Two-Stage Nonnegative Sparse Representation for Large-Scale Face Recognition

Ran He, *Member, IEEE*, Wei-Shi Zheng, *Member, IEEE*, Bao-Gang Hu, *Senior Member, IEEE*, and Xiang-Wei Kong, *Member, IEEE*

***Abstract*— This paper proposes a novel nonnegative sparse representation approach, called two-stage sparse representation (TSR), for robust face recognition on a large-scale database. Based on the divide and conquer strategy, TSR decomposes the procedure of robust face recognition into outlier detection stage and recognition stage. In the first stage, we propose a general multisubspace framework to learn a robust metric in which noise and outliers in image pixels are detected. Potential loss functions, including L_1, $L_{2,1}$, and correntropy are studied. In the second stage, based on the learned metric and collaborative representation, we propose an efficient nonnegative sparse representation algorithm to find an approximation solution of sparse representation. According to the L_1 ball theory in sparse representation, the approximated solution is unique and can be optimized efficiently. Then a filtering strategy is developed to avoid the computation of the sparse representation on the whole large-scale dataset. Moreover, theoretical analysis also gives the necessary condition for nonnegative least squares technique to find a sparse solution. Extensive experiments on several public databases have demonstrated that the proposed TSR approach, in general, achieves better classification accuracy than the state-of-the-art sparse representation methods. More importantly, a significant reduction of computational costs is reached in comparison with sparse representation classifier; this enables the TSR to be more suitable for robust face recognition on a large-scale dataset.**

***Index Terms*— Correntropy, L_1 regularization, large-scale, nonnegative sparse representation, robust face recognition.**

I. Introduction

AUTOMATIC face recognition has been a popular research area in computer vision and machine learning. It is also one of the most successful applications of image analysis and understanding. A face recognition system identifies one person by comparing a query facial image with the registered images in a face database [1]. Two major concerns in designing a face recognition system are: 1) the query images are subject to changes in illumination as well as occlusion [2]–[4], and 2) the number of the recorded images is often tens of thousands [5]. To address these concerns, we have to focus on two issues: 1) how to yield a robust representation for a query image, and 2) how to classify a query image as fast as possible.

In recent decades, a considerable amount of research has been reported on robust representations. To alleviate the problem of occlusion, modular eigenspaces are developed in [6]. In [7], the eigenwindow method is proposed to classify partially occluded objects. In [8], the subspace coefficients are computed by substituting the mean square errors with a conventional robust M-estimator. In [9], a subsampling and hypothesize-and-test approach is proposed to reject outliers and learn the coefficients of subspace. In [10] and [11], robust component analysis methods are developed to learn robust components and coefficients. A main limitation of these methods is that they detect the whole example as an outlier and discard it from the learning process. In [3] and [12], occlusion masks are incorporated to learn a robust classifier. However, occlusion masks may discard useful redundant information [13].

Recently, the sparse signal theory has shown that sparse representation can be robust as well as discriminative, and thus can play an important role in computer vision problems [14]. Wright *et al*. [14] proposed a sparse representation classifier (SRC) for robust face recognition, which opens a new direction to deal with occlusion and corruption in face recognition. Impressive results were reported against many well-known face recognition methods [15]. Many variations of SRC were also developed. From the viewpoint of optimization, most of those variations can be categorized into iterative shrinkage-threshold and iteratively reweighted least squares (IRLS)-based methods. Yang *et al*. [16] gave a review of iterative shrinkage-threshold based sparse representation methods for robust face recognition. Gabor features [17] and Markov random fields [13] are used to further improve performance. Based on robust M-estimators, the methods in [18]–[20] compute robust sparse representation via IRLS. Other variations of SRC are nonnegative sparse representation methods [18], [21], [22]. Although SRC and its variations significantly improve the robustness of face recognition, they still need to solve an L_1 minimization problem on the whole dataset, which makes the computation expensive for large-scale datasets.[1]

Manuscript received April 21, 2011; revised October 19, 2012; accepted October 21, 2012. Date of publication November 29, 2012; date of current version December 18, 2012. This work was supported in part by the Natural Science of Foundation of China under Grant 61075051, Grant 60971095, Grant 61103155, and Grant 61102111, and the Guangdong Provincial Government of China through the Computational Science Innovative Research Team Program.

R. He and B.-G. Hu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: rhe1979@gmail.com; hubg@nlpr.ia.ac.cn).

W.-S. Zheng is with the School of Information Science and Technology, and the Guangdong Province Key Laboratory of Computational Science, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: wszheng@ieee.org).

X.-W. Kong is with the School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: kongxw@dlut.edu.cn).

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

Digital Object Identifier 10.1109/TNNLS.2012.2226471

[1] As in [5], large-scale face recognition systems need to deal with recognition on more than 10 000 images in a face database.

In order to tackle the two basic issues above in a unified framework, we propose a two-stage sparse representation (TSR) framework based on the divide and conquer strategy. Inspired by the iteratively reweighted procedure in [18], the procedure of robust recognition is decomposed into the outlier detection stage and recognition stage. In the first stage, to deal with varying illumination as well as occlusion, we propose a general multisubspace framework to learn a robust metric. More specifically, we learn an appearance space on the training set, and segment the space into several subspaces with the same number of columns. The combined rows from their subspaces constitute the original space (see an example of the segmentation in Fig. 2). We first detect noise and outliers in these subspaces and then learn a robust metric on the appearance space. Potential loss functions, including L_1, $L_{2,1}$, and correntropy [23] are studied. In the second stage, to reduce computational costs, we propose an approximation algorithm of the nonnegative sparse representation based on the L_1 ball theory [14]. Then, based on the learned metric, we filter the large-scale dataset into a small subset according to the nearest neighbor criterion such that we can compute a sparse representation on the filtered subset. This nonnegative sparse solution is unique and can be optimized efficiently. Extensive experiments validate the proposed model in extracting sparse and robust representations, and demonstrate that the proposed model significantly reduces computational costs and even leads to better recognition results as compared to SRCs.

Compared with previous work, the major contributions of this paper lie in three-fold.

1) A divide and conquer strategy is developed for robust face recognition, which makes the computation of robust sparse representation on a large-scale face database possible.
2) An effective and simple outlier detection framework is proposed. Experimental results show that the outlier detection framework can efficiently detect continuous occlusion and corruption in face recognition. Compared with robust sparse representation methods, it can also efficiently deal with the large variations incurred by glasses and mustache in the training set.
3) Based on the analysis of L_1 regularization and L_1 ball theory [14], an approximation algorithm of the nonnegative sparse representation and a filtering strategy are developed. Finally, the necessary condition for nonnegative least squares technique to find a sparse solution is also given. Theoretical analysis and experimental results show that the filtering step plays a similar role of L_1 regularization in the nonnegative sparse representation.

This paper is a complete and systemic work of our previous conference papers [24] and [25], which focused on robust recognition and discriminative semi-supervised learning, respectively. Both theory and algorithm are significantly enhanced. From algorithmic viewpoint, we extend the outlier detection stage of [24] to a general framework, in which potential loss functions, such as L_1, $L_{2,1}$, and correntropy are studied. We also extend the outlier detection in [24] to a multisubspaces algorithm. Experimental results show that the multisubspaces strategy can further improve accuracy especially for L_1, $L_{2,1}$ loss functions. From the theoretic viewpoint, we harness the theory in [25] to show that the proposed TSR method actually finds an approximation solution of the nonnegative sparse representation. The combination of [24] and [25] makes the theory of TSR method solid and complete. The further analysis based on [25] also gives the necessary condition for the nonnegative least squares techniques in [18], [21], and [24] to find a sparse solution. In addition, more experiments on large-scale PEAL Database [26] and comparisons with robust sparse representation methods [20], [18] are conducted to evaluate the proposed model for large-scale face recognition.

The remainder of this paper is organized as follows. In Section II, we begin with a brief review of sparse representation and nonnegative sparse representation for face recognition. Then we discuss the L_1-norm technique and nonnegative least squares technique, and present an efficient algorithm in Section III. In Section IV, we detail the outlier detection framework and the TSR method for face recognition. A comparison between the proposed method and the state-of-the-art methods is conducted in Section V. Finally, we draw conclusions in Section VI.

II. Sparse Representation for Face Recognition

A. Sparse Representation-Based Methods

In machine learning and computer vision, one aims to seek a suitable sparse solution from the whole training set $X = [X_1, X_2, \ldots, X_k] \in \mathbb{R}^{d\times n}$ for k classes

$$\min ||\beta||_0 \quad \text{s.t.} \quad y = X\beta \tag{1}$$

where $||.||_0$ denotes the L_0-norm, which counts the number of nonzero entries in a vector, and $y \in \mathbb{R}^{d\times 1}$ is an input sample. However, the problem of finding the sparse solution of (1) is NP-hard, and is thus difficult to solve.

The theory of compressive sensing [27], [28] reveals that if the solution β is sparse enough, we can solve the following convex relaxed optimization problem to obtain an approximate solution:

$$\min ||\beta||_1 \quad \text{s.t.} \quad y = X\beta \tag{2}$$

where $||.||_1$ denotes the L_1-norm.

To deal with occlusions and corruptions, Wright *et al.* [14] further proposed a robust linear model as $y = X\beta + e$ where $e \in \mathbb{R}^d$ is an error item. Assuming that the noise item e has also a sparse representation, one can compute a robust sparse representation as follows:

$$\min ||\beta||_1 + ||e||_1 \quad \text{s.t.} \quad ||y - (X\beta + e)||_2 \le \varepsilon. \tag{3}$$

We denote the algorithm using (3) by SRC1.

However, in many applications, the noise level ε is unknown beforehand. In such cases, the Lasso optimization algorithm [29] can be used to recover the sparse solution from

$$\min ||y - (X\beta + e)||_2^2 + \lambda(||\beta||_1 + ||e||_1) \tag{4}$$

where λ can be viewed as an inverse of the Lagrange multiplier in (3). It has been shown in [14] that ε can be interpreted as a pixel level noise whereas λ cannot be. We denote the algorithm using (4) for face recognition by SRC2.

Many methods have been developed to solve (3) and (4). From the viewpoint of optimization, those variations are either iterative shrinkage-threshold based or IRLS based. Reference [16] gives a review of iterative shrinkage-threshold based sparse representation methods. Although SRCs indeed improve the classification rate of robust face recognition against the traditional methods in most cases [14], [16], their computational cost is still high.

B. Nonnegative Sparse Representation-Based Methods

In face recognition, one also aims to seek the sparsest nonnegative solution from the whole training set X

$$\min ||\beta||_0 \quad \text{s.t.} \quad y = X\beta \quad \text{and} \quad \beta \geq 0. \tag{5}$$

The problem of finding the sparse solution of (5) is NP-hard [30], [31], and very difficult to solve in general. Fortunately, we can replace the L_0-norm by an L_1-norm [30], [31] if the solution β is sparse enough. Then we can solve the following linear problem to obtain an approximation solution:

$$\min ||\beta||_1 \quad \text{s.t.} \quad y = X\beta \quad \text{and} \quad \beta \geq 0. \tag{6}$$

Orthogonal matching pursuit algorithm [31], second-order cone programming [22], and nonnegative least squares [21], [24] were proposed to solve the model in (6). He *et al*. [18] further combined nonnegative sparse coding and maximum correntropy criterion to deal with occlusion and corruption problems in robust face recognition. Guan *et al*. [32] adopted robust stochastic approximation for online nonnegative factorization.

Recently, Slawski and Hein [33] showed that nonnegative least squares technique with thresholding was resistant to overfitting and experimentally outperformed L_1 minimization. And extensive experimental observations [18], [21], [24] also showed that, without harnessing the L_1-norm technique, the nonnegative least squares technique can also learn a sparse representation for image-based object recognition. However, a theoretical investigation is still needed for supporting the sparse idea and discussing its relationship with the L_1 minimization technique [21], [24]. Moreover, finding sparse representation and nonnegative sparse representation remains as a difficult problem, and is an open research topic [14].

III. Nonnegative Sparse Coding Algorithms

In this section, we first propose an L_1 regularized nonnegative sparse coding algorithm based on the L_0-L_1 equivalence theory [30], [31]. Then we analyze the effectiveness of the L_1 regularized item, and discuss the necessary condition for nonnegative least squares technique to find a sparse solution. Finally, based on collaborative representation, we propose an efficient nonnegative sparse coding algorithm, which computes an approximate sparse coding on the nearest subset instead of the whole dataset.

A. L_1 Regularized Nonnegative Sparse Coding Algorithm

The nonnegative sparse coding algorithm aims to find the sparsest solution of an underdetermined and nonnegative linear system. Based on the L_0-L_1 equivalence theory and Lagrange multiplier method, we can rewrite (6) as

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1 \quad \text{s.t.} \quad \beta \geq 0. \tag{7}$$

We denote the method to solve (7) as nonnegative sparse representation (NSR). The optimal problem in (7) can be re-formulated as the following quadratic program:

$$\min_{\beta} \left(\frac{\lambda}{2} - X^T y\right)^T \beta + \frac{1}{2}\beta^T X^T X\beta \quad \text{s.t.} \quad \beta \geq 0. \tag{8}$$

Since $X^T X$ is a positive semidefinite matrix, this quadratic program in (8) is convex. Based on the Karush–Kuhn–Tucker optimal conditions, the following linear complementary problem (LCP) is derived[2] [34]:

$$\alpha = X^T X\beta - X^T y + \frac{\lambda}{2}, \quad \alpha \geq 0, \quad \beta \geq 0, \quad \beta^T \alpha = 0 \tag{9}$$

where (α, β) is often said to be a complementary solution of (8). If the matrix X has full column rank ($\text{rank}(X) = n$), the convex program in (8), and the LCP in (9) have unique solutions for each vector y.

In order to solve the above criterion, we are now introducing an active set-based optimization technique. The KKT tells us that not all coefficients in α are active. Hence, we divide them into two sets. Let F and G be two subsets of $\{1, \ldots, n\}$ such that $F \cup G = \{1, \ldots, n\}$ and $F \cap G = \phi$. Let F and G be the working set and inactive set in the active set algorithm, respectively. Considering the following column partition of the matrix $X = [X_F, X_G]$ where $X_F \in \mathbb{R}^{m \times |F|}$, $X_G \in \mathbb{R}^{m \times |G|}$, and $|F|$, $|G|$ are the numbers of F and G, respectively, we can rewrite (9) as:

$$\begin{bmatrix} \alpha_F \\ \alpha_G \end{bmatrix} = \begin{bmatrix} X_F^T X_F & X_F^T X_G \\ X_G^T X_F & X_G^T X_G \end{bmatrix} \begin{bmatrix} \beta_F \\ \beta_G \end{bmatrix} - \begin{bmatrix} X_F^T y \\ X_G^T y \end{bmatrix} + \frac{\lambda}{2}$$

where $\beta_F,\ \alpha_F \in \mathbb{R}^{|F|}$, $\beta_G,\ \alpha_G \in \mathbb{R}^{|G|}$, $\beta = (\beta_F, \beta_G)$, and $\alpha = (\alpha_F, \alpha_G)$. Then we can compute values of the variables

Algorithm 1 L_1 Regularized Nonnegative Sparse Coding Algorithm

1: **Input:** data matrix X, test sample y, $F = \phi$, $G = \{1, \ldots, n\}$, $\beta = \mathbf{0}$, and $\alpha = -X^T y$.
2: **Output:** sparse code β.
3: Normalize the columns of X and y to have unit l_2-norm.
4: Compute $r = \arg\min\{\alpha_i : i \in G\}$. If $\alpha_r < 0$, set $F = F \cup r$, $G = G - r$.
Otherwise stop: $\beta^* = \beta$ is the optimal solution.
5: Compute β_F^* by solving (10). If $\beta_F^* \geq 0$, set $\beta^t = (\beta_F^*, 0)$ and go to Step 4. Otherwise let r satisfy

$$\theta = \frac{-\beta_r}{\beta_r^* - \beta_r} = \min_i \left\{ \frac{-\beta_i}{\beta_i^* - \beta_i} : i \in F \text{ and } \beta_i^* < 0 \right\}$$

and set $\beta^t = ((1-\theta)\beta_F + \theta\beta_F^*,\ 0)$, $F = F - r$, $G = G \cup r$. Return to Step 5.
6: Compute α according to (11) and return to Step 4.

[2]Since the solution β is assumed to be sparse, we can use LCP to efficiently find a sparse active set.

TABLE I
RELATIONSHIP BETWEEN THE VALUE OF α_i AND THE DISTANCE FROM x_i TO y

	$-x_i^T y$	$x_i^T \hat{y}$ ($\hat{y} \doteq X_F \beta_F$)	$\alpha_i = x_i^T \hat{y} - x_i^T y$
Case 1	$\|\|x_i - y\|\|_2 \leq \|\|x_j - y\|\|_2$	$\|\|x_i - \hat{y}\|\|_2 \geq \|\|x_j - \hat{y}\|\|_2$	$\alpha_i \leq \alpha_j$
Case 2	$\|\|x_i - y\|\|_2 \leq \|\|x_j - y\|\|_2$	$\|\|x_i - \hat{y}\|\|_2 \leq \|\|x_j - \hat{y}\|\|_2$	$\alpha_i \leq \alpha_j$ or $\alpha_i \geq \alpha_j$
Case 3	$\|\|x_i - y\|\|_2 \geq \|\|x_j - y\|\|_2$	$\|\|x_i - \hat{y}\|\|_2 \geq \|\|x_j - \hat{y}\|\|_2$	$\alpha_i \leq \alpha_j$ or $\alpha_i \geq \alpha_j$
Case 4	$\|\|x_i - y\|\|_2 \geq \|\|x_j - y\|\|_2$	$\|\|x_i - \hat{y}\|\|_2 \leq \|\|x_j - \hat{y}\|\|_2$	$\alpha_i \geq \alpha_j$

β_F and α_G by the following iterative procedure:

$$\min_{\beta_F \in R^{|F|}} ||X_F \beta_F - y||_2^2 + \lambda \sum_{i \in F} \beta_i \tag{10}$$

$$\alpha_G = X_G^T (X_F \beta_F - y) + \tfrac{\lambda}{2}. \tag{11}$$

And the optimal solution is given by $\beta = (\beta_F, 0)$ and $\alpha = (0, \alpha_G)$. Algorithm 1 summarizes the optimal procedure.

B. Efficient Nonnegative Sparse Coding Algorithm

In sparse code algorithms for computer vision and pattern recognition [14], one often normalizes each column of the dataset X to have unit l_2-norm, which forms an L_1 ball to make the recovery of arbitrary corruption possible [14]. Another merit of this normalization step is to easily determine the L_1 regularization item λ. In this section, we study Algorithm 1 under the L_1 ball and develop an efficient algorithm.

In Algorithm 1, α_i controls the working set F. In each iteration, the index r corresponding to the minimum α_r is added to the working set F. Looking at (11), there are three parts in α_i. The first two parts of α_i are $x_i^T X_F \beta_F$ and $x_i^T y$, respectively. Here, we denote $X_F \beta_F$ by $\hat{y}$. Proposition 1 shows the relationship between the value of $x_i^T y$ and the value of the l_2 distance $||x_i - y||_2$. If $||x_i||_2^2 = 1$, $||x_j||_2^2 = 1$, and $x_i^T y \geq x_j^T y$, x_j will be far away from y than x_i (i.e., $||x_i - y||_2 \leq ||x_j - y||_2$). Based on Proposition 1, we categorize the relationship between the value of α_i and the distance from x_i to y into four cases in Table I.

Proposition 1: For $\forall\, x_i, x_j$, and y, if $||x_i||_2^2 = 1$, $||x_j||_2^2 = 1$, and $x_i^T y \geq x_j^T y$, then the inequality $||x_i - y||_2 \leq ||x_j - y||_2$ holds true.

Proof Sketch: Given that $||x_i||_2^2 = 1$, $||x_j||_2^2 = 1$, and $x_i^T y \geq x_j^T y$, we have $(x_i^T x_i - 2x_i^T y + y^T y) \leq (x_j^T x_j - 2x_j^T y + y^T y)$. Hence, $||x_i - y||_2^2 \leq ||x_j - y||_2^2$.

For Cases 1 and 4 in Table I, the λ in Algorithm 1 plays a role of a truncation function. Considering the inequality $\alpha_r < 0$ in Step 4 of algorithm and $\lambda > 0$, the inequality $x_r^T (X_F \beta_F - y) + (\lambda/2) < 0$ can be written as $x_r^T (X_F \beta_F - y) < -(\lambda/2)$. This means that there may be a sample x_i that corresponds to a large α_i value ($\alpha_i < 0$) and can further reduce the objective. But the nonnegative regularization item λ will restrict this sample from the working set F. In Cases 1 and 4, we learn if x_i is nearer to y than x_j, α_i will be smaller than α_j. Hence, λ plays a role of a truncation function and always removes faraway samples ($-(\lambda/2) \leq \alpha_i < 0$).

For Cases 2 and 3 in Table I, λ in Algorithm 1 plays the role of a discrimination function. In (11), there are two items that decide the value of α_i. The sample x_i that is near to the test sample y may have a large value α_i so that it does not satisfy the inequality in Step 4. However, α_j with respect to a faraway sample x_j can also have a small value. If a sample is redundant in the dataset X, it will be potentially restricted from the working set F. Hence, λ potentially makes Algorithm 1 compute a discriminate code.

Based on the above four cases, we learn that the L_1 regularizer λ plays an important role in finding a nonnegative sparse solution, and also plays a role of hard thresholding in [33]. If a dataset tends to be large, the solution computed by the nonnegative least squares technique in [18] and [21] without harnessing the L_1 regularization may not be sparse. However, according to Cases 1 and 4, we observe that the L_1 regularization item λ plays the role of a truncation function to remove faraway samples. This indicates that if the nonnegative least squares technique is used to compute a solution on the nearest dataset, the solution will be an approximation of nonnegative sparse solution. Since the datasets used in [18] and [21] are not very large and the dimension of dataset is often larger than the size of dataset, the nonnegative least squares techniques in [18] and [21] can find a nonnegative sparse solution. Hence, the necessary condition for nonnegative least squares technique to find a sparse solution is that the nearest dataset is used. Experimental results in Section V-G1 also confirm this finding.

If we directly make use of a truncation function to remove faraway samples, we will lose useful information in Cases 2 and 3. Fortunately, we aim to perform classification in which the data from one class are often assumed to be clustered. Hence, we can develop an efficient nonnegative sparse coding algorithm. First, we consider Cases 1 and 2 and compute a nonnegative sparse code from the nearest dataset. The nearest neighbor parameter n_{knn} is used to substitute the regularization λ. Second, based on the collaborative character [35] in sparse representation classification (i.e., sparse coding is performed collaboratively over relative datasets), we assume that only a test sample can be expressed by the datum from its relative and collaborative classes. Hence, we compute an informative and sparse code only from its relative classes.

IV. TSR

Motivated by the efficient nonnegative sparse coding algorithm, we present a TSR algorithm for robust face recognition in this section.

A. Learning a Robust Metric

In real-world face recognition, facial images are often corrupted by noise or outliers, that is, some pixels do not

TABLE II
POTENTIAL LOSS FUNCTIONS AND WEIGHTING FUNCTIONS

	Name	Loss Function $\phi()$	Weighting Function $\delta()$
ϕ_1	$L_{2,1}$	$\sqrt{\varepsilon + x^2}$	$1/\sqrt{\varepsilon + x^2}$
ϕ_2	L_1	$\lvert x\rvert$	$1/\lvert x\rvert$
ϕ_3	MCC	$1 - \exp(-(x^2/\varepsilon))$	$\exp(-(x^2/\varepsilon))$
ϕ_4	$\hat{L}_1$	$\begin{cases} x^2/2 & \lvert x\rvert \le \varepsilon \\ \varepsilon\lvert x\rvert - \frac{\varepsilon^2}{2} & \lvert x\rvert > \varepsilon \end{cases}$	$\begin{cases} 1 & \lvert x\rvert \le \varepsilon \\ \varepsilon/\lvert x\rvert & \lvert x\rvert > \varepsilon \end{cases}$

belong to facial images. One expects to learn a metric M through which outliers are efficiently detected and rejected so that classification algorithms can work on the uncorrupted subsets of pixels in facial images. Generally, M is assumed to be a diagonal matrix [36].

To deal with outliers and for the following recognition, we define the metric M as a function of a test sample y, a subspace $U \in \mathbb{R}^{d\times m}$ that models variation of the dataset X,[3] and a projection coefficient vector $\xi \in \mathbb{R}^{m\times 1}$, i.e., $M \triangleq M(U, y, \xi)$ where $M_{jj} \triangleq \phi(y_j - \sum_{i=1}^{m} U_{ij}\xi_i)$ and $\phi(x)$ is a robust function listed in Table II. One expects to find a metric that has the minimum matrix norm

$$M^* = \arg\min_{M(y,U,\xi)} ||M(y, U, \xi)||_1 \tag{12}$$

where $||M||_1 \triangleq \sum_{i=1}^{d}\sum_{j=1}^{d} |M_{ij}|$. Then we obtain the following optimization problem:

$$M^* = \arg\min_{M(y,U,\xi)} \sum_{j=1}^{d} \phi\left(y_j - \sum_{i=1}^{m} U_{ij}\xi_i\right). \tag{13}$$

The problem in (13) can be optimized in an alternative minimum way [23]

$$M_{jj}^t = \delta\left(y_j - \sum_{i=1}^{m} U_{ij}\xi_i^{t-1}\right) \tag{14}$$

$$\xi^t = \arg\min_{\xi} (y - U\xi)^T \left(M^t\right)(y - U\xi) \tag{15}$$

where $\delta(.)$ is the weighting function corresponding to a specific loss function $\phi(.)$. The optimization problem in (15) is a weighted linear regression problem, and its analytical solution can be directly computed by $\xi^t = (U^T M^t U)^{-1} U^T M^t y$.

Table II lists some potential loss functions ($\phi(.)$) and their corresponding weighting functions ($\delta(.)$). When substituting $\phi_2(.)$ into (13), the objective function of (13) is L_1-norm. Since the weighting function of L_1-norm is unpredictable around the origin, $\phi_1(.)$ and $\phi_4(.)$ are often used to replace L_1-norm. When ε tends to be zero, $\phi_1(.)$ and $\phi_4(.)$ become $\phi_2(.)$. When substituting $\phi_3(.)$ into (13), the optimization problem in (13) is actually a maximum correntropy (MCC) problem [23]. It has been shown in [23] that MCC can efficiently deal with nonGaussian noise and large outliers.

Fig. 1 further shows the shapes of different loss functions. The robust function $\phi_3(.)$ in MCC lies in the functions of $L_{2,1}$ and L_1 within the range $[-1, 1]$. Moreover, $\phi_3(.)$ gives outliers the same loss value (=1) regardless of whether where outliers are arbitrarily far away from the origin. In Fig. 1(b), we observe that the weighting function $\delta_3(.)$ corresponding to $\phi_3(.)$ assigns the outliers zero weight whereas other weighting functions do not. Hence, the robust function $\phi_3(.)$ in MCC can efficiently deal with outliers in the alternated minimization procedure.

[3]In this paper, the subspace U is composed of the eigenvectors computed by principal component analysis [3].

Algorithm 2 Learning a Robust Metric

1: **Function:** LRM(Subspace U, test data y, small positive value ε).
2: **Output:** M and ξ.
3: **repeat**
4: Initialize converged = FALSE.
5: Update M according to (14).
6: Update ξ according to (15).
7: **if** the variation of entropy is smaller than ε **then**
8: converged = TRUE.
9: **end if**
10: **until** converged == TRUE

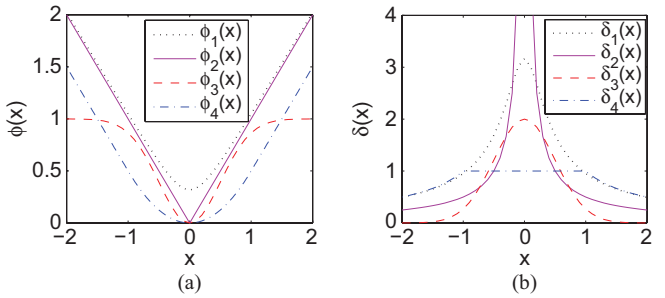

Fig. 1. Robust loss functions and their corresponding weighting functions. The analytic functions of $\phi(x)$ and $\delta(x)$ are listed in Table II. In $\phi_1(x)$, $\varepsilon = 0.1$. In $\phi_3(x)$, $\varepsilon = 0.5$. In $\phi_4(x)$, $\varepsilon = 1$. (a) Robust loss functions. (b) Weighting functions.

Algorithm 2 outlines the optimal procedure. According to the half-quadratic optimization [37], Algorithm 2 alternately minimizes the objective by (14) and (15) until Algorithm 2 converges. Since outliers are significantly far away from uncorrupted pixels, the M-estimator will punish the outliers during alternative minimum procedure. When the algorithm converges, we obtain a robust metric in which the diagonal entries corresponding to outliers will have small values.

1) Multisubspace Approach: In real-world face recognition, the errors incurred by face disguise are often not random. Face images are hardly occluded by a monkey image or corrupted by random noise. The errors are often due to the corruption of a large region of the human face, e.g., by a scarf. To deal with this real-world scenario, we develop a multisubspace approach in this section.

Fig. 2(e) shows an example of the robust metric learned by Algorithm 2. We observe that although Algorithm 2 accurately detects scarf occlusion, it also treats other important areas around two eyes as outliers. As a result, those areas have smaller values (darker pixels) in the learned metric. Looking at the iterative procedure in (14) and (15), we find that M_{jj}^t is

Fig. 2. Example of multisubspace when the number of subspaces is two. (a) 1-D of the subspace U learning by PCA. (The vector is reshaped to an image for show.) (b) and (c) Subspaces from U, and $\cup U^l = U$. (d) Input face with scarf occlusion. (e) Robust metric learned by Algorithm 2. (Darker pixels correspond to lower weights.) (f) Robust metric learned by Algorithm 3.

determined by the reconstructed data $\hat{y} = \sum U_{ij}\xi_i^{t-1}$. If $\hat{y}$ is well reconstructed without the affection of outliers, outliers can be accurately detected. However, since $\hat{y}$ is iteratively computed, it will be potentially affected by outliers especially when there are large corruptions.

To alleviate this problem, we extend Algorithm 2 to a multisubspace algorithm. The simple idea is to segment U into $\mathbf{n_u}$ subspaces with the same number of columns ($\cup U^l = U$) where $\mathbf{n_u}$ is the number of subspaces. Finally, we find a confidence subspace in which outliers only corrupt a small part and can be easily detected. Then an optimal linear coefficient ξ^* is learned on the confidence subset U^l, and used to reconstruct $\hat{y}$. The final robust metric can be calculated as follows:

$$M_{jj}^* = \phi\left(y_j - \sum_{i=1}^{m} U_{ij}\xi_i^*\right) \tag{16}$$

where M_{jj} denotes the jth diagonal element of subspace M. Assuming that M^l corresponds to U^l and is a subset of M that is computed on U, the confidence subset is determined as follows:

$$l^* = \arg\max_l \sum_j M_{jj}^l \tag{17}$$

where M_{jj}^l denotes the jth diagonal element of the lth subspace M. In other words, if a subset is less corrupted, it will receive larger weights than other subsets.

Algorithm 3 summarizes the procedure of the proposed multisubspace algorithm. In Step 3, we learn a metric M on the whole subspace and compute M^l for each U^l. In Step 4, we find the confidence subspace U^l according to (17). In Step 5, a confidence coefficient ξ^* is computed on the confidence subspace U^l. In Step 6, we obtain the robust learning metric according to (16). An intuitive way to determine the partition of the whole space U is based on the face structure. It is suggested to partition different facial organs into different subspaces. Fig. 2 shows an example of multisubspace when the number of partition $\mathbf{n_u} = 2$. Fig. 2 (f) shows an example of the robust metric learned by Algorithm 3. Compared with Algorithms 2 and 3 can detect the occlusion on the whole face more accurately.

B. Algorithm of Two-Stage Sparse Representation

Based on the analysis of the role of regularizer in Section III-B, we first filter the database into a small subset according to the nearest neighbor criterion in the learned robust metric. We denote the number of the nearest neighbors by n_{knn} and let $n_{\text{knn}} < \min(n, d)$ (The n_{knn} is the k in KNN.). The motivation of this filtering step lies on three folds: 1) it ensures that the following optimization problem in (18) has a unique solution; 2) it plays a similar role as λ in Lasso to remove some samples corresponding to small coefficients[4]; and 3) it significantly reduces computational costs so that TSR can deal with large-scale problems. The efficiency of this filtering step will be further corroborated in experiment Sections V-E and V-G1. Unless otherwise stated, we set n_{knn} to 300 throughout this paper.

Algorithm 3 Learning a Robust Metric via Multisubspaces

1: **Input:** Subspace U and $U^l(\bigcup_l U^l = U)$, a test visual data y, and a small positive value ε.
2: **Output:** M.
3: Set $M = LRM(U, y, \varepsilon)$.
4: Find the subspace U^l according to (17).
5: Set $\xi^* = LRM(U^l, y, \varepsilon)$.
6: Compute M according to (16).

Second, a nonnegative representation is computed on the subset by solving (7). For convenience, we set $\lambda = 0$ [21]. Then we have

$$\min ||y - X\beta||^2 \quad \text{s.t.} \quad \beta \geq 0. \tag{18}$$

If the matrix X has full column rank ($\text{rank}(X) = n$), the matrix $X^T X$ is positive definite so that the strictly convex program in (18) has unique solutions for each vector y [38]. Our implementation of (18) is based on an active set algorithm for linear programming [34], [38].

Inspired by the sparse classifier proposed in [14], we classify y as follows. For each class c, let $\delta_c : \mathbb{R}^n \to \mathbb{R}^{n_c}$ be a function, which selects the coefficients belonging to class c, i.e., $\delta_c(\beta) \in \mathbb{R}^{n_c}$ is a vector whose entries are the entries in β corresponding to class c. Utilizing only the coefficients associated to class c, the given sample y is reconstructed as $\hat{y}_c = X_c\delta_c(\beta)$. Then y is classified based on these reconstructions by assigning it to the class that minimizes the residual between y and $\hat{y}_c$

$$\min_c r_c(y) \doteq ||y - \hat{y}_c||_M \tag{19}$$

where $||.||_M$ is the correntropy induced diagonal metric in (13).

Algorithm 4 summarizes the procedure of the TSR. In Step 3, we make use of Algorithm 2 (or Algorithm 3) to learn a robust metric M. And then we set $\hat{X} = \sqrt{M}X$ and $\hat{y} = \sqrt{M}y$ so that we can perform classification under M in the following steps. In Step 4, to reduce computational costs, the large-scale dataset $\hat{X}$ is filtered into a small subset $\hat{X}^1$. In Step 5, a sparse representation is computed by NSR on the $\hat{X}^1$. In Step 6, considering that each object class often has several instances, i.e., $n_c \geq 1$, we select all the instances in the class corresponding to the nonzero coefficients in Step 6 so that we select the most competitive class. In Step 7, the query image y is classified according to the residuals.

[4]The active set algorithm [3], [8] to solve (21) selects the sample that can significantly reduce the objective step by step. The last selected samples often correspond to smallest coefficients and are far away from the query y.

Algorithm 4 TSR

1: **Input:** data matrix X, a test sample y, the number of the nearest neighbor n_{knn}.
2: **Output:** identity(y).
3: Compute a robust diagonal metric M according to Algorithm 2 (or Algorithm 3), and set $\hat{X} = \sqrt{M}X$ and $\hat{y} = \sqrt{M}y$.
4: Compute a nearest subset I^1 in $\hat{X}$ to $\hat{y}$ according to the nearest neighbor criterion, and set $\hat{X}^1 = \{x_i | i \in I^1\}$.
5: Solve the nonnegative least squares problem

$$\beta^* = \arg\min_\beta ||\hat{X}^1\beta - \hat{y}|| \quad \text{s.t. } \beta \geq 0. \tag{20}$$

6: Set $I^2 = \{i | \beta_i > 0 \text{ and } i \in I^1\}$ and $\hat{X}^2 = \{\hat{X}_c | \ \hat{x}_i \in \hat{X}_c \text{ and } i \in I^2\}$, solve the nonnegative problem

$$\beta^* = \arg\min_\beta ||\hat{X}^2\beta - \hat{y}|| \quad \text{s.t. } \beta \geq 0. \tag{21}$$

7: Compute the residuals $r_c(\hat{y}) = ||\hat{y} - \hat{X}^2\delta_c(\beta^*)||_2$, for $c = 1, \ldots, k$.
8: identity(y) $= \arg\min_c r_c(\hat{y})$.

Fig. 3. Samples of cropped faces in the AR database.

V. Experimental Verification

In this section, the proposed method is systematically compared with the state-of-the-art methods: reconstructive and discriminative subspace method (LDAonK) [3], linear regression classification (LRC) [39], sparse representation-based classification (SRC) [14], correntropy-based sparse representation (CESR) [18], and robust sparse coding (RSC) [20]. Experiments were performed on an AMD Quad-Core 1.80-GHz Windows XP machines with 2-GB memory.

A. Experimental Setting and Databases

Database: Two face databases and one large-scale database are selected to evaluate different methods. All the images are converted to grayscale and the facial image are aligned by fixing the locations of two eyes. The descriptions of four databases are as follows.

1) *AR Database [40]:* The AR database consists of over 4000 facial images of 126 subjects (70 men and 56 women). For each subject, 26 facial images are taken in two separate sessions. We select a subset of the dataset consisting of 65 male subjects and 54 female subjects. The images are cropped with dimension 112×92. Fig. 3 shows some cropped images in this dataset.
2) *Extended Yale B Database [41]:* The Extended Yale B database consists of 2414 frontal-face images of 38 subjects. The cropped 192×168 face images are captured under various laboratory-controlled lighting conditions and different facial expressions. For each subject, half of the images are randomly selected for training (i.e., about 32 images per subject), and the left half for testing. Fig. 4 shows some cropped images in this database.

Fig. 4. Samples of cropped faces in the extended Yale B database.

Fig. 5. Samples of cropped faces in the CAS-PEAL database. (a) Cropped facial images in the training set. (b) Occluded facial images in the testing set.

3) *PEAL Database [26]:* The CAS-PEAL database is a large-scale Chinese face database. For the training set, we select all frontal facial images that have undergone expression and lighting variations, where the pose degrees of all frontal facial images are less than or equal to 22°. Then the training set contains 7448 images of 1038 individuals. For the first testing set, we select 261 images of 261 individuals occluded by sunglasses in the accessory dataset. For the second testing set, we select 784 images of 262 individuals occluded by hat in the accessory dataset. The images are cropped with dimension 32×32. Fig. 5 shows some cropped images in this database.

Algorithm Setting: The details of compared techniques are as follows.

1) *SRC:* we compare its two robust models, which are different in the aspects of robustness and computational strategy [14].
SRC1: the implementation minimizes the L_1-norm in (22) via a primal-dual algorithm for linear programming based on [27][5]

$$\min ||\beta||_1 + ||e||_1 \quad \text{s.t.} \quad ||y - X\beta + e||_2 \leq \varepsilon \tag{22}$$

where ε is a given nonnegative error tolerance.
SRC2: the implementation minimizes the L_1-norm in (23) via an active set algorithm based on [42][6]

$$\min ||y - X\beta + e||_2 + \lambda(||\beta||_1 + ||e||_1) \tag{23}$$

where λ is a given sparsity penalty.
2) *TSR*[7]: The n_{knn} in Step 2 is set to 300. The subspace U in Algorithm 2 is composed of the eigenvectors corresponding to the five largest eigenvalues as suggested in [3] and [8]. Algorithm 2 is used to learn a robust metric.
3) *MTSR:* Algorithm 3 is used to learn a robust metric.
4) *TSR1:* TSR1 denotes the TSR without the filtering steps (Steps 4 and 5), i.e., the TSR1 directly computes a nonnegative sparse representation on the whole dataset

[5]The source code. Available at: http://www.acm.caltech.edu/l1magic/.
[6]The source code. Available at: http://redwood.berkeley.edu/bruno/sparsenet/.
[7]The source code. Available at: http://www.openpr.org.cn/index.php/All/63-Two-stage-Sparse-Representation/View-details.html.

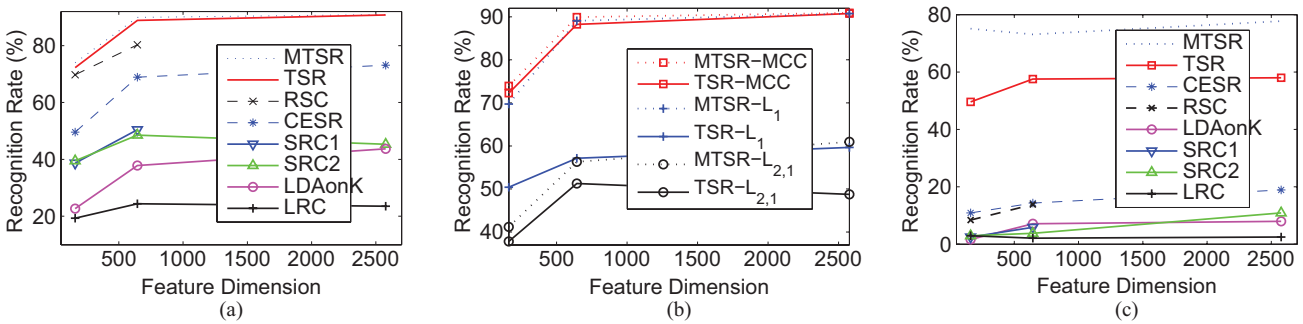

Fig. 6. Recognition rates for the AR dataset. (a) Recognition rates of different methods under sunglasses occlusion. (b) Recognition rates of different loss functions under sunglasses occlusion. (c) Recognition rates of different methods under scarf occlusion.

$\hat{X}$ via nonnegative least squares technique. Algorithm 2 is used to learn a robust metric.

5) *CESR:* Setting of the parameters of CESR follows the suggestion in [18].[8]
6) *RSC:* Setting of the parameters of RSC follows the suggestion in [20].[9]
7) *LDAonK:* Setting of the parameters of LDAonK follows the suggestion in [3].

In order to estimate the parameter ε of SRC1 and λ of SRC2, five-fold cross-validation on each dimension of data for each training set was used, where the candidate value set for both ε and λ is {1, 0.5, 0.25, 0.1, 0.075, 0.05, 0.025, 0.01, 0.005, 0.001, 0.0005, and 0.0001}. Note that due to large computational cost of SRC1, SRC2, and RSC, exhaustive search of the parameter value is not applicable. Also for the same reason, we can only report the experimental results of SRCs in the lower dimensional feature space. Since the multisubspace strategy in Algorithm 3 is developed only for real-world occlusion, we only report its results on real-world occlusion experiments.

B. Recognition Under Sunglasses Disguise

In this section, we evaluate the robustness of different methods and performance of different loss functions under sunglasses disguise. For training, we use 952 images (about eight for each subject) of un-occluded frontal views with varying facial expression. And for testing, we use images with sunglasses. Fig. 6(a) shows the recognition performance of different methods based on different downsampled images of dimensions 154, 644, and 2576 [14]. Those numbers correspond to downsampling ratios of 1/8, 1/4, and 1/2, respectively.

We observe from the numerical simulations that the methods can be ranked in descending recognition accuracy as MTSR, TSR, RSC, CESR, SRC1, SRC2, LDAonK, and LRC. The sparse representation-based methods outperform two nonsparse ones. If occlusions exist, it is unlikely that the test image will be very similar to the subspace of the same class, so that LRC performs poorly. In this case, TSR and RSC, CESR significantly perform better than two SRC methods. As in [20], iteratively reweighted methods seem to deal with outliers better than other methods. MTSR achieves the highest recognition rates. This is because the outlier detection stage of MTSR can efficiently detect the sunglasses occlusion. The MCC in MTSR is robust to nonGaussian noise [18].

Although TSR, CESR, and RSC are all based on M-estimators, they deal with outliers in different ways. Both CESR and RSC find robust weights and solve an L_1 minimization subproblem. Since RSC uses complexity strategy with a large computational cost to detect outliers, it may be more robust than CESR in some cases. Although both TSR and CESR are based on correntropy to detect outliers, TSR can deal with outliers better due to its independent outlier detection step. As a result, TSR outperforms CESR.

Fig. 6(b) further shows the recognition rates of TSR and MTSR based on different loss functions. We observe that the loss function will significantly affect the recognition accuracy. When $\hat{L}_1$ and $L_{2,1}$ loss functions are used, the recognition rates of TSR are close to those of SRC1 and SRC2, and are lower than those of RSC. These experimental results indicate that the accuracy of outlier detection depends on loss functions. Since MCC is robust to nonGaussian noise and large outliers, TSR based on MCC can accurately detect outliers and hence achieve highest recognition rates.

Fig. 6(b) also shows that the multisubspace strategy in MTSR can further improve recognition accuracy especially when using $\hat{L}_1$ and $L_{2,1}$ loss functions. We observe that the recognition rate of MTSR-$\hat{L}_1$ is close to those of TSR-MCC and MTSR-MCC. They all achieve their highest recognition rates when the dimension is 2576. It is due to that the weighting function of $\hat{L}_1$ is based on a threshold ε as shown in Fig. 1(b). If a given value is smaller than ε, its corresponding weight will be 1. The confidence subset makes multisubspaces algorithm find an accurate reconstruction of $\hat{y}$ and an appropriate value of ε such that the algorithm can accurately detect outliers. Hence, the recognition rate of MTSR-$\hat{L}_1$ is significantly higher than that of TSR-$\hat{L}_1$. In robust statistics, MCC can efficiently deal with large corruptions and nonGaussian noise such that both TSR-MCC and MTSR can accurately detect sunglasses occlusions and learn similar robust metrics in this experiment. Hence, it is possible for TSR-MCC to achieve higher recognition rates than MTSR-$\hat{L}_1$.

C. Recognition Under Scarf Occlusion

Face recognition under scarf occlusion can be used to evaluate the robustness of different methods. For sparse

[8]The source code. Available at: http://www.openpr.org.cn/index.php/All/69-CESR/View-details.html.

[9]The source code. Available at: http://www4.comp.polyu.edu.hk/~cslzhang.

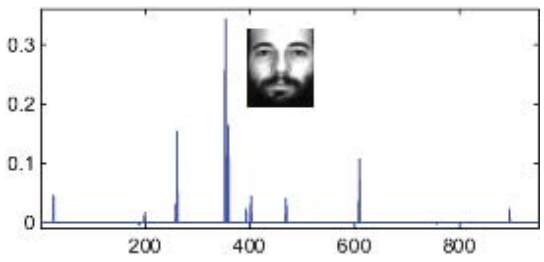

Fig. 7. Sparse representation learned by SRC1 under scarf occlusion.

representation methods, partition scheme is often used to tackle scarf occlusion [14], [18], [39]. In real-world problems, occlusion and corruption are often known such that we cannot make use of partition scheme in all scenarios. In this section, we evaluate the robustness of different methods under scarf occlusion on facial images. For training, we select 952 images (about eight for each subject) of un-occluded frontal views with varying facial expressions. For testing, we use images with scarf as shown in Fig. 2(d). Note that our purpose is to fairly evaluate different methods rather than achieve the highest recognition rate on this dataset.

Fig. 6(c) shows the recognition rates of eight compared methods. We observe that MTSR significantly outperforms other methods. Although RSC and CESR perform better than two SRC methods, they still fail to detect the errors incurred by scarf occlusion. The reason for the low recognition rates is that there is mustache on some facial images. Since the pixels in the mustache area are similar to the pixels corrupted by scarf occlusion, SRCs fail to detect the occluded area and always select the facial images with mustache as the most similar images. Fig. 7 shows such an example. As a result, the recognition rates of SRCs, RSC, and CESR are quite low. As discussed in Section V-G2, MTSR and TSR make use of PCA subspace in the outlier detection stage instead of using the whole training samples. This subspace can model facial variations and alleviate the affection of individual sample. Hence, MTSR and TSR accurately detect the corrupted region and achieve the highest recognition rates. This experiment also confirms that accurate outlier detection is important for robust learning methods.

D. Recognition Under Random Pixel Corruption

In some scenarios, the query image y can be partially corrupted [14], [18]. We testify the robustness of TSR on the extended Yale B face database. For each subject, half of the images are randomly selected for training, and the left half are for testing. The training and testing sets contain 1205 and 1209 images, respectively. All the images are downsampled to 48×42. Each test image is corrupted by replacing a percentage of randomly selected pixels with random pixel value, which follows a uniform distribution over $[0, 255]$. The percentage of corruption is from 10% to 80%. Since the recognition rates of different sparse representation methods are similar under random corruptions, we only report the results of SRC1 and SRC2 here.

Fig. 8 plots the recognition accuracy of five methods under different levels of corruptions. When the level of corruptions tends to be high, the three sparse methods begin to significantly outperform the two nonsparse ones. Recognition rates of the two SRC methods are very close. Finally, TSR performs slightly worse than the SRC methods when the level of corruptions is small. Although TSR only obtains similar recognition rates as compared with the SRC methods, it can significantly reduce computational cost.

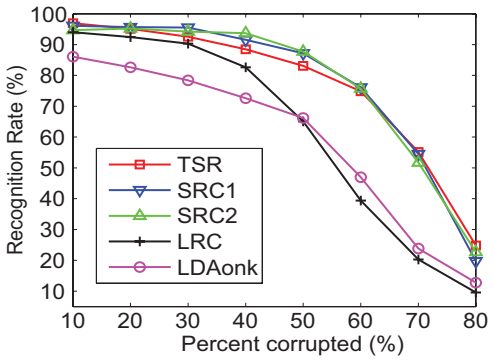

Fig. 8. Recognition rates under random pixel corruption.

TABLE III
RECOGNITION RATES (%) ON LARGE-SCALE PEAL DATASET

Algorithm	LRC	MTSR	TSR	TSR1
Glasses Occlusion	30.3	**44.0**	40.2	35.3
Hat Occlusion	23.1	**54.2**	53.8	45.2

E. Recognition on Large-Scale Dataset

In this section, we further evaluate the performance of TSR on a large-scale face database. We selected the large-scale PEAL dataset where there are 7448 facial images in the training set. Table III shows experimental results for the two testing sets. Since there are large variations of facial images in the training set and occlusions in the testing set, the recognition task is difficult. Hence, recognition rates of different methods are low.

For the first testing set of glasses occlusion, both MTSR and TSR perform better than LRC and TSR1. Although the robust metric in TSR1 is the same as that in TSR, TSR learns an approximation solution of nonnegative sparse representation. The coefficients learned by TSR are more informative than those learned by TSR1. As a result, there is a significant improvement in terms of recognition rate.

For the second testing set of hat occlusion, the proposed methods significantly outperform LRC. As expected, MTSR achieves the highest recognition rate. Looking at samples of images occluded by hat in Fig. 5, we observe that the hat occlusion incurs large variations of the pixels beyond two eyes. Although the occluded pixels are significantly different from those unconcluded pixels, they only corrupt small parts of one cropped face image. LRC fails to deal with these small occlusions such that it obtains a low recognition rate. However, TSR can accurately detect these occlusions and hence significantly outperforms LRC. Since the errors incurred by hat occlusion only occupy a small area of the whole face

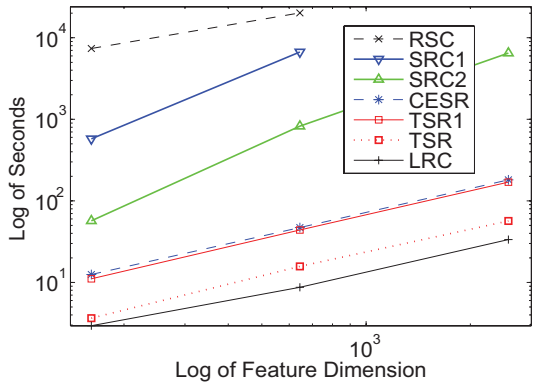

Fig. 9. Computation time (S) on various feature spaces on the AR dataset.

image, both MTSR and TSR can accurately detect the errors. As a result, the recognition rates of the two methods are similar.

F. Computational Costs

Computational complexity is an important issue for sparse representation methods. In this section, we compare the computation costs of different methods. The computational complexity of Algorithm 4 mainly depends on the KNN filtering step. Since the computational complexity of quick sort is $O(nlog(n))$, the computational complexity of Algorithm 4 is $O(nlog(n))$.

Fig. 9 shows the overall computation time for various features on AR database, with the same experiment setting as that in Section V-B. When the feature dimension is 644, RSC, SRC1, SRC2, CESR, TSR1, and TSR take 170, 56, 6.9, 0.40, 0.37, and 0.13 s for each test image, respectively. (In [14], SRC1 takes about 75 s per test image on a PowerMac G5.) All these compared methods can be ranked in terms of computational cost from large to small as follows. RSC, SRC1, SRC2, CESR, TSR1, TSR, and LRC. It is clear that TSR is better than SRC1, SRC2, and RSC due to less computational cost and better performance. Note that the computational cost is significantly reduced by over 50 times using TSR as compared with the SRC2. In addition, since RSC must solve an L_1 minimization problem in each reweighting iteration, RSC takes nearly k times longer than other SRC methods (k is the number of iterations).

It is interesting to observe that the computational cost of TSR1 and CESR is similar. This is because they both compute a nonnegative sparse representation on the whole dataset although they deal with outliers in different ways. This also indicates that the computational cost of solving an L_1 minimization is too large. Since TSR makes use of an approximation solution and works on a collaborative subset, TSR can further reduce computation time.

G. Parameter Selection

1) Parameter n_{knn} and Sparsity: The number of the nearest neighbor n_{knn} is a key parameter to control computational cost, recognition rates, and the number of nonzero coefficients (L_0 norm). In this section, we study how the n_{knn} affects the performance of TSR. The first experimental setting is the same as that of the AR dataset in Section V-B and the feature dimension is 644. The second experimental setting is the same as that on glasses occlusion in Section V-E.

TABLE IV
RECOGNITION RATES (%) AND THE TOTAL NUMBER OF NONZERO COEFFICIENTS (L_0 NORM) FOR VARIOUS NUMBER OF n_{knn} OF TSR ON THE AR DATASET

n_{knn}	100	200	300	400	500	952
Recognition Rate	84.0	84.0	86.6	87.4	**88.2**	86.6
L_0 Norm	19	22	23	24	25	28

TABLE V
RECOGNITION RATES (%) AND THE TOTAL NUMBER OF NONZERO COEFFICIENTS (L_0 NORM) FOR VARIOUS NUMBER OF n_{knn} OF TSR ON GLASSES OCCLUSION ON THE PEAL DATASET

n_{knn}	100	200	300	400	500	7448
Recognition Rate	39.1	39.5	**40.2**	39.5	39.1	35.3
L_0 Norm	29	31	33	34	35	87

Table IV shows recognition rates and the number of nonzero coefficients for various number of the nearest neighbors on the AR dataset. The total number of coefficients of SRC1 and SRC2 is 40 and 25, respectively. We observe from the numerical simulations that TSR can yield a sparse representation, and both recognition rates and the total number of nonzero coefficients increase as the number of the nearest neighbor increases. It is also interesting to observe that the recognition rate at 500 is even higher than that at 952. This coincides with the results in Section V-E.

Table V further shows recognition rates and the number of nonzero coefficients for different numbers of nearest neighbors on the PEAL glasses occlusion dataset. The highest recognition rate is achieved when n_{knn} is 300. We observe that only using part of the facial images to compute a sparse representation can yield a higher recognition rate. This means that n_{knn} plays a key role of L_1-regularization and affects both recognition rates and sparsity. Experimental results corroborate that the filtering large-scale database in TSR is reasonable.

Moreover, the experimental results demonstrate that even if a robust metric has been learned, a test image may not be very close to the same class in the training set. The filtering step (or L_1-regularization) is important for sparse representation-based methods to learn an informative and discriminative representation for robust face recognition.

2) Subspace Selection: In real-world face recognition, there are often glasses or mustache on registered facial images. Since those images cannot be treated as outliers, they potentially affect the recognition accuracy of sparse representation methods as discussed in Section V-C. Hence, the subspace selection is also an important issue for the outlier detection in TSR, which alleviates the affection of individual sample in training set. In this section, simulations are run to show how to determine an appropriate subspace for TSR. The setting is the same as that in Section V-B. The dimension of facial images is 644. The eigenvectors learned by principal component analysis (PCA) are used as the subspace U in TSR.

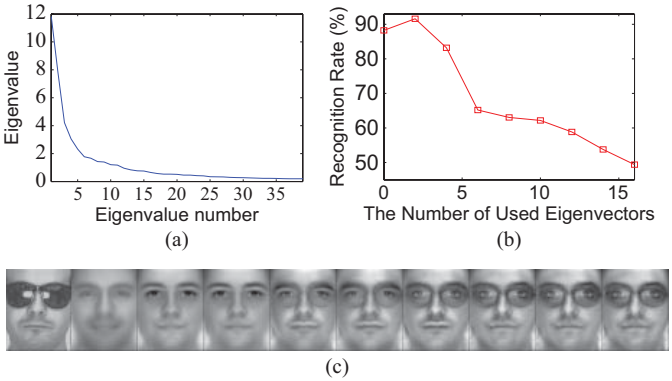

Fig. 10. Relationship between subspace selection and robustness. (a) Eigenvalues of the data matrix X calculated using the AR dataset. (b) Recognition rates as a function of the number of eigenvectors. (c) Reconstructed images by using different numbers of eigenvectors. The leftmost facial image is the input occluded image. The number of eigenvectors ranges from 1 to 41 at each five interval.

We plot the eigenvalues of the training data matrix X from the AR dataset in Fig. 10(a). We can see that the first few eigenvalues are significantly larger than the remainder. Given an occluded face image, we show the reconstructed images by using different numbers of eigenvectors in Fig. 10(c). The left facial image is the input occluded image and the number of eigenvectors ranges from 1 to 41 with interval 5. We observe that when the number of eigenvectors is larger than 16, there are obvious shadows around two eyes in the reconstructed face images. This means that although using more eigenvectors can achieve smaller reconstruction errors, the reconstructed images will be dominated by outliers (occluded pixels). Fig. 10(a) and (c) indicates that the first several eigenvectors dominate the major variations of training set and the remaining eigenvectors often affect local details.

Fig. 10(b) further shows recognition accuracy as a function of the number of eigenvectors used. The subspace U in learned robust metric is composed of different numbers of eigenvectors. The "0" indicates that we use mean vector as the input 1-D subspace U. We observe that recognition rates decrease significantly as more eigenvectors are used. This means that the outlier detection stage cannot accurately detect outliers. Hence, selection of an appropriate subspace to detect outliers is important for robust face recognition. From Fig. 10(a)–(c), it is suggested that the first several eigenvectors of a given training dataset is good enough for the outlier detection stage of TSR.

VI. Conclusion

In order to alleviate the high computational cost of sparse representation for large-scale face recognition, this paper decomposed the computation of a robust sparse representation into two stages. In the first stage, a general multisubspace framework was proposed to learn a robust metric in which noise and outliers (image pixels) were detected. In the second stage, based on the learned metric and collaborative representation, we proposed an efficient nonnegative sparse code algorithm to find an approximation solution of sparse representation. Then a filtering strategy was developed to avoid the computation of the sparse representation on the whole large-scale dataset. According to the L_1 ball theory, the approximated solution is unique and can be optimized efficiently. Theoretical analysis also showed that the nonnegative least squares technique only finds an approximate solution of nonnegative sparse representation technique. Extensive experiments demonstrated that the proposed framework not only significantly reduces computational costs but also can yield better results as compared to the state-of-the-art methods.

Experimental results also showed that robust algorithms will fail to detect the occluded region if the variations in the training set are partially similar to occlusions. Potential future work includes studying a robust model that can deal with errors in both training set and testing set. In addition, the analysis in Section IV-A showed that correntropy has similar properties with $L_{2,1}$-norm around the origin. Another future work is to make use of correntropy as an approximation of $L_{2,1}$-norm [43] to solve the unpredictable problem of $L_{2,1}$-norm in Fig. 1(b), and to study feature grouping [44].

Acknowledgment

The authors would like to thank the Associate Editor and the reviewers for their valuable comments and advice.

References

[1] 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.

[2] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 748–763, Jun. 2002.

[3] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 337–350, Mar. 2006.

[4] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.

[5] Z. Zhou, S. Chindaro, and F. Deravi, "A classification framework for large-scale face recognition systems," in *Proc. Int. Conf. Biometrics*, 2009, pp. 337–346.

[6] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," Vision & Modeling Group, Massachusetts Inst. Technology, Cambridge, Tech. Rep. 245, 1994.

[7] K. Ohba and K. Ikeuchi, "Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 9, pp. 1043–1047, Sep. 1997.

[8] M. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.

[9] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *Comput. Vis. Image Understand.*, vol. 78, no. 1, pp. 99–118, 2000.

[10] F. De la Torre and M. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, nos. 1–3, pp. 117–142, 2003.

[11] R. He, B.-G. Hu, and X. Yuan, "Robust discriminant analysis based on nonparametric maximum entropy," in *Proc. Asian Conf. Mach. Learn.*, Nanjing, China, 2009, pp. 120–134.

[12] H. Jia and A. M. Martinez, "Support vector machines in face recognition with occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2009, pp. 136–141.

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

[14] 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.

[15] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and S. Z. Li, "Ensemble-based discriminant learning with boosting for face recognition," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 166–178, Jan. 2006.

[16] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma, "Fast ℓ_1-minimization algorithms and an application in robust face recognition: A reivew," in *Proc. IEEE Int. Conf. Image Process.*, May 2010, pp. 1849–1852.

[17] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

[18] 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.

[19] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "A regularized correntropy framework for robust pattern recognition," *Neural Comput.*, vol. 23, no. 8, pp. 2074–2100, 2011.

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

[21] N. Vo, B. Moran, and S. Challa, "Nonnegative-least-square classifier for face recognition," in *Proc. Int. Symp. Neural Netw., Adv. Neural Netw.*, 2009, pp. 449–456.

[22] Y. Ji, T. Lin, and H. Zha, "Mahalanobis distance based non-negative sparse representation for face recognition," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2009, pp. 41–46.

[23] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.

[24] R. He, B. G. Hu, W. S. Zheng, and Y. Q. Guo, "Two-stage sparse representation for robust recognition on large-scale database," in *Proc. AAAI Conf. Artif. Intell.*, 2010, pp. 475–480.

[25] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Nonnegative sparse coding for discriminative semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2011, pp. 2849–2856.

[26] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man Cybern. A, Syst. Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[27] E. Candes and J. Romberg. (2005). *ℓ_1-Magic: Recovery of Sparse Signals via Convex Programming*. [Online]. Available: http://www.acm.caltech.edu/l1magic/

[28] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[30] D. Donoho and J. Tanner, "Sparse nonnegative solutions of underdetermined linear equations by linear programming," *Proc. Nat. Acad. Sci.*, vol. 102, no. 27, pp. 9446–9451, 2005.

[31] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4813–4820, Nov. 2008.

[32] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1087–1099, Jul. 2012.

[33] M. Slawski and M. Hein, "Sparse recovery by thresholded non-negative least squares," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2011.

[34] L. F. Portugal, J. J. Judice, and L. N. Vicente, "A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables," *Math. Comput.*, vol. 63, no. 208, pp. 625–643, 1994.

[35] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2011, pp. 1–8.

[36] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, vol. 15. Cambridge, MA: MIT Press, 2002, pp. 505–512.

[37] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.

[38] A. Bjorck, "A direct method for sparse least-squares problems with lower and upper bounds," *Numer. Math.*, vol. 54, no. 1, pp. 19–32, 1988.

[39] 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.

[40] A. M. Martinez and R. Benavente, "The AR face database," Computer Vision Center (CVC), Univ. Autònoma de Barcelona, Barcelona, Spain, Tech. Rep. 24, 1998.

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

[42] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*, vol. 19. Cambridge, MA: MIT Press, 2006, pp. 801–808.

[43] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[44] L. W. Zhong and J. T. Kwok, "Efficient sparse modeling with automatic feature grouping," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 9, pp. 1436–1447, Sep. 2012.

Ran He (M'10) received the B.S. degree in computer science from the Dalian University of Technology, Dalian, China, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009.

He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science. He serves as an Associate Editor of *Neurocomputing* (Elsevier). His current research interests include information theoretic learning and computer vision.

Wei-Shi Zheng (M'08) received the Ph.D. degree in applied mathematics with Sun Yat-Sen University, Guangzhou, China, 2008.

He was a Post-Doctoral Researcher on the European SAMURAI Research Project with the Queen Mary University of London, London, U.K. He joined Sun Yat-Sen University as an Associate Professor under the One-Hundred People Program of Sun Yat-Sen University. His current research interests include object association, categorization for visual surveillance, feature extraction, kernel methods, transfer learning, and face image analysis.

Bao-Gang Hu (M'94–SM'99) received the M.Sc. degree from the University of Science and Technology, Beijing, China, in 1983, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, in 1993, both in mechanical engineering.

He was a Research Engineer and Senior Research Engineer with C-CORE, Memorial University of Newfoundland, St. John's, NL, Canada, from 1994 to 1997. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing. From 2000 to 2005, he was the Chinese Director of LIAMA, the Chinese-French Joint Laboratory for Computer Science, Control and Applied Mathematics. His current research interests include intelligent systems, pattern recognition, and plant growth modeling.

Xiang-Wei Kong (M'04) received the B.E. and M.Sc. degrees from Harbin Shipbuilding Engineering Institute, Harbin, China, in 1985 and 1988, respectively, and the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2003.

She is a Professor with the Department of Electronic and Information Engineering, and the Vice Director of the Information Security Research Center, Dalian University of Technology. Her current research interests include multimedia security and forensics, digital image processing, and pattern recognition.