Nuclear-L_1 Norm Joint Regression for Face Reconstruction and Recognition

Lei Luo, Jian Yang, Jianjun Qian, and Ying Tai

Nanjing University of Science and Technology, Nanjing 210094, P.R. China

Abstract. Recognizing a face with significant lighting, disguise and occlusion variations is an interesting and challenging problem in pattern recognition. To address this problem, many regression based methods, represented by sparse representation classifier (SRC), are presented recently. SRC uses the L_1-norm to characterize the pixel-level sparse noise but ignore the spatial information of noise. In this paper, we find that nuclear-norm is good for characterizing image-wise structural noise, and thus we use the nuclear norm and L_1-norm to jointly characterize the error image in regression model. Our experimental results demonstrate that the proposed method is more effective than state-of-the-art regression methods for face reconstruction and recognition.

1 Introduction

Face recognition is closely related to our life, which has been applied widely to information security, law enforcement and surveillance, smart cards, access control, etc. However, recognizing a face with significant lighting, disguiseand occlusion variations is still a challenging problem in pattern recognition.

Recently, a number of methods have been developed to address this problem. Among them, sparse Representation Coding (SRC) [1] is the most attractive and receiving more and more attention. In fact, SRC can be considered as a generalization of nearest feature classifiers, which strikes a balance between NN [2] and NFS [3]. Differing from these classifiers, the representation of SRC is global, using all the training data as a dictionary, and the classification is performed by checking which class yields the least coding error. Because of its simplicity and effectiveness, SRC has been applied and investigated extensively. In order to further improve the robustness of sparse coding, an extended SRC [4] and some re-weighted L_1 or L_2 minimization algorithms [5], [6] were presented. Zhang et al. [7] have shown that it is the collaborative representation (CR) but not the L_1-norm sparse constraint that truly improves the FR performance. Yang et al. [8] re-examined the role of L_1-optimizer and found that for pattern recognition tasks, L_1-optimizer provides more meaningful classification information (e.g. closeness) than L_0-optimizer does. Meanwhile, integrating sparse coding with other methods is also a meaningful effort. For example, Yang et al. [9] proposed sparse representation classifier steered discriminative projection. Zheng et al. [10] performed SRC in low rank projection with discrimination.

The characterization of the residual term plays a key role in regression model based face recognition Methods. Linear regression based classifier (LRC) [11] uses L_2- norm to characterize the coding residual, while SRC uses L_1- norm. Yang et al. [12] presented robust sparse coding (RSC), which uses an M-estimator to fit the general noises. He et al. [13] proposed a correntropy based sparse representation (CESR) scheme by virtue of the correntropy induced metric for describing residual. Essentially, the core idea of Yang et al. [12] and He et al. [13] is to use a robust estimator to generate the new variables in accordance with the known distributions. Li et al. [14] explored the structure of the error incurred by occlusion and measured errors by the weighted L_1 metric. K. Jia et al. [15] introduced a class of structured sparsity-inducing norms into the SRC framework to fit these structural noises.Yang et al. [16] used nuclear norm to describe the residual term and proposed a nuclear norm based matrix regression (NMR) model, which has been shown that NMR is robust to face recognition with occlusion and illumination changes.

From the probability distribution point of view, we know that L_1-norm provides an optimal characterization for errors with the Laplace distribution. Therefore, SRC (with L_1-norm) generally performs well for the sparse noise which on the whole follows the Laplace distribution. However, in practice, some noises caused by occlusions, disguise or illumination does not follow Laplacian distribution (see the example in Fig.1). Thus, L_1-norm is not enough for error characterization. In this paper, we find that the singular values of the error image fit Laplace distribution well in real-world disguises, occlusions, or illumination induced error images. Thus, we can use L_1-norm of the singular value vector, i.e., nuclear norm of the error image, to characterize this kind of structural noises. To handle the pixel-level sparse noise and image-level structural noise together, we will use two norms, i.e., nuclear norm and L_1-norm, to jointly characterize the error image in our regression model.

The proposed nuclear-L_1 norm joint regression model can be solved by using alternating direction method of multipliers (ADMM). In each step of the algorithm, a closed-form solution can be obtained by fixing the other variables. In general, the complexity of the proposed algorithm is much lower than SRC or RSC. In addition, nuclear norm is used as a metric to characterize the distance between test samples and classes, which is different from the previous methods using of the Euclidean (L_2)-norm.We perform experiments on the Extended Yale B and AR databases, the results demonstrate that the proposed method is more effective than state-of-the-art regression methods for face reconstruction and recognition.

The remainder of this paper is structured as follows: In Section 2, we first introduce our model, i.e., the nuclear-L_1 norm joint regression (NL_1R) model. In Section 3, we solve the proposed model by virtue of ADMM. In Section 4, we study the complexity of our algorithm. In Section 5, we design the NL_1R based classifier. In Section 6, we present a series of experiments to demonstrate the robustness effectiveness of the proposed algorithm. In Section 7, we conclude the paper with a brief conclusion.

2 Problem Formulation

Given a set of n observed 2D data matrices $\mathbf{A}_1, \cdots, \mathbf{A}_n \in R^{p\times q}$ and a matrix $\mathbf{D} \in R^{p\times q}$, let us represent $\mathbf{B}$ linearly using $\mathbf{A}_1, \cdots, \mathbf{A}_n$, i.e., $\mathbf{D} = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 +, \cdots, + x_n\mathbf{A}_n + \mathbf{E}$, where $x_1, x_2, \cdots, x_n$ is a set of representation coefficients, $x_1\mathbf{A}_1 + x_2\mathbf{A}_2 +, \cdots, + x_n\mathbf{A}_n$ is the reconstructed image and $\mathbf{E}$ is the representation residual. Let us denote the following linear mapping from R^n to $R^{p\times q}$: $\mathrm{A}(\mathbf{x}) = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 +, \cdots, + x_n\mathbf{A}_n$, where $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$. Then, we will consider the following model

$$\min_{\mathbf{E},\mathbf{x}} \|\mathbf{E}\| \quad \text{s.t.} \quad \mathrm{A}(\mathbf{x}) - \mathbf{D} = \mathbf{E}, \tag{1}$$

where $\|\cdot\|$ is a norm.

It is crucial that which norm should be chosen to characterize the error matrix $\mathbf{E}$ better. It's well-known that if errors are independently and identically distributed with Laplacian(or Gaussian), then L_1(or L_2)-norm is optimal for characterizing the errors (a proof is in [12]). This means there must exist some close relationship between the error (or residual) metric and error distribution. Figure 1(a) shows an original image with scarf. One can decompose (a) into the recovered term (b) and noise term (c). Figure 2(a) delineates the empirical and fitted distributions of noise term $\mathbf{E}$ by using Gaussian or Laplacian distribution model. We can see that Gaussian and Laplacian distribution are far away from the empirical distribution. So, L_2-norm (or L_1-norm) based method can not describe the noise matrix effectively. Instead, Figure 2(b) shows that singular values of noise matrix $\mathbf{E}$ follow Laplacian distribution well. And this trend becomes more stable and evident with the increase of the size of images. In addition, for other noises caused by occlusion and illumination, we obtain the similar result.

Thus, it's reasonable that we assume that the singular values of error matrix are independently and identically distributed with Laplacian distribution, i.e.,

$$\delta_i \sim p_\theta(\delta_i) = \tfrac{1}{2b} \exp(-|\delta_i - \mu|/b), \tag{2}$$

where $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_n$ are the all singular values of error matrix $\mathbf{E}$, $\theta = (\mu, b)$.

Thus, the likelihood function of the estimator is that

$$\prod_{i=1}^{n} p_\theta(\delta_i) = \tfrac{1}{(2b)^n} \exp(-\sum_{i=1}^{n} |\delta_i - \mu|/b). \tag{3}$$

By taking the logarithm, we obtain that

$$\ln \prod_{i=1}^{n} p_\theta(\delta_i) = -\sum_{i=1}^{n} |\delta_i - \mu|/b + \ln \tfrac{1}{(2b)^n}. \tag{4}$$

For convenience, we can assume $\mu = 0,\ b = 1$. According to the maximum likelihood criterion, we need to maximizing $\ln \prod\limits_{i=1}^{n} p_\theta(\delta_i)$, which is equal to minimizing

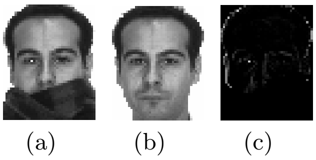
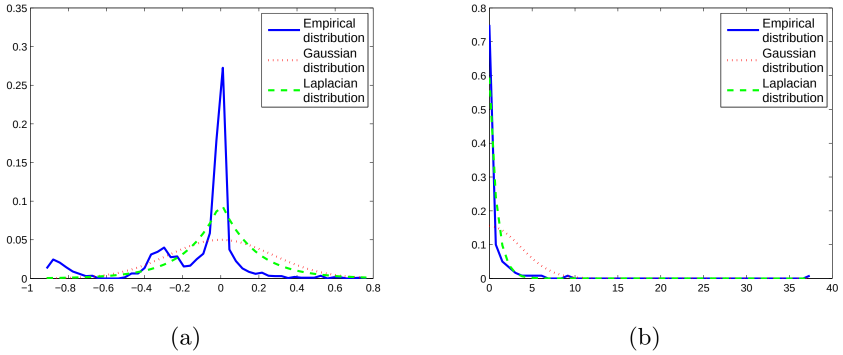

(a) (b) (c)

Fig. 1. (a) Original image; (b) recovered image; (c) noise image **E**



(a) (b)

Fig. 2. (a) The empirical distribution and the fitted distribution of the noise image **E**; (b) The empirical distribution and the fitted distribution of the singular value vector of noise image **E**

$\sum_{i=1}^{n} |\delta_i|$, thus,

$$\min \sum_{i=1}^{n} |\delta_i| = \min |\delta|_1 = \min \|\mathbf{E}\|_* . \qquad (5)$$

Therefore, under this assumption, nuclear norm can be chosen as a proper descriptor to characterize structural noises. Certainly, for some sparse noises, which follow Laplacian distribution, the L_1-norm is an optimal choice. In order to keep the advantage of L_1-norm, we can use L_1-norm as a regularized term to further improve the performance of nuclear norm, which yields the following model:

$$\min_{\mathbf{E},\mathbf{x}} \|\mathbf{E}\|_* + \alpha \|\mathbf{E}\|_1, \quad \text{s.t.} \quad \mathrm{A}(\mathbf{x}) - \mathbf{D} = \mathbf{E}, \qquad (6)$$

where $\alpha > 0$ is a parameter. It is used to balance the nuclear norm and L_1-norm.

The advantage of this method is that the metric based on different norms can complement each other long for short, which prevents the limitation of the single metric. Thus, the collaborative effect of nuclear-L_1 norm will be suitable for characterizing the reconstruction error (or the difference between occluded face image and its ground truth)if we choose a proper parameter α. In Section 6, we will further see that the joint use of two norms is robust for the characterization of noises, and this method will fit more complicated noises.

Furthermore, borrowing the idea of the ridge regression, we would like to add a similar regularization term to Eq. (6) and obtain the regularized matrix regression model based on nuclear-L_1 norm:

$$\min_{\mathbf{E},\mathbf{x}} \|\mathbf{E}\|_* + \alpha\|\mathbf{E}\|_1 + \frac{1}{2}\beta \left\|\mathbf{x}\right\|_2^2, \quad \text{s.t.} \quad \mathrm{A}\left(\mathbf{x}\right) - \mathbf{D} = \mathbf{E}. \tag{7}$$

The new objective function is non-smooth, but continuous and convex. For the convenience, we introduce an auxiliary variable $\mathbf{Z}$ for the splitting, thus, (7) is converted to the following equivalent problem:

$$\min_{\mathbf{E},\mathbf{Z},\mathbf{x}} \|\mathbf{E}\|_* + \alpha\|\mathbf{Z}\|_1 + \frac{1}{2}\beta \left\|\mathbf{x}\right\|_2^2, \quad \text{s.t.} \quad \mathrm{A}\left(\mathbf{x}\right) - \mathbf{D} = \mathbf{E}, \ \mathbf{E} = \mathbf{Z}. \tag{8}$$

In (8), the new constraint $\mathbf{E} = \mathbf{Z}$ guarantees the identity of $\mathbf{E}$ and $\mathbf{Z}$, thus, $\mathbf{Z}$ can be regarded as a proxy for $\mathbf{E}$. We will discuss how to solve this model in the following section.

3 Proposed Algorithm

The alternating direction method of multipliers (ADMM) or the augmented Lagrange multipliers (ALM) method was presented originally in [17], [18], which has been studied extensively in the theoretical frameworks of Lagrangian functions [19]. Recently, ADMM has been applied to the nuclear norm optimization problems [20], [21], which updates the variables alternately by minimizing the augmented Lagrangian function with respect to the variables in a Gauss-Seidel manner. Here, we provide the process of using ADMM to solve the problem (8), which is equal to minimizing the following augmented Lagrangian function:

$$\begin{aligned} L_\mu = \quad & \|\mathbf{E}\|_* + \alpha\|\mathbf{Z}\|_1 + \tfrac{1}{2}\beta \left\|\mathbf{x}\right\|_2^2 + tr\left({\mathbf{Y}_1}^T\left(\mathrm{A}\left(\mathbf{x}\right) - \mathbf{D} - \mathbf{E}\right)\right) \\ & + tr\left({\mathbf{Y}_2}^T\left(\mathbf{E} - \mathbf{Z}\right)\right) + \tfrac{\mu}{2}\left(\left\|\mathrm{A}\left(\mathbf{x}\right) - \mathbf{D} - \mathbf{E}\right\|_F^2 + \left\|\mathbf{E} - \mathbf{Z}\right\|_F^2\right), \end{aligned} \tag{9}$$

where μ is a penalty parameter, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are the Lagrange multipliers. We proceed by alternately fixing one variable and solving for the other, and iterating. Then, the detailed algorithm for solving nuclear-L_1 norm joint regression (NL_1R) model is summarized in Algorithm 1.

The key steps are to solve the optimization problems in step 2, 3 and 4. For step 4, by taking the derivative w.r.t $\mathbf{x}$ for the objective function, and setting the derivative to zero, we have the optimal solution of the sub-problem in step 4:

$$\mathbf{x} = (\mu\mathbf{M}^T\mathbf{M}{+}\beta\mathbf{I})^{-1}\mathbf{M}^T\mathrm{Vec}(\mu\mathbf{D}{+}\mu\mathbf{E} - \mathbf{Z}_1), \tag{10}$$

where $\mathbf{M} = [\mathrm{Vec}(\mathbf{A}_1), \cdots, \mathrm{Vec}(\mathbf{A}_n)]$, Vec is an operator converting a matrix into a vector.

For step 3, we need to introduce the following soft-thresholding (shrinkage) operator:

$$S_\varepsilon\left[x\right] = \begin{cases} x - \varepsilon, \ \text{if } x < \varepsilon, \\ x + \varepsilon, \ \text{if } x > -\varepsilon, \\ 0, \text{otherwise}, \end{cases} \tag{11}$$

Algorithm 1 Solving NL_1R by ADMM

```
Input: A set of matrices $\mathbf{A}_1, \cdots, \mathbf{A}_n$ and a matrix $\mathbf{D} \in R^{p\times q}$, the model parameters $\alpha, \beta$ and the value of $\mu$.
while not converged do
1. Initialize $\mathbf{Z} = \mathbf{Y_1} = \mathbf{Y_2} = \mathbf{0},\ \mathbf{x} = \mathbf{0}$;
2. fix the others and update $\mathbf{E}$ by
   $\mathbf{E} = \arg\min_{\mathbf{E}} \frac{1}{\mu}\|\mathbf{E}\|_* + \frac{1}{2}\left(\left\|\mathbf{E} - \left(\mathrm{A}(\mathbf{x}) - \mathbf{D} + \frac{1}{\mu}\mathbf{Y}_1\right)\right\|_F^2 + \left\|\mathbf{E} - \left(\mathbf{Z} - \frac{1}{\mu}\mathbf{Y}_2\right)\right\|_F^2\right);$
3. fix the others and update $\mathbf{Z}$ by
   $\mathbf{Z} = \arg\min_{Z} \frac{\alpha}{\mu}\|\mathbf{Z}\|_1 + \frac{1}{2}\left\|\mathbf{Z} - \left(\mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2\right)\right\|_F^2;$
4. fix the others and update $\mathbf{x}$ by
   $\mathbf{x} = \arg\min_{\mathbf{x}} \frac{1}{2}\beta\|\mathbf{x}\|_2^2 + \frac{\mu}{2}\left\|\mathrm{A}(\mathbf{x}) - \mathbf{D} - \mathbf{E} + \frac{1}{\mu}\mathbf{Y}_1\right\|_F^2;$
5. update the multipliers $\mathbf{Y}_1 = \mathbf{Y}_1 + \mu(\mathrm{A}(\mathbf{x}) - \mathbf{D} - \mathbf{E})$ and $\mathbf{Y}_2 = \mathbf{Y}_2 + \mu(\mathbf{E} - \mathbf{Z})$.
end while
Output: Optimal representation coefficient $\mathbf{x}$ and $\mathbf{E}, \mathbf{Z}$
```

where $x \in R,$ and $\varepsilon > 0$, if we extend soft-thresholding operator to vectors or matrices, then we have

$$S_\varepsilon[\mathbf{W}] = \arg\min_{\mathbf{X}} \varepsilon\|\mathbf{X}\|_1 + \tfrac{1}{2}\|\mathbf{X} - \mathbf{W}\|_F^2. \tag{12}$$

That is, the optimal solution of the sub-problem in step 3 is that:

$$\mathbf{Z} = sgn(\mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2) \circ max\{\left|\mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2\right| - \frac{\alpha}{\mu}, 0\}, \tag{13}$$

where the symbolic function $sgn(\cdot)$ and the absolute value $|\cdot|$ act on the each element of the matrix $\mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2$, and $\circ$ is the Hadamard product.

In the following, we consider how to solve the sub-problem in step 2.

Given a matrix $\mathbf{Q} \in R^{p\times q}$ of rank r, the singular value decomposition (SVD) of $\mathbf{X}$ is

$$\mathbf{Q} = \mathbf{U}_{p\times r}\mathbf{\Sigma}\mathbf{V}_{q\times r}^T, \quad \mathbf{\Sigma} = diag(\sigma_1, \cdots, \sigma_r), \tag{14}$$

where $\sigma_1, \cdots, \sigma_r$ 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

$$\mathrm{D}_\tau(\mathbf{Q}) = \mathbf{U}_{p\times r}\mathrm{diag}\left(\{\max(0, \sigma_j - \tau)\}_{1\le j\le r}\right)\mathbf{V}_{q\times r}^T. \tag{15}$$

Theorem 1. For each $\mathbf{A}, \mathbf{B} \in R^{p \times q}$ and $\tau > 0$, the singular value shrinkage operator in (15) obeys

$$\frac{1}{2}\mathrm{D}_\tau (\mathbf{A} + \mathbf{B}) = \arg\min_E \tau \|\mathbf{E}\|_* + \frac{1}{2}\left(\|\mathbf{E} - \mathbf{A}\|_F^2 + \|\mathbf{E} - \mathbf{B}\|_F^2\right). \tag{16}$$

Proof: Since the function $\mathrm{h}_0(\mathbf{E}) = \tau\|\mathbf{E}\|_* + \frac{1}{2}\left(\|\mathbf{E} - \mathbf{A}\|_F^2 + \|\mathbf{E} - \mathbf{B}\|_F^2\right)$ is strictly convex, it is easy to see that there exists a unique minimizer, and we thus need to prove that it is equal to $\frac{1}{2}\mathrm{D}_\tau(\mathbf{A} + \mathbf{B})$. To do this, recall the definition of a sub-gradient of a convex function $\mathrm{f} : R^{n_1 \times n_2} \to R$ .We say that $\mathbf{J}$ is a sub-gradient of f at $\mathbf{E}_0$, denoted $\mathbf{J} \in \partial\mathrm{f}(\mathbf{E}_0)$, if

$$\mathrm{f}(\mathbf{E}) \geq \mathrm{f}(\mathbf{E}_0) + \langle \mathbf{Z}, \mathbf{E} - \mathbf{E}_0 \rangle \tag{17}$$

for all $\mathbf{E}$. Now $\widehat{\mathbf{E}}$ minimizes h_0 if and only if 0 is a sub-gradient of the functional h_0 at the point $\widehat{\mathbf{E}}$, i.e.

$$0 \in \widehat{\mathbf{E}} - \mathbf{A} + \widehat{\mathbf{E}} - \mathbf{B} + \tau\partial\left\|\widehat{\mathbf{E}}\right\|_* = 2\widehat{\mathbf{E}} - (\mathbf{A} + \mathrm{B}) + \tau\partial\left\|\widehat{\mathbf{E}}\right\|_*, \tag{18}$$

where $\partial\left\|\widehat{\mathbf{E}}\right\|_*$ is the set of sub-gradients of the nuclear norm. Let $\mathbf{E} \in R^{n_1 \times n_2}$ be an arbitrary matrix and $\mathbf{U\Sigma V}^T$ be its SVD. It is known that

$$\partial\|\mathbf{E}\|_* \in \left\{\mathbf{UV}^T + \mathbf{W} : \mathbf{W} \in R^{n_1 \times n_2}, \mathbf{U}^T\mathbf{W} = 0, \mathbf{WV} = 0, \|\mathbf{W}\|_2 \leq 1\right\}. \tag{19}$$

Set $\widehat{\mathbf{E}} = \frac{1}{2}\mathrm{D}_\tau(\mathbf{A} + \mathbf{B})$ for short. In order to show that $\widehat{\mathbf{E}}$ obeys (17), decompose the SVD of $\mathbf{A} + \mathbf{B}$ as

$$\mathbf{A} + \mathbf{B} = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0{}^* + \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*, \tag{20}$$

where $\mathbf{U}_0,\ \mathbf{V}_0$ (resp. $\mathbf{U}_1,\ \mathbf{V}_1$) are the singular vectors associated with singular values greater than τ (resp. smaller than or equal to τ). With these notations, we have

$$\widehat{\mathbf{E}} = \frac{1}{2}\mathbf{U}_0(\mathbf{\Sigma}_0 - \tau\mathbf{I})\mathbf{V}_0^*, \tag{21}$$

therefore,

$$\mathbf{A} + \mathbf{B} - 2\widehat{\mathbf{E}} = \tau(\mathbf{U}_0\mathbf{V}_0^* + \mathbf{W}), \mathbf{W} = \tau^{-1}\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*. \tag{22}$$

By definition, $\mathbf{U}_0^*\mathbf{W} = 0, \mathbf{WV}_0 = 0$ and since the diagonal elements of $\mathbf{\Sigma}_1$ have magnitudes bounded by τ , we also have $\|\mathbf{W}\|_2 \leq 1$. Hence $\mathbf{A}+\mathbf{B}-2\widehat{\mathbf{E}} \in \tau\partial\left\|\widehat{\mathbf{E}}\right\|_*$, which concludes the proof.

Therefore, for the sub-problem in step 2, the optimal solution is that:

$$\mathbf{E} = \frac{1}{2}\mathrm{D}_{\frac{1}{\mu}}\left(\mathrm{A}(\mathbf{x}) - \mathbf{D} + \frac{1}{\mu}\mathbf{Y}_1 + \mathbf{Z} - \frac{1}{\mu}\mathbf{Y}_2\right) \tag{23}$$

It should be noted that (8) is different from low rank representation (LRR) [22], because both original variable $\mathbf{E}$ and the auxiliary variable $\mathbf{Z}$ brought in are all in the objective function for model (8). The aim of LRR is subspace segmentation, and the nuclear norm is the replacement of rank. But our model is based on face reconstruction and recognition, which is from the view of regression. And the nuclear norm is used to characterize the distribution of the singular values.

4 Complexity and Convergence Analysis

In this part, we discuss the time complexity of the proposed algorithm. It is easy to see that the main running time of the proposed algorithm is consumed by performing SVD on the small matrix of the size $p \times q$, and some matrix multiplications. In step 2, the time complexity of performing SVD is $O\left(pq^2\right)$ (we can assume that $q \leq p$). The time complexity of matrix multiplications is $O\left(npq + n^2\right)$. Thus, the total time complexity of the proposed algorithm is $O\left(pq^2 + npq + n^2\right)$. It is also reported that the commonly used L_1-minimization solvers have an empirical complexity of $O\left(p^2q^2n^{1.3}\right)$ and the complexity of RSC with $\beta = 1$ is about $O\left(p^2q^2n\right)$, where n is the sample number [12]. Now, we compare the complexity of RSC with Algorithm 1. Firstly, we can obtain that $\frac{pq^2+npq+n^2}{p^2q^2n} = \frac{1}{pn} + \frac{1}{pq} + \frac{n}{p^2q^2}$. In general, in the face recognition experiments, $pq \geq 10$, $pn \geq 10$, thus, if $\frac{n}{p^2q^2} \leq \frac{4}{5}$, i.e., $n \leq \frac{4}{5}p^2q^2$, then, our algorithm in this paper will have much lower complexity. It is evident that this condition $n \leq \frac{4}{5}p^2q^2$ can be easily satisfied, for example, in our experiments, $pq \geq 10$, $pn \geq 10$. This main reason for the lower complexity is that our model is based on matrix computation directly, e.g., in step 2, we don't need to convert the each sample of train image into a vector.

The convergence properties of the ADMM have been generally discussed. For more details, one may refer to [21], [23]. But in this paper, it is enough that we only need to choose a proper termination parameters ε, and use the following termination conditions:

$$\left\|\mathrm{A}\left(\mathbf{x}\right) - \mathbf{D} - \mathbf{E}\right\|_\infty \leq \varepsilon \text{ and } \left\|\mathbf{E} - \mathbf{Z}\right\|_\infty \leq \varepsilon. \tag{24}$$

5 The Design of the Classifier

For the design of classifier, some new ideas should be noted, for example, Luan et al. [24] introduced two descriptors, i.e., sparsity and smoothness, to represent characteristic of the sparse error component, and applied them to face recognition. Li and Lu [25] proposed a new decision rule, i.e., sum of coefficient (SoC) to match better with SRC. That is, they make full use of the information of the objective function.

In this section, we will use nuclear norm as a metric to characterize the distance between test samples and classes. This is because nuclear norm is more robust than Frobenius (L_2)-norm as a distance metric [26]. Meanwhile, since reconstruction image of all training images can be regarded as the denoised image, thus, we use it as the reference image of classification. That is, we first use Algorithm 1 to obtain the optimal representation coefficients $\mathbf{x}^*$ for a test image $\mathbf{D}$, then use the reconstruction image $\mathrm{A}\left(\mathbf{x}^{*}\right)$ of all training images as the new reference image of classification. In addition, let $\sigma_i : R^n \rightarrow R^n$ be the characteristic function that selects the coefficients associated with the i-th class. For $\mathbf{x} \in R^n$, $\sigma_i(\mathbf{x})$ 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 $\mathbf{D}$ in class i as $\hat{\mathbf{D}}_i = \mathrm{A}\left(\sigma_i(\mathbf{x}^{*})\right)$. Finally, the nuclear norm of the representation residual is used to characterize the distance between reconstruction image and classes, that is, $\mathrm{r}_i\left(\mathbf{D}\right) = \left\|\mathrm{A}\left(\mathbf{x}^{*}\right) - \mathrm{A}\left(\sigma_i\left(\mathbf{x}^{*}\right)\right)\right\|_{*}$ for $i = 1, \cdots k$. Thus, we can define the following decision rule: if $r_l(\mathbf{D}) = \min_i r_i(\mathbf{D})$, then $\mathbf{D}$ belongs to Class l.

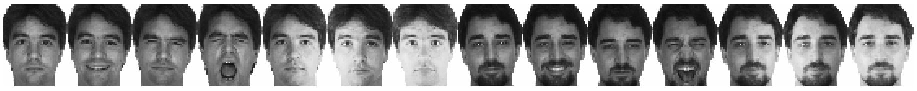
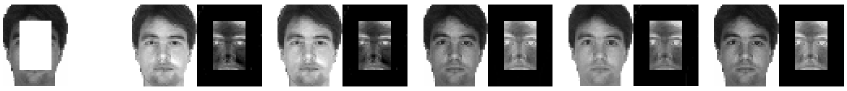
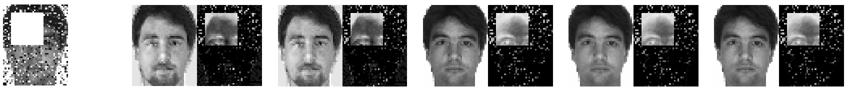


Fig. 3. Fourteen samples of cropped images of one person for training on AR database



(a) Test image (b) SRC (c) CRC (d) RSC (e) CESR (f) NL_1R

Fig. 4. Recovered clean image and occluded part via five methods for the image **B** with white block image



(a) Test image (b) SRC (c) CRC (d) RSC (e) CESR (f) NL_1R

Fig. 5. Recovered clean image and occluded part via five methods for the image **B** with composite noise

Table 1. The comparison of the error rates(%) for face reconstruction via five methods for the image B with two different cases (a) and (b) corresponding to Fig 4. and Fig. 5, respectively

Cases	SRC	CRC	RSC	CESR	NL_1R
(a)	33.43	33.42	0.38	15.30	3.5491×10^{-9}
(b)	31.84	31.84	0.18	4.02	8.7172×10^{-4}

6 Experiment and Analysis

In this Section, we perform experiments on public face image databases and compare the proposed model with state-of-the-art methods. Our aim is to demonstrate the robustness NL_1R to disguise, occlusion and illumination. Note that here in SRC and RSC, the matlab function "l_1-ls" [6] is used to calculate the sparse representation coefficients.

6.1 Databases

The AR face database [27] contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting 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 (65 men and 55 women) participated in both sessions. The images of these 120 individuals were selected and used in our experiment. We manually cropped the face portion of the image and then normalized it to 50 × 40 pixels.

The extended Yale B face database [28] contains 38 human subjects under nine poses and 64 illumination conditions the light source direction and the camera axis. The 64 images of a subject in a particular pose are acquired at camera frame rate of 30 frames/s, so there is only small change in head pose and facial expression for those 64 images. All frontal-face images marked with P00 are used, and each image is resized to 96 × 84 pixels and 42 × 48 pixels (only in Subsection (6.4)), respectively.

6.2 Face Reconstruction

To evaluate the method proposed in this paper, some experiments for face reconstruction will perform on AR face database. Given fourteen face images selected from the AR face database, as shown in Fig. 3, which are used for training. We choose the first image from training images denoted by $\mathbf{B}$ as the original image. The original image with artificial occlusion is used as the testing image. For the artificial occlusion, we choose the cases: (a) white block image, (b) random sparse noise plus white block. For these cases, a comparison of Sparse representation (SRC), Collaborative representation based classification (CRC), Robust sparse coding (RSC), correntropy-based sparse representation (CESR), and our method is shown in Fig. 4 and 5, and the comparison of the reconstruction error rates is shown in Table 1. We compute the face reconstruction error rates by $E_{error} = \|\mathbf{X} - \mathbf{B}\|_F / \|\mathbf{B}\|_F$, where $\mathbf{X}$ is the reconstruction image. From Table 1, we can find that the reconstruction performance of NL_1R is superior to the other methods evidently.

Meanwhile, we can also see that SRC and CRC are not fit to recover the clean images for the images with block image or composite noise. Thus, a suitable error metric is very important for face reconstruction.

Table 2. The maximal recognition rates(%) of SRC, LRC, CRC, RSC, CESR and NL_1R on AR database

Cases	SRC	LRC	CRC	RSC	CESR	NL_1R
Clear	99.2	86.8	98.9	99.0	92.2	99.7
Glasses	95.1	93.2	92.9	96.7	95.0	96.7
Scarf	66.2	30.7	63.7	64.3	33.5	73.3

Table 3. The maximal recognition rates(%) of SRC, LRC, CRC, RSC, CESR and NL_1R on the extended Yale B face database

SRC	LRC	CRC	RSC	CESR	NL_1R
94.0	94.3	81.9	94.2	68.8	97.5

6.3 Recognition with Real Face Disguise

In this experiment, we mainly test the robustness of NL_1R in dealing with real disguise on the AR database. Twenty-six face images of these 120 individuals are selected and used in our experiment. Eight images of them are used for training, which vary as follows: (a) neutral expression, (b) smiling, (c) angry, (d) screaming, (e)-(h) are taken under the same conditions. Eighteen images of them are used for testing, but we will set three different cases: (1) **face images without occlusion** (or **clear images**): Images from the testing set vary as follows: (i) right light on (j) left light on (k) all sides light, (l)-(n) are taken under the same conditions. (2) **face images with glasses**: Images from the testing set vary as follows: (i) wearing sun glasses (j) wearing sun glasses and left light on (k) wearing sun glasses and right light on, and (l)-(n) are taken under the same conditions as (i)-(k). (3) **face images with scarf**: Images from the testing set vary as follows: (i) wearing scarf (j) wearing scarf and left light on (k) wearing scarf and right light on, and (l)-(n) are taken under the same conditions as (i)-(k). Thus, for each case, the total number of training samples is 840.

In all cases mentioned above, SRC, LRC , CRC , RSC , CESR and the NL_1R proposed are, respectively, used for image classification. Here, we can choose the balance factor $\alpha \in [0.00001,\ 0.5]$ and the regularized parameter $\beta \in [0.5,\ 3]$.The maximal recognition rate of each method is compared in Table 3, where the second, third and forth line correspond to the cases (1), (2) and (3), respectively. From Table 2, we can find that NL_1R gets the better performance than state-of-the-art methods. For example, CESR only achieves 33.5% for the facial images with scarves, but our method is 73.3%. This experiment means that nuclear-L_1 norm fits better to characterize real disguises.

6.4 Recognition with Illumination

In this Subsection we test the advantage of our algorithm for illumination. The first 16 images per subject are used for training, and the remaining images for testing on the extended Yale B face database, where α is 1 and the regularized parameter β is set to 0.05. Table 3 shows the results of some latest approaches and our method.We can find NL_1R achieves much higher recognition rates than the other methods. The maximal recognition rates of SRC, LRC, CRC, RSC, CESR and NL_1R are 94.0%, 94.3%, 81.9%, 94.2%, 68.8% and 97.5%, respectively. Compared to RSC, at least 3.3% improvement is achieved by NL_1R, which demonstrates NL_1R is more effective to illumination for face recognition.

6.5 Recognition with Different Random Occlusion

In the first experiment, we use the same experiment setting as in [8], [12] to test the robustness of NSC. Subsets 1 and 2 of Extended Yale B are used for training and subset 3 is used for testing. The face images are resized to 96×84. Subset 3 with the unrelated randomly block image is used for testing (see Figure 6(a)). Here, α and β is set to 0.00001 and 0.05, respectively. Figure 6(c) shows recognition rates curve of SRC, CRC, RSC, CESR and NL_1R versus the various levels of occlusion (from 10 percent to 50 percent). From Figure 6(c), we can see that the advantage of the proposed NL_1R is more evident with the level of occlusion increasing. Especially, when the occlusion percentage is 50%, NL_1R achieves the best recognition rate 95.2%, compared to 65.3% for SRC, 48.5% for CRC, 87.6%for RSC and 57.4% for CESR. And for other occlusion percents, RSC and our method achieve the similar results. But the performance of CESR and CRC is very poor when the block is large, which shows these methods are not suit to deal with this block occlusion case.

In the second experiment, we use the composite noise (pixel corruption plus unrelated randomly block occlusion)(Fig. 6(b)) to further evaluate the robustness of our method.We choose the optimal $\alpha = 200$ and $\beta = 0.00005$, respectively. Figure 6(d) shows recognition rates curve of SRC, CRC, RSC, CESR and NL_1R versus the levels of composite noise (from 10 percent to 50 percent). From Figure 6(d), we can see that when the occlusion percentage is 50%, NL_1R achieves the best recognition rate 40.8%, compared to 28.1% for SRC, 24.1% for CRC, 23.7%for RSC and 22.1% for CESR. The above experiments also illuminate that nuclear-L_1 norm is more suitable for large block occlusion and composite noise.

7 Conclusions

The characterization of noises is a significant problem in regression model based face recognition. This paper presents a nuclear-L_1 norm joint regression model. Since L_1-norm is good at characterizing sparse noises with the Laplace distribution, and nuclear norm is suitable for characterizing image-wise structural noises, our model fits more kinds of noises. This problem is solved by virtue of

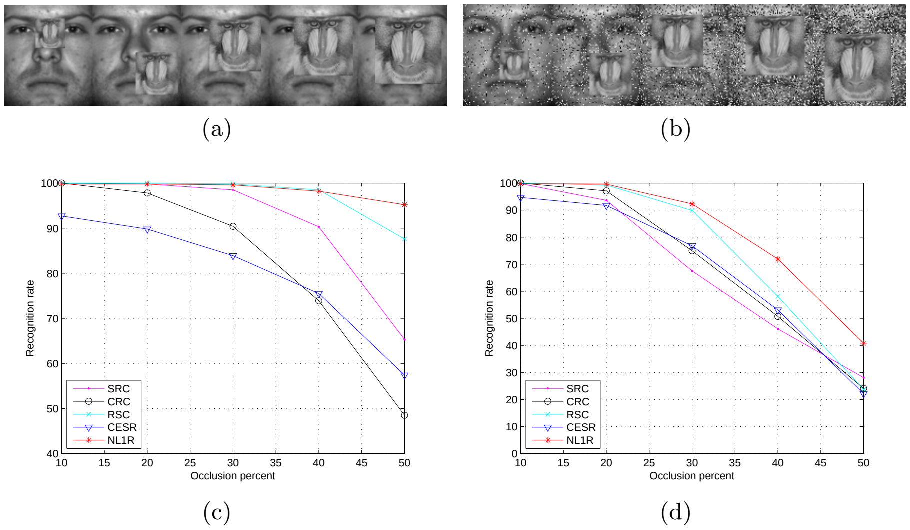


Fig. 6. (a) The face images with unrelated block occlusion; (b) the face images with Composite noise; (c) the recognition rates(%) of SRC, CRC, RSC, CESR and NL_1R under the unrelated block occlusion percentage from 10 to 50; (d) the recognition rates(%) of SRC, CRC, RSC, CESR and NL_1R under the composite noise percentage from 10 to 50.

ADMM. In addition, nuclear norm is employed as a metric to characterize the distance between test samples and classes. Our experiments demonstrate that the proposed method is more effective than state-of-the-art regression methods for face reconstruction and recognition.

References

1. Wright, J., Yang, A., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE PAMI **31** (2009) 210–227
2. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13** (1967) 21–27
3. Li, S., Lu, J.: Face recognition using the nearest feature line method. IEEE Trans. Neural Networks **10** (1999) 439–443
4. Lu, C.Y., Min, H., J. Gui, L.Z., Lei, Y.K.: Face recognition via weighted sparse representation. J. Vis. Commun. Image Represent **24** (2003) 111–116
5. Daubechies, I., Devore, R., Fornasier, M., Gunturk, C.: Iteratively re-weighted least squares minimization for sparse recovery. arXiv: 0807.0575 (2008)
6. Cands, E., Wakin, M., Boydg, S.: Enhancing sparsity by reweighted l1 minimization. J. Fourier Anal. Appl. **14** (2008) 877–905
7. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: which helps face recognition? In ICCV (2011)
8. Yang, J., Zhang, L., Xu, Y., Yang, J.Y.: Beyond sparsity: The role of l1-optimizer in pattern classification. Pattern Recognition **45** (2012) 1104–1118

9. Yang, J., Chu, D., Zhang, L., Xu, Y.: Sparse representation classifier steered discriminative projection with applications to face recognition. IEEE Trans. Neural Netw. Learn. Syst. **24** (2013) 1023–1035
10. Zheng, Z., Zhang, H., Jia, J., Zhao, J., Guo, L., Fu, F., Yu, M.: Low-rank matrix recovery with discriminant regularization. In Advances in Knowledge Discovery and Data Mining (2013) 473–448
11. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. IEEE PAMI **32** (2010) 2106–2112
12. Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In CVPR (2011)
13. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. IEEE PAMI (2011) 1753–1766
14. Li, X.X., Dai, D.Q., Zhang, X.F., , Ren, C.X.: Structured sparse error coding for face recognition with occlusion. IEEE Trans. on Image Processing **22** (2013) 1889–1990
15. Jia, K., Chan, T.H., Ma, Y.: Robust and practical face recognition via structured sparsity. In Computer Vision-ECCV (2012) 331–344
16. Yang, J., Qian, J.J., Luo, L., Zhang, F.L., , Gao, Y.C.: Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. arXiv:1405.1207 (2014)
17. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximations. IEEE Trans. on Image Processing **22** (1976) 17–140
18. Gabay, D.: Applications of the method of multipliers to variational inequalities. In Fortin, M., Glowinski,R. (eds.) Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems (1983) 299–331
19. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statisticallearning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3** (2011) 1–112
20. Hansson, A., Liu, Z., Vandenberghe, L.: Subspace system identification via weighted nuclear norm optimization. In CDC (2012) 3439–3444
21. Lin, Z., Chen, M., Ma, Y.: Multiplier method for exact recovery of corrupted low-rank matrices. UIUC Technical Report UILU-ENG-09-2215 (2009)
22. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Patt. Anal. Mach. Intell. (2013) 171–184
23. He, B., Tao, M., Yuan, X.: Alternating direction method with gaussian back substitution for separable convex programming. SIAM Journal on Optimization **22** (2012) 313–340
24. Luan, X., Liu, B., Yang, L., Qian, J.: Extracting sparse error of robust pca for face recognition in the presence of varying illumination and occlusion. Pattern Recognition **47** (2014) 495–508
25. Li, J., Lu, C.Y.: A new decision rule for sparse representation based classification for face recognition. Neurocomputing **116** (2013) 265–271
26. Gu, Z.H., Shao, M., Li, L.Y.: Discriminative metric: Schatten norm vs. vector norm. (ICPR 2012, Tsukuba, Japan)
27. Martinez, A., benavente, R.: The ar face database. Tech-nical Report 24, CVC (1998)
28. Lee, K., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE PAMI **27** (2005) 684–698