Neurocomputing 175 (2016) 715–722

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Locality-preserving low-rank representation for graph construction from nonlinear manifolds

Liansheng Zhuang [a,*], Jingjing Wang [a], Zhouchen Lin [b,c], Allen Y. Yang [d], Yi Ma [e], Nenghai Yu [a]

[a] School of Information Science and Technology, USTC, Hefei, China
[b] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, China
[c] Cooperative Medianet Innovation Center, Shanghai Jiaotong University, China
[d] Department of EECS, University of California, Berkeley, US
[e] ShanghaiTech University, Shanghai, China

ARTICLE INFO

*Article history:*
Received 8 August 2015
Received in revised form
14 October 2015
Accepted 19 October 2015
Communicated by Zhang Zhaoxiang
Available online 10 November 2015

*Keywords:*
Nonlinear manifold clustering
Graph construction
Low-rank representation

ABSTRACT

Building a good graph to represent data structure is important in many computer vision and machine learning tasks such as recognition and clustering. This paper proposes a novel method to learn an undirected graph from a mixture of nonlinear manifolds via Locality-Preserving Low-Rank Representation ($L^2R^2$), which extents the original LRR model from linear subspaces to nonlinear manifolds. By enforcing a locality-preserving sparsity constraint to the LRR model, $L^2R^2$ guarantees its linear representation to be nonzero only in a local neighborhood of the data point, and thus preserves the intrinsic geometric structure of the manifolds. Its numerical solution results in a constrained convex optimization problem with linear constraints. We further apply a linearized alternating direction method to solve the problem. We have conducted extensive experiments to benchmark its performance against six state-of-the-art algorithms. Using nonlinear manifold clustering and semi-supervised classification on images as examples, the proposed method significantly outperforms the existing methods, and is also robust to moderate data noise and outliers.



## 1. Introduction

Graph-based methods have attracted a lot of attention over the last decade in the field of computer vision and machine learning. Various graph-based algorithms have been successfully applied in diverse scenarios, such as image segmentation [1–3], semi-supervised learning [4], and dimensionality reduction [5,6]. Their core idea is to learn a discriminative graph to characterize the relationship among the data samples. However, how to learn a good graph to accurately capture the underlying structure from the observed data is still a challenging problem. In this paper, we propose a novel method to address the graph construction problem for nonlinear manifolds based on some emerging tools in low-rank representation and sparse optimization.

Conceptually, a good graph should reveal the true intrinsic complexity or dimensionality of the data points (say through local linear relations), and also capture certain global structures of the data as a whole (i.e. multiple clusters, subspaces, or manifolds). Traditional methods, such as *k-nearest neighbors* and *Locally Linear Embedding* (LLE) [7,8], mainly rely on pair-wise Euclidean distances to build a graph by a family of overlapping local patches. Since pair-wise distances only characterize the local geometry of these patches by linearly reconstructing each data point from its neighbors, these graphs can only capture local structures, and are very sensitive to local data noise and errors as well. Moreover, traditional methods only work well for a single manifold, and often fail when data points arise from multiple manifolds.

Most recently, in order to capture the global structure of the data, several methods [9–13] have been proposed to construct a sparse and block-diagonal graph with new mathematical tools (such as *sparse representation* [14] and *Low-Rank Representation* [12]) from high-dimensional statistics and covex optimization. Different from traditional methods, these methods represent each datum as a linear combination of all the remaining samples (such as in [9,10]) or all the whole data (such as in [11–13]). Here, we call them *Representation-based methods*. By solving a high-dimensional convex optimization problem, these methods automatically select the most informative neighbors for each datum, and

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.10.119
0925-2312/

simultaneously obtain the graph adjacency structure and graph weights in nearly a parameter-free way. Benefitting from the new mathematical tools, these methods are able to generate a block-diagonal graph, and are robust to data noise. However, the block-diagonal structures obtained by these methods are often fragile, because they hold the hypothesis that the manifold can be embedded linearly or almost linearly in the ambient space. Unfortunately, in real applications, this hypothesis may not be always true. It has been proven that many high-dimensional data usually exhibit significant nonlinear structure, where these representation-based methods often fail to deliver satisfactory performance. As a result, the block-diagonal structures cannot be enforced strictly in this case.

In fact, studies on manifold learning have shown that, to deal with data sampled from nonlinear manifolds [15–17], one has to exploit the local geometrical structure of manifold, or use a nonlinear mapping to "flatten" the data (such as kernel methods [18,19]). In order to preserve the local geometrical structure embedded in high-dimensional space, some graph regularizers are readily imposed on the linear combination representation of the data. For example, Zheng et al. proposed a method called *Low-Rank Representation with Local Constraint* (LRRLC) [15] by incorporating the so-called *local consistency assumption* into the original *Low-Rank Representation* (LRR) model, with the hope that if two samples are close in the intrinsic geometry of the data distribution, they will have a large similarity coefficient. LRRLC introduced a weighted sparsity term (i.e. graph regularization term) with data-dependent weights into the original LRR model. The weights changed with the distance between samples. However, the graph regularization term could not guarantee that close samples would have large similarity coefficient. It only enforced the coefficients between faraway points to be small. Essentially, the LRRLC model only used the relationship among points to re-weight the linear representation.

Our goal is to preserve both the global structure *and* the local structure in our constructed graph. To capture the global structure, the linear representation $Z$ should be block diagonal, which means that the coefficient $Z_{ij}$ should be zero if data point $x_i$ and $x_j$ are not in the same cluster.[1] Since the local consistency assumption only encourages the coefficients between close samples to be nonzero, it will not necessarily lead to being block diagonal. On the contrary, the LRRLC model often fails to accurately represent the geometric structure of manifolds because of two drawbacks. First, LRRLC directly uses affine subspaces to find neighborhoods of points, and thus are likely to select faraway points as neighbors. This may cause a coefficient between two faraway points to be nonzero, even if they belong to different manifolds or are well separated by other points on the same manifold. Second, LRRLC uses the non-negativity constraint to define neighborhood for every point. When a point is on a boundary, this constraint may choose points from other manifold/subspace as its neighbors, or the boundary point will be isolated. These two drawbacks often violate the block diagonalization of its solution in the LRRLC model. As a result, the LRRLC model often obtains a dense graph that negatively affects its performance.

### 1.1. Contributions

Inspired by the above insights, we propose to extend the LRR model to construct an informative graph called *Locality-Preserving Low-Rank Representation Graph* ($L^2R^2$-graph). Specially, given a set of data points, we represent each data point as a linear combination of all the other points. For each point, we determine its neighbors according to the pair-wise distance. By restricting the coefficient $Z_{ij}$ for non-neighbors to be zero and imposing the affine constraint, we approximate the nonlinear manifold by a collection of affine subspaces. Since we require that data vector on the same affine subspace can be clustered in the same cluster, we require that the coefficient vectors of all data points collectively form a low-rank matrix. By imposing the low-rank constraint, the $L^2R^2$-graph can better capture the global cluster or subspace structures of the whole data, and is more robust to noise and outliers.

It is worthwhile to highlight several advantages of $L^2R^2$-graph over the existing works:

1. Compared with traditional methods, since $L^2R^2$-graph imposes the low-rank constraint, it can better capture the global structure. Moreover, as shown in later experiments, though $L^2R^2$-graph uses pair-wise distance to define the graph adjacent structure, it is insensitive to the global parameters, while traditional methods are more sensitive to the global parameters.
2. Compared with other representation-based methods based on the hypothesis of linear subspaces, $L^2R^2$-graph explicitly considers the local structure of manifolds, and preserve it during graph construction. Such local structure preservation makes the learned $L^2R^2$-graph more sparse than these representation-based methods.
3. Compared with LRRLC-graph [15], $L^2R^2$-graph can better preserve the geometric structure of manifolds. In LRRLC-graph, the local structure is used to re-weight the linear combination coefficients, which compromises the block diagonality assumption of the representation. While in $L^2R^2$-graph, the local structure is used to define the neighborhood of each point. Since restricting the coefficient $Z_{ij}$ for non-neighbors to be zero may not affect the block diagonality of the representation $Z$, the resulting $Z$ could still be block diagonal in ideal cases.

We conduct extensive experiments on simulation data and public databases for two typical tasks, namely nonlinear manifolds clustering and semi-supervised classification. The experimental results clearly demonstrate that the $L^2R^2$-graph can significantly improve the learning performance, and is more informative and discriminative than other graphs constructed by conventional methods.

The remainder of this paper is organized as follows. In Section 2, we give the details of how to construct a locality-preserving low-rank graph. Our experiments and analysis are presented in Section 3. Finally, Section 4 concludes our paper.

## 2. Graph building via locality-preserving low-rank representation

### 2.1. Low-rank representation: an overview

Low-Rank Representation (LRR) was proposed to segment data drawn from a union of multiple linear (or affine) subspaces. Given a set of sufficiently dense data vectors $X=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]\in\mathbb{R}^{d\times n}$ (each column is a sample) drawn from a union of $k$ subspaces, LRR seeks the lowest-rank representation that represent all the vectors as the linear combination of the data themselves, and solves the following convex optimization problem:

$$\begin{aligned}&\min_{Z,E}\quad \|Z\|_*+\lambda\|E\|_{2,1},\\ &\text{s.t.}\quad X=XZ+E,\end{aligned}\tag{1}$$

[1] Note that it is a misconception for $Z$ to be block diagonal that $Z_{ij}$ should be nonzero if $x_i$ and $x_j$ are in the same cluster.

where $Z=[\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_n]$ is the coefficient matrix with each $\boldsymbol{z}_i$ being the coefficient vector of $\boldsymbol{x}_i$ based on the data matrix $X$ and $\|\cdot\|_*$ is the nuclear norm, i.e., sum of singular values. The optimal solution $Z^*$ of the above problem is called the "lowest-rank representation" of data $X$. $\|E\|_{2,1}=\sum_{j=1}^{n}\sqrt{\sum_{i=1}^{d}(e_{ij})^2}$ is called the $\ell_{2,1}$-norm, where $e_{ij}$ is the $(i,j)$-th element of matrix $E$. It is used to model the corruption error. Finally, the parameter $\lambda>0$ is used to balance the two terms, which could be chosen according to the properties of the two norms, or tuned empirically.

As shown in [20,12], $Z^*$ is block-diagonal when data are clean and sampled from independent subspaces. LRR better captures the global structure of the data, and is more effective for robust subspace segmentation from corrupted data. After solving problem (1), we can define the affinity matrix $W$ of an undirected graph by $W=(|Z^*|+|Z^{*T}|)/2$. Therefore, the undirected graph (called the LRR-graph) also captures the global structure of the whole data. It could be integrated into any spectral clustering algorithm (such as Normalized Cuts [21]) to segment the linear (or affine) subspaces. In the next subsection, we will generalize the LRR model to construct an undirected graph to approximate nonlinear manifolds.

### 2.2. Locality-preserving low-rank representation

In this section, we propose a novel approach to construct a graph, called *Locality-preserving Low Rank Representation* ($L^2R^2$). The key idea is based on the well-known observation that a nonlinear manifold can be approximated by a collection of piecewise affine subspaces. Therefore, the neighborhood of each data point can be fit by an affine subspace model. This task is essentially an LRR problem. However, we need to further constrain that in the low-rank coefficients $\boldsymbol{z}_i$ of $\boldsymbol{x}_i$, those coefficients that correspond to the sufficiently faraway data points in $X$ (e.g., in terms of Euclidian distance or $K$-nearest neighbors) from $\boldsymbol{x}_i$ should be zero, because the overall distribution of the samples on the manifold is nonlinear and therefore faraway points should not belong to the affine subspace.

Since we require the samples on the same affine subspace to be clustered in one cluster, the collection of all coefficient vectors $Z=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n]$ should remain a low-rank matrix just as in LRR. Therefore, the $L^2R^2$ model solves the following problem:

$$\begin{aligned}&\min_{Z,E}\quad \|Z\|_*+\lambda\|E\|_{2,1},\\ &\text{s.t.}\quad X=XZ+E,\\ &Z^T\mathbf{1}=\mathbf{1},\\ &Z_{ij}=0,(i,j)\in\overline{\Omega},\end{aligned}\tag{2}$$

where $\mathbf{1}$ is an all-one vector, $\Omega$ is a set of edges between the samples in $X$ that define an adjacency graph, and $\overline{\Omega}$ is the complement of $\Omega$, whereby $(i,j)\in\overline{\Omega}$ indicates that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are not neighbors. In this paper, we use the $K$-nearest neighbor method to determine the graph adjacency structure, where $K$ is user specified. It turns out that using $L^2R^2$ for clustering nonlinear manifolds is quite stable with respect to a reasonable choice of $K$. More discussion about this issue are presented in Section 3.1.

As we can see from (2), although $L^2R^2$ seeks the lowest-rank representation among all the data points, it preserves the local geometric structure in its solution by enforcing $Z_{ij}=0,(i,j)\in\overline{\Omega}$. So we call the optimal solution $Z^*$ of (2) as "Locality-Preserving Low-Rank Representation." Preserving locality brings an important benefit for $L^2R^2$, namely the final solution $Z^*$ is guaranteed to be sparse ($Z^*_{ij}\neq 0$ only if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are neighbors). In Fig. 1, we show the sparsity pattern of the obtained coefficients matrices from USPS database. As we can see, the coefficients matrix obtained by $L^2R^2$ model is sparser than that obtained by the original LRR model. According to [22], low sparsity is one of the basic characteristics of an informative graph. Therefore, $L^2R^2$ is more suitable for constructing an informative graph.

### 2.3. Solving $L^2R^2$ via ADMM

In this section, we detail an efficient convex optimization algorithm to solve the $L^2R^2$ problem (2). Note here that (2) is a convex optimization problem with linear constraints with respect

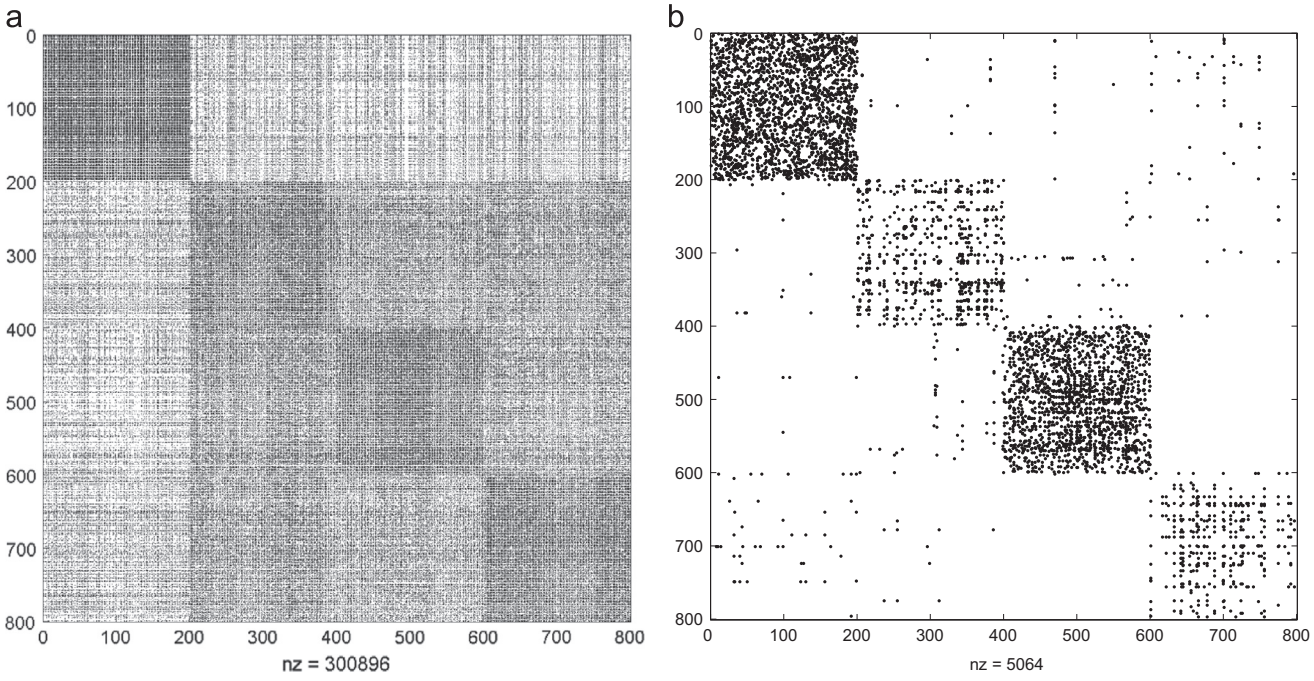



**Fig. 1.** Visualize the sparsity pattern of the coefficients matrices from USPS database obtained by LRR model and $L^2R^2$ model. (a) is the sparsity pattern of the coefficients obtained by LRR method. (b) is the sparsity pattern of the coefficients obtained by $L^2R^2$ model. Both LRR model and $L^2R^2$ model share the same parameter $\lambda=0.1$. The number of nearest neighbor is set to 10.

to $Z$ and $E$. Therefore, it can be written in the standard form as

$$\min_{\mathcal{P}_{\overline{\Omega}}(Z)=\mathbf{0},E} \quad \|Z\|_*+\lambda\|E\|_{2,1} \quad \text{s.t.,} \quad A(Z)+B(E)=\boldsymbol{c}, \tag{3}$$

where $\mathcal{P}_{\overline{\Omega}}(Z)$ is a projection operator that maps elements $z_{ij}$ of $Z$ in a vector form when $(i,j)\in\overline{\Omega}$, $\mathbf{0}$ is an all-zero vector. It is easy to verify that the linear mapping operators $A(Z)$ and $B(E)$ take the following form:

$$A(Z)=\begin{pmatrix}\mathrm{vec}(XZ)\\ Z^T\mathbf{1}\end{pmatrix}, \quad B(E)=\begin{pmatrix}\mathrm{vec}(E)\\ \mathbf{0}\end{pmatrix}, \tag{4}$$

where $\mathrm{vec}(\cdot)$ is the vectorization operator that stacks the columns of a matrix into a vector. The $\boldsymbol{c}$ vector takes the form:

$$\boldsymbol{c}=\begin{pmatrix}\mathrm{vec}(X)\\ \mathbf{1}\end{pmatrix}. \tag{5}$$

We recognize the main difficulty in optimizing (3) is three folds: first, the objective function $f(Z,E)\doteq\|Z\|_*+\lambda\|E\|_{2,1}$ contains two convex but nonsmooth functions, i.e., the nuclear norm of $Z$ and the $\ell_{2,1}$-norm of $E$. Second, the domain of matrix $Z$ is constrained by the condition that $\mathcal{P}_{\overline{\Omega}}(Z)=\mathbf{0}$. Third, $Z$ and $E$ have to simultaneously satisfy the equality constraint $h(Z,E)\doteq A(Z)+B(E)-\boldsymbol{c}=0$.

In convex optimization, a popular approach to handle these challenges is known as the *alternating direction method of multipliers* (ADMMs) method. ADMM has been developed to address large-scale distributed optimization problems [23]. More recently, it has been successfully applied to sparse optimization problems such as basis pursuit [24,25] and Robust PCA [26], which are intimately related to the problem in this paper. Therefore, we choose ADMM to solve the $L^2R^2$ problem.

First, we observe that in the constrained convex optimization problem (3), the most expensive computation is the evaluation of the nuclear norm $\|Z\|_*$, which is equal to the $\ell_1$-norm of all the singular values of $Z$. Therefore, we first introduce a new matrix $J$ and an equality constraint $J=Z$, such that the minimization of $\|J\|_*$ and the evaluation of the constraint $h(Z,E)=0$ can be separated. Hence, we convert (3) to the following equivalent problem:

$$\min_{\mathcal{P}_{\overline{\Omega}}(Z)=\mathbf{0},E,J} \quad \|J\|_*+\lambda\|E\|_{2,1} \quad \text{s.t.,} \quad h(Z,E)=0,\ J=Z. \tag{6}$$

Second, we derive the augmented Lagrangian function $\mathcal{L}_\mu(Z,E,J)$ of (6) over the real space and ignore the constraint $\mathcal{P}_{\overline{\Omega}}(Z)=\mathbf{0}$.

$$\mathcal{L}_\mu(Z,E,J)=\|J\|_*+\|E\|_{2,1}+<Y_1,X-XZ-E>+<Y_2,\mathbf{1}^T-\mathbf{1}^TZ>+<Y_3,Z-J>+\frac{\mu}{2}\left(\|X-XZ-E\|_F^2+\|\mathbf{1}^T-\mathbf{1}^TZ\|_F^2+\|Z-J\|_F^2\right), \tag{7}$$

where $Y_1$, $Y_2$, and $Y_3$ are the matrices of the Lagrange multipliers that correspond to the three equality constraints in (6), respectively.

Third, according to ADMM, the updates of the variables $Z$, $E$ and $J$ go as follows by minimizing over the augmented Lagrangian function $\mathcal{L}_\mu(Z,E,J)$ alternately. More specifically, assuming $E^k$, $Z^k$, and $Y_i^k$ are fixed,

$$J^{k+1}=\min_J \|J\|_*+<Y_3^k,Z^k-J>+\frac{\mu^k}{2}\|Z^k-J\|_F^2. \tag{8}$$

Next, assuming that $J^{k+1}$, $Z^k$, and $Y_i^k$ are fixed,

$$E^{k+1}=\min_E \|E\|_{2,1}+<Y_1^k,X-XZ^k-E>+\frac{\mu^k}{2}\|X-XZ^k-E\|_F^2. \tag{9}$$

Each of the above subproblems has a closed-form proximal function solution as detailed in [27,12].

In this paper, we need to derive a new update rule for the low-rank representation $Z$ subject to the locality constraint $\mathcal{P}_{\overline{\Omega}}(Z)=\mathbf{0}$. More specifically, we solve the following problem:

$$\min_{\mathcal{P}_{\overline{\Omega}(Z)}=0} <Y_1^k,X-XZ-E^{k+1}>+\frac{\mu^k}{2}\|X-XZ-E^{k+1}\|_F^2+<Y_2^k,\mathbf{1}^T-\mathbf{1}^TZ>+\frac{\mu^k}{2}\|\mathbf{1}^T-\mathbf{1}^TZ\|_F^2+<Y_3^k,Z-J^{k+1}>+\frac{\mu^k}{2}\|Z-J^{k+1}\|_F^2.$$

It is equivalent to:

$$\min_{\mathcal{P}_{\overline{\Omega}(Z)}=0}\frac{1}{2}\|X-XZ-E^{k+1}+\frac{1}{\mu^k}Y_1^k\|_F^2+\frac{1}{2}\|\mathbf{1}^T-\mathbf{1}^TZ+\frac{1}{\mu^k}Y_2^k\|_F^2+\frac{1}{2}\|Z-J^{k+1}+\frac{1}{\mu^k}Y_3^k\|_F^2.$$

We linearize the above equation with respect to $Z$ at $Z^k$:

$$\min_{\mathcal{P}_{\overline{\Omega}(Z)}=0} <-X^T\left(X-XZ^k-E^{k+1}+\frac{1}{\mu^k}Y_1^k\right)-\mathbf{1}\left(\mathbf{1}^T-\mathbf{1}^TZ^k+\frac{1}{\mu^k}Y_2^k\right)+\left(Z^k-J^{k+1}+\frac{1}{\mu^k}Y_3^k\right),Z-Z^k>+\frac{\eta}{2}\|Z-Z^k\|_F^2,$$

where $\eta=\|X\|_2^2+\|\mathbf{1}^T\|_2^2+1$.

Let

$$H^k=-X^T\left(X-XZ^k-E^{k+1}+\frac{1}{\mu^k}Y_1^k\right)-\mathbf{1}\left(\mathbf{1}^T-\mathbf{1}^TZ^k+\frac{1}{\mu^k}Y_2^k\right)+\left(Z^k-J^{k+1}+\frac{1}{\mu^k}Y_3^k\right).$$

Then we have

$$\min_{\mathcal{P}_{\overline{\Omega}}(Z)=0} <H^k,Z-Z^k>+\frac{\eta}{2}\|Z-Z^k\|_F^2, \tag{10}$$

and

$$\min_{\mathcal{P}_{\overline{\Omega}}(Z)=0}\|Z-Z^k+\frac{1}{\eta}H^k\|_F^2=P_\Omega\left(Z^k-\frac{1}{\eta}H^k\right). \tag{11}$$

Finally, assuming that $E^{k+1}$, $Z^{k+1}$, and $J^{k+1}$ are fixed. The update rules for $Y_1$, $Y_2$, and $Y_3$ follow a simple dual ascend step [28]:

$$Y_1^{k+1}=Y_1^k+\mu^k(X-XZ^{k+1}-E^{k+1}),$$
$$Y_2^{k+1}=Y_2^k+\mu^k(\mathbf{1}^T-\mathbf{1}^TZ^{k+1}),$$
$$Y_3^{k+1}=Y_3^k+\mu^k(Z^{k+1}-J^{k+1}).$$

### 2.4. Constructing $L^2R^2$-graph

Given a data matrix $X$, let $G=(V,E)$ be a graph associated with a weight matrix $W=\{w_{ij}\}$, where $V=\{\boldsymbol{x}_i\}_{i=1}^n$ is the vertex set, and $E=\{e_{ij}\}$ is the edge set, each edge $e_{ij}$ associating nodes $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ with a weight $w_{ij}$. The problem of graph construction is to determine the graph weight matrix $W$. In this paper, we are primarily concerned about the estimation of an undirected graph with nonnegative weight coefficients.

After solving the problem (2), we may obtain the optimal coefficient matrix $Z^*$. Since each data point is represented by its neighborhood, $Z^*$ naturally characterizes how other samples contribute to the reconstruction of $\boldsymbol{x}_i$. Such information is useful for recovering the clustering relation among samples. The low-rank constraint guarantees that the coefficients of the samples coming from the same affine subspace are highly correlated and fall into the same cluster, so that $Z^*$ can capture the global structure (i.e. the clusters) of the whole data. Moreover, since $Z^*$ is automatically sparse, the graph derived from $Z^*$ is naturally sparse. After obtaining $Z^*$, we can derive the graph weight matrix $W$ as:

$$W=(|Z^*|+|Z^*|^T)/2. \tag{12}$$

The method for constructing an $L^2R^2$-graph is summarized in Algorithm 1.[2]

**Algorithm 1.** Graph construction via $L^2R^2$.

**Input:** Data matrix $X=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]\in\mathbb{R}^{d\times n}$, balance parameter $\lambda$, and $k$-nearest neighbors parameter $K$.

**Steps:**

1: Normalize all the samples $\hat{\boldsymbol{x}}_i=\boldsymbol{x}_i/\|\boldsymbol{x}_i\|_2$ to obtain $\hat{X}=\{\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\ldots,\hat{\boldsymbol{x}}_n\}$.

2: Find neighbors in $X$ using $K$-nearest neighbor method, and assign adjacency graph $\Omega$.

3: Solve the following problem according to Section 2.3:

$$(Z^*,E^*)=\arg\min_{Z,E}\ \|Z\|_*+\lambda\|E\|_1$$
$$\text{s.t.}\quad \hat{X}=\hat{X}Z+E,\ \mathbf{1}^T=\mathbf{1}^TZ,$$
$$Z_{ij}=0,(i,j)\in\overline{\Omega}$$

4: Construct the graph weight matrix $W$ by $W=\frac{|Z^*|+|Z^{*T}|}{2}$.

**output**: The weight matrix $W$ of $L^2R^2$-graph.

## 3. Experiments

In this section, we demonstrate the performance of $L^2R^2$-graph, and compare it with several state-of-the-art graph learning algorithms, which include both traditional graphs ($k$NN-graph, LLE-graph) and optimization-based graphs ($\ell_1$-graph [9], LRR-graph [12], NNLRS-graph [11], and LRRLC-graph [15]). For $k$NN-graph, we adopt Euclidean distance as our similarity measure, and use a Gaussian kernel to re-weight the edges. The Gaussian kernel parameter $\sigma$ is set to 1. For a fair comparison, we share the same graph adjacency structure $\Omega$ among $k$NN-graph, LLE-graph and $L^2R^2$-graph. The algorithms are benchmarked in two applications, namely manifold clustering on synthetic data and semi-supervised classification on image data.

### 3.1. Manifold clustering on synthetic data

In this subsection, we consider the manifold clustering application via standard spectral clustering on $W$. The experiment is conducted on a series of synthetic data sets with added Gaussian noise and data corruption.

First, we evenly sample 800 noise-free points in total from 4 half-circle manifolds (a.k.a. moon manifolds). Then the sample points are embedded into a 100-dimensional space and occupy the first two dimensions. Finally, we add Gaussian noise with mean 0 and variance 0.01 in all the 100-D coordinates, and further randomly select 10% of the samples in $X$ to corrupt with much higher Gaussian noise with zero mean and variance $0.7*\|x\|_2$. One example of the noisy data is shown in Fig. 2.

We apply all the aforementioned graph-learning methods on the synthetic data set, and use Normalized Cuts [21] to generate the final segmentation. Examples of the clustering results are shown in Fig. 3. As clustering methods cannot predict the class label of each cluster, we use a postprocess step to assign each cluster a label: given ground truth classification results, the label of a cluster is the index of the ground truth class that contributes the maximum number of samples to the cluster. And then we can obtain the segmentation accuracy by computing the percentage of correctly classified data vectors. The segmentation accuracy is shown in Fig. 4.

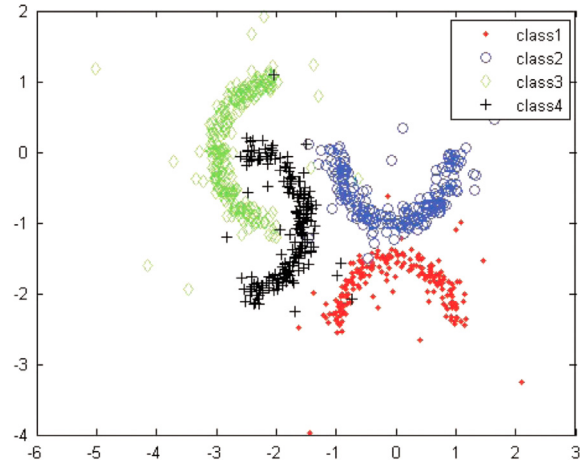



**Fig. 2.** An example of 800 points sampled in $\mathbb{R}^2$ and embedded in 100-D space with added noise and corruption.

We draw the following observations:

1. The state-of-the-art representation-based methods (i.e. $\ell_1$-graph, LRR, NNLRS) are inferior to the traditional methods ($k$NN and LLE). This should not come as a surprise because they aim to preserve only global linear structures of the data. When the data are drawn from nonlinear manifolds, the existing representation methods often violate the local geometric structure, and result in poor performance.
2. All methods ($k$NN, LLE, LRRLC, $L^2R^2$) that utilize the local geometric information significantly outperforms representation-based methods. It shows that exploring the local structure is very important for graph construction from nonlinear manifolds.
3. Though both $L^2R^2$ and LRRLC utilize the local structure, the segmentation accuracy of $L^2R^2$ is more than 12% higher than that of LRRLC. This shows that $L^2R^2$ is more effective in utilizing the local structure information.
4. Though sharing the same graph adjacency structure, $L^2R^2$ markedly outperforms LLE and $k$-NN. This shows that the global low-rank constraint is also crucial to construct an informative graph.

*Evaluation of parameter sensitivity:* There are two parameters in the $L^2R^2$ model: $\lambda$ and $K$. The balance parameter $\lambda$ is to deal with the gross corruption errors in data. The number of nearest neighbors $K$ has an important impact to the graph adjacency structure. In this experiment, we evaluate the change of the segmentation accuracy using $L^2R^2$-graph when the neighborhood parameter $K$ varies. Since the percentage of gross corruption errors in data is fixed, we set $\lambda=0.05$ empirically. The results are shown in Table 1.

From this table, we can see that the performance of $L^2R^2$-graph in manifold clustering is rather stable when $K$ varies from 5 to 20. When $K$ reaches 50, the performance eventually decreases. This is due to the fact that the samples in a 50-point nearest neighborhood do not fit well in a unique affine subspace anymore. In real applications, we can set $K$ between 5 and 10 so as to expect each point and its neighbors to lie on or close to a locally affine patch of the manifold.

### 3.2. Semi-supervised classification on real data

In this subsection, we choose a popular graph-based semi-supervised learning method, namely, *local and global consistency* (LGC) [4], to benchmark the performance of the different graph-

[2] The ADMM implementation can be further accelerated via a *linearized alternating direction method with adaptive penalty* (LADMAP) method. For brevity, refer to [26].

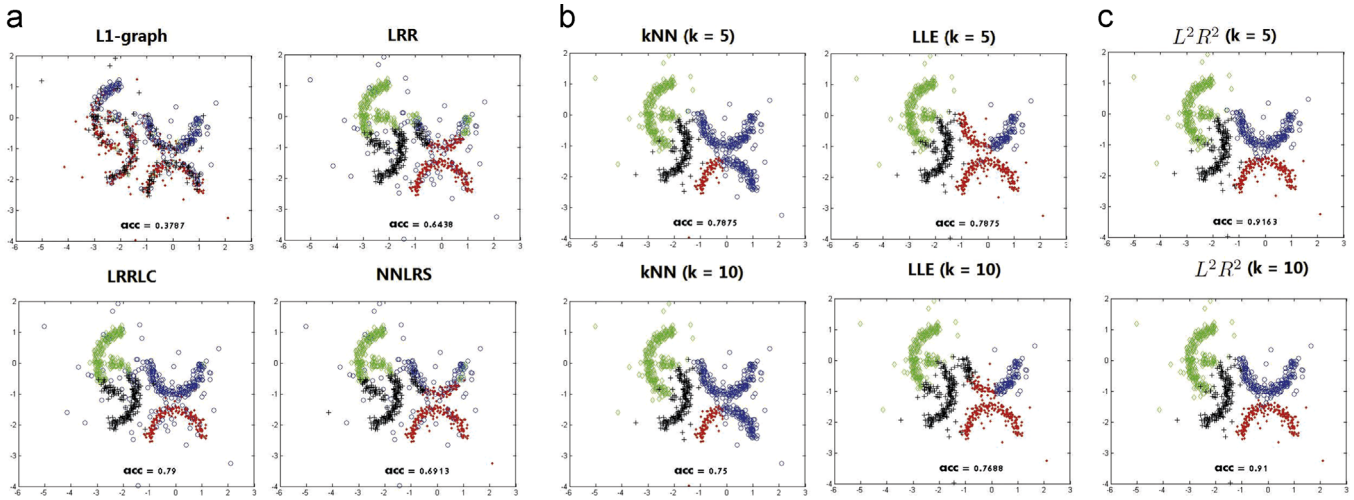



**Fig. 3.** Nonlinear manifold clustering results. (a) Representation-based graph methods; (b) traditional graph methods; (c) $L^2R^2$-graph. The traditional graph and $L^2R^2$-graph share the same local neighborhood structure.

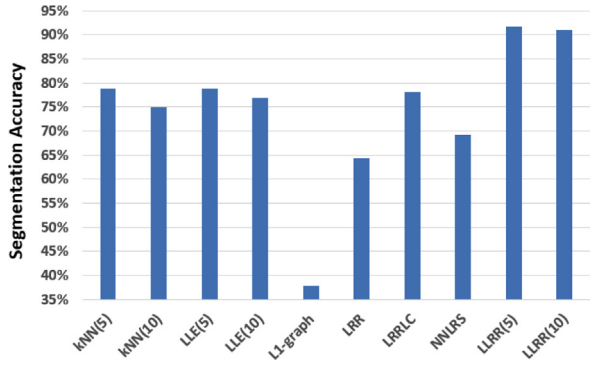



**Fig. 4.** Segmentation accuracy on synthetic data.

**Table 1**
Segmentation accuracy (%) on synthetic data using $L^2R^2$-graph with different nearest neighborhoods number. The parameter $\lambda$ is fixed to 0.05.

| $K$ | 5 | 10 | 20 | 50 |
|---|---|---|---|---|
| Acc. (%) | 91.6 | 91.0 | 91.1 | 80.3 |

learning approaches. LGC is built on an undirected graph, and utilizes the graph and known labels to recovery a continuous classification function $F \in \mathbb{R}^{|V| \times c}$ by optimizing the following energy functions:

$$\min_{F \in \mathbb{R}^{|V| \times c}} \mathrm{tr}\{F^T \tilde{L}_W F + \mu (F-Y))^T (F-Y)\}, \tag{13}$$

where $Y \in \mathbb{R}^{|V| \times c}$ is the label matrix, in which $Y_{ij} = 1$ if sample $\boldsymbol{x}_i$ is associated with label $j$ for $j \in \{1, 2, \ldots, c\}$ and $Y_{ij} = 0$ otherwise. $\tilde{L}_W$ is the normalized graph Laplacian $\tilde{L}_W = D^{-1/2}(D-W)D^{-1/2}$, in which $D$ is a diagonal matrix with $D_{ij} = \sum_j W_{ij}$. The weight $\mu \in [0, \infty)$ balances the tradeoff between the local fitting and the global smoothness of the function $F$. In all experiments, $\mu$ is simply fixed to 0.99.

To comprehensively estimate the performance of $L^2R^2$-graph, we conduct our experiments on two well-known public databases (USPS and COIL20). Sample images from both databases are shown in Fig. 5. Experimentally, USPS roughly has linear subspace structure, while COIL20 lies on nonlinear manifolds. So, we choose these two databases to form our datasets. For USPS database, we only use the images of digits 1, 2, 3 and 4 as four classes, each having 1296, 926, 824 and 852 samples, respectively. So, there are 3874 images in total. For COIL20 database, it contains 20 objects. The images of each objects were taken 5° apart as the object is rotated on a turntable and each object has 72 images. The size of each grayscale image is $32 \times 32$ pixels.

We combine different graphs with the LGC frameworks, and quantitatively evaluate their performance by following the approaches in [9,22,11]. For USPS database, we randomly select 200 images for each category, and randomly label them. The percentage of labeled samples ranges from 10% to 60%. The final results are reported in Tables 2 and 3.

From these two tables, we draw the following conclusions:

1. On the USPS database, NNLRS achieves the best performance when more than 20% of the samples are labeled, while $L^2R^2$-graph is the second best. With 10% labeling, $L^2R^2$-graph has the best performance. The result shows NNLRS has a slight edge in modeling linear subspaces, while $L^2R^2$ remains effective only second to NNLRS.
2. On the COIL20 database, $L^2R^2$ using five nearest neighbors to construct the adjacency graph achieves the best performance across the board. The result shows that $L^2R^2$ is most effective in the case of modeling nonlinear manifolds. In this case, $L^2R^2(5)$ significantly outperforms NNLRS.
3. $L^2R^2$-graph is significantly superior to LRRLC-graph. This means that preserving the local structure is more efficient to explore the local geometric structure than re-weighting the linear coefficients.
4. $L^2R^2$-graph also markedly outperforms LLE-graph in all cases. This one again proves that preserving the global structure is important to construct an informative graph.

## 4. Conclusion

This paper proposes a novel informative graph, called Locality-preserving Low-Rank Representation graph ($L^2R^2$-graph), for graph-based learning methods. $L^2R^2$-graph represents each data point with its neighbors, and approximate the manifolds with piece-wise affine subspaces. By imposing a low-rank constraint on the coefficient matrix, $L^2R^2$-graph can jointly compute the linear coefficients of all points, and better capture the global structure of the whole data on nonlinear manifolds. Meanwhile, by preserving the local structure, $L^2R^2$-graph is automatically guaranteed to be sparse. Our synthetic and real experiments on both manifold clustering and semi-supervised classification have showed that $L^2R^2$-graph is more informative and more suitable for learning nonlinear manifold structures than the existing graph-learning methods.

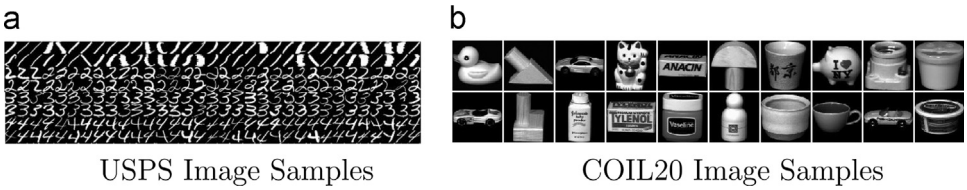



**Fig. 5.** Sample images in our experiments.

**Table 2**
Classification error rates (%) on USPS database under different labeled percentages.

| Methods | 10% | 20% | 30% | 40% | 50% | 60% |
|---|---|---|---|---|---|---|
| $\ell_1$-graph | 33.5 | 26.4 | 18.9 | 16.6 | 11.7 | 8.89 |
| LRR | 3.49 | 1.83 | 1.22 | 0.92 | 0.61 | 0.49 |
| NNLRS | 2.80 | **1.62** | **1.13** | **0.88** | **0.59** | **0.48** |
| LRRLC | 3.40 | 3.06 | 3.02 | 2.80 | 2.68 | 2.58 |
| $k$NN(5) | 3.13 | 2.22 | 1.55 | 1.20 | 0.82 | 0.65 |
| LLE(5) | 3.69 | 2.85 | 2.41 | 2.10 | 1.96 | 1.92 |
| $L^2R^2(5)$ | **2.42** | 1.64 | 1.26 | 0.92 | 0.61 | 0.52 |
| $k$NN(10) | 4.53 | 4.28 | 4.01 | 3.95 | 3.81 | 3.69 |
| LLE(10) | 4.83 | 3.84 | 3.57 | 3.35 | 3.30 | 3.15 |
| $L^2R^2(10)$ | 2.66 | 1.88 | 1.38 | 1.05 | 0.74 | 0.66 |

**Table 3**
Classification error rates (%) on COIL20 database under different labeled percentages.

| Method | 10% | 20% | 30% | 40% | 50% | 60% |
|---|---|---|---|---|---|---|
| $\ell_1$-graph | 29.0 | 23.1 | 17.8 | 15.8 | 14.4 | 12.0 |
| LRR | 16.8 | 13.2 | 10.7 | 10.4 | 9.35 | 8.41 |
| NNLRS | 12.1 | 9.33 | 8.78 | 6.99 | 6.87 | 6.10 |
| LRRLC | 8.53 | 6.49 | 5.24 | 4.35 | 4.32 | 3.58 |
| $k$NN(5) | 12.6 | 10.9 | 10.5 | 10.4 | 9.40 | 8.85 |
| LLE(5) | 6.49 | 4.01 | 2.77 | 2.12 | 2.06 | 1.54 |
| $L^2R^2(5)$ | **4.39** | **2.07** | **1.71** | **1.20** | **1.15** | **0.96** |
| $k$NN(10) | 18.2 | 16.1 | 15.4 | 15.1 | 14.8 | 14.4 |
| LLE(10) | 13.6 | 11.2 | 10.0 | 9.40 | 8.49 | 8.34 |
| $L^2R^2(10)$ | 11.3 | 9.12 | 7.83 | 7.34 | 7.31 | 5.72 |

## Acknowledgments

This work is supported by National Science Foundation (NSF) of China (Nos. 61472379 and 61371192). Zhouchen Lin is supported by National Basic Research Program of China (973 Program) (Grant no. 2015CB352502), National Natural Science Foundation (NSF) of China (Grant nos. 61272341 and 61231002), and Microsoft Research Asia Collaborative Research Program.

## References

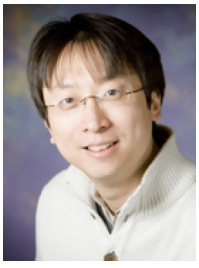
[1] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 888–905.
[2] T. Li, S. Yan, T. Mei, X.-S. Hua, I.-S. Kweon, Image decomposition with multilabel context: algorithms and applications, IEEE Trans. Image Process. 20 (2011) 2301–2314.
[3] T. Li, X. Wu, B. Ni, K. Lu, S. Yan, Weakly-supervised scene parsing with multiple contextual cues, Inf. Sci. (2015) 59–72.
[4] D. Zhou, O. Bousquet, T. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: NIPS, 2004, pp. 595–602.
[5] Y. Pang, Z. Ji, P. Jing, X. Li, Ranking graph embedding for learning to rerank, IEEE Trans. Neural Netw. Learn. Syst. 24 (8) (2013) 1292–1303.
[6] Y. Pang, S. Wang, Y. Yuan, Learning regularized lda by clustering, IEEE Trans. Neural Netw. Learn. Syst. 25 (12) (2014) 2191–2201.
[7] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locality linear embedding, Science 290 (5500) (2000) 2323–2326.
[8] T. Jebara, J. Wang, S. Chang, Graph construction and b-matching for semi-supervised learning, ICML (2009) 441–448.
[9] B. Cheng, J. Yang, S. Yan, Y. Fu, T. Huang, Learning with $\ell_1$-graph for image analysis, IEEE Trans. Image Process 19 (4) (2010) 858–866.
[10] R. He, W.-S. Zheng, B.-G. Hu, X.-W. Kong, Nonnegative sparse coding for discriminative semi-supervised learning, in: CVPR, 2011, pp. 792–801.
[11] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, N. Yu, Non-negative low rank and sparse graph for semi-supervised learning, in: CVPR, 2012.
[12] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Trans. PAMI 35 (1) (2013) 171–184.
[13] L. Zhuang, S. Gao, J. Tang, J. Wang, Z. Lin, Y. Ma, N. Yu, Constructing a non-negative low-rank and sparse graph with data-adaptive features, IEEE Trans. Image Process. 24 (11) (2015) 3717–3728.
[14] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. PAMI 31 (2) (2008) 210–227.
[15] Y. Zheng, X. Zhang, S. Yang, L. Jiao, Low-rank representation with local constraint for graph construction, Neurocomputing 122 (24) (2013) 398–405.
[16] X. Lu, Y. Wang, Y. Yuan, Graph-regularized low-rank representation for describing of hyperspectral images, IEEE Trans. Geosci. Remote Sens. 51 (7–1) (2013) 4009–4018.
[17] Y. Peng, B.-L. Lu, S. Wang, Enhanced low-rank representation via sparse manifold adaption for semi-supervised learning, Neural Netw. 65 (2015) 1–17.
[18] S. Yang, Z. Feng, Y. Ren, H. Liu, L. Jiao, Semi-supervised classification via kernel low-rank representation graph, Knowl. Based Syst. 69 (2014) 150–158.
[19] S. Xiao, M. Tan, D. Xu, Z.Y. Dong, Robust kernel low-rank representation, IEEE Trans. Neural Netw. Learn. Syst. PP (2015) 1.
[20] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: ICML, 2010.
[21] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. PAMI 22 (8) (2000) 888–905.
[22] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.
[23] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.
[24] J. Yang, Y. Zhang, Alternating direction algorithms for $\ell_1$-problems in compressive sensing, SIAM J. Sci. Comput. 33 (1) (2011) 250–278.
[25] A. Yang, Z. Zhou, A. Balasubramanian, S. Sastry, Y. Ma, Fast $\ell_1$-minimization algorithms for robust face recognition, IEEE Trans. Image Process. 22 (8) (2013) 3234–3246.
[26] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low rank representation, in: NIPS, 2011.
[27] E. Candés, X. Li, Y. Ma, J. Wright, Robust principal component analysis, J. ACM 58 (3) (2011).
[28] D. Bertsekas, Nonlinear Programming, Athena Sci. (2003).

**Liansheng Zhuang** received his Ph.D. degree and bachelor's degree respectively in 2006 and 2001, from the University of Science and Technology of China (USTC), China. Since 2006, he has served as a lecturer in the School of Information Science and Technology, USTC. During 2012–2013, he was a visiting research scientist in EECS in the University of California, Berkeley. His main research interesting is in computer vision, and machine learning. He is a member of ACM, IEEE, and CCF.

**Jingjing Wang** received the B.S. degree in the Department of Electronic Engineering and Information Science (EEIS) from the University of Science and Technology of China (USTC), Hefei, China, in 2010. He is now pursuing the Ph.D. degree in EEIS at USTC. His main research interests are machine learning and computer vision.

**Zhouchen Lin** (M'00–SM'08) received the Ph.D. degree in applied mathematics from Peking University in 2000. He is currently a professor with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University. He is also a Chair Professor with Northeast Normal University. He was a Guest Professor with Shanghai Jiaotong University, Beijing Jiaotong University, and Southeast University. He was also a Guest Researcher with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer vision, image processing, machine learning, pattern recognition, and numerical optimization. He is area chair of CVPR 2014, ICCV 2015, NIPS 2015, and AAAI 2016. He is an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and the International Journal of Computer Vision.

**Allen Y. Yang** is the co-founder and CTO of Grafty, Inc. He is also a principal investigator at UC Berkeley. Previously he served as CTO and acting COO of Atheer Inc. His primary research areas include high-dimensional pattern recognition, computer vision, image processing, and applications in motion segmentation, image segmentation, face recognition, and sensor networks. He has published three books/chapters, 13 journal papers and more than 30 conference papers. He is also an inventor of 12 US patents/applications. Allen received his BEng degree in Computer Science from the University of Science and Technology of China (USTC) in 2001. From the University of Illinois at Urbana-Champaign (UIUC), he received two MS degrees in Electrical Engineering and Mathematics in 2003 and 2005, respectively, and a Ph.D. in Electrical and Computer Engineering in 2006. Among the awards he received are a Best Bachelor's Thesis Award from USTC in 2001, a Henry Ford II Scholar Award from UIUC in 2003, a Best Paper Award from the International Society of Information Fusion and a Best Student Paper Award from Asian Conference on Computer Vision in 2009. He is a senior member of IEEE.



**Yi Ma** (F'13) received bachelors degree in Automation and Applied Mathematics from Tsinghua University, Beijing, China, in 1995. He received the MS degree in electrical engineering and computer science (EECS) in 1997, the MA degree in mathematics in 2000, and the PhD degree in EECS in 2000, all from the University of California, Berkeley. From 2000 to 2011, he was an associate professor of the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign (UIUC). From 2009 to 2014, he was a principal researcher and group manager of the Visual Computing Group at Microsoft Research Asia. He is currently a professor and the executive dean of the School of Information Science and Technology of ShanghaiTech University in China. His main research interests include computer vision and data science, and has written two textbooks "An Invitation to 3D Vision" and "Generalized Principal Component Analysis, all published by Springer. He is a Fellow of the IEEE. He has served as associate editor for International Journal on Computer Vision, IEEE Trans. on Pattern Analysis and Machine Intelligence, IEEE Trans. on Information Theory, IEEE Signal Processing Magazine, SIAM Journal on Imaging Sciences, and IMA Journal on Information and Inference. He was the recipient of the David Marr Best Paper Prize at the International Conference on Computer Vision in 1999 and Honorable Mention for the Longuet-Higgins Best Paper Award at the European Conference on Computer Vision in 2004. He received the CAREER Award from the US National Science Foundation in 2004 and the Young Investigator Program Award from the US Office of Naval Research in 2005. He has served as Area Chair for NIPS, CVPR and ICCV, Program Chair for ICCV 2013, and General Chair for ICCV 2015.

**Nenghai Yu** received his B.S. degree in 1987 from Nanjing University of Posts and Telecommunications, M.E. degree in 1992 from Tsinghua University and Ph.D. degree in 2004 from University of Science and Technology of China, where he is currently a professor. His research interests include multimedia security, multimedia information retrieval, video processing and information hiding.