IEEE Transactions on Image Processing

Robust Multi-View Graph Recovery with Applications to Unsupervised and Semi-Supervised Learning

Journal:	*Transactions on Image Processing*
Manuscript ID	TIP-17512-2017
Manuscript Type:	Regular Paper
Date Submitted by the Author:	27-Jul-2017
Complete List of Authors:	fang, xiaozhao; Shenzhen Graduate, Han, Na teng, shaohua Wu, Jigang Xu, Yong; Harbin Institute of Technology, the Bio-Computing Research Center, Shenzhen Graduate School Li, Xuelong; Chinese Academy of Sciences,
EDICS:	33. ARS-IIU Image and Video Interpretation and Understanding < Image and Video Analysis, Synthesis and Retrieval, 4. SMR-Rep Image and Video Representation < Image & Video Sensing, Modeling, and Representation, 23. ELI-STE Stereoscopic and Multiview Processing and Display < Electronic Imaging

Robust Multi-View Graph Recovery with Applications to Unsupervised and Semi-Supervised Learning

Xiaozhao Fang, *Member, IEEE,* Na Han, Shaohua Teng, Jigang Wu, *Member, IEEE,* Yong Xu, *Senior Member, IEEE,* and Xuelong Li, *Fellow, IEEE*

Abstract—Graph based methods have been widely applied in unsupervised and semi-supervised learning. The performance of these methods highly depends on the quality of the graph. In the real-world applications, the same object is commonly represented by different features, i.e., multi-view features, which lead to multiple graphs in multi-view learning. However, we usually do not know what kind of graph is important for the task in advance, and existing multi-view learning methods become weak in dealing with noisy graphs when the data is corrupted by noise. In this paper, we address the problems by first observing that the noises of each graph have specific structures. Then, based on this observation we propose a robust multi-view graph recovery (RMGR) method in which the specific structures are used to clean the multiple input noisy graphs and these cleaned graphs are simultaneously aggregated into a consensus graph by adaptively assigning great weighted coefficients for important graphs. In addition, the rank constraint is imposed on the Laplacian matrix of the consensus graph such that it has exactly c connected components, i.e., the c clusters. In doing so, the graph structure is adaptively adjusted during optimization to more accurately partition data. We propose an alternating optimization strategy to solve the optimization problem. Extensive experiments on synthetic and several benchmark data sets demonstrate the effectiveness of the proposed method.

Index Terms—Multi-view learning, Laplacian matrix, noisy graph, alternating optimization

1 Introduction

In many real-world applications, the same sample is commonly represented by multiple different features, which often provides information complementary to each other. For example, an image can be described by different features such as pixels, context and its labels. A webpage can be represented by its context, the text of webpage linking to the page. The challenge in multi-view learning is to fuse or integrate multiple features from different views to obtain better performance for tasks such as data analysis, clustering and classification [1].

Graph based learning methods exploit the relationships among samples to partition samples into different groups such that samples in the same group have high similarity to each other [2]. In the past decades, many graph based clustering methods have been proposed, such as flexible manifold embedding (FME) [3], Laplacian regularized least squares (LapRLS) [4], and semi-supervised learning using gaussian fields and harmonic functions (GFHF) [5], etc. The underly mechanism of graph based methods is to partition the data into respective groups based on the input graph/similarity matrix and thus the clustering results highly depend on the learning process of similarity matrix [2]. Although the remarkable clustering results achieved by using single feature representation in some cases, single feature representation cannot handle the realistic tasks to a satisfactory extent because different variations in lighting conditions, complex backgrounds, and scale changes may become obstacles for similarity learning [6]. To solve the issue, multiple features representation is introduced into the different learning tasks. Samples from different views representation are different in density of distribution, noise level, and neighborhoods. With these multiple features, the sample can be depicted comprehensively. However, these differences may bring about disagreement in different clustering tasks, which leads a challenging that how to efficiently reconcile the inharmonious information when we apply the multi-view learning methods [1]. Recently, lots of different methods for multi-view learning have been proposed among which two popular lines are 1) to combine many similarity matrices [7], [8], [9], multiple kernel [10], [11], [12], [13], [14], or multiple graphs [15], [16], [17], [18] together with an optimal weight. 2) to project samples from different views into a common latent subspace in which the conflict information is well reconciled [19], [20]. This paper mainly focus on the robust multi-view graph recovery (RMGR) from noisy data by the first line and apply the obtained graph to unsupervised and semi-supervised learning. Next, we will review some related methods in multiple graph

- *X. Fang, N. Han, S. Teng and J. Wu are with the School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China, (e-mail: xzhfang168@126.com, hannagdut@126.com, shteng@gdut.edu.cn, asjgwu@gmail.com,).*
- *Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, Guangdong, P. R. China. He is 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 (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China (e-mail: xuelong_li@opt.ac.cn).*
- *Corresponding author: N. Han (e-mail: hannagdut@126.com).*

Manuscript received January 08, 2016; revised April 17, 2016.

learning and multi-view clustering.

2 Related Works

Spectral clustering methods have been achieved the surprised performance such as normalized cut [21] and ratio cut [22]. In essence, most of spectral clustering methods are based on the manifold information of data/data similarity matrix (graph). When accessing to multi-view data, some researches based on multiple graph learning have been proposed to learn an optimal graph for capturing data similarity structure. Niu et al., [23] proposed a multiple non-redundant spectral clustering views method in which the non-redundant subspace was learned to provide multiple clustering solutions for the original problem. Karasuyama et al., [16] exploited multiple graph to perform label propagation. With the sparse integration, the multiple graph is properly aggregated by learning an adaptive weight. Nie et al., [24] proposed a parameter-free auto-weighted multiple graph learning method in which an optimal weight for each graph automatically was learned without introducing an additive parameter as previous methods do. Chaudhuri et al., [25] projected the samples from different views into a lower dimensional subspace and clustered these samples via canonical correlation analysis. Cai et al. [26] proposed a image clustering method by integrating heterogenous image features with graph. Some methods such as Xia et al., [27] and Li et al., [28] adaptively learn weights for graphs during optimization and thus these graphs can be suitably integrated. Similar methods can be found in [18], [15] and [14]. Zhou et al., [6] proposed to capture the structures of noises in each kernel and integrate them into a robust and consensus framework to learn a low-rank matrix. Most of these kind methods try to learn a consensus clustering by using a linear combination of multiple input graphs/similarity matrices. It is well known that the clustering results depend on the quality of the input graph. When the original data contain noises and outliers, such simple integration has no mechanism to address noises and outliers in input graphs. Especially, in the multi-view learning setting, if the data is corrupted by noises and outliers, all view data are inaccurate and then all input graphs are biased estimations of similarity relationship among data. Thus, the final clustering results are not satisfactory. Moreover, these methods first construct the graphs and then perform clustering over the graphs. It is obvious that the graph structure cannot be changed during clustering process, which, of course, is not reasonable. It seems plausible that the optimization phase should be allowed to change the graph structure such that it can better partition the data.

To solve these problems, in this paper, we propose a robust multi-view graph recovery (RMGR) method whose objective is to learn a consensus graph from multiple input graphs. We first observe that the noises of each input graphs have specific structures, Then, based on the analysis, we use a structured sparse matrix to model the noises such that the graphs are cleaned. To learn the consensus graph from multiple input graphs, we minimize the average disagreements between them. Difference from previous works which learn the consensus graph by using a simple linear combination, RMGE imposes exponential constraint on the weighted coefficients which not only identifies the graphs that are important for clustering but also provides more freedom for a better integration. To optimize the structure of consensus graph, RMGR imposes a rank constraint on the Laplacian matrix of the consensus matrix such that the final consensus graph has exactly c connected components (where c is the number of clusters). In other word, the graph structure can be automatically adjusted during learning such that it can exactly partition data points to respective group. The underlying optimization problem is difficult and we show it can be solved via alternating minimization. Experimental results on synthetic and benchmark data sets exhibit the effectiveness of the proposed method.

The rest of paper is structured as follows. Section 3 introduces our proposed method and gives the corresponding optimization algorithm. In Section 4, we demonstrate how to apply our method to unsupervised clustering and semi-supervised classification. We give the algorithm analysis in Section 5. In Section 6, we present our experimental results. Section 7 concludes the paper.

3 Robust multi-view graph recovery

In this section, we present a framework for robust multi-view graph recovery (RMGR), and then give the corresponding optimization algorithm.

3.1 Formulation

Suppose that we have n data points $X = [x_1, ..., x_n]^T \in \Re^{m\times n}$ in which m is the dimension of feature and n is the number of data points. Given the whole data set X, the adjacent matrix $W = \{w_{ij}\} \in \Re^{n\times n}, i = 1, .., n; j = 1, ..., n$ and the corresponding degree matrix D ($d_{ij} = \Sigma_{j=1}^{n} w_{ij}$) can be constructed. We define the cluster indicator matrix $F = [f_1, ..., f_n]^T \in \Re^{n\times c}$ (c is the number of classes). The classical Ratio Cut clustering can be written as

$$\min_{F^T F=I} Tr(F^T LF) \tag{1}$$

where $L = D - W$ is the graph Laplacian. (1) can be solved by eigenvalues calculating.

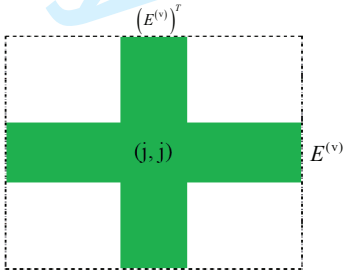


Fig. 1. Noise structure in the adjacent matrix corresponds to vth view.

For multi-view data, let M be the number of views and $X^{(1)}, ..., X^{(M)}$ be the data matrix of each view in which $X^{(v)} \in \Re^{m^{(v)}\times n}$ ($v = 1, ..., M$ and $m^{(v)}$ is the feature dimension of the the vth view). For each view, an adjacent matrix (graph) is constructed and thus we have $W^{(1)}, W^{(2)}, ..., W^{(M)}$ for all views. The task of our method is to learn a consensus graph W from $W^{(1)}, ..., W^{(M)}$. In

real-world applications, data points may contain noises and outliers (in the next sections, we uniformly represent noises and outliers as noises for simplification), and thus the adjacent matrix contains certain degree noises. In next, we will analyze the structure of noises. We have the following observation: when the jth data point is corrupted by noise, both the jth row and the jth column of the adjacent matrix are simultaneously contaminated (see Figure 1). To reduce the negative effect of noises, we, according to Liu et al. [29], introduce a structured sparse matrix to model noises so that we finally obtain a noise-free graph. First, we introduce a row-wise sparse matrix $E^{(v)}$ to model the corrupted components on the rows of the vth ($v = 1, 2, ..., M$) original input adjacent matrix. It is clear that $(E^{(v)})^T$ is a column-wise sparse matrix modeling the corrupted components on the columns of the vth original input adjacent matrix. For each view, we use $E^{(v)}$ and $(E^{(v)})^T$ to clean the graph $W^{(v)}$, thus we have the following objective.

$$W^{(v)} - (E^{(v)} + (E^{(v)})^T) \tag{2}$$

Therefore, the final consensus graph W can be learned from $W^{(v)}$ by using the following optimization objective

$$\begin{aligned} \min_{W,E^{(v)},\alpha,\rho} & \sum_{v=1}^{M} \alpha_{(v)} (\|W - (W^{(v)} - E^{(v)} - (E^{(v)})^T)\|_F^2 \\ & + \lambda_3 \|E^{(v)}\|_{2,1}) + \lambda_1 \|W\|_F^2 \\ s.t.\ & W \geq 0,\ w_i^T \mathbf{1} = 1,\ \sum_{v=1}^{M} \alpha_{(v)}^{\rho} = 1,\ \alpha_{(v)} \geq 0 \end{aligned} \tag{3}$$

where $\|E\|_{2,1}$ is the $\ell_{2,1}$-norm of matrix E and $\|E\|_{2,1} = \sum_{i=1}^{n} \|e^i\|_2$ (e^i is the ith row of E). $\mathbf{1}$ is a vector with all elements are 1. λ_3 and λ_1 are the regularization parameters and the $\|W\|_F^2$ is used to add a prior of uniform distribution. $\sum_{v=1}^{M} \alpha_{(v)}^{\rho} = 1$ ($\rho \in (0, 1)$) is to avoid trivial solution. The first term linearly combines M graphs using weighting coefficients $\alpha = (\alpha_{(1)}, \alpha_{(2)}, ..., \alpha_{(M)})^T$. $W \geq 0$ is used to ensure that the resulted W can be directly used as a real graph. We expose the exponential constraint $\sum_{v=1}^{M} \alpha_{(v)}^{\rho} = 1$ on weighted coefficients $\alpha_{(v)}$, which can guarantee that the important graph $W^{(v)}$ can be finely assigned to a great value of $\alpha_{(v)}$, i.e., $\alpha_{(v)}$ has more freedom to better accomplish the weight assignment.

If the data contain c classes, an ideal graph learning scheme should be has the ability to automatically adjust the graph structure such that the obtained graph has exactly c connected components. To achieve the goal, a direct method is to constraint the rank of Laplacian matrix of the resulted graph, such that the connected components are equal to the cluster number. If the resulted graph W is nonnegative, then the Laplacian matrix has the following property.

THEOREM 1 [30], [31] The multiplicity c of the eigenvalue 0 of the Laplacian matrix L_W is equal to the number of the connected components in the graph with the matrix W.

Theorem 1 indicates that if $rank(L_W) = n - c$ ($L_W = D_W - \frac{W+W^T}{2}$, D_W is a diagonal matrix whose jth diagonal element is $D_{Wjj} = \sum_j \frac{W_{ij}+W_{ji}}{2}$), then we can cluster data points into c clusters based on W. Thus, to learn an ideal consensus graph, we propose the following optimization problem.

$$\begin{aligned} \min_{W,E^{(v)},\alpha,\rho} & \sum_{v=1}^{M} \alpha_{(v)} (\|W - (W^{(v)} - E^{(v)} - (E^{(v)})^T)\|_F^2 \\ & + \lambda_3 \|E^{(v)}\|_{2,1}) + \lambda_1 \|W\|_F^2 \\ s.t.\ & W \geq 0, w_i^T \mathbf{1} = 1, rank(L_W) = n - c, \\ & \sum_{v=1}^{M} \alpha_{(v)}^{\rho} = 1, \alpha_{(v)} \geq 0 \end{aligned} \tag{4}$$

Suppose $\sigma_i(L_W) \geq 0$ (L_W is positive semi-definite) is the ith smallest eigenvalue of L_W, for a large enough value λ_2, the following problem holds.

$$\begin{aligned} \min_{W,E^{(v)},\alpha,\rho} & \sum_{v=1}^{M} \alpha_{(v)} (\|W - (W^{(v)} - E^{(v)} - (E^{(v)})^T)\|_F^2 \\ & + \lambda_3 \|E^{(v)}\|_{2,1}) + \lambda_1 \|W\|_F^2 + \lambda_2 \sum_{i=1}^{c} \sigma_i(L_W) \\ s.t.\ & W \geq 0, w_i^T \mathbf{1} = 1, \\ & \sum_{v=1}^{M} \alpha_{(v)}^{\rho} = 1,\ \alpha_{(v)} \geq 0 \end{aligned} \tag{5}$$

According to the Ky Fan's Theorem [32], we have $\sum_{i=1}^{c} \sigma_i(L_W) = \min_{F^T F = I} Tr(F^T L_W F)$. Therefore, the problem (5) is further rewritten as the following optimization objective

$$\begin{aligned} \min_{W,E^{(v)},\alpha,\rho} & \sum_{v=1}^{M} \alpha_{(v)} (\|W - (W^{(v)} - E^{(v)} - (E^{(v)})^T)\|_F^2 \\ & + \lambda_3 \|E^{(v)}\|_{2,1}) + \lambda_1 \|W\|_F^2 + \lambda_2 Tr(F^T L_W F) \\ s.t.\ & W \geq 0, w_i^T \mathbf{1} = 1,\ \sum_{v=1}^{M} \alpha_{(v)}^{\rho} = 1, \\ & \alpha_{(v)} \geq 0,\ \ F^T F = I \end{aligned} \tag{6}$$

where λ_2 and λ_1 are the balancing parameters. The key of our method is to use of error matrices $E^{(v)} + (E^{(v)})^T$ to clean the consensus graph W, which leads to that our formulation (6) is robust. Meanwhile, by using regularization term $Tr(F^T L_W F)$, we constraint the rank of Laplacian matrix L_W so that the connected component in the data points are exact the cluster number and each connected component corresponds to one cluster, and thus the cluster structure is very remarkable in final consensus graph W. It can be seen from (6) that our method simultaneously optimizes the consensus graph W and the clustering structure to obtain the optimal W.

The optimization problem in (6) involves four variables, i.e., W, $E^{(v)}$, α, and ρ. The direct optimization is difficult, thus we alternately minimize the objective function with respective to each variable.

First, by fixing the other variables, we optimize $E^{(v)}$ by solve the following problem

$$\min_{E^{(v)}} \|W - (W^{(v)} - E^{(v)} - (E^{(v)})^T)\|_F^2 + \lambda_3 \|E^{(v)}\|_{2,1} \tag{7}$$

By setting the derivative of (7) w.r.t $E^{(v)}$ to zero, we have

$$\begin{aligned}&(2I+\lambda_3 S^{(v)})E^{(v)}+2(E^{(v)})^T\\&=W^{(v)}-W-W^T+(W^{(v)})^T\end{aligned}\tag{8}$$

where $S^{(v)}$ is a diagonal matrix and the jth diagonal element is $\frac{1}{2\|E_j^{(v)}\|_2}$. Denote $A^{(v)}=W^{(v)}-W-W^T+(W^{(v)})^T$, $s_{jj}^{(v)}$ as the (j,j)th element of $S^{(v)}$ and $e_{jk}^{(v)}$ as the (k,j)th element in $E^{(v)}$. By considering the (j,k)th and (k,j) elements on both sides of (8), we have

$$W_{ij}=\begin{cases}e^{-\frac{\|x_i-x_j\|^2}{\sigma}} & \text{if } x_i \text{ and } x_j \text{ have the same labels}\\ 0 & \text{otherwise}\end{cases}\tag{9}$$

The solutions of $e_{jk}^{(v)}$ and $e_{kj}^{(v)}$ can be obtained by solving (**??**).

$$\begin{cases}e_{kj}^{(v)}=\frac{(2+\lambda_3 s_{jj}^{(v)})A_{kj}^{(v)}-2A_{jk}^{(v)}}{\lambda_3^2 s_{kk}^{(v)}+2\lambda_3 s_{kk}^{(v)}+2\lambda_3 s_{jj}^{(v)}}\\ e_{jk}^{(v)}=\frac{(2+\lambda_3 s_{kk}^{(v)})A_{jk}^{(v)}-2A_{kj}^{(v)}}{\lambda_3^2 s_{jj}^{(v)}+2\lambda_3 s_{jj}^{(v)}+2\lambda_3 s_{kk}^{(v)}}\end{cases}\tag{10}$$

It is well known that W, $W^{(v)}$ and $A^{(v)}$ are symmetric. Therefore, we written solution of $e_{kj}^{(v)}$ as

$$e_{kj}^{(v)}=\frac{s_{jj}^{(v)}A_{kj}^{(v)}}{\lambda_3 s_{kk}^{(v)}s_{jj}^{(v)}+2s_{kk}^{(v)}+2s_{jj}^{(v)}}\tag{11}$$

Second, by by fixing the other variables, we optimize W by solve the following problem

$$\begin{aligned}&\min_W\|W-\frac{\sum_{v=1}^M\alpha_{(v)}(W^{(v)}-E^{(v)}-(E^{(v)})^T)}{\sum_{v=1}^M\alpha_{(v)}}\|_F^2\\&+\lambda_1\|W\|_F^2+\lambda_2\sum_{i=1}^n\sum_{j=1}^n\|F_i-F_j\|_F^2W_{ij}\\&s.t.\ \sum_j w_{ij}=1, w_{ij}\geq 0\end{aligned}\tag{12}$$

Let $B=\frac{\sum_{v=1}^M\alpha_{(v)}(W^{(v)}-E^{(v)}-(E^{(v)})^T)}{\sum_{v=1}^M\alpha_{(v)}}$, we have the following optimization objective

$$\min_{\sum_j w_{ij}=1,w_{ij}\geq 0}\|w_{ij}-b_{ij}\|_2^2+\lambda_1 w_{ij}^2+\lambda_2 v_{ij}^T w_{ij}\tag{13}$$

where $v_{ij}=\|f_i-f_j\|_2^2$. Denote v_i as a vector with the jth element equal to v_{ij} (similarly for w_i and b_i), the problem (13) can be decomposed into n sub-problems and each sub-problem can be written as follows

$$\min_{w_i\mathbf{1}=1,w_i\geq 0}\frac{1+\lambda_1}{\lambda_2}\|w_i-(\frac{1}{1+\lambda_1}b_i-\frac{v_i}{\frac{2(1+\lambda_1)}{\lambda_2}})\|_2^2\tag{14}$$

The solution for (14) is introduced in the appendix A.

Third, by by fixing the other variables, we optimize F by solving the following problem

$$\min_{F^TF=I}Tr(F^TL_WF)\tag{15}$$

We obtain the solution F formed by the c eigenvectors of L_W corresponding to the c smallest eigenvalues.

Fourth, by by fixing the other variables, we optimize α by solve the following problem

$$\min_\alpha\sum_{v=1}^M\alpha_{(v)}q_{(v)},\ s.t.\ \sum_{v=1}^M\alpha_{(v)}^\rho=1,\ \forall i,\ \alpha_{(v)}\geq 0\tag{16}$$

where $q_{(v)}=\|W-(W^{(v)}-E^{(v)}-(E^{(v)})^T)\|_F^2+\lambda_3\|E^{(v)}\|_{2,1}$. By solving problem (16), we obtain the solution of $\alpha_{(v)}$, i.e., $\alpha_{(v)}=\frac{q_{(v)}^{\frac{1}{\rho-1}}}{(\sum_{(v)}q_{(v)}^{\frac{\rho}{\rho-1}})^{\frac{1}{\rho}}}$.

The detailed process of solving problem (16) can be found in the appendix B.

The detailed optimization algorithm to solve problem (6) is summarized in **Algorithm** 1.

Algorithm 1 :Algorithm to solve problem (6)

Input: M adjacent matrices $W^{(v)}$(v=1,2,...,M); Parameters λ_1, λ_2; λ_3; ρ; Cluster number c.
Initialization: $E=\mathbf{0}$; W is the mean value of M similarity matrices of all views based on the k nearest neighbor graph; F is formed by the c eigenvectors of $L=D-\frac{W+W'}{2}$ corresponding to the c smallest eigenvalues.
while not converged **do**
1. Update $E^{(v)}$ by (11);
2. Update W by solving (14);
3. Update F by solving (15);
4. Update α by solving (16)

end while
Output: Consensus graph W; Cluster indicator matrix F

4 RMGR for Unsupervised Clustering and Semi-supervised Classification

For simplicity, we take the classic Ratio Cut in the following statement. When we obtain the solutions of W and F following the optimization in **Algorithm** 1, we can perform unsupervised clustering and semi-supervised classification by them.

4.1 RMGR for Unsupervised Clustering

In spectral clustering, it is well known that the cluster indicator matrix F must satisfy (1). When we obtain the final solution of F by **Algorithm** 1, then we treat each row of F as a new representation of each data point and compute the prediction labels by using k-means algorithm. The detailed process of RMGR for unsupervised clustering is summarized in **Algorithm** 2.

4.2 RMGR for Semi-supervised Classification

The goal of our method is to learn a robust consensus graph W from multiple input matrices and the semi-supervised classification results highly depend on the quality of W. Thus, in this section we focus on evaluate the quality of W by conducing semi-supervised classification experiments based on learned W.

For simplicity, we take the single view data in the following statement. Given data points $X=$

Algorithm 2 :Algorithm of RMGR for unsupervised clustering

Input: M adjacent matrices $W^{(v)}$(v=1,2,...,M); Cluster number c.
Steps:
1: Solve problem (6) by **Algorithm** 1 and obtain an optimal solution F^*.
2: Treat each row of F as a new representation of each data point and compute the prediction labels by using k-means algorithm.
Output: Cluster indictor labels of each data point.

$[x_1, x_2, ..., x_u, x_{u+1}, ..., x_n] \in \Re^{m\times n}$, where $x_i|_{i=1}^{u}$ and $x_i|_{u+1}^{n}$ are the labeled and unlabeled data points, respectively. The goal of semi-supervised classification is to accurately classify these unlabeled data points into their respective classes. The labels of labeled data points are denoted as $y_i \in \{1, 2, ..., c\}$. The binary label matrix $Y = [y_1, y_2, ..., y_n]^T$ is defined as follows: for each data point x_i $(i = 1, 2, ..., n)$, $y_i \in \Re^c$ is its label vector, if x_i is from the kth class $(k = 1, 2, ..., c)$, then only the kth entry of y_i is one and all the other entries are zero.

Gaussian fields and harmonic functions (GFHF) [5] estimates a prediction label matrix $F \in \Re^{n\times c}$ on the learned graph $W^{n\times n}$ with respect to the label fitness and the manifold smoothness. The label fitness means that F should be close to the given labels for the labeled nodes, *i.e.*, $\sum_{i=1}^{u} \|F_i - Y_i\|_F^2$ and the manifold smoothness means that F should be smooth on the whole graph of both labeled and unlabeled nodes, *i.e.*, $\sum_{i,j=1}^{n} \|F_i - F_j\|^2 W_{ij}$. Therefore, the objective function of GFHF is

$$\min_F \frac{1}{2} \sum_{i,j=1}^{n} \|F_i - F_j\|^2 W_{ij} + \lambda_\infty \sum_{i=i}^{u} \|F_i - Y_{i.}\|_F^2 \qquad (17)$$

where λ_∞ is a very large number such that $\sum_{i=1}^{u} \|F_i - Y_i\|_F^2$ is approximately satisfied. According to [3], the objective function of GFHF can be rewritten as

$$\min_F Tr(F^T L F) + Tr(F - Y)^T U (F - Y) \qquad (18)$$

where L is the graph Laplacian matrix which is denoted as $L = D - W$, where D is a diagonal matrix with the diagonal elements as $D_{ii} = \sum_j W_{ij}$. U is also a diagonal matrix with the first u and the rest $n - u$ diagonal elements as λ_∞ and 0, respectively. F is solved by setting the derivative of (17) with respect to F to zero.

$$F = (L + U + \varepsilon I)^{-1}(UY) \qquad (19)$$

where ε is a small positive constant whose goal is to obtain numerically more stable solution for F. The final classification results are obtained by using the KNN algorithm on the obtained labels F.

The detailed process of applying RMGR to semi-supervised classification is presented in **Algorithm** 3.

Algorithm 3 :Algorithm of RMGR for semi-supervised classification

Input: M adjacent matrices $W^{(v)}$(v=1,2,...,M).
Steps:
1: Construct binary label matrix Y.
2: Solve problem (6) by **Algorithm** 1 and obtain an optimal solution W^*.
3: Carry out the semi-supervised classification on the obtained W using the existing graph based semi-supervised classification method GFHF.
4: Use KNN algorithm to perform classification on the obtained F.
Output: Classification results.

5 Algorithm Analysis

In this section, we first verify that the updating $E^{(v)}$ as (7) can monotonically decreases the objective function (6). Then, we analyze the convergence behavior of the proposed optimization algorithm in **Algorithm** 1. Finally, we give the complexity analysis of algorithm.

5.1 Convergence Analysis

To prove the convergence behavior, we need the following lemmas.

Lemma 1. For any two non-zero constants a and b, the following inequality holds [33].

$$\sqrt{a} - \frac{a}{2\sqrt{b}} \leq \sqrt{b} - \frac{b}{2\sqrt{b}} \qquad (20)$$

Proof. The detailed proof is similar as that in [33].

Lemma 2. The following inequality holds provided that $\nu_t^i|_{i=1}^{r}$ are non-zero vectors, where ν is an arbitrary number [33].

$$\begin{aligned} &\sum_i \|\nu_{t+1}^i\|_2 - \sum_i \frac{\|\nu_{t+1}^i\|_2^2}{2\|\nu_t^i\|_2} \\ &\leq \sum_i \|\nu_t^i\|_2 - \sum_i \frac{\|\nu_t^i\|_2^2}{2\|\nu_t^i\|_2} \end{aligned} \qquad (21)$$

Proof. By respectively replacing $\|\nu_{t+1}^i\|_2^2$ and $\|\nu_t^i\|_2^2$ with a and b, the following inequality holds for any i

$$\|\nu_{t+1}^i\|_2 - \frac{\|\nu_{t+1}^i\|_2}{2\|\nu_t^i\|_2} \leq \|\nu_t^i\|_2 + \frac{\|\nu_t^i\|_2^2}{2\|\nu_t^i\|_2} \qquad (22)$$

Jointly considering (21) and (22) over i, (21) holds. □

Theorem 1. Updating $E^{(v)}$ as (7) can monotonically decreases the objective function (6).

Proof. According to the definition of E_t (For simplicity, we only consider the single view data) in the (8), we can see that

$$\begin{aligned} E_t &= \arg\min \|W' + E + E^T\|_F^2 + \triangle + \lambda_3 \|E\|_{2,1} \\ &= \arg\min \|W' + E + E^T\|_F^2 + \triangle + \lambda_3 \sum_{i=1}^{n} \|e^i\|_2 \end{aligned} \qquad (23)$$

where $W' = W - W^{(v)}$, $\triangle = \lambda_1 \|W\|_F^2 + \lambda_2 Tr(F^T L_W F)$ and e^i is the ith row of E. For the single view data, $\alpha_{(v)}$ is a constant and thus we omit it in (23).

Thus, we have

$$\begin{aligned}&\|W'+E_t+E_t^T\|_F^2 \le \|W'+E_{t-1}+E_{t-1}^T\|_F^2\\ &\Rightarrow \|W'+E_t+E_t^T\|_F^2+\lambda_3\sum_i\frac{\|e_t^i\|_2^2}{2\|e_{t-1}^i\|_2}\\ &\le \|W'+E_{t-1}+E_{t-1}^T\|_F^2+\lambda_3\sum_i\frac{\|e_{t-1}^i\|_2^2}{2\|e_{t-1}^i\|_2}\end{aligned} \tag{24}$$

Then, the following inequality holds

$$\begin{aligned}&\|W'+E_t+E_t^T\|_F^2+\lambda_3\sum_i\|e_t^i\|_2\\ &-\lambda_3(\sum_i\|e_t^i\|_2-\sum_i\frac{\|e_t^i\|_2^2}{2\|e_{t-1}^i\|_2})\\ &\le \|W'+E_{t-1}+E_{t-1}^T\|_F^2+\lambda_3\sum_i\|e_{t-1}^i\|_2\\ &-\lambda_3(\sum_i\|e_{t-1}^i\|_2-\sum_i\frac{\|e_{t-1}^i\|_2^2}{2\|e_{t-1}^i\|_2})\end{aligned} \tag{25}$$

Meanwhile, according to Lemma 2, we have $\sum_i\|e_t^i\|_2-\sum_i\frac{\|e_t^i\|_2^2}{2\|e_{t-1}^i\|_2}\le\sum_i\|e_{t-1}^i\|_2-\frac{\|e_{t-1}^i\|_2^2}{2\|e_{t-1}^i\|_2}$. Therefore, we have the following inequality

$$\begin{aligned}&\|W'+E_{t+1}+E_{t+1}^T\|_F^2+\triangle+\lambda_3\sum_i\|e_{t+1}^i\|_2\\ &\le \|W'+E_t+E_t^T\|_F^2+\triangle+\lambda_3\sum_i\|e_t^i\|_2\end{aligned} \tag{26}$$

which indicates that in each iteration, the objective function value of $\arg\min_E\|W'+E+E^T\|_F^2+\triangle+\lambda_3\sum_{i=1}^n\|e^i\|_2$ monotonically decreases using our proposed updating rule. □

Theorem 2. The iterative optimization in **Algorithm** 1 converges.

Proof. Theorem 1 shows that updating E^v can monotonically decreases the objective function value of (6). When updating the other variables, we find that they have analytical solution and thus the global optima of each sub-problem can monotonically decreases the objective function. In additions, the objective function value of (6) is always greater than 0. Therefore, **Algorithm** 1 converges. □

5.2 Complexity Analysis

The main computation cost of RMGR is to solve the eigenequation of (15). Thus, the complexity is $\mathcal{O}(n^3)$ for each iteration. When updating $E^{(v)}$, the complexity is only $\mathcal{O}(n^2m)$ since there are only element-wise optimizations for each iteration. When updating W, the complexity is $\mathcal{O}(mn)$. The complexity of computing α is just costs $\mathcal{O}(m)$. To sum up, in one iteration, the time complexity is $\mathcal{O}(n^3+n^2m+nm+m)$. If the algorithm needs φ iterations steps, then the total computational cost is $\mathcal{O}(\varphi(n^3+n^2m+nm+m))$. As shown in the experimental section, the proposed optimization algorithm usually converges very fast and, thus φ is a small number, which indicates that the proposed optimization algorithm is still efficient in most cases.

6 Experiments

In this section, we evaluate our method on the tasks of unsupervised clustering and semi-supervised classification. First, we conduct experiments on the synthetic data. Second, we conduct the unsupervised clustering experiments to compare our method with state-of-the-art methods on four multi-view data. Third, we conduct the semi-supervised experiments to compare different methods. Lastly, we study the performance variation of our method with respect to the different regularization parameters and algorithmic convergence. Our code and related data will be released online (http://www.yongxu.org/lunwen.html) if our paper is accepted.

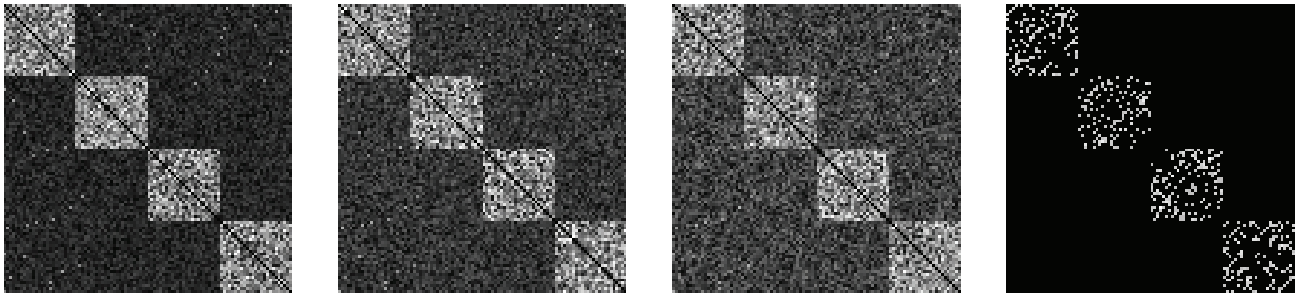
6.1 Experiments on the synthetic data

The first experiment is to test the robustness of our method. The synthetic data set we used is a 100×100 matrix with four 25×25 block matrices diagonally arranged. These points within each block denotes the similarity (affinity) of points on one cluster, while the points outside all blocks denotes noise. These points within each block is randomly generated in the range of 0 and 1, while the noisy points is randomly generated in the range of 0 and c. In our experiments, c is set as 0.4, 0.5 and 0.6, respectively. We directly use these original graphs as different view graphs and the final output W is the recovery graph. Moreover, to make the graph recovery task more challenging, we randomly pick out 25 noisy points and set their values to be 1. Figure 2 shows the original random matrices/graphs and the graph recovery results. We can notice that our method exhibits good performance in the graph recovery task.

We also test the quality of the graph learned from different view graphs. To simulate, we sample three sets (three view data sets) from the standard Gaussian. We also add different noises on these data sets. The weight matrix $W^{(v)}$ $(v=1,2,3)$ of each view data is defined as $W_{ij}^{(v)}=e^{-\frac{||x_i^{(v)}-x_j^{(v)}||_2^2}{\sigma}}$, where $x_i^{(v)}$ and $x_j^{(v)}$ denote two data points from the vth view data set and σ denotes the width of the heat kernel. We select the optimal value of σ by using ridge search method. The similarity matrices (graphs) of these three data sets are shown in Figure 3 (a), (b) and (c) from which we can see that the block diagonal structures are not obvious, particularly in Figure 3 (b) and (c). The graph in Figure 3 (d) is the graph learned by our method. We also see that although the block diagonal structure of each view graph is very obscure, we can finally learn a graph with block diagonal structures by our proposed method.

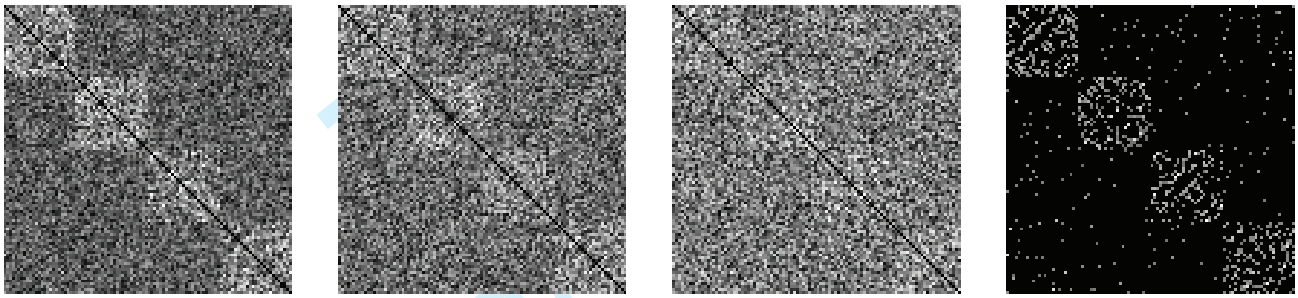
6.2 Experiments of unsupervised clustering on the multi-view data

For all unsupervised clustering, we use two metrics, purity and normalized mutual information (NMI) to evaluate all compared methods. All experiments are run 50 times (unless otherwise stated) and then the mean result and standard deviation are reported. Besides comparing single view methods, i.e., K-means (Euclidean distance) and spectral clustering (SC) [36], we also compare with some state-of-the-art multi-view methods: co-regularized spectral clustering (CoregSC) [8], multi-model spectral clustering



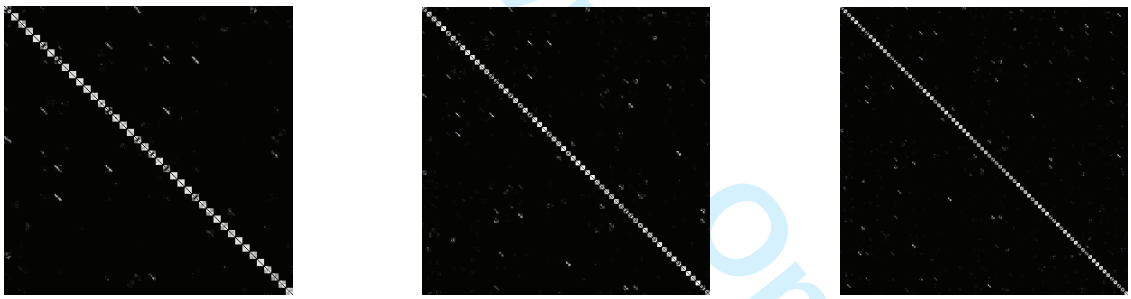
(a) Original graph, noise=0.4 (b) Original graph, noise=0.5 (c) Original graph, noise=0.6 (d) Graph recovery result

Fig. 2. Graph recovery from noisy graphs. Please note that we use the three original input matrices as different view graphs.



(a) The first view graph (b) The second view graph (c) The third view graph (d) Graph recovery result of our method

Fig. 3. Graph recovery from different view graphs.



(a) #40 clusters (b) #60 clusters (c) #80 clusters

Fig. 4. Graph recovery from the COIL100 data set.

(MMSC) [7], multi-view spectral clustering (MVSC) [17], and auto-weighted multiple graph learning (AMGL) [24]. For each single view clustering method (K-means, SC), we give the experimental results of the best view in the following experiments. The clustering experiments are conducted with a range of number of cluster c. So we use the first c classes in the data set for testing.

The first multi-view data used in our experiments is the COIL100 [1] data set. The COIL100 data set contains 100 objects and each object has 72 images. The images of each objects were taken 5 degrees apart as the object is rotated on a turntable. The size of each image is 32×32 pixels, with 256 grey levels per pixel. Thus, each image is represented by a 1024-dimensional vectors. In the experiment, we partition the data set into 9 subsets COIL1, COIL2,...and COIL9. COIL1, COIL2,...,COIL9 contain all images respectively taken in the directions of

$$
\begin{aligned}
COIL1 &\subset [0^0, 15^0] \cup [340^0, 355^0] \\
COIL2 &\subset [20^0, 35^0] \cup [320^0, 335^0] \\
COIL3 &\subset [40^0, 55^0] \cup [300^0, 315^0] \\
&\dots \\
COIL7 &\subset [120^0, 135^0] \cup [220^0, 235^0] \\
COIL8 &\subset [140^0, 155^0] \cup [200^0, 215^0] \\
COIL9 &\subset [160^0, 175^0] \cup [180^0, 195^0]
\end{aligned}
$$

In this way, we construct nine subsets with relatively different views. And each view have 800 images and each class has 8 images. Table 1 shows the clustering purity. Our method constantly outperforms all compared methods. We also find that K-means outperforms many multi-view methods in most cases. The reason may be that these data

1. Available at http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php.

TABLE 1
Clustering purity of different methods on the COIL100 data set

Clusters #	K-means	SC	CoregSC	MMSC	MVSC	AMGL	RMGR
20	85.39±3.46	77.37±0.05	74.96±2.01	76.64±1.22	79.00±0.78	77.50±0.87	**94.94±0.35**
40	83.44±2.41	73.17±0.04	78.56±2.13	78.89±1.34	80.81±0.98	79.81±1.01	**87.10±1.73**
60	80.11±1.17	68.69±0.03	74.24±1.67	75.50±2.01	74.90±0.96	72.23±1.21	**83.69±1.46**
80	77.53±1.75	68.24±0.03	72.23±2.50	71.20±1.29	72.78±0.95	70.63±0.96	**81.00±1.41**
100	77.27±1.56	66.90±0.02	67.89±1.87	69.50±1.30	72.22±0.79	69.35±1.33	**79.00±1.54**
Avg.	80.74±2.07	70.87±0.03	73.58±2.04	74.35±1.43	75.94±0.89	73.90±1.07	**85.14±1.30**

TABLE 2
Clustering purity of different methods on the CMU PIE data set

Clusters #	K-means	SC	CoregSC	MMSC	MVSC	AMGL	RMGR
4	56.62±3.66	68.81±0.07	79.96±0.17	84.20±0.17	88.85±0.12	86.23±0.08	**100.00±0.00**
12	41.53±2.84	59.46±0.02	79.87±0.18	82.56±0.15	86.37±0.14	84.43±0.06	**100.00±0.00**
20	36.55±1.87	50.72±0.02	78.68±0.21	80.25±0.09	83.90±0.23	83.43±0.04	**88.44±2.00**
28	33.50±1.58	48.70±0.01	75.85±0.30	78.89±0.11	82.32±0.11	81.74±0.05	**86.63±2.05**
36	33.84±1.29	49.42±0.01	74.38±0.24	78.86±0.08	82.06±0.15	80.26±0.04	**85.30±1.43**
44	32.02±1.31	49.48±0.03	74.30±0.13	79.90±0.13	82.56±0.14	81.83±0.03	**85.87±1.17**
52	31.27±1.11	49.23±0.01	75.75±0.20	77.63±0.21	80.80±0.12	79.91±0.03	**84.36±1.83**
60	30.04±1.18	46.98±0.01	71.50±0.22	76.20±0.12	77.90±0.13	78.03±0.02	**82.16±1.12**
68	28.98±0.90	46.01±0.01	72.47±0.26	75.60±0.23	77.85±0.15	78.31±0.03	**83.14±1.35**
Avg.	36.03±1.75	51.79±0.02	75.86±0.21	80.45±0.14	82.51±0.14	81.58±0.04	**88.43±1.21**

TABLE 3
Clustering purity of different methods on the MSRC-V1 data set

Clusters #	K-means	SC	CoregSC	MMSC	MVSC	AMGL	RMGR
2	82.35±15.43	84.08±0.08	84.78±0.36	85.67±0.22	90.22±0.21	88.78±0.17	**98.33±0.03**
3	80.59±13.15	80.11±0.02	79.80±0.45	80.63±0.30	**83.70±0.26**	81.25±0.12	79.26±1.39
4	70.75±1.29	70.08±0.07	66.71±0.37	65.32±0.42	69.00±0.42	67.44±0.06	**92.41±0.58**
5	68.46±2.21	70.53±0.06	67.90±0.78	68.80±0.38	73.99±0.50	70.96±0.08	**90.66±0.02**
6	72.16±2.77	72.64±0.06	73.80±0.67	74.50±0.46	79.89±0.44	76.22±0.07	**86.44±0.67**
7	72.60±2.34	71.17±0.04	72.46±0.62	75.08±0.34	80.54±0.53	78.23±0.05	**85.49±0.52**
Avg.	74.48±6.19	74.77±0.06	74.24±0.54	75.00±0.35	79.56±0.39	77.15±0.09	**88.77±0.53**

TABLE 4
Clustering purity of different methods on the CiteSeer data set

Clusters #	K-means	SC	CoregSC	MMSC	MVSC	AMGL	RMGR
2	48.25±11.74	49.25±0.10	48.85±0.11	50.46±0.21	51.90±0.20	50.06±0.07	**53.25±0.01**
3	55.78±5.71	54.47±0.22	52.78±0.20	56.30±0.18	56.63±0.15	55.89±0.05	**61.33±0.02**
4	53.60±5.10	52.67±0.26	52.36±0.25	55.87±0.23	56.84±0.16	54.23±0.03	**62.09±0.11**
5	18.56±2.33	20.35±0.36	18.30±0.18	21.10±0.26	21.40±0.23	20.76±0.07	**23.89±0.37**
6	33.58±2.30	42.80±0.16	41.80±0.24	44.69±0.19	46.96±0.15	42.89±0.12	**58.08±0.36**
Avg.	41.95±5.44	43.91±0.22	35.68±0.19	45.73±0.34	46.75±0.18	44.77±0.07	**51.72±0.17**

points in the COIL100 meet the convex shape of cluster and thus the Euclidean distance can be used as a suitable measurement for clustering them. The graph recovery results from the COIL100 data set are shown in Figure 4 from which we again see that our method can recover an ideal graph for clustering task.

The second multi-view data used in this experiments is the CMU PIE [2] data set which contains 41368 face images from 68 persons and these images have "pose", "illumination", and "expression" changes. We select five subsets (each subset corresponding to a distinct pose) from the CMU PIE data set as multi-view data. These five subsets are C05 (left pose), C07 (upward pose), C09 (downward pose), C27 (front pose), and C29 (right pose). These images in each subset are taken under different illumination and expression conditions. The resolution of images in our experiments is of 64×64. We randomly select 20 images from each class and thus the number of all images in each view is $20 \times 68 = 1360$. In doing so, each view data follows different poses. The experimental results are shown in Table 2. Again, our method obtains the best clustering results and the clustering purity reaches 100% in the first two cases. In general, almost every multi-view clustering methods obtains the better clustering results than single view learning clustering results.

The third multi-view data is the MSRC-v1 data set [24] which contains 240 images from 8 classes. Following Lee et al. [34], 7 classes composed of airplane, tree, cow, face, bicycle, building, and car are selected in our experiments. Each class has 30 images. To construct multi-view data, we extract 5 visual features from each image: 576 histogram of oriented gradient, 24 color moment, 256 local binary pattern, 512 gist, and 254 centrist features. The experimental results are shown in Table 3. It can be seen that our method outperforms the other methods in almost all of cases.

The fourth multi-view data is the CiteSeer data set[3] [35] , which consists of 3,312 documents that are about scientific publications. These documents can be further classified into

2. Available at http://www.ri.cmu.edu/research_project_detail.html?project_id=418.

3. Available at http://archive.ics.uci.edu/ml/.

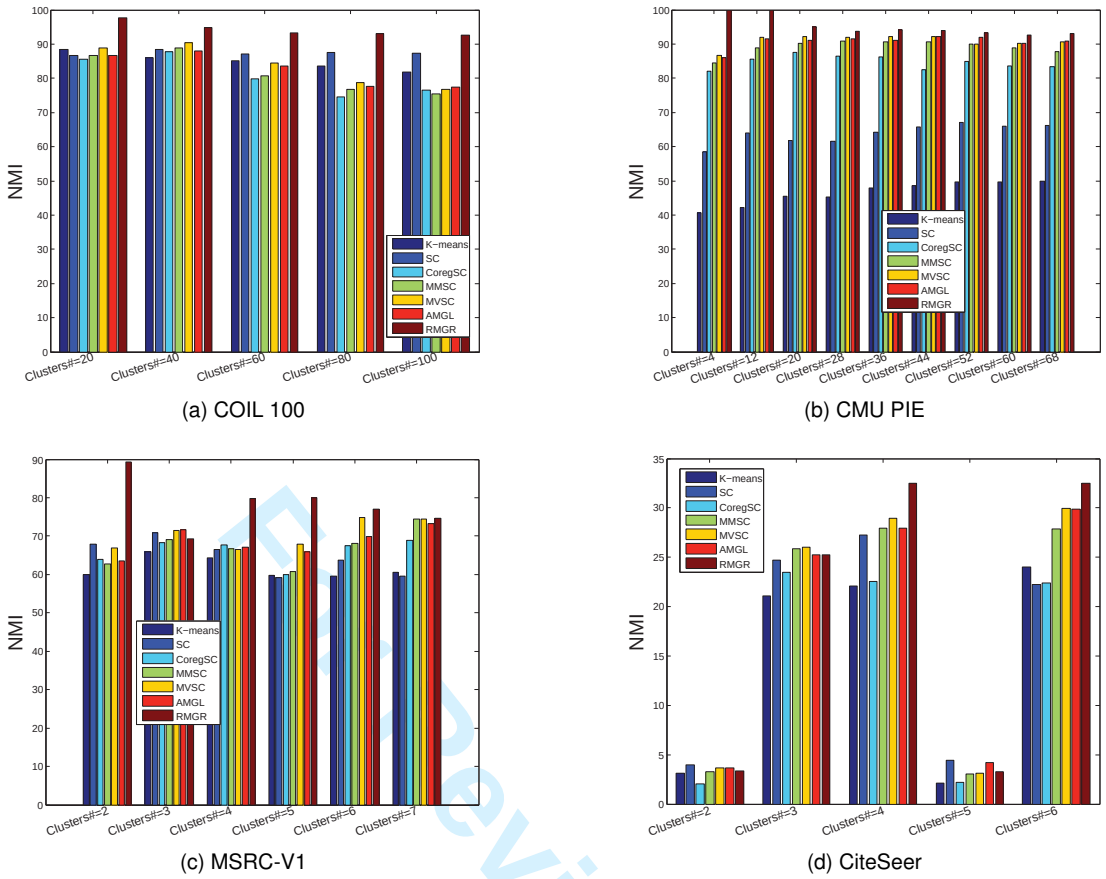


Fig. 5. The NMI of different methods on these four data sets.

6 classes: Agents, AI, DB, IR, ML and HCI. In our experiments, a 3,703-dimensional vector representing whether the key words are included for the text view, and the other 3279-dimensional vector that records the citing relations between every two documents are built up. The experimental results are shown in Table 4, in which we can see that our method obtains the best performance in unsupervised clustering.

The experimental results of NMI (NMI$\times$100) on the four data sets are shown in Figure 5. Please note that the greater of the number of NMI the better of the clustering performance. From the results, our method again obtains the best performance. Especially, the improvements of NMI on many cases are very obvious.

6.3 Experiments of semi-supervised classification on the multi-view data

To test the semi-supervised classification ability of different methods, we conduct the semi-supervised classification experiments on these four data sets. We compare our method with the single view method GFHF (the result is obtained from the best view) and multi-view methods: adaptive model semi-supervised classification (AMMSS) [26], sparse multiple graph integrations (SMGI) [16] and AMGI [24]. For each data set, we randomly select Labeled#=τ samples per subject as the labeled samples and remaining samples are used as unlabeled samples. The goal of semi-supervised classification is to reduce the number of labeled samples. So we are more interested in the performance of these methods with low labeling ratio such as $\tau = 2$. We report the mean classification accuracy (%) over 30 random splits on the labeled and unlabeled samples. The final experimental results are reported in Figure 6, which shows that our method obtains the good performance. For example, the improvement of classification accuracies(%) are very obvious on the CMU PIE and CiteSeer data sets.

6.4 Experiments analyses

From the experimental results of unsupervised clustering and semi-supervised classification, we can obtain the following observations.

1) In general, our method outperforms almost all of compared methods in the unsupervised clustering setting. In many cases, the improvements are very significant. This means that the graph recovery results of our proposed method are effective. For example, face images in the CMU PIE data set have strong variations (e.g., poses, illuminations and expressions et al.), all the other methods learn inappropriate graphs and thus the clustering performance are significant deteriorated. However, our method can recovery a good consensus graph by considering the cluster structure and reducing the negative effect of noises. Such observations indicate that the cluster structure is very significant for recovering an ideal graph structure from the multiple input graphs. In other words, graph structure should be dynamic during learning such that the structure of graph can coincides with the requirement of sample partitioning.

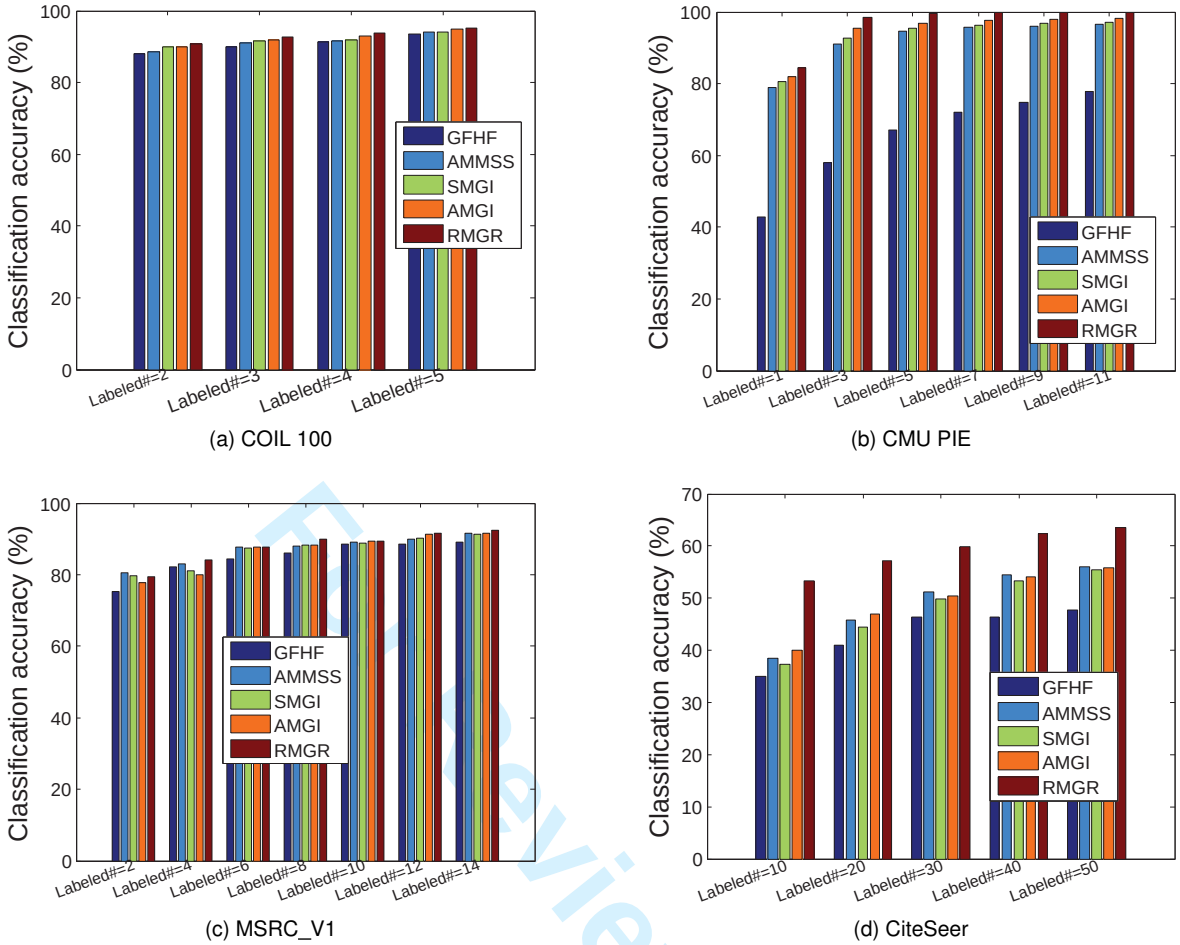


Fig. 6. Semi-supervised classification results of different methods on these four data sets.

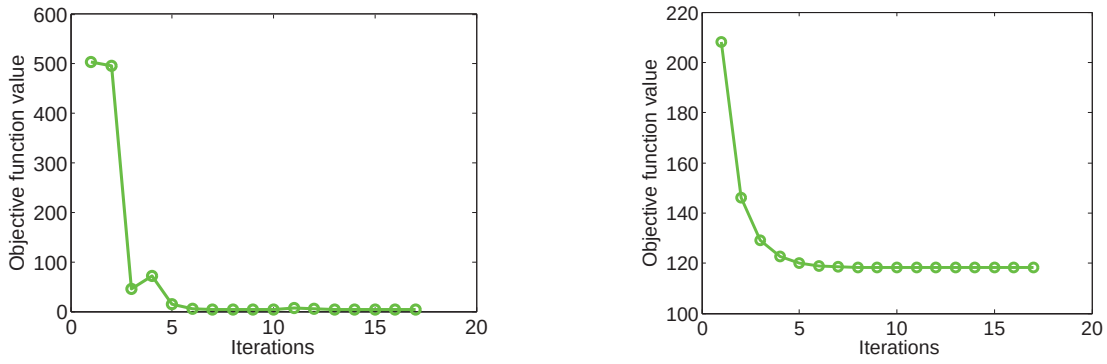
2) From the results in the Table 2 and Table 3, it can be seen that the performance of many multi-view clustering methods is inferior to that of the single clustering methods with the best view result. For example, when we use the first four class samples to test different methods, the clustering results of CoregSC and MMSC are not satisfactory. The reason may be that these samples in the fourth class contain large variations which makes the graph learning process of CoregSC and MMSC subject to these variations of samples. As a result, the weighting coefficient corresponding to the important view may be assigned a small value owing to the simple linear weighted coefficients combination. Thus, the graphs obtained by these methods are the biased estimations of the optimal graphs, which degrades the performance of clustering. However, our proposed method overcomes the problem by introducing a structured sparse matrix to compensate noises and using the exponential weighted coefficients combination which can identify the graphs of important view. Thus, our method obtains competitive clustering results by reducing the effect of negative factors.

3) The proposed method outperforms other methods in terms of semi-supervised classification accuracy (%), which indicates that the graph structure recovered by our method encodes more discriminant information and thus can more effectively propagate labels from the labeled samples to unlabeled samples.

6.5 Algorithmic convergence and parameter sensitivity

To solve the optimization objective in (6), we propose an iterative update rule. The theory of convergence is proven in the Section 5. The convergence behaviors of our method on the MSRC and CMU PIE data sets are shown in Figure 7. From the results in Figure 7, we can see that objective values of RMGR usually converge in less than 10 iterations, which indicates that the optimization algorithm is effective. We have similar observations on other data sets.

In our proposed method, three parameters λ_1, λ_2 and λ_3 are required to be set in advance. To validate how the parameters affect the performance of our proposed method, we conduct experiments to evaluate their sensitivity. Figure 7 shows the the performance variance *w.r.t.* λ_1, λ_2 and λ_3 in terms of unsupervised clustering result, e.g., purity on the COIL 100 and CMU PIE data sets. It is observed that the clustering performance is sensitive to λ_1 when $\lambda_1 \leq 10^3$. However, when we set $\lambda_1 > 10^3$ the clustering performance is good. For example, when the value of λ_1 is set to 10^5, the clustering performance of our method is good on the these two data sets. We also test sensitivities of λ_2 and λ_3 when we fix the value of λ_1. It is observed that the clustering performance varies corresponding to different values of λ_2 and λ_3. When the value of λ_3 in the range of $[10^0, 10^5]$, our method is not sensitive to λ_3 and the performance is good.



(a) MSRC (#4 clusters) (b) PIE (#20 clusters)

Fig. 7. Convergence curves of our method (unsupervised clustering).

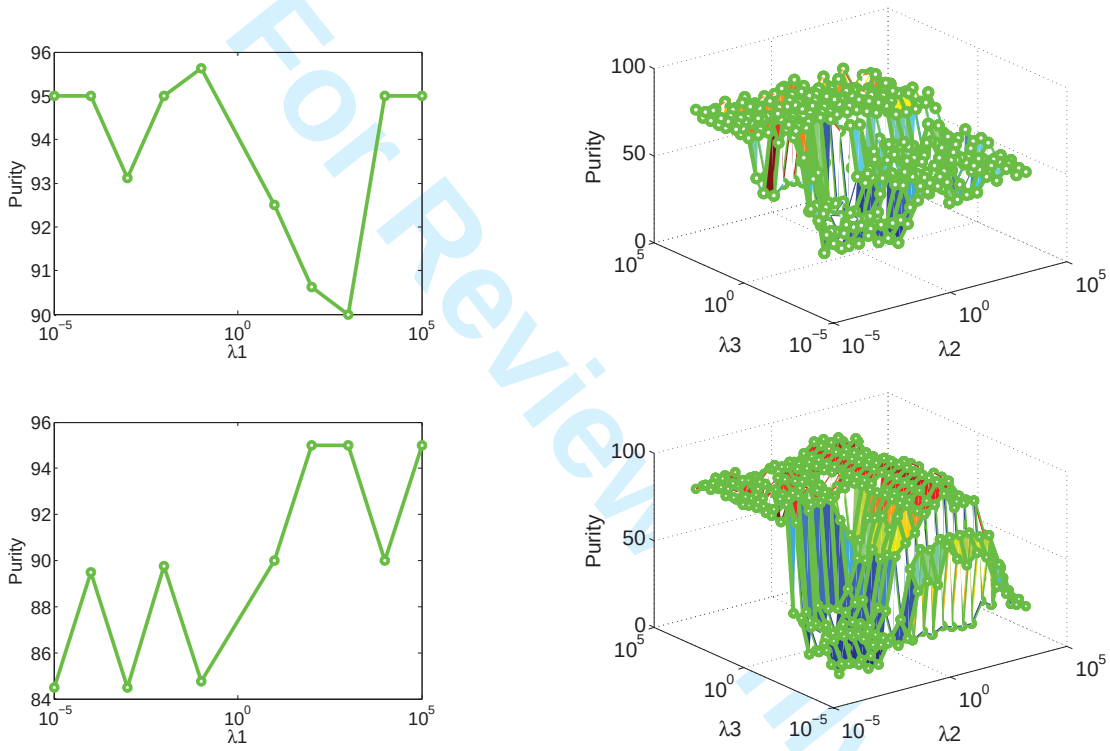


Fig. 8. Parameters sensitivity (λ_1,λ_2 and λ_3) of our method on the COIL100 and CMU PIE data sets, where we set $c = 20$. The graphs in first and second rows are the experimental results on the COIL100 and CMU PIE data sets, respectively.

However, we also observe that our method is somewhat sensitive to λ_2. When the value of λ_2 is not large or small, the clustering performance is good. This demonstrates the necessity of the cluster structure in unsupervised clustering task. Especially, our method has different sensitivity level to λ_2 on the CMU PIE and COIL 100 data sets. Therefore, how to select the optimal values of different parameters is data set dependent and still an open problem, which will be studied in our future work. In our experiments, we first set $\lambda_1 = 10^5$ owing its insensitivity. Then, we make an attempt to find a candidate interval where the optimal λ_2 and λ_3 may exist. Finally, we find the optimal values of these three parameters in the 3D candidate space of (λ_1, λ_2, and λ_3) with a fixed search step length.

7 Conclusion

In this paper, we propose a robust multi-view graph recovery method. We first analyze the noises of graphs have specific structures (symmetric, column-wise and row-wise). Based on this observation, we introduce both column-sparse and row-sparse matrices to model the noises, such that robust consensus graph can be recovered by minimizing the disagreement over the cleaned graphs. To capture the data structure better, we enforce the connected components in obtained consensus graph equal to the cluster number by using the dynamic graph learning strategy. We develop an iterative algorithm to solve the hard optimization problem. Experimental results indicate that our method achieves the best unsupervised clustering and semi-supervised classifi-

cation results. We will explore many common structures of noises in multiple input graphs in the future.

APPENDIX A
SOLVING PROBLEM (14)

Problem (14) can be simply written as

$$\min_{r^T\mathbf{1}=1, r\geq 0} \|r-c\|_2^2 \tag{27}$$

where $r = w_i$ and $c = (\frac{1}{1+\lambda_1}b_i - \frac{v_i}{2(1+\lambda_1)})$. The Lagrangian function of problem (27) is as follows

$$\mathcal{L} = \frac{1}{2}\|r-c\|_2^2 - \alpha(r^T\mathbf{1}-1) - \beta^T r \tag{28}$$

where α and β are Lagrangian coefficients. According to the KKT condition, we have the following equations:

$$\begin{cases} \forall i, \;\; r_i - c_i - \alpha - \beta_i = 0 \\ \forall i, \;\; r_i \geq 0 \\ \forall i, \;\; \beta_i \geq 0 \\ \forall i, \;\; r_i\beta_i = 0 \end{cases} \tag{29}$$

Thus, we have $r-c-\alpha\mathbf{1}-\beta = 0$. According to the constraint $r^T\mathbf{1}=1$, we have $\alpha = \frac{1-\mathbf{1}^T c-\mathbf{1}^T\beta}{n}$. So, $r = (c - \frac{\mathbf{1}\mathbf{1}^T}{n}c + \frac{1}{n}\mathbf{1} - \frac{\mathbf{1}^T\beta}{n}\mathbf{1})$.

Denote $\bar{\beta} = \frac{\mathbf{1}^T\beta}{n}$ and $u = c - \frac{\mathbf{1}\mathbf{1}^T}{n}c + \frac{1}{n}\mathbf{1}$, then we can write $r = u + \beta - \bar{\beta}\mathbf{1}$. So, we have the following equation

$$r_i = u_i + \beta_i - \bar{\beta} \tag{30}$$

From (29) and (30), we know that if $\beta = 0$, the $r_i = u_i + \beta_i - \bar{\beta} = (u_i - \bar{\beta})_+$, where $(h)_+ = \max(0, h)$. We also know that if $\beta > 0$, $r_i = u_i + \beta_i - \bar{\beta} = 0$ and $u_i - \bar{\beta} < 0$, then $r_i = (u_i - \bar{\beta})_+ = 0$. Finally, we have

$$r_i = (u_i - \bar{\beta})_+ \tag{31}$$

Thus, we can obtain the optimal solution r_i if we know $\bar{\beta}$. We write (30) as $\beta_i = r_i + \bar{\beta} - u_i$, and we know, according to (29), $\beta_i = (\bar{\beta} - u_i)_+$. Since c is a n dimensional vector, then we have $\bar{\beta} = \frac{1}{n}\sum_{i=1}^n(\bar{\beta} - u_i)_+$. By defining $f(\bar{\beta}) = \frac{1}{n}\sum_{i=1}^n(\bar{\beta} - u_i)_+ - \bar{\beta}$, we know $f(\bar{\beta}) = 0$ is a piece-wise linear function. It is easy to find the roots by examining the $n+1$ lines to check for intersections with $r = 0$.

APPENDIX B
SOLVING PROBLEM (16)

The Lagrangian function of problem (16) is as follows

$$\mathcal{L} = \min_{\alpha} \sum_{v=1}^M \alpha_{(v)} q_{(v)} + \epsilon(\sum_{v=1}^M \alpha_{(v)}^\rho - 1) \tag{32}$$

Taking the derivation of problem (32) w.r.t $\alpha_{(v)}$ and setting the derivation to zero, we have

$$\sum_{v=1}^M (q_{(v)} + \rho\epsilon\alpha_v^{\rho-1}) = 0 \tag{33}$$

Further, we obtain

$$\sum_{v=1}^M \epsilon^{\frac{\rho}{\rho-1}}\alpha_{(v)}^\rho = -\sum_{v=1}^M \frac{1}{\rho}\alpha_{(v)}^{\frac{\rho}{\rho-1}} \tag{34}$$

Considering $\sum_{v=1}^M \alpha_{(v)}^\rho = 1$, we have $\epsilon = -\frac{1}{\rho}(\sum_{v=1}^M \alpha_{(v)}^{\frac{\rho}{\rho-1}})^{\frac{\rho}{\rho-1}}$ and then we obtain

$$\alpha_{(v)} = \frac{q_{(v)}^{\frac{1}{\rho-1}}}{(\sum_{(v)} q_{(v)}^{\frac{\rho}{\rho-1}})^{\frac{1}{\rho}}} \tag{35}$$

by substituting solution of ϵ into (33).

REFERENCES

[1] Z. Y. Zhang, Z. Zhai, and L. M. Li. Uniform Projection for Multi-view Learning, *IEEE Trans. Patten Anal. Mach. Intell.*, DOI 10.1109/TPAMI.2016.2601608, 2016.

[2] F. P. Nie, X. Q. Wang, and H. Huang. Clustering and projected clustering with adaptive neighbors, *The 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, USA, 2014.

[3] F. P. Nie, D. Xu, I. W. Tsang, and C. 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.

[4] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, *J. Mach. Learn. Res.*, vol. 7, pp. 2399-2434, Nov. 2006.

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

[6] P. Zhou, L. Du, L. Shi, H. Wang, Y. D. Shen. Recovery of corrupted multiple kernels for clustering, *IJCAI*, 2015.

[7] X. Cai, F. P. Nie, H. Huang, and F. Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering, *2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1977-1984.

[8] A. Kumar, P. Rai, and Hal Daume. Co-regularized multi-view spectral clustering, *In Advances in Neural Information Processing Systems*, 2011, pp. 1413-1421.

[9] J. Yu, M. Wang, and D. C. Tao. Semisupervised multiview distance metric learning for cartoon synthesis, *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4636-4648, 2012.

[10] Y. Y. Lin, T. L. Liu, C. S. Huh. Multiple kernel learning for dimenisonality reduction, *IEEE Trans. Patten Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147-1160, 2011.

[11] M. Gonen and E. Alpaydin. Multiple kernel learning algorithms, *Journal of Machine Learning Research*, vol. 12, pp. 2211-2268, 2011.

[12] M. Gonen and A. A. Margolin. Localized Data Fusion for Kernel K-Means Clustering with Application to Cancer Biology *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[13] V. Sindhwani and D.S. Rosenberg. An RKHS for multi-view learning and manifold co-regularization, *Proceedings of the 25th international conference on machine learning*, pp. 976-983. ACM, 2008.

[14] S. Yu, L.C. Tranchevent, X. Liu, W. Glanzel, J. A. K. Suykens, B. De Moor, and Y. Moreau. Optimized data fusion for kernel k-means clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp.1031-1039, 2012

[15] W. Tang, Z. Lu, and I. S. Dhillon. Clustering with multiple graphs, *In Proceedings of the 9th IEEE International Conference on Data Mining*, 2009

[16] M. Karasuyama, H. Mamitsuka. Multiple graph label propagation by sparse integration. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 12, pp. 1999-2012, 2013.

[17] Y. Q. Li, F. P. Nie, H. Huang, and J. Z. Huang. Large-scale multi-view spectral clustering via bipartite graph. *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[18] M. Wang, X. S. Hua, R. C. Hong, J. H. Tang, G. J. Qi, and Y. Song. Unified video annotation via multigraph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, pp. 733-746, 2009.

[19] C. Xu, D. C. Tao, and C. Xu. Multi-view Intact Space Learning. *IEEE Trans. Patten Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2531-2544, 2015.

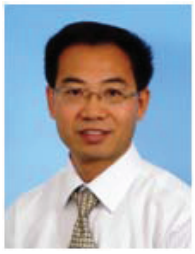
[20] M. N. Kan, S. G. Shan, H. H. Zhang, S. H. Lao, and X. L. Chen. Multi-view Discriminant Analysis. *IEEE Trans. Patten Anal. Mach. Intell.*, DOI 10.1109/TPAMI.2015.2435740, 2015.

[21] J. B. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.
[22] L. Hagen and A. B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-aided design of integrated circuits and systems*, vol. 11, no. 9, pp. 1074-1085, 1992.
[23] D. N. Niu, J. G Dy, and M. I Jordan. Multiple non-redundant spectral clustering views. *In Proceedings of the 27th international conference on machine learning*, 2010, pp. 831-838. 2010.
[24] F. P. Nie, J. Li, and X. L. Li. Parameter-Free Auto-Weighted Multiple Graph Learning: A Framework for Multiview Clustering and Semi-supervised Classification. *IJCAI*, 2015.
[25] K. Chaudhuri, S. M Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. *In Proceedings of the 26th annual international conference on machine learning*, pp. 129-136. ACM, 2009.
[26] X. Cai, F. P. Nie, and H. Huang. Heterogeneous image feature integration via multi-modal semi-supervised learning model. *2013 IEEE International Conference on Computer Vision*, pp. 1737-1744. 2013.
[27] T. Xia, D. Tao, T. Mei, and Y. Zhang. Multiview spectral embedding. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 6, 2010, pp. 1438-1446.
[28] Y. Li, F. Nie, H. Huang, and J. Huang. Large-scale multi-view spectral clustering via bipartite graph. *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
[29] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Patten Anal. Mach. Intell.*, vol. 35, no. 1, 2013, pp. 171-174.
[30] B. Mohar. The laplacian spectrum of graphs. *In Graph Theory, Combinatorics, and Applications*, pp. 871-898. Wiley, 1991.
[31] F. R. K. Chung. Spectral Graph Theory. *CBMS Regional Conference Series in Mathematics*, No. 92, American Mathematical Society, February, 1997.
[32] K. Fan. On a theorem of weyl concerning eigenvalues of linear transformations. *Proceedings of the National Academy of Sciences*, vol. 35, no. 11, pp. 652-655, 1949.
[33] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *NIPS*, 2010.
[34] Y. J. Lee and K. Grauman. Foreground focus: Unsupervised learning from partially matching images. *International Journal of Computer Vision*, vol. 85, no. 2, 2009, pp. 143-166.
[35] A. Asuncion and D. Newman. *UCI machine learning repository*, 2007.
[36] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, vil. 2, pp. 849-856, 2002.

Xiaozhao Fang received his M.S.degree in 2008, and the Ph.D.degree in computer science and technology at Shenzhen Graduate School, HIT, Shenzhen (China) in 2016. He is currently with the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include pattern recognition, data mining and machine learning.

Na Han received her B.S.degree in computer science and technology at HIT in 2004. She is currently pursuing the Ph.D.degree in computer science and technology with the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include pattern recognition and machine learning.

Shaohua Teng is a Professor of Guangdong University of Technology in China. He was born on January in 1962. He is responsible for teaching data mining in School of Computer Science and Technology. He is engaged in education and technology transfer on knowledge discovery issues, and research in big data, network security, cooperative work, machine learning and statistical pattern recognition. Dr. Teng earned a Ph.D. in Industry Engineering at Guangdong University of Technology. He has applied for 6 patents on his invention. He has published 300 papers on computer magazines and international conferences and 2 books. He earned the Provincial Science and Technology Award, and Guangdong outstanding teacher.



Jigang Wu received B.Sc. degree from Lanzhou University, China in 1983, and doctoral degree from the University of Science and Technology of China in 2000. He was with the Center for High Performance Embedded Systems, Nanyang Technological University, Singapore, from 2000to 2010, as a research fellow. He was Dean, Tianjin distinguished professor of School of Computer Science and Sofware, Tianjin Polytechnic University, China, from 2010 to 2015. Now he is distinguished professor of School of Computer Science and Technology, Guangdong University of Technology. He has published more than 200 papers in IEEE TOC, TPDS, TVLSI, TNNLS, TSMC JPDC, PARCO, JSA, and international conferences. His research interests include network computing, cloud computing, machine intelligence, reconfigurable architecture. He is a member of the IEEE. He serves in China Computer Federation as technical committee member in the branch committees, High Performance Computing, Theoretical Computer Science, and Fault Tolerant Computing.

Yong Xu was born in Sichuan,China,in1972.He received his B.S.degree, M.S. degree in 1994 and 1997, respectively. He received the Ph.D.degree in Pattern Recognition and Intelligence system at NUST(China) in 2005. Now he works at Shenzhen Graduate School, Harbin Institute of Technology. His current 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, Xian, China.