# Robust Semi-Supervised Subspace Clustering via Non-Negative Low-Rank Representation

Xiaozhao Fang, *Student Member, IEEE*, Yong Xu, *Senior Member, IEEE*,
Xuelong Li, *Fellow, IEEE*, Zhihui Lai, and Wai Keung Wong

***Abstract*—Low-rank representation (LRR) has been successfully applied in exploring the subspace structures of data. However, in previous LRR-based semi-supervised subspace clustering methods, the label information is not used to guide the affinity matrix construction so that the affinity matrix cannot deliver strong discriminant information. Moreover, these methods cannot guarantee an overall optimum since the affinity matrix construction and subspace clustering are often independent steps. In this paper, we propose a robust semi-supervised subspace clustering method based on non-negative LRR (NNLRR) to address these problems. By combining the LRR framework and the Gaussian fields and harmonic functions method in a single optimization problem, the supervision information is explicitly incorporated to guide the affinity matrix construction and the affinity matrix construction and subspace clustering are accomplished in one step to guarantee the overall optimum. The affinity matrix is obtained by seeking a non-negative low-rank matrix that represents each sample as a linear combination of others. We also explicitly impose the sparse constraint on the affinity matrix such that the affinity matrix obtained by NNLRR is non-negative low-rank and sparse. We introduce an efficient linearized alternating direction method with adaptive penalty to solve the corresponding optimization problem. Extensive experimental results demonstrate that NNLRR is effective in semi-supervised subspace clustering and robust to different types of noise than other state-of-the-art methods.**

***Index Terms*—Affinity matrix, low-rank representation (LRR), subspace clustering, supervision information.**

## I. Introduction

Subspace analysis is an important technology in signal processing and pattern recognition. An underlying assumption of subspace analysis is that the data often contain some types of structure [1]. Subspace has been successfully applied to different visual data such as face [2] and motion [3] for data visual analysis and clustering [4], [5]. Subspace methods can be roughly divided into three categories. The first one is unsupervised subspace learning and typical examples include principal component analysis (PCA) [6], clustering and projected clustering with adaptive neighbors (CAN) [7], locally line embedding [8], and locality preserving projections [9]. The second category is supervised subspace learning, in which the label information is used to capture discriminant feature representation. The most popular methods of this category are the well-known linear discriminant analysis (LDA) [10], 2-D LDA [11], marginal fisher analysis (12), and neighborhood minmax projections [13]. The third category is semi-supervised subspace learning [14]–[16], which utilizes relatively limited labeled data and sufficient unlabeled data to obtain the subspace.

Subspace clustering is an important clustering problem which attracts much attention in recent years. Generalized PCA (GPCA) is a typical subspace method for clustering data, which transforms the subspace clustering into the problem of how to fit the data with polynomials [17]. Sparse subspace clustering (SSC) method has been proposed to cluster datapoints that lie in a union of low-dimensional subspaces [18]. SSC can be used as spectral clustering method, which first learns an affinity matrix from the training datapoints and then obtains the final clustering results based on the constructed affinity matrix by using the corresponding clustering method such as normalized cuts [19] and Gaussian fields and harmonic functions (GFHF) [15]. Random sample consensus clusters the datapoints by modeling mixed data as a set of independent datapoints drawn from a mixture of probabilistic distributions [20]. Unfortunately, some existing subspace clustering methods (i.e., GPCA and matrix recovery [21]) assume that the data strictly drawn from a single subspace. However, in some real-world applications, the data cannot be characterized by a single subspace. Thus, it is reasonable to consider that the data are approximately drawn from a mixture of

Manuscript received February 2, 2015; revised April 11, 2015; accepted July 3, 2015. Date of publication August 3, 2015; date of current version July 15, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61370163, Grant 61233011, and Grant 61332011, in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20130329151843309, Grant JCYJ20130329151843309, Grant JCYJ20140417172417174, and Grant CXZZ20140904154910774, in part by the China Post-Doctoral Science Foundation under Project 2014M560264, and in part by the Shaanxi Key Innovation Team of Science and Technology under Grant 2012KCT-04. This paper was recommended by Associate Editor Y. Jin. *(Corresponding author: Yong Xu.)*

X. Fang and Z. Lai are with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China (e-mail: xzhfang168@126.com; lai_zhi_hui@163.com).

Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China, and also with the Key Laboratory of Network Oriented Intelligent Computation, Shenzhen 518055, Guangdong, China (e-mail: yongxu@ymail.com).

X. Li is with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an 710119, Shaanxi, China (e-mail: xuelong_li@opt.ac.cn).

W. K. Wong is with the Institute of Textiles & Clothing, Hong Kong Polytechnic University, Hong Kong, and also with the Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518055, China. (e-mail: calvin.wong@polyu.edu.hk).

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/TCYB.2015.2454521

several low-dimensional subspaces [1]. Recovering such subspace structure naturally imposes a challenging to the subspace clustering. With this view, given a set of datapoints, which may be corrupted by errors and approximately drawn from the subspaces, a good subspace clustering should try to correct the possible errors and at the same time to cluster data into their respective subspaces with each clustering corresponding to a subspace [1].

During the past two decades, a number of robust subspace clustering methods, which are mainly based on low-rank representation (LRR) [1] and sparse representation theories have been proposed. The well-known robust PCA (RPCA) [21], [22] can efficiently seize the low-dimensional subspace structure by seeking a low-rank component and an error component to approximate the original data. Latent LRR (LatLRR) [23] and its robust version (RobustLatLRR) [24] seamlessly integrate subspace clustering and feature exaction into a unified framework. LatLRR can learn two low-rank matrices one of which is used for robust subspace clustering and the other is able to robustly extract salient features from the observation data. Latent SSC [25] integrates dimensionality reduction and subspace clustering into a framework. The use of the projection can reduce the influence of noise to some extent. Non-negative low-rank and sparse (NNLRS) [26] graph for semi-supervised learning learns the weights of edges in graph by seeking a NNLRS matrix that represents each datapoint as a linear combination of others. The obtained graph structure can capture both global mixture of subspaces structure and locally linear structure of the data. Despite their great success based on LRR and sparse representation theories, these methods have two obvious disadvantages.

1) In most of these methods, robust subspace clustering can be performed by first learning an affinity matrix from the given data and then clustering the datapoints to respective subspaces by using the corresponding clustering methods. It is evident that these two steps are independent and thus an overall optimum cannot be guaranteed.
2) The labeled training samples are always insufficient due to the expensive labeling cost. In the LRR-based subspace clustering, the affinity matrix plays a significant role in exploiting the subspace structure. Thus, it is necessary to use the limited label information to guide the affinity matrix construction so that it can deliver strong discriminant information. However, in the conventional LRR-based subspace clustering methods, the label information is not used to guide the affinity matrix construction.

Inspired by the above insights, we propose a robust semi-supervised subspace clustering via non-negative LRR (NNLRR) method. By combining the LRR framework and the GFHF method in a single optimization problem, the supervision information is explicitly incorporated to guide the affinity matrix construction and the affinity matrix construction and subspace clustering are accomplished in one step to guarantee the overall optimum. More specifically, the affinity matrix is obtained by seeking a nonnegative low-rank matrix that represents each data sample as a linear combination of others. We also explicitly impose the sparse and non-negative constraints on the affinity matrix such that it is sparse and the elements in the affinity matrix can be directly used to cluster data. Benefiting from the breakthroughs in high-dimensional optimization [27]–[29], the optimization problem can be solved by convex relaxation. The convex optimization associated with the NNLRR model can be efficiently solved by the linearized alternating direction method with adaptive penalty (LADMAP) [26], [27], which can efficiently use less auxiliary matrix and matrix inversion [26] and thus it can effectively reduce the computation cost. We conduct extensive experiments of semi-supervised subspace clustering and the validity of NNLRR is demonstrated by the experiment results. In summary, the contributions of this paper includes the following.

1) The label information is explicitly incorporated to guide the affinity matrix construction so that the affinity matrix can effectively exploit the subspace structure of data. Thus the data can be accurately clustered to respective subspaces.
2) Unlike previous semi-supervised subspace clustering methods which separately treat the affinity matrix construction and clustering algorithm, NNLRR integrates these two tasks into one single optimization framework to guarantee the overall optimum.

The remaining of this paper is organized as follows. Section II briefly reviews some methods that are closely related to our method. Section III introduces the basis idea of NNLRR and some related discussions. Extensive experiments are conducted in Section IV. Finally, we conclude this paper in Section V.

## II. Related Works

Since the method proposed in this paper is based on LRR [1] and GFHF [15], we briefly review them to help reading this paper. Before delving in, we list some notations in the following. Sample set matrix is denoted as $X = [x_1, \ldots, x_u, x_{u+1}, \ldots, x_n] \in \Re^{m\times n}$, where $x_i|_{i=1}^{u}$ and $x_j|_{j=u+1}^{n}$ are the labeled and unlabeled samples, respectively. The labels of labeled samples are denoted as $y_i \in \{1, 2, \ldots, c\}$, where $c$ is the total number of classes. The label indicator binary matrix $Y \in \Re^{n\times c}$ is defined as follows: for each training sample $x_i (i = 1, 2, \ldots, n)$, $y_i \in \Re^{c}$ is its label vector. If $x_i$ is from the $k$th $(k = 1, 2, \ldots, c)$ class, then only the $k$th entry of $y_i$ is one and all the other entries are zero.

### A. Gaussian Fields and Harmonic Functions

GFHF is a well-known semi-supervised learning method, in which the predicted label matrix $F \in \Re^{n\times c}$ is estimated on the graph with respect to the label fitness and manifold smoothness [30]. Let us denote $F_i$ and $Y_i$ as the $i$th rows of $F$ and $Y$, respectively. GFHF minimizes the following objective function:

$$\min_{F} \frac{1}{2}\sum_{i,j=1}^{n}\left\|F_i - F_j\right\|^2 S_{ij} + \lambda_\infty \sum_{i=1}^{u}\|F_i - Y_i\|^2 \tag{1}$$

where $\lambda_\infty$ is a very large number such that $\sum_{i=1}^u \|F_i - Y_i\|^2 = 0$ is approximately satisfied and $F$ is the predicted labels for all the samples. $S^{n\times n}$ is the graph weight matrix which represents the similarity of a pair of training samples. As shown in [30], (1) can be reformulated as

$$\min_F \frac{1}{2}Tr\big(F^TLF\big) + Tr\big((F-Y)^TU(F-Y)\big) \qquad (2)$$

where $L \in \Re^{n\times n}$ is the graph Laplacian matrix and calculated as $L = D - S$, where $D_{ii} = \sum_j S_{ij}$ is a diagonal matrix. $U \in \Re^{n\times n}$ is also a diagonal matrix with the first $u$ and the rest $n-u$ diagonal elements as $\lambda\infty$ and 0, respectively.

### B. Low-Rank Representation

We assume that the observed data matrix $X \in \Re^{m\times n}$ is approximately drawn from a union of $c$ low-dimensional subspaces $\{\prod_i\}_{i=1}^c$ contaminated by error $E$. The objective function of LRR can be formulated as

$$\min_{Z,E} \quad \text{rank}(Z) + \gamma\|E\|_0, \qquad \text{s.t.}\ \ X = AZ + E \qquad (3)$$

where $\gamma$ is a parameter and $A$ is the dictionary that spans the union of subspace $\bigcup_{i=1}^c \prod_i$. Minimizer $Z$ is the lowest rankness representation of $Z$ with respect to the dictionary $A$. $E$ is the matrix that characterizes the error in the original $X$. $\|\cdot\|_0$ is the sparsity measure and is defined as the number of nonzero entries. Obviously, direct optimization of (3) is NP-hard [31]. Thus the optimization problem of (3) is relaxed into the following optimization problem:

$$\min_{Z,E} \|Z\|_* + \gamma\|E\|_1, \qquad \text{s.t.}\ \ X = AZ + E \qquad (4)$$

where $\|Z\|_*$ is the nuclear norm (i.e., the sum of the singular values) of $Z$ which can approximate the rank of $Z$. $\|E\|_1$ is a good relaxation of $\|E\|_0$. When $A = I$, LRR degenerates to RPCA, which is suitable for recovering a matrix drawn from a single subspace. Generally, dictionary $A$ is replaced by the original matrix $X$. Thus, (4) can be written as

$$\min_{Z,E} \|Z\|_* + \gamma\|E\|_1, \qquad \text{s.t.}\ \ X = XZ + E. \qquad (5)$$

When (5) is applied to subspace clustering, the obtained $Z$ is used to define an affinity matrix $(|Z| + |Z^T|)$, then the clustering results are obtained by applying the corresponding clustering method to the defined affinity matrix. In robust semi-supervised subspace clustering, the affinity matrix is constructed by (5), and then the clustering algorithm such as GFHF is directly applied to the constructed affinity matrix.

## III. NNLRR

### A. Motivations of NNLRR

In this paper, we focus on robust semi-supervised subspace clustering. How to reasonably exploit the limited label information and to ensure the algorithmic overall optimum are two important issues in machine learning and computer vision fields. The label information is very effective to improve the discriminant ability of the affinity matrix [30]. However, as shown in Section II-B, in LRR, it is evident that the limited label information is not exploited to guide the affinity matrix construction. Moreover, in LRR, the affinity matrix is first constructed and then the clustering algorithm is applied to the constructed affinity matrix. Such independent two steps cannot guarantee the overall optimum.

GFHF uses a concise way to incorporate the label information into semi-supervised leaning, which provides a feasible solution to address these two problems. Specifically, we integrate LRR and GFHF into a unified framework so that the label information can be introduced to guide the affinity matrix construction and the affinity matrix construction and subspace clustering are simultaneously performed in one step in order to guarantee the overall optimum.

### B. Model of NNLRR

The problem at hand is to design a compact model that naturally unifies LRR and GFHF. Our idea is to simultaneously perform the affinity matrix construction and semi-supervised subspace clustering. Therefore, the model of NNLRR is given as follows:

$$\begin{aligned}\min_{F,Z,E} \quad & \sum_{i=1}^n\sum_{j=1}^n \|F_i - F_j\|^2 Z_{ij} + Tr\big((F-Y)^TU(F-Y)\big)\\ & + \gamma\|Z\|_* + \beta\|E\|_{2,1}\\ \text{s.t.} \quad & X = AZ + E, \quad Z \geq 0, \quad \|Z\|_0 \leq T \end{aligned} \qquad (6)$$

where $\|E\|_{2,1} = \sum_{j=1}^n \sqrt{\sum_{i=1}^m (E_i^j)^2}$ is the $\ell_{2,1}$-norm of $E$ [32], $E_i^j$ is the $(i,j)$th entry of $E$. $\gamma$ and $\beta$ are the parameters to balance the importance of the corresponding terms. The other variables have the same definitions as (2) and (4). The nonnegative sparse constraint ($Z \geq 0$ and $\|Z\|_0 \leq T$) is to ensure that the obtained low-rank and sparse matrix can be directly used as the affinity matrix. $\|E\|_{2,1}$ encourages the columns in $E$ to be zero, which assumes the error is "sample-specific," i.e., some samples are corrupted and the others are clean [1]. There of course are many choices to match the error term, such as $\|E\|_F^2$ for the small Gaussian noise and $\|E\|_1$ for the random corruption. In this paper, we focus on the $\ell_{2,1}$-norm. The first term evaluates the label fitness. The role of the second term is the same as that in (2). The third term ensures that $Z$ can capture the global mixture of subspaces via low-rank constraint. The fourth term tries to fit the error in the original data. After we obtain the minimizer $(Z^*, E^*)$, $AZ^*$ (or $X - E^*$) can be used to obtain a low-rank recovery of matrix $X$. The obtained $F$ can be directly used to perform semi-supervised subspace clustering by using the following way. If $h = \arg\max_\kappa F_i^\kappa (i = u+1, \ldots, n; \kappa = 1, \ldots, c)$, then the $i$th sample is assigned to the $h$th class, where $F_i^\kappa$ denotes the $(i,\kappa)$th entry of $F$.

### C. Solution to NNLRR

The difficulty to solve the NNLRR problem (6) is that there are three terms closely related to $Z$. A feasible way is to make the objective function (6) separable. To this end, an auxiliary matrix $W$ is introduced into (6). We first convert (6) to the

**Algorithm 1** : Solving NNLRR by LADMAP

```
Input: Data set matrix $X$; Label indicator matrix $Y$;
Matrix $U$; Parameters $\gamma$ and $\beta$;
Initialization: $Z_0 = W_0 = \mathbf{O}$; $E_0 = \mathbf{O}$; $F_0 = \mathbf{O}$; $Y_{1,0} = \mathbf{O}$;
$Y_{2,0} = \mathbf{O}$; $\mu_0 = 0.1$, $\mu_{max} = 10^7$, $\rho_0 = 1.01$, $\gamma = 1$,
$\epsilon_1 = 10^{-7}$, $\epsilon_2 = 10^{-6}$, $\theta = \|A\|_F^2$, $k = 0$;
while not converged do
    1. Fix the others and update $Z$ by solving (9)
    2. Fix the others and update $F$ by solving (10).
    3. Fix the others and update $E$ by solving (11).
    4. Fix the others and update $W$ by solving (12).
    5. Update the multipliers as follows:
    $\begin{cases} Y_1^{k+1} \leftarrow Y_1^k + \mu^k(X - AZ^k - E^k) \\ Y_2^{k+1} \leftarrow Y_2^k + \mu^k(Z^k - W^k) \end{cases}$
    6. Update the parameter $\mu$ follows:
    $\mu^{k+1} = min(\mu_{max}, \rho\mu^k)$, where
    $\rho = \begin{cases} \rho_0 & \text{if } \mu^k\Omega/\|X\|_F \le \epsilon_2 \\ 1 & \text{otherwise} \end{cases}$
    7. Check the convergence conditions
    $\begin{cases} \|X - AZ^k - E^k\|_F/\|X\|_F \le \epsilon_1 \quad or \\ \mu^k\Omega/\|X\|_F \le \epsilon_2 \end{cases}$
    where $\Omega = \max\Big(\sqrt{\theta}\|Z^k - Z^{k+1}\|_F, \|W^k - W^{k+1}\|_F,$
    $\|E^k - E^{k+1}\|_F, \|F^k - F^{k+1}\|_F\Big)$
    8. Update $k$: $k \leftarrow k + 1$.
end while
Output: $F$, $Z$, $E$.
```

following equivalent problem:

$$\min_{F,Z,E,W} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|F_i - F_j\right\|^2 W_{i,j} + Tr\big((F - Y)^T U(F - Y)\big) + \gamma\|Z\|_* + \beta\|E\|_{2,1}$$
$$\text{s.t.} \quad X = AZ + E, \quad Z = W, \quad W \ge 0, \quad \|W\|_0 \le T. \tag{7}$$

This problem can be solved by the LADMAP which can use less auxiliary matrix and matrix inversion ($Z$ is not replaced by another auxiliary matrix), hence computation cost can be reduced. Equation (7) can be converted into the following augmented Lagrangian function:

$$\begin{aligned}
&\Gamma(Z, W, F, E, Y_1, Y_2, \mu) \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|F_i - F_j\right\|^2 W_{i,j} + Tr\big((F - Y)^T U(F - Y)\big) \\
&\qquad + \gamma\|Z\|_* + \beta\|E\|_{2,1} \\
&\qquad + \langle Y_1, X - AZ - E\rangle + \langle Y_2, Z - W\rangle \\
&\qquad + \frac{\mu}{2}\Big(\|X - AZ - E\|_F^2 + \|Z - W\|_F^2\Big) \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|F_i - F_j\right\|^2 W_{i,j} + Tr\big((F - Y)^T U(F - Y)\big) \\
&\qquad + \gamma\|Z\|_* + \beta\|E\|_{2,1} \\
&\qquad + \psi(Z, W, E, Y_1, Y_2, \mu) - \frac{1}{2\mu}\Big(\|Y_1\|_F^2 + \|Y_2\|_F^2\Big) \\
&\text{s.t.} \quad W \ge 0
\end{aligned} \tag{8}$$

where $\psi(Z, W, E, Y_1, Y_2, \mu) = \mu/2(\|X - AZ - E + (Y_1/\mu)\|_F^2 + \|Z - W + (Y_2/\mu)\|_F^2)$ and $\langle A, B\rangle = Tr(A^T B)$. $Y_1$ and $Y_2$ are Lagrange multipliers and $\mu \ge 0$ is a penalty parameter. The LADMAP updates the variables $Z, W, E$, and $F$ alternately, by minimizing $\Gamma$ with other variables fixed and then updates $Y_1$ and $Y_2$. With some algebra, the updating scheme can be designed as follows:

$$\begin{aligned}
Z^{k+1} &= \arg\min_Z \gamma\|Z\|_* \\
&\quad + \Big\langle \nabla_Z\varphi\Big(Z^k, W^k, E^k, Y_1^k, Y_2^k, \mu^k\Big), Z - Z^k \Big\rangle \\
&\quad + \frac{\mu^k\theta}{2}\left\|Z - Z^k\right\|_F^2 \\
&= \arg\min_Z \gamma\|Z\|_* + \frac{\mu^k\theta}{2}\left\| Z - Z^k + \frac{\left[-X^T\left(X - AZ^k - E^k + \frac{Y_1^k}{\mu^k}\right) + \left(Z^k - W^k + \frac{Y_2^k}{\mu^k}\right)\right]}{\theta} \right\|_F^2
\end{aligned} \tag{9}$$

$$\begin{aligned}
F^{k+1} &= \arg\min_F \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|F_i - F_j\right\|^2 W_{ij}^k \\
&\quad + Tr\Big((F - Y)^T U(F - Y)\Big) \\
&= \arg\min_F Tr\Big(F^T L^k F\Big) + Tr\Big((F - Y)^T U(F - Y)\Big)
\end{aligned} \tag{10}$$

$$E^{k+1} = \arg\min_E \beta\|E\|_{2,1} + \frac{\mu^k}{2}\left\|X - AZ^{k+1} + \frac{Y_1^k}{\mu^k} - E\right\|_F^2 \tag{11}$$

$$\begin{aligned}
W^{k+1} &= \arg\min_{W \ge 0, \|W\|_0 \le T} Tr(\Theta(R \odot W)) \\
&\quad + \frac{\mu^k}{2}\left\|W - \left(Z^{k+1} + \frac{Y_2^k}{\mu^k}\right)\right\|_F^2
\end{aligned} \tag{12}$$

where $\nabla_Z\varphi$ is the partial differential of $\varphi$ with respect to $Z$, $\theta = \|A\|_F^2$ in (9). $L \in \Re^{n\times n}$ in (10) is the graph Laplacian matrix and calculated as $L^k = D^k - W^k$, where $D_{ii}^k = \sum_j W_{ij}^k$ is the diagonal matrix. In (12), $R_{ij} = \frac{1}{2}\|F_i^{k+1} - F_j^{k+1}\|^2$, $\odot$ is the Hadamard product operator of matrices and $\Theta$ is a matrix with all elements are ones. The complete algorithm is outlined in Algorithm 1. Note that steps 1–3 of Algorithm 1 are convex problems and both have closed form solutions. Step 1 is solved via the singular value thresholding [33], it can be computed in the closed form

$$Z^{k+1} = \jmath_{\frac{\gamma}{\theta\mu^k}}\left( Z^k - \frac{\left[-X^T\left(X - AZ^k - E^k + \frac{Y_1^k}{\mu^k}\right) + \left(Z^k - W^k + \frac{Y_2^k}{\mu^k}\right)\right]}{\theta} \right) \tag{13}$$

where $\jmath$ is the thresholding operator with respect to the singular value $\gamma/(\theta\mu^k)$. It can be found that $Z$ is solved by a proximal method.

For the problem in step 2, it is straightforward to set the derivative of (10) with respect to $F$ to zero, namely

$$\frac{\partial\left(Tr\left(F^T L^k F\right) + Tr\left((F-Y)^T U(F-Y)\right)\right)}{\partial F} = 0 \tag{14}$$

then, we have

$$F^{k+1} = \left(L^k + U\right)^{-1} UY. \tag{15}$$

Step 3 is solved by via the following Lemma 1.

*Lemma 1 [1], [34]:* Let $Q$ be a given matrix. If the optimal solution to

$$\min_P \alpha\|P\|_{2,1} + \frac{1}{2}\|Q-P\|_F^2 \tag{16}$$

is $P^*$, then the $i$th column of $P^*$ is

$$P_i^* = \begin{cases} \frac{\|P_i\|_2-\alpha}{\|P_i\|_2} Q_i & \text{if } \|Q_i\|_2 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where $P_i$ and $Q_i$ are the $i$th columns of matrices $P$ and $Q$, respectively.

To ensure the solution of determining $W$ is sparse, we impose constraint of $\|W_i\|_0 \leq T$ on $W$. We decompose problem (12) into $n$ independent subproblems each of which can be formulated as a weighted nonnegative sparse coding problem, namely

$$\begin{aligned} \min_{W_i} \quad & W_i \odot R_i + \frac{\mu^k}{2}\left\| W_i - \left(Z^{k+1} + \frac{Y_2^k}{\mu^k}\right)_i \right\|_2^2 \\ \text{s.t.} \quad & W_i \geq 0, \|W_i\|_0 \leq T \end{aligned} \tag{18}$$

where $W_i$ and $R_i$ are the $i$th ($i = 1, 2, \ldots, n$) columns of matrices $W$ and $R$, respectively. And (18) has a closed form solution [35], [36].

### *D. Discussion*

In essence, the goal of our method is to estimate a function $F$ on a NNLRS graph. The integration of the affinity matrix learning and semi-supervised clustering can guarantee that the estimated function $F$ and the learned affinity matrix are perfectly matched, i.e., the algorithmic optimum can be guaranteed. In addition, the label information of labeled samples can enable the learned affinity matrix to have strong discriminant ability. Generally, our method is based on two basic assumptions: 1) local consistency and 2) manifold assumptions. The former implies that nearby samples are likely to have the same label, whereas the latter says samples lying in the same manifold tend to have the same label. In our method, we use a NNLRS graph (affinity matrix) to approximate the underlying manifold and simultaneously propagate labels to unlabeled samples along the learned graph.

A vector $F_i \in F$ ($i = 1, 2, \ldots, n$) corresponds to a classification function. $\forall F_i \in F$ assigns $c$ real values to sample $x_i$ and the maximal value of $c$ real values denotes the class of sample $x_i$ belongs. To find the optimal vector $F_i$ to accurately classify sample $x_i$, the objective function of GFHF [30] is used as the cost function. The first term of (6) is the smoothness cost, which means that a good classification function not changes too much between nearly samples. In other words, samples that are close nearby tend to have the nearly same labels. By using the constraint of $X = AZ + E$, the nearby samples are selected to reconstruct the original samples. The second term of (6) means a good classification function should not change too much from the labels of the labeled samples. Note that this term is only used on the labeled samples. The goal of the third term of (6) is to enforce $Z$ to have block-wise structure, which explicitly represents the neighborhood to neighborhood reconstruction. That is to say that the obtained affinity matrix $Z$ can better characterize the similarity of samples and thus the label information can be accurately propagated by the learned graph. The goal of the fourth term of (6) is to filter out the noisy information.

### *E. Difference From NNLRS [26]*

To our best knowledge, NNLRS is the most similar one to ours. NNLRS is the originally designed for the semi-supervised clustering problem by using the affinity matrix, namely, NNLRS solves the following problem:

$$\begin{aligned} \min_{Z,E} \quad & \|Z\|_* + \beta\|Z\|_1 + \lambda\|E\|_{2,1} \\ \text{s.t.} \quad & X = AZ + E, \quad Z \geq 0. \end{aligned} \tag{19}$$

Once obtaining the optimal $Z^*$, the column vectors of $Z^*$ are normalized by $z_i^* = z_i^* / \|z_i\|_2^*$ and the elements in each column vector are pruned by a predefined threshold $\varrho$

$$z_{ij}^* = \begin{cases} z_{ij}^*, & \text{if } \ z_{ij}^* \geq \varrho \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Although we use the affinity matrix to perform the semi-supervised clustering, our NNLRR is quite different from NNLRS in three aspects.

1) In NNLRS, the NNLRS graph (the affinity matrix) is firstly learned and then the clustering algorithms are performed on the learned graph. In contrast, by integrating GFHF and LRR into a unified framework, the affinity matrix learning and semi-supervised clustering are simultaneously accomplished in a step. Such integration can guarantee the overall optimum.
2) In NNLRS, the limited label information is not exploited to guide the affinity matrix construction, while in our NNLRR, the label information is used to endow the affinity matrix with strong discriminant ability.
3) In NNLRS, to obtain the optimal graph, the learned affinity matrix needs to prune, i.e., some elements of the affinity matrix should be set to 0 by a given threshold value $\varrho$. However, in practice, how to estimate the optimal threshold value is dataset dependent. In NNLRR, by integrating GFHF and LRR, the learned affinity matrix is optimal without any extra operation.

### *F. Connection Between NNLRR and Other Methods*

*1) Connection Between NNLRR and GHFH [15]:* If we set $U \neq 0$, $\gamma = 0$, and $\beta \to \infty$, then the objective function of

NNLRR in (6) reduces the following problem:

$$\min_{F,Z} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|F_i - F_j\right\|^2 Z_{ij} + Tr\big((F-Y)^T U(F-Y)\big)$$
$$\text{s.t.} \quad X = AZ, \quad Z \geq 0, \quad \|Z\|_0 \leq T \qquad (21)$$

which can be written as

$$\min_{F,Z} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|F_i - F_j\right\|^2 Z_{ij} + Tr\big((F-Y)^T U(F-Y)\big) + \nu\|X - AZ\|_F^2,$$
$$\text{s.t.} \quad Z \geq 0, \quad \|Z\|_0 \leq T \qquad (22)$$

where $\nu$ is the parameter to balance the different terms. If we further set $\nu = 0$ and discard the constraints, then (22) becomes the objective function of GFHF.

*2) Connection Between NNLRR and CAN [7]:* If we set $U = 0$, $\gamma = 0$, and $\beta \to \infty$, then (6) reduces

$$\min_{F,Z} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|F_i - F_j\right\|^2 Z_{ij} + \nu\|X - AZ\|_F^2$$
$$\text{s.t.} \quad Z \geq 0, \quad \|Z\|_0 \leq T \qquad (23)$$

which can be somewhat seen as the formulation of CAN in the purpose of assigning the adaptive neighbors for each sample without using the constraints $Z \geq 0$ and $\|Z\|_0 \leq$ T. Please note that in CAN the affine constraint and rank constraint are, respectively, imposed on $Z$ and $L$ (graph Laplacian matrix) to ensure that the connected components in the resulted affinity matrix are exactly equal to the cluster number [7] and the orthogonal constraint $F^T F = I$ is used to avoid a trivial solution.

*3) Semi-Supervised CAN:* If we set $U \neq 0$, $\gamma = 0$, and $\beta \to \infty$. then (6) reduces (22) which is a formulation of semi-supervised CAN by integrating GHFH and sample reconstruction into a unified framework. Such integration guarantees that for each sample the neighbor assignment is an adaptive process.

### *G. Convergence and Complexity Analyses*

The convergence of the exact augmented lagrange multipliers algorithm has been proven in the condition that the objective function is smooth [37]. It is very difficult to prove the convergence of the more complicated LADMAP. In RPCA [22], the convergence of inexact augmented lagrange multipliers has been proven in which two variable matrices were iterated alternately. The proposed NNLRR in this paper involves four iterating variable matrices $(Z, W, E, F)$ and the objective of the optimization problem (6) is not smooth. Thus it is not easy to prove the convergence in theory. According to the theoretical results in [1], three conditions are sufficient (but may not necessary) for Algorithm 1 that has a good convergence properties which are as follows.

1) The parameter $\mu$ in step 6 is needed to be upper bounded.
2) The so-called dictionary $A$ ($A = X$ in our method) is of full column rank.
3) In each iteration step, the residual produced by $\eta = \|(Z^k, F^k, W^k) - (Z, F, W)\|_F^2$ is monotonically decreasing, where $Z^k$ and $W^k$, respectively, denote the solution produced at the $k$th iteration and $(Z, F, W) = \arg\min \Gamma$ whose value is more than that of $(Z^k, F^k, W^k)$.

It has been shown in [1] that the above conditions can be approximately satisfied. Condition 1 is easy to be guaranteed by step 6 in the proposed method. The dictionary $A$ can be substituted by the orthogonal basis of this dictionary in practice and thus condition 2 is easy to obey. As discussed in [1] and [38], the convexity of the Lagrangian function could guarantee condition 3 satisfied to some extent, although it is not easy to strictly prove the monotonically decreasing condition. Therefore, Algorithm 1 can be expected to have good convergence properties.

Generally speaking, the major computation burden of NNLRR lies in step 1 since it involves the singular value decomposition (SVD). Specifically, in step 1, the SVD is operated on an $n \times n$ matrix, which is time consuming if the number of samples (i.e., $n$) is very large. As it is referred in [1], by substituting $A$ with the orthogonal basis of the dictionary, the computation complexity of step 1 is $\mathcal{O}(nr_A^2)$, where $r_A$ is the rank of the dictionary $A$. The computation complexity of step 2 is about $\mathcal{O}(n^3)$. The computation complexity of step 3 is about $\mathcal{O}(n^2 r_A)$. The computation complexity of step 4 is trivial owing its simple closed solution. Thus, the computation complexity of NNLRR is $\mathcal{O}(\tau(nr_A^2 + n^3 + n^2 r_A))$ in general, where $\tau$ is the number of iterations. The iteration number $\tau$ depends on the choice of $\rho$; $\tau$ is small while $\rho$ is large, and vice versa.

## IV. Experiments

In the experiments, the proposed NNLRR method will be tested using four datasets: 1) Yale [9]; 2) AR [2]; 3) Extended YaleB [23]; and 4) COIL20 [30]. The compared methods include LRR [1], SSC [18], LatLRR [23], RobustLatLRR [24], Local subspace Analysis (LSA) [39], and NNLRS [26]. Since the affinity matrix in NNLRR is nonnegative, we also compare the performance of NNLRR and the non-negative sparse graph (SPG) [40] in terms of semi-supervised clustering performance. For the sake of fair comparison, apart from NNLRR, all the other methods firstly construct an affinity matrix by respective corresponding technique and then GFHF is performed on the constructed affinity matrix to obtain the semi-supervised clustering results. In addition, we modify all of the compared methods to the same noise norm, i.e., the $\ell_{2,1}$-norm for fair comparison. For each dataset, we randomly select different samples from per subject as labeled samples and used the remaining as unlabeled samples and all experiments are run five times (unless otherwise stated) and then the mean classification result and standard deviation (%) are reported. The parameters of all these methods are carefully adjusted in order to obtain the best clustering results. From the experiment, we can find the performance of NNLRR is robust to parameter $\gamma$ (see Fig. 3). Thus, we set $\gamma = 1$ in all the experiments which can guarantee a satisfactory result. Better results may be achieved with tuning it. Parameter $\beta$ of NNLRR controls the

TABLE I
CLUSTERING RESULTS (%) ON THE YALE DATASET. NOTE THAT # TR DENOTES THE NUMBER OF LABELED SAMPLES OF ONE SUBJECT

| # Tr. | LRR | LatLRR | RobustLatLRR | SPG | LSA | SSC | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|---|
| 2 | 53.43±3.26 | 56.24±4.23 | 55.09±3.25 | 51.25±2.36 | 51.23±4.20 | 55.85±3.65 | 56.66±3.51 | **58.29±3.14** |
| 3 | 59.58±3.90 | 62.52±4.95 | 64.80±3.30 | 55.16±4.27 | 55.78±2.82 | 59.33±3.34 | 67.00±4.10 | **67.76±3.60** |
| 4 | 63.67±2.84 | 66.67±3.13 | 67.50±4.00 | 58.85±2.96 | 58.90±2.46 | 69.95±3.18 | 71.83±3.56 | **73.76±3.40** |
| 5 | 67.33±2.25 | 71.63±3.56 | 73.33±3.52 | 66.45±5.52 | 64.29±3.00 | 68.11±3.95 | 76.89±4.05 | **78.44±2.56** |
| 6 | 71.39±3.14 | 75.78±3.47 | 77.05±3.26 | 71.46±2.42 | 66.38±2.05 | 72.44±4.95 | 81.54±4.03 | **82.33±4.80** |

Fig. 1. Images of one person from the Yale face dataset.

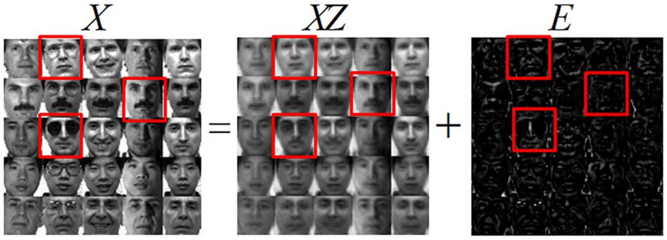
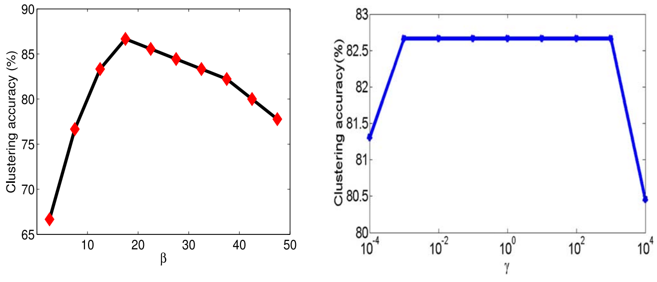


Fig. 2. Some samples of using NNLRR to correct the errors in the Yale face dataset. Left: the original data matrix $X$. Middle: the corrected data $XZ$. Right: the error $E$.



(a) (b)

Fig. 3. Clustering results (%) versus parameter. (a) $\beta$. (b) $\gamma$.

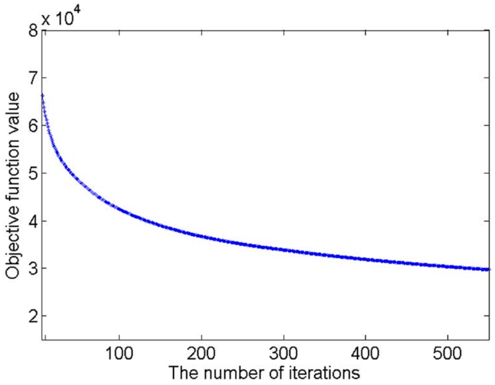


Fig. 4. Convergence process for NNLRR.

tradeoff between the error term and other terms. The selection of parameter $\beta$ is usually based on the prior of the error level of data. In our experiments, we used the grid-search strategy to conduct parameter selection for each algorithm that we implemented. The MATLAB code of NNLRR is publicly available at http://www.yongxu.org/lunwen.html.

### A. Experiment on the Yale Face Dataset

The Yale face dataset (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 images of 15 individuals and each person provides 11 different images with various facial expressions and lighting conditions. In our experiments, each image was manually cropped and resized to 32 × 32 pixels. Fig. 1 shows the sample images of a person from the Yale face dataset.

In this experiment, 2–6 images per person are randomly selected as labeled samples and the remaining are regarded as unlabeled samples. Parameter $\beta$ of NNLRR is set to 34. The clustering results are shown in Table I, in which NNLRR outperforms other state-of-the-art algorithms. For example, when we select 5 and 6 images per person as labeled samples, the classification accuracy of NNLRR is 78.44% and 82.33% which are, respectively, higher NNLRS (the second best algorithm) by 1.56% and 0.80%, respectively. We randomly select some images which contain different error: glasses, beard, and expression, to demonstrate the performance of corrupted images recovery by NNLRR. The recovery results are shown in Fig. 2, in which these images with corruption are approximately recovered (indicated with the red boxes). Fig. 3 shows the clustering results (%) versus the variations of parameter $\beta$ and $\gamma$, respectively, in which the first six images per person were labeled and the remaining were unlabeled. Please note that the experiment ($\gamma$) is just only to verify the algorithmic robustness to $\gamma$. Fig. 4 shows the convergence process of algorithm.

### B. Experiment on the AR Face Dataset

AR face dataset contains over 4000 images corresponding to 126 persons. These images were captured under different facial expressions, illuminations, and occlusions (sun glasses and scarf). The pictures were taken under strictly controlled conditions. In this experiment, we take all the images of the first 30 persons from a subset which provides 3120 gray images from 120 subjects with each subject providing 26 images. Thus, there are 718 images in total are selected in this experiment. Each image is cropped and resized to 32 × 32 pixels. Fig. 5 shows some images of one person from the AR dataset. In this experiment, 2, 5, 8, 11, and 14 images per person are randomly selected as labeled samples and the remaining images are regarded as unlabeled samples. Parameter $\beta$ of NNLRR was set to 93.

TABLE II
CLUSTERING RESULTS (%) ON THE AR DATASET. NOTE THAT # TR DENOTES THE NUMBER OF LABELED SAMPLES PER SUBJECT

| # Tr. | LRR | LatLRR | RobustLatLRR | SPG | LSA | SSC | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|---|
| 2 | 75.09±2.52 | 78.78±3.00 | 80.49±2.57 | 59.38±2.64 | 66.69±2.86 | 73.83±1.47 | 82.25±1.80 | **85.39±1.65** |
| 5 | 91.41±0.96 | 90.46±2.93 | 94.26±2.78 | 77.39±2.56 | 83.87±2.90 | 86.97±1.51 | 93.34±2.25 | **95.29±0.47** |
| 8 | 95.65±1.75 | 94.42±2.35 | 93.00±2.74 | 86.81±1.09 | 87.66±2.38 | 93.62±1.25 | 95.95±1.83 | **96.85±1.27** |
| 11 | **97.23±0.72** | 95.75±2.88 | 96.66±2.96 | 91.46±1.38 | 92.93±2.94 | 95.00±1.52 | 97.20±1.76 | **97.33±0.41** |
| 14 | **97.88±0.24** | 96.40±2.21 | 97.00±3.02 | 95.00±1.48 | 93.78±2.33 | 96.89±2.03 | **97.69±1.68** | **97.56±0.60** |

TABLE III
CLUSTERING RESULTS ON THE FIRST SUBSET. NOTE THAT # TR DENOTES THE NUMBER OF LABELED SAMPLES OF SUBJECT

| # Tr. | LRR | LatLRR | RobustLatLRR | SPG | LSA | SSC | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|---|
| 1 | 79.58±1.64 | 82.22±1.29 | 84.04±1.55 | 79.23±1.36 | 74.65±3.30 | 78.74±2.32 | **87.34±1.90** | **87.64±1.56** |
| 2 | 89.93±1.59 | 89.90±1.58 | 92.50±1.82 | 87.00±2.21 | 81.96±2.75 | 86.61±2.08 | 92.47±1.73 | **93.44±1.97** |
| 3 | 93.23±1.73 | 94.97±1.36 | 95.56±1.93 | 87.85±3.41 | 88.31±2.50 | 92.48±1.10 | 95.59±1.67 | **96.36±1.60** |

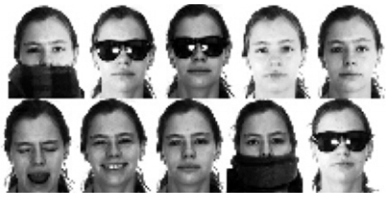


Fig. 5. Some face images of one person from the AR dataset.

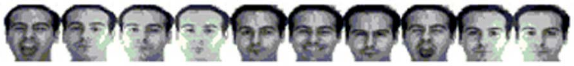


Fig. 6. Some face images of one person from the first subset of the AR dataset.

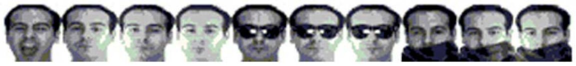


Fig. 7. Some face images of one person from the second subset of the AR dataset.

Table II shows the clustering results on the AR dataset. It can be seen that the clustering performance of NNLRR is almost better than all the compared algorithms. Especially, when the size of labeled samples is small, the superiority of NNLRR is very obvious. As the labeled samples increase, LRR and NNLRR have almost similar performance.

In order to elaborate NNLRRs validity on different noises, following [41], we test NNLRR and other algorithms on two subsets of the AR face dataset. The first subset excludes the images wearing glasses or scarf and thus the errors in this subset are mainly shadows and expression (see Fig. 6). The second subset includes only the first 13 face images and thus the time-caused error is excluded (see Fig. 7). For the first subset, 1–3 images per subject are randomly selected as labeled samples and the remaining images are regarded as unlabeled samples. For the second subset, the first seven images (no glasses and scarf occlusions) per person are used as labeled samples and the remaining images (glasses or scarf occlusions) are used as unlabeled samples. Table III shows the clustering results on the first subset, in which NNLRR outperforms all the other methods (parameter $\beta$ of NNLRR is set to 120). Table IV shows the clustering results on the second subset (parameter $\beta$ of NNLRR is set to 160). From this table, we can see that NNLRR outperforms other algorithms when dealing occlusion. Although RobustLatLRR and NNLRS have strong power in handing occlusion, the label information is not used to guide the affinity matrix construction. Thus the improvement of performances of them are not obvious.

## C. Experiment on the COIL20 Dataset

The COIL20 dataset (http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php) contains 1440 images of 20 objects and each object provides 72 images. The images of each subject were taken at pose intervals of $5^0$. The original images were normalized to 128 × 128 pixels. In this experiment, each image was converted to a gray-scale image of 32 × 32 pixels for computational efficiency in the experiments. Fig. 8 shows some images from the COIL20 dataset. In this experiment, 2, 4, 6, 8, and 10 images per subject are randomly selected as labeled samples and the remaining images are used as unlabeled samples. Table V shows the clustering results on the COIL20 dataset (parameter $\beta$ of NNLRR is set to 0.2). Again, NNLRR performs better than the other methods.

## D. Experiment on the Extended Yale B Dataset

The Extent Yale B dataset (http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html) consists of 2432 human face images of 38 subjects. Each subject contains about 64 images taken under different illuminations. Half of the images are corrupted by shadows or reflection. Each image is cropped and resized to 32 × 32 pixels. Fig. 9 shows some images of one person from the Extended Yale B dataset. As with the AR dataset, we use the first 18 persons and 1134 images in total in the Extended Yale B to test different methods.

In this experiment, 4, 7, 10, 13, and 16 images per subject are randomly selected as labeled samples and the remaining images are regarded as unlabeled samples. It is obvious that NNLRR outperforms other methods as elaborated in Table VI (parameter $\beta$ of NNLRR was set to 145). The classification accuracy of NNLRR is higher than other methods on the most of cases.

TABLE IV
CLUSTERING RESULTS (%) ON THE SECOND SUBSET

| LRR | LatLRR | RobustLatLRR | SPG | LSA | SSC | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|
| 92.78% | 91.67% | 93.07% | 90.56% | 86.67% | 91.11% | 92.82% | **94.44%** |

TABLE V
CLUSTERING RESULTS (%) ON THE COIL20 DATASET. NOTE THAT # TR DENOTES THE NUMBER OF LABELED SAMPLES PER SUBJECT

| # Tr. | LRR | LatLRR | RobustLatLRR | LSA | SSC | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|
| 2 | 68.97±3.30 | 72.08±3.10 | 74.26±2.88 | 66.67±3.61 | 74.35±2.90 | 74.90±2.95 | **76.37±4.56** |
| 4 | 74.61±2.28 | 78.55±3.02 | 81.04±2.59 | 77.20±2.57 | 81.76±3.35 | 83.29±2.67 | **84.41±2.45** |
| 6 | 82.00±2.46 | 82.89±2.97 | 83.93±3.10 | 79.58±2.96 | 84.51±1.48 | 84.50±2.81 | **86.78±3.22** |
| 8 | 84.62±1.86 | 85.95±1.78 | 87.76±2.42 | 84.22±2.51 | 86.66±2.05 | 88.86±2.53 | **89.09±1.72** |
| 10 | 85.55±1.08 | 88.87±2.06 | **90.54±2.75** | 85.00±2.90 | **90.46±0.56** | 89.43±2.99 | **90.80±1.62** |

TABLE VI
CLUSTERING RESULTS (%) ON THE EXTENDED YALE B DATASET. NOTE THAT # TR DENOTES THE NUMBER OF LABELED SAMPLES PER SUBJECT

| # Tr. | LRR | LatLRR | RobustLatLRR | SPG | SSC | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|
| 4 | 70.80±3.69 | 70.01±3.12 | 70.43±3.30 | **74.10±2.63** | 67.10±2.20 | **75.52±2.59** | 72.01±1.65 |
| 7 | 83.94±0.58 | 82.00±3.18 | 83.67±2.86 | 81.84±1.12 | 84.07±2.25 | 84.87±2.30 | **85.57±1.94** |
| 10 | 86.94±1.35 | 85.44±1.75 | 85.90±2.92 | 86.49±0.44 | 87.65±1.56 | **89.00±2.96** | **88.00±0.73** |
| 13 | 89.11±1.89 | 89.05±2.40 | 89.66±1.79 | 88.13±2.12 | 90.35±1.23 | 90.90±1.93 | **91.71±1.56** |
| 16 | 91.00±1.95 | 91.06±2.21 | 90.38±2.55 | 90.76±1.40 | 91.26±1.17 | 92.71±1.75 | **93.67±0.79** |

Fig. 8. Some images from the COIL20 dataset.

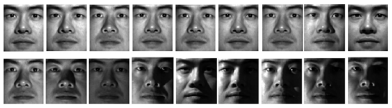


Fig. 9. Some images of one person from the Extended Yale B dataset.

### *E. Experiment on Contiguous Occlusions and Random Pixel Corruptions*

In this section, we selected the first 15 persons from the Extended Yale B face dataset in order to test the robustness of NNLRR to different corruptions. We simulate various levels of contiguous occlusions and random pixel corruptions as follows.

1) *Contiguous Occlusions:* The block occlusions are randomly added to different locations in the labeled and unlabeled images with block size of 5 × 5, 10 × 10, 15 × 15, and 20 × 20 (see Fig. 10).
2) *Random Pixel Corruptions:* We randomly choose pixels from labeled and unlabeled samples and corrupt them by salt and pepper noises. The rates of corrupted pixels are 5%, 10%, 15%, and 20% (see Fig. 10).

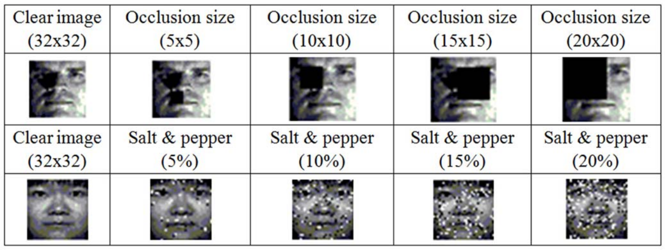


Fig. 10. Some examples of original and corrupted images under varying level of contiguous occlusions and different percentages of salt and pepper noises.

Thirty images per person are randomly selected as labeled samples and the remaining images are used as unlabeled samples. Since SPG is somewhat not suitable to address the corruptions and thus we do not compare it in this experiment. Tables VII and VIII, respectively, show the clustering results of different algorithms on the contiguous occlusions and random pixel corruptions. Obviously, the clustering performance of NNLRR is better than other methods on these two cases, which shows the robustness of NNLRR for the contiguous occlusions and random pixel corruptions. Recovering a clear face image from the images contaminated by different level errors is not an easy job [41]. From Fig. 11, we can see NNLRR can remove block and salt and pepper noises well.

### *F. Discussion*

Based on the experimental results shown in the above sections, the following observations can be concluded.

1) In most previous low-rank-based semi-supervised subspace clustering methods in which the label information is not used so that the constructed affinity matrix cannot deliver enough discriminant information and thus, the performance of them cannot be obviously improved. However, in NNLRR, the label information (i.e., label matrix $Y$) is used to guide the affinity matrix construction. It can be found from the experiments that NNLRR performs better than the state-of-the-art semi-supervised

TABLE VII
CLUSTERING RESULTS (%) ON THE EXTENDED YALE B DATASET WITH CONTIGUOUS OCCLUSIONS

| Occ.size | LRR | LatLRR | RobustLatLRR | SSC | LSA | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|
| 5×5 | 88.89±1.07 | 88.43±2.36 | 89.60±2.50 | 89.82±1.07 | 82.63±3.88 | **90.31±3.22** | **90.92±0.41** |
| 10×10 | 87.85±2.69 | 87.82±2.66 | 88.02±1.87 | 88.36±1.62 | 80.34±4.35 | 88.89±2.62 | **89.57±3.28** |
| 15×15 | 85.30±2.33 | 84.10±2.74 | 85.00±1.39 | 83.40±1.38 | 75.92±3.45 | 85.83±2.44 | **86.82±1.47** |
| 20×20 | 82.39±2.91 | 82.24±3.02 | 82.80±2.12 | 80.68±2.16 | 72.87±3.53 | **84.53±2.01** | **84.10±1.08** |

TABLE VIII
CLUSTERING RESULTS (%) ON THE EXTENDED YALE B DATASET WITH RANDOM PIXEL CORRUPTIONS

| Cor.rate | LRR | LatLRR | RobustLatLRR | SSC | LSA | NNLRS | NNLRR |
|---|---|---|---|---|---|---|---|
| 5% | 82.30±2.01 | 84.33±2.06 | 86.03±2.20 | 78.53±3.75 | 74.03±4.58 | 87.76±3.23 | **89.48±1.03** |
| 10% | 82.67±0.79 | 82.87±2.25 | 82.96±2.37 | 76.35±5.21 | 72.88±2.82 | 84.25±3.58 | **84.50±1.28** |
| 15% | 72.17±0.93 | 71.10±2.58 | 71.50±2.45 | 70.00±4.79 | 68.41±3.95 | **73.06±3.89** | **72.45±0.49** |
| 20% | 67.13±1.49 | 66.54±2.00 | 67.06±2.11 | 65.20±3.75 | 63.32±3.09 | 67.28±3.75 | **68.20±1.40** |

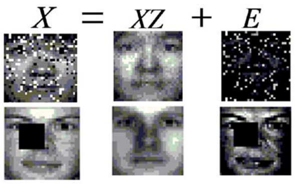


Fig. 11. Two examples of using NNLRR to recovery the corrupted Extended Yale B face images. Left: the contaminated matrix *X*. Middle: the corrected data *XZ*. Right: the error *E*.

subspace clustering methods on the mean classification accuracy by making effective use of the label information in the process of the affinity matrix construction. In most cases, the improvement of classification accuracy is significant such as on the Yale and AR face datasets.

2) NNLRR is more robust than the other compared methods, especially on the dataset with multiple noises. For example, the images in the AR dataset involve different errors such as occlusions (glasses and scarf), illuminations, and expressions, it is not easy job to cluster them. However, the affinity matrix construction and subspace clustering are integrated into one step in NNLRR so that NNLRR can find an overall optimum. In other words, NNLRR can find an optimal balance between the error fitting (i.e., *E*) and subspace clustering. Thus, NNLRR outperforms the similar methods such as LRR, LatLRR, RobustLatLRR, and NNLRS in most cases.
3) Although SPG imposes nonnegative sparse constraint on the affinity matrix, such constraint only captures locally linear structure of the data but the global mixture of subspaces structure may be lost [26]. NNLRR can capture the global mixture of subspaces structure via the explicit low-rank constraint. As shown in all the experiments, NNLRR has obvious advantages in terms of clustering performance than SPG. Similarly, the global mixture of subspaces structure is not captured in SSC. Although NNLRS imposes non-negative sparse and low-rank constraints on the affinity matrix, the label information is not used to guide the affinity matrix construction. Thus, the improvement of clustering accuracy is not obvious.
4) NNLRR can properly deal with both contiguous occlusions and random pixel corruptions (see Fig. 11).
5) From the deduction of algorithm in Section III-C, we also find the main limitation of NNLRR is that the computation cost is still high since it involves to the SVD. We would like to speed up our algorithm in the future. In addition, a issue in NNLRR is how to estimate the parameter $\beta$, especially when the data are contaminated by different level errors such as outlier, noise, and corruption, the selection of $\beta$ is quite challenging.

## V. CONCLUSION

In this paper, we propose a novel semi-supervised subspace clustering method called NNLRR, in which the label information is used to guide the affinity construction. Moreover, NNLRR integrates the affinity matrix construction and subspace clustering into one step to guarantee an overall optimum. An associated efficient iteratively LADMAP is introduced to solve the optimization problem, which uses less auxiliary variables and matrix inversion. We conduct adequate experiments to verify that NNLRR is superior to the state-of-the-art methods. In the future, we will explore the applications of this idea on other methods.

## REFERENCES

[1] G. C. Liu *et al.*, "Robust recovery of subspace structure by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[2] Y. Xu *et al.*, "Data uncertainty in face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1950–1961, Oct. 2014.

[3] C. Xu, D. C. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[4] Y. Luo, D. C. Tao, B. Geng, C. Xu, and S. J. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 523–536, Feb. 2013.

[5] C. Gong, D. C. Tao, K. Fu, and J. Yang, "ReLISH: Reliable label inference via smoothness hypothesis," in *Proc. AAAI*, Quebec, QC, Canada, 2014, pp. 1840–1846.

[6] Y. Xiao *et al.*, "Topographic NMF for data representation," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1762–1771, Oct. 2014.

[7] F. P. Nie, X. Q. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. KDD*, New York, NY, USA, 2014, pp. 977–986.

[8] X. Z. Fang *et al.*, "Locality and similarity preserving embedding for feature selection," *Neurocomputing*, vol. 128, pp. 304–315, Mar. 2014.

[9] F. P. Nie, S. M. Xiang, Y. Q. Song, and C. S. Zhang, "Orthogonal locality minimizing globality maximizing projections for feature extraction," *Opt. Eng.*, vol. 48, no. 1, 2009, Art. ID 017202.

[10] T. K. Kim, B. Stenger, J. Kittler, and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning sets and its application," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 216–232, 2011.

[11] K. Liu, Y. Q. Cheng, and J. C. Yang, "Algebraic feature extraction for image recognition based on an optimal discriminant criterion," *Pattern Recognit.*, vol. 26, no. 6, pp. 903–911, 1993.

[12] S. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[13] F. P. Nie, S. M. Xiang, and C. S. Zhang, "Neighborhood minmax projections," in *Proc. IJCAI*, Hyderabad, India, 2007, pp. 993–998.

[14] M. Belkin and P. Niyógi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learn.*, vol. 56, no. 1, pp. 209–239, 2004.

[15] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, Washington, DC, USA, 2003, pp. 912–919.

[16] Y. Luo *et al.*, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013.

[17] R. Vidál, Y. Ma, and S. Sastry, "Generalized principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.

[18] E. Elhamifar and R. Vidál, "Sparse subspace clustering," in *Proc. CVPR*, Miami, FL, USA, 2009, pp. 2790–2797.

[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[20] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[21] A. Ganesh, K. Min, J. Wright, and Y. Ma, "Principal component pursuit with reduced linear measurements," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Cambridge, MA, USA, 2012, pp. 1281–1285.

[22] J. Wright, Y. G. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *Proc. NIPS*, Whistler, BC, Canada, 2009, pp. 1–9.

[23] G. C. Liu and S. C. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. ICCV*, Barcelona, Spain, 2011, pp. 1615–1622.

[24] H. Y. Zhang, Z. C. Lin, C. Zhang, and J. B. Gao, "Robust latent low-rank representation for subspace clustering," *Neurocomputing*, vol. 145, pp. 369–373, Dec. 2014.

[25] V. M. Patel, H. Nguyen, and R. Vidal, "Latent space sparse subspace clustering," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 225–232.

[26] L. S. Zhuang *et al.*, "Non-negative low-rank and sparse graph for semi-supervised learning," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2328–2335.

[27] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. NIPS*, Granada, Spain, 2011, pp. 612–620.

[28] J. Yang and Y. Zhang, "Alternating direction algorithms for L1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[29] T. Y. Zhou, W. Bian, and D. C. Tao, "Divide-and-conquer anchoring for near-separable nonnegative matrix factorization and completion in high dimensions," in *Proc. ICDM*, Dallas, TX, USA, 2013, pp. 917–926.

[30] F. P. Nie, D. Xu, I. W.-H. Tsang, and C. S. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[31] F. P. Nie, J. J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proc. ICML*, Beijing, China, 2014, pp. 1062–1070.

[32] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

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

[34] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 569–592, 2009.

[35] J. Yang, D. L. Chu, L. Zhang, Y. Xu, and J. Y. Yang, "Sparse representation classifier steered discriminative projection with application to face recognition," *IEEE Trans. Neural Netw.*, vol. 24, no. 7, pp. 1023–1035, Jul. 2013.

[36] L. Zhang *et al.*, "Kernel sparse representation-based classifier," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1684–1695, Apr. 2012.

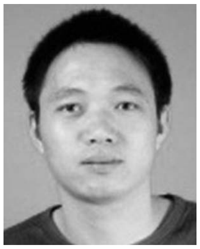
[37] D. Bertsekas, "Nonquadratic penalty functions: Convex programming" in *Constrained Optimization and Lagrange Multiplier Methods*. Belmont, MA, USA: Athena Scientific, 1996, ch. 5, sec. 5.3, pp. 326–340.

[38] J. Eckstein and D. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, pp. 293–318, Apr. 1992.

[39] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate," in *Proc. ECCV*, Graz, Austria, 2006, pp. 94–106.

[40] R. He, W. S. Zheng, B. G. Hu, and X. W. Kong, "Non-negative sparse coding for discriminative semi-supervised learning," in *Proc. CVPR*, Providence, RI, USA, 2011, pp. 2849–2856.

[41] J. H. Chen and J. Yang, "Robust subspace segmentation via low-rank representation," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1432–1445, Aug. 2013.



**Xiaozhao Fang** (S'15) received the M.S. degree in computer science from the Guangdong University of Technology, Guangzhou, China, in 2008. He is currently pursuing the Ph.D. degree in computer science and technology with Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China.

His current research interests include pattern recognition and machine learning. He has published over 15 journal papers.

**Yong Xu** (M'06–SM'15) was born in Sichuan, China, in 1972. He received the B.S. and M.S. degrees from the PLA University of Science and Technology, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2005.

He is currently with Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, biometrics, machine learning, and video analysis.

**Xuelong Li** (M'02–SM'07–F'12) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China.

He is a Full Professor with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

**Zhihui Lai** received the B.S. degree in mathematics from South China Normal University, Guangdong, China, in 2002, the M.S. degree from Jinan University, Guangdong, in 2007, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2011.

He was a Research Associate and a Post-Doctoral Fellow with the Hong Kong Polytechnic University, Hong Kong, from 2010 to 2013. He is currently a Post-Doctoral Fellow with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Harbin, China. His current research interests include face recognition, image processing and content-based image retrieval, pattern recognition, and compressive sense. He has authored over 30 scientific papers in pattern recognition and computer vision.

**Wai Keung Wong** received the Ph.D. degree from the Hong Kong Polytechnic University, Hong Kong.

He is currently with the Institute of Textiles & Clothing, Hong Kong Polytechnic University, Hong Kong, and Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China. His current research interests include artificial intelligence, pattern recognition, and optimization of manufacturing scheduling, planning, and control. He has published over 50 scientific articles in refereed journals, including the IEEE Transactions on Neural Networks and Learning Systems, the *Pattern Recognition*, the *International Journal of Production Economics*, the *European Journal of Operational Research*, the *International Journal of Production Research*, the *Computers in Industry*, and the IEEE Transactions on Systems, Man, and Cybernetics.