Spectral Clustering by Joint Spectral Embedding and Spectral Rotation

Yanwei Pang, *Senior Member, IEEE*, Jin Xie, Feiping Nie, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Spectral clustering is an important clustering method widely used for pattern recognition and image segmentation. Classical spectral clustering algorithms consist of two separate stages: 1) solving a relaxed continuous optimization problem to obtain a real matrix followed by 2) applying *K*-means or spectral rotation to round the real matrix (i.e., continuous clustering result) into a binary matrix called the cluster indicator matrix. Such a separate scheme is not guaranteed to achieve jointly optimal result because of the loss of useful information. To obtain a better clustering result, in this paper, we propose a joint model to simultaneously compute the optimal real matrix and binary matrix. The existing joint model adopts an orthonormal real matrix to approximate the orthogonal but nonorthonormal cluster indicator matrix. It is noted that only in a very special case (i.e., all clusters have the same number of samples), the cluster indicator matrix is an orthonormal matrix multiplied by a real number. The error of approximating a nonorthonormal matrix is inevitably large. To overcome the drawback, we propose replacing the nonorthonormal cluster indicator matrix with a scaled cluster indicator matrix which is an orthonormal matrix. Our method is capable of obtaining better performance because it is easy to minimize the difference between two orthonormal matrices. Experimental results on benchmark datasets demonstrate the effectiveness of the proposed method (called JSESR).**

***Index Terms*—Normalized cut (Ncut), spectral clustering, spectral rotation.**

I. INTRODUCTION

CLUSTERING plays an important role in machine learning, data mining, image segmentation, and pattern classification [1], [2]. The goal of clustering is to classify elements into clusters on the basis of their similarity [3].

A large number of clustering methods have been brought forward. Classical methods include hierarchical clustering [4]; *K*-means clustering [5]; spectral clustering [3], [6]; support vector clustering [7]; multiview clustering [8]; genetic clustering [9]; etc., Among these clustering methods, spectral clustering has become one of the popular methods because of its robustness and effectiveness. Generally, the performance of the spectral clustering is better than other methods. Spectral clustering is able to seek the optimal partitioning of data based on the spectral graph theory. Traditional clustering algorithms such as *K*-means can only perform clustering with convex distribution. If the sample spaces are nonconvex, *K*-means would fall into a local optimal solution. Compared with *K*-means, spectral clustering can perform clustering with nonconvex sphere of sample spaces and obtain the globally optimal solution in a relaxed continuous domain.

Although many spectral clustering methods have been proposed, such as Min Cut [10], Ratio Cut (Rcut) [11], Normalized Cut (Ncut) [12], Min–Max Cut [13], Spectral Embedded Clustering [14], *K*-way Rcut [15], and *K*-way Ncut [16], all of these methods adopt a two-stage process. The first stage is to learn the relaxed continuous spectral vectors and the second stage is usually to employ *K*-means or spectral rotation to post-process the continuous spectral vectors in order to obtain the final binary cluster indicator matrix. In practice, the manner of separately performing the two stages is not able to jointly obtain the optimal solution.

In this paper, in order to overcome the aforementioned drawback of spectral clustering, we propose a new spectral clustering framework (called JSESR) that jointly performs spectral embedding and spectral rotation. That is, the real-valued cluster indicator matrix usually obtained by conducting spectral embedding in the intermediate stage and the binary cluster indicator matrix usually obtained by conducting spectral rotation in the last stage are iteratively computed in our method.

Recently, Yang *et al.* [17] proposed a unified framework for discrete spectral clustering (UFDSC). The UFDSC is able to obtain the final clustering results by one step and results in significant improvement of clustering performance. Nevertheless, the objective function of UFDSC has a term which employs an orthonormal matrix (in this paper, orthonormal matrix denotes the matrix whose columns or rows are orthonormal vectors,[1] that is, $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ or $\mathbf{F}\mathbf{F}^T = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. The orthogonal matrix denotes the matrix whose columns or rows are orthogonal vectors but not necessarily orthogonal unit vectors) to approximate a nonorthonormal matrix. The

Manuscript received May 15, 2018; revised August 22, 2018; accepted August 27, 2018. Date of publication October 3, 2018; date of current version October 22, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61632081, in part by the National Program on Key Basic Research Project (973 Program) under Grant 2014CB340400, and in part by the National Natural Science Foundation of China under Grant 61871470 and Grant 61761130079. This paper was recommended by Associate Editor R. Tagliaferri. *(Corresponding author: Feiping Nie.)*

Y. Pang and J. Xie are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: pyw@tju.edu.cn; jinxie@tju.edu.cn).

F. Nie is with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi’an 710071, China (e-mail: feipingnie@gmail.com).

X. Li is with the Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an 710119, China (e-mail: xuelong_li@opt.ac.cn).

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

Digital Object Identifier 10.1109/TCYB.2018.2868742

---

[1] orthogonal unit vectors.

approximation cannot be very precise in theory. As will be shown in the toy example in Fig. 1, this method tends to generate incorrect clustering results for unbalanced data where the underlying numbers of clusters are far from uniform. By contrast, the proposed JSESR is capable of overcoming the drawback because approximation is conducted in-between two orthonormal matrices.

In summary, the novelty, contribution, and characteristic of the proposed JSESR are as follows.

1) A joint model is proposed to simultaneously and iteratively perform spectral embedding and spectral rotation with spectral embedding generating a real-valued cluster indicator matrix and spectral rotation generating a binary cluster indicator matrix. Compared to the classical spectral clustering methods, the proposed joint model is able to overcome the drawbacks of the information loss and the risk of the discrete clustering deviation.
2) In the spectral rotation part of the proposed joint model, approximation is conducted in-between two orthonormal matrices: a) a matrix generated by spectral embedding followed by a rotation operation and b) a scaled cluster indicator matrix. Therefore, the proposed method is able to obtain an accurate clustering result. In addition, the proposed method is able to overcome the problem of the unbalance of UFDSC.
3) The physical meaning of the scaled cluster indicator matrix is interpreted. Moreover, the theoretical derivation of the scaled cluster indicator matrix is given. The insight in the scaled cluster indicator matrix is helpful to understand the proposed method and developing a new method.
4) The proposed method cannot only achieve an accurate clustering result but also be implemented very efficiently. The optimization process of the proposed convergences is in about three iterations.

The rest of this paper is organized as follows. In Section II, the related work is discussed. Classical spectral embedding and spectral rotation are described in Section III. The proposed method is presented in Section IV. The experimental results are presented in Section V. Finally, Section VI concludes this paper.

II. RELATED WORKS

There are many clustering methods [18]. According to different criterion, these clustering methods can be divided into different categories. The categories may be overlapping and a clustering method can belong to two or more categories. Generally, the clustering methods can be divided into hierarchical methods [4]; partition methods [5]; density methods [19], [20]; kernel methods [21]; and spectral methods [6].

The hierarchical methods produce a hierarchy of nested clusterings. There are two main categories of hierarchical methods: 1) the agglomerative methods and 2) the divisive hierarchical methods.

Given the number of partitions, the partitioning method creates an initial partitioning. Then the method adopts an iterative relocation technique that attempts to improve the partitioning by moving samples from one group to another group. *K*-means and its variants [5], [22] are the typical instances of partitioning methods. The performance of *K*-means is unsatisfactory when the problem of the curse-of-dimensionality is severe. In addition, most of this kind of methods is sensitive to initialization.

Density methods assume that the samples of each cluster are drawn from the probability distribution. The main tasks of the density method include determining and analyzing local density of data points, the distribution parameters, and identifying the clusters. DBSCAN [19] and OPTICS [20] are two representative density methods. DBSCAN is sensitive to parameters, such as the radius and the number of points within the radius. In OPTICS, the clusters are identified as local density maxima that are far away from any points of higher density. Because the OPTICS depends on the relative densities rather than their absolute values, OPTICS is more robust than DBSCAN.

Kernel methods implicitly map the data into low dimensional space where clustering is conducted. Many clustering methods can be extended to their kernel versions by the kernel technique [21], [23]. A representative method is kernel *K*-means. As kernel methods in the fields of subspace analysis and classifier learning, the kernel methods for clustering also encounter the problem of determining the type of kernels and the parameters of the selected kernel function.

Spectral clustering methods are closely related to this paper and were successfully applied in segmentation [12], [16]; semisupervised learning [24]; multitask learning [25]; scene detection [26]; and so on [27]–[29]. Representative spectral clustering methods include Min Cut [10], Rcut [11], Ncut [12], and Min–Max Cut [13]. Generally speaking, spectral clustering methods employ spectral graph theory [30] and formulate the clustering task as an eigen-decomposition problem [31], [32]. The core of the spectral clustering methods is optimally partitioning a graph with a criterion under some constraints. Different spectral clustering methods adopt different objective functions and/or different constraints. The goal of Min Cut is to partition a graph into *k*-subgraphs such that the maximum cut across the subgroups (the maximum cut problem is to find a subgroup of the vertex to make the number of edges between the subgroup and the complementary subgroup is as large as possible) is minimized [10]. However, the minimum cut criterion favors cutting small sets of isolated nodes in the graph that bisect the existing segments. Ncut [12] overcomes the drawback of Min Cut by computing the cut cost as a fraction of the total edge connections to all the nodes in the graph. Rcut allows freedom to find natural partitions: the numerator captures the Min Cut criterion and the denominator favors an even partition [11]. In Min–Max Cut, the similarity between two subgraphs is minimized and at the same time the similarity within each subgraph is maximized [13]. When clusters overlap heavily, Min–Max Cut tends to give more compact and balanced clusters.

Though many variants of Min Cut, Rcut, Ncut, and Min–Max Cut were developed [15], [16], these methods first conduct spectral embedding to form a real-valued cluster indicator matrix and binarize the real-valued cluster indicator matrix to form the final binary cluster indicator matrix. These methods cannot directly compute the final clustering result.

As mentioned in Section I, the UFDSC [17] can directly obtain the binary cluster indicator matrix. Despite its success, the accuracy of the method is limited because it employs an orthonormal matrix to approximate a nonorthonormal matrix. Our method is able to overcome the drawback and achieves more accurate clustering results.

III. CLASSICAL NORMALIZED CUT AND THE SPECTRAL ROTATION

The classical spectral clustering called k-way Ncut [16] and spectral rotation [33] are the basis of the proposed method. In this section, we describe the two methods. Spectral embedding (k-way Ncut) yields real-valued cluster indicator matrix. Taking the real-valued cluster indicator matrix as input, the spectral rotation results in discrete-valued indicator matrix.

A. Spectral Embedding

Let the dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be clustered into K distinct clusters. Suppose that X contains N samples $\mathbf{x}_i \in \mathbb{R}^M$, $i = 1, \ldots, N$. The N samples form a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N\times M}$. The K clusters $C_1, \ldots, C_K$ meet three conditions: $\bigcup_{i=1}^{K} C_i = X$; $C_i \neq \varnothing$, $i = 1, \ldots, K$; and $C_i \bigcap C_j = \varnothing$, $i \neq j$, $i, j = 1, \ldots, K$. Let the cluster indicator matrix $\mathbf{Y} = [\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \ldots, \bar{\mathbf{y}}_N]^T \in \mathbb{B}^{N\times K}$ with each vector $\bar{\mathbf{y}}_i \in \mathbb{B}^K$ be a cluster indicator of the corresponding sample $\mathbf{x}_i \in \mathbb{R}^M$. If $\mathbf{x}_i$ is considered in the k cluster C_k, then the kth element y_{ik} of $\bar{\mathbf{y}}_i$ is 1 and other elements are all 0.

Let a_{ij} be the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$. The set of a_{ij} defines the affinity matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$. A common choice of a_{ij} is

$$a_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{t}\right) & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where t is a real-valued parameter. The degree matrix $\mathbf{D}$ is derived from the affinity matrix $\mathbf{A}$. The off-diagonal elements of $\mathbf{D}$ are zero and the ith element d_i of the diagonal is

$$d_i = \sum\nolimits_{j=1}^{N} a_{ij}. \tag{2}$$

The value of d_i measures the significance of a sample $\mathbf{x}_i$. The degree matrix can be expressed as $\mathbf{D} = \mathrm{diag}\{d_1, d_2, \ldots, d_N\}$.

The goal of minimum k-way Ncut is to simultaneously minimize the sum $f(\mathbf{Y})$

$$f(\mathbf{Y}) = \sum_{x_i \in \mathbf{C}_k} \sum_{x_j \notin \mathbf{C}_k} a_{ij} \tag{3}$$

and maximize the sum $g(\mathbf{Y})$ of weighted volume $V(\mathbf{C}_k)$ of each cluster C_k

$$g(\mathbf{Y}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} D_{ii} \tag{4}$$

where $D_{ii} = \sum_{j=1}^{N} a_{ij}$ measures the significance of a sample $\mathbf{x}_i$.

The effect of minimizing the sum of similarity [i.e., (3)] is to let samples in different clusters have the least similarity. Defining the Laplacian matrix $\mathbf{L}$

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \tag{5}$$

Equation (3) can be written as

$$\begin{aligned} f(\mathbf{Y}) &= \sum_{x_i \in \mathbf{C}_k} \sum_{x_j \notin \mathbf{C}_k} a_{ij} = \sum_{k=1}^{K} {\mathbf{y}_k}^T (\mathbf{D} - \mathbf{A}) \mathbf{y}_k \\ &= \sum_{k=1}^{K} {\mathbf{y}_k}^T \mathbf{L} \mathbf{y}_k \end{aligned} \tag{6}$$

where $\mathbf{y}_k$ is the kth column of $\mathbf{Y}$.

The effect of maximizing the sum of weighted volume [i.e., (4)] is to let samples in the same clusters have the largest similarity. $g(\mathbf{Y})$ can be equivalently expressed in the matrix form

$$g(\mathbf{Y}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} D_{ii} = \sum_{k=1}^{K} {\mathbf{y}_k}^T \mathbf{D} \mathbf{y}_k. \tag{7}$$

Therefore, the problem of k-way Ncut can be formulated as minimizing $J(\mathbf{Y})$

$$J(\mathbf{Y}) = \frac{1}{K} \sum_{k=1}^{K} \frac{{\mathbf{y}_k}^T \mathbf{L} \mathbf{y}_k}{{\mathbf{y}_k}^T \mathbf{D} \mathbf{y}_k}. \tag{8}$$

Define scaled partition matrix $\mathbf{Z}$

$$\mathbf{Z} = \mathbf{Y}\left(\mathbf{Y}^T \mathbf{D} \mathbf{Y}\right)^{-1/2} \tag{9}$$

then (8) becomes

$$\begin{aligned} J(\mathbf{Y}) &= \frac{1}{K} \sum_{k=1}^{K} {\mathbf{y}_k}^T \mathbf{L} \mathbf{y}_k \left({\mathbf{y}_k}^T \mathbf{D} \mathbf{y}_k\right)^{-1} \\ &= \frac{1}{K} \mathrm{tr}\Big(\left(\mathbf{Y}^T \mathbf{L} \mathbf{Y}\right)\left(\mathbf{Y}^T \mathbf{D} \mathbf{Y}\right)^{-1}\Big) \\ &= \frac{1}{K} \mathrm{tr}\Big(\left(\mathbf{Y}^T \mathbf{D} \mathbf{Y}\right)^{-1/2} \mathbf{Y}^T \mathbf{L} \mathbf{Y} \left(\mathbf{Y}^T \mathbf{D} \mathbf{Y}\right)^{-1/2}\Big) \\ &= \frac{1}{K} \mathrm{tr}\Big(\mathbf{Y}\left(\mathbf{Y}^T \mathbf{D} \mathbf{Y}\right)^{-1/2}\Big)^T \mathbf{L} \mathbf{Y} \left(\mathbf{Y}^T \mathbf{D} \mathbf{Y}\right)^{-1/2}\Big) \\ &= \frac{1}{K} \mathrm{tr}\left(\mathbf{Z}^T \mathbf{L} \mathbf{Z}\right) \\ &= J(\mathbf{Z}). \end{aligned} \tag{10}$$

Defining $\bar{\mathbf{F}} = \mathbf{D}^{1/2} \mathbf{Y} \left(\mathbf{Y}^T \mathbf{D} \mathbf{Y}\right)^{-1/2} = \mathbf{D}^{1/2} \mathbf{Z}$, we have $\mathbf{Z} = \mathbf{D}^{-1/2} \bar{\mathbf{F}}$. Therefore, it holds that

$$\begin{aligned} J(\mathbf{Y}) = J(\mathbf{Z}) &= \frac{1}{K} \mathrm{tr}\left(\mathbf{Z}^T \mathbf{L} \mathbf{Z}\right) \\ &= \frac{1}{K} \mathrm{tr}\Big(\left(\mathbf{D}^{-1/2} \bar{\mathbf{F}}\right)^T \mathbf{L} \left(\mathbf{D}^{-1/2} \bar{\mathbf{F}}\right)\Big) \\ &= \frac{1}{K} \mathrm{tr}\left(\bar{\mathbf{F}}^T \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} \bar{\mathbf{F}}\right) \\ &= \frac{1}{K} \mathrm{tr}\left(\bar{\mathbf{F}}^T \tilde{\mathbf{L}} \bar{\mathbf{F}}\right) \\ &= J\left(\bar{\mathbf{F}}\right) \end{aligned} \tag{11}$$

where $\tilde{\mathbf{L}} = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$ is known as the normalized Laplacian matrix and $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix.

Note that, $J(\bar{\mathbf{F}}) = (1/K)\mathrm{tr}(\bar{\mathbf{F}}^T \tilde{\mathbf{L}} \bar{F})$ [i.e., (11)] is hard to solve, because the elements of $\bar{\mathbf{F}}$ are constrained to be discrete values. The solution of this problem is to relax the matrix $\bar{\mathbf{F}}$ from discrete values to continuous ones. Then the problem (11) becomes

$$J(\mathbf{F}) = \frac{1}{K} \mathrm{tr}\left(\mathbf{F}^T \tilde{\mathbf{L}} F\right) \tag{12}$$

where $\mathbf{F} \in \mathbb{R}^{N \times K}$.

Accordingly, the problem of spectral clustering can be formulated as

$$\mathbf{F}^* = \underset{\mathbf{F}^T\mathbf{F}=\mathbf{I}}{\arg\min}\, \mathrm{tr}\Big(\mathbf{F}^T\tilde{\mathbf{L}}F\Big). \tag{13}$$

The optimal $\mathbf{F}^*$ of $\mathbf{F}$ consists of the eigenvectors of $\tilde{\mathbf{L}}$. The process of computing $\mathbf{F}^*$ is called spectral embedding because it makes use of the spectrum of the affinity (similarity) matrix of the data to perform dimensionality reduction before clustering.

B. Classical Spectral Rotation

The optimal $\mathbf{F}^*$ which is obtained by solving the optimization problem (13) is not a zero-one valued matrix. Therefore, to get the final clustering result, it is common to apply K-means or spectral rotation [33] to transform $\mathbf{F}^*$ to a zero-one valued matrix so that it approaches the underlying cluster indicator matrix. It is known that the underlying cluster indicator matrix $\mathbf{Y}$ is binary and its element is either 0 or 1.

Spectral rotation [33] is an algorithm for optimally transforming the real-valued cluster indicator matrix $\mathbf{F}^*$ to a binary matrix $\mathbf{Y}$.

Prior to describing the spectral rotation method, it should be noted that the optimal $\mathbf{F}^*$ is not unique in the sense of minimizing the trace $\mathrm{tr}(\mathbf{F}^T\tilde{\mathbf{L}}F)$ expressed in (13).

Theorem 1: If $\mathbf{F}^*$ is an optimal solution for minimizing $\mathrm{tr}(\mathbf{F}^T\tilde{\mathbf{L}}F)$ [i.e., (13)] and $\mathbf{R} \in \mathbb{R}^{K\times K}$ is a rotation matrix (nonsingular matrix) satisfying $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, then $\mathbf{F}^*\mathbf{R}$ is also an optimal solution for minimizing $\mathrm{tr}(\mathbf{F}^T\tilde{\mathbf{L}}F)$

$$\mathrm{tr}\Big[\big(\mathbf{F}^*\mathbf{R}\big)^T\tilde{\mathbf{L}}\big(\mathbf{F}^*\mathbf{R}\big)\Big] = \mathrm{tr}\Big(\mathbf{F}^{*T}\tilde{\mathbf{L}}\mathbf{F}^*\Big). \tag{14}$$

Proof: Because $(\mathbf{F}^*\mathbf{R})^T\tilde{\mathbf{L}}(\mathbf{F}^*\mathbf{R}) = \mathbf{R}^T(\mathbf{F}^{*T}\tilde{\mathbf{L}}\mathbf{F}^*)\mathbf{R} = \mathbf{R}^{-1}(\mathbf{F}^{*T}\mathbf{L}\mathbf{F}^*)\mathbf{R}$, it holds that $\mathbf{R}^{-1}(\mathbf{F}^{*T}\tilde{\mathbf{L}}\mathbf{F}^*)\mathbf{R}$ and $\mathbf{F}^{*T}\mathbf{L}\mathbf{F}^*$ are similar. It is known that similar matrices have the same trace. Therefore, $\mathrm{tr}[\mathbf{R}^{-1}(\mathbf{F}^{*T}\mathbf{L}\mathbf{F}^*)\mathbf{R}] = \mathrm{tr}[\mathbf{F}^{*T}\mathbf{L}\mathbf{F}^*]$, meaning that both $\mathbf{F}^*\mathbf{R}$ and $\mathbf{F}^*$ are optimal estimators. ■

Taking into account that both $\mathbf{F}^*\mathbf{R}$ and $\mathbf{F}^*$ are optimal estimators, the goal of spectral rotation is to minimize the distance between $\mathbf{F}^*\mathbf{R}$ and a binary matrix $\mathbf{Y} \in \mathrm{Ind}$

$$\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I},\mathbf{Y}\in\mathrm{Ind}} \big\|\mathbf{F}^*\mathbf{R} - \mathbf{Y}\big\|_F^2. \tag{15}$$

$\mathbf{Y} \in \mathrm{Ind}$ denotes $\mathbf{Y}$ is an indicator matrix, $\mathbf{Y} = [\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \ldots, \bar{\mathbf{y}}_N]^T \in \mathbb{B}^{N\times K}$ and the unique 1 in $\bar{\mathbf{y}}_i$ indicate the cluster membership of the corresponding sample $\mathbf{x}_i$.

IV. PROPOSED METHOD

As stated in Section III, spectral embedding followed by spectral rotation yields the clustering result. Spectral embedding yields real-valued cluster indicator matrix which is then used to obtain cluster indicator matrix by spectral rotation. However, successively and independently performing spectral embedding and spectral rotation are not guaranteed to yield globally optimal solution. To obtain better result, in this paper, we propose a new framework to simultaneously perform spectral embedding and a variant of spectral rotation.

A. Objective Function

A straightforward method for simultaneously obtaining real-valued cluster indicator matrix and discrete-valued cluster indicator matrix is combining (13) and (15)

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I},\mathbf{R}^T\mathbf{R}=\mathbf{I},\mathbf{Y}\in\mathrm{Ind}} \Big[\mathrm{tr}\Big(\mathbf{F}^T\tilde{\mathbf{L}}F\Big) + \alpha\|\mathbf{F}\mathbf{R} - \mathbf{Y}\|_F^2\Big]_F^2 \tag{16}$$

where α is a weight parameter. In fact, (16) is the mathematical formulation of the UFDSC proposed in [17]. However, it is challenging for the spectral rotation part $||\mathbf{F}\mathbf{R} - \mathbf{Y}||_F^2$ to approximate the zero-one discrete matrix $\mathbf{Y}$ with the real matrix $\mathbf{F}\mathbf{R}$. The reason is as follows. It can be proved that $\mathbf{F}\mathbf{R}$ is an orthonormal matrix. It means that not only the columns of $\mathbf{F}\mathbf{R}$ are orthogonal but also the magnitude of each column is one. However, $\mathbf{Y}$ is very different from $\mathbf{F}\mathbf{R}$. The main difference is that $\mathbf{Y}$ is not necessarily an orthonormal matrix though it is an orthogonal matrix. It is difficult to approximate an orthonormal matrix to a nonorthonormal matrix. Note that, the cluster indicator matrix $\mathbf{Y}$ can be an orthogonal matrix with a constant scale only when all the clusters contain the same number of samples. For example, the following $\mathbf{Y}$ given in (17) consists of three clusters with the first, second, and third cluster containing 3, 2, and 1 samples

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{17}$$

In this case, $\mathbf{Y}$ is not a scaled orthonormal matrix. It is impossible to perfectly approximate such $\mathbf{Y}$ with any orthonormal matrix. The matrix $\mathbf{Y}$ given in (18) is an orthonormal matrix divided by scaler

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \sqrt{2}\begin{bmatrix} \sqrt{1/2} & 0 & 0 \\ \sqrt{1/2} & 0 & 0 \\ 0 & \sqrt{1/2} & 0 \\ 0 & \sqrt{1/2} & 0 \\ 0 & 0 & \sqrt{1/2} \\ 0 & 0 & \sqrt{1/2} \end{bmatrix} \tag{18}$$

where each cluster contains two samples. Only in this special case, it is possible to approximate the cluster indicator matrix with an orthonormal matrix.

To overcome the above-mentioned difficulty, we propose to replace $||\mathbf{F}\mathbf{R} - \mathbf{Y}||_F^2$ with $||\mathbf{F}\mathbf{R} - \mathbf{D}^{1/2}\mathbf{Y}(\mathbf{Y}^\mathbf{T}\mathbf{D}\mathbf{Y})^{-1/2}||_F^2$. The term $\mathbf{D}^{1/2}\mathbf{Y}(\mathbf{Y}^\mathbf{T}\mathbf{D}\mathbf{Y})^{-1/2}$ is called the scaled cluster indicator matrix $\mathbf{Y}_s$

$$\mathbf{Y}_s = \bar{\mathbf{F}} = \mathbf{D}^{1/2}\mathbf{Y}\Big(\mathbf{Y}^\mathbf{T}\mathbf{D}\mathbf{Y}\Big)^{-1/2}. \tag{19}$$

Theorem 2: The scaled cluster indicator matrix $\mathbf{Y}_s = \mathbf{D}^{1/2}\mathbf{Y}(\mathbf{Y}^\mathbf{T}\mathbf{D}\mathbf{Y})^{-1/2}$ is an orthonormal matrix.

Proof: The product of ${\mathbf{Y}_s}^T$ and $\mathbf{Y}_s$ is

$$\begin{aligned} {\mathbf{Y}_s}^T\mathbf{Y}_s &= \Big(\mathbf{D}^{1/2}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-1/2}\Big)^T\Big(\mathbf{D}^{1/2}\mathbf{Y}\Big(\mathbf{Y}^\mathbf{T}\mathbf{D}\mathbf{Y}\Big)^{-1/2}\Big) \\ &= \Big(\Big(\mathbf{Y}^\mathbf{T}\mathbf{D}\mathbf{Y}\Big)^{-1/2}\Big)^T\mathbf{Y}^T\Big(\mathbf{D}^{1/2}\Big)^T\Big(\mathbf{D}^{1/2}\mathbf{Y}\Big(\mathbf{Y}^\mathbf{T}\mathbf{D}\mathbf{Y}\Big)^{-1/2}\Big) \end{aligned}$$

$$
\begin{aligned}
&= \left(\left(\left(\mathbf{Y^TDY}\right)^{-1/2}\right)^T \mathbf{Y}^T\left(\left(\mathbf{D}^{1/2}\right)^T\mathbf{D}^{1/2}\right)\mathbf{Y}\left(\mathbf{Y^TDY}\right)^{-1/2}\right)\\
&= \left(\left(\left(\mathbf{Y^TDY}\right)^{-1/2}\right)^T \left(\mathbf{Y}^T\mathbf{DY}\right)\left(\mathbf{Y^TDY}\right)^{-1/2}\right)\\
&= \mathbf{I}. \qquad (20)
\end{aligned}
$$

■

Because both $\mathbf{FR}$ and $\mathbf{Y}_s$ (i.e., $\mathbf{D^{1/2}Y(Y^TDY)^{-1/2}}$) are orthonormal matrices, it is reasonable to approximate $\mathbf{Y}_s$ with $\mathbf{FR}$. It is relatively easy to minimize the difference between $\mathbf{FR}$ and $\mathbf{D^{1/2}Y(Y^TDY)^{-1/2}}$. Formally, in our method the spectral clustering is formulated as the following optimization problem:

$$
\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I},\mathbf{R}^T\mathbf{R}=\mathbf{I},\mathbf{Y}\in\text{Ind}} \left[\begin{array}{c} \text{tr}\left(\mathbf{F}^T\tilde{\mathbf{L}}\mathbf{F}\right) \\ +\alpha\left\|\mathbf{FR}-\mathbf{D^{1/2}Y}\left(\mathbf{Y^TDY}\right)^{-1/2}\right\|_F^2 \end{array}\right]. \qquad (21)
$$

The Physical Meaning of the Scaled Cluster Indicator Matrix: Now, we describe the physical meaning of the scaled cluster indicator matrix $\mathbf{Y}_s = \mathbf{D^{1/2}Y(Y^TDY)^{-1/2}}$. Equation (2) shows how the degree d_i of the degree matrix is computed. Denote the degree vector $\mathbf{d} \in \mathbb{R}^N$ by $\mathbf{d} = [d_1, d_2, \ldots, d_N]^T$. Let $\mathbf{y}_k$ be the kth column of $\mathbf{Y}$ and y_{ik} is the ik-entry of $\mathbf{Y}$. Then, it can be verified that the ik-entry y_{ik}^s of $\mathbf{Y}_s$ is

$$
y_{ik}^s = y_{ik}\sqrt{\frac{d_i}{\sum_{j=1}^N d_j y_{jk}}} = y_{ik}\sqrt{\frac{d_i}{\mathbf{d}^T\mathbf{y}_k}}. \qquad (22)
$$

Equation (22) means that the ik-entry of $\mathbf{Y}_s$ is the degree d_i normalized by the sum of weighted degrees. Intuitively, the larger the y_{ik}^s is, the larger probability for the sample i to belong to cluster k. For example, suppose that the cluster indicator matrix $\mathbf{Y}$ and the degree matrix $\mathbf{D}$ are, respectively,

$$
\mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad (23)
$$

$$
\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 11 \end{bmatrix} \qquad (24)
$$

then the scaled indicator matrix is

$$
\mathbf{Y}_s = \begin{bmatrix} \sqrt{\frac{1}{1+3+5}} & 0 & 0 \\ \sqrt{\frac{3}{1+3+5}} & 0 & 0 \\ \sqrt{\frac{5}{1+3+5}} & 0 & 0 \\ 0 & \sqrt{\frac{7}{7+9}} & 0 \\ 0 & \sqrt{\frac{9}{7+9}} & 0 \\ 0 & 0 & \sqrt{\frac{11}{11}} \end{bmatrix}. \qquad (25)
$$

From (25), one can find that the scaled cluster indicator matrix is a real-valued and orthonormal matrix. It is easy to minimize the difference between $\mathbf{FR}$ and the scaled cluster indicator matrix.

The Relationship Between $\mathbf{F}$ *and the Scaled Cluster Indicator Matrix:* $\mathbf{D^{1/2}Y(Y^TDY)^{-1/2}}$. Suppose that $\mathbf{F}^*$ is composed of the eigenvectors of the normalized Laplacian matrix $\tilde{\mathbf{L}}$ (i.e., $\mathbf{F}^*$ is the solution to minimize tr($\mathbf{F}^T\tilde{\mathbf{L}}F$). In theory, $\mathbf{F}^*$ is closely related to $\mathbf{D^{1/2}Y(Y^TDY)^{-1/2}}$. The relationship can be discovered by analyzing the objective function [i.e., (11) and (12)] of the spectral embedding. It is noted that the matrix $\mathbf{F}$ is the relaxed solution to the scaled cluster indicator matrix $\mathbf{Y_s}$.

B. Optimization

The proposed optimization formulation [i.e., (21)] contains not only real variables (i.e., $\mathbf{F}$ and $\mathbf{R}$) but also a zero-one variable (i.e., $\mathbf{Y}$). It is challenging to find the globally optimal solution to the complex problem. We propose an alternative algorithm to solve the optimization problem. Specifically, the proposed algorithm iteratively performs three steps: $\mathbf{R}$-step, $\mathbf{Y}$-step, and $\mathbf{F}$-step.

***R**-Step:* The goal of $\mathbf{R}$-step is to seek the optimal rotation matrix $\mathbf{R}$ when $\mathbf{F}$ and $\mathbf{Y}$ are fixed. Omitting terms irrelevant to solve $\mathbf{R}$, the problem expressed in (21) is reduced to the following optimization problem:

$$
\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I}}\left\|\mathbf{FR}-\mathbf{D^{1/2}Y}\left(\mathbf{Y^TDY}\right)^{-1/2}\right\|_F^2. \qquad (26)
$$

The minimum problem (26) is equivalent to

$$
\max_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \text{Tr}\left(\mathbf{R^TF^TD^{1/2}Y}\left(\mathbf{Y^TDY}\right)^{-1/2}\right) = \max_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \text{Tr}\left(\mathbf{R}^T\mathbf{M}\right) \qquad (27)
$$

where $\mathbf{M} \triangleq \mathbf{F}^T\mathbf{D^{1/2}Y(Y^TDY)^{-1/2}}$.

Theorem 3 tells how to obtain the optimal solution $\mathbf{R}^*$ to (27).

Theorem 3: Let the single value decomposition (SVD) of $\mathbf{M}$ be $\mathbf{M} = \mathbf{USV}^T$. Then the optimal solution $\mathbf{R}^*$ to the problem of $\max_{\mathbf{R}^T\mathbf{R}=I}$ Tr($\mathbf{R}^T\mathbf{M}$) [i.e., (27)] is

$$
\mathbf{R}^* = \mathbf{UV}^T. \qquad (28)
$$

Proof:

$$
\begin{aligned}
\text{tr}\left(\mathbf{R}^T\mathbf{M}\right) &= \text{tr}\left(\mathbf{R}^T\mathbf{USV}^T\right)\\
&= \text{tr}\left(\mathbf{SV}^T\mathbf{R}^T\mathbf{U}\right)\\
&= \text{tr}(\mathbf{SG})\\
&= \sum_i s_{ii}g_{ii} \qquad (29)
\end{aligned}
$$

where $\mathbf{G} = \mathbf{V}^T\mathbf{R}^T\mathbf{U}$, s_{ii} and g_{ii} are the (i, i) elements of $\mathbf{S}$ and $\mathbf{G}$. Because $\mathbf{GG}^T = \mathbf{V}^T\mathbf{R}^T\mathbf{U}(\mathbf{V}^T\mathbf{R}^T\mathbf{U})^T = \mathbf{V}^T\mathbf{R}^T\mathbf{U}\mathbf{U}^T\mathbf{RV} = \mathbf{I}$, it is true that $\mathbf{G}$ is an orthonormal matrix. According to the property of orthonormal matrix, one can find $-1 \le g_{ij} \le 1$. Moreover, $s_{ii} \ge 0$ holds because s_{ii} is a non-negative singular value. Therefore, it is true that

$$
\text{tr}\left(\mathbf{R}^T\mathbf{M}\right) = \sum_i s_{ii}g_{ii} \le \sum_i s_{ii}. \qquad (30)
$$

Investigating (30), one can find that $\mathrm{tr}(\mathbf{R}^T\mathbf{M})$ equals to its upper bound $\sum_i s_{ii}$ (i.e., $\mathrm{tr}(\mathbf{R}^T\mathbf{M}) = \sum_i s_{ii}$) when $g_{ii} = 1$ (i.e., $\mathbf{G}$ is an identity matrix). Therefore, $\mathbf{R}$ gets its optimal value when

$$\mathbf{G} = \mathbf{V}^T\mathbf{R}^T\mathbf{U} = \mathbf{I} \tag{31}$$

holds.

Equation (31) means that the optimal $\mathbf{R}$ is $\mathbf{R}^* = \mathbf{U}\mathbf{V}^T$. ■

$\mathbf{Y}$-Step: The goal of $\mathbf{Y}$-step is to seek the optimal $\mathbf{Y}$ when $\mathbf{R}$ and $\mathbf{F}$ are fixed. When $\mathbf{R}$ and $\mathbf{F}$ are fixed, the optimization problem (21) is reduced to

$$\begin{aligned}&\min_{\mathbf{Y}\in\mathrm{Ind}}\left\|\mathbf{FR}-\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\right\|_F^2\\ \Rightarrow&\min_{\mathbf{Y}\in\mathrm{Ind}}\left\|\mathbf{F}-\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\mathbf{R}\right\|_F^2.\end{aligned} \tag{32}$$

The minimum of (32) is zero when $\mathbf{F} = \mathbf{D}^{1/2}\mathbf{Y}(\mathbf{Y}^{\mathbf{T}}\mathbf{DY})^{-1/2}\mathbf{R}$. According to (22), the ikth element of $\mathbf{D}^{1/2}\mathbf{Y}(\mathbf{Y}^{\mathbf{T}}\mathbf{DY})^{-1/2}$ is $y_{ik}\sqrt{d_i/\mathbf{d}^T\mathbf{y}_k}$. Therefore, the optimal element of $\mathbf{Y}$ is

$$Y_{ij} = \begin{cases}1, & j = \arg\min_k \left\|\mathbf{f}_i - \sqrt{\frac{d_i}{\mathbf{d}^T\mathbf{y}_k}}\mathbf{r}_k\right\|_F^2\\ 0, & \text{else}\end{cases} \tag{33}$$

where $\mathbf{f}_i$ is the ith row of the matrix $\mathbf{F}$ and $\mathbf{r}_k$ is the kth row of the matrix $\mathbf{R}$.

$\mathbf{F}$-Step: The goal of $\mathbf{F}$-step is to seek the optimal $\mathbf{F}$ when $\mathbf{R}$ and $\mathbf{Y}$ are fixed. When $\mathbf{R}$ and $\mathbf{Y}$ are fixed, the optimization problem (21) becomes

$$\begin{aligned}&\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\begin{bmatrix}\mathrm{tr}\left(\mathbf{F}^T\tilde{\mathbf{L}}F\right)\\ +\,\alpha\left\|\mathbf{FR}-\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\right\|_F^2\end{bmatrix}\\ \Rightarrow&\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\begin{bmatrix}\mathrm{tr}\left(\mathbf{F}^T\tilde{\mathbf{L}}F\right)\\ +\alpha\left(\mathrm{tr}\left(\left(\mathbf{FR}-\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\right)^T\right.\right.\\ \left.\left.\left(\mathbf{FR}-\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\right)\right)\right)\end{bmatrix}\\ \Rightarrow&\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\begin{bmatrix}\mathrm{tr}\left(\mathbf{F}^T\tilde{\mathbf{L}}F\right)\\ -2\alpha\left(\mathrm{tr}\left((\mathbf{FR})^T\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\right)\right.\end{bmatrix}\\ \Rightarrow&\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\left[\mathrm{tr}\left(\mathbf{F}^T\tilde{\mathbf{L}}F\right)-2\alpha\left(\mathrm{tr}\left(\mathbf{F}^T\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\mathbf{R}^T\right)\right]\\ \Rightarrow&\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\left[\mathrm{tr}\left(\mathbf{F}^T\tilde{\mathbf{L}}F-2\alpha\mathbf{F}^T\mathbf{C}\right)\right].\end{aligned} \tag{34}$$

In the last line of (34), the matrix $\mathbf{C}$ is defined as

$$\mathbf{C} = \mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^{\mathbf{T}}\mathbf{DY}\right)^{-1/2}\mathbf{R}^T. \tag{35}$$

The problem of (34) can be further relaxed into

$$\max_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\left[\mathrm{tr}\left(\mathbf{F}^T\mathbf{BF}\right)+2\alpha\left(\mathrm{tr}\left(\mathbf{F}^T\mathbf{C}\right)\right]\right. \tag{36}$$

where $\mathbf{B} = \lambda\mathbf{I} - \tilde{\mathbf{L}} \in \mathbb{R}^{N\times N}$. λ is an arbitrary constant to ensure that $\mathbf{B}$ is a positive definite matrix. Theorem 4 tells how to obtain the optimal solution $\mathbf{F}^*$ to (36).

Algorithm 1: Algorithm to Solve the Problem (36)

```
Input : The matrix $\mathbf{Y}$. The matrix $\mathbf{R}$. The affinity matrix
        $\mathbf{A}$. The degree matrix $\mathbf{D}$. The parameter $\alpha$. The
        maximum number of iteration $T_1$.
Output: $\mathbf{F}$
Initialization:
Compute the parameter $\lambda$ via power method [34], the
normalized Laplacian matrix $\tilde{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, $\mathbf{C}$
according to (35), and $\mathbf{B} = \lambda\mathbf{I} - \tilde{\mathbf{L}}$.
Randomly initialize $\mathbf{F}$.
while convergence criteria not satisfied and number of
iteration $\leq T_1$ do
    Update $\mathbf{E} = \mathbf{BF} + \alpha\mathbf{C}$.
    Calculate $\tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T = \mathbf{E}$ via compact SVD of $\mathbf{E}$.
    Update $\mathbf{F} = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T$.
end
```

Algorithm 2: Proposed Clustering Algorithm

```
Input : The $N$ samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^T \in \mathbb{R}^{N\times M}$.
        The number of clusters $K$. The parameter $\alpha$. The
        maximum number of iteration $T_2, T_3$.
Output: Cluster indicator matrix $\mathbf{Y}$.
Initialization:
Compute the affinity matrix $\mathbf{A}$ according to (1) and the
degree matrix $\mathbf{D}$ according to (2).
Randomly initialize $\mathbf{F}$ and $\mathbf{Y}$.
while convergence criteria not satisfied and number of
iteration $\leq T_3$ do
    For fixed $\mathbf{F}$ and $\mathbf{Y}$, compute the rotation matrix $\mathbf{R}$
    according to $\mathbf{R}^* = \mathbf{U}\mathbf{V}^T$(i.e., eq. (28)).
    For fixed $\mathbf{Y}$ and $\mathbf{R}$, update $\mathbf{F}$ according to
    Algorithm 1.
    while convergence criteria not satisfied and number
    of iteration $\leq T_2$ do
        For fixed $\mathbf{F}$ and $\mathbf{R}$, update $\mathbf{Y}$ according to (33)
    end
end
```

Theorem 4: Let the compact SVD of $\mathbf{E} = \mathbf{BF} + \alpha\mathbf{C}$ be $\mathbf{E} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T$, where $\tilde{\mathbf{U}} \in \mathbb{R}^{N\times K}$, $\tilde{\mathbf{S}} \in \mathbb{R}^{K\times K}$, and $\tilde{\mathbf{V}} \in \mathbb{R}^{K\times K}$. Then the optimal solution $\mathbf{F}^*$ to the problem $\max_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\ [\mathrm{tr}(\mathbf{F}^T\mathbf{BF}) + 2\alpha(\mathrm{tr}(\mathbf{F}^T\mathbf{C})]$ [i.e., (36)] is

$$\mathbf{F}^* = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T. \tag{37}$$

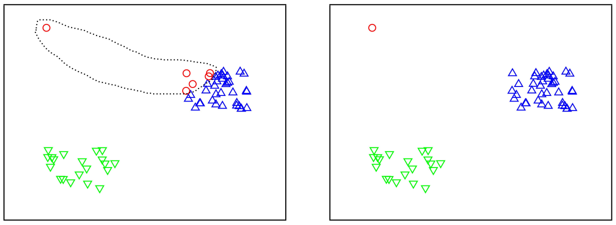
The detailed proof can be found in Appendix A.

Algorithm 1 gives the details of the solution to the problem (36).

Algorithm 2 gives the proposed algorithm, where the $\mathbf{R}$-step, $\mathbf{Y}$-step, and $\mathbf{F}$-step are iteratively conducted.

C. Complexity Analysis

The traditional spectral clustering method consists of spectral embedding and spectral rotation. The time computational complexity of spectral embedding is $O(n^3)$, and the time cost of spectral rotation is $O(K^3+tnK^2)$, where n is the number



(a) (b)

Fig. 1. Comparison of UFDSC and the proposed JSESR on the toy data. (a) UFDSC. (b) Proposed JSESR.

of samples, K is the number of clusters, and t is the number of iteration of spectral rotation. The time cost of traditional spectral clustering is $O(n^3 + K^3 + tnK^2)$.

The time cost of the proposed Algorithm 1 is $O(t_1 nK^2)$. The time computational complexity of the proposed JSESR is $O(t_3(t_1 nK^2 + K^3 + t_2 nK^2))$, where t_1 is the number of iteration of Algorithm 1, t_2 is the number of iteration to update $\mathbf{Y}$, and t_3 is the number of iteration of the proposed JSESR.

For large scale data, the number of data is much larger than the number of clusters (i.e., $n \gg K$). Therefore, the time cost of the proposed JSESR is $O(t_3(t_1 nK^2 + t_2 nK^2))$ and the time cost of the traditional spectral clustering method is $O(n^3 + tnK^2)$. From Figs. 2 and 3, we can find that the proposed method can converge very fast, t_1, t_2, and t_3 are usually small. In addition, we know that $n \gg K$. Therefore, compared with traditional spectral clustering methods, JSESR has much less computation complexity for large scale data.

D. Convergence Analysis

The convergence of Algorithm 1 has been proved in [34]. We introduce the convergence proof of Algorithm 1 according to [34]. Next, we prove the convergence of proposed Algorithm 2.

Theorem 5: Algorithm 1 will monotonically increase the objective of the problem (36) in each iteration until the algorithm converges.

The detailed proof of Theorem 5 is in Appendix B.

Theorem 6: Algorithm 2 will monotonically decrease the objective of the problem (21) in each iteration until the algorithm converges.

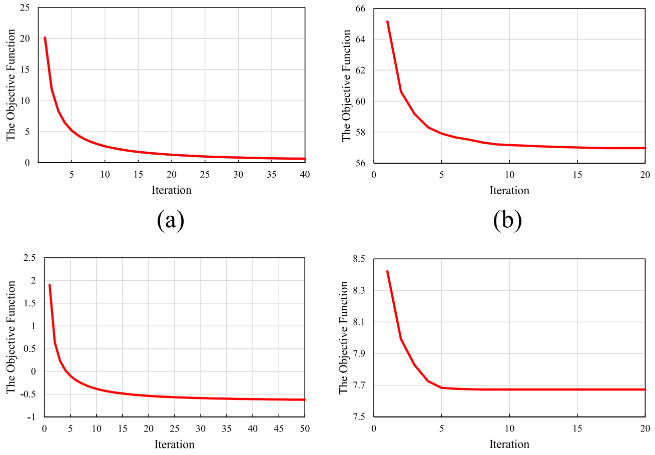
The detailed proof of Theorem 6 can be found in Appendix C.

V. EXPERIMENTAL RESULTS

In this section, we compare the proposed JSESR with K-means [5], Ncut [12], [16], Rcut [11], and UFDSC [17]. Note that, spectral rotation is employed to transform the relaxed continuous results of Ncut and Rcut to binary cluster indicator matrices. Therefore, we denote the Ncut and Rcut algorithms as Ncut+SR and Rcut+SR, respectively.

We begin by giving a toy example in order to show the superiority of the proposed JSESR against the UFDSC.

A. Toy Example for Comparison of UFDSC and the Proposed JSESR



(a) (b) (c) (d)

Fig. 2. (a) and (c) Curves of the objective function in (34) versus iteration number t_1. (b) and (d) Curves of the objective function in (32) versus iteration number t_2.

The toy example shown in Fig. 1 contains three clusters. In the top left of Fig. 1(a) [also Fig. 1(b)], there is a cluster consisting of only 1 sample. In the bottom left of Fig. 1(a) [also Fig. 1(b)], there is a cluster consisting of 20 samples. The cluster in the right of Fig 1(a) [also Fig. 1(b)] consists of 30 samples. It is obvious that the number of each cluster is of unbalance.

Fig. 1(a) and (b) shows the clustering results of the UFDSC and the proposed JSESR, respectively. The samples that are classified as the same cluster are marked with a unique color. From Fig. 1(a), one can find that the UFDSC incorrectly classifies the single sample in the top left of Fig. 1(a) and the five samples in the top left corner of the right of Fig. 1(a) as the same cluster. It is noted that the Ncut itself is able to solve such unbalanced samples. However, the UFDSC is not able to deal with such severely unbalanced situation. The reason is as follows. In the spectral rotation part of the UFDSC, the nonorthonormal cluster indicator matrix $\mathbf{Y}$ is used for approximating the orthonormal matrix $\mathbf{FR}$. When the samples are extremely unbalanced, the degree of orthonormalization of $\mathbf{Y}$ is very low, resulting in large approximation error. From Fig. 1(b), it is observed that the proposed JSESR is capable of perfectly dealing with the unbalanced problem.

B. Datasets

Besides the above-mentioned toy example, 19 benchmark datasets are used for evaluation. Among the 19 datasets, there are 16 image datasets: 1) AR face dataset [35]; 2) AT&T [36]; 3) Binalpha [37]; 4) COIL20 [38]; 5) COIL100 [39]; 6) Jaffe [40]; 7) MPEG7 [41]; 8) MSRA [42]; 9) GeorgiaTech [43]; 10) PIE [44]; 11) UMIST [45]; 12) Yale [46]; 13) Extended Yale Face Database B (Yale B) [46]; 14) C-Cube [47]; 15) FERET [48]; and 16) MNIST [49]. The other ones are from UCI machine learning repository [50]: 1) control; 2) dermatology; and 3) movements. Table I summarizes the characteristics (number of samples, dimension, number of classes) of datasets used in the experiments.

C. Parameter Setting and Convergence Property

The self-tuning spectral clustering method [51] is adopted to determine the parameter t in (1). The number of nearest neighbors used in (1) is set to be five for all the algorithms. In

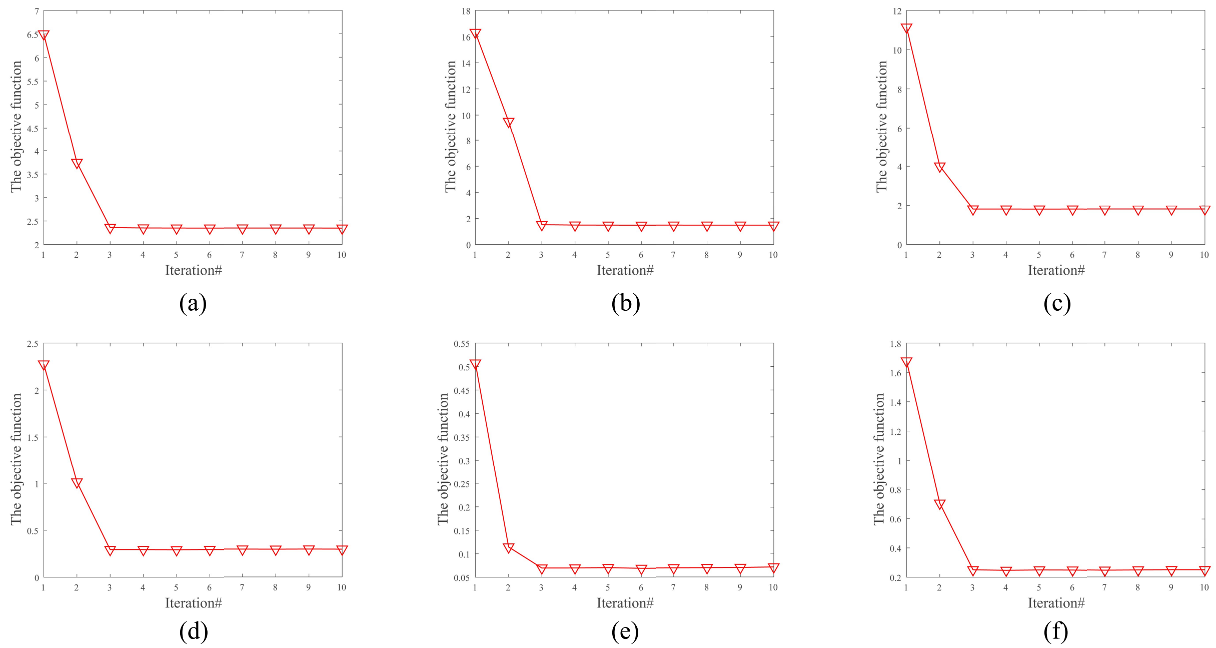


Fig. 3. Curves of the objective function in (21) versus iteration number t_3 on the (a) Binalpha, (b) Coil100, (c) Mpeg7, (d) UMIST, (e) Dermatology, and (f) Movements datasets.

TABLE I
DESCRIPTION OF DATASETS

Datasets	Num of Instances	Dimensions	Classes
AR	2600	792	100
AT&T	400	168	40
Binalpha	1404	320	36
COIL20	1440	1024	20
COIL100	7200	1024	100
Jaffe	213	1024	10
Mpeg7	1400	6000	70
MSRA	1499	256	12
GeorgiaTech	750	1800	50
PIE	3332	256	68
UMIST	360	168	20
Yale	165	256	15
YaleB	2414	2016	38
Control	600	60	6
Dermatology	366	34	6
Movements	360	90	15
C-Cube	38160	34	46
FERET	1400	6400	200
MNIST	70000	784	10

our method, the tradeoff parameter α balances the part of spectral embedding and the part of spectral rotation. The value of α is chosen from the set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$, for different α, the best performance is reported. All the clustering algorithms run 20 times and the average results are reported. The maximum number of iterations T_1, T_2, T_3 are set to be 100, 20, 10, respectively.

Before comparison with different clustering methods, the convergence property of Algorithm 1 and updating $\mathbf{Y}$ according to (33) are shown in Fig. 2. The results are obtained when the tradeoff parameter is set to be $\alpha = 0.1$. The convergence property of the proposed JSESR is shown in Fig. 3. The results on the Binalpha, Coil100, Mpeg7, UMIST, Dermatology, and Movements datasets are obtained when the tradeoff parameter α is set to be 0.1. It is observed that the proposed method convergences in about three iterations. The fast convergence property implies that the proposed JSESR is of high efficiency.

D. Parameter Sensitivity

As mentioned before, we tune tradeoff parameter α in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. The effects of α is shown in Fig. 4. As shown in Fig. 4, it can be found that the clustering accuracy on all eight datasets can get better performance when a small value is used for α. In addition, when $\alpha = 10^{-2}$, our proposed JSESR on all eight datasets can get relatively great performance.

E. Comparison With Other Methods

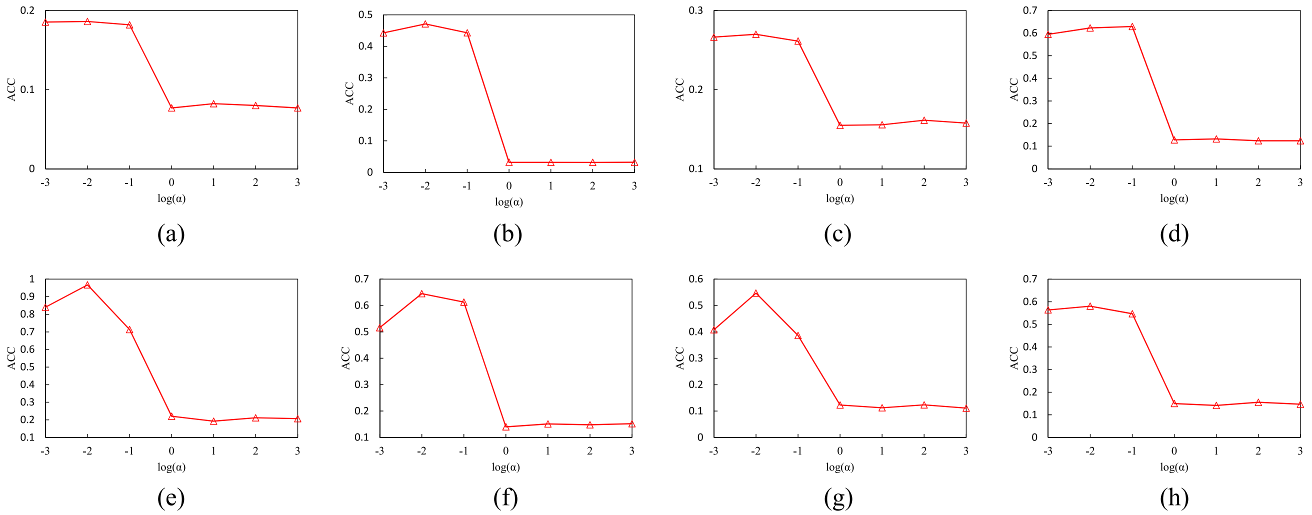
The experimental results of different methods on the 18 benchmark datasets are given in Tables II–VI. Table II gives the ACC performance. It can be seen from Table II that the proposed JSESR achieves the highest ACC for all the 18 datasets.

Table III compares the proposed JSESR with *K*-means, Ncut+SR, and Rcut+SR in terms of NMI. It can be seen from Table III that the proposed JSESR achieves the highest NMI performance for all the 18 datasets.

In Table IV, the Purity of different methods is given. For all the 18 datasets, similar to the phenomena observed from Tables II and III, the proposed method performs better than *K*-means, Ncut+SR, and Rcut+SR.

Tables V and VI give the Homogeneity [52] and Jaccard Index performance of different methods. As shown in Tables V and VI, the proposed JSESR achieves the highest Homogeneity and Jaccard Index performance of all the 18 datasets.

In addition to the *K*-means, Ncut+SR, and Rcut+SR methods, the proposed JSESR is also compared with the UFDSC on



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. Clustering accuracy of the proposed JSESR with different α on (a) AR, (b) C-cube, (c) FERET, (d) GeorigiaTech, (e) Jaffe, (f) MNIST, (g) MSRA, and (h) UMIST datasets.

TABLE II

COMPARISON IN TERMS OF ACC (%) AND THEIR VARIATIONS. THE NUMBER IN BRACKETS IS THE STANDARD DEVIATION (%)

Datasets	AR	AT&T	Binalpha	COIL20	COIL100	GeorgiaTech	Jaffe	Mpeg7	MSRA
K-means	10.26(0.56)	57.65(3.77)	41.36(2.1)	54.84(5.20)	41.83(1.50)	42.11 (2.14)	73.10(8.74	47.18(2.56)	45.85(2.83)
Ncut+SR	13.83(0.21)	73.46(3.32)	43.53(0.93)	80.44(6.05)	72.34(2.23)	62.88(1.04)	90.28(8.94)	50.98(0.30)	47.29(1.66)
Rcut+SR	13.66(0.39)	73.63(1.89)	45.83(0.93)	78.01(6.95)	70.74(1.76)	62.11(1.07)	92.35(5.28)	50.76(0.59)	46.50(2.47)
Ours (JSESR)	**18.53(0.40)**	**76.40(1.21)**	**47.42(0.71)**	**82.93(4.48)**	**83.49(1.75)**	**63.87(0.60)**	**96.06(0.42)**	**52.60(1.12)**	**53.85(1.75)**
Datasets	PIE	UMIST	Yale	YaleB	Control	Dermatology	C-cube	FERET	MNIST
K-means	11.13(0.68)	46.42(1.98)	53.88(4.48)	9.36(0.51)	58.22(3.41)	74.67(7.00)	33.03(1.45)	24.14(0.46)	51.19(4.74)
Ncut+SR	40.33(0.82)	53.42(0.70)	66.45(0.95)	34.79(0.76)	56.33(7.34)	77.13(7.69)	42.77(0.96)	25.50(0.39)	58.01(0.18)
Rcut+SR	40.66(0.55)	53.15(0.94)	65.61(2.41)	34.38(0.36)	59.90(6.81)	81.26(4.53)	44.98(0.97)	25.86(0.41)	57.88(0.21)
Ours (JSESR)	**42.79(0.01)**	**57.44(3.41)**	**66.79(0.27)**	**36.26(0.86)**	**71.63(7.86)**	**83.64(3.48)**	**47.11(0.75)**	**27.01(0.21)**	**61.70(0.57)**

TABLE III

COMPARISON IN TERMS OF NMI (%) AND THEIR VARIATIONS. THE NUMBER IN BRACKETS IS THE STANDARD DEVIATION (%)

Datasets	AR	AT&T	Binalpha	COIL20	COIL100	GeorgiaTech	Jaffe	Mpeg7	MSRA
K-means	37.43 (0.36)	77.11 (0.48)	56.99(1.04)	70.63(3.10)	73.36(0.65)	50.43(1.97)	85.59(0.56)	68.25(1.34)	53.43(2.18)
Ncut+SR	37.52(0.17)	88.08(1.57)	60.55(0.32)	90.05(2.80)	90.47(0.69)	72.01(0.46)	92.30(5.60)	70.54(0.13)	63.72(2.05)
Rcut+SR	37.85(0.25)	88.13(0.93)	60.13(0.38)	89.84(3.30)	89.98(0.58)	71.64(0.55)	93.42(3.39)	70.40(0.19)	61.10(4.39)
Ours (JSESR)	**39.32(0.18)**	**89.12(0.44)**	**61.91(0.22)**	**90.95(1.93)**	**93.38(0.38)**	**75.67(0.22)**	**95.24(0.38)**	**71.56(0.48)**	**68.17(0.92)**
Datasets	PIE	UMIST	Yale	YaleB	Control	Dermatology	C-cube	FERET	MNIST
K-means	61.55(1.59)	55.93(2.94)	61.36(3.82)	12.31(0.74)	69.14(2.08)	81.15(6.49)	44.59(0.30)	65.21(0.23)	47.82(2.64)
Ncut+SR	55.77(0.46)	72.25(0.32)	68.14(1.11)	42.71(0.50)	70.08(7.41)	78.51(7.15)	55.44(0.50)	66.71(0.39)	62.13(0.52)
Rcut+SR	55.91(0.52)	72.01(0.64)	68.83(1.47)	42.24(0.61)	73.90(6.99)	82.48(5.69)	55.85(0.58)	66.85(0.33)	61.95(0.06)
Ours (JSESR)	**63.93(0.47)**	**73.62(1.27)**	**69.29(0.43)**	**44.36(0.52)**	**74.93(3.62)**	**83.71(1.76)**	**56.50(0.99)**	**67.66(0.16)**	**62.65(0.54)**

TABLE IV

COMPARISON IN TERMS OF PURITY (%) AND THEIR VARIATIONS. THE NUMBER IN BRACKETS IS THE STANDARD DEVIATION (%)

Datasets	AR	AT&T	Binalpha	COIL20	COIL00	GeorgiaTech	Jaffe	Mpeg7	MSRA
K-means	10.52(0.82)	60.51(0.98)	44.34(1.77)	58.31(4.81)	50.75(1.30)	44.81(1.60)	74.93(5.63)	50.19(2.14)	47.84(2.38)
Ncut+SR	18.56(0.23)	77.53(3.46)	46.75(0.80)	84.23(6.04)	76.30(2.21)	64.23(1.03)	90.54(8.75)	55.66(0.28)	53.30(2.66)
Rcut+SR	18.28(0.40)	77.79(1.96)	48.49(0.83)	82.61(7.11)	74.80(1.83)	64.40(1.08)	92.35(5.28)	55.36(0.53)	51.83(4.05)
Ours (JSESR)	**18.96(0.42)**	**80.25(0.73)**	**50.09(0.88)**	**85.60(3.45)**	**86.41(1.02)**	**66.00(0.37)**	**96.06(0.42)**	**57.33(1.14)**	**57.03(1.79)**
Datasets	PIE	UMIST	Yale	YaleB	Control	Dermatology	C-Cube	FERET	MNIST
K-means	11.61(0.71)	49.85(2.84)	55.30(3.64)	10.08(0.54)	66.85(2.12)	83.52(5.89)	42.74(1.17)	26.71(0.42)	55.16(3.51)
Ncut+SR	43.12(1.11)	56.86(0.75)	66.56(0.95)	35.93(0.77)	62.92(6.93)	79.54(6.95)	52.12(0.82)	26.50(0.46)	67.39(0.41)
Rcut+SR	43.35(0.87)	56.64(0.86)	66.21(2.53)	35.84(0.39)	66.50(6.53)	83.06(5.15)	53.71(0.85)	26.86(0.42)	67.13(0.07)
Ours (JSESR)	**44.18(0.99)**	**60.06(1.88)**	**67.27(0.00)**	**37.12(0.93)**	**75.20(5.19)**	**86.01(2.58)**	**54.39(0.97)**	**28.29(0.19)**	**68.60(1.06)**

the AR, Control, Dermatology, Movements, and Yale datasets. The results are shown in Fig. 5 where Fig. 5(a)–(c) adopts ACC, NMI, and Purity for comparison. One can find that the proposed JSESR is superior to UFDSC no matter which evaluation metrics are employed.

VI. CONCLUSION

In this paper, we have presented a novel spectral clustering method called JSESR. The method simultaneously performs spectral embedding (i.e., computing a real-valued cluster indicator matrix) and spectral rotation (transforming the

TABLE V

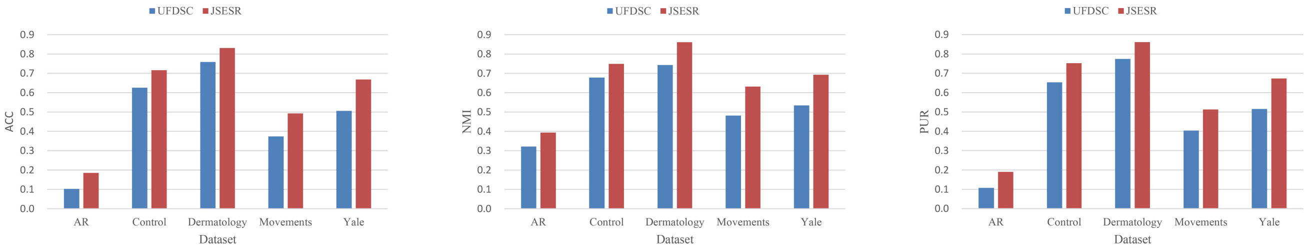
COMPARISON IN TERMS OF HOMOGENEITY (%) AND THEIR VARIATIONS. THE NUMBER IN BRACKETS IS THE STANDARD DEVIATION (%)

Datasets	AR	AT&T	Binalpha	COIL20	COIL00	GeorgiaTech	Jaffe	Mpeg7	MSRA
K-means	36.36(0.54)	74.62(1.65)	56.36(0.78)	74.32(2.11)	73.36(0.65)	65.49(1.15)	79.09(6.05)	65.34(1.19)	59.28(3.78)
Ncut+SR	36.52(0.19)	82.62(2.16)	60.26(0.43)	87.56(2.77)	88.80(0.73)	74.63(0.31)	94.73(6.20)	69.04(0.31)	63.49(2.27)
Rcut+SR	36.31(0.21)	86.09(1.28)	59.65(0.63)	87.26(2.67)	88.61(0.85)	73.85(0.49)	95.55(3.87)	67.59(0.51)	63.38(2.51)
Ours (JSESR)	**36.87(0.30)**	**89.72(0.77)**	**61.44(0.83)**	**88.86(3.55)**	**91.40(0.72)**	**74.73(0.66)**	**96.23(0.45)**	**69.81(0.91)**	**65.18(1.56)**
Datasets	PIE	UMIST	Yale	YaleB	Control	Dermatology	C-cube	FERET	MNIST
K-means	35.61(0.86)	64.94(1.65)	59.96(3.33)	13.450.58	66.331.77	80.766.55	46.49(0.69)	65.68(0.23)	47.82(2.64)
Ncut+SR	55.18(0.41)	71.93(1.32)	65.51(2.67)	42.980.46	67.265.93	85.634.57	55.85(0.53)	66.71(0.36)	60.76(0.52)
Rcut+SR	55.71(0.46)	71.96(0.71)	66.43(1.03)	42.230.42	71.090.99	84.135.67	55.95(0.65)	66.85(0.33)	61.83(0.06)
Ours (JSESR)	**56.06(0.50)**	**75.80(1.32)**	**69.90(0.70)**	**43.63(0.45)**	**71.95(0.91)**	**88.17(1.86)**	**56.44(0.99)**	**67.66(0.16)**	**62.65(2.06)**

TABLE VI

COMPARISON IN TERMS OF JACCARD INDEX (%) AND THEIR VARIATIONS. THE NUMBER IN BRACKETS IS THE STANDARD DEVIATION (%)

Datasets	AR	AT&T	Binalpha	COIL20	COIL00	GeorgiaTech	Jaffe	Mpeg7	MSRA
K-means	5.960.41	40.872.66	26.57(1.71)	42.05(3.31)	30.591.29	29.18(1.26)	53.92(9.92)	28.94(1.63)	28.91(3.54)
Ncut+SR	10.24(0.12)	56.335.43	28.24(0.76)	66.25(8.15)	56.07(3.01)	45.20(0.70)	85.95(9.29)	34.43(0.43)	31.05(0.43)
Rcut+SR	10.21(0.07)	57.713.57	28.96(1.01)	67.07(7.88)	55.76(2.55)	45.50(1.55)	89.06(9.89)	34.23(0.47)	30.79(1.49)
Ours(JSESR)	**10.26(0.05)**	**61.951.07**	**29.46(0.49)**	**67.86(4.89)**	**60.71(2.35)**	**45.92(0.44)**	**93.63(1.02)**	**35.01(0.61)**	**34.45(2.11)**
Datasets	PIE	UMIST	Yale	YaleB	Control	Dermatology	C-cube	FERET	MNIST
K-means	8.49(0.28)	29.01(2.56)	33.50(4.08)	5.18(0.41)	43.714.91	60.219.03	19.09(0.60)	13.73(0.30)	34.52(4.19)
Ncut+SR	25.49(0.51)	36.11(1.96)	43.45(1.21)	21.08(0.28)	41.164.57	69.255.77	27.20(0.79)	14.61(0.26)	40.86(0.18)
Rcut+SR	25.51(0.38)	35.93(0.95)	49.47(1.66)	20.87(0.09)	42.860.90	69.275.56	29.02(0.83)	14.85(0.27)	40.53(0.21)
Ours(JSESR)	**26.28(0.59)**	**40.91(1.21)**	**51.01(0.54)**	**21.14(0.25)**	**49.93(3.78)**	**76.56(7.72)**	**30.81(1.27)**	**15.61(0.14)**	**44.61(0.59)**



(a) (b) (c)

Fig. 5. Comparison of the proposed JSESR and the UFDSC on the AR, Control, Dermatology, Movements, and Yale datasets. (a) ACC. (b) NMI. (c) Purity.

real-valued cluster indicator matrix into binary cluster indicator matrix). The objective function is a tradeoff between the spectral embedding (k-way Ncut) and a variant of spectral rotation. The proposed method employs a scaled cluster indicator matrix to approximate the rotated embedding matrix. Because both the scaled cluster indicator matrix and the embedding matrix are orthonormal matrices, the approximation is precise. We have developed an effective and efficient algorithm to solve the corresponding optimization problem.

APPENDIX A

PROOF OF THEOREM 4

With the technique of Lagrangian multiplier, the constrained maximum optimization problem expressed in (36) can be converted to an unconstrained problem with its objective function being $L(\mathbf{F}, \mathbf{B}, \mathbf{C}, \mathbf{\Lambda})$

$$\begin{aligned} L(\mathbf{F}, \mathbf{B}, \mathbf{C}, \mathbf{\Lambda}) = {} & \operatorname{tr}\big(\mathbf{F}^T\mathbf{B}\mathbf{F}\big) + 2\alpha\big(\operatorname{tr}\big(\mathbf{F}^T\mathbf{C}\big)\big) \\ & - \operatorname{tr}\big(\mathbf{\Lambda}\big(\mathbf{F}^T\mathbf{F} - \mathbf{I}\big)\big). \end{aligned} \tag{38}$$

In (38), the symmetric matrix $\mathbf{\Lambda}$ is the Lagrangian multipliers. Computing the derivative of $L(\mathbf{F}, \mathbf{B}, \mathbf{C}, \mathbf{\Lambda})$ and then setting the derivative to zero yields

$$\frac{\partial L}{\partial \mathbf{F}} = 2\mathbf{B}\mathbf{F} + 2\alpha\mathbf{C} - 2\mathbf{F}\mathbf{\Lambda} = 0. \tag{39}$$

First derive how to compute the matrix $\mathbf{\Lambda}$ of the Lagrangian multipliers and then we describe how to compute the optimal matrix $\mathbf{F}$.

Compute $\mathbf{\Lambda}$*:* Defining $\mathbf{E} = \mathbf{B}\mathbf{F} + \alpha\mathbf{C}$, (39) can be written as

$$\mathbf{F}\mathbf{\Lambda} = \mathbf{B}\mathbf{F} + \alpha\mathbf{C} = \mathbf{E}. \tag{40}$$

The matrix $\mathbf{E}$ can be constructed by compact SVD

$$\mathbf{E} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T \tag{41}$$

where $\tilde{\mathbf{U}} \in \mathbb{R}^{N\times K}$, $\tilde{\mathbf{S}} \in \mathbb{R}^{K\times K}$, and $\tilde{\mathbf{V}} \in \mathbb{R}^{K\times K}$.

Multiplying $(\mathbf{F}\mathbf{\Lambda})^T$ on the left-hand side of (40) and simultaneously multiplying $\mathbf{E}^T$ on the right-hand side of (40) result in

$$\begin{aligned} (\mathbf{F}\mathbf{\Lambda})^T(\mathbf{F}\mathbf{\Lambda}) &= \mathbf{E}^T\mathbf{E} \\ \mathbf{\Lambda}^T\mathbf{F}^{\mathbf{T}}\mathbf{F}\mathbf{\Lambda} &= \left(\tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T\right)^T\left(\tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T\right) \\ \mathbf{\Lambda}^T\mathbf{\Lambda} &= \tilde{\mathbf{V}}\tilde{\mathbf{S}}^2\tilde{\mathbf{V}}^T. \end{aligned} \tag{42}$$

Because $\mathbf{\Lambda}$ is symmetric, $\mathbf{\Lambda} = \mathbf{\Lambda}^T$ and $\mathbf{\Lambda}^T\mathbf{\Lambda} = \mathbf{\Lambda}^2$ hold. Therefore, $\mathbf{\Lambda}$ can be obtained according to the last line of (42)

$$\mathbf{\Lambda} = \tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T. \tag{43}$$

Compute $\boldsymbol{F}$*:* Substitute (43) and (41) into (40), we have

$$\mathbf{F}\left(\tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T\right) = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T \tag{44}$$

and

$$\mathbf{F}^* = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T. \tag{45}$$

Equation (45) gives the optimal solution of $\mathbf{F}$.

APPENDIX B
PROOF OF THEOREM 5

The solution to the problem (36) can be solved by following the problem $\max_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} \operatorname{tr}(\mathbf{F}^T\mathbf{E})$, where $\mathbf{E} = \mathbf{B}\mathbf{F} + \alpha\mathbf{C}$.

Suppose $\tilde{\mathbf{F}}$ is the optimized solution of the problem (36), then

$$\operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{E}\right) \geq \operatorname{tr}\left(\mathbf{F}^T\mathbf{E}\right). \tag{46}$$

We substitute $\mathbf{E} = \mathbf{B}\mathbf{F} + \alpha\mathbf{C}$ into (46), then

$$\operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{B}\tilde{\mathbf{F}}\right) - 2\operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{B}\mathbf{F}\right) + \operatorname{tr}\left(\mathbf{F}^T\mathbf{B}\mathbf{F}\right) \geq 0. \tag{47}$$

$\mathbf{B}$ is positive definite, so we could find $\mathbf{B} = \mathbf{L}^T\mathbf{L}$ vis Cholesky factorization. Because $\| \bullet \|_F^2 \geq 0$, we have

$$\begin{aligned}
&\left\|\mathbf{L}\tilde{\mathbf{F}} - \mathbf{L}\mathbf{F}\right\|_F^2 \geq 0\\
&\Rightarrow \operatorname{tr}\left(\left(\mathbf{L}\tilde{\mathbf{F}} - \mathbf{L}\mathbf{F}\right)^T\left(\mathbf{L}\tilde{\mathbf{F}} - \mathbf{L}\mathbf{F}\right)\right) \geq 0\\
&\Rightarrow \operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{B}\tilde{\mathbf{F}}\right) - 2\operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{B}\mathbf{F}\right) + \operatorname{tr}\left(\mathbf{F}^T\mathbf{B}\mathbf{F}\right) \geq 0.
\end{aligned} \tag{48}$$

Based on (47) and (48), we could infer that

$$\operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{B}\tilde{\mathbf{F}}\right) + 2\alpha\operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{C}\right) \geq \operatorname{tr}\left(\mathbf{F}^T\mathbf{B}\mathbf{F}\right) + 2\alpha\operatorname{tr}\left(\mathbf{F}^T\mathbf{C}\right). \tag{49}$$

Therefore, Algorithm 1 increases the value of the objective function in (36) monotonically in each iteration until it converges.

APPENDIX C
PROOF OF THEOREM 6

Suppose $\tilde{\mathbf{F}}$, $\tilde{\mathbf{R}}$, and $\tilde{\mathbf{Y}}$ are the optimized solution of the problem (21).

Based on Theorem 5, we substitute $\mathbf{B} = \lambda\mathbf{I} - \bar{\mathbf{L}}$ and $\mathbf{C} = \mathbf{D}^{1/2}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-1/2}\mathbf{R}^T$ into (49). Equation (49) can be rewritten as

$$\begin{aligned}
&\operatorname{tr}\left(\tilde{\mathbf{F}}^T\left(\lambda\mathbf{I} - \bar{\mathbf{L}}\right)\tilde{\mathbf{F}}\right) + 2\alpha\operatorname{tr}\left(\tilde{\mathbf{F}}^T\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\mathbf{R}^T\right)\\
&\geq \operatorname{tr}\left(\mathbf{F}^T\left(\lambda\mathbf{I} - \bar{\mathbf{L}}\right)\mathbf{F}\right) + 2\alpha\operatorname{tr}\left(\mathbf{F}^T\mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\mathbf{R}^T\right)\\
&\Rightarrow -\operatorname{tr}\left(\tilde{\mathbf{F}}^T\bar{\mathbf{L}}\tilde{\mathbf{F}}\right) - \alpha\left\|\tilde{\mathbf{F}}^T\mathbf{R} - \mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\right\|_F^2\\
&\geq -\operatorname{tr}\left(\mathbf{F}^T\bar{\mathbf{L}}\mathbf{F}\right) - \alpha\left\|\mathbf{F}^T\mathbf{R} - \mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\right\|_F^2\\
&\Rightarrow \operatorname{tr}\left(\tilde{\mathbf{F}}^T\bar{\mathbf{L}}\tilde{\mathbf{F}}\right) + \alpha\left\|\tilde{\mathbf{F}}^T\mathbf{R} - \mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\right\|_F^2\\
&\leq \operatorname{tr}\left(\mathbf{F}^T\bar{\mathbf{L}}\mathbf{F}\right) + \alpha\left\|\mathbf{F}^T\mathbf{R} - \mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\right\|_F^2.
\end{aligned} \tag{50}$$

According to (33), the problem (50) becomes

$$\begin{aligned}
&\operatorname{tr}\left(\tilde{\mathbf{F}}^T\bar{\mathbf{L}}\tilde{\mathbf{F}}\right) + \alpha\left\|\tilde{\mathbf{F}}^T\mathbf{R} - \mathbf{D}^{1/2}\tilde{\mathbf{Y}}\left(\tilde{\mathbf{Y}}^T\mathbf{D}\tilde{\mathbf{Y}}\right)^{-1/2}\right\|_F^2\\
&\quad\leq \operatorname{tr}\left(\mathbf{F}^T\bar{\mathbf{L}}\mathbf{F}\right) + \alpha\left\|\mathbf{F}^T\mathbf{R} - \mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\right\|_F^2.
\end{aligned} \tag{51}$$

As proved in Theorem 3, $\tilde{\mathbf{R}}$ is the optimal solution to the problem (27). The problem (51) becomes

$$\begin{aligned}
&\operatorname{tr}\left(\tilde{\mathbf{F}}^T\bar{\mathbf{L}}\tilde{\mathbf{F}}\right) + \alpha\left\|\tilde{\mathbf{F}}^T\tilde{\mathbf{R}} - \mathbf{D}^{1/2}\tilde{\mathbf{Y}}\left(\tilde{\mathbf{Y}}^T\mathbf{D}\tilde{\mathbf{Y}}\right)^{-1/2}\right\|_F^2\\
&\quad\leq \operatorname{tr}\left(\mathbf{F}^T\bar{\mathbf{L}}\mathbf{F}\right) + \alpha\left\|\mathbf{F}^T\mathbf{R} - \mathbf{D}^{1/2}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{D}\mathbf{Y}\right)^{-1/2}\right\|_F^2.
\end{aligned} \tag{52}$$

Therefore, Algorithm 2 decreases the value of the objective function in (21) monotonically in each iteration until it converges.

REFERENCES

[1] L. Guo *et al.*, "Two-stage local constrained sparse coding for fine-grained visual categorization," *Sci. China Inf. Sci.*, vol. 61, no. 1, 2018, Art. no. 018104.

[2] Y. Pang, L. Ye, X. Li, and J. Pan, "Incremental learning with saliency map for moving object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 640–651, Mar. 2018.

[3] G. Cui, X. L. Li, and Y. Dong, "Subspace clustering guided convex nonnegative matrix factorization," *Neurocomputing*, vol. 292, pp. 38–48, May 2018.

[4] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Stat. Probab.*, 1967, pp. 281–297.

[6] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[7] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, no. 2, pp. 125–137, 2002.

[8] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. 4th IEEE Int. Conf. Data Min.*, 2004, pp. 19–26.

[9] F. Huang, X. Li, S. Zhang, and J. Zhang, "Harmonious genetic clustering," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 199–214, Jan. 2018.

[10] Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1101–1113, Nov. 1993.

[11] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 11, no. 9, pp. 1074–1085, Sep. 1992.

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

[13] C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *Proc. IEEE Int. Conf. Data Min.*, San Jose, CA, USA, 2001, pp. 107–114.

[14] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1796–1808, Nov. 2011.

[15] P. K. Chan, M. D. F. Schlag, and J. Y. Zien, "Spectral K-way ratio-cut partitioning and clustering," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 13, no. 9, pp. 1088–1096, Sep. 1994.

[16] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 313–319.

[17] Y. Yang, F. Shen, Z. Huang, and H. T. Shen, "A unified framework for discrete spectral clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2273–2279.

[18] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[19] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Int. Conf. Knowl. Disc. Data Min.*, 1996, pp. 226–231.

[20] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, 1999.

[21] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recognit.*, vol. 41, no. 1, pp. 176–190, 2008.

[22] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.

[23] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[24] S. Mehrkanoon, C. Alzate, R. Mall, R. Langone, and J. A. K. Suykens, "Multiclass semisupervised learning based upon kernel spectral clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 720–733, Apr. 2015.

[25] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1083–1094, May 2015.

[26] R. Panda, S. K. Kuanar, and A. S. Chowdhury, "Nyström approximated temporally constrained multisimilarity spectral clustering approach for movie scene detection," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 836–847, Mar. 2018.

[27] Y. Pang, S. Wang, and Y. Yuan, "Learning regularized LDA by clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2191–2201, Dec. 2014.

[28] Z. Li, J. Zhang, K. Zhang, and Z. Li, "Visual tracking with weighted adaptive local sparse appearance model via spatio-temporal context learning," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4478–4489, Sep. 2018.

[29] C. Luo, Z. Li, K. Huang, J. Feng, and M. Wang, "Zero-shot learning via attribute regression and class prototype rectification," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 637–648, Feb. 2018.

[30] F. R. K. Chung, *Spectral Graph Theory*, Amer. Math. Soc., 1997.

[31] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM J. Res. Develop.*, vol. 17, no. 5, pp. 420–425, 1973.

[32] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Math. J.*, vol. 23, no. 2, pp. 298–305, 1973.

[33] J. Huang, F. Nie, and H. Huang, "Spectral rotation versus K-means in spectral clustering," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 431–437.

[34] F. Nie, R. Zhang, and X. Li, "A generalized power iteration method for solving quadratic problem on the Stiefel manifold," *Sci. China Inf. Sci.*, vol. 60, no. 11, 2017, Art. no. 112101.

[35] A. M. Martinez, "The are face database," CVC, New Delhi, India, Rep. #24, 1998.

[36] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 1994, pp. 138–142.

[37] *Binary Alphadigits Database*. Accessed: Sep. 10, 2015. [Online]. Available: https://cs.nyu.edu/∼roweis/data.html.

[38] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Rep. CUCS-006-96, 1996.

[39] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Rep. CUCS-006-96, 1996.

[40] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.

[41] L. J. Latecki, R. Lakamper, and T. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2000, pp. 424–429.

[42] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[43] *Georgia Tech Face Database*. Accessed: Feb. 1, 2001. [Online]. Available: http://www.anefian.com/research/face_reco.htm

[44] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

[45] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*. Heidelberg, Germany: Springer, 1998, pp. 446–456.

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

[47] F. Camastra, M. Spinetti, and A. Vinciarelli, "Cursive character challenge: A new database for machine learning and pattern recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 385–411.

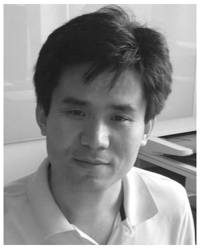
[48] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.

[49] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[50] D. Dheeru and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[51] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1601–1608.

[52] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn.*, 2007, pp. 410–420.



Yanwei Pang (M'07–SM'09) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2004.

He is currently a Professor with Tianjin University, Tianjin, China. He has published over 100 scientific papers, including over 30 IEEE TRANSACTIONS papers. His current research interests include object detection, image recognition, image processing, and deep learning and their applications in self-driving cars, unmanned surface vessel, visual surveillance, human–machine interaction, and biometrics.

Jin Xie received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2016, where he is currently pursuing the Ph.D. degree under the supervisor of Prof. Y. Pang.

His current research interests include machine learning and computer vision.

Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He has published over 100 papers in the top journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, SIGIR, and ACM MM. His current research interest includes machine learning and its application fields.

Dr. Nie is serving as an associate editor or a program committee member for several prestigious journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and conferences in the related fields.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.