Pattern Recognition 88 (2019) 50–63

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-view Subspace Clustering with Intactness-Aware Similarity

Xiaobo Wang[a], Zhen Lei[b,*], Xiaojie Guo[c], Changqing Zhang[c], Hailin Shi[a], Stan Z. Li[b,d]

[a] JD AI Research, China
[b] CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences, China
[c] Tianjin University, China
[d] Faculty of Information Technology, Macau University of Science and Technology, China

ARTICLE INFO

Article history:
Received 14 May 2017
Revised 12 August 2018
Accepted 10 September 2018
Available online 11 September 2018

Keywords:
Intact space
Intactness-aware similarity
Multi-view subspace clustering

ABSTRACT

Multi-view subspace clustering, which aims to partition a set of multi-source data into their underlying groups, has recently attracted intensive attention from the communities of pattern recognition and data mining. This paper proposes a novel multi-view subspace clustering model that attempts to form an informative intactness-aware similarity based on the intact space learning technique. More specifically, we learn an intact space by integrating encoded complementary information. An informative similarity matrix is simultaneously constructed, which enforces the constructed similarity to have maximum dependence with its latent intact points by adopting the Hilbert–Schmidt Independence Criterion (HSIC). A new explanation on the advantages of such intactness-aware similarity has been provided (*i.e.*, the similarity is learned according to the local connectivity). To effectively and efficiently seek the optimal solution of the associated problem, a new ADMM based algorithm is designed. Moreover, to show the merit of the proposed joint optimization, we also conduct the clustering in two separated steps. Extensive experimental results on six benchmark datasets are provided to reveal the effectiveness of the proposed algorithm and its superior performance over other state-of-the-art alternatives.

© 2018 Published by Elsevier Ltd.

1. Introduction

Clustering data points into different groups such that the objects in the same group are highly similar to each other is one of the most fundamental topics in data mining and pattern recognition [1–6]. In the last two decades, a number of clustering approaches have been developed, such as hierarchical clustering based methods [7,8], k-means based methods [9,10], the iteration based methods [11–13], the collaboration based methods [14–16], the factorization based methods [17–21], spectral-based clustering [22–26], and of which, the spectral-based clustering method is arguably the most popular one.

Spectral-based clustering mainly contains two steps, *i.e.* firstly constructing the similarity matrix and then performing spectral clustering on it. Arguably, the first step to learn the similarity matrix is of utmost importance, as the success of spectral clustering algorithm is largely dependent on constructing an informative similarity matrix. To learn such a good similarity matrix, Ng et al. [27] try to learn it based on the data locality. Nie et al. [28] develop a more sophisticated method to build the similarity matrix according to the distance of raw data points. For simplicity, Euclidean distance is used. To learn the similarity matrix with structure priors, SSC [29] and LRR [30] try to find a sparse and a low-rank self-representation, and directly use it to construct the similarity matrix. Lu et al. [31] propose a least square regression based method to take advantages of data correlation for the similarity matrix construction. To further refine the representation, Feng et al. [32] impose a block-diagonal prior on the self-representation, which makes the learned similarity matrix to be exactly block-diagonal. Guo et al. [22] simultaneously learn the self-representation and the similarity matrix of self-representation into a unified framework, which shows a reasonable improvement on the clustering results. You et al. [33] adopt the orthogonal matching pursuit to make the subspace-preserving representation under broad conditions. Additionally, Li et al. [34] propose a novel structured sparse norm on the learned self-representation, and achieve promising results by integrating the sparse self-representation learning and the spectral clustering into one framework. In fact, most of these methods can achieve promising performance for single view data. However, for multi-view one, the above methods are usually difficult to find good clusters due to the potential presence of view insufficiency or high dimensionality

* Corresponding author at: Center for Biometrics and Security Research & National Laboratory of Pattern Rec, 95 Zhongguancun East Road, Beijing 100190, China.
E-mail addresses: wangxiaobo8@jd.com (X. Wang), zlei@nlpr.ia.ac.cn (Z. Lei), xiaojie.guo@tju.edu.cn (X. Guo), changqingzhang@tju.edu.cn (C. Zhang), shihailin@jd.com (H. Shi), szli@nlpr.ia.ac.cn (S.Z. Li).

https://doi.org/10.1016/j.patcog.2018.09.009
0031-3203/© 2018 Published by Elsevier Ltd.

of data. Generally, they can not be directly applied to multi-view cases. This paper concentrates on multi-view subspace clustering.

In practice, many kinds of real-world data appear in multiple views. For instance, web pages contain texts, hyperlinks and possibly existing visual information. As another example, images and videos are often described by different kinds of features, such as color, texture and edge. In general, these multi-view representations can seamlessly capture the rich information from multiple data cues as well as the complementary information among different cues, and can be beneficial to clustering task. To incorporate different views, early methods focus on the setting of two views. De et al. [35] utilize a bipartite similarity matrix to connect two types of features and adopt the standard spectral clustering to obtain the final results. Bickel et al. [10] extend k-means to handle the cases with two conditionally independent views. However, these methods depend on the assumption that there are only two views, and can not handle the cases of three or more views. To incorporate more views, Cai et al. [36] propose a multi-view spectral clustering model to integrate heterogeneous visual descriptors for image categorizations. Tang et al. [37] fuse the multiple graphs information with linked matrix factorization. The co-regularized multi-view spectral clustering is introduced in [38] to perform clustering on different views simultaneously with a co-regularization constraint. Collins et al. [39] learn a common representation under the spectral clustering framework by combining Laplacians of different views. Gao et al. [40] unify the representation learning and spectral clustering into one framework. To capture the high-order cross information among multiple views, Zhang et al. [24] propose a multiple features clustering model with low-rank tensor constraint. Although various existing methods indeed improve the spectral clustering performance for multi-view data, they mainly suffer from constructing an invalid similarity graph due to the insufficient information between different views or without considering the high-order dependence between the intact information and the constructed similarity.

In this paper, we propose a novel multi-view subspace clustering model, termed as Multi-view Subspace Clustering with Intactness-Aware Similarity (MSC_IAS), which intends to construct an *intactness-aware* similarity matrix under the assumption that the similarity should have maximum dependence with the corresponding points in the intact space. Specifically, to avoid the common issue in most of existing multi-view subspace clustering methods, *i.e.* the information loss from insufficient views, we try to recover an intact space from multi-view data. As indicated by the work [41], "intact" means *complete* and *not damaged* in Merriam-Webster, which are the favorable properties for similarity construction. Meanwhile, rather than directly using k-means or standard spectral clustering on the latent intact space, we adopt the Hilbert–Schmidt Independence Criterion (HSIC) to guide the intactness-aware similarity matrix building. More concretely, the contributions of this paper can be summarized as follows:

- We propose a novel multi-view subspace clustering model, namely Multi-view Subspace Clustering with Intactness-Aware Similarity (MSC_IAS), that constructs the similarity based on the intact space learning technique, and unifies them into one framework. Moreover, two separated steps are conducted to validate the superiority of such joint optimization.
- We construct the similarity based on the assumption that the constructed similarity has maximum dependence with its corresponding intact space, which can be measured by the Hilbert–Schmidt Independence Criterion (HSIC). Moreover, a new explanation (local connectivity) on the similarity has been provided, that is, the learned intactness-aware similarity has a larger value if their data points in the intact space have a small ℓ_1 distance.
- We demonstrate the efficacy and the superior performance of our proposed framework over the state-of-the-art alternatives by conducting experimental results on six benchmark datasets.

The rest of this paper is organized as follows. In Section 2, we briefly review the background of multi-view subspace clustering. Then we revisit the preliminary knowledge in Section 3. In Section 4, we introduce the proposed Multi-view Subspace Clustering with Intactness-Aware Similarity (MSC_IAS) algorithm. Section 5 provides a new ADMM based solution. Experiments and analysis on the benchmark datasets are conducted in Section 6. Finally, Section 7 gives the conclusion to this paper.

2. Related work

In this section, we briefly introduce the background of multi-view subspace clustering. To perform clustering by integrating multi-view information, according to the works [42–44], there exist three distinctions, *i.e.* late integration, intermediate integration and early integration.

1) *Late integration applies a clustering algorithm to each individual view and subsequently combines the results.* Bruno and Marchand-Maillet [45] treat the optimal clustering as hidden factors to produce the clustering of different views. Greene and Cunningham [46] concatenate the clustering results of different views into one matrix, and then perform non-negative matrix factorization to obtain the final results. Xia et al. [47] first learn from each view and then recover a shared low-rank transition probability matrix as a crucial input to the standard Markov chain method for clustering. In the work [48], authors use mapping functions to make clusters of different views comparable and learn the best clusters. Besides these, the fuzzy clustering approaches [15,49] that generalize the three fusion strategies have also been developed.
2) *Intermediate integration computes separate similarity matrices on different views and produces a fused pairwise representation which is then passed to a clustering algorithm.* In this regards, most of spectral-based methods [24–26,37,38,40,42,50–52] belong to this category. For example, Kumar et al. [38] co-regularize the clustering hypotheses to exploit the complementary information between different views. Tang et al. [37] fuse the multiple graphs information with linked matrix factorization. Li et al. [50] introduce a subspace technique to address the transformation invariant issue. Wang et al. [52] propose an angular based regularization to coordinate all views to yield a correlations consensus. Cao et al. [51] exploit the complementary information between different views to form an informative representation. Zhang et al. [24] consider the self-representation of each view as a tensor equipped with low-rank constraints, and directly construct the similarity matrix on the self-representation.
3) *Early integration involves the direct combination of data from views into a single view representation before data clustering.* Guo et al. [53] formulate the subspace learning with multiple views as a joint optimization problem with a common subspace representation matrix and a group sparsity inducing norm. White et al. [54] explicitly learn a common representation based on multiple views as a joint optimization problem with a common subspace representation matrix. Lu et al. [3] try to find a low-dimensional embedding of data by

computing the eigenvectors of the normalized Laplacian matrix. Zhang et al. [25] seek the underlying latent representation and simultaneously perform data reconstruction based on the learned latent representation.

However, despite the promising results have achieved by the above methods, they mainly suffer from two shortcomings, one is that making the similarity consensus may greatly degrade the performance due to the fact that although the block structures in different similarity matrix are similar, their values can be dramatically different. In other words, the assumption of similarity consensus may not hold. The other one is that the similarity constructed by the above methods is less informative since most of them ignore the local connectivity. In this paper we report the development of a new method for early data fusion to solve the above shortcomings based on the intact space learning technique [41].

3. Preliminary knowledge

3.1. Notation

Throughout the paper, if not specified, we write scalar as lowercase letter u and vector as bold lowercase letter $\mathbf{u}$. Bold uppercase letter $\mathbf{U}$ stands for a matrix. The trace, transpose and inverse of $\mathbf{U}$ are represented by $\mathrm{tr}(\mathbf{U})$, $\mathbf{U}^T$ and $\mathbf{U}^{-1}$, respectively. Particularly, given a matrix $\mathbf{U}=[\mathbf{U}_{ij}]$, its ith row is denoted as $\mathbf{U}^i$ while its jth column $\mathbf{U}_j$. The ℓ_1-norm and the Frobenius norm of $\mathbf{U}$ is designated as $||\mathbf{U}||_1$, $\|\mathbf{U}\|_F$, respectively. In addition, $\mathbf{1}$ and $\mathbf{I}$ are all-ones matrix and identity matrix with appropriate sizes, respectively.

3.2. Hilbert–Schmidt Independence Criterion

The Hilbert–Schmidt Independence Criterion (HSIC) is proposed in the work [55] to measure the (in)dependence of two random variables $\mathcal{X}$ and $\mathcal{Y}$, and has been widely used in many applications, including feature selection [56], matching [57] and multi-view subspace clustering [51]. To introduce the HSIC, we first revisit the definition of cross-covariance $C_{\mathbf{xy}}$. Let us define mapping $\phi(\mathbf{x})$ from $\mathbf{x}\in\mathcal{X}$ to kernel space $\mathcal{F}$, such that the inner product between vectors in that space is given by a kernel function $k_1(\mathbf{x}_i,\mathbf{x}_j)=\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle$. Let $\mathcal{G}$ be a second kernel space on $\mathcal{Y}$, with a kernel function $k_2(\mathbf{y}_i,\mathbf{y}_j)=\langle\varphi(\mathbf{y}_i),\varphi(\mathbf{y}_j)\rangle$. The cross-covariance between two random variables can be defined as:

$$C_{\mathbf{xy}}=E_{\mathbf{xy}}[(\phi(\mathbf{x})-\mu_{\mathbf{x}})\otimes(\varphi(\mathbf{y})-\mu_{\mathbf{y}})] \tag{1}$$

where $\mu_{\mathbf{x}}=E(\phi(\mathbf{x}))$ and $\mu_{\mathbf{y}}=E(\varphi(\mathbf{y}))$, and $\otimes$ is the tensor product. Then, according to the work [55], we have:

Definition 1. Given two separable reproducing kernel Hilbert spaces $\mathcal{F}$, $\mathcal{G}$ and a joint distribution $p_{\mathbf{xy}}$, we define the HSIC as the Hilbert–Schmidt norm of the associated cross-covariance operator $C_{\mathbf{xy}}$:

$$\mathrm{HSIC}(p_{\mathbf{xy}},\mathcal{F},\mathcal{G}):=||C_{\mathbf{xy}}||^2_{\mathbf{HS}}, \tag{2}$$

where $||\cdot||_{\mathbf{HS}}$ denotes the Hilbert–Schmidt norm of a matrix.

However, the joint distribution $p_{\mathbf{xy}}$ is usually unknown or hard to estimate. For practical uses, HSIC has to be estimated using a finite number of data samples. As a consequence, we have the following empirical definition.

Definition 2. (HSIC) Consider a series of n independent observations drawn from $p_{\mathbf{xy}}$, $\mathcal{Z}:=\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)\}\subseteq\mathcal{X}\times\mathcal{Y}$, an empirical estimator of $\mathrm{HSIC}(\mathcal{Z},\mathcal{F},\mathcal{G})$, is given by:

$$\mathrm{HSIC}(\mathcal{Z},\mathcal{F},\mathcal{G})=(n-1)^{-2}\,\mathrm{tr}(\mathbf{K}_1\mathbf{H}\mathbf{K}_2\mathbf{H}), \tag{3}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are the Gram matrices with $k_{1,ij}=k_1(\mathbf{x}_i,\mathbf{x}_j)$, $k_{2,ij}=k_2(\mathbf{y}_i,\mathbf{y}_j)$. $k_1(\mathbf{x}_i,\mathbf{x}_j)$ and $k_2(\mathbf{y}_i,\mathbf{y}_j)$ are the kernel functions defined in the kernel space $\mathcal{F}$ and $\mathcal{G}$, respectively. $\mathbf{H}=\mathbf{I}-n^{-1}\mathbf{1}\mathbf{1}^T$ centers the Gram matrix to have zero mean.

It is important to note that according to Eq. (3), to maximize the dependency between two random variables $\mathcal{X}$ and $\mathcal{Y}$, the empirical estimate of HSIC, *i.e.*, $\mathrm{tr}(\mathbf{K}_1\mathbf{H}\mathbf{K}_2\mathbf{H})$ should be maximized. For more details about the HSIC, please refer to the papers [55,58].

4. Problem formulation

4.1. Intact space learning

As described in the work [41], in intact space learning, it is practical to assume that each individual view only captures partial information while all the views together possess redundant information about the latent intact representation (please see Fig. 1 for example). Therefore, integrating multi-view information is valuable and necessary. Supposing that $\mathbf{X}_i\in\mathbb{R}^d$ is a sample of the latent intact space and is represented by V-view features $\mathbf{F}_i^v\in\mathbb{R}^{d_v}$, where $v\in\{1,2,\ldots,V\}$, d_v denotes the feature dimension of the vth view. Each view $\mathbf{F}_i^v$ can be obtained from a proper view generation function $\mathbf{W}^v\in\mathbb{R}^{d_v\times d}$, *i.e.* $\mathbf{W}^v\mathbf{X}_i$, which could thus be understood as a particular reflection of the sample. Consequently, the intact space learning is to recover the latent intact space from the following constraint:

$$\min_{\mathbf{X},\mathbf{W}^v}\ \frac{1}{V}\sum_{v=1}^{V}||\mathbf{F}^v-\mathbf{W}^v\mathbf{X}||, \tag{4}$$

where $\mathbf{F}^v=[\mathbf{F}_1^v,\mathbf{F}_2^v,\ldots,\mathbf{F}_n^v]\in\mathbb{R}^{d_v\times n}$ and $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n]\in\mathbb{R}^{d\times n}$, and n is the total amount of data points. In addition, $\|\cdot\|$ represents a certain norm to penalize the error. It is sure that the choice of penalty (*i.e.*, the norm) is important to different tasks. The norm $\|\cdot\|$ on such error term (4) depends upon the prior knowledge about the pattern of noise, corruptions or outliers.[1] Moreover, for the sake of stability, two additional terms to regularize the desired intact space $\mathbf{X}$ and the linear sample matrix $\mathbf{W}^v$ are adopted. Specifically, we adopt the Frobenius norm for the intact space $\mathbf{X}$ and the additional constraints (*i.e.*, $\forall i,\ \ ||\mathbf{W}_i^v||_2\le 1$) for the linear sample matrix $\mathbf{W}^v$. In the sequel, the intact space learning model can be formulated as:

$$\min_{\mathbf{X},\mathbf{W}^v}\ \frac{1}{V}\sum_{v=1}^{V}||\mathbf{F}^v-\mathbf{W}^v\mathbf{X}||+\lambda_1||\mathbf{X}||_F^2,\quad \text{s.t.}\quad \forall i,\ \ ||\mathbf{W}_i^v||_2\le 1 \tag{5}$$

where λ_1 is a non-negative weight to balance the corresponding term.

To recover the intact space, Xu et al. [41] have proved two facts based on information theory. One is that more views will bring in more information with respect to the intact space, we can learn the latent intact space by exploiting the complementarity between multiple views, although each individual view is insufficient. The other one is that we may not obtain all the necessary views to learn the latent intact space, but we can approximately restore it when provided with enough views. These two rules[2] give the theoretical guarantee for successfully recovering the intact space.

Although in the work [41], the authors have analyzed that the Cauchy distance is better than ℓ_2 distance for handling large corruptions such as outliers. In this paper, we do not focus on the

[1] For instance, according to the works [34,59], the Frobenius norm will be used if the data are contaminated with dense noise; the ℓ_1-norm will be adopted if the data are contaminated with sparse corruptions; the ℓ_{21}-norm will be used if the data are contaminated with gross corruptions over a few columns; or the combination of these norms (*e.g.*, GMM) will be used for mixed patterns of noise and corruptions. In the work [41], the authors employed the Cauchy norm for the case of outliers.

[2] One may refer to [41] for detailed proof.

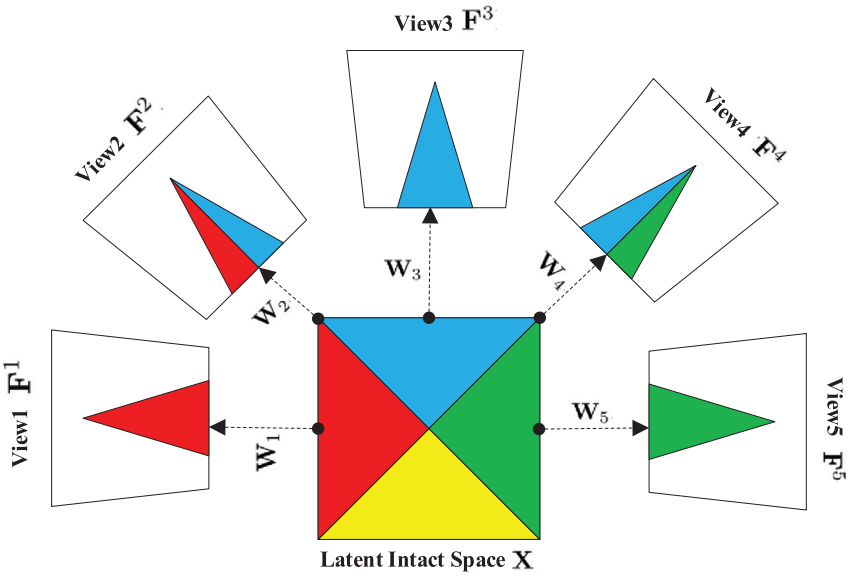

Fig. 1. View insufficiency assumption provided in the work [41] . Each individual view only captures partial information but all views involve the intact information. So intact space learning aims to recover the latent intact space $\mathbf{X}$ from its partial views $\mathbf{F}^v$, where $v \in \{1, \ldots, V\}$. Specifically, the work [41] assumes that the partial view $\mathbf{F}^v$ at hand is obtained from its latent intact space $\mathbf{X}$ through a linear sample matrix $\mathbf{W}^v$. Details are provided in Section 4.1.

"better" choice of the penalty (*i.e.*, Cauchy loss) and simply employ the ℓ^2 loss, say $\|\cdot\|_2^2$ (or the square of Frobenius norm $\|\cdot\|_F^2$), to do the job. The reasons are mainly summarized in the following three aspects.

The first one is that the square of Frobenius norm $\|\cdot\|_F^2$ is more efficient to solve than the Cauchy loss provided in the work [41]. The second one is that the employed norm depends on the prior knowledge of data. In practice, it is hard to decide which norm is the best. Thus we simply adopt the square of Frobenius norm. Last but not least, even if the data contain outliers, please note that the aim of this paper is to learn an *intactness-aware similarity* to do the clustering, the robustness to outliers can also be handled in the process of similarity construction[3] In other words, this paper mainly focuses on the informative similarity construction for multi-view clustering task. Finally, the intact space learning model can be simply formulated as:

$$\min_{\mathbf{X},\mathbf{W}_v} \quad \frac{1}{V}\sum_{v=1}^{V} ||\mathbf{F}^v - \mathbf{W}^v\mathbf{X}||_F^2 + \lambda_1||\mathbf{X}||_F^2. \qquad \text{s.t.} \quad \forall i, \quad ||\mathbf{W}_i^v||_2 \le 1 \tag{6}$$

Furthermore, it is interesting to note that the objective of intact space learning in Eq (6) is almost the same as the non-negative matrix factorization works like [20,60,61] or the multi-view dictionary learning works [62]. But differently, the intact space learning model is more elastic since it does not contain non-negative penalty on the learned matrix. Here, we take the objective (6) directly and mainly focus on the following informative similarity construction.

4.2. Intactness-aware similarity construction

After obtaining the intact space $\mathbf{X}$, the next step is to gather the data into their underlying clusters. In general, the spectral-based clustering methods like normalized cuts [63] usually show better performance than the k-means method, due to the manifold information utilized in these clustering models. As pointed out in [64], *the key in spectral clustering is the similarity graph construction*. So before applying the spectral clustering algorithm, we should construct an informative similarity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$. In this paper, we aim to learn an intactness-aware similarity $\mathbf{S}$ that has maximum dependence with the intact space $\mathbf{X}$ by resorting to the Hilbert–Schmidt Independence Criterion (HSIC). Specifically, given a set of intact space data $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$, we adopt the inner kernel $\mathbf{K}_1 = \mathbf{X}^T\mathbf{X}$ for the intact space $\mathcal{X}$, and simply employ the linear kernel $\mathbf{K}_2 = \mathbf{S} - \mathbf{D}$ for the similarity space $\mathcal{S}$, where $\mathbf{S}\succeq 0$ and $\mathbf{D}$ is defined as a diagonal matrix where the ith diagonal entry is $\Sigma_j \mathbf{S}_{ij}$. Then according to the empirical estimate of HSIC, we have the following constraint:

$$\begin{aligned} &\max_{\mathbf{S}} \quad \text{HSIC}(\mathbf{X}, \mathbf{S}) = \max_{\mathbf{S}} \\ &\text{tr}(\mathbf{K}_1\mathbf{H}\mathbf{K}_2\mathbf{H}) = \max_{\mathbf{S}} \quad \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{H}(\mathbf{S}-\mathbf{D})\mathbf{H}) \\ &\quad = -\max_{\mathbf{S}} \ \text{tr}(\mathbf{X}\mathbf{H}\mathbf{L}(\mathbf{X}\mathbf{H})^T) = \min_{\mathbf{S}} \ \text{tr}(\mathbf{X}\mathbf{H}\mathbf{L}(\mathbf{X}\mathbf{H})^T), \end{aligned} \tag{7}$$

where $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$ is a centering matrix to make the intact space to be centered, $\mathbf{L} = \mathbf{D} - \mathbf{S}$ can be viewed as the Laplacian matrix of $\mathbf{S}$. Moreover, to make the points of the constructed similarity lie in a union of affine subspaces, we also enforce $\mathbf{S}_i^T\mathbf{1} = 1$. Thus, we should minimize the following objective to get the informative similarity:

$$\min_{\mathbf{S}} \quad \text{tr}(\mathbf{X}\mathbf{H}\mathbf{L}(\mathbf{X}\mathbf{H})^T) \quad \text{s.t.} \ \mathbf{S}_i^T\mathbf{1} = 1, \quad \mathbf{S}_i \succeq 0. \tag{8}$$

4.3. New insight on intactness-aware similarity

Prior to giving the solution of the proposed intactness-aware model (8), we further discuss its underlying implication. Specifically, if we assume that the intact space $\mathbf{X}$ is centered, then multiplying the intact space $\mathbf{X}$ by the centering matrix $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$ does not make any change. In other words, $\mathbf{X} = \mathbf{X}\mathbf{H}$. Thus, the objective of Eq. (8) can be rewritten as follows:

$$\begin{aligned} &\min_{\forall i \ \mathbf{S}_i^T\mathbf{1}=1, \ \mathbf{S}_i\succeq 0} \text{tr}(\mathbf{X}\mathbf{H}\mathbf{L}(\mathbf{X}\mathbf{H})^T) = \text{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^T) \\ &\quad = \min_{\forall i \ \mathbf{S}_i^T\mathbf{1}=1, \ \mathbf{S}_i\succeq 0} \sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{X}_i - \mathbf{X}_j||_2^2\mathbf{S}_{ij} \end{aligned} \tag{9}$$

[3] The details will be provided in Section 4.3.

Thus, this constraint is consistent with the intuition, *i.e.*, data points should have a larger similarity (probability) to be in the same cluster if they have a smaller distance, or vice versa. However, simply solving the problem (9) may result in that only the nearest data is assigned as the neighbor of $\mathbf{X}_i$ with probability 1 and all the others with probabilities 0 (*i.e.*, the learned similarity matrix $\mathbf{S}$ is an identity matrix). So we enforce $||\mathbf{S}||_F^2$ to prevent from the trivial solution like:

$$\min_{\mathbf{S}} \quad \lambda_2 \sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{X}_i - \mathbf{X}_j||_2^2 \mathbf{S}_{ij} + \gamma ||\mathbf{S}||_F^2 \quad \text{s.t.} \quad \mathbf{S}_i^T \mathbf{1} = 1,\ \mathbf{S}_i \succeq 0. \tag{10}$$

where λ_2 and γ are non-negative trade-off parameters.

Furthermore, remember that in intact space learning model, we adopt the square of Frobenius norm to handle the case of dense noise, while in large corruptions such as outliers, the recovered intact space may not be perfect. We address this issue by replacing the ℓ_2 distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ into ℓ_1 distance in Eq. (10). Finally, our proposed *intactness-aware similarity learning model* can be formulated as follows:

$$\min_{\mathbf{S}} \quad \lambda_2 \sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{X}_i - \mathbf{X}_j||_1 \mathbf{S}_{ij} + \gamma ||\mathbf{S}||_F^2 \quad \text{s.t.} \quad \mathbf{S}_i^T \mathbf{1} = 1,\ \mathbf{S}_i \succeq 0. \tag{11}$$

Note that constructing similarity based on distance, the most popular way is the Gaussian similarity function:

$$\mathbf{S}_{ij} = \exp\left(\frac{-||\mathbf{X}_i - \mathbf{X}_j||_2^2}{2\sigma^2}\right), \tag{12}$$

which results in local neighborhoods are connected with relatively high weights, while edges between far away points have positive, but negligible weights. The trade-off parameter σ can be tuned in practice.

Comparing our intactness-aware similarity with the Gaussian similarity, our advantages can be mainly summarized in three aspects. First, the intactness-aware similarity is *adaptively* learned based on the local distance of data. Second, the employed ℓ_1 distance is more robust than ℓ_2 distance to outliers. Last but not least, empirically, considering all the connections like the Gaussian similarity may not work very well in practice. Usually we can achieve better performance if putting focus on the locality of data. In other words, it is preferred to learn a sparse $\mathbf{S}_i$, *i.e.*, only the k nearest neighbors of $\mathbf{X}_i$ could have chance to connect to $\mathbf{X}_i$, rather than all the data points.[4]

4.4. Multi-view Subspace Clustering with Intactness-Aware Similarity

A straightforward way to accomplish the grouping is firstly recovering the intact space by Eq. (6) and then building the similarity graph according to Eq. (12) (in this paper we call it **MSC+Gauss**) or Eq. (11) (**MSC+IAS**), in two separated steps. However, separately conducting the intact space recovery and the similarity measurement may not be optimal for constructing the similarity matrix, since the similarity is constructed based on the recovered intact space. To address this issue, we further empirically propose to simultaneously learn the intact space and the similarity matrix in a unified fashion. Specifically, combining (6) and (11), and discarding the stability term $\lambda_1||\mathbf{X}||_F^2$ in Eq. (6) since $\mathbf{X}$ has been regularized by Eq. (11), leads to our final Multi-view Subspace Clustering with Intactness-Aware Similarity (**MSC_IAS**) model as follows:[5]

$$\min_{\mathbf{W}_v, \mathbf{X}, \mathbf{S}} \underbrace{\frac{1}{V}\sum_{v=1}^{V} ||\mathbf{F}^v - \mathbf{W}^v\mathbf{X}||_F^2}_{\text{Intact Space Learning}} + \underbrace{\lambda_2 \sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{X}_i - \mathbf{X}_j||_1 \mathbf{S}_{ij} + \gamma ||\mathbf{S}||_F^2}_{\text{Intactness-Aware Similarity}}$$
$$\text{s.t.} \quad \forall i, \quad ||\mathbf{W}_i^v||_2 \le 1; \quad \mathbf{S}_i^T \mathbf{1} = 1, \quad \mathbf{S}_i \succeq 0. \tag{13}$$

Fig. 2 shows the main framework of the proposed method.

5. Optimization

Prior to giving the solution of the proposed MSC_IAS model (13), we simplify it to be more concise. Specifically, according to the work [65], we know that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{X}_i - \mathbf{X}_j||_1 \mathbf{S}_{ij} = ||\mathbf{X}\mathbf{A}_{\mathbf{S}}||_1, \tag{14}$$

where $\mathbf{A}_{\mathbf{S}} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$, $\mathbf{U}$ and $\mathbf{\Sigma}$ are the eigen decomposition of the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$.

For solving the MSC_IAS model (13), we observe that the objective function is not jointly convex to the variables $\{\mathbf{W}^v, \mathbf{X}, \mathbf{S}\}$, but convex with respect to each of them when the others are fixed. Therefore, we adopt the ADMM algorithm [66] to solve the associated optimization problem, which has proven to be an efficient and effective solver of problems like (13). To make the subproblem easy to solve, we introduce one auxiliary variable $\mathbf{Q}$ to replace the spare term $\mathbf{X}\mathbf{A}_{\mathbf{S}}$. Accordingly, $\mathbf{Q} = \mathbf{X}\mathbf{A}_{\mathbf{S}}$ acts as the additional constraint. Note that the probability properties of every $\mathbf{S}_i$ are enforced as hard constraints. The augmented Lagrangian function of (13) is:

$$\mathcal{L}_{\{\forall i,\ ||\mathbf{W}_i^v||_2 \le 1;\ \mathbf{S}_i^T\mathbf{1}=1,\ \mathbf{S}_i \succeq 0\}}(\mathbf{W}, \mathbf{X}, \mathbf{Q}, \mathbf{S}) = \frac{1}{V}\sum_{v=1}^{V} ||\mathbf{F}^v - \mathbf{W}^v\mathbf{X}||_F^2 + \lambda_2||\mathbf{Q}||_1 + \gamma||\mathbf{S}||_F^2 + \Phi(\mathbf{Z}, \mathbf{Q} - \mathbf{X}\mathbf{A}_{\mathbf{S}}), \tag{15}$$

with the definition $\Phi(\mathbf{Z}, \mathbf{C}) = \frac{\mu}{2}||\mathbf{C}||_F^2 + \langle \mathbf{Z}, \mathbf{C} \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the matrix inner product, μ is a positive penalty scalar and, $\mathbf{Z}$ is the Lagrangian multiplier. Therefore, there are mainly four variables, including **W, X, Q** and **S**, to solve. The designed solver iteratively updates one variable at a time by fixing the others. The solutions of the subproblems are described below.

X-subproblem: By leaving only terms in (15) that depend on $\mathbf{X}$, we obtain:

$$\min_{\mathbf{X}} \quad \frac{1}{V}\sum_{v=1}^{V} ||\mathbf{F}^v - \mathbf{W}^v\mathbf{X}||_F^2 + \Phi(\mathbf{Z}, \mathbf{Q} - \mathbf{X}\mathbf{A}_{\mathbf{S}}) \tag{16}$$

Taking the derivative of the above objective respect to $\mathbf{X}$ and setting it to zero, leads to the following equation:

$$\left(\frac{2}{V}\sum_{v=1}^{V}\mathbf{W}^{vT}\mathbf{W}^v\right)\mathbf{X} + \mathbf{X}(\mu\mathbf{A}_{\mathbf{S}}\mathbf{A}_{\mathbf{S}}^T) = \frac{2}{V}\sum_{v=1}^{V}\mathbf{W}^{vT}\mathbf{F}^v + \mathbf{Z}\mathbf{A}_{\mathbf{S}}^T + \mu\mathbf{Q}\mathbf{A}_{\mathbf{S}}^T \tag{17}$$

This is a standard Sylvester equation, which is solved by using the Bartels–Stewart algorithm [67].

[4] We will discuss this fact in the Section 5.

[5] Empirically, we found that adding an extra $\lambda_1||\mathbf{X}||_F^2$ term in MSC_IAS makes little difference to the clustering performance, but results in more complex solver and higher computational cost. While for the MSC+Gauss and MSC+IAS, this term is needful to make the intact space learning stabilized.

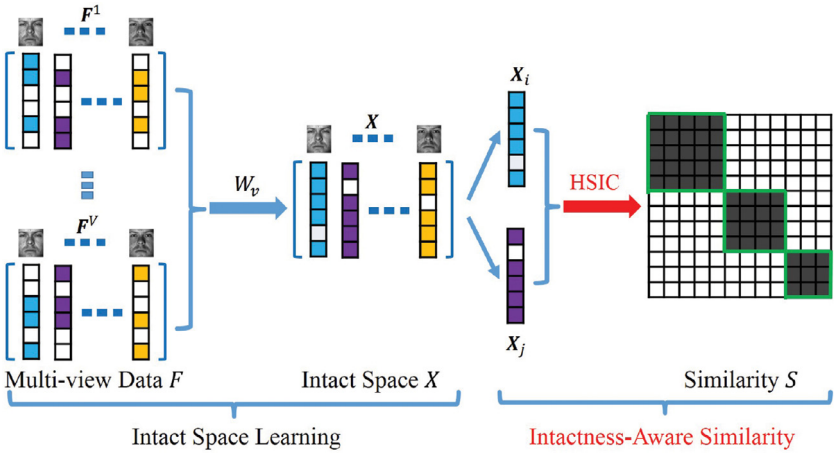

Fig. 2. The framework of Multi-view Subspace Clustering with Intactness-Aware Similarity (MSC_IAS). Given a collection of data points with multiple features, *e.g.*, $\mathbf{F}^1, \ldots, \mathbf{F}^V$, the proposed MSC_IAS integrates all the multi-view information to learn an intact space $\mathbf{X}$, and constructs the intactness-aware similarity matrix $\mathbf{S}$ via the HSIC, into a unified optimization framework. Moreover, the proposed HSIC can be explained as that the similarity is learned by assigning the adaptive and optimal neighbors for each intact data point based on the local connectivity. Afterwards, the spectral clustering algorithm such as the normalized cuts is employed on the learned similarity to obtain the final clustering results.

W-subproblem: It is obvious that $\mathbf{W}$ can be solved by views. Dropping the unrelated terms and taking derivative of $\mathcal{L}$ with respect to $\mathbf{W}^v$ reads:

$$\min_{\mathbf{W}^v} \ ||\mathbf{F}^v - \mathbf{W}^v\mathbf{X}||_F^2, \quad \forall i, \quad ||\mathbf{W}_i^v||_2 \le 1. \tag{18}$$

According to the work [68], we know that the above subproblem can be effectively solved by the following ADMM algorithm:

$$\begin{cases} \mathbf{W}^{v(r+1)} = \underset{\mathbf{W}^v}{\operatorname{argmin}} \ ||\mathbf{F}^v - \mathbf{W}^v\mathbf{X}||_F^2 + \tau||\mathbf{W}^v - \mathbf{M}^{(r)} + \mathbf{T}^{(r)}||_F^2 \\ \qquad\qquad = (\mathbf{F}^v\mathbf{X}^T + \tau(\mathbf{M}^{(r)} - \mathbf{T}^{(r)}))(\mathbf{X}\mathbf{X}^T + \tau\mathbf{I})^{-1} \\ \mathbf{M}^{(r+1)} = \underset{\mathbf{M}}{\operatorname{argmin}} \ \tau||\mathbf{W}^{v(r+1)} - \mathbf{M} + \mathbf{T}^{(r)}||_F^2, \ \text{s.t.} \ ||\mathbf{M}_i||_2^2 \le 1 \\ \mathbf{T}^{(r+1)} = \mathbf{T}^{(r)} + \mathbf{W}^{v(r+1)} - \mathbf{M}^{(r+1)}, \end{cases} \tag{19}$$

where τ is a penalty scalar and is updated ($\tau^{(r+1)} = 1.2\tau^{(r)}$) if appropriate. r is the iteration index and the maximum number of iterations is 100. We use the same settings as the work [68] does. For more details, one may refer to [68].

Q-subproblem: The corresponding subproblem is given by:

$$\min_{\mathbf{Q}} \ \lambda_2||\mathbf{Q}||_1 + \Phi(\mathbf{Z}, \mathbf{Q} - \mathbf{X}\mathbf{A}_\mathbf{S}). \tag{20}$$

We can use the soft-thresholding operator to get $\mathbf{Q}$ as:

$$\mathbf{Q} = \operatorname{sign}\left(\mathbf{X}\mathbf{A}_\mathbf{S} - \frac{\mathbf{Z}}{\mu}\right)\max\left(|\mathbf{X}\mathbf{A}_\mathbf{S} - \frac{\mathbf{Z}}{\mu}| - \frac{\lambda_3}{\mu}\right). \tag{21}$$

S-subproblem: The associated problem is written as:

$$\min_{\mathbf{S}} \ \lambda_2\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{X}_i - \mathbf{X}_j||_1\mathbf{S}_{ij} + \gamma||\mathbf{S}||_F^2 \ \text{s.t.} \ \forall i, \quad \mathbf{S}_i^T\mathbf{1} = 1; \ \mathbf{S}_i \succeq 0. \tag{22}$$

It can be separated into a set of smaller independent problems,

$$\forall i \quad \mathbf{S}_i =: \underset{\mathbf{S}_i^T\mathbf{1}=1; \ \mathbf{S}_i \succeq 0}{\operatorname{argmin}} ||\mathbf{S}_i + \mathbf{d}_i^X||_2^2, \tag{23}$$

where $\mathbf{d}_i^X \in \mathbb{R}^n$ is a vector, the jth element of which is $\mathbf{d}_{ij}^X = \frac{\lambda_3||\mathbf{X}_i - \mathbf{X}_j||_1}{2\gamma}$. For each $\mathbf{S}_i$, the closed-form solution can be easily obtained [22,28]:

$$\mathbf{S}_i = \left(\frac{1 + \sum_{j=1}^{k}\tilde{\mathbf{d}}_{ij}^X}{k}\mathbf{1} - \mathbf{d}_i^X\right)_+, \tag{24}$$

where the operator $(\mathbf{u})_+$ turns negative elements in $\mathbf{u}$ to 0 while keeping the rest. Please notice that the parameter $k \in \{1, \ldots, n\}$ is introduced to control the number of nearest neighbors of $\mathbf{X}_i$ that could have chance to connect to $\mathbf{X}_i$. Therefore, the local connectivity is balanced by the parameter k, and shows more superior than the popular Gaussian similarity in general. In addition, the elements of $\tilde{\mathbf{d}}_{ij}^X$ are those of $\mathbf{d}_{ij}^X$ but with the ascending order. Since the graph constructed according to $\mathbf{S}$ obtained by (24) is generally an unbalanced digraph, we employ $\frac{\mathbf{S}+\mathbf{S}^T}{2}$ to achieve the balance. Moreover, the parameter γ can be determined by:

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}\mathbf{d}_{i,k+1}^X - \frac{1}{2}\sum_{j=1}^{k}\mathbf{d}_{ij}^X\right). \tag{25}$$

Multiplier: Besides, the multiplier $\mathbf{Z}$ needs to be updated, which can be simply done through:

$$\mathbf{Z} = \mathbf{Z} + \mu(\mathbf{Q} - \mathbf{X}\mathbf{A}_\mathbf{S}); \quad \mu = \mu\rho. \tag{26}$$

The entire algorithm is summarized in Algorithm 1, which terminates when the maximal number (in all the experiments conducted in this paper, 100 is used) of iterations is reached or the objective value $f(t)$ in the tth iteration meets the stop criteria $\frac{|f(t+1) - f(t)|}{f(t)} \le 10^{-2}$.

6. Experiments

6.1. Experimental settings

6.1.1. Dataset description

Six benchmark datasets adopted in the experiments are those widely used in recent works [24,25,29] for face and image clustering, including:

Extended Yale-B[6] consists of 2414 face images of 38 individuals. Each individual has 64 near frontal images under different illuminations. Similar to [24], we select the first 10 classes as the final dataset, which has 640 frontal face images in total.

Yale[7] is composed of 165 grayscale images of 15 individuals. Each individual has 11 images, with different facial expression and configuration.

[6] http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html.
[7] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

Algorithm 1: **M**ulti-view **S**ubspace **C**lustering with **I**ntactness-**A**ware **S**imilarity (MSC_IAS).

Input: Multi-view data matrices: $\mathbf{F}^1, \mathbf{F}^2, \ldots, \mathbf{F}^V$, latent intact space data dimension d, cluster number c, nearest neighbor number k, parameter λ_2.
Initialize: $\mathbf{X}^{(0)}$, $\mathbf{W}^{v(0)}$, $\mathbf{Q}^{(0)}$, $\mathbf{S}^{(0)}$ and $\mathbf{Z}^{(0)}$, $\mu^{(0)} = 1.25$, $\rho > 1$, $t = 1$.
while *not converged* **do**
  $t = t + 1$;
  Update $\mathbf{X}^{(t)}$ via Eq. (17);
  for *v from 1 to V* **do**
    Update $\mathbf{W}^{v(t)}$ via Eq. (19);
  end
  Update $\mathbf{Q}^{(t)}$ via Eq. (21);
  for *i from 1 to n* **do**
    Update $\mathbf{S}_i^{(t)}$ via Eq. (24);
  end
  Balance $\mathbf{S}$ by $\frac{\mathbf{S}+\mathbf{S}^T}{2}$;
  Update the multipliers $\mathbf{Z}^{(t)}$ via Eq. (26);
  $\mu^{(t+1)} = \mu^{(t)}\rho$;
end
Apply the normalized cuts algorithm [63] on the intactness-aware similarity matrix $\mathbf{S}$.
Output: Final data clustering results.

ORL[8] contains 400 face images of 40 distinct subjects. Each subject has 10 different face images, which were taken at different times, changing with the lighting, facial expressions and facial details.

COIL-20 [9] consists of 1440 images of 20 object categories. Each class has 72 images. All images are normalized to 32×32 pixel arrays with 256 gray levels per pixel.

MSRCV1 [10] consists of 240 images of 9 object classes. We select 7 classes, *i.e.*, tree, building, airplane, cow, face, car and bicycle.

BBCSport [11] contains the documents from the BBC Sport website corresponding to the sports news in 5 topical areas.

For the datasets Extended Yale-B, Yale, ORL and COIL-20, three types of features, *i.e.* intensity, LBP [69] and Gabor [70] are extracted to form the multi-view information. For the dataset MSRCV1, six types of features are extracted from each image to construct different view features. For the document dataset BBC-Sport, two views are associated. All the processed datasets are provided by the works [24,25]. For more details, please refer to them.

6.1.2. Compared methods

We compare our method with recently proposed state-of-the-art alternatives, including 3 single-view methods and 8 multi-view ones.

SPC [27]: We select the most informative view to perform with the standard spectral clustering scheme.

LRR [30]: Low-rank constraint and the best performed single-view feature are used in this method.

RTC [71]: The method utilizes tensor to represent images and it is robust to outliers.

PCA+LRR: We first concatenate all the types of multiple features and employ PCA to reduce the feature dimension to 300, then do the clustering with LRR method.

Min-Disagreement [72]: This method creates a bipartite graph and tries to minimize the disagreement. Then the final result is obtained by spectral clustering.

Co-Reg SPC [38]: The method co-regularizes the clustering hypotheses to enforce that corresponding data points should be in the same cluster.

Co-Training SPC [73]: The co-training is adopted within the spectral clustering framework.

RMSC [47]: This method first clusters on each view and then integrates them to exploit a shared low-rank transition probability matrix.

ConvexReg SPC [39]: A common representation for all views is first learned. Then the standard spectral clustering is carried out on the similarity matrix.

LT-MSC [24]: Low-rank tensor constraint is enforced to directly construct the similarity matrix and then perform the spectral clustering.

LMSC [25]: This method seeks the underlying latent representation and simultaneously performs data reconstruction based on the learned latent representation.

As aforementioned, one might wonder the performance of first recovering $\mathbf{X}$ by optimizing the model of (6) and then somehow constructing the similarity matrix on the recovered $\mathbf{X}$, in two separated steps. To verify the advantages of our intactness-aware similarity and the gain of our unified method, we add two extra methods, the difference between which comes from the way of measuring the distance between data points to construct the similarity. One employs the commonly used Gaussian similarity function (12) as the baseline, called **MSC+Gauss**, the other adopts the manner of (11), termed as **MSC+IAS**.

6.1.3. Evaluation metrics

To assess the performance, six metrics including Normalized Mutual Information (**NMI**), Accuracy (**ACC**), Adjusted Rand Index (**ARI**), **F-score, Precision** and **Recall** are utilized, of which, the **ACC** and **NMI** are the most two popular metrics and have been adopted in the literature, such as [71,74,75]. For other metrics **ARI**, ***F*-score, Precision** and **Recall**, detailed definitions can be found in [76,77], and also have been widely used in [24,78]. Overall, these six metrics favor different properties in the clustering. For all the metrics, a higher value indicates a better clustering quality. We report the average accuracy and standard derivation of all the competitors over 30 independent trials.

6.2. Experimental results

6.2.1. Parameters effect

We test the parameters effect on Extended Yale-B dataset as an example, and the influence of ACC and NMI are displayed. For the baseline **MSC+Gauss**, there are three parameters, the regularizer weights λ_1 and the latent dimension d for learning the intact space in step one, and the trade-off parameter σ for the Gaussian similarity function (12) in step two. We tune one parameter by fixing the others, as shown in Fig. 3. We can see that all the four parameters have a certain effect on the Extended Yale-B. For all the datasets, the space of parameter λ_1 is {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1}, parameter $d \in \{100, 200, \ldots, 2000\}$ and $\sigma \in \{0.01, 0.05, 0.1, 0.5, 1\}$, respectively. The best performance is reported after tuning the parameters in their corresponding spaces. For **MSC+IAS**, based on the above analysis, we know that in step two, the parameters λ_2 and γ can be determined by the parameter k. Thus this model has three major parameters including λ_1, d and k that need to be tuned. Specifically, we tune the parameter $\lambda_1 \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$, parameter $d \in \{100, 200, \cdots, 2000\}$ and k from 1 to 15, respectively. The parameter effects are displayed in Fig. 4. Similarly, the best performance are reported by

[8] http://www.uk.research.att.com/facedatabase.html .
[9] http://www.cs.columbia.edu/CAVE/software/softlib/ .
[10] http://research.microsoft.com/en-us/projects/objectclassrecognition/ .
[11] http://mlg.ucd.ie/datasets/ .

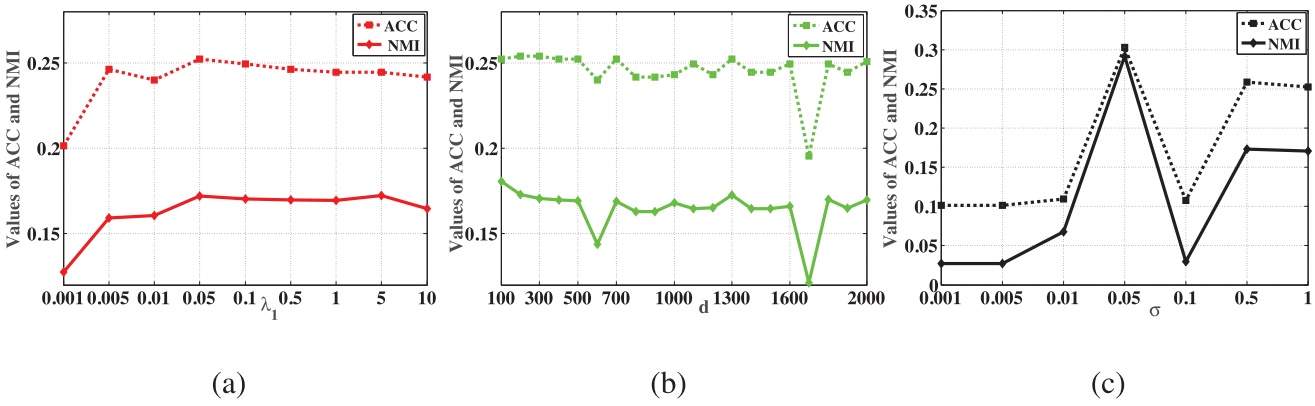

Fig. 3. MSC+Gauss: From left to right, the parameter effect of λ_1, d and σ on Extended Yale-B dataset with ACC (dashed) and NMI (solid) metrics, respectively.

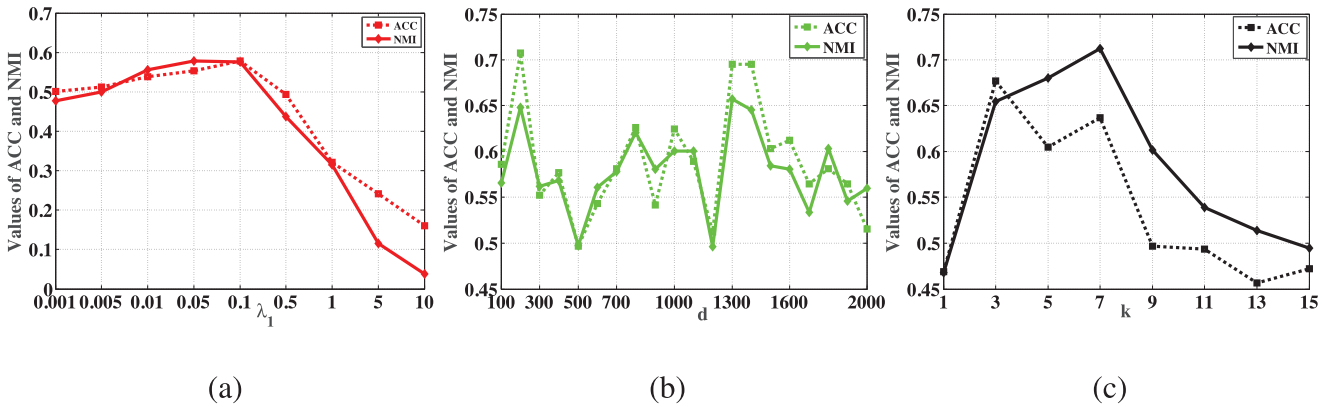

Fig. 4. MSC+IAS: From left to right, the parameter effect of λ_1, d and k on Extended Yale-B dataset with ACC (dashed) and NMI (solid) metrics, respectively.

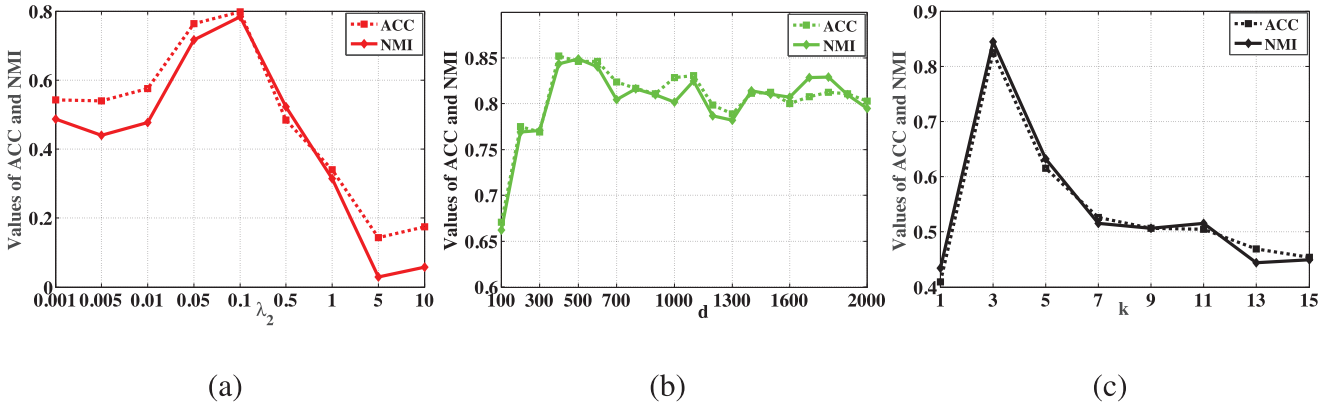

Fig. 5. MSC_IAS: From left to right, the parameter effect of λ_2, d and k on Extended Yale-B dataset with ACC (dashed) and NMI (solid) metrics, respectively.

tuning the parameters in their corresponding spaces. For all other compared methods, their parameters tuning is done in the work [24], here we simply copy the corresponding results from the paper [24]. For **MSC_IAS**, there are also three parameters λ_2, d and k need to be tuned. Specifically, we tune the parameter λ_2 by fixing the other parameters. As shown in Fig. 5, the λ_2 influence with ACC and NMI metrics is displayed in the first sub-graph of Fig. 5. From the curves, we can observe that, the ACC and NMI dramatically rises as the value increases from 0.01 to 0.1, while it starts to drop afterwards. This is reasonable since the spectral-based clustering is a little sensitive to the input similarity matrix. The second sub-figure of Fig. 5 gives the effect of the intact space dimension d on Extended Yale-B, the ACC and NMI dramatically rises as the dimension increases from 100 to 500, and drops afterwards. The last plots of Fig. 5 shows the effect of the nearest neighbor number k, the trend of which shows a similar phenomenon with that of the dimension, *i.e.* the accuracy is improved by a larger k in a certain range, but degenerates when k is too large. Both the above two experiments indicate that the performance gains from the sufficiency of information but would be hurt by the redundancy.

6.2.2. Performance comparison

We report the detailed clustering results on six benchmark datasets in Tables 1–5. The values in bold-italic represent the best performance while the bold values are the second best performance.

Table 1 provides the quantitative comparison among the competitors on Extended Yale-B dataset. It can be observed that most

Table 1
Results (mean ± standard deviation) on **Extended Yale-B**. We set $d = 500, k = 3, \lambda_2 = 0.1$ in MSC_IAS.

	Method	NMI	ACC	ARI	F-score	Precision	Recall
	SPCbest [27]	0.360 ± 0.016	0.366 ± 0.059	0.225 ± 0.018	0.303 ± 0.011	0.296 ± 0.010	0.310 ± 0.012
Single-view	LRRbest [30]	0.625 ± 0.004	0.615 ± 0.013	0.451 ± 0.002	0.508 ± 0.004	0.481 ± 0.002	0.539 ± 0.001
	RTC [71]	0.373 ± 0.001	0.360 ± 0.000	0.215 ± 0.005	0.291 ± 0.003	0.287 ± 0.005	0.294 ± 0.002
	PCA+LRR	0.568 ± 0.005	0.569 ± 0.012	0.400 ± 0.003	0.463 ± 0.002	0.433 ± 0.002	0.498 ± 0.002
	Min-Disagreement [72]	0.186 ± 0.003	0.242 ± 0.018	0.088 ± 0.001	0.181 ± 0.001	0.174 ± 0.001	0.189 ± 0.002
	Co-Reg SPC [38]	0.151 ± 0.001	0.224 ± 0.000	0.066 ± 0.001	0.160 ± 0.000	0.157 ± 0.001	0.162 ± 0.000
Multi-view	Co-Train SPC [73]	0.302 ± 0.007	0.186 ± 0.001	0.043 ± 0.001	0.140 ± 0.001	0.137 ± 0.001	0.143 ± 0.002
	RMSC [47]	0.157 ± 0.019	0.210 ± 0.013	0.060 ± 0.014	0.155 ± 0.012	0.151 ± 0.012	0.159 ± 0.013
	ConvexReg SPC [39]	0.163 ± 0.022	0.216 ± 0.019	0.072 ± 0.012	0.164 ± 0.010	0.163 ± 0.010	0.165 ± 0.011
	LT-MSC [24]	0.637 ± 0.003	0.626 ± 0.010	0.459 ± 0.030	0.521 ± 0.006	0.485 ± 0.001	0.539 ± 0.002
	LMSC [25]	**0.715 ± 0.011**	**0.736 ± 0.010**	0.578 ± 0.021	0.618 ± 0.009	**0.654 ± 0.010**	0.786 ± 0.009
Baseline	MSC+Gauss	0.441 ± 0.005	0.451 ± 0.007	0.266 ± 0.005	0.348 ± 0.005	0.309 ± 0.004	***0.961 ± 0.003***
	MSC+IAS	0.702 ± 0.010	0.718 ± 0.009	**0.597 ± 0.005**	**0.637 ± 0.008**	0.615 ± 0.008	0.716 ± 0.012
Proposed	MSC_IAS	***0.809 ± 0.008***	***0.783 ± 0.006***	***0.701 ± 0.008***	***0.735 ± 0.003***	***0.698 ± 0.012***	**0.802 ± 0.011**

Table 2
Results (mean ± standard deviation) on **Yale**. We set $d = 500, k = 3, \lambda_2 = 0.05$ in MSC_IAS.

	Method	NMI	ACC	ARI	F-score	Precision	Recall
	SPCbest [27]	0.654 ± 0.009	0.616 ± 0.030	0.440 ± 0.011	0.475 ± 0.011	0.457 ± 0.011	0.495 ± 0.010
Single-view	LRRbest [30]	0.709 ± 0.011	0.697 ± 0.000	0.515 ± 0.004	0.547 ± 0.007	0.529 ± 0.003	0.567 ± 0.004
	RTC [71]	0.607 ± 0.013	0.594 ± 0.016	0.371 ± 0.005	0.412 ± 0.012	0.384 ± 0.005	0.443 ± 0.025
	PCA+LRR	0.632 ± 0.006	0.582 ± 0.038	0.353 ± 0.009	0.396 ± 0.008	0.360 ± 0.007	0.441 ± 0.008
	Min-Disagreement [72]	0.645 ± 0.005	0.615 ± 0.043	0.433 ± 0.006	0.470 ± 0.006	0.446 ± 0.005	0.496 ± 0.006
	Co-Reg SPC [38]	0.648 ± 0.002	0.564 ± 0.000	0.436 ± 0.002	0.466 ± 0.000	0.455 ± 0.004	0.491 ± 0.003
Multi-view	Co-Train SPC [73]	0.672 ± 0.006	0.630 ± 0.011	0.452 ± 0.010	0.487 ± 0.009	0.470 ± 0.010	0.505 ± 0.007
	RMSC [47]	0.684 ± 0.033	0.642 ± 0.036	0.485 ± 0.046	0.517 ± 0.043	0.500 ± 0.043	0.535 ± 0.044
	ConvexReg SPC [39]	0.673 ± 0.023	0.611 ± 0.035	0.466 ± 0.032	0.501 ± 0.030	0.476 ± 0.032	0.532 ± 0.029
	LT-MSC [24]	0.765 ± 0.008	0.741 ± 0.002	0.570 ± 0.004	0.598 ± 0.006	0.569 ± 0.004	0.629 ± 0.005
	LMSC [25]	0.754 ± 0.009	0.768 ± 0.010	**0.659 ± 0.011**	**0.649 ± 0.008**	**0.623 ± 0.006**	0.701 ± 0.003
Baseline	MSC+Gauss	0.562 ± 0.009	0.442 ± 0.008	0.286 ± 0.010	0.338 ± 0.005	0.284 ± 0.009	***0.881 ± 0.006***
	MSC+IAS	**0.769 ± 0.011**	**0.773 ± 0.009**	0.582 ± 0.008	0.603 ± 0.006	0.578 ± 0.009	0.683 ± 0.012
Proposed	MSC_IAS	***0.821 ± 0.008***	***0.823 ± 0.006***	***0.709 ± 0.005***	***0.706 ± 0.007***	***0.698 ± 0.010***	**0.759 ± 0.011**

Table 3
Results (mean ± standard deviation) on **ORL**. We set $d = 500, k = 7, \lambda_2 = 0.05$ in MSC_IAS.

	Method	NMI	ACC	ARI	*F*-score	Precision	Recall
	SPCbest [27]	0.884 ± 0.002	0.726 ± 0.025	0.655 ± 0.005	0.664 ± 0.005	0.610 ± 0.006	0.728 ± 0.005
Single-view	LRRbest [30]	0.895 ± 0.006	0.773 ± 0.003	0.724 ± 0.020	0.731 ± 0.004	0.701 ± 0.001	0.754 ± 0.002
	RTC [71]	0.792 ± 0.001	0.601 ± 0.000	0.450 ± 0.002	0.465 ± 0.002	0.388 ± 0.003	0.581 ± 0.001
	PCA+LRR	0.867 ± 0.003	0.750 ± 0.033	0.650 ± 0.007	0.658 ± 0.007	0.624 ± 0.007	0.696 ± 0.008
	Min-Disagreement [72]	0.876 ± 0.002	0.748 ± 0.051	0.654 ± 0.004	0.663 ± 0.004	0.615 ± 0.004	0.718 ± 0.003
	Co-Reg SPC [38]	0.853 ± 0.003	0.715 ± 0.000	0.602 ± 0.004	0.615 ± 0.000	0.567 ± 0.004	0.666 ± 0.004
Multi-view	Co-Train SPC [73]	0.901 ± 0.003	0.730 ± 0.005	0.656 ± 0.007	0.665 ± 0.007	0.612 ± 0.008	0.727 ± 0.006
	RMSC [47]	0.872 ± 0.012	0.723 ± 0.025	0.645 ± 0.029	0.654 ± 0.028	0.607 ± 0.033	0.709 ± 0.027
	ConvexReg SPC [39]	0.883 ± 0.013	0.734 ± 0.031	0.668 ± 0.032	0.676 ± 0.035	0.628 ± 0.041	0.731 ± 0.030
	LT-MSC [24]	0.930 ± 0.002	0.795 ± 0.007	0.750 ± 0.003	**0.768 ± 0.007**	**0.766 ± 0.009**	0.837 ± 0.004
	LMSC [25]	**0.931 ± 0.011**	**0.819 ± 0.017**	***0.988 ± 0.002***	0.758 ± 0.009	0.735 ± 0.008	0.832 ± 0.006
Baseline	MSC+Gauss	0.791 ± 0.006	0.610 ± 0.008	0.494 ± 0.012	0.507 ± 0.009	0.469 ± 0.007	***0.891 ± 0.009***
	MSC+IAS	0.912 ± 0.011	0.805 ± 0.006	0.732 ± 0.013	0.742 ± 0.009	0.687 ± 0.008	0.809 ± 0.010
Proposed	MSC_IAS	***0.939 ± 0.007***	***0.845 ± 0.005***	**0.806 ± 0.009**	***0.802 ± 0.010***	***0.772 ± 0.011***	**0.851 ± 0.013**

Table 4
Results (mean ± standard deviation) on **COIL-20**. We set $d = 700, k = 3, \lambda_2 = 1$ in MSC_IAS.

	Method	NMI	ACC	ARI	F-score	Precision	Recall
	SPCbest [27]	0.806 ± 0.008	0.661 ± 0.061	0.619 ± 0.018	0.640 ± 0.017	0.596 ± 0.021	0.692 ± 0.013
Single-view	LRRbest [30]	0.829 ± 0.006	0.761 ± 0.003	0.719 ± 0.020	0.734 ± 0.004	0.717 ± 0.001	0.751 ± 0.002
	RTC [71]	0.755 ± 0.002	0.654 ± 0.021	0.543 ± 0.001	0.568 ± 0.005	0.522 ± 0.002	0.623 ± 0.000
	PCA+LRR	0.832 ± 0.004	0.770 ± 0.031	0.718 ± 0.007	0.732 ± 0.011	0.725 ± 0.004	0.739 ± 0.011
	Min-Disagreement [72]	0.789 ± 0.002	0.661 ± 0.052	0.597 ± 0.005	0.619 ± 0.005	0.579 ± 0.007	0.666 ± 0.003
	Co-Reg SPC [38]	0.765 ± 0.001	0.560 ± 0.000	0.568 ± 0.003	0.593 ± 0.000	0.558 ± 0.003	0.627 ± 0.002
Multi-view	Co-Train SPC [73]	0.813 ± 0.005	0.648 ± 0.016	0.604 ± 0.012	0.625 ± 0.011	0.588 ± 0.016	0.671 ± 0.005
	RMSC [47]	0.801 ± 0.018	0.685 ± 0.045	0.637 ± 0.044	0.656 ± 0.042	0.620 ± 0.057	0.698 ± 0.026
	ConvexReg SPC [39]	0.815 ± 0.023	0.693 ± 0.049	0.647 ± 0.055	0.666 ± 0.051	0.622 ± 0.071	0.720 ± 0.033
	LT-MSC [24]	0.862 ± 0.002	0.804 ± 0.011	0.748 ± 0.004	0.761 ± 0.007	0.741 ± 0.009	0.776 ± 0.006
	LMSC [25]	0.879 ± 0.011	0.802 ± 0.009	**0.821 ± 0.017**	0.794 ± 0.006	0.783 ± 0.008	0.898 ± 0.012
Baseline	MSC+Gauss	0.910 ± 0.010	0.801 ± 0.007	0.783 ± 0.005	0.793 ± 0.009	0.707 ± 0.007	***0.961 ± 0.009***
	MSC+IAS	**0.914 ± 0.009**	**0.811 ± 0.008**	0.801 ± 0.010	**0.804 ± 0.011**	**0.798 ± 0.014**	0.823 ± 0.007
Proposed	MSC_IAS	***0.958 ± 0.005***	***0.845 ± 0.009***	***0.849 ± 0.010***	***0.839 ± 0.012***	***0.803 ± 0.008***	***0.910 ± 0.006***

Table 5
Results (mean ± standard deviation) on **MSRCV1** and **BBCSport**. We set $\{d = 700, k = 5, \lambda_2 = 1\}$ and $\{d = 500, k = 3, \lambda_2 = 0.1\}$ in MSC_IAS, respectively.

		MSRCV1			BBCSport		
	Method	ACC	NMI	*F*-score	ACC	NMI	F-score
Single-view	SPCbest [27]	0.668 ± 0.050	0.574 ± 0.031	0.535 ± 0.043	0.836 ± 0.035	0.715 ± 0.060	0.767 ± 0.038
	LRRbest [30]	0.593 ± 0.013	0.492 ± 0.011	0.453 ± 0.044	0.787 ± 0.002	0.692 ± 0.001	0.769 ± 0.002
	RTC [71]	0.457 ± 0.017	0.329 ± 0.003	0.298 ± 0.007	0.710 ± 0.005	0.654 ± 0.005	0.740 ± 0.009
	PCA+LRR	0.557 ± 0.009	0.483 ± 0.012	0.415 ± 0.011	0.763 ± 0.004	0.674 ± 0.001	0.743 ± 0.007
Multi-view	Min-Disagreement [72]	0.692 ± 0.034	0.606 ± 0.032	0.574 ± 0.004	0.797 ± 0.049	0.776 ± 0.001	0.769 ± 0.006
	Co-Reg SPC [38]	0.653 ± 0.016	0.569 ± 0.012	0.537 ± 0.021	0.733 ± 0.058	0.717 ± 0.008	0.766 ± 0.013
	Co-Train SPC [73]	0.601 ± 0.007	0.582 ± 0.006	0.554 ± 0.003	0.732 ± 0.006	0.702 ± 0.009	0.698 ± 0.008
	RMSC [47]	0.691 ± 0.007	0.585 ± 0.006	0.576 ± 0.016	0.857 ± 0.009	0.812 ± 0.012	0.866 ± 0.009
	ConvexReg SPC [39]	0.575 ± 0.011	0.523 ± 0.012	0.508 ± 0.009	0.798 ± 0.016	0.776 ± 0.013	0.743 ± 0.015
	LT-MSC [24]	0.783 ± 0.012	0.692 ± 0.015	0.672 ± 0.017	0.851 ± 0.615	0.810 ± 0.012	0.820 ± 0.012
	LMSC [25]	0.805 ± 0.012	0.653 ± 0.010	0.651 ± 0.017	***0.900*** ± ***0.004***	**0.825** ± **0.006**	***0.886*** ± ***0.007***
Baseline	MSC+Gauss	0.723 ± 0.004	0.654 ± 0.003	0.621 ± 0.006	0.754 ± 0.018	0.732 ± 0.011	0.750 ± 0.006
	MSC+IAS	**0.806** ± **0.012**	**0.701** ± **0.011**	**0.679** ± **0.016**	0.846 ± 0.007	0.821 ± 0.009	0.835 ± 0.018
Proposed	MSC_IAS	***0.857*** ± ***0.050***	***0.777*** ± ***0.050***	***0.737*** ± ***0.050***	**0.892** ± **0.008**	***0.856*** ± ***0.012***	**0.882** ± **0.009**

of the comparisons have relatively low performances. The major reason is the large variation of illumination of this dataset. Our proposed MSC_IAS algorithm still achieves significant improvements around 9.4%, 4.7%, 12.3%, 11.7%, 4.4% and 1.6% over the most competitive method LMSC [25] in terms of NMI, ACC, AR, *F*-score, Precision and Recall, respectively. Moreover, we can see that the MSC+Gauss generally has poor performance due to the connections of all data points for constructing the similarity. But interesting, we find that the MSC+Gauss has high values on the metric of Recall. For the MSC+IAS, it outperforms the state-of-the-art method LT-MSC [24], and also achieves better performance than MSC+Gauss. The reason of this is likely that the intactness-aware similarity is learned by adaptively according to the local connectivity. For the proposed MSC_IAS, it achieves the highest performance on this dataset, which shows not only the powerful of local connectivity in constructing the similarity graph but also the superiority of the unified optimization.

Table 2 displays the clustering results on Yale dataset. Similar trend to Table 1, MSC+Gauss achieves lower performance than most of compared methods. However, from the values, we can see that MSC+IAS excels all the baselines, both single-view and multi-view methods. The main reason is that the *intactness-aware* similarity we constructed is based on the local connections. Moreover, the more robust to outliers ℓ_1 distance has been adopted. The most three competitive multi-view clustering methods RMSC, LT-MSC and LMSC, have achieved a relatively promising results. The single-view method LRR also has competitive results. However, MSC+IAS is still comparable with them. As a comparison, the proposed MSC_IAS algorithm further gains significant (at least 5%) improvements over MSC+IAS, which has validated the superiority of the unified optimization of the proposed MSC_IAS.

Table 3 shows the performance comparison on ORL dataset, from which we notice that all of SPC, Co-Reg SPC, Co-Train SPC, RMSC, ConvexReg SPC perform relatively poorly. The methods LRR, LT-MSC and LMSC produce more promising results on this dataset. It seems that the self-representation models are more suitable for constructing the similarity matrix on this dataset. The improvement of MSC_IAS over these two methods is not very significant. However, our method introduces k nearest neighbor concept (with relatively small k) in the similarity matrix, which may connect the objects of the same subject with slight pose changes. Thus, our method achieve higher performance based on this informative similarity matrix.

Table 4 gives the clustering results on COIL-20 dataset. Consistently, the proposed MSC_IAS outperforms the other competitors thanks to the intactness-aware similarity matrix constructed on the latent intact space. There are at least 3% improvements over MSC+IAS in terms of all the six evaluation metrics. In addition, from the numbers in the four tables, MSC+Gauss competes very favorably with LT-MSC, which implies that the recovery of the intact space is beneficial. MSC+IAS continuously improves MSC+Gauss. That is to say, the similarity construction strategy (11) is more powerful than simply using the Gaussian distance. By comparing our MSC_IAS with MSC+IAS, we can clearly observe the advantages of jointly recovering the intact space and optimizing the similarity.

Table 5 gives the clustering results on MSRCV1 and BBCSport, respectively. On MSRCV1 dataset, we can see that our MSC_IAS outperforms the compared methods by a large margin. The reason behind this is that the learned intact space can benefit from more views. As a consequence, the similarity learned by our model Eq. (13) is more reliable for spectral clustering. On BBCSport dataset, the non-image processing case, we can see that the proposed MSC_IAS achieves at least 2% improvements over the competitors RMSC and LT-MSC. Comparing with the recent proposed LMSC, our method MSC_IAS still achieves higher performance on most of evaluation metrics. Moreover, on these two datasets, we can see a similar trend as that shown on the former four datasets, that is, the jointly recovering the intact space and optimizing the similarity MSC_IAS is generally better than the two separated steps MSC+IAS and the similarity constructed using the Gaussian distance.

6.3. Exploratory experiments

6.3.1. Visualization

Fig. 6 gives the similarity matrices of our method on Extended Yale-B, Yale, ORL and COIL-20 datasets, respectively. We plot these similarity matrices according to the intended clusters. It can be seen that our method can reveal the underlying diagonal-block structures very well (please see the zoomed-in regions shown in Fig. 6 for details), which are desired by spectral-based clustering methods. This further validates the advantages of our unified model.

To further validate the advantages of our proposed MSC_IAS algorithm, we also give the visualization of clustering results. The clustering examples are shown in Fig. 7. We select the first 3 clusters, and for each cluster, 10 images are randomly chosen to show. As shown in Fig. 7(a), each row contains incorrect faces and the accuracy on the Extended Yale-B dataset is not very high. In Fig. 7(b), each cluster consists few errors, which has demonstrated that the proposed MSC_IAS is good for face clustering. The clustering results on the ORL dataset are also promising. As shown in Fig. 7(c), the clustering accuracies of the first 3 clusters (top-down) are: 80%, 90% and 90%, respectively. On the COIL-20 dataset, our

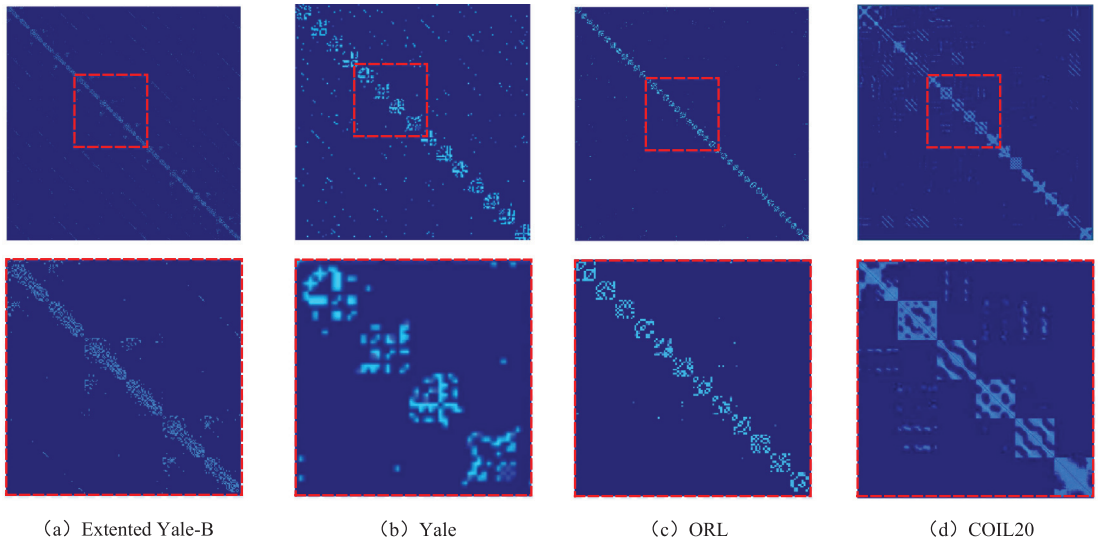

Fig. 6. From left to right: Visualization of similarity matrices by MSC_IAS on Extended Yale-B, Yale, ORL and COIL-20, respectively. The zoomed-in patches in the lower row correspond to the regions bounded by red boxes in the upper row. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

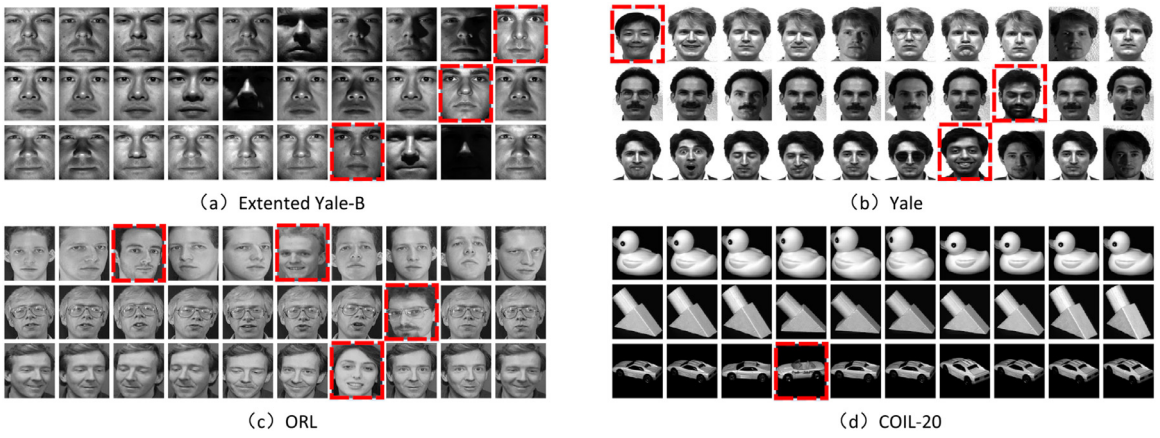

Fig. 7. Visualization of the MSC_IAS clustering results on four datasets.

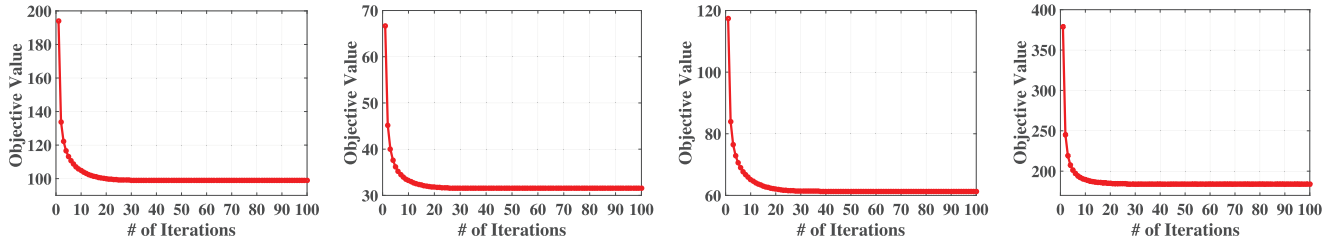

Fig. 8. From left to right: convergence speed of the proposed MSC_IAS algorithm on Extended Yale-B, Yale, ORL and COIL-20, respectively.

method achieves a significantly better clustering result, as shown in Fig. 7(d), only the third row contains incorrect images.

6.3.2. Computational complexity

The associated problem of our model (Eqs. (13) and (15)) is solved by mainly dividing it into four optimisation subproblems and solving them alternatingly. Among those subproblems, there are three operations that give rise to high computational cost. Specifically, the three operations are: (1) The eigen decomposition of Laplacian matrix $\mathbf{L}$, which has a complexity of $O(n^3)$. However, since the Laplacian matrix $\mathbf{L}$ is sparse, by employing the sparse analysis methods [79], the complexity can be reduced to $O(rn^2)$, where r is the ratio of nonzero samples in $\mathbf{L}$ to the total number of samples n; (2) The solving of Sylvester equation. As analysis in [67], its complexity is of $O(n^2)$; (3) The matrix multiplication operation of $\mathbf{A_S}\mathbf{A_S}^T$. By leveraging the eigen decomposition of the $\mathbf{L}$ matrix, *i.e.*, we choose the first K largest eigenvalues to get $\mathbf{A_S}$. With this reduced $\mathbf{A_S}$, the cost is $O(Kn^2)$. All in all, the fact that all three operations have a complexity of $O(n^2)$ brings down the complexity of the whole algorithm to $O(n^2)$.

6.3.3. Convergence analysis

It is worth noting that there is no established theory of global convergence in literature for ADMM algorithms applied to non-convex problems as the one solved in this work. Similar to [22,24,78], we show the convergence speed empirically. More specifically, the convergence speed of our algorithm on the four datasets is given in the Fig. 8. It suggests that the proposed algorithm has very strong and stable convergence behavior. It can be seen that the MSC_IAS converges in less than 40 iterations, and running the MSC_IAS algorithm on Extended Yale-B, Yale, ORL and COIL-20 only takes about 118 s, 76 s, 98 s, and 179 s (seconds) on a laptop with 3.20 GHz Intel Core i5 CPU and 4GB RAM, respectively. To allow more experimental verifications, our code can be downloaded from http://www.cbsr.ia.ac.cn/users/xiaobowang.

7. Conclusion

This paper has proposed a novel multi-view subspace clustering model to construct an intactness-aware similarity based on the recent intact space learning technique. Specifically, the intact space is recovered based on multi-view information and the intactness-aware similarity is constructed based on the local connectivity of intact space. For the similarity construction, different from previous models, in our method, the data similarity matrix is constructed by adaptively assigning the neighbors for each intact space data point according to the local connectivity. Moreover, the robustness to outliers ℓ_1 distance is employed. Finally, we have formulated the overall problem into a unified optimization framework and have designed an effective and efficient ADMM based algorithm to seek the solution. The experimental results on several commonly adopted benchmark datasets, in comparison with the state-of-the-art alternatives, have demonstrated the significant improvement of the proposed method. In the future, it is interesting to extend MSC_IAS to handle incomplete multi-view data.

Acknowledgments

This work was supported by National Key Research and Development Plan (grant no. 2016YFC0801002), the Chinese National Natural Science Foundation Projects #61876178, #61473291, #61572501, #61502491, #61572536, the Science and Technology Development Fund of Macau (no. 151/2017/A and 152/2017/A), and AuthenMetric R&D Funds.

References

[1] A. Zimek, J. Vreeken, The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives, Mach. Learn. 98 (1) (2015) 121–155.
[2] S. Kanaan-Izquierdo, A. Ziyatdinov, A. Perera-Lluna, Multiview and multifeature spectral clustering using common eigenvectors, Pattern Recognit. Lett. 102 (2018) 30–36.
[3] C. Lu, S. Yan, Z. Lin, Convex sparse spectral clustering: single-view to multi-view, IEEE Trans. Image Process. 25 (6) (2016) 2833–2843.
[4] Y. Wang, L. Wu, X. Lin, J. Gao, Multiview spectral clustering via structured low-rank matrix factorization, IEEE Trans. Neural Netw. Learn. Syst. 99 (2018) 1–11.
[5] L. Houthuys, R. Langone, J.A. Suykens, Multi-view kernel spectral clustering, Inf. Fusion 44 (2018) 46–56.
[6] X. He, L. Li, D. Roqueiro, K. Borgwardt, Multi-view spectral clustering on conflicting views, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 826–842.
[7] S.C. Johnson, Hierarchical clustering schemes., Psychometrika 32 (3) (1967) 241–254.
[8] J. Lee, D. Yeung, C. Tsang, Hierarchical clustering based on ordinal consistency, Pattern Recognit. 38 (11) (2005) 1913–1925.
[9] J. MacQueen, Some methods for classification and analysis of multivariate observations., in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
[10] S. Bickel, T. Scheffer, Multi-view clustering., in: Proceedings of Industrial Conference on Data Mining, 2004, pp. 19–26.
[11] P. Tseng, Nearest q-flat to m points, J. Optim. Theory Appl. 105 (1) (2000) 249–252.
[12] J. Ho, M.H. Yang, J. Lim., Clustering appearances of objects under varying illumination conditions., in: Proceedings of IEEE Computer Vision and Pattern Recognition, 2003.
[13] Y. Wang, W. Zhang, L. Wu., Iterative views agreement: an iterative low-rank based structured optimization method to multi-view spectral clustering., in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2016.
[14] J. Sublime, B. Matei, G. Cabanes., Entropy based probabilistic collaborative clustering., Pattern Recognit. 72 (2017) 144–157.
[15] W. Pedrycz, Collaborative fuzzy clustering., Pattern Recognit. Lett. 23 (14) (2002) 1675–1686.
[16] A. Cornuéjols, C. Wemmert, P. Gançarski., Collaborative clustering: why, when, what and how., Inf. Fusion 39 (2018) 81–95.
[17] J.P. Costeira, T. Kanade, A multibody factorization method for independently moving objects, Int. J. Comput. Vis. 29 (3) (1998) 159–179.
[18] X. Wang, Z. Lei, H. Shi, X. Guo, X. Zhu, S.Z. Li, Co-referenced subspace clustering, in: IEEE International Conference on Multimedia and Expo, 2018.
[19] K. Kanatani, Motion segmentation by subspace separation and model selection, in: Proceedings of IEEE Conference on Computer Vision, 2001, pp. 586–591.
[20] Z. Akata, C. Thurau, C. Bauckhage., Non-negative matrix factorization in multimodality data for segmentation and label prediction., in: Proceedings of the 16th Computer vision winter workshop, 2011.
[21] Y. Liu, L. Jiao, F. Shang, An efficient matrix factorization based low-rank representation for subspace clustering, Pattern Recognit 46 (1) (2013) 284–292.
[22] X. Guo, Robust subspace segmentation by simultaneously learning data representations and their affinity matrix., in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2015, pp. 119–128.
[23] X. Wang, X. Guo, Z. Lei, C. Zhang, S.Z. Li, Exclusivity-consistency regularized multi-view subspace clustering., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 923–931.
[24] C. Zhang, H. Fu, S. Liu, C. Liu, X. Cao, Low-rank tensor constrained multiview subspace clustering., in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1582–1590.
[25] C. Zhang, Q. Hu, H. Fu, P. Zhu, X. Cao, Latent multi-view subspace clustering., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4279–4287.
[26] M. Brbic, I. Kopriva, Multi-view low-rank sparse subspace clustering, Pattern Recognit. 73 (1) (2018) 247–258.
[27] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm., in: Proceedings of Advances in Neural Information Processing Systems, 2002, pp. 849–856.
[28] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors., in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 977–986.
[29] E. Elhamifar, R. Vidal, Sparse subspace clustering., in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2790–2797.
[30] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 171–184.
[31] C. Lu, H. Min, Z. Zhao, Robust and efficient subspace segmentation via least squares regression, in: Proceedings of IEEE Conference on European Conference on Computer Vision, 2012, pp. 347–360.
[32] J. Feng, Z. Lin, H. Xu, Robust subspace segmentation with block-diagonal prior., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3818–3825.
[33] C. You, R. Daniel, V. Rene, Scalable sparse subspace clustering by orthogonal matching pursuit., in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 3918–3927.
[34] C. Li, V. Rene, Structured sparse subspace clustering: a unified optimization framework., in: Proceedings of IEEE Computer Vision and Pattern Recognition., 2015, pp. 277–286.
[35] S. De, R. Virginia, Spectral clustering with two views., in: Proceedings of ICML Workshop on Learning with Multiple Views, 2005, pp. 20–27.
[36] X. Cai, F. Nie, W. Cai, Heterogeneous image features integration via multi–modal semi-supervised learning model., in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1737–1744.
[37] W. Tang, Z. Lu, I.S. Dhillon, Clustering with multiple graphs., in: Proceedings of Industrial Conference on Data Mining, 2009, pp. 1016–1021.
[38] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering., in: Proceedings of Advances in neural information processing systems, 2011, pp. 1413–1421.
[39] M. Collins, J. Liu, J. Xu, Spectral clustering with a convex regularizer on millions of images., in: Proceedings of IEEE Conference on European Conference on Computer Vision, 2014, pp. 282–298.
[40] H. Gao, F. Nie, X. Li, Multi-view subspace clustering., in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4238–4246.
[41] C. Xu, D. Tao, C. Xu, Multi-view intact space learning., IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 114–127.
[42] P. Pavlidis, J. Weston, J. Cai, W.S. Noble, Learning gene functional classifications from multiple data types, J. Comput. Biol. 9 (2) (2002) 401–411.
[43] Y. Zhang, X. Hu, X. Jiang, Multi-view clustering of microbiome samples by robust similarity network fusion and spectral clustering, IEEE/ACM Trans. Comput. Biol. Bioinf. 14 (2) (2017) 264–271.
[44] M. Žitnik, B. Zupan, Data fusion by matrix factorization, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 41–53.

[45] E. Bruno, S. Marchand-Maillet, Multiview clustering: a late fusion approach using latent models., in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, 2009, pp. 736–737.
[46] D. Greene, P. Cunningham, A Matrix Factorization Approach for Integrating Multiple Data Views, Springer Machine Learning and Knowledge Discovery in Databases, Berlin Heidelberg, 2009.
[47] R. Xia, Y. Pan, L. Du, Robust multi-view spectral clustering via low-rank and sparse decomposition., in: Proceedings of the AAAI Conference on Artificial Intelligence, 2014, pp. 2149–2155.
[48] B. Long, S. Philip, Z. Zhang, A general model for multiple view unsupervised learning., in: Proceedings of the Structural Dynamics and Materials, 2008, pp. 822–833.
[49] G. Cleuziou, M. Exbrayat, L. Martin, Cofkm: a centralized method for multiple-view clustering., in: Proceedings of the IEEE International Conference on Data Mining, 2009, pp. 752–757.
[50] Q. Li, Z. Sun, Z. Lin, Transformation invariant subspace clustering., Pattern Recognit. 59 (2016) 142–155.
[51] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering., in: Proceedings of IEEE Computer Vision and Pattern Recognition., 2015, pp. 586–594.
[52] Y. Wang, X. Lin, L. Wu, Robust subspace clustering for multi-view data by exploiting correlation consensus., IEEE Trans. Image Process. 24 (11) (2015) 3939–3949.
[53] Y. Guo, Convex subspace representation learning from multi-view data., in: AAAI, Vol. 1, 2013, p. 2.
[54] M. White, X. Zhang, D. Schuurmans, Y.-l. Yu, Convex multi-view subspace learning, in: Advances in Neural Information Processing Systems, 2012, pp. 1673–1681.
[55] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with Hilbert–Schmidt norms, in: International conference on algorithmic learning theory, Springer, 2005, pp. 63–77.
[56] L. Song, J.B. aand K M. Borgwardt, Gene selection via the bahsic family of algorithms., Bioinformatics 23 (13) (2007) 1490–1498.
[57] N. Quadrianto, L. Song, A.J. Smola, Kernelized sorting, in: Advances in Neural Information Processing Systems, 2009, pp. 1289–1296.
[58] A. Gretton, K. M.Borgwardt, M. Rasch, A kernel method for the two-sample-problem., in: Advances in Neural Information Processing Systems, 2006, pp. 513–520.
[59] Q. Zhao, D. Meng, Z. Xu, W. Zuo, L. Zhang, Robust principal component analysis with complex noise., in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 55–63.
[60] J. Liu, C. Wang, J. Gao., Multi-view clustering via joint nonnegative matrix factorization., in: Proceedings of the Structural Dynamics and Materials, 2013, pp. 252–260.
[61] H. Lu, Z. Fu, X. Shu, Non-negative and sparse spectral clustering, Pattern Recognit. 47 (1) (2014) 418–426.
[62] X. Jing, R. Hu, F. Wu., Uncorrelated multi-view discrimination dictionary learning for recognition, in: Proceedings of the Association for the Advancement of Artificial Intelligence, 2014, pp. 2787–2795.
[63] J. Shi, J. Malik, Normalized cuts and image segmentation., IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[64] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.
[65] K. Elyor, T. Xiang, Z.Y. Fu, S.H. Gong, Person re-identification by unsupervised l1 graph learning, in: Proceedings of IEEE Conference on European Conference on Computer Vision, 2016, pp. 283–299.
[66] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. learn. 3 (1) (2011) 1–122.
[67] R. Bartels, G.W. Stewart., Solution of the matrix equation $ax + xb = c$, Commun. ACM 15 (9) (1972) 820–826.
[68] S. Gu, L. Zhang, W. Zuo, X. Feng., Projective dictionary pair learning for pattern classification., in: Advances in Neural Information Processing Systems, 2014, pp. 793–801.
[69] T. Ojala, M. Pietikainen, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns., IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[70] M. Lades, J. Vorbruggen, J. Buhmann, Distortion invariant object recognition in the dynamic link architecture., IEEE Trans. Comput. 42 (3) (1993) 300–311.
[71] X. Cao, X. Wei, Y. Han, Robust tensor clustering with non-greedy maximization., in: Proceedings of the 25th International Joint Conference on Artificial Intelligence., 2013.
[72] S. De, R. Virginia, Spectral clustering with two views., in: Proceedings of International Conference on Machine Learning, 2005, pp. 20–27.
[73] A. Kumar, H. Daume, A co-training approach for multi-view spectral clustering., in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 393–400.
[74] X. Cao, C. Zhang, C. Zhou, Constrained multi-view video face clustering., IEEE Trans. Image Process. 24 (11) (2015) 4381–4393.
[75] J. Huang, F. Nie, H. Huang, Spectral rotation versus k-means in spectral clustering., in: Proceedings of the AAAI Conference on Artificial Intelligence, 2013.
[76] C.D. Manning, P. Raghavan, H. Schutze, Introduction to information retrieval., Cambridge: Cambridge University Press, Cambridge, 2008.
[77] L. Hubert, P. Arabie, Comparing partitions., J. Classif. 2 (1) (1985) 193–218.
[78] X. Cao, C. Zhang, H. Fu, Diversity-induced multi-view subspace clustering., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 586–594.
[79] M. Aharon, M. Elad, A. Bruckstein, K-svd: an algorithm for designing overcomplete dictionaries for sparse representation., IEEE Trans. Image Process. 54 (11) (2006) 4311–4322.

Xiaobo Wang received the B.S. and M.E. degrees from the School of Science, Tianjin University, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include machine learning, data mining, and computer vision.

Zhen Lei received the B.S. degree in automation from the University of Science and Technology of China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently an Associate Professor. He has published over 100 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an Area Chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.

Xiaojie Guo received the B.E. degree in software engineering from the School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China, in 2008, and the M.S. and Ph.D. degrees in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2010 and 2013, respectively. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He was a recipient of the Piero Zamperoni Best Student Paper Award in the International Conference on Pattern Recognition (International Association on Pattern Recognition), in 2010.

Changqing Zhang received the B.S. and M.S. degrees from the College of Computer Science, Sichuan University, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from Tianjin University, China, in 2016. He is currently an Assistant Professor with the School of Computer Science and Technology, Tianjin University. His current research interests include machine learning, data mining, and computer vision.

Hailin Shi is currently research scientist at JD AI Research, where he is in charge of R&D in face detection and recognition. Before that, he worked at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science. He receives his MS degrees from University of Paris 6, France. He receives his PhD degree from University of Chinese Academy of Science. His research interests include deep learning, computer vision, face recognition, person re-identification.

Stan Z. Li received his B.Eng. from Hunan University, China, M.Eng. from National University of Defense Technology, China, and PhD degree from Surrey University, UK. He is currently a professor and the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). He worked at Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor at Nanyang Technological University, Singapore. He was elevated to IEEE Fellow for his contributions to the fields of face recognition, pattern recognition and computer vision. His research interest includes pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has published over 200 papers in international journals and conferences, and authored and edited 8 books. He was an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and is acting as the editor-in-chief for the Encyclopedia of Biometrics. He served as a program co-chair for the International Conference on Biometrics 2007, 2009 and 2015, and has been involved in organizing other international conferences and workshops in the fields of his research interest.