IEEE TRANSACTIONS ON

IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

MARCH 2019 VOLUME 28 NUMBER 3 IIPRE4 (ISSN 1057-7149)

For the March 2019 issue, see p. 1049 for Table of Contents.

IEEE

IEEE TRANSACTIONS ON

IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

MARCH 2019 VOLUME 28 NUMBER 3 IIPRE4 (ISSN 1057-7149)

PAPERS

(Contents Continued on Page 1052)

IEEE TRANSACTIONS ON IMAGE PROCESSING (ISSN 1057-7149) is published monthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $641.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in print, electronic, and CD-ROM. **Postmaster**: Send address changes to IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

(Contents Continued from Page 1051)

(Contents Continued on Page 1053)

(Contents Continued from Page 1052)

EDICS ..... Available online at https://signalprocessingsociety.org/publications-resources/ieee-transactions-image-processing

Information for Authors ..... Available online at https://signalprocessingsociety.org/publications-resources/information-authors

IEEE TRANSACTIONS ON

IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

MARCH 2019 VOLUME 28 NUMBER 3 IIPRE4 (ISSN 1057-7149)

PAPERS

(Contents Continued on Page 1050)

IEEE TRANSACTIONS ON IMAGE PROCESSING (ISSN 1057-7149) is published monthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $641.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in print, electronic, and CD-ROM. **Postmaster**: Send address changes to IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

(Contents Continued from Page 1049)

EDICS ..... Available online at https://signalprocessingsociety.org/publications-resources/ieee-transactions-image-processing

Information for Authors ..... Available online at https://signalprocessingsociety.org/publications-resources/information

Multiview Consensus Graph Clustering

Kun Zhan, Feiping Nie, Jing Wang, and Yi Yang

***Abstract*— A graph is usually formed to reveal the relationship between data points and graph structure is encoded by the affinity matrix. Most graph-based multiview clustering methods use predefined affinity matrices and the clustering performance highly depends on the quality of graph. We learn a consensus graph with minimizing disagreement between different views and constraining the rank of the Laplacian matrix. Since diverse views admit the same underlying cluster structure across multiple views, we use a new disagreement cost function for regularizing graphs from different views toward a common consensus. Simultaneously, we impose a rank constraint on the Laplacian matrix to learn the consensus graph with exactly k connected components where k is the number of clusters, which is different from using fixed affinity matrices in most existing graph-based methods. With the learned consensus graph, we can directly obtain the cluster labels without performing any post-processing, such as k-means clustering algorithm in spectral clustering-based methods. A multiview consensus clustering method is proposed to learn such a graph. An efficient iterative updating algorithm is derived to optimize the proposed challenging optimization problem. Experiments on several benchmark datasets have demonstrated the effectiveness of the proposed method in terms of seven metrics.**

***Index Terms*— Unsupervised learning, multiview clustering, image retrieval, graph learning.**

I. Introduction

MULTIVIEW data are pervasive in many domains. For example, an image can be represented by different types of handcrafted descriptors; A news can be reported by multiple articles in different languages; and a person can be identified by face, fingerprint, iris, and signature. With the increasing amount of multiview data, it is very important to exploit the mutual agreement of diverse views information to obtain better clustering performance than using any single view data [1].

Manuscript received March 12, 2018; revised August 16, 2018 and September 17, 2018; accepted October 13, 2018. Date of publication October 22, 2018; date of current version November 7, 2018. This work was supported in part by the National Science Foundation of China under Grant 61201422 and in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20120211120013. This work was done when K. Zhan was visiting the University of Technology Sydney. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xudong Jiang. *(Corresponding author: Jing Wang.)*

K. Zhan is with the School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: kzhan@lzu.edu.cn).

F. Nie with the School of Computer Science and Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: feipingnie@gmail.com)

J. Wang is with the Department of Creative Informatics, The University of Tokyo, Tokyo 113-8654, Japan (e-mail: jing_wang@mist.i.u-tokyo.ac.jp).

Y. Yang is with the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: yi.yang@uts.edu.au).

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/TIP.2018.2877335

Data clustering is a fundamental topic of unsupervised learning. Since a particular single view cannot provide comprehensive information to fully reflect the essence of cluster structure, it has become popular to develop unsupervised multiview learning methods by integrating heterogeneous and complementary information. Due to the well-defined mathematical framework of spectral clustering (SC) [2]–[4], numerous methods extend SC to unsupervised multiview learning. Co-training SC [5] searches for the clusterings that agree across the views. Co-regularized SC [6] maximizes the agreement between different views. Robust multiview SC [7] separates noise from graphs and learns a shared low-rank transition probability matrix. Multiview SC via bipartite graph [8] mainly attains high accuracy with low computational complexity but its performance highly relies on the selection of the number of salient points. These SC-based methods use the predefined similarity matrix and the clustering performance of these methods highly depends on the quality of the similarity matrix.

Different from using a predefined graph, clustering with adaptive neighbors (CAN) learns the graph structure and the embedding matrix simultaneously [9] and it obtains a higher clustering performance than SC. CAN can learn to obtain a graph with exactly k connected components where k is the number of clusters. Since the graph has an ideal graph structure, CAN does not require to perform post-processing SC and k-means clustering algorithms. Some graph-based methods use a weighted-sum rule to extend these single view CAN to multiview learning [10], [11]. However, none of CAN-based has exploited agreement between diverse views efficiently. The objective function of [11] learns the weights and a global graph simultaneously. Its final learned global graph can be regarded as the weighted sum of different graphs. Different from [11], we use a specific disagreement cost to explore the consistency between different views. With the disagreement cost, graphs from different views regularize each other towards a common consensus.

In this paper, a multiview consensus graph clustering method is proposed based on two motivations: using a new disagreement cost function and a rank constraint on the Laplacian matrix. Since different views admit the same underlying cluster structure of the data, 1) we leverage the common consensus information derived from the connections between diverse views to better exploit the cluster structure and 2) we learn a unified affinity graph with a rank constraint on the Laplacian matrix so that the graph has k connected components—k clusters, *i.e.*, each cluster belongs to one connected component of the graph. The two motivations are combined into an overall objective function. To solve the

challenging optimization problem, we propose an optimization procedure according to two theorems related to the graph Laplacian matrix. To evaluate the effectiveness of the proposed method, we conduct experiments on three benchmark datasets in comparison to state-of-the-art approaches. Experimental results demonstrate that the proposed method performs better than other approaches consistently.

The remainder of this paper is organized as follows. In Section II, we introduce some related works. In Section III, an objective function is proposed for optimizing a global graph constrained by the rank of the Laplacian matrix. In Section V, we propose a novel algorithm to optimize the well-designed objective function. In Section VI, numerical experiments are conducted. We use three datasets and compare with seven state-of-the-art methods. Section VII concludes with some discussion.

II. Related Works

Multiview learning stems from the co-training in 1998 [1]. Co-training is based on assumptions of sufficiency, compatibility, and conditional independence. In unsupervised learning, Kumar *et al.* [5], [6] adopt co-training strategies to minimize the disagreement between diverse views. Later, inspired by co-training, many unsupervised multiview learning methods use the Hilbert-Schmidt independence criterion (HSIC) [12], [13] to learn a global similarity graph [14]–[18] or an embedding matrix [19]. Besides using co-training strategy and HSIC, canonical correlation analysis is used to exploit feature correlation between different views [20], [21].

Most multiview learning methods are similarity-based approaches. A graph or a kernel can be used to characterize the pairwise affinity between data points and an element of the graph or the kernel can be regarded as the similarity between two data points [22]. Since most multiview learning methods need to exploit the similarity, k-means clustering [23], non-negative matrix factorization (NMF) [24], SC [3], graph embedding [25], subspace learning [26], and multiple-kernel learning [27] are usually extended to multiview setting. The objective functions of k-means clustering, NMF, and SC exploit similar semantic data structure and are equivalent under certain transforms [28]–[30]. Besides extending k-means clustering to multiview setting directly [31], there are two ways using k-means clustering in multiview learning, one way is multiview discriminatively embedded k-means clustering [32], [33], and the other is multiple kernels k-means clustering [34]–[36]. Since SC-based methods need to predefine a similarity matrix, subspace learning methods [15]–[17], [37]–[41] are used to learn a better similarity matrix than using a fixed function. Multiple kernel learning methods [42]–[45] construct different kernels in different views to integrate these kernels in linear or non-linear manners. Since graphs or kernels can be concatenated to be a tensor, tensor computation is applied to multiview learning [46]–[48].

The aforementioned methods need to predefine graphs or kernels and to learn the clustering results. In SC, define a graph, learn an embedding matrix, and then perform k-means clustering algorithm to obtain the final clustering results. Different from these three steps methods, Nie *et al.* [9] recently propose CAN to learn the similarity matrix and the embedding matrix simultaneously, which does not need to perform post-processing k-means clustering. Inspired by CAN, multiview clustering methods [10], [11] use a rank constraint on the Laplacian matrix to learn a global graph with k number of connected components but does not exploit the common consensus in diverse views well.

III. Preliminaries

A dataset is denoted by $\mathcal{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n_v)}\}$ and it has n_v-view features. Without loss of generality, data matrix is represented by $\mathbf{X}^{(v)} = [\mathbf{X}_1^{(v)}, \mathbf{X}_2^{(v)}, \ldots, \mathbf{X}_k^{(v)}] = [\boldsymbol{x}_1^{(v)}, \boldsymbol{x}_2^{(v)}, \ldots, \boldsymbol{x}_n^{(v)}] \in \mathbb{R}^{d^{(v)} \times n}$, where $\mathbf{X}_c^{(v)}$ denotes the data matrix belonging to the c-th cluster, $\boldsymbol{x}_i^{(v)}$ denotes a data point in v-th view, n is the number of data points, $d^{(v)}$ is the dimension, and k is the number of clusters.

The goal of spectral clustering is to partition data points into k weakly inter-connected clusters [2]–[4]. First, an affinity matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n \times n}$ is constructed to model the similarity w_{ij} between pairwise $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Second, the k number of eigenvectors $\mathbf{H}$ of the normalized Laplacian matrix $\mathbf{L}$ corresponding to the top k smallest eigenvalues are regarded as the low dimensional embedding of the raw data $\mathbf{X}$. Third, k-means clustering algorithm is performed by using the rows of $\mathbf{H}$ as feature vectors to partition data points into k clusters. The objective function is given by,

$$\begin{aligned} &\min_{\mathbf{H}} \ \mathrm{Tr}(\mathbf{H}^\top \mathbf{L}\mathbf{H}) \\ &\text{s.t. } \mathbf{H} \in \mathbb{R}^{n \times k}, \quad \mathbf{H}^\top \mathbf{H} = \mathbf{I}, \end{aligned} \tag{1}$$

where $\mathbf{H}$ is the low dimensional embedding matrix, $\mathrm{Tr}(\cdot)$ is the trace operator, $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$ is the normalized Laplacian matrix, $\mathbf{D}$ is a diagonal matrix with each diagonal element $d_{jj} = \sum_{i=1}^{n} w_{ij}$, and $\mathbf{I}$ is an identity matrix.

It is straightforward to check that the performance of the spectral clustering highly depends on the quality of the predefined affinity matrix $\mathbf{W}$. Usually, the graph structure may vary with varying the graph construction methods and the number of the connected components of $\mathbf{W}$ is unknown and uncertain. Assuming that the graph has exactly k connected components, clustering results can be obtained from the graph itself since each component belongs to one cluster. The graph structure can be learned to obtain such an ideal structure adaptively until the sum of top k smallest eigenvalues of $\mathbf{L}$ is zeroed [9], [11], which can be achieved by the following objective function,

$$\begin{aligned} &\min_{\mathbf{H},\mathbf{W}} \ \mathrm{Tr}(\mathbf{H}^\top \mathbf{L}\mathbf{H}) + \alpha \|\mathbf{W}\|_{\mathrm{F}}^2 \\ &\text{s.t. } \mathbf{H} \in \mathbb{R}^{n \times k}, \quad \mathbf{H}^\top \mathbf{H} = \mathbf{I}, \\ &\qquad \mathbf{W} \geq \mathbf{0}, \quad \mathbf{W}\mathbf{1} = \mathbf{1}, \end{aligned} \tag{2}$$

where α is a regularization parameter, and $\mathbf{W}$ is constrained by $\mathbf{W}\mathbf{1} = \mathbf{1}$ so that a normalized Laplacian matrix $\mathbf{L} = \mathbf{I} - \mathbf{W}$ is obtained.

Eq. (2) can be solved alternately by optimizing two subproblems until the sum of the top k smallest eigenvalues of $\mathbf{L}$ is zeroed [11]. $\mathbf{W}$ optimized by Eq. (2) has an ideal neighbors assignment and the data points are already partitioned

into k clusters [4], [11], [49], which is inspired by the following two theorems related to $\mathbf{L}$.

Theorem 1: The number k of connected components of the graph is equal to the multiplicity of 0 as an eigenvalue of $\mathbf{L}$.

Proof: Since $\mathbf{L}$ is positive semi-definite [4], [49], all eigenvalues of $\mathbf{L}$ are non-negative. It is straightforward to check that $\sum_{i,j=1}^{n} w_{ij}(h_i - h_j)^2 = 0$ while $w_{ij} \geq 0$, if and only if h is constant on each connected component. ∎

Without loss of generality, eigenvalues of $\mathbf{L}$ are arranged in ascending order: $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Theorem 1 indicates that if the sum of the top k smallest eigenvalues satisfies the constraint $\sum_{i=1}^{k} \lambda_i = 0$, i.e., the rank of the Laplacian matrix is $\text{rank}(\mathbf{L}) = n - k$, the graph $\mathbf{W}$ has an ideal neighbors assignment and has exactly k connected components.

Theorem 2: Eigenvalues of $\mathbf{L}$ are ordered by $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and the corresponding eigenvectors are $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_n$. The inequality $\sum_{i=1}^{k} \lambda_i \leq \sum_{i=1}^{k} \boldsymbol{h}_i^\top \mathbf{L} \boldsymbol{h}_i$ holds for any orthogonal vectors $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_k$.

Proof: Since both $\boldsymbol{h}_i$ and $\boldsymbol{\varphi}_i$ are orthogonal vectors and $k \leq n$, we have $\sum_{i=1}^{k} (\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2 \leq \sum_{i=1}^{n} (\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2 = 1$. For each j, we have

$$\begin{aligned}
\boldsymbol{h}_j^\top \mathbf{L} \boldsymbol{h}_j &= \boldsymbol{h}_j^\top \Big(\sum_{i=1}^{n} \lambda_i \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^\top\Big) \boldsymbol{h}_j \\
&= \sum_{i=1}^{n} \lambda_i (\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2 \\
&= \lambda_k \sum_{i=1}^{n} (\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2 + \sum_{i=1}^{k} (\lambda_i - \lambda_k)(\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2 \\
&\quad + \sum_{i=k+1}^{n} (\lambda_i - \lambda_k)(\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2 \\
&\geq \lambda_k + \sum_{i=1}^{k} (\lambda_i - \lambda_k)(\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2.
\end{aligned}$$

Thus,

$$\sum_{i=1}^{k} \lambda_i - \sum_{j=1}^{k} \boldsymbol{h}_j^\top \mathbf{L} \boldsymbol{h}_j \leq \sum_{i=1}^{k} (\lambda_i - \lambda_k)\Big(1 - \sum_{j=1}^{k} (\boldsymbol{h}_j^\top \boldsymbol{\varphi}_i)^2\Big) \leq 0.$$

Then, the inequality $\sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} \boldsymbol{\varphi}_i^\top \mathbf{L} \boldsymbol{\varphi}_i = \min_{\mathbf{H}^\top \mathbf{H} = \mathbf{I}} \text{Tr}(\mathbf{H}^\top \mathbf{L} \mathbf{H}) \leq \sum_{i=1}^{k} \boldsymbol{h}_i^\top \mathbf{L} \boldsymbol{h}_i$ holds. ∎

Theorem 2 is a special case of Fan's theorem [50] for $\mathbf{L}$. Theorem 2 indicates that $\text{rank}(\mathbf{L}) = n - k$ can be achieved by minimizing Eq. (1) until $\sum_{i=1}^{k} \lambda_i = 0$. Since a fixed $\mathbf{W}$ can hardly obtain such a result, $\mathbf{W}$ and $\mathbf{H}$ can be learned by solving Eq. (2) simultaneously [11].

In [11], Eq. (2) is solved alternatively by optimizing two subproblems. One subproblem is an eigenvalue decomposition problem as solving Eq. (1), and the other is a Euclidean projection problem on the simplex space [51]–[53]. Since each column $\boldsymbol{w}_j$ of $\mathbf{W}$ is independent of each other, the projection problem is given by,

$$\begin{aligned}
&\min_{\boldsymbol{w}_j} \ \boldsymbol{w}_j^\top \boldsymbol{g} + \alpha \boldsymbol{w}_j^\top \boldsymbol{w}_j \\
&\text{s.t. } \boldsymbol{w}_j \geq 0, \quad \mathbf{1}^\top \boldsymbol{w}_j = 1,
\end{aligned} \tag{3}$$

where $\boldsymbol{g}_j$ is a vector and its element is $g_{ij} = \|h_i - h_j\|_2^2$.

According to [9] and [11], the projection problem Eq. (3) is used to tune the graph structure. A trade-off between two graph structures can be achieved by tuning α, the first case is that one vertex is connected with only one other vertex, and the second case is that all vertices are connected with each other by the same weight $\frac{1}{n}$. The trade-off renders $\sum_{i=1}^{k} \lambda_i$ close to zero.

The first case is to optimize the objective function,

$$\begin{aligned}
&\min_{\boldsymbol{w}_j} \ \boldsymbol{w}_j^\top \boldsymbol{g}_j \\
&\text{s.t. } \boldsymbol{w}_j \geq 0, \quad \mathbf{1}^\top \boldsymbol{w}_j = 1.
\end{aligned} \tag{4}$$

Eq. (4) returns a minimum value $g_{ij} = \min(\boldsymbol{g}_j)$ and the solution of Eq. (4) is that the i-th element of $\boldsymbol{w}_j$ is assigned to one and others are zeros, *i.e.*, the j-th vertex is only connected to only one other the i-th vertex with the weight of $w_{ij}^\star = 1$ in graph.

The second case is given by,

$$\begin{aligned}
&\min_{\boldsymbol{w}_j} \ \boldsymbol{w}_j^\top \boldsymbol{w}_j \\
&\text{s.t. } \boldsymbol{w}_j \geq 0, \quad \mathbf{1}^\top \boldsymbol{w}_j = 1.
\end{aligned} \tag{5}$$

The solution of Eq. (5) is $w_{ij}^\star = \frac{1}{n}$ for all elements in $\boldsymbol{w}_j$, which indicates that the j-th vertex is connected with all vertices in graph.

In practice, the structure of the graph $\mathbf{W}$ can be tuned by α in a heuristic way to accelerate the procedure [9]. α is changed during the iteration. It is initialized by $\alpha = 1$, then it is increased if the number of connected components of graph $\mathbf{S}$ is larger than k and α is decreased if the number is smaller than k in each iteration. According to Theorem 1, $\sum_{i=1}^{k+1} \lambda_i \leq 0$ implies that the number of the connected components of graph $\mathbf{W}$ is larger than k and $\sum_{i=1}^{k} \lambda_i > 0$ implies that the number of components is smaller than k.

IV. Multiview Consensus Clustering

According to Theorem 1, clustering results can be obtained from the graph itself directly without having to perform k-means clustering or other thresholding algorithms. In an ideal case, $\mathbf{H}\mathbf{H}^\top$ is a strictly block diagonal matrix. We learn a consensus graph $\mathbf{S}$ that best approximates $\mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top$ from different views, so we define the following cost function as a disagreement measure between each individual view and the global view,

$$m^{(v)} = \|\mathbf{S} - \beta \mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top\|_{\text{F}}^2, \tag{6}$$

where β is a scaling constant.

We suppose that $\mathbf{S}$ can be denoted by $\mathbf{S} = \mathbf{H}^{(*)}(\mathbf{H}^{(*)})^\top$. We obtain $\|\mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top\|_{\text{F}}^2 = \|\mathbf{H}^{(*)}(\mathbf{H}^{(*)})^\top\|_{\text{F}}^2 = k$ due to the constraint $(\mathbf{H}^{(v)})^\top \mathbf{H}^{(v)} = (\mathbf{H}^{(*)})^\top \mathbf{H}^{(*)} = \mathbf{I}$. Substituting it into Eq. (6) and ignoring the constant additive and scaling terms, we have,

$$m^{(v)} = -\text{Tr}\big((\mathbf{H}^{(v)})^\top \mathbf{S} \mathbf{H}^{(v)}\big) = -\langle \mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top, \mathbf{S} \rangle, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product.

Eqs. (6) and (7) are co-regularization terms between single view and global view [6] and these terms regularize diverse views to a common consensus. Co-regularization has *compatibility* and *independence* assumptions for its success [1], [6]. Compatibility means that different views usually together admit the same underlying clustering across multiple views, and independence is that features in different views are conditional independent of each other.

Each view graph $\mathbf{W}^{(v)}$ are learned by Eq. (2). First, using these graphs $\mathbf{W}^{(v)}$ learns the different embedding matrices $\mathbf{H}^{(v)}$. Second, a global affinity graph $\mathbf{S}$ is learned to obtain an ideal graph structure with the constraint $\text{rank}(\mathbf{L}_S) = n - k$. Third, we minimize the disagreement between each $\mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top$ and the global graph $\mathbf{S}$. Then, we optimize the overall objective function,

$$\begin{aligned}\min_{\mathbf{H}^{(v)},\mathbf{S}} \ & \sum_{v=1}^{n_v} \text{Tr}\Big((\mathbf{H}^{(v)})^\top \mathbf{L}^{(v)} \mathbf{H}^{(v)}\Big) \\ & + \sum_{v=1}^{n_v} \|\mathbf{S} - \beta \mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top\|_{\text{F}}^2 \\ \text{s.t. } & \forall v, \quad \mathbf{H}^{(v)} \in \mathbb{R}^{n\times k}, \quad (\mathbf{H}^{(v)})^\top \mathbf{H}^{(v)} = \mathbf{I}, \\ & \mathbf{S} \geq \mathbf{0}, \quad \mathbf{S1} = \mathbf{1}, \quad \text{rank}(\mathbf{L}_S) = n - k, \end{aligned} \tag{8}$$

where $\mathbf{L}_S$ is the normalized Laplacian matrix since $\mathbf{S}$ is constrained by $\mathbf{S1} = \mathbf{1}$.

In the objective function, Eq. (8), the co-regularization term inspired by [6] makes different views to agree with each other and regularizes each view-specific set of kernels $\mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top$ towards a common consensus graph $\mathbf{S}$. Since Eq. (8) has the co-regularization term, the objective function is different from the one in [11]. In [11], they only sum up different graphs linearly but in Eq. (8) the term regularizes diverse views towards a common consensus. We learn a consensus graph with exactly k connected components where k is the number of clusters. The clustering results can be obtained by the learned graph $\mathbf{S}$ directly without further post-processing steps because the rank of the normalized Laplacian matrix $\mathbf{L}_S$ is constrained by $\text{rank}(\mathbf{L}_S) = n - k$. By minimizing Eq. (8), embedding matrices $\mathbf{H}^{(v)}$ and the affinity graph $\mathbf{S}$ are learned simultaneously. We propose a novel algorithm to optimize the objective function Eq. (8) in the following section.

V. Optimization

The objective function Eq. (8) is divided into two subproblems and is alternately solved them effectively.

The first subproblem is to fix $\mathbf{S}$, updating $\mathbf{H}^{(v)}$. Then, Eq. (8) becomes

$$\begin{aligned}\min_{\mathbf{H}^{(v)}} \ & \sum_{v=1}^{n_v} \text{Tr}\Big((\mathbf{H}^{(v)})^\top (\mathbf{L}^{(v)} - 2\beta\mathbf{S})\mathbf{H}^{(v)}\Big) \\ \text{s.t. } & \forall v, \quad \mathbf{H}^{(v)} \in \mathbb{R}^{n\times k}, \quad (\mathbf{H}^{(v)})^\top \mathbf{H}^{(v)} = \mathbf{I}. \end{aligned} \tag{9}$$

Here the hyper parameter β trades off SC objective and spectral embedding agreement.

Note that the problem Eq. (9) is independent between different v, then, we have,

$$\begin{aligned}\max_{\mathbf{H}^{(v)}} \ & \text{Tr}\Big((\mathbf{H}^{(v)})^\top (\mathbf{W}^{(v)} + 2\beta\mathbf{S})\mathbf{H}^{(v)}\Big) \\ \text{s.t. } & \mathbf{H}^{(v)} \in \mathbb{R}^{n\times k}, \quad (\mathbf{H}^{(v)})^\top \mathbf{H}^{(v)} = \mathbf{I}. \end{aligned} \tag{10}$$

The optimal $\mathbf{H}^{(v)}$ of Eq. (10) is formed by the k eigenvectors corresponding to the top k largest eigenvalues of the matrix $[\mathbf{W}^{(v)} + 2\beta\mathbf{S}]$. Since each $\mathbf{W}^{(v)}$ is fixed, each embedding matrix $\mathbf{H}^{(v)}$ is regularized by the global graph $\mathbf{S}$.

The second subproblem is to fix $\mathbf{H}^{(v)}$, updating $\mathbf{S}$. Then, Eq. (8) becomes

$$\begin{aligned}\min_{\mathbf{S}} \ & \|\mathbf{S}\|_{\text{F}}^2 - 2\beta\Big\langle \sum_{v=1}^{n_v} \mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top, \mathbf{S}\Big\rangle \\ \text{s.t. } & \mathbf{S} \geq \mathbf{0}, \quad \mathbf{S1} = \mathbf{1}, \quad \text{rank}(\mathbf{L}_S) = n - k. \end{aligned} \tag{11}$$

Denoting $\sum_{v=1}^{n_v} 2\beta\mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top$ by a matrix $\mathbf{Q}$, we have

$$\begin{aligned}\min_{\mathbf{S}} \ & \|\mathbf{S}\|_{\text{F}}^2 - \text{Tr}(\mathbf{Q}\mathbf{S}^\top) \\ \text{s.t. } & \mathbf{S} \geq \mathbf{0}, \quad \mathbf{S1} = \mathbf{1}, \quad \text{rank}(\mathbf{L}_S) = n - k. \end{aligned} \tag{12}$$

The rank constraint, $\text{rank}(\mathbf{L}_S) = n - k$, can be achieved by $\sum_{i=1}^{k} \lambda_i = 0$. According to Theorem 2, $\sum_{i=1}^{k} \lambda_i = 0$ is equalize to minimizing $\text{Tr}(\mathbf{H}^\top \mathbf{L}_S \mathbf{H})$ subject to $\mathbf{H} \in \mathbb{R}^{n\times k}$ and $\mathbf{H}^\top\mathbf{H} = \mathbf{I}$. Then, referring to Eq. (12), we fix $\mathbf{H}$ and update $\mathbf{S}$ by,

$$\begin{aligned}\min_{\mathbf{S}} \ & \|\mathbf{S}\|_{\text{F}}^2 - \text{Tr}(\mathbf{Q}\mathbf{S}^\top) + \gamma\,\text{Tr}(\mathbf{H}^\top \mathbf{L}_S \mathbf{H}) \\ \text{s.t. } & \mathbf{S} \geq \mathbf{0}, \quad \mathbf{S1} = \mathbf{1}, \end{aligned} \tag{13}$$

where γ is a trade-off parameter.

Note that the problem Eq. (13) is independent between different j, then we have,

$$\begin{aligned}\min_{\boldsymbol{s}_j} \ & \sum_{i=1}^{n} \Big(\gamma \|\boldsymbol{h}_i - \boldsymbol{h}_j\|_2^2 - q_{ij}\Big) s_{ij} + \boldsymbol{s}_j^\top \boldsymbol{s}_j \\ \text{s.t. } & \boldsymbol{s}_j \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{s}_j = 1, \end{aligned} \tag{14}$$

where $\boldsymbol{s}_j$ denotes a column of $\mathbf{S}$.

Denoting $\boldsymbol{p}_j$ as a vector with the i-th element equal to $p_{ij} = \gamma \|\boldsymbol{h}_i - \boldsymbol{h}_j\|_2^2 - q_{ij}$, then solving Eq. (14) is equal to optimizing the following objective function,

$$\begin{aligned}\min_{\boldsymbol{s}_j} \ & \frac{1}{2}\left\|\boldsymbol{s}_j + \frac{\boldsymbol{p}_j}{2}\right\|_2^2 \\ \text{s.t. } & \boldsymbol{s}_j \geq \mathbf{0}, \quad \mathbf{1}^\top \boldsymbol{s}_j = 1. \end{aligned} \tag{15}$$

Eq. (15) is a Euclidean projection problem on the simplex space. The Lagrangian function of Eq. (15) is

$$\mathcal{L}(\boldsymbol{s}_j, \eta, \boldsymbol{\rho}) = \frac{1}{2}\left\|\boldsymbol{s}_j + \frac{\boldsymbol{p}_j}{2}\right\|_2^2 - \eta(\mathbf{1}^\top \boldsymbol{s}_j - 1) - \boldsymbol{\rho}^\top \boldsymbol{s}_j \tag{16}$$

where η and $\boldsymbol{\rho}$ are the Lagrangian multipliers.

According to the Karush-Kuhn-Tucker condition [51], it can be verified that the optimal solution $\boldsymbol{s}_j^\star$ is

$$\boldsymbol{s}_j^\star = \left(-\frac{\boldsymbol{p}_j}{2} + \eta\mathbf{1}\right)_+. \tag{17}$$

Algorithm 1 Multiview Consensus Clustering Algorithm

input : Dataset $\mathcal{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n_v)}\}$, the cluster number k, and parameter β.
output : $\mathbf{S}$ with exactly k connected components.
initialize: $\mathbf{W}^{(v)}$ is initialized by Eq. (2) where we replace $\mathbf{H}^{(v)}$ by $(\mathbf{X}^{(v)})^\top$ and then both $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$ are optimized by Eq. (2), $\mathbf{S}$ is initialed by $[\sum_{v=1}^{n_v} \mathbf{H}^{(v)}(\mathbf{H}^{(v)})^\top]$, and $\mathbf{H}$ is formed by k number of eigenvectors corresponding to the top k smallest eigenvalues of the Laplacian matrix $\mathbf{L}_S$.

```
while not converge do
    for v ∈ {1, 2, ..., n_v} do
        Update H^(v) by solving Eq. (10), i.e., H^(v) is formed by k eigenvectors with the top k largest eigenvalues of [W^(v) + 2βS];
    end
    repeat
        for j ∈ {1, 2, ..., n} do
            Update s_j by using Eq. (17);
        end
        S = (S + S^⊤)/2;
        Form H by k eigenvectors with the top k smallest eigenvalues of L_S;
    until S has k connected components;
end
```

Referring to [9] and [11], once we learn to obtain the graph $\mathbf{S}$ with optimizing Eq. (13), we need to compute eigenvalues of $\mathbf{L}_S$ in order to calculate $\sum_{i=1}^{k} \lambda_i$.

The detailed algorithm is summarized in Algorithm 1.

According to Theorems 1 and 2, the stopping condition of the second subproblem of the algorithm is $\sum_{i=1}^{k} \lambda_i = 0$ [9] so that it achieves the rank constraint $\text{rank}(\mathbf{L}_S) = n - k$. Following [9], parameters α of Eq. (2) and γ of Eq. (13) are used to achieve the constraint $\text{rank}(\mathbf{L}) = n - k$, so they are set to one initially and tune them according to $\sum_{i=1}^{k} \lambda_i$ during iteration until $\sum_{i=1}^{k} \lambda_i = 0$.

Convergence and Complexity Analysis: Since the second order derivative of Eq. (15) with respect to $\boldsymbol{w}_j$ is equal to $1 \geq 0$, Eq. (15) is a convex problem. Because optimizing $-(\mathbf{H}^{(v)})^\top \mathbf{S}\mathbf{H}^{(v)}$ in Eq. (9) is equal to minimizing $(\mathbf{H}^{(v)})^\top \mathbf{L}_S \mathbf{H}^{(v)}$ and the Laplacian matrix is a positive semi-definite, Eq. (9) is a convex optimization problem. Optimizing $\mathbf{S}$ and $\mathbf{H}^{(v)}$ alternately, both of them decrease monotonically. As a result, the overall objective function value of Eq. (8) decreases monotonically in each iteration until Algorithm 1 converges.

The first step of the objective function Eq. (8) is to solve Eq. (9). It is an eigen-decomposition procedure and the complexity of the generalized eigenvector problem is $O((n+k)n^2)$. In each individual view, we need to solve Eq. (10) and to calculate the largest k eigenvectors of $[\mathbf{W}^{(v)} + 2\beta\mathbf{S}]$, so solving Eq. (9) costs $O(kn_v n^2)$. The second step is a Euclidean projection problem on the simplex space. We need $O(n)$ time to compute $\boldsymbol{s}_j$, and $O(t_1 n)$ to solve Eq. (15) where t_1 is the iteration number. n times are needed to calculate each $\boldsymbol{s}_j, \forall j$, so the complexity of the first step of Eq. (11) is $O((t_1+1)n^2)$. Thus, the total complexity of Eq. (8) is

$$O(((t_1 + 1 + kn_v)n^2)t_o), \tag{18}$$

where t_o is the number of iterations of the two steps.

Fig. 1. Example images of different datasets.

VI. Experimental Results

A. Datasets

Three benchmark datasets are used to demonstrate the effectiveness of the proposed method, including

COIL-20 dataset is from the Columbia object image library [54]. COIL-20 has 1440 images of 20 classes and each class contains 72 images. The first view is the 1024-D intensity feature, the second view is the 3304-D LBP feature, and the third view is the 6750-D Gabor feature [37].

UCI digits dataset has 10 classes digits, each class has 200 different handwritten digits, and there are 2000 data points. The first view is the 216-D profile-correlation feature, the second is the 76-D Fourier-coefficient feature, the third is the 64-D Karhunen-Loeve-coefficient feature, the fourth is the 240-D intensity-averaged feature in 2×3 windows, the fifth is the 47-D Zernike moment feature, and the sixth is the 6-D morphological feature.

MSRC-v1 dataset contains 240 images in eight classes and each class has 30 images. Following [10], we select seven classes: tree, building, airplane, cow, face, car, and bicycle. The first view is the 1302-D CENTRIST feature, the second is the 48-D color moment, the third is the 512-D GIST feature, the fourth is the 100-D HOG feature, the fifth view is the 256-D LBP feature, and the sixth view is 200-D SIFT feature.

A small fraction of samples in these image datasets are shown in Fig. 1.

B. Experimental Setup

We denote the proposed multiview consensus graph clustering as MCGC for short. We compare MCGC with seven baselines:

1) SC-Best: The standard spectral clustering (SC) [3] is performed on every single view and we report the best single view result. The graph is constructed by following [55].

TABLE I
CLUSTERING PERFORMANCE

Methods	ACC	NMI	Purity	Precision	Recall	F-score	ARI
COIL-20							
SC-best	73.35±1.08	82.69±0.85	75.96±1.24	66.99±1.40	71.64±1.31	69.23±1.35	67.58±1.42
CAN-best	91.46±0.00	94.79±0.00	91.81±0.00	80.30±0.00	96.32±0.00	87.58±0.00	86.88±0.00
CRSC	75.10±1.15	84.06±0.53	76.27±1.06	70.98±1.05	72.91±1.19	71.93±1.10	70.45±1.15
MKKM	77.64±0.00	84.37±0.00	77.71±0.00	73.88±0.00	75.19±0.00	74.53±0.00	73.20±0.00
AMGL	74.89±2.04	84.51±1.02	77.69±1.96	67.14±2.89	76.00±1.46	71.26±1.72	69.67±1.85
MLAN	84.44±0.00	92.44±0.00	87.92±0.00	72.82±0.00	92.82±0.00	81.61±0.00	80.54±0.00
MVGL	92.50±0.00	97.69±0.00	95.00±0.00	90.58±0.00	97.46±0.00	93.89±0.00	93.57±0.00
MCGC	**99.51±0.00**	**99.45±0.00**	**99.51±0.00**	**99.02±0.00**	**99.11±0.00**	**99.06±0.00**	**99.01±0.00**
UCI digits							
SC-best	89.40±0.04	80.00±0.12	89.40±0.04	79.95±0.07	80.57±0.07	80.26±0.07	78.07±0.08
CAN-best	96.30±0.00	91.64±0.00	96.30±0.00	92.65±0.00	92.79±0.00	92.72±0.00	91.92±0.00
CRSC	91.46±0.04	83.99±0.05	91.46±0.04	83.40±0.06	84.23±0.07	83.81±0.07	82.01±0.07
MKKM	89.45±0.00	81.74±0.00	89.45±0.00	80.64±0.00	81.19±0.00	80.92±0.00	78.80±0.00
AMGL	86.93±0.93	87.02±0.99	86.93±0.93	83.23±1.50	84.21±1.48	83.72±1.49	81.91±1.66
MLAN	97.30±0.00	93.90±0.00	97.30±0.00	87.25±0.00	46.15±0.00	60.37±0.00	45.96±0.00
MVGL	94.20±0.00	89.05±0.00	94.20±0.00	87.57±0.00	89.15±0.00	88.35±0.00	94.20±0.00
MCGC	**97.55 ± 0.00**	**94.22 ± 0.00**	**97.55 ± 0.00**	**95.14 ± 0.00**	**95.20 ± 0.00**	**95.17 ± 0.00**	**94.64 ± 0.00**
MSRC-v1							
SC-best	73.81±0.00	61.79±0.00	73.81±0.00	58.45±0.00	61.44±0.00	59.91±0.00	53.26±0.00
CAN-best	71.90±0.00	61.96±0.00	71.90±0.00	51.52±0.00	59.70±0.00	55.31±0.00	47.49±0.00
CRSC	89.52±0.00	80.74±0.00	89.52±0.00	80.15±0.00	82.36±0.00	81.24±0.00	78.17±0.00
MKKM	73.81±0.00	64.45±0.00	75.71±0.00	61.35±0.00	63.28±0.00	62.30±0.00	56.12±0.00
AMGL	90.48±0.00	82.90±0.00	90.48±0.00	79.80±0.00	82.79±0.00	81.27±0.00	78.19±0.00
MLAN	75.24±0.00	76.19±0.00	81.43±0.00	66.32±0.00	78.56±0.00	71.93±0.00	66.94±0.00
MVGL	91.43±0.00	84.44±0.00	91.43±0.00	82.87±0.00	84.33±0.00	83.59±0.00	80.92±0.00
MCGC	**92.38±0.00**	**84.70±0.00**	**92.38±0.00**	**84.74±0.00**	**85.71±0.00**	**85.22±0.00**	**82.83±0.00**

Note: The best results are highlighted in bold.

2) CAN-Best: The clustering with adaptive neighbors (CAN) is performed on single view and we report the best result. We optimize Eq. (2) instead of the standard CAN [9].

3) CRSC: Co-regularized multi-view spectral clustering (CRSC) [6] uses a co-regularization term for maximizing the agreement between the different individual view and the global view.

4) MKKM: Multiple kernels k-means (MKKM) [35] clustering reduces the redundancy and enhances the diversity of the multiple kernels through a matrix-induced regularization.

5) AMGL: Auto-weighted multiple graph learning (AMGL) [56] learns the view weight adaptively and sums up different Laplacian matrices from different views to calculate the embedding matrix in subspace.

6) MLAN: Multiview learning with adaptive neighbors (MLAN) [10] extends CAN to the multiview setting.

7) MVGL: Multiview clustering with graph learning (MVGL) [11] learns individual graph and then integrate the learned multiple graphs into a global graph with exactly k components.

The default parameters of each compared method are adopted in our experiments. Without loss of generality, for all these methods, we run each method 10 times and report the mean of performance as well as the standard deviation in Table I. SC, CRSC, MKKM, and AMGL require k-means clustering after they obtain the embedding data representation. k-means clustering is sensitive to initial values, so we run k-means clustering processing 30 times and report the result with the minimum value for the objective function of k-means clustering among results of these 30 times. CAN, MLAN, MVGL, and MCGC obtain the clustering indicators using the learned graph directly. Since each connected component belongs to one cluster, the clustering labels are obtained directly by the learned global graph $\mathbf{S}$ according to Tarjan's [57] strongly connected component algorithm.

C. Evaluation Metrics

Seven metrics are used to evaluate the performance: clustering accuracy (ACC), normalized mutual information (NMI), Purity, Precision, Recall, F-score, and adjusted rand index (ARI). For these widely used metrics, the larger value indicates the better clustering performance. These metrics are calculated by comparing the obtained label of each sample with the ground-truth labels provided in datasets.

ACC measures clustering accuracy and is defined by

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(\tau_i, \mathrm{map}(r_i))}{n},$$

where n data points are belonging to k clusters, τ_i denotes the ground-truth label of the i-th sample, r_i denotes the corresponding learned clustering label, and $\delta(\cdot,\cdot)$ denotes the Dirac delta function

$$\delta(x, y) = \begin{cases} 1, & \text{if } x = y; \\ 0, & \text{otherwise,} \end{cases}$$

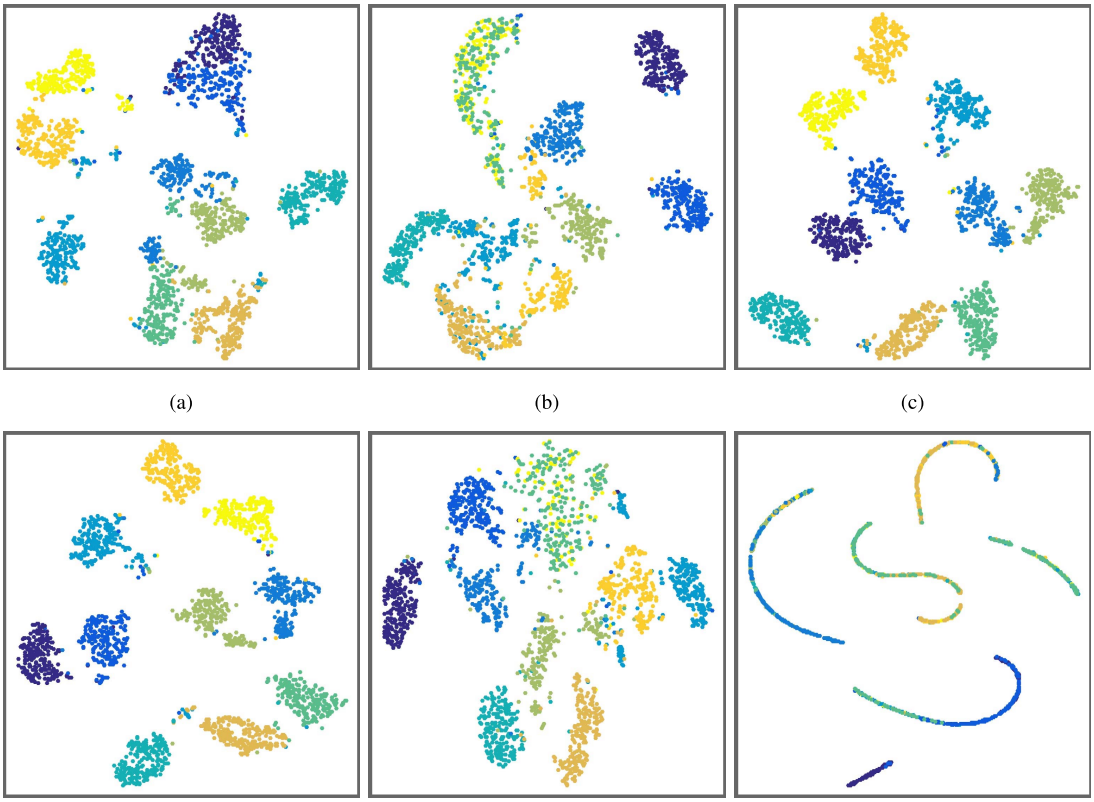
and $\mathrm{map}(r_i)$ is the optimal mapping function that permutes the obtained labels to match the ground-truth labels. The best mapping is found by the Kuhn-Munkres algorithm [58].

NMI measures the similarity between τ_i and r_i and is defined by

$$\mathrm{NMI}(\tau_i, r_i) = \frac{I(\tau_i, r_i)}{\sqrt{E(\tau_i)E(r_i)}},$$

(a) (b) (c)

Fig. 2. Visualization of the clustering results of COIL-20 with t-SNE in different views. (a) View 1. (b) View 2. (c) View 3.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Visualization of the clustering results of UCI digits with t-SNE in different views. (a) View 1. (b) View 2. (c) View 3. (d) View 4. (e) View 5. (f) View 6.

where $I(\tau_i, r_i)$ is mutual information between τ_i and r_i and $E(\cdot)$ returns the information entropy.

Let n_i^r be the item number in the i-th cluster ($1 \le i \le k$) obtained by using the clustering algorithms and n_i^τ be the number of the i-th cluster in the ground-truth label. Then, NMI is given by

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} \log\left(\frac{n \cdot n_{ij}}{n_i^r n_j^\tau}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i^r \log\frac{n_i^r}{n}\right)\left(\sum_{i=1}^{k} n_i^\tau \log\frac{n_i^\tau}{n}\right)}},$$

where n_{ij} is the item number which is in the intersection between τ_i and r_i.

Purity is the percentage of correct labels and is defined by

$$\mathrm{Purity} = \frac{1}{n}\sum_{i=1}^{k} \max_{1 \le j \le k} |\mathrm{map}(r_i) \cap \tau_j|.$$

Precision and Recall are defined by

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{FN}},$$

where TP, FP, and FN denote the number of items correctly labeled as belonging to the positive cluster, wrongly labeled as belonging to positive cluster, and wrongly labeled as belonging to negative cluster, respectively.

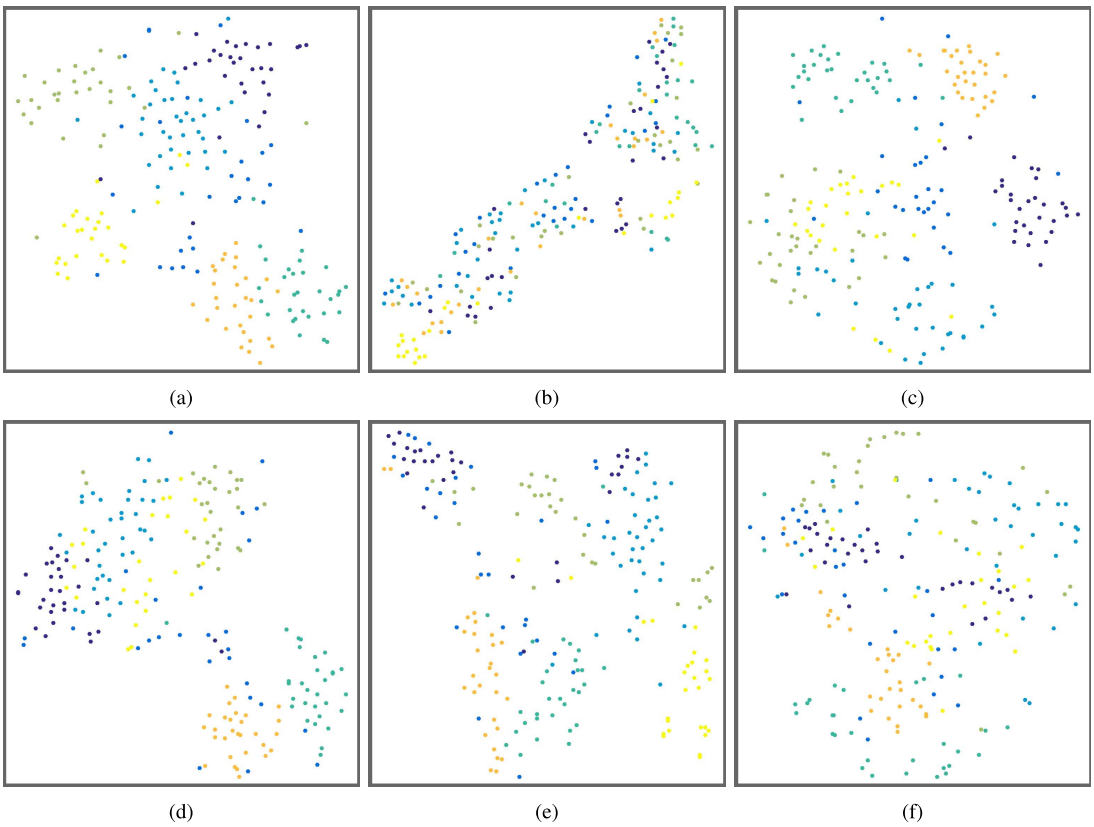


Fig. 4. Visualization of the clustering results of MSRC-v1 with t-SNE in different views. (a) View 1. (b) View 2. (c) View 3. (d) View 4. (e) View 5. (f) View 6.

F-score is then defined by calculating the harmonic mean of Precision and Recall,

$$F\text{-score} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

ARI is defined by

$$\text{ARI} = \frac{\sum_{i,j=1}^{k} C_{n_{ij}}^2 - \frac{\sum_{i=1}^{k} C_{n_i^\tau}^2 \sum_{i=1}^{k} C_{n_i^r}^2}{C_n^2}}{\frac{1}{2}\left(\sum_{i=1}^{k} C_{n_i^\tau}^2 + \sum_{i=1}^{k} C_{n_i^r}^2\right) - \frac{\sum_{i=1}^{k} C_{n_i^\tau}^2 \sum_{i=1}^{k} C_{n_i^r}^2}{C_n^2}},$$

where combination operation C_n^m is defined as a selection of m items from a collection n.

D. Performance Evaluation

The clustering performance is listed in Table I. It can be seen from Table I that all multiview learning methods obtain better performance than SC-best, the proposed MCGC obtains better results than CAN-best, and MCGC achieves better performance than other state-of-the-art methods in almost all experiments. Since CRSC [6] minimizes the diversity between different views, it obtains a better results than SC-best. Performance of CAN-based methods is higher than SC-based methods because CAN-based methods learn to obtain a better structured graph. However, performance of MLAN [10] and MVGL [11] is lower than the proposed MCGC since they fuse information from different views linearly. Besides learning a better structured graph, MCGC uses the disagreement cost to minimize the diversity between different views, so it obtains a better results than other CAN-based methods.

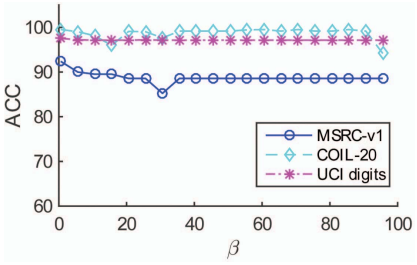


Fig. 5. The performance of MCGC is stable with respect to β.

To be more intuitive, we visualize the data points and the clustering results with t-distributed stochastic neighbor embedding (t-SNE) [59] in different views as shown in Figs. 2, 3, and 4. In Figs. 2, 3, and 4, each data point is visualized by t-SNE and different colors denote different cluster labels obtained by MCGC. Since data points in different views reflect different characteristics and data points of some clusters are mixed with each other in the individual view as shown in Figs. 2, 3, and 4, it is clearly difficult to obtain an ideal clustering result with individual view feature. It can be seen from Table I that MCGC obtains good clustering results due to minimizing disagreement of diverse views and constraining the rank of the Laplacian matrix.

E. Parameter Sensitivity

There is one parameter β in the objective function Eq. (8) of MCGC. Fig. 5 shows accuracy varies with β on three datasets. It can be seen from Fig. 5 that the performance is stable when

its value varies in a range of [0.6, 100] with 5 intervals. In this paper, we use $\beta = 0.6$ for all the datasets.

VII. Conclusions

We presented a novel multiview consensus clustering method based on two motivations: minimizing disagreement between diverse views and constraining the rank of the Laplacian matrix. We designed a cost function for minimizing disagreement to make the graph structure in diverse views agree with each other and a rank constraint was imposed in the objective for learning a global graph with exactly k connected components. We learned the graph structure and the embedding matrix, simultaneously. Different from existing graph-based methods, MCGC obtained the cluster assignment directly from the graph itself without any post-processing steps such as thresholding on the embedding matrix or k-means clustering algorithms. The efficient optimization algorithm was presented after plentiful analysis. Experiments on three benchmarks had demonstrated the superiority of MCGC.

Since MCGC needs to calculate the eigenvalues during iteration to achieve the constraint $\text{rank}(\mathbf{L}_S) = n - k$, CAN-based algorithms spent more time than SC-based. The multiplicity of 0 as an eigenvalue of $\mathbf{L}_S$ is equal to the number of connected components, so we will consider the use of Tarjan's [57] strongly connected component algorithm to achieve the constraint in future.

References

[1] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. COLT*, 1998, pp. 92–100.

[2] 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.

[3] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NIPS*, 2002, pp. 849–856.

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

[5] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proc. ICML*, 2011, pp. 393–400.

[6] A. Kumar, P. Rai, and H. Daumé, "Co-regularized multi-view spectral clustering," in *Proc. NIPS*, 2011, pp. 1413–1421.

[7] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proc. AAAI*, 2014, pp. 2149–2155.

[8] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. AAAI*, vol. 29, 2015, pp. 2750–2756.

[9] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. KDD*, 2014, pp. 977–986.

[10] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. AAAI*, 2017, pp. 2408–2414.

[11] K. Zhan, C. Zhang, J. Guan, and J. Wang, "Graph learning for multiview clustering," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2887–2895, Oct. 2018.

[12] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert–Schmidt norms," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2005, pp. 63–77.

[13] L. Song, A. Smola, A. Gretton, and K. M. Borgwardt, "A dependence maximization view of clustering," in *Proc. ICML*, vol. 24, 2007, pp. 815–822.

[14] H. Wang, C. Weng, and J. Yuan, "Multi-feature spectral clustering with minimax optimization," in *Proc. CVPR*, 2014, pp. 4106–4113.

[15] X. Cao, C. Zhang, C. Zhou, H. Fu, and H. Foroosh, "Constrained multi-view video face clustering," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4381–4393, Nov. 2015.

[16] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. CVPR*, 2015, pp. 586–594.

[17] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao, "Flexible multi-view dimensionality co-reduction," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 648–659, Feb. 2017.

[18] K. Zhan, J. Shi, J. Wang, H. Wang, and Y. Xie, "Adaptive structure concept factorization for multiview clustering," *Neural Comput.*, vol. 30, no. 4, pp. 1080–1103, 2018.

[19] J. Wang, F. Tian, H. Yu, C. H. Liu, K. Zhan, and X. Wang, "Diverse non-negative matrix factorization for multiview data representation," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2620–2632, Sep. 2018.

[20] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *Proc. COLT*, 2007, pp. 82–96.

[21] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. ICML*, vol. 26, 2009, pp. 129–136.

[22] C. Deng, R. Ji, D. Tao, X. Gao, and X. Li, "Weakly supervised multi-graph learning for robust image reranking," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 785–795, Apr. 2014.

[23] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Oakland, CA, USA, vol. 1, no. 14, 1967, pp. 281–297.

[24] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[25] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[26] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[27] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Jul. 2011.

[28] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: Spectral clustering and normalized cuts," in *Proc. KDD*, vol. 10, 2004, pp. 551–556.

[29] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proc. ICDM*, 2005, pp. 606–610.

[30] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[31] X. Cai, F. Nie, and H. Huang, "Multi-view K-means clustering on big data," in *Proc. IJCAI*, vol. 23, 2013, pp. 2598–2604.

[32] J. Xu, J. Han, and F. Nie, "Discriminatively embedded K-means for multi-view clustering," in *Proc. CVPR*, 2016, pp. 5356–5364.

[33] J. Xu, J. Han, F. Nie, and X. Li, "Re-weighted discriminatively embedded K-means for multi-view clustering," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 3016–3027, Jun. 2017.

[34] L. Du *et al.*, "Robust multiple kernel K-means using $\ell_{2,1}$-norm," in *Proc. IJCAI*, 2015, pp. 3476–3482.

[35] X. Liu, Y. Dou, J. Yin, L. Wang, and E. Zhu, "Multiple kernel k-means clustering with matrix-induced regularization," in *Proc. AAAI*, 2016, pp. 1888–1894.

[36] Y. Wang, X. Liu, Y. Dou, and R. Li, "Approximate large-scale multiple kernel k-means using deep neural network," in *Proc. IJCAI*, vol. 26, 2017, pp. 3006–3012.

[37] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. ICCV*, 2015, pp. 1582–1590.

[38] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *Proc. ICCV*, 2015, pp. 4238–4246.

[39] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Latent multi-view subspace clustering," in *Proc. CVPR*, vol. 30, 2017, pp. 4333–4341.

[40] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Proc. AAAI*, vol. 32, 2018, pp. 3730–3737.

[41] X. Dong, Y. Yan, M. Tan, Y. Yang, and I. W. Tsang, "Late fusion via subspace search with consistency preservation," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 518–528, Jan. 2019.

[42] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, Jul. 2006.

[43] C. Cortes, M. Mohri, and A. Rostamizadeh, "Learning non-linear combinations of kernels," in *Proc. NIPS*, 2009, pp. 396–404.

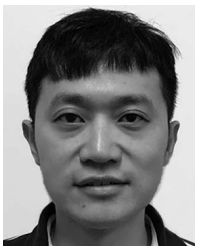
[44] M. Kloft and G. Blanchard, "The local Rademacher complexity of ℓ_p-norm multiple kernel learning," in *Proc. NIPS*, 2011, pp. 2438–2446.

[45] J. He, C. Du, C. Du, F. Zhuang, Q. He, and G. Long, "Nonlinear maximum margin multi-view learning with adaptive kernel," in *Proc. IJCAI*, vol. 28, 2017, pp. 1830–1836.
[46] X. Liu, S. Ji, W. Glänzel, and B. De Moor, "Multiview partitioning via tensor methods," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 5, pp. 1056–1069, May 2013.
[47] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3111–3124, Nov. 2015.
[48] X. Cao, X. Wei, Y. Han, and D. Lin, "Robust face clustering via tensor decomposition," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2546–2557, Nov. 2015.
[49] B. Mohar, Y. Alavi, G. Chartrand, and O. Oellermann, "The Laplacian spectrum of graphs," *Graph Theory, Combinat., Appl.*, vol. 2, no. 12, pp. 871–898, 1991.
[50] K. Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations I," *Proc. Nat. Acad. Sci. USA*, vol. 35, no. 11, pp. 652–655, 1949.
[51] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[52] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. ICML*, vol. 25, 2008, pp. 272–279.
[53] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Math. Program.*, vol. 158, no. 1, pp. 575–585, 2016.
[54] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.
[55] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. NIPS*, 2005, pp. 1601–1608.
[56] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proc. IJCAI*, 2016, pp. 1881–1887.
[57] R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM J. Comput.*, vol. 1, no. 2, pp. 146–160, 1972.
[58] L. Lovász and M. D. Plummer, *Matching Theory*. Amsterdam, The Netherlands: Elsevier, 1986.
[59] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Kun Zhan received the B.S. and Ph.D. degrees from the School of Information Science and Engineering, Lanzhou University, China, in 2005 and 2010, respectively. He was a Visiting Student with the Department of Electrical and Computer Engineering, Dalhousie University, Halifax, Canada, from 2009 to 2010. He is currently an Associate Professor with Lanzhou University and a Visiting Scholar with the University of Technology Sydney. His main research interests include machine learning and data science.

Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, China, in 2009. He is currently a Professor with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, Shaanxi, China. His research interests are machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Jing Wang received the B.Eng. degree in electronics and information technology from the Anhui University of Technology, Ma'anshan, China, in 2010, the M.Sc. degree in multimedia information technology from the City University of Hong Kong, Hong Kong, in 2012, and the Ph.D. degree with the Faculty of Science and Technology, Bournemouth University, Poole, U.K., 2018. She is currently a Post-Doctoral Fellow with the Department of Creative Informatics, The University of Tokyo. Her current research interests include machine learning, computer vision, and data mining.



Yi Yang received the Ph.D. degree in computer science from Zhejiang University. He was a Post-Doctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, from 2011 to 2013. He is currently a Professor with the Centre for Artificial Intelligence, School of Software, Faculty of Engineering and Information Technology, University of Technology Sydney. His research interests include multimedia, computer vision, and data science.