathbf{Y}^\star\Big\|_F^2 + \Big\|\mathbf{x}_\gamma^\star - \mathbf{x}^\star\Big\|_2^2 = 0.$$

The proof of Theorem 3 is given in supplemental materials, available online.

4.2 Fast ADMM Algorithm

Following the derivation of ADMM as shown in Section 3.1, it is easy to obtain the ADMM algorithm for model (22). Specifically, in the step of updating $\mathbf{x}$, since $\mathbf{Y}$ and $\mathbf{Z}$ are fixed, the additional term $\gamma\|\mathbf{Y}\|_{\mathrm{F}}^2$ does not affect the optimization of $\mathbf{x}$. Thus, we can use Step 3 in Algorithm 1 for updating $\mathbf{x}$, as long as λ is replaced by θ in the definition of $\mathbf{M}$, i.e., $\mathbf{M} = (\mathbf{H}^T\mathbf{H} + \frac{\theta}{\mu}\mathbf{I})^{-1}\mathbf{H}^T$.

In the step of updating $\mathbf{Y}$ for solving Model (22), since $\mathbf{x}$ and $\mathbf{Z}$ are fixed, we have

$$\begin{aligned}\mathbf{Y}_{\gamma}^{k+1} &= \arg\min_{\mathbf{Y}}\Big(||\mathbf{Y}||_{*} + \mathrm{Tr}\big(\mathbf{Z}^{T}(\mathbf{A}(\mathbf{x}) - \mathbf{Y} - \mathbf{B})\big) \\ &\quad + \tfrac{\mu}{2}||\mathbf{A}(\mathbf{x}) - \mathbf{Y} - \mathbf{B}||_{F}^{2} + \gamma||\mathbf{Y}||_{F}^{2}\Big) \\ &= \arg\min_{\mathbf{Y}}\Big(||\mathbf{Y}||_{*} + \tfrac{\mu}{2}||\mathbf{A}(\mathbf{x}) - \mathbf{Y} - \mathbf{B} + \tfrac{1}{\mu}\mathbf{Z}||_{F}^{2} + \gamma||\mathbf{Y}||_{F}^{2} - \tfrac{1}{2\mu}||\mathbf{Z}||_{F}^{2}\Big) \\ &= \arg\min_{\mathbf{Y}}\Big(||\mathbf{Y}||_{*} + (\tfrac{\mu}{2} + \gamma)||\mathbf{Y} - \tfrac{\mu}{\mu+2\gamma}\Big(\tfrac{1}{\mu}\mathbf{Z} + \mathbf{A}(\mathbf{x}) - \mathbf{B}\Big)||_{F}^{2} + const\Big) \\ &= \arg\min_{\mathbf{Y}}\Big(\tfrac{1}{\mu+2\gamma}||\mathbf{Y}||_{*} + \tfrac{1}{2}||\mathbf{Y} - \tfrac{\mu}{\mu+2\gamma}\Big(\mathbf{A}(\mathbf{x}) - \mathbf{B} + \tfrac{1}{\mu}\mathbf{Z}\Big)||_{F}^{2}\Big).\end{aligned} \tag{23}$$

Recalling that in the step of updating $\mathbf{Y}$ for solving the original model (6), we have

$$\mathbf{Y}^{k+1} = D_{\varsigma}(\mathbf{Q}) = \mathbf{U}_{p\times r}\,\mathrm{diag}(\{\max(0, \sigma_j - \varsigma)\}_{1\le j\le r})\mathbf{V}_{q\times r}^{T},$$

where $\mathbf{Q} = (\mathbf{A}(\mathbf{x}) - \mathbf{B} + \frac{1}{\mu}\mathbf{Z})$ and $\varsigma = \frac{1}{\mu}$.

Let us define $\mathbf{Q}^{\gamma} = \frac{\mu}{\mu+2\gamma}\mathbf{Q} = \frac{\mu}{\mu+2\gamma}(\mathbf{A}(\mathbf{x}) - \mathbf{B} + \frac{1}{\mu}\mathbf{Z})$, $\varsigma^{\gamma} = \frac{\mu}{\mu+2\gamma}\varsigma = \frac{\mu}{\mu+2\gamma}\frac{1}{\mu}$. It is easy to know that $\mathbf{Q}^{\gamma}$ and $\mathbf{Q}$ share the same singular vectors and the corresponding singular values satisfy $\sigma_j^{\gamma} = \frac{\mu}{\mu+2\gamma}\sigma_j$, $j = 1, \ldots, r$. Therefore, using Theorem 1, it can be derived from Eq. (23) that

$$\begin{aligned}\mathbf{Y}_{\gamma}^{k+1} &= D_{\varsigma}(\mathbf{Q}^{\gamma}) = \tfrac{\mu}{\mu+2\gamma}\mathbf{U}_{p\times r}\,\mathrm{diag}\left(\{\max(0, \sigma_j - \varsigma)\}_{1\le j\le r}\right)\mathbf{V}_{q\times r}^{T} \\ &= \tfrac{\mu}{\mu+2\gamma}D_{\varsigma}(\mathbf{Q}).\end{aligned} \tag{24}$$

From Eq. (24), it is easy to understand the conclusion of Theorem 4, because

$$\mathbf{Y}_{\gamma}^{k+1} \to \mathbf{Y}^{k+1}, \text{when } \gamma \to 0.$$

Based on the above analysis, we know that we can use a similar ADMM algorithm, a modified version of Algorithm 1, to solve the approximate NMR model (22). The modifications include: (i) In Step 1, $\mathbf{M}$ is redefined as $\mathbf{M} = (\mathbf{H}^{T}\mathbf{H} + \frac{\theta}{\mu}\mathbf{I})^{-1}\mathbf{H}^{T}$, and (ii) In Step 4 for updating $\mathbf{Y}$: a multiplier $\frac{\mu}{\mu+2\gamma}$ is added, i.e., the operator in Step 4 is replaced by $\mathbf{Y}^{k+1} = \frac{\mu}{\mu+2\gamma}D_{\frac{1}{\mu}}(\mathbf{A}(\mathbf{x}^{k+1}) - \mathbf{B} + \frac{1}{\mu}\mathbf{Z}^{k})$.

An acceleration scheme was originally presented by Nesterov [48] for solving a convex programming problem with a convergence rate of $O(1/k^2)$. Subsequently, much work has been done applying Nesterov's concept to other first-order methods. More recently, the Nesterov-type scheme was also applied to accelerate the alternating direction methods [46], [50]. Based on Goldstein et al.'s work [46], we develop the fast ADMM Algorithm (Algorithm 2) for our approximate NMR model.

In the accelerated case the primal residual is unchanged, $\mathbf{r}_{pri}^{k} = \mathbf{A}(\mathbf{x}^{k}) - \mathbf{Y}^{k} - \mathbf{B}$. A simple derivation yields the new dual residual $\mathbf{s}_{dual}^{k} = \mu\,\mathrm{A}(\mathbf{x}^{k} - \mathbf{x}^{k-1})$. Based on Lemma 6 in [47], we can use the following termination criterion:

$$||\mathbf{r}_{pri}^{k}||_2 \le \varepsilon_{pri} \text{ and } ||\mathbf{s}_{dual}^{k}||_2 \le \varepsilon_{dual}, \tag{25}$$

where $\varepsilon_{pri} = \sqrt{pq}\varepsilon_{abs} + \varepsilon_{rel}\max\{||\mathrm{A}(\mathbf{x})||_F, ||\mathbf{Y}||_F, ||\mathbf{B}||_F\}$, and $\varepsilon_{dual} = \sqrt{pq}\varepsilon_{abs} + \varepsilon_{rel}||\mathbf{Z}||_F$.

Algorithm 2. Fast ADMM Algorithm for NMR

```
Input: A set of image matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and an image matrix $\mathbf{B} \in R^{p\times q}$, the model parameters $\lambda$, $\mu$ and $\gamma$, the termination condition parameters $\varepsilon_{abs}$ and $\varepsilon_{rel}$.
1: Let $\mathbf{H} = [\mathrm{Vec}(\mathbf{A}_1), \ldots, \mathrm{Vec}(\mathbf{A}_n)]$. Compute
   $\mathbf{M} = (\mathbf{H}^{T}\mathbf{H} + \frac{\theta}{\mu}\mathbf{I})^{-1}\mathbf{H}^{T}$, where $\theta = \lambda(1+\gamma)$;
2: $\mathbf{x}^0 = \hat{\mathbf{x}}^0 = \mathbf{0}, \mathbf{Z}^0 = \hat{\mathbf{Z}}^0 = \mathbf{0}, \alpha^0 = 1, \mathrm{k} = 0$;
3: Updating $\mathbf{Y}$: $\mathbf{Y}^{k+1} = \frac{\mu}{\mu+2\gamma}D_{\frac{1}{\mu}}(\mathbf{A}(\hat{\mathbf{x}}^{k}) - \mathbf{B} + \frac{1}{\mu}\hat{\mathbf{Z}}^{k})$;
4: Updating $\mathbf{x}$: Let $\mathbf{g} = \mathrm{Vec}\big(\mathbf{B} + \mathbf{Y}^{k+1} - \frac{1}{\mu}\hat{\mathbf{Z}}^{k}\big)$. $\mathbf{x}^{k+1} = \mathbf{Mg}$;
5: Updating $\mathbf{Z}$: $\mathbf{Z}^{k+1} = \hat{\mathbf{Z}}^{k} + \mu\big(\mathbf{A}(\mathbf{x}^{k+1}) - \mathbf{Y}^{k+1} - \mathbf{B}\big)$;
6: Updating $\alpha$: $\alpha^{k+1} = \frac{1+\sqrt{1+4(\alpha^k)^2}}{2}$;
7. Updating $\hat{\mathbf{x}}$: $\hat{\mathbf{x}}^{k+1} = \mathbf{x}^{k+1} + \frac{\alpha^k-1}{\alpha^{k+1}}\big(\mathbf{x}^{k+1} - \mathbf{x}^{k}\big)$;
8. Updating $\hat{\mathbf{Z}}$: $\hat{\mathbf{Z}}^{k+1} = \mathbf{Z}^{k+1} + \frac{\alpha^k-1}{\alpha^{k+1}}\big(\mathbf{Z}^{k+1} - \mathbf{Z}^{k}\big)$;
9. If Eq. (25) is not satisfied go to Step 3.
Output: Optimal regression coefficient vector $\hat{\mathbf{x}}^{k+1}$
```

4.3 Convergence Analysis

To discuss the convergence of the Fast ADMM Algorithm for NMR, we need to introduce the concept of the strongly convex function. A function $f(\mathbf{x})$ is called a strongly convex function with parameter $\eta_f > 0$ if the following inequality holds for all $\mathbf{x}, \mathbf{y}$ in its domain and $t \in [0, 1]$:

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \le tf(\mathbf{x}) + (1-t)f(\mathbf{y}) - \tfrac{1}{2}\eta_f t(1-t)\|\mathbf{x} - \mathbf{y}\|_2^2.$$

Intuitively, strong convexity means that a function lies above its local quadratic approximation. From the definition of the strongly convex function, it is not hard to derive that

Lemma 1. *A function $f(\mathbf{x})$ is strongly convex with parameter η_f if and only if the function $\mathbf{x} \mapsto f(\mathbf{x}) - \frac{\eta_f}{2}\|\mathbf{x}\|_2^2$ is convex.*

The proof is given in supplemental materials, available online.

For Model (22), let $\mathrm{P}(\mathbf{Y}) = ||\mathbf{Y}||_* + \gamma|\mathbf{Y}||_F^2$, $\mathrm{Q}(\mathbf{x}) = \frac{1}{2}\theta||\mathbf{x}||_2^2$. Since $||\mathbf{Y}||_*$ is convex, by Lemma 1, we know both $\mathrm{P}(\mathbf{Y})$ and $\mathrm{Q}(\mathbf{x})$ are strongly convex functions. Meanwhile, it is easy to see that the strong convexity parameters η_P and η_Q of $\mathrm{P}(\mathbf{Y})$ and $\mathrm{Q}(\mathbf{x})$ are 2γ and θ, respectively. In addition, the conjugate of a convex function F, denoted F^*, is defined as $\mathrm{F}^*(\mathbf{p}) = \sup_{\mathbf{u}}\langle\mathbf{u}, \mathbf{p}\rangle - \mathrm{F}(\mathbf{u})$. From [46], we know that the fast ADMM can achieve $O(1/k^2)$ convergence under the condition that both objective-function terms are strongly convex. Therefore, we have the following convergence theorem for Algorithm 2:

Theorem 4. *If we choose $\mu \le \frac{2\gamma\theta^2}{\rho(\mathbf{H}^T\mathbf{H})^2}$, then the iterates $\{\mathbf{Z}^k\}$ generated by Algorithm 2 satisfies*

$$D(\mathbf{Z}^{\star}) - D(\mathbf{Z}^{k}) \le \frac{2\|\hat{\mathbf{Z}}^{1} - \mathbf{Z}^{\star}\|}{\mu(k+2)^2},$$

where $D(\mathbf{Z}) = -\mathrm{P}^*(\mathbf{Z}) - \mathrm{Q}^*\big(-\mathbf{H}^T\mathrm{vec}(\mathbf{Z})\big) - \langle\mathbf{Z}, \mathbf{B}\rangle$ is dual to problem (22), $\mathbf{Z}^{\star}$ is a Lagrange multiplier that maximizes the dual, and $\rho\big(\mathbf{H}^T\mathbf{H}\big)$ is the spectral radius of the matrix $\mathbf{H}^T\mathbf{H}$.

Theorem 4 means that the Fast ADMM Algorithm for NMR has a convergence rate of $O(1/k^2)$, where k is the number of iterations. Thus, the accelerated variant of ADMM, Algorithm 2, exhibits faster convergence speed than the conventional ADMM, Algorithm 1. Fig. 1 shows

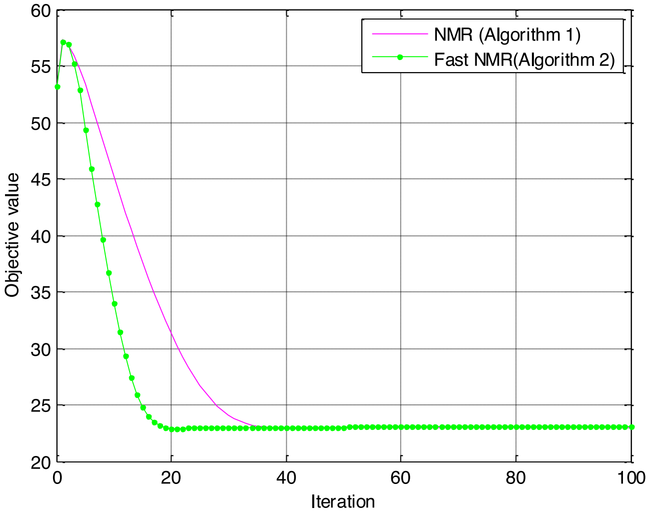


Fig. 1. Comparison of convergence rate of the NNR (Algorithm 1) and its fast version (Algorithm 2).

an example of removing the white-block occlusion of an image via NMR, where ADMM (Algorithm 1) converges after nearly 40 iterations while Fast ADMM (Algorithm 2) only needs 20 iterations. Since Fast ADMM consumes almost the same computation as ADMM in each iteration step, Algorithm 2 is nearly two times faster than Algorithm 1. Here, the CPU time of Fast ADMM is 0.0211 second, while that of ADMM is 0.0412 second.

5 NMR Based Robust Classification

NMR uses the nuclear norm to characterize the residual image (error image) $\mathbf{E}$. In this section, we first give justifications for why a nuclear norm is suitable for characterizing the error image caused by occlusions or illumination variations. We then present the NMR based classifier, which can handle well-aligned image recognition problems. We further extend the NMR model to do face alignment and classification simultaneously.

5.1 Justification of Claim that Nuclear Norm is Robust to Occlusions and Illumination

In this section, we will provide a probabilistic explanation for why we use nuclear norm to characterize the error images caused by occlusions and illumination changes. Noticing that in our motivation of modeling, we use the term of "low-rank" because it is an intuitive concept for describing the error image caused by occlusion. Actually, our model is not limited to "low-rank" error images, because it does not optimize the rank function directly but optimizes the "nuclear norm" of the error image instead. Nuclear norm provides a more flexible characterization of the error image than "rank" function, because it is still useful for characterizing an "approximately low-rank" error image, which may be algebraically full-rank, but many singular values of it are very close to zero. For example, in the case of illumination changes, assuming that the faces are Lambertian and generally of smooth geometry, the elements in the error image are highly correlated. Therefore, the error image caused by illumination changes is generally "approximately low-rank".

We know that the nuclear norm of a matrix is the sum of all singular values of the matrix, which is actually the L_1 norm of the singular value vector. From the probability distribution point of view, we know that the L_1 norm provides an optimal characterization for random variables with the Laplacian distribution, while the L_2 norm is optimal for a Gaussian distribution [6], [7]. Therefore, if the singular values of an error image satisfy the Laplacian distribution, the nuclear norm will provide a good characterization of the error image.

Fig. 2 shows examples of the error images caused by occlusions and illumination changes, where (a1) and (a2) are the original images which can be viewed as the expected, reconstructed image; (b1) and (b2) are, respectively, the corresponding occluded image and the image with a different illumination. The error image (c1) is the difference between (a1) and (b1), and the error image (c2) is the difference between (a2) and (b2). The former is low-rank, while the latter is approximately low-rank. Fig. 2 e1 and e2 illustrate the error image fitted by different distributions, where Gaussian and Laplacian distributions are far away from the empirical distribution. That is to say, the pixel-level errors do not follow Laplace or Gaussian distribution, either for occlusion or illumination change. However, Fig. 2 f1 and f2 show that singular values of the error image fit a Laplacian distribution well, both for occlusion and illumination changes. These examples show that the nuclear norm is more suitable for characterizing the error image than either the L_1 or L_2 norm. This observation provides us a probabilistic justification for using a nuclear norm under occlusions and illumination variations.

More instances of showing the robustness of nuclear-norm-based matrix regression to illumination changes and occlusions (including artificial occlusions and real-world occlusions caused by glasses and scarfs) are given in supplemental materials, available online. In addition, since the nuclear norm is more suitable for characterizing the error image than other norms, we will use the nuclear norm of the residual image as a similarity measure to design the rule for classification.

5.2 NMR Classifier

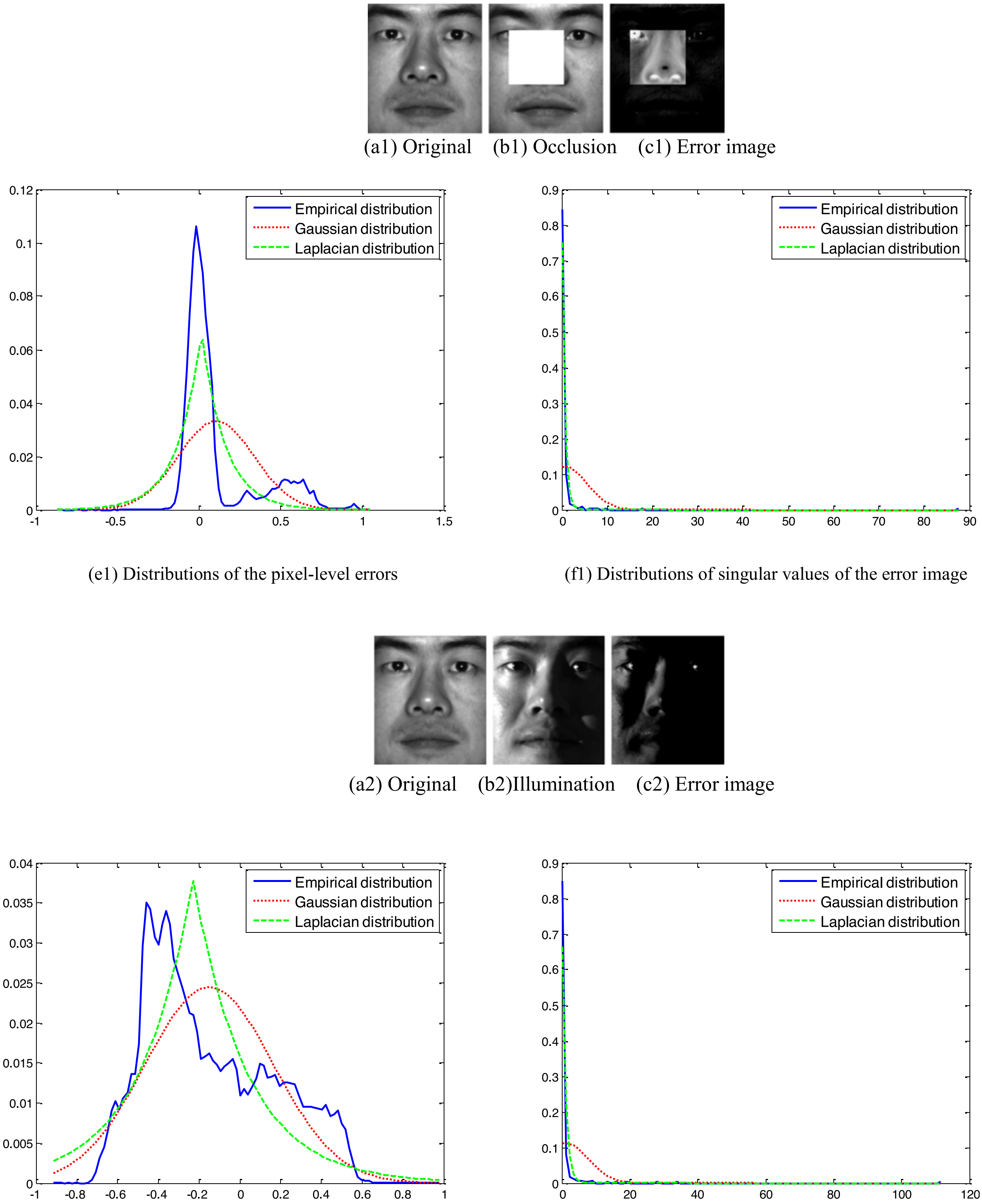
Similar to the strategy of SRC, we use the training samples of all classes to form the set of regressors. Let $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be training sample images of all classes. For a given test image B, we use all training samples to represent it and obtain the representation coefficient vector by solving the NMR model (or the approximate NMR Model) via Algorithm 1 (or Algorithm 2) and obtain the optimal solution.

Based on the optimal solution $\mathbf{x}^\star$, we get the reconstructed image of $\mathbf{B}$ as $\hat{\mathbf{B}} = \mathrm{A}(\mathbf{x}^\star)$, and the residual image $\mathbf{E} = \mathbf{B} - \hat{\mathbf{B}}$.

Let $\delta_i : R^n \to R^n$ be the characteristic function that selects the coefficients associated with the ith class. For $\mathbf{x} \in R^n$, $\delta_i(\mathbf{x})$ is a vector whose only nonzero entries are the entries in $\mathbf{x}$ that are associated with Class i. Using the coefficients associated with the ith class, one can get the reconstruction of $\mathbf{B}$ in Class i as $\hat{\mathbf{B}}_i = \mathrm{A}(\delta_i(\mathbf{x}^\star))$. The corresponding class reconstruction error is defined by

$$e_i(\mathbf{B}) = ||\hat{\mathbf{B}} - \hat{\mathbf{B}}_i||_* = \|\mathrm{A}(\mathbf{x}^\star) - \mathrm{A}(\delta_i(\mathbf{x}^\star))\|_*. \tag{26}$$

The decision rule is defined as: if $e_l(\mathbf{B}) = \min_i e_i(\mathbf{B})$, then $\mathbf{B}$ is assigned to Class l.



(a1) Original (b1) Occlusion (c1) Error image

(e1) Distributions of the pixel-level errors

(f1) Distributions of singular values of the error image

(a2) Original (b2)Illumination (c2) Error image

(e2) Distributions of the pixel-level errors

(f2) Distributions of singular values of the error image

Fig. 2. Example images with occlusion and illumination changes and the corresponding distributions of the pixel-level errors and singular values of the error image.

5.3 Robust Face Alignment and Classification with NMR

As mentioned above, NMR mainly concentrates on the classification of the aligned face images, which needs images in both the training set and the test set to be well-aligned. However, practically, the observed test image $\mathbf{B}'$ is always misaligned. In this section, we present an extended model of NMR to deal with the face alignment and classification problems simultaneously. Specifically, let us transform the misaligned image $\mathbf{B}'$ to a well aligned image $\mathbf{B}$ by

$\mathbf{B} = \mathbf{B}' \circ \tau$, where τ is a nonlinear transformation which is introduced to describe image deformation in the alignment process [10]. Motivated by the work in [10], [51], [52], we formulate our extended model as

$$\min||\mathbf{Y}||_* + \gamma||\mathbf{Y}||_F^2 + \tfrac{1}{2}\theta||\mathbf{x}||_2^2 \quad subject\ to \quad \tilde{\mathrm{A}}(\mathbf{x}) - \mathbf{B}' \circ \tau = \mathbf{Y}, \tag{27}$$

where $\tilde{\mathrm{A}}$ is the aligned training set, which can be calculated through an alignment method as suggested in [52]. However, Eq. (27) is hard to solve since it is a non-convex optimization problem. Using a similar strategy as in [10], [51], [52], we solve the problem via an iterative convex optimization framework, which iteratively linearizes the current estimate of τ and seeks for representations like

$$\begin{aligned} &\min||\mathbf{Y}||_* + \gamma||\mathbf{Y}||_F^2 + \tfrac{1}{2}\theta||\mathbf{x}||_2^2 \\ &subject\ to \quad \tilde{\mathrm{A}}(\mathbf{x}) - (\mathbf{B}' \circ \tau + \mathrm{Mat}(J\Delta\tau)) = \mathbf{Y}, \end{aligned} \tag{28}$$

where $J = \frac{\partial}{\partial \tau}\mathrm{Vec}(\mathbf{B}') \circ \tau \in R^{d \times v}$ is the Jacobian of $Vec(\mathbf{B}') \circ \tau$ with respect to the transformation parameters τ, $d = p \times q$ is the dimension of the image, and $\mathrm{Mat}(\cdot)$ is an operator converting the vector R^d into a matrix $R^{p\times q}$ and $\Delta\tau$ is the step in τ.

Algorithm 3. Fast ADMM Algorithm for Solving the Model in (28)

```
Input: A set of aligned training image matrices $\tilde{\mathbf{A}}_1, \ldots, \tilde{\mathbf{A}}_n$ and an image matrix $\mathbf{B}' \in R^{p\times q}$, the model parameters $\lambda$, $\mu$, $\gamma$ and initial transformation $\tau_0$ of $\mathbf{B}'$, the termination condition parameters $\varepsilon_{abs}$ and $\varepsilon_{rel}$.
Let $\mathbf{H} = [\mathrm{Vec}(\tilde{\mathbf{A}}_1), \ldots, \mathrm{Vec}(\tilde{\mathbf{A}}_n)]$. Compute
  $\mathbf{M} = (\mathbf{H}^T\mathbf{H} + \frac{\theta}{\mu}\mathbf{I})^{-1}\mathbf{H}^T$, where $\theta = \lambda(1+\gamma)$;
$\mathbf{x}^0 = \hat{\mathbf{x}}^0 = \mathbf{0}, \Delta\tau^0 = \Delta\hat{\tau}^0 = 0, \mathbf{Z}^0 = \hat{\mathbf{Z}}^0 = \mathbf{0}, \alpha^0 = 1, \mathrm{k} = 0$;
While not converged ($k = 0, 1, \ldots$) do
  Updating $\mathbf{Y}$: $\mathbf{Y}^{k+1} = \frac{\mu}{\mu+2\gamma}D_{\frac{1}{\mu}}(\tilde{\mathbf{A}}(\hat{\mathbf{x}}^k) - (\mathbf{B}' \circ \tau + \mathrm{Mat}(J\Delta\hat{\tau}^k))$
    $+\frac{1}{\mu}\hat{\mathbf{Z}}^k)$;
  Updating $\mathbf{x}$: Let $\mathbf{g} = \mathrm{Vec}((\mathbf{B}' \circ \tau + \mathrm{Mat}(J\Delta\hat{\tau}^k)) + \mathbf{Y}^{k+1}$
    $-\frac{1}{\mu}\hat{\mathbf{Z}}^k)$. $\mathbf{x}^{k+1} = \mathbf{Mg}$;
  Updating $\Delta\tau$: Let $f = \mathrm{Vec}\Big(\tilde{\mathbf{A}}(\mathbf{x}^{k+1}) - \mathbf{Y}^{k+1} - \mathbf{B}' \circ \tau + \frac{1}{\mu}\hat{\mathbf{Z}}^k\Big)$.
    $\Delta\tau^{k+1} = (J^TJ)^{-1}J^Tf$;
  Updating $\mathbf{Z}$: $\mathbf{Z}^{k+1} = \hat{\mathbf{Z}}^k + \mu(\tilde{\mathbf{A}}(\mathbf{x}^{k+1}) - \mathbf{Y}^{k+1} - (\mathbf{B}' \circ \tau$
    $+\mathrm{Mat}(J\Delta\tau^{k+1})))$;
  Updating $\alpha$: $\alpha^{k+1} = \frac{1+\sqrt{1+4(\alpha^k)^2}}{2}$;
  Updating $\hat{\mathbf{x}}$: $\hat{\mathbf{x}}^{k+1} = \mathbf{x}^{k+1} + \frac{\alpha^k-1}{\alpha^{k+1}}(\mathbf{x}^{k+1} - \mathbf{x}^k)$;
  Updating $\Delta\hat{\tau}$: $\Delta\hat{\tau}^{k+1} = \Delta\tau^{k+1} + \frac{\alpha^k-1}{\alpha^{k+1}}(\Delta\tau^{k+1} - \Delta\tau^k)$;
  Updating $\hat{\mathbf{Z}}$: $\hat{\mathbf{Z}}^{k+1} = \mathbf{Z}^{k+1} + \frac{\alpha^k-1}{\alpha^{k+1}}(\mathbf{Z}^{k+1} - \mathbf{Z}^k)$;
End while
Output: Solution $\mathbf{Y}^{k+1}, \mathbf{x}^{k+1}, \Delta\tau^{k+1}$ to the model in (28)
```

The linearized formulation in Eq. (28) is a convex programming problem over the deformation step $\Delta\tau$, the error matrix $\mathbf{Y}$ and the coefficient vector $\mathbf{x}$. A sub-problem is the optimization of the deformation step $\Delta\tau$, which can solved via least squares estimation. Algorithm 3 presents the fast ADMM Algorithm for solving the model in (28). For more details on the derivation of the algorithm, please refer to our supplemental material, available online.

Algorithm 4 summarizes the procedure of our robust face alignment and classification method. As we can see, after solving problem (27), the optimized $\mathbf{x}^\star, \mathbf{Y}^\star, \tau^\star$ are obtained. Then, the test image $\mathbf{B}'$ is assigned to the class where the reconstruction error is the smallest.

Algorithm 4. Fast ADMM for Robust Face Alignment and Classification with NMR

```
Input: A set $\tilde{\mathbf{A}}$ of aligned training image matrices and an image matrix $\mathbf{B}' \in R^{p\times q}$, the model parameters $\lambda$, $\mu$, $\gamma$ and initial transformation $\tau_0$ of $\mathbf{B}'$.
While not converged do
  Compute an optimal step $\Delta\tau^\star$ by solving the model in (28) via Algorithm 3.
  Update $\tau \leftarrow \tau + \Delta\tau^\star$.
End while
Get the optimal solution $\mathbf{x}^\star, \mathbf{Y}^\star, \tau^\star$ to the model in (27).
Compute the class reconstruction error: $e_i(\mathbf{B}') = ||\hat{\mathbf{B}} - \hat{\mathbf{B}}_i||_*$
  $= \|\tilde{\mathrm{A}}(\mathbf{x}^\star) - \tilde{\mathrm{A}}(\delta_i(\mathbf{x}^\star))\|_*$.
Output: $identity(\mathbf{B}') = \arg\min_i e_i(\mathbf{B}')$
```

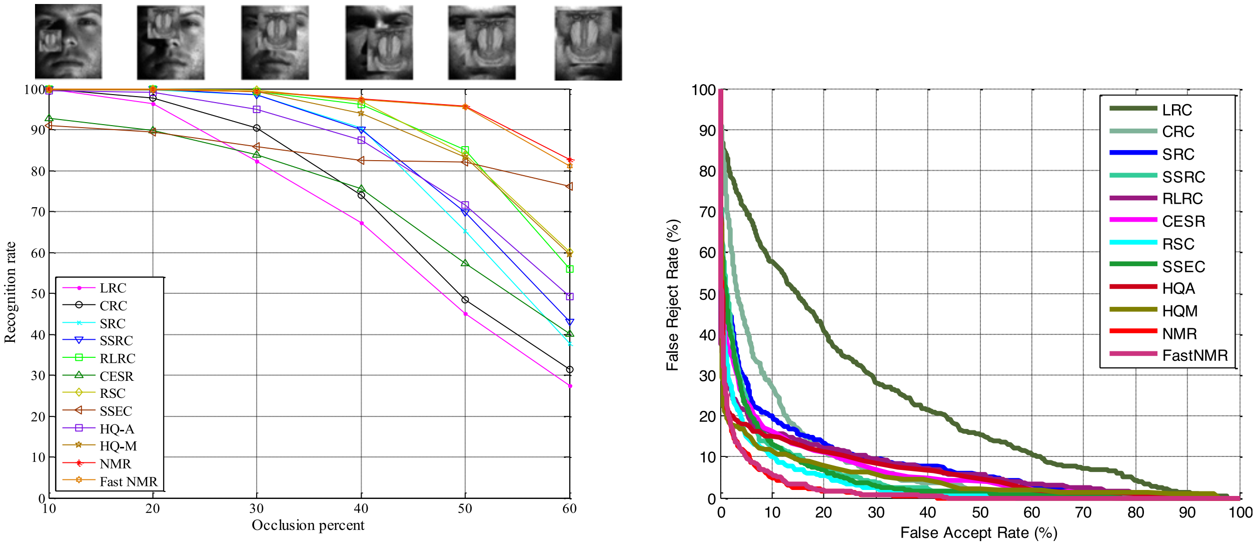
6 Experiments

Five publicly available databases, the Extended Yale B database [35], the AR database [36], the Multi-PIE database [37], the FRGC Database [38], and the EURECOM Kinect database [65] are used in our experiments. The description of these databases is given in supplemental materials, available online. The proposed methods, NMR (Algorithm 1) and Fast NMR (Algorithm 2), are tested and compared with state-of-the-art linear representation related classifiers: LRC [1], CRC [16], SRC [4], SSRC [53], RLRC [17], CESR [9], RSC [6], SSEC [12], half-quadratic with the additive form (HQ_A), half-quadratic with the multiplicative form (HQ_M) [34]. LRC, CRC and RLRC are tuned to achieve their best performance by choosing the optimal parameters, and the parameter settings of the other methods follow the authors' suggestions. The regression parameter for NMR is chosen as $\lambda = 1$ and $\gamma = 0.001$ for Fast NMR. The penalty parameter in all of our algorithms is chosen as $\mu = 1$. It should be mentioned that all experiments are done on the original face images, without any image preprocessing and feature extraction step.

6.1 Recognition and Verification with Random Occlusions

In the first experiment, we use the similar experiment setting as in [4] to test the performance of the proposed model. Subsets 1 and 2 of the Extended Yale B are used for training and Subset 3 for testing. Each test image is corrupted by a randomly located square block of a "baboon" image with varying block sizes. The block size determines the occlusion level of an image. We conduct face recognition tests first and shows the recognition rates of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR under different occlusion levels in Fig. 3a. The images on the top of Fig. 3a illustrate the occlusion levels varying from 10 to 60 percent. We then perform face verification tests and show the DET curve (a plot of false reject rate against false accept rate) in Fig. 3b.

From Fig. 3a, we can see that the proposed NMR and Fast NMR achieve very close results, and they significantly outperform other robust methods such as SRC, SSRC, RSC,



(a)

(b)

Fig. 3. (a) Recognition rates (percent) of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR under different levels of occlusion; (b) DET curves of all methods when the occlusion level is 50 percent.

RLRC, HQ_M and SSEC, when the occlusion level is equal to or larger than 50 percent. When the occlusion level is no more than 30 percent, SRC, RSC, RLRC and HQ_M achieve similar results with NMR. The performance of SSEC is good when the occlusion level becomes high, but it has no advantage when the occlusion level is relative low. The recognition rates of LRC and CRC drop fast with the increase of occlusion levels; thus the two methods are sensitive to the level of structural noise. From Fig. 3b, we can see the proposed methods still achieve the leading results among all methods in face verification task.

In the second experiment, we also use Subsets 1 and 2 for training and Subset 3 for testing, but with occlusions of different kinds of objects: cup, dollar, cartoon mask, book, flower and puzzle in test images (as shown in Fig. 4). The recognition rate of each method is shown in Fig. 5. The proposed NMR and Fast NMR achieve the best results among all methods. This experiment demonstrates that NMR is more robust than the others for face recognition with different, contiguous occlusions.

In the third experiment, for the test images in Subset 3, we impose another two special occlusions: a square black block and a square block whose elements are random numbers between 0 and 255. Fig. 6 shows the recognition rates of each method under various occlusion levels with the black block and the random block. In general, the results in Fig. 6 are consistent with those in Fig. 3. NMR and Fast NMR always achieve robust performance and outperform state-of-the-art methods in both occlusion cases. In Fig. 6a, the performance difference between NMR and RSC (or SSEC) is not as remarkable as that shown in Fig. 3 when the occlusion level is over 50 percent. The recognition rate of NMR is 57.3, 6.2, 4.0 percent higher than SRC, RSC and SSEC when the occlusion level is 60 percent. In Fig. 6b, the performance difference between NMR and RSC (or RLRC) is remarkable when the occlusion level is larger than 40 percent. NMR still achieves a recognition rate of 86.4 percent when the occlusion level is 60 percent, which is 4.1, 22.8 percent higher than SSEC and RSC.

Finally, for all of the above mentioned face recognition experiments, we conduct the corresponding face verification tests. The verification accuracy is measured in terms of the DET curve and the equal error rate (EER), i.e., the point where the false accept rate is equal to the false reject rate. The EERs of all methods are shown in Table 1 and the corresponding DET curves are shown in Fig. S-9 in supplemental materials, available online. From Table 1 and Fig. S-9, available online, we can see that the verification performances of different methods are generally consistent with their recognition performances. Our methods

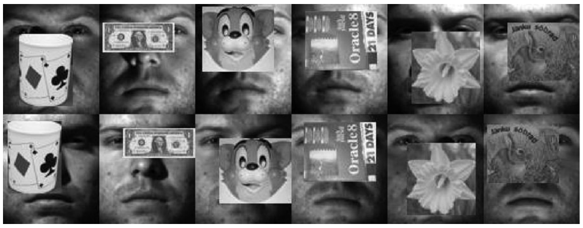


Fig. 4. Sample images of one person with occlusions of different kinds of objects.

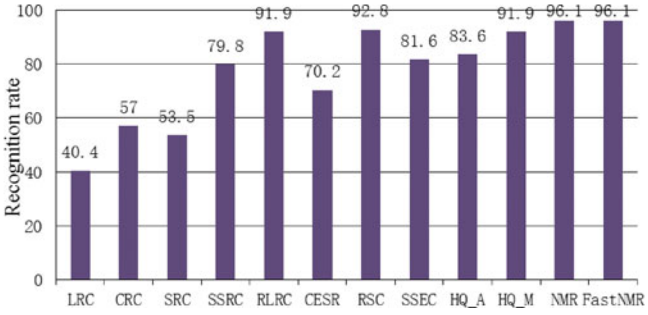
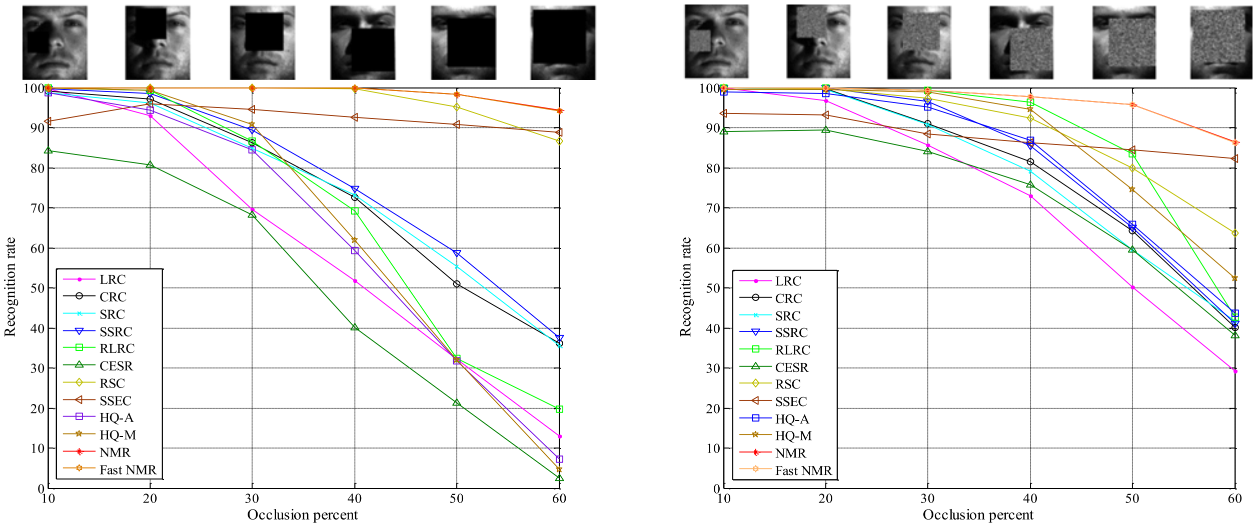


Fig. 5. Recognition rates (percent) of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, RSC, SSEC, HQ_A, HQ_M, NMR and fast NMR under different, contiguous occlusions.



(a)

(b)

Fig. 6. Recognition rates (percent) of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR under the different occlusion levels. (a) the case that test images are with the occlusion of black block; (b) the case that test images are with the occlusion of random block.

always achieve the best verification accuracy among all methods. These results further demonstrate the robustness of the proposed methods.

6.2 Recognition with Real World Occlusions

In the first experiment, we evaluate the robustness of our methods in dealing with a real disguise on the AR database. Here, we select eight frontal face images without occlusion, i.e., the first four images of Sessions 1 and 2 for training. We construct two test sets: (i) Six images with sunglasses from both sessions, and (ii) Six images with scarves from both sessions. The classification results of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR are listed in Table 2. From Table 2, we observe that NMR and Fast NMR achieve the highest recognition rate for each test set. For test images with sunglasses, where the occlusion level is relatively low, the sparseness assumption holds so SRC can achieve good results. Besides, in this case, SSRC, CESR and HQ_M achieve encouraging results. There is no significant performance difference between NMR and these methods. However, when the occlusion level becomes larger, as in the case of images with scarves, the performance advantage of NMR becomes evident.

In the second experiment, we evaluate our methods on another database with real world occlusions: the EURECOM Kinect database [65]. Here, 10 images without occlusion in Sessions 1 and 2 are used for training, and the rest 6 images with occlusion caused by sunglasses, hand, and paper are for test. The recognition rates and equal error rates of all the methods are listed in Table 3 (the corresponding DET curves are shown in Fig. S-10 in supplemental materials, available online). From Table 3, we can see clearly that the proposed NMR and Fast NMR still achieve the best results among all methods. This demonstrates that our NMR is more robust than others to occlusions caused by hands with characteristics similar to the face.

TABLE 1
Equal Error Rates of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR under Different Occlusion Cases on the Extended Yale B Database

EER (%)	“baboon” Block	Different objects	Black Block	Random Block
LRC	29.46	39.52	40.12	28.58
CRC	15.72	23.22	19.81	14.32
SRC	15.39	25.63	13.77	15.36
SSRC	11.93	21.20	13.28	8.78
RLRC	14.34	14.92	40.18	13.20
CESR	13.45	21.27	41.68	11.42
RSC	10.06	11.06	5.89	8.78
SSEC	11.73	18.79	12.24	8.66
HQA	13.56	20.42	33.82	13.88
HQM	10.74	16.47	30.24	9.40
NMR	7.31	9.10	1.72	7.81
Fast NMR	7.53	10.32	2.72	7.28

6.3 Recognition with Illumination Changes

In this section, we test the proposed method under different illumination conditions. In the first experiment we choose Subset 1 of the Extended Yale B database for training. As we know, the extreme illumination change is a challenging task for most face recognition methods. Therefore, Subsets 4 and 5 with extreme lighting conditions are used for testing. Fig. 7 shows the recognition rates of all methods tested on Subset 4 and Subset 5. For both subsets, NMR and Fast NMR achieves the best results among all methods. Some robust sparse representation methods like CESR, HQ_A, HQ_M seem not very robust to extreme illumination changes. SSEC, as a method designed exclusively for contiguous occlusion, is not suitable for extreme illumination changes either. However, the classical linear regression based method LRC and its robust version RLRC seem more insensitive to illumination changes than robust sparse representation methods.

TABLE 2
Recognition Rates (percent) of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR on the AR Database

	LRC	CRC	SRC	SSRC	RLRC	CESR	RSC	SSEC	HQ_A	HQ_M	NMR	Fast NMR
Sunglass	92.8	93.5	94.4	95.4	94.6	95.0	89.2	79.0	94.7	95.0	***96.9***	***96.9***
Scarf	30.7	63.6	57.6	66.7	53.3	33.5	66.8	49.1	48.7	50.1	***73.5***	73.3

TABLE 3
Recognition Rates (RRs) and Equal Error Rates of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR on the EURECOM Kinect Database

	LRC	CRC	SRC	SSRC	RLRC	CESR	RSC	SSEC	HQ_A	HQ_M	NMR	Fast NMR
RR	48.4	59.0	59.0	69.2	67.3	62.5	69.9	61.5	69.9	70.8	**75.0**	**75.0**
EER	30.19	16.93	16.05	15.56	27.91	21.62	14.43	33.67	17.53	15.93	13.71	**13.06**

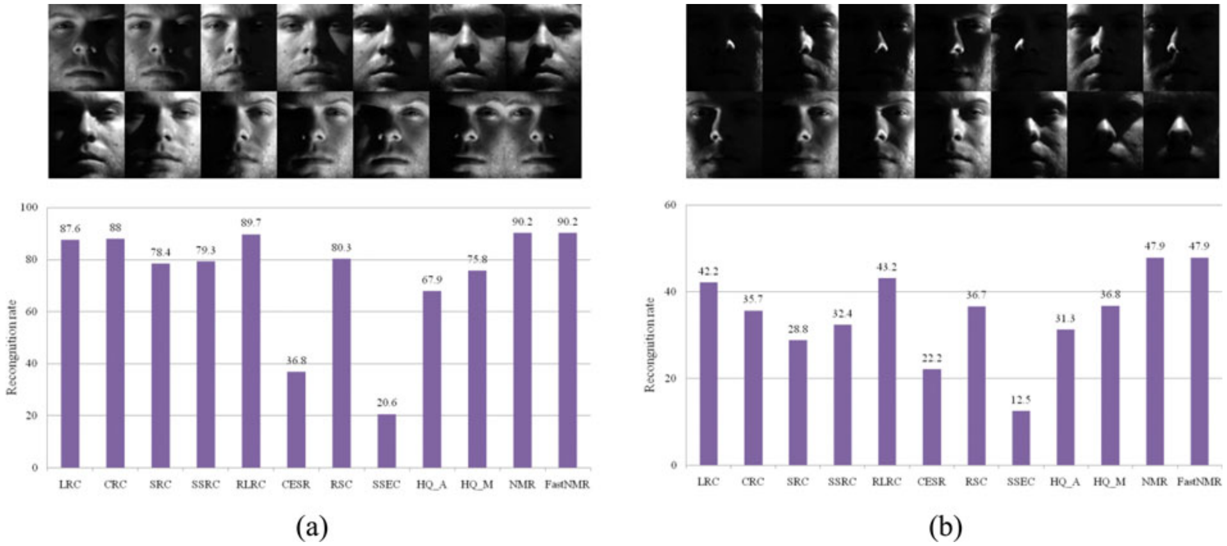
We conducted the second experiment on the Multi-PIE database. There are 249 subjects in Session 1, and 166, 160, 175 subjects in Sessions 2, 3 and 4, respectively. All subjects of Session 1, each having 8 frontal neutral images with slight illumination changes are used for training. All subjects of Sessions 2, 3 and 4, each having 10 frontal neutral images with different illumination variations, are used for testing. Table 4 lists recognition rates of all methods for the three test sets. NMR and Fast NMR always achieve the best results, but the robust sparse representation methods like SRC, SSRC, RSC and HQ_M also achieve competitive results in these tests. The performance of LRC, however, is not very good. Note that the illumination conditions of images in the Multi-PIE database are much better than those in the Extended Yale B database as used in the preceding experiment. It seems that SRC, SSRC, RSC and HQ_M are insensitive to relatively slight illumination changes.

6.4 Experiment on the FRGC database

In this section, we choose a subset of the FRGC database, which contains 220 persons and each person has 20 images. These images are taken in different conditions such as large illumination variations, low resolution of the face region, and possible blurring. We use the first 10 images per class for training, and the remaining for the tests. Here, the face region of each image is first cropped from the original high-resolution images and resized to a spatial resolution of 32×32. The classification results of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR are shown in Table 5. These results demonstrate the effectiveness of the proposed method for face recognition in the different conditions. SSEC is designed exclusively for face recognition with contiguous occlusion, but its performance is not competitive in general cases without occlusion. In contrast, some other methods like CRC, RSC and HQ_M achieve very good results in this experiment.

6.5 Comparison Analysis of Computation Time

In this section, we compare the running time of the proposed NMR with state-of-the-art methods. Our programming environment is Matlab 2011, and all algorithms are implemented on a Core Duo 2.93 GHz with 4 G RAM desktop. We conduct face recognition experiments with "Baboon" block occlusion at a 50 percent occlusion level on the Extended Yale B database. The number of training



(a)

(b)

Fig. 7. Recognition rates (percent) of each classifier under different illumination conditions on the extended yale B database. (a) on the subset 4, (b) on the subset 5.

TABLE 4
Recognition Rates (percent) of LRC, CRC, SRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M and NMR on the Multi-PIE Database under Different Illuminations

	Session 2	Session 3	Session 4
LRC	76.4	67.0	74.2
CRC	82.4	71.8	80.2
SRC	82.7	73.6	82.0
SSRC	84.9	77.5	84.0
RLRC	80.8	70.9	79.3
CESR	76.6	64.9	76.2
SSEC	66.2	53.6	59.1
RSC	82.8	75.3	81.8
HQ_A	79.5	68.6	77.7
HQ_M	82.7	74.2	83.2
NMR	***85.8***	77.9	84.5
Fast NMR	***85.8***	***78.2***	***84.6***

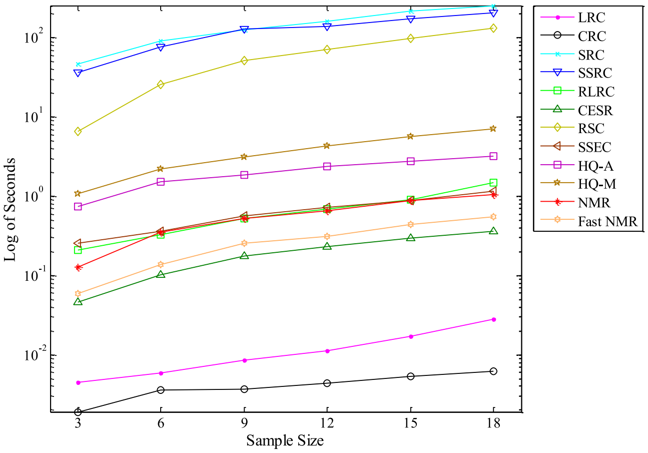


Fig. 8. Illustration of the average running time (base-10 log of seconds) of recognizing one testing sample for each method on the extended yale B database.

samples of each class varies from 3 to 18, with an interval of 3. The average running time (base-10 log of seconds) of recognizing one testing sample for each method is illustrated in Fig. 8.

From Fig. 8, we can see that LRC and CRC are the fastest methods, because they only involve a linear regression problem which has a close-form solution. But, the two methods are not very robust, particularly when there are high occlusion levels. CESR is also faster than NMR, but its recognition performance is always significantly lower than NMR. SSEC and RLRC perform as fast as NMR, but SSEC is sensitive to illumination changes while RLRC is sensitive to occlusions as demonstrated in foregoing experiments. In contrast, the proposed NMR is a more general face recognition algorithm. Compared to the halt-quadratic based sparse representation methods HQ_A and HQ_M, NMR is faster and more robust to occlusion and illumination changes. The other robust methods such as RSC, SRC and SSRC, are significantly more time-consuming than NMR. The empirical computational complexity of RSC is $O(k(n^2m^{1.3}))$, where k is the number of iterations, while that of SRC is $O(n^2(m+n)^{1.3}))$ because it needs to use an extra identity matrix in order to represent the occluded or corrupted pixels [7], [4]. NMR has a computational complexity of $O(k(m^{1.5}+mn))$, which is much lower than those of RSC and SRC. SSRC involves a structured sparsity-inducing norm based optimization problem, which seems to be computationally more expensive than a nuclear norm based one. Fast NMR further improves the convergence rate of NMR, and generally needs half the number of iterations required by NMR. Since Fast NMR consumes almost the same computation as NMR in each iteration step, Fast NMR achieves a factor of two speed improvement compared to NMR.

6.6 Experiments on Simultaneous Face Alignment and Recognition

In this section, we conduct experiments to exhibit the robustness of our method against misalignment, illumination variation, and contiguous occlusion. Here the CMU Multi-PIE database is employed; it should be noted that the test images used are all misaligned in the original resolution of 640×480. All of the 249 subjects present in Session 1 are used as training subjects. We compare with three closely related methods: robust registration and illumination via sparse representation (RASR) [10], misalignment robust representation (MRR) [51], and robust face alignment and structured sparse representation classification (RA-SSRC) [53]. In addition, with all of the experiments, we manually click outer-eye corners in all training images and crop them to the size of 60 × 45 for all methods. The distance between the two outer eye corners is normalized to be 37 pixels.

In the first experiment, we evaluate the robustness of our method to deal with various levels of contiguous occlusion. As described in [10], we choose the same frontal images of seven illuminations {0, 1, 7, 13, 14, 16, 18} with neutral facial expression from each subject as training images. Frontal images of illumination {10} from Session 1 (the same session used for training) are used as test images. A randomly located block of each face image is replaced by the image Baboon and we simulate various levels of contiguous block occlusion from 10 to 50 percent. The left part of Table 6 shows the recognition rates of the four methods varied using different occlusion levels. We can see that our method performs best in all of the cases. Moreover, as the occlusion rate gets higher, the recognition rates of the other three methods drop rapidly, while ours seems to be more stable.

The second experiment is conducted to evaluate the robustness of our method to deal with illumination

TABLE 5
Recognition Rates (percent) of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, SSEC, HQ_A, HQ_M, NMR and Fast NMR on the FRGC Database

LRC	CRC	SRC	SSRC	RLRC	CESR	RSC	SSEC	HQ_A	HQ_M	NMR	Fast NMR
77.0	92.2	89.2	77.5	77.5	81.9	92.0	70.5	84.7	91.9	***93.3***	93.2

TABLE 6
The Recognition Rates (percent) of Different Methods Versus Different Occlusion Levels and Illumination Variations on the Multi-PIE Database

method	occlusion level					illumination variation			
	10%	20%	30%	40%	50%	Session 1	Session 2	Session 3	Session 4
RASR	98.4	94.4	79.5	43.4	20.5	77.9	55.6	50.4	49.6
MRR	98.0	94.8	79.9	42.2	21.3	78.4	55.5	51.3	49.7
RA-SSRC	***98.8***	95.2	82.7	52.2	24.5	81.1	60.0	54.8	52.6
Algorithm 4	***98.8***	***96.0***	***83.5***	***71.5***	***50.2***	***87.1***	***64.8***	***61.2***	***60.2***

variations. To fully exhibit the ability of our method with respect to illumination variation, we choose seven frontal images with slight illumination variation {05, 06, 07, 08, 15, 16, 17} from each subject as training samples. Another seven frontal images with severe illumination variation {00, 01, 02, 11, 12, 13, 19} per subject from Sessions 1–4 are used as test samples. The recognition rates of the four methods are shown in the right part of Table 6. As we can see, our method achieves remarkable advantages over other methods in all cases, which demonstrates the effectiveness of our method for dealing with the task of illumination changes.

7 Conclusions and Future Work

This paper investigates using nuclear norm to characterize the occlusion and illumination variation caused error image which has a two-dimensional structure and is generally low-rank (or nearly low-rank). A nuclear norm based matrix regression model is introduced, and the augmented Lagrange multipliers method, and its accelerated version, are developed for solving the model. The proposed NMR classifier is examined on four popular face image databases: the Extended Yale B, AR, EURECOM, Multi-PIE and FRGC, and experimental results indicate that NMR is more robust than state-of-the-art regression based methods for face recognition with occlusions and illumination changes. Although Fast NMR is faster than most robust regression methods, a computationally more efficient algorithm is still required for its real-world application. In addition, the question of whether the nuclear norm based model is effective for more complex noise and the question of how to extend the model for general noise need further investigation.

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Science Fund of China under Grant Nos. 91420201, 61472187, 61502235, 61233011 and 61373063, the 973 Program No. 2014CB349303, and Program for Changjiang Scholars and Innovative Research Team in University. J. Yang is the corresponding author. Jian Yang, Lei Luo, and Jianjun Qian are co-first authors.

Reference

[1] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.

[2] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. Neural Netw.*, vol. 10, no. 2, pp. 439–443, Mar. 1999.

[3] J. T. Chien and C.-C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition,"*IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1644–1649, Dec. 2002.

[4] J. Wright, A. Y. Yang, A. Ganesh, S. 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.

[5] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Commun. Statistics: Theory Methods*, vol. A6, pp. 813–827, 1977.

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

[7] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[8] R. He, W. S. Zheng, B. G. Hu, and X. W. Kong, "A regularized correntropy framework for robust pattern recognition," *Neural Comput.*, vol. 23, pp. 2074–2100, 2011.

[9] R. He, W. S. Zheng, and B. G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[10] A. Wagner, J. Wright, A. Ganesh, Z. H. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust registration and illumination via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2011.

[11] 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.*, 2009, pp. 1050–1057.

[12] X.-X. Li, D.-Q. Dai, X.-F. Zhang, and C.-X. Ren, "Structured sparse error coding for face recognition with occlusion," *IEEE Trans. Image Process.*, vol. 22 , no. 5, pp. 1889–1999, May 2013.

[13] R. Rigamonti, M. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1545–1552.

[14] Q. Shi, A. Eriksson, A. Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 553–560.

[15] J. Yang, L. Zhang, Y. Xu, and J. Y. Yang, "Beyond sparsity: The role of L1-optimizer in pattern classification," *Pattern Recog.*, vol. 45, pp. 1104–1118, 2012.

[16] L. Zhang, M. Yang, and X. C. Feng, "Sparse representation or collaborative representation which helps face recognition?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 471–478.

[17] I. Naseem, R. Togneri, and M. Bennamoun, "Robust regression for face recognition," *Pattern Recog.*, vol. 45, pp. 104–118, 2012.

[18] M. Fazel, Matrix rank minimization with applications, PhD dissertation, Stanford Univ., Stanford, CA, 2002.

[19] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. Am. Control Conf.*, 2001, vol. 6, pp. 4734–4739.

[20] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[21] B. Recht, W. Xu, and B. Hassibi, "Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization," in *Proc. 47th IEEE Conf. Decision Control*, 2008, pp. 3065–3070

[22] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[23] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

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

[25] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Neural Inform. Process. Syst.*, Dec. 2009.

[26] R. He, Z. N. Sun, T. N. Tan, and W. S. Zheng, "Recovery of corrupted low-rank matrices via half-quadratic based nonconvex minimization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 2889–2896.

[27] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1235–1256, 2009.

[28] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," arXiv:1009.5055v3, 2013.

[29] A. Hansson, Z. Liu, and L. Vandenberghe, "Subspace system identification via weighted nuclear norm optimization," arXiv:1207.0023, 2012.

[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, p. 1122, 2011.

[31] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[32] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximations," *Comput. Math. Appl.*, vol. 2, pp. 17–40, 1976.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, p. 1122, 2011.

[34] R. He, W.-S. Zheng, T. Tan, and Z. Sun, "Half-quadratic based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 261–275, Feb. 2014.

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

[36] A. M. Martinez and R. Benavente, "The AR face database," Computer Vision Center, Barcelona, Spain, Tech. Rep. #24, Jun. 1998.

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

[38] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, vol. 1, pp. 947–954.

[39] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[40] X. Yuan and J. F. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *Pacific Journal of Optimization*, vol. 9, no. 1, pp. 167–180, 2013.

[41] S. Kontogiorgis and R. Meyer, "A variable-penalty alternating direction method for convex optimization," *Math. Program.*, vol. 83, pp. 29–53, 1989.

[42] P. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numerical Anal.*, vol. 16, no. 6, pp. 964–979, 1979.

[43] E. Esser, "Applications of Lagrangian-based alternating direction methods and connections to split Bregman," UCLA CAM, Los Angeles, CA, Tech. Rep. 09-31, 2009.

[44] B. He and H. Yang, "Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities," *Operations Res. Lett.*, vol. 23, pp. 151–161, 1998.

[45] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," UCLA CAM, Los Angeles, CA, Tech. Rep. 12-52, pp. 12–52, 2012.

[46] T. Goldstein, B. O'Donoghue, and S. Setzer, R. Baraniuk, "Fast alternating direction optimization methods," *SIAM J. Imag. Sci.*, vol. 7, no. 3, pp. 1588–1623, 2014.

[47] B. He and X. Yuan, "On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers," *Numerische Mathematik*, vol. 130, no. 3, pp. 567–577, 2015.

[48] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o(1 = k∧2)," *Soviet Math. Dokl.*, vol. 27, pp. 372–376, 1983.

[49] Y. Nesterov and A. Nemirovski, "On first-order algorithms for l1/nuclear norm minimization," *Acta Numerica*, vol. 22, pp. 509–575, 2013.

[50] D. Goldfarb, S. Ma, and K. Scheinberg, "Fast alternating linearization methods for minimizing the sum of two convex functions, *Math. Programming*, vol. 141, no. 1, pp. 349–382, 2013.

[51] M. Yang, L. Zhang, D. Zhang, "Efficient misalignment robust representation for real-time face recognition," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, vol. 7572, pp. 850–863.

[52] A. Peng Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 763–770.

[53] K. Jia, T. Chan, and Y. Ma, "Robust and practical face recognition via structured sparsity," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 331–344.

[54] J. Huang, Z. Zhang, and D. Metaxas, "Learning with structured sparsity," in *Proc. Int. Conf. Mach. Learn.*, 2009.

[55] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, 2011.

[56] R. Jenatton, G. Obozinski, and F. Bach. "Structured sparse principal component analysis," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2010.

[57] J. Mairal, R. Jenatton, G. Obozinski, F. Bach, "Learning hierarchical and topographic dictionaries with structured sparsity," SPIE Wavelets and Sparsity XIV, San Diego, United States. SPIE, 8138, Aug. 2011.

[58] Q. Zhou, L. Zhang, L. Ma, "Learning topographic sparse coding through similarity function," in *Proc. 4th Int. Conf. Nat. Comput.*, 2008, pp. 241–245.

[59] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1605–1612.

[60] M. Savvides, B. V. K. Vijay Kumar, and P. K. Khosla, "Corefaces—robust shift invariant PCA-based correlation filter for illumination tolerant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, pp. II-834–II-841.

[61] W. Chen, M. Joo, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Trans. Syst., Man Cybern.*, vol. 36, no. 2, pp. 458–464, Apr. 2006.

[62] J. Yang and C. Liu, "Horizontal and vertical 2DPCA-based discriminant analysis for face verification on a large-scale database," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 4, pp. 781–792, Dec. 2007.

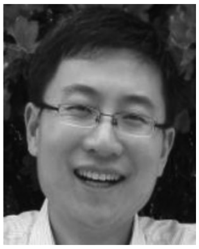
[63] J. Yang, J. Qian, L. Luo, F. Zhang, and Y. Gao, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," arXiv:1405.1207, 2014.

[64] L. Luo, J. Yang, and J. Qian, "Nuclear norm regularized sparse coding," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 1834–1839.

[65] R. Min, N. Kose, J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.

Jian Yang received the BS degree in mathematics from the Xuzhou Normal University in 1995, the MS degree in applied mathematics from the Changsha Railway University in 1998, and the PhD degree from the Nanjing University of Science and Technology (NUST), Nanjing, China, on the subject of pattern recognition and intelligence systems in 2002. In 2003 to 2007, he was a postdoctoral fellow at the University of Zaragoza, Hong Kong Polytechnic University and New Jersey Institute of Technology, respectively. He is currently a professor in the School of Computer Science and Technology of NUST. He is the author of more than 100 scientific papers in pattern recognition and computer vision. His journal papers have been cited more than 4,000 times in the ISI Web of Science, and 8,000 times in the Web of Scholar Google. His research interests include pattern recognition, computer vision, and machine learning. He is currently an associate editor of Pattern Recognition Letters and IEEE Transactions Neural Networks and Learning Systems, respectively.

Lei Luo received the BS degree from Xinyang Normal University, Xinyang, China, in 2008, the MS degree from Nanchang University, Nanchang, China, in 2011. He is currently working towards the PhD degree in pattern recognition and intelligence system from the school of computer science and engineering, Nanjing University of Science and Technology, Nanjing, China. His current research interests include pattern recognition and optimization algorithm.



Jianjun Qian received the BS and MS degrees in 2007 and 2010, respectively, and the PhD degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), in 2014. Now, he is an assistant professor in the school of computer science and engineering of NUST. His research interests include pattern recognition, computer vision, and face recognition in particular.

Ying Tai received the BS degree in the school of computer science and engineering from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2012. He is currentlyworking towards the PhD degree from NUST. His current research interests include pattern recognition, computer vision, and especially face recognition.

Fanlong Zhang received the BS and MS degrees in 2007 and 2010, respectively. He is currently working towards the PhD degree with the school of computer science and engineering, Nanjing University of Science and Technology (NUST), Nanjing, China. His current research interests include pattern recognition and optimization.

Yong Xu received the BS and MS degrees from the Air Force Institute of Meteorology, China, in 1994 and 1997, respectively, and the PhD degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, in 2005. He is currently with the Shenzhen Graduate School, Harbin Institute of Technology. His current interests include pattern recognition, biometrics, machine learning, video analysis and bioinformatics. He has authored over 100 scientific papers.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.