Robust matrix regression for illumination and occlusion tolerant face recognition

Jianchun Xie, Jian Yang, Jianjun Qian, Ying Tai
School of Computer Science and Engineeting, Nanjing University of Science and Technology
{jcxie,csjyang,csjqian}@njust.edu.cn, tyshiwo@gmail.com

Abstract

Face recognition (FR) via regression analysis based classification has been widely applied in the past several years. In the existing regression methods, the testing image is represented as a linear combination of the training samples and the error image is converted into vector which is characterized by l_1-norm or l_2-norm. Therefore the two-dimensional structure of the error image is neglected in practice. In this paper, we operate on the two-dimensional image matrix directly, and propose a new face recognition method, namely Robust Matrix Regression (RMR). We perform the minimal weighted nuclear norm constraint on the representation error image as criterion to make full use of the low rank structural information. The proposed model is efficiently solved by an alternating direction method of multipliers (ADMM) and experimental results on public face databases demonstrate the effectiveness of our model in dealing with variations of occlusion and illumination.

1. Introduction

Face recognition (FR) is one of the most visible and challenging problems in computer vision and has been received significant attention [22]. In the past two decades, numerous face classification methods have been developed by many scholars around the world. Among these methods, linear regression analysis based methods have achieved promising results.

Naseem *et al*. presented the linear regression classifier (LRC)[15] and the robust linear regression classifier for face classification (RLRC)[16]. Wright *et al*. presented a sparse representation based classification (SRC) method [17], which codes a query image as a sparse linear combination of the training samples via the l_1-norm minimization. It's interesting that SRC gives remarkable performance when dealing with FR with block occlusion and random pixel corruption.

Since SRC has got the exciting breakthroughs in visual identification and classification tasks, many researchers began to investigate the role of the sparse coding for image classification. Yang *et al*. found that the l_1-constraint is more meaningful than l_0-constraint for classification, and provided some theoretical supports for SRC [19]. Besides, Zhang *et al*. proposed the l_2-norm regularization on the representation coefficients, which is named as collaborative representation classifier (CRC) [21]. CRC achieves similar results as SRC. However, CRC is not a robust method for FR. In order to deal with the practical noises, Yang *et al*. proposed a robust sparse coding (RSC) based on robust linear regression model and fit the character of the noises with an adaptive distribution. RSC is robust to various kinds of outliers in FR [20]. Recently, He *et al*. presented the correntropy based sparse representation (CESR) algorithm for FR [6, 7], which enables one to perform both error correction and error detection. Motivated by recent advances in 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 [10].

However, these former regression based methods usually stretch the image matrix into a vector. That is to say, the error image is characterized pixel by pixel, which ignores the structural information in the original data. These methods actually assume that the pixels in the error image are independent identically distributed (i.i.d.). However, this assumption dose not hold in real scenarios (*e.g*. occlusion and illumination). Compared with the previous methods, Yang *et al*. proposed nuclear norm based matrix regression (NMR) [18], which employed nuclear norm constraint as a criterion to make full use of the low rank structural information. They reveal that the nuclear norm constraint is better than l_1-norm or l_2-norm constraint when the images are with structural noises. More recently, Chen *et al*. believed that the error image followed matrix elliptically contoured distribution and employed l_1-norm constraint on the representation coefficients [2].

Some recent work pointed out that the visual data had low-rank structure [12]. However, the rank minimization problem is NP hard in general. To address this problem, Fazel *et al*. applied the nuclear norm heuristic to solve the rank minimization problem [3]. Gu *et al*. introduced

weighted nuclear norm minimization in image denoising area [5].

To make full use of the low-rank structural information of the error image, Our method no longer convert the image matrix into a vector in advance, but search for the representation coefficients based on the image matrix directly. The proposed model is named as robust matrix regression (RMR), which is efficiently solved by an alternating direction method of multipliers (ADMM). Experiments performed on the benchmark face databases show that RMR achieves much better performance than existing regression based methods for FR. In next section, we briefly review NMR [18] and then point out the pros and cons of the nuclear norm as the error description.

1.1. Nuclear Norm based Matrix Regression

Suppose an image data $\mathbf{Y} \in \mathbf{R}^{m\times n}$ to be represented linearly using a set of k image matrices $\mathbf{A}_1, \cdots, \mathbf{A}_k \in \mathbf{R}^{m\times n}$ *i.e.*,

$$\mathbf{Y} = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 \cdots + x_k\mathbf{A}_k + \mathbf{E}, \tag{1}$$

where $x_1, x_2, \cdots, x_k$ is a set of representation coefficients. For convenience, we denote the following linear mapping $\mathbf{A}(\mathbf{x}) = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 \cdots + x_k\mathbf{A}_k$ and then represent the error matrix as: $\mathbf{E} = \mathbf{Y} - \mathbf{A}(\mathbf{x})$.

NMR uses l_2-norm to constraint the representation coefficients, the objective function is formulated as the following minimization problem

$$\min_{\mathbf{x}} ||\mathbf{Y} - \mathbf{A}(\mathbf{x})||_* + \frac{\lambda}{2}||\mathbf{x}||_2^2, \tag{2}$$

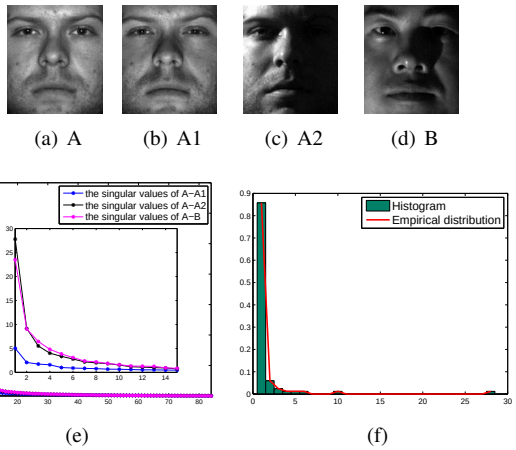
where $||\mathbf{E}||_* = \sum_{i=1}^{q} d_i$ is the nuclear norm of matrix $\mathbf{E}$, $q = min(m, n)$, d_i is the i^{th} biggest singular values of $\mathbf{E}$. Eq. (2) is named as nuclear norm based matrix regression.

1.2. Pros and Cons of Nuclear Norm based Error Description

NMR achieves many merits in FR. First, NMR is more robust than vector regression based models and achieve satisfying results when handling the general FR. Second, NMR is more powerful than the structured sparse error coding models.

As we know, the nuclear norm is the sum of the magnitudes of the singular values, as a result, the larger singular values will have more heavily influence on the nuclear norm than the small ones. In FR, when the testing and training images are clean and without extreme illumination, the nuclear norm minimization can get perfect classification results. However, when the images are under the extreme lighting condition or corruption, the classification result based on nuclear norm is unsatisfactory.

Let's take an example to illustrate this scenario. Figs.1(a) and (b) show clean face images, and (a), (b), (c) come from the same class. Figs.1 (c) and (d) show the face images with extreme illumination condition from different classes. From Fig.1 (e) (including the local enlarge figure) and Fig.1(f), we find that the singular values of error images have two major characters: (1) the distribution of the singular values of error image (A-A1) is more stable and the singular values are much smaller than others; (2) the nuclear norm of A-A2, and A-B is 76.57 and 76.45, which is influenced by extreme illumination seriously. According to the minimization of the nuclear norm, the image B will be miss-classified to A, this case is led by some bigger singular values.



(a) A (b) A1 (c) A2 (d) B

(e) (f)

Figure 1. (a) and (b) are clean face image, (c) and (d) with extreme illumination changes. (e) The singular values of different error image matrices. (f) The empirical distribution curve and histogram of the singular values of matrix (A-A2).

2. Proposed method

Motivated by the idea of robust regression, we define a robust nuclear norm of the error matrix, and adopt Majorization-Minimization (MM) algorithm to develop a reweighted nuclear norm minimization algorithm.

2.1. Robust Nuclear Norm: An non-convex function

As we know, the rank of a matrix equals the number of nonzero singular values of a matrix, we can relate the matrix rank to the l_0-norm of a vector [13]. This relation encourages us to utilize a non-convex function for dealing with the rank minimization problems of the error matrix.

Motivated by the recent developments of non-convex penalties in sparsity model, we propose a non-convex optimization model for handing the minimization of the low rank error matrix. The robust nuclear norm is defined by

$$F_\theta(\mathbf{d}) = \sum_{i=1}^{q} \rho_\theta(d_i). \tag{3}$$

d_i is the i^{th} biggest singular values of the error matrix $\mathbf{E}$, the function ρ_θ gives the contribution of each singular value d_i to the $F_\theta(\mathbf{d})$. Based on the characteristic of $\mathbf{d}$, a reasonable ρ_θ should be a nondecreasing function, *i.e.* $\rho_\theta(d_i) \leq \rho_\theta(d_{i'})$ for $d_i \leq d_{i'}$, and the first order derivative exist. For example, we can define $\rho_\theta(d_i) = (d_i + \gamma)^\theta$, where $0 < \theta \leq 1$ and γ is small positive value; or $\rho_\theta(d_i) = \log(d_i + \gamma)$.

It was shown by Chartrand [1] that a nonconvex function ρ_θ could produce exact result with fewer measurements. Recently, there has been an explosion of research on this topic in matrix norm case.

In this paper, we let $\rho_\theta(d_i)$ be a non-convex function. Since the function ρ_θ satisfies the properties of nondecreasing and non-convex, the first-order Taylor expansion of $F_\theta(\mathbf{d})$ in the neighborhood of $\mathbf{d}^{(t)}$ satisfies

$$F_\theta(\mathbf{d}) \leq F_\theta(\mathbf{d}^{(t)}) + F_\theta^{'}(\mathbf{d}^{(t)})^T(\mathbf{d} - \mathbf{d}^{(t)}), \tag{4}$$

where $F_\theta^{'}(\mathbf{d}^{(t)}) = [\rho_\theta^{'}(d_1^{(t)}), \rho_\theta^{'}(d_2^{(t)}), \cdots, \rho_\theta^{'}(d_q^{(t)})]^T$ is the derivative of $F_\theta(\mathbf{d})$ respect to $\mathbf{d}$. We employ a bound optimization strategy that is a special instance of the Majorization-Minimization (MM) algorithms [9], which alternates between the construction of a convex upper bound of the objective function and the minimization of that bound.

Let $J_\theta(\mathbf{d}|\mathbf{d}^{(t)}) = F_\theta(\mathbf{d}^{(t)}) + F_\theta^{'}(\mathbf{d}^{(t)})^T(\mathbf{d} - \mathbf{d}^{(t)})$. For any $\mathbf{d}$ and $\mathbf{d}^{(t)}$, MM algorithm proceeds by repeating two steps: (1) construct a convex upper bound for non-convex prior terms; (2) minimize the upper bound until the algorithm converges. Particularly, it computes a sequence of successive iterations as follows: given a singular value vector $\mathbf{d}^{(t)}$ at the t^{th} iteration, the algorithm computes $\mathbf{d}^{(t+1)}$ in the sequence as: $\mathbf{d}^{(t+1)} = \arg\min_{\mathbf{d}} J_\theta(\mathbf{d}|\mathbf{d}^{(t)})$, which guarantees that the current iteration is tight at the previous iteration minimizer.

It can be proved that the objective function monotonically decreases, since

$$F_\theta(\mathbf{d}^{(t+1)}) \leq J_\theta(\mathbf{d}^{(t+1)}|\mathbf{d}^{(t)}) \leq J_\theta(\mathbf{d}^{(t)}|\mathbf{d}^{(t)}) = F_\theta(\mathbf{d}^{(t)}), \tag{5}$$

Therefore, we omit constant terms from the upper bounding function that do not effect the optimization. The solution of $\min F(\mathbf{d})$ equates to $\min F^{'}(\mathbf{d}^{(t)})^T\mathbf{d} = \min \sum_{i=1}^{m_2} \rho^{'}(d_i^{(t)})d_i$. We can view $\rho^{'}(d_i^{(t)})$ as the t-th iteration weight of d_i, so

$$\min F^{'}(d^{(t)})^T d = \min \sum_{i=1}^{q} w_i^{(t)} d_i \triangleq \|\mathbf{E}\|_{w*}, \tag{6}$$

we call the reweighted form of Eq. (6) as Robust Matrix Regression (RMR) model.

2.2. Reweighting Scheme

In this section, we discuss the weighted function $\mathbf{w}^{(t)} = (w_1^{(t)}, \cdots, w_q^{(t)})^T$, which is assigned to the singular values of the error image $\mathbf{E}$.

Low rank structure of the error image is an important assumption in FR. However, the low rank matrix minimization problem is non-convex and intractable, hence approximation is required. Function $\rho_\theta = \sum_{i=1}^{q}(d_i + \gamma)^\theta \to rank(\mathbf{E})$ as $\theta \to 0$ and $\lim_{\theta \to 0} \frac{(d_i+\gamma)^\theta}{\theta} = \log(d_i + \gamma)$, which means that the function $\log(d_i + \gamma)$ is a smooth surrogate for $rank(\mathbf{E})$ [4]. In our paper, we choose $\rho_\theta = \log(d_i + \gamma)$, the low rank matrix minimization problem can be approximated as the minimization the robust nuclear norm

$$\min \sum_{i=1}^{q} \log(d_i + \gamma), \tag{7}$$

where γ is set to be a small positive value to provide stability for the log function. The first-order Taylor expansion of $\log(d_i + \gamma)$ in the neighborhood of $d_i^{(t)}$ satisfies

$$\log(d_i + \gamma) \leq \log(d_i^{(t)} + \gamma) + \frac{d_i + \gamma}{d_i^{(t)} + \gamma}, \tag{8}$$

we employ MM algorithms [9] to optimize Eq. (7)

$$\min \sum_{i=1}^{q} (d_i^{(t)} + \gamma)^{-1} d_i, \tag{9}$$

if the initial weight is chosen as $\mathbf{w}^{(0)} = (1, \cdots, 1)^T$, the first iteration will minimize the nuclear norm of matrix $\mathbf{E}$, the other iterative procedure shows that in each step, a weighted nuclear norm is minimized.

For the sake of simplicity and computational efficiency, l_2-norm is chosen to be constraint on coefficients.

3. Optimization of RMR

In this section, we use the alternating direction method of multipliers (ADMM) to solve the optimization problem (9). ADMM was presented in [12], which has been studied extensively in the theoretical frameworks of Lagrangian functions. First, the objective function of RMR is formulated as

$$\min_{\mathbf{x}} \|\mathbf{E}\|_{w*} + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 \quad s.t. \quad \mathbf{E} = \mathbf{Y} - \mathbf{A}(\mathbf{x}), \tag{10}$$

where the variable $\mathbf{E}$ and $\mathbf{x}$ are separable in the objective function. The augmented Lagrangian function L_μ is defined by

$$\begin{aligned} L_\mu = \|\mathbf{E}\|_{w*} + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \mathrm{Tr}(\mathbf{Z}^T(\mathbf{A}(\mathbf{x}) + \mathbf{E} - \mathbf{Y})) \\ + \frac{\mu}{2}\|\mathbf{A}(\mathbf{x}) + \mathbf{E} - \mathbf{Y}\|_F^2, \end{aligned} \tag{11}$$

where $\mu > 0$ is a penalty parameter and $\mathbf{Z}$ is the Lagrangian multipliers. ADMM utilizes the separability of Eq. (11) and substitutes the joint minimization over $\mathbf{E}$ and $\mathbf{x}$ with two sub-problems. We minimize the Lagrangian function with respect to each variable $\mathbf{E}$ and $\mathbf{x}$ one at a time while fixing the others, *i.e.*, the update of the variables goes as follows:

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}} L_\mu(\mathbf{E}^{(t)}, \mathbf{x}, \mathbf{Z}^{(t)}); \tag{12}$$

$$\mathbf{E}^{(t+1)} = \arg\min_{\mathbf{E}} L_\mu(\mathbf{E}, \mathbf{x}^{(t+1)}, \mathbf{Z}^{(t)}); \tag{13}$$

$$\mathbf{Z}^{(t+1)} = \mathbf{Z}^{(t)} + \mu(\mathbf{A}(\mathbf{x}^{(t+1)}) + \mathbf{E}^{(t+1)} - \mathbf{Y}). \tag{14}$$

Noting that Eq. (12) and Eq. (13) are the critical optimization problems and Eq. (14) is a proximal minimization step of the Lagrange multipliers $\mathbf{Z}$.

Solving of x: Fixing $\mathbf{E}^{(t)}$ and $\mathbf{Z}^{(t)}$, the objective function of Eq. (12) is reformulated as

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}}(\frac{\mu}{2}\|\mathbf{A}(\mathbf{x}) - (\mathbf{Y} - \mathbf{E}^{(t)} - \frac{1}{\mu}\mathbf{Z}^{(t)})\|_F^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2) \tag{15}$$

$\mathbf{A}(\mathbf{x})$ can be written as $\mathbf{Hx}$ in the context of Frobenius norm, where $\mathbf{H} = [vec(\mathbf{A}_1), \cdots, vec(\mathbf{A}_k)]$ and $vec(\cdot)$ denotes the vectorization of a matrix into a vector by stacking all its columns below each other. Letting $\mathbf{b}^{(t)} = vec(\mathbf{Y} - \mathbf{E}^{(t)} - \frac{1}{\mu}\mathbf{Z}^{(t)})$, thus the solution of Eq. (15) is

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}}(\|\mathbf{Hx} - \mathbf{b}^{(t)}\|_F^2 + \frac{\lambda}{\mu}\|\mathbf{x}\|_2^2), \tag{16}$$

this problem is a standard ridge regression model, we can get its closed-form solution as

$$\mathbf{x}^{(t+1)} = (\mathbf{H}^T\mathbf{H} + \frac{\lambda}{\mu}\mathbf{I})^{-1}\mathbf{H}^T\mathbf{b}^{(t)}. \tag{17}$$

Solving of E: First, we computer weight $\mathbf{w}$. Using the singular value decomposition $\mathbf{E}^{(t)} = \mathbf{U}^{(t)}\mathbf{D}^{(t)}\mathbf{V}^{(t)T}$, we can get $d(\mathbf{E}^{(t)}) = (d_1^{(t)}, \cdots, d_q^{(t)})$. Update $w_i^{(t)} = (d_i^{(t)} + \gamma)^{-1}$, the weight values are satisfied with $w_1^{(t)} \leq w_2^{(t)} \leq \cdots \leq w_q^{(t)}$.

Second, fixing $\mathbf{x}^{(t+1)}$ and $\mathbf{Z}^{(t)}$, the Eq. (13) is expressed as:

$$\begin{aligned}\mathbf{E}^{(t+1)} = \arg\min_{\mathbf{E}}[\frac{1}{2}\|\mathbf{E} - (\mathbf{A}(\mathbf{x}^{(t+1)}) - \mathbf{Y} \\ + \frac{1}{\mu}\mathbf{Z}^{(t)})\|_F^2 + \frac{1}{\mu}\|\mathbf{E}\|_{w^{(t)}*}].\end{aligned} \tag{18}$$

The above optimal problem is normally solved via the singular value thresholding (SVT) operator with weighted vector $\mathbf{w} = (w_1, \cdots, w_q)^T$ [5].

Theorem 1 *[5] For each $0 \leq w_1 \leq \cdots \leq w_q$ and $\boldsymbol{Y} \in R^{m\times n}$, the weighted nuclear norm minimization problem $\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 + \tau\|\boldsymbol{X}\|_{w*}$ has the solution*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}_{w,\tau}(\boldsymbol{D})\boldsymbol{V}^T, \tag{19}$$

Algorithm 1 RMR Algorithm via ADMM

Input: an image matrix $\mathbf{Y} \in R^{m\times n}$, a set of training images $\mathbf{A}_1, \cdots, \mathbf{A}_k$ and convert to matrix $\mathbf{H}$. The parameters λ, μ, γ and the termination condition parameters $\varepsilon_1, \varepsilon_2 > 0$. The iteration counter $t = 0$ and the maximum number of iterations $t_{\max}$.

Initialize: $\mathbf{Z} = \mathbf{E} = \mathbf{0} \in R^{m\times n}$, $\mathbf{w} = \mathbf{1} \in R^q$.

While not converged **do**

1. Update $\mathbf{x}^{(t+1)}$ by Eq. (17).

2. Update weight $w_i^{(t)} : w_i^{(t)} = (d_i^{(t)} + \gamma)^{-1}$.

3. Update $\mathbf{E}^{(t+1)}$ with SVT operator by Eqs.(19, 20),

$$\mathbf{E}^{(t+1)} = \mathbf{U}\mathbf{S}_{w^{(t)},\frac{1}{\mu}}(\mathbf{A}(\mathbf{x}^{(t+1)}) - \mathbf{Y} + \tfrac{1}{\mu}\mathbf{Z}^{(t)})\mathbf{V}^T.$$

4. Update $\mathbf{Z}^{(t+1)}$ by Eq. (14).

Terminate on convergence or when t attains a specified maximum number of iterations t_{max}, output solution, otherwise $t = t + 1$, and go to step 1.

Output: Optimal regression coefficient vector $\mathbf{x}^*$

where $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ is the SVD of $\boldsymbol{Y}$, and $\boldsymbol{S}_{w,\tau}(\boldsymbol{D})$ is the SVT operator with weighted vector $\boldsymbol{w}$,

$$\boldsymbol{S}_{w,\tau}(\boldsymbol{D}) = diag(max(d_i - w_i\tau, 0)). \tag{20}$$

The detailed algorithm for solving RMR is summarized in Algorithm 1. The termination conditions are given as follows with the proper termination parameters ε_1 and ε_2.

$$\|\mathbf{A}(\mathbf{X}^{(t+1)}) + \mathbf{E}^{(t+1)} - \mathbf{Y}\|_F^2 / \|\mathbf{Y}\|_F^2 < \varepsilon_1, \tag{21}$$

$$\max(\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|_2^2, \|\mathbf{E}^{(t+1)} - \mathbf{E}^{(t)}\|_F^2) / \|\mathbf{Y}\|_F^2 < \varepsilon_2. \tag{22}$$

We now discuss the computational complexity of the proposed Algorithm 1. Given the training sample size k and the image size $m \times n$, the computational complexity of matrix multiplications in step 1 is $O(k{\cdot}mn)$; the computational complexity of the SVD in step 2 and step 3 is $O(mn^2)$ (assuming $n \leq m$). Therefore, the total computational complexity of the optimal algorithm is $O(k \cdot mn + 2mn^2)$ for one time iteration.

4. RMR Classifier

Similar to the strategy of SRC, we use the training samples of all classes to form the set of regressors. Given sufficient training samples $\mathbf{A}_1, \cdots, \mathbf{A}_k$ from different classes, a new testing sample $\mathbf{Y} \in R^{m\times n}$ can be approximated by the linear span $\mathbf{Y} = x_1\mathbf{A}_1 + \cdots + x_k\mathbf{A}_k$, where $\mathbf{x}$ is the repre-

sentation coefficient by solving the following RMR model

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{Y} - \mathbf{A}(\mathbf{x})\|_{w*} + \frac{\lambda}{2}\|\mathbf{x}\|_2^2. \quad (23)$$

Based on the optimal solution $\mathbf{x}^*$, we get the reconstructed image of $\mathbf{Y}$ as $\hat{\mathbf{Y}} = \mathbf{A}(\mathbf{x}^*)$ and the error image $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$. Let $\delta_i : R^k \to R^k$ be the characteristic function that selects the coefficients associated with the i^{th} class. For $\mathbf{x} \in R^k$, $\delta_i(\mathbf{x})$ 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 i^{th} class, one can get the reconstruction of $\mathbf{Y}$ in class i as $\hat{\mathbf{Y}}_i = \mathbf{A}(\delta_i(\mathbf{x}^*))$. The corresponding class reconstruction error is defined by

$$e_i(\mathbf{Y}) = \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_i\|_{w^{final}*} = \|\mathbf{A}(\mathbf{x}^*) - \mathbf{A}(\delta_i(\mathbf{x}^*))\|_{w^{final}*}, \quad (24)$$

where $\|\mathbf{E}\|_{w^{final}*} = \sum_{i=1}^{q} w_i^{final} d_i$, w^{final} is the final weighted vector. The decision rule is

$$\text{identity}(\mathbf{Y}) = \arg\min_i e_i(\mathbf{Y}). \quad (25)$$

5. Experiments

In this section, the proposed method RMR is evaluated and compared with LRC [15], CRC [21], SRC [17], SSRC [10], RLRC [16], CESR [6], RSC [20], half-quadratic with the additive form (HQ_A), half-quadratic with the multiplicative form (HQ_M) [8] and NMR [18]. The parameter settings of all the methods follow the author's suggestions and all experiments are done on the original face images. We choose parameters $\lambda \in [0.01, 200]$, $\gamma = 10^{-6}$, $\mu = 1$.

5.1. Experiments on the Extended Yale B Database

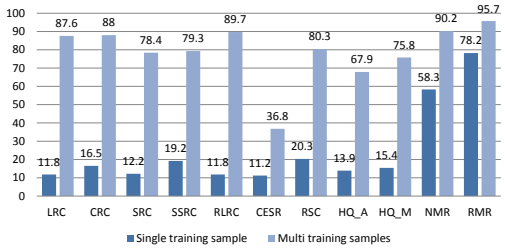
The Extended Yale B database contains 38 classes [11] and each class contains 64 images taken under different illumination conditions. The 64 images of a subject are required at camera frame rate of 30 frames/second, so there is only small change in head pose and facial expression for those 64 images. Each image is resized to 96×84 pixels in our experiment.

We design four different experiments. In the first two experiments, we set two training modes: the "single training sample" protocol and the "multi training samples" protocol. It should also be noted that for the "single training sample" protocol, we choose the first image of each person from Subset 1 for training.

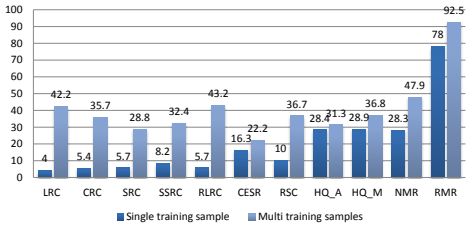
In the first experiment, Subset 1 is chosen for training, Subset 4 and Subset 5 as shown in Fig.2 with extreme lighting conditions are used for testing. The experimental results of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, HQ_A, HQ_M, NMR and RMR are shown in Fig. 3.

From Fig.3, we find that RMR achieves the best results than the other methods no matter the training sample is single or not, Even the testing images are from Subset 4 or

Figure 2. Sample images with different illumination conditions. Top: Subset 4, Bottom: Subset 5



(a)



(b)

Figure 3. Recognition rates (%) of each classifier under different illumination conditions on Extended Yale B database. (a) on the Subset 4, (b) on the Subset 5.

Subset 5, and Especially, from Subset 5, the recognition rate of our method is 92.52% which achieves 200% improvement over NMR. Some robust sparse representation methods work well for pixel-level noise, but they seem to be very sensitive to the extreme illumination changes.

In the second experiment, we investigate the robustness of the proposed method to different level of contiguous noise. We use the similar experiment setting as in [17], and Subset 3 for testing, but with different kinds of occlusions: cup, dollar, cartoon mask, book, flower and puzzle in testing images (as show in Fig. 4). We compare our method with several related methods. The recognition results of each method are displayed in Fig. 5

The proposed RMR achieves the best result among all methods. This experiment demonstrates that RMR is more robust than the other methods for FR with different contiguous occlusions even when the training set is not adequate.

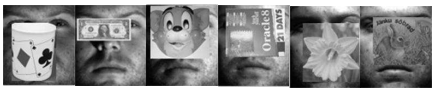


Figure 4. Sample images of one person with different occlusion.

Table 1. Recognition rates (%) of each classifier under different levels of occlusion

Occlusion	10%	20%	30%	40%	50%	60%
LRC	100	96.3	82.2	67.3	45.0	27.4
CRC	100	97.8	90.4	73.9	48.5	31.3
SRC	100	99.8	98.5	90.3	65.3	37.5
SSRC	100	100	98.5	90.0	69.8	43.0
RLRC	100	100	99.7	96.2	85.0	55.9
CESR	92.7	89.8	83.9	75.5	57.4	40.1
RSC	100	100	99.8	96.9	87.6	60.1
HQ_A	99.6	99.1	95.0	87.5	71.5	49.3
HQ_M	99.8	99.1	99.3	93.9	83.3	59.6
NMR	100	100	99.3	97.6	95.8	82.7
RMR	**100**	**100**	**100**	**100**	**100**	**100**

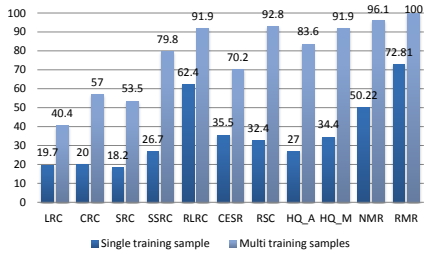


Figure 5. Recognition rates (%) of each classifier under different occlusion.

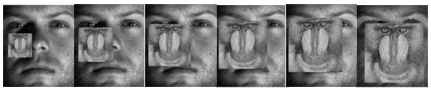
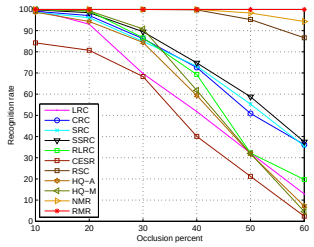


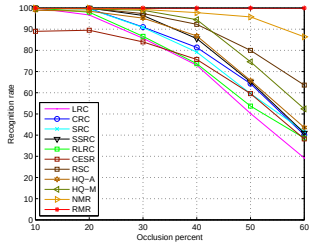
Figure 6. Sample images of one person with different levels of occlusion.

In the third experiment, we use the similar training samples as mentioned in the first experiment, and Subset 3 for testing. Here, each testing image is corrupted by randomly located square block of "baboon" with varying block sizes. The block size determines the occlusion level of an image. Fig. 6 illustrates the occlusion levels varying from 10% to 60% . Table 1 shows the recognition rates of LRC, CRC, SRC, SSRC, RLRC, CESR, RSC, HQ_A, HQ_M, NMR and RMR under different occlusion levels. From Table 1, we can see that the recognition rates of RMR are 100%, when the occlusion levels from 10% to 60% , which exhibits better performance than the compared methods.

In the fourth experiment, for the testing images in subset 3, we impose two special occlusions: a square black block and a square random block. The occlusion levels varying from 10% to 60% and the recognition rates of each method under various occlusion levels are illustrated in Figs. 7 (a) and (b) respectively. In general, the recognition rates of RMR always achieve 100%, which outperforms those state-of-the-art methods in different levels of different occlusion.



(a)



(b)

Figure 7. Recognition rates (%) of each classifier under different occlusion levels.

5.2. Experiments on the AR Database

The AR face database [14] contains over 4000 face images of 126 people, including frontal views of faces with different facial expressions, illumination conditions and occlusions. The pictures of most persons were taken in two sessions, separated by two weeks. Each section contains 13 color images and 120 individuals participated in both sessions. The images of these 120 individuals were selected and used in our experiments. We manually cropped the face portion of the image to 50×40 pixels and the images of one person are shown in Fig. 8.

First, we divide the images of one person into four sets: (1) the first 4 images of sessions 1 and 2; (2) the first 4 images of sessions 1 and 2 corrupted by a square block of "baboon" with a 60% occlusion level; (3) 6 images with sunglasses from both sessions; (4) 6 images with scarves from both sessions. In this experiment, we still set two training modes. The first mode is that we use the eight images of the first set for training and the others for testing to verify the effectiveness of our method in dealing with real disguise and high level occlusion. The second mode is that we use only the first image of session 1 for training and the test setting remains the same as the previous mode.

The results of FR are listed in Table 2. As it is shown in Table 2, compared with the other methods, the proposed method achieves the highest recognition rate for each test, no matter the training samples are adequate or not.

Table 2. Recognition rates (%) of each classifier on AR face database.

Model	Multi training samples			Single training sample		
Case	"baboon" Occlusion	Sunglasses	Scarf	"baboon" Occlusion	Sunglasses	Scarf
LRC	52.3	92.8	30.7	38.4	70.6	29.4
CRC	68.6	93.5	63.6	39.5	73.8	40.1
SRC	70.0	94.4	57.6	39.9	72.4	43.8
SSRC	94.0	95.4	66.7	53.4	74.6	45.1
RLRC	95.1	94.6	53.3	60.0	75.1	38.2
CESR	88.9	95.0	33.5	51.1	72.5	21.8
RSC	94.8	94.7	66.8	41.4	75.3	46.0
HQ_A	94.2	94.7	48.7	47.6	74.2	35.8
HQ_M	94.6	95.0	50.1	56.3	75.5	35.8
NMR	95.1	96.9	73.5	64.8	75.9	50.0
RMR	**96.9**	**97.1**	**78.1**	**65.3**	**82.1**	**54.0**

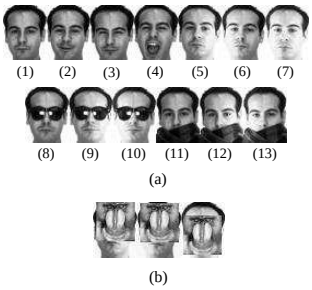
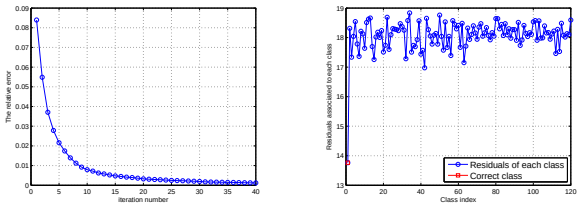


Figure 8. Example images from AR face database. (a) Sample images of a person in session 1. (b) Sample images with 60% "baboon" occlusion.



(a) (b)

Figure 9. An example of face recognition with sunglasses using RMR. (a) The convergence curve of ADMM. (b) The residual of each class by RMR.

More specifically, given adequate training samples, when the faces are disguised with sunglasses, our method shows some improvement with the recognition rate to be 97.1%; when the faces are disguised with scarves, in which the occlusion level becomes larger, the advantage of RMR becomes evident. It achieves 78.1% recognition rate that improves 4.6% compared with NMR which ranks second; when the testing images were corrupted by a very high level artificial occlusion (e.g. "baboon" with a 60% occlusion level), RMR can still achieve 96.9% recognition rate. We think the inherent difference between the natural and artificial occlusions that plays a dominant role in the recognition task, and the natural occlusion is more complicated and more difficult than the artificial one.

The convergence of the ADMM process is shown in Fig.9 (a) and Fig.9 (b) plots the residuals of each class.

6. The performance of difference parameter

In this section, we investigate how the parameter λ affect the performance of our method in different sceneries. For FR with occlusion, Subsets 1 and 2 of Extended Yale B database are used for training and Subset 3 with block occlusion (the square black 60% occlusion level) are used for testing. For FR with real-disguise, the first 4 images of Sessions 1 and 2 are used for training and the images with real face disguise (glasses, scarf) of each session are used for testing. For FR with illumination, the Subset 1 of the Extended Yale B database is used for training, and the images with extreme lighting conditions (Subset 4, Subset 5) are used for testing. We compare the performance of parameters varying in the range of [0.01, 200]. As show in Fig. 10, in the general case, the recognition rates of RMR with respect to different sceneries always achieve its optimal or nearly optimal performance when the parameter is set to be 10. Hence, we set the parameter $\lambda = 10$ in the experiment for convenience. when the testing images are extremely influenced by illumination variations (Subset 4 and Subset 5 of Extended Yale B database), our experiment results show that we get higher recognition rates by adjusting the parameter. More specifically, we set parameter $\lambda = 1$ for Subset 4 of Extended Yale B database and $\lambda = 200$ for Subset 5.

7. Conclusions

This paper presents a novel robust matrix regression model, and applies the alternating direction method of multipliers to solve it. The main merit of RMR is its robustness

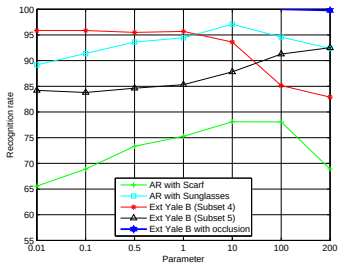


Figure 10. Recognition rates (%) with different parameters

to occlusions and illumination changes in FR. The reason lies in the fact that RMR adopts the minimal weighted nuclear norm constraint on the representation error image and takes advantage of the low rank structural information of image matrix. Extensive experimental results demonstrate that the proposed method is more robust than state-of-the-art regression based methods for FR with the variations of occlusion and illumination.

Acknowledgement

This work was partially supported by the National Science Fund for Distinguished Young Scholars under Grant Nos.61125305, 91420201, 61472187, 61233011, 61502235 and 61373063, the Key Project of Chinese Ministry of Education under Grant No.313030, the 973 Program No. 2014CB349303, Fundamental Research Funds for the Central Universities No.30920140121005, and Program for Changjiang Scholars and Innovative Research Team in University No. IRT13072.

References

[1] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *Signal Processing Letters, IEEE*, 14(10):707–710, 2007.

[2] J.-J. Chen, J. Yang, L. Luo, J. Qian, and W. Xu. Matrix variate distribution-induced sparse representation for robust image classification. 2015.

[3] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

[4] M. Fazel, H. Hindi, and S. P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *American Control Conference, 2003. Proceedings of the 2003*, volume 3, pages 2156–2162. IEEE, 2003.

[5] S. Gu, L. Zhang, W. Zuo, and X. Feng. Weighted nuclear norm minimization with application to image denoising. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2862–2869. IEEE, 2014.

[6] R. He, W.-S. Zheng, and B.-G. Hu. Maximum correntropy criterion for robust face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1561–1576, 2011.

[7] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong. A regularized correntropy framework for robust pattern recognition. *Neural Computation*, 23(8):2074–2100, 2011.

[8] R. He, W.-S. Zheng, T. Tan, and Z. Sun. Half-quadratic-based iterative minimization for robust sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(2):261–275, 2014.

[9] D. R. Hunter and K. Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[10] K. Jia, T.-H. Chan, and Y. Ma. Robust and practical face recognition via structured sparsity. In *Computer Vision–ECCV 2012*, pages 331–344. Springer, 2012.

[11] K.-C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):684–698, 2005.

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

[13] C. Lu, J. Tang, S. Yan, and Z. Lin. Generalized nonconvex nonsmooth low-rank minimization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4130–4137. IEEE, 2014.

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

[15] I. Naseem, R. Togneri, and M. Bennamoun. Linear regression for face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):2106–2112, 2010.

[16] I. Naseem, R. Togneri, and M. Bennamoun. Robust regression for face recognition. *Pattern Recognition*, 45(1):104–118, 2012.

[17] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[18] 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 preprint arXiv:1405.1207*, 2014.

[19] J. Yang, L. Zhang, Y. Xu, and J.-y. Yang. Beyond sparsity: The role of l 1-optimizer in pattern classification. *Pattern Recognition*, 45(3):1104–1118, 2012.

[20] M. Yang, D. Zhang, and J. Yang. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 625–632. IEEE, 2011.

[21] D. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 471–478. IEEE, 2011.

[22] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003.