Flexible Multi-View Dimensionality Co-Reduction

Changqing Zhang, Huazhu Fu, Qinghua Hu, *Senior Member, IEEE*, Pengfei Zhu, *Member, IEEE*, and Xiaochun Cao, *Senior Member, IEEE*

***Abstract*— Dimensionality reduction aims to map the high-dimensional inputs onto a low-dimensional subspace, in which the similar points are close to each other and vice versa. In this paper, we focus on unsupervised dimensionality reduction for the data with multiple views, and propose a novel method, called *Multi-view Dimensionality co-Reduction*. Our method flexibly exploits the complementarity of multiple views during the dimensionality reduction and respects the similarity relationships between data points across these different views. The kernel matching constraint based on Hilbert-Schmidt Independence Criterion enhances the correlations and penalizes the disagreement of different views. Specifically, our method explores the correlations within each view independently, and maximizes the dependence among different views with kernel matching jointly. Thus, the locality within each view and the consistence between different views are guaranteed in the subspaces corresponding to different views. More importantly, benefiting from the kernel matching, our method need not depend on a common low-dimensional subspace, which is critical to reduce the influence of the unbalanced dimensionalities of multiple views. Specifically, our method explicitly produces individual low-dimensional projections for individual views, which could be applied for new coming data in the out-of-sample manner. Experiments on both clustering and recognition tasks demonstrate the advantages of the proposed method over the state-of-the-art approaches.**

***Index Terms*— High dimensional, multi-view dimensionality co-reduction, kernel matching.**

I. Introduction

DIMENSIONALITY reduction (DR) has steadily been a fundamental technique for high-dimensional data. Although unsupervised dimensionality reduction can leverage abundant unlabeled data, it is still challenging due to the lack of label information to guide the reduction process. One effective way for improving the unsupervised approach is to introduce additional constraints, which can assist the unsupervised procedure. In this paper, we focus on performing dimensionality reduction for the data with multiple views, which can serve as references for each other.

In real world applications, data are often represented with multiple views, since data are usually collected from diverse sources or different feature extractors. Take images/videos for example, they are often described with different visual descriptors, such as SIFT [1], Gabor [2] and LBP [3]. These different types of features may characterize different specific information. Therefore, better performance could be expected by exploring the complementarity among different views. Recently, multi-view approaches have demonstrated the effectiveness in many applications (e.g., clustering [4], [5], classification [6], metric learning [7] and outlier detection [8]). They mainly benefit from the complementarity of multiple views, which can improve the quality of the performance.

The existing dimensionality reduction methods [9] [16]–[18] for single view are not applicable for extending to the multi-view setting directly, since they cannot exploit the intrinsic correlations between different views. Thus designing a suitable multi-view regularization is the key point of multi-view dimensionality reduction. A common assumption of multi-view learning is that *different views should be complementary with each other* [4], [19], [20]. Based on this role, a few works aiming for multi-view dimensionality reduction are proposed. For example, the method in [21] learns a common subspace to compare two different views. However, it is hard to deal with the data with more than two views. Some methods [10], [22], [23] break the view number limitation by using Canonical Correlation Analysis (CCA). They compare different views on a learned low-dimensional common subspace. Unfortunately, these methods may produce unsatisfactory results, especially when the dimensionalities of different views are unbalanced.

To address the view number limitation and unbalanced dimensionality challenges, in this paper, we propose a novel multi-view dimensionality reduction method, called **M**ulti-view **D**imensionality **c**o-**R**eduction (**MDcR**), as shown in Fig. 1. Our method utilizes the kernel matching to regularize the dependence across multiple views, and simultaneously, obtains the low-dimensional projection for each view. Specifically, the Hilbert-Schmidt Independence Criterion (HSIC) for kernel matching is employed to explore the correlations of different views, which avoids the restriction of reducing all the views to a common low-dimensional subspace [10], [22], [23]. The proposed method both explores the correlations within each view independently, and maximizes the dependence among different views with kernel matching jointly. Instead of learning a common subspace, our method utilizes kernel

Manuscript received April 26, 2016; revised August 14, 2016; accepted October 27, 2016. Date of publication November 10, 2016; date of current version December 6, 2016. This work was supported in part by the National Natural Science Foundation of China under Grants 61602337, Grant 61432011, and Grant 61602345, and in part by the National Program on Key Basic Research Project under Grant 2013CB329304. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ling Shao. *(Corresponding author: Xiaochun Cao.)*

C. Zhang, Q. Hu, and P. Zhu are with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: zhangchangqing@tju.edu.cn; huqinghua@tju.edu.cn; zhupengfei@tju.edu.cn).

H. Fu is with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632 (e-mail: huazhufu@gmail.com).

X. Cao is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: caoxiaochun@iie.ac.cn).

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

Digital Object Identifier 10.1109/TIP.2016.2627806

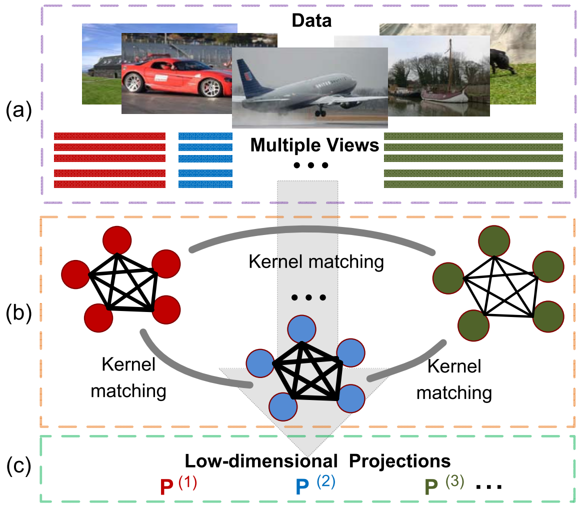


Fig. 1. Given the multi-view data (a), with the kernel matching constraint (b) our MDcR jointly learns low-dimensional projections (c). Benefiting from the HSIC, our MDcR does not require to compare different views in a common subspace.

matching to relieve the restriction of equal dimensionality for different views. This is especially important for unbalanced dimensionalities of different views. Finally, our method can be solved efficiently by using the alternating direction optimization strategy and converges to local optimal.

The remainder of the paper is organized as follows. In Section II, we briefly review the related work to the proposed method. The details of MDcR are elaborated in Section III. To verify the efficacy of our method, extensive experimental results are shown in Section IV. Finally, the conclusion of the paper is drawn in Section V.

II. Related Work

A. Dimensionality Reduction

Generally, most existing dimensionality reduction methods can be divided into two main categories, i.e., projection-based and manifold-based methods. The projection-based techniques [9], [24] map the high-dimensional data onto the low-dimensional subspace by using a learned projection. For example, Principal Component Analysis (PCA) [9] is an unsupervised linear method, which reduces dimensionality by embedding the data into a linear low-dimensional subspace. Metric Multidimensional Scaling (MDS) [24] maps high-dimensional data to a low-dimensional linear subspace with pairwise distances preserved. Specifically, given Euclidean distances, metric MDS is equivalent to PCA up to scaling and rotation [25]. Manifold-based techniques consider that data points are nonlinear and lie on an unknown manifold of lower dimensionality [16], [26], [27]. They execute dimensionality reduction on a batch of given data by modeling a manifold in terms of global [26] or local [16] intrinsic geometry. However, unlike projection-based methods, they are difficult to be applied for new data due to the lack of low-dimensional projection. Overall, the traditional dimensionality reduction methods are not applicable for multi-view data since there is no constraint for exploring correlations among multiple views.

B. Multi-View Learning

Multi-view data are ubiquitous in real world. Note that, the view concept is different from the works [53], [54] which define view in a geometry way. The existing multi-view learning methods can be divided into two categories, i.e., supervised/semi-supervised [7], [28]–[30] and unsupervised methods [10], [22], [31]–[33]. For unsupervised multi-view dimensionality reduction, Canonical Correlation Analysis (CCA) and its variants [10], [23] are widely utilized as a joint dimensionality reduction technique. Partial Least Squares (PLS) [34] maps different views to a common linear subspace in the regression manner and has been empirically proved to be superior to CCA [13]. Distributed Spectral Embedding (DSE) [15] tries to find a low-dimensional physically meaningful embedding with the distribution of each view being sufficiently smooth. The Multiple Kernel Learning (MKL) based method [14] learns a low-dimensional common representation with/without supervised information, however, the kernel mapping may be risky especially when the number of high dimensional data is small and it is not guaranteed to converge. The method in [12] utilizes a structured sparsity-inducing norm for the projection matrix to map different view to a common low-dimensional space. However, the dimensions of the patterns corresponding to different views in [12] must be the same to construct a 2-D grid, which limits its generation. Tensor CCA (TCCA) [11] generalizes CCA to handle the data of an arbitrary number of views by analyzing the covariance tensor of the different views, however, the main disadvantage of the TCCA lies in the rather high computational cost.

Generally, these existing approaches usually suffer from the following main disadvantages. First, these methods only focus on exploring the correlations among different views, but without exploring the relationships within each view itself (e.g., graph regularization for smoothness). Second, these methods usually reduce different views to a common low-dimensional space, which is unreasonable especially when the dimensionalities of different views are unbalanced. By contrast, our method simultaneously explores each view independently and enforces the dependence across multiple views. Furthermore, our method avoids suffering from the unbalanced dimensionalities of different views due to the kernel matching. For clarification, we give a brief overview of existing unsupervised multi-view dimensionality reduction methods in Table I.

C. Kernel Matching

The kernel matching criterion is developed to map each instance of the target domain into the corresponding instance of the source domain according to their geometric similarities expressed in kernel matrices [35]. The early work [36] employs HSIC to measure the independence between given random variables in Reproducing Kernel Hilbert Spaces. The work [37] proposes an unsupervised kernel sorting algorithm to match object pairs from two sources of observations based on the HSIC. The work [38] aims to find different novel clustering results and the novelty among each other

TABLE I

COMPARISON OF *UNSUPERVISED MULTI-VIEW DIMENSIONALITY REDUCTION* METHODS. INTRAVIEW: EXPLORING THE CORRELATIONS WITHIN EACH VIEW, INTERVIEW: EXPLORING THE CORRELATIONS ACROSS THE DIFFERENT VIEWS, MULTIPLEVIEW: APPLIED FOR THE DATA WITH MORE THAN 2 VIEWS, DIMFREE: FREE OF REDUCING EACH VIEW TO A COMMON SPACE ('√' AND '×' INDICATE PRESENCE AND ABSENCE OF PROPERTY, RESPECTIVELY)

Methods	IntraView	InterView	MultipleView	DimFree
PCA[9]	√	×	×	√
CCA[10]	×	√	√	×
TCCA [11]	×	√	√	×
SSMVD [12]	×	√	√	×
PLS[13]	×	√	×	√
MKL-DR[14]	√	√	√	×
DSE[15]	√	√	√	×
NaMDR	√	×	×	√
Our MDcR	√	√	√	√

is measured by HSIC. It employs the inner product kernel as well as two nonlinear kernels. However, the optimization for nonlinear case is not such efficient. In the work [35], the authors exploit the criterion HSIC in a semi-supervised manner to map pairs of instances to each other without exact correspondence requirement. The work [32] performs self-based subspace clustering with multiple views, while the HSIC is used for diversity measure. In our work, we make our goal to perform multi-view dimensionality reduction jointly. To this end, we learn the low-dimensional projection for each view by maximizing HSIC over the kernel matrices to explore the correlations across these multiple views.

III. THE PROPOSED METHOD

We firstly give a brief introduction for graph embedding dimensionality reduction [39]. Then we will detail the HSIC for kernel matching and induce the proposed MDcR method.

A. Graph Embedding Dimensionality Reduction

Graph embedding dimensionality reduction [39] aims to perform dimensionality reduction with local relationships of points preserved. Given $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$ with each column being a sample vector, the goal of dimensionality reduction is obtaining the corresponding low-dimensional representation $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{K\times N}$, where $K \ll D$ and $\mathbf{z}_i$ is the low-dimensional representation of $\mathbf{x}_i$. For the graph embedding based dimensionality reduction method [39], the objective is designed to ensure the sufficiently smooth on the the data manifold. The intuitive explanation is that if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close, then their low-dimensional representations $\mathbf{z}_i$ and $\mathbf{z}_j$ should be similar to each other. The distance of two low-dimensional representations $\mathbf{z}_i$ and $\mathbf{z}_j$ is defined as

$$d(\mathbf{z}_i, \mathbf{z}_j) = \|\frac{\mathbf{z}_i}{\sqrt{d_{ii}}} - \frac{\mathbf{z}_j}{\sqrt{d_{jj}}}\|^2, \tag{1}$$

where the distance is normalized by d_{ii} and d_{jj} in order to reduce the impact of popularity of nodes as in traditional graph-based learning [40], [41], and $\mathbf{D}$ is a diagonal matrix with element $d_{ii} = \sum_{j=1}^{n} w_{ij}$. $\mathbf{W} = (w_{ij})$ is the affinity matrix, which is often constructed by the original data matrix $\mathbf{X}$. Then, the graph-regularized representation can be learned by the following objective function

$$\min_{\mathbf{Z}\in\mathbb{R}^{K\times N}} \sum_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|^2 w_{ij} = \max_{\mathbf{Z}\in\mathbb{R}^{K\times N}} tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathsf{T}}), \tag{2}$$

where $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ is the normalized graph Laplacian matrix, and $tr(\cdot)$ denotes the trace of a matrix.

For the multi-view setting, a naive way is incorporating multiple views directly as

$$\max_{\mathbf{Z}^{(v)}\in\mathbb{R}^{K^{(v)}\times N}} \sum_{v=1}^{V} tr(\mathbf{Z}^{(v)}\mathbf{L}^{(v)}{\mathbf{Z}^{(v)}}^{\mathsf{T}}), \tag{3}$$

where v denotes the view index. $\mathbf{Z}^{(v)}$ and $\mathbf{L}^{(v)}$ denote the learned low-dimensional representation and the normalized graph Laplacian matrix corresponding to the v^{th} view, respectively. For explicitly learning the projections between high-dimensional and low-dimensional spaces, the projection $\mathbf{P}^{(v)}$ is introduced as $\mathbf{Z}^{(v)} = \mathbf{P}^{(v)}\mathbf{X}^{(v)}$ for the v^{th} view. Accordingly, the objective function in Eq. (3) is derived as

$$\begin{aligned} &\max_{\mathbf{P}^{(v)}\in\mathbb{R}^{K^{(v)}\times D^{(v)}}} \sum_{v=1}^{V} tr(\mathbf{P}^{(v)}\mathbf{X}^{(v)}\mathbf{L}^{(v)}{\mathbf{X}^{(v)}}^{\mathsf{T}}{\mathbf{P}^{(v)}}^{\mathsf{T}}) \\ &\text{s.t. } \mathbf{P}^{(v)}{\mathbf{P}^{(v)}}^{\mathsf{T}} = \mathbf{I}, v = 1, \ldots, V, \end{aligned} \tag{4}$$

where $D^{(v)}$ and $K^{(v)}$ correspond to the dimensionalities of the original (high-dimensional) space and the corresponding reduced (low-dimensional) subspace, respectively. The constraint $\mathbf{P}^{(v)}{\mathbf{P}^{(v)}}^{\mathsf{T}} = \mathbf{I}$ prevents the trivial solution. Intuitively, this naive way reduces the dimensionality of each view independently and does not exploit the correlations of these multiple views.

B. Kernel Matching via HSIC

In this work, we introduce Hilbert Schmidt Independence Criterion (HSIC) [36] to explore the correlations among different views, which has the following advantages [38]. First, HSIC measures dependence of the reduced subspaces of different views by mapping variables into a reproducing kernel Hilbert space (RKHS) such that these views need not depend on a common low-dimensional subspace, which is critical to reduce the influence of the unbalanced dimensionalities of multiple views. Second, in this manner, we can estimate dependence between different views without explicitly estimating the joint distribution of the random variables, which makes the algorithm computationally efficient. Last but not least, the empirical HSIC turns out to be equal to the trace of product of the data matrix with inner product kernel, which makes our problem solvable.

Suppose that two sets of observations $\mathbf{X}$ and $\mathbf{Y}$ are drawn jointly from a probability distribution $P_{\mathbf{xy}}$. The HSIC measures the dependence between $\mathbf{x}$ and $\mathbf{y}$ by computing the norm of the cross-covariance operator over the domain $\mathcal{X} \times \mathcal{Y}$ in Hilbert space. Given the universal Hilbert space is universal, this norm vanishes if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent. A large value suggests strong dependence with respect to the choice of kernels. Let us define a 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 $k_1(\mathbf{x}_i, \mathbf{x}_j) = < \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >$. Let $\mathcal{G}$ be a second kernel space on $\mathcal{Y}$, with kernel function $k_2(\mathbf{y}_i, \mathbf{y}_j) =< \varphi(\mathbf{y}_i), \varphi(\mathbf{y}_j) >$. The cross-covariance is a function that gives the covariance of two random variables and defined as

$$C_{\mathbf{xy}} = E_{\mathbf{xy}}[(\phi(\mathbf{x}) - \mu_{\mathbf{x}}) \otimes (\varphi(\mathbf{y}) - \mu_{\mathbf{y}})], \tag{5}$$

where $\mu_{\mathbf{x}} = E(\phi(\mathbf{x}))$ and $\mu_{\mathbf{y}} = E(\varphi(\mathbf{x}))$, and $\otimes$ is the tensor product. Then, we have the following definition of HSIC [36] as the Hilbert-Schmidt norm of the associated cross-covariance operator $C_{\mathbf{xy}}$

$$\text{HSIC}(P_{\mathbf{xy}}, \mathcal{F}, \mathcal{G}) := ||C_{\mathbf{xy}}||^2_{\text{HS}}, \tag{6}$$

where $||\mathbf{A}||_{\text{HS}}$ denotes the Hilbert-Schmidt norm of a matrix as

$$||\mathbf{A}||_{\text{HS}} = \sqrt{\sum_{i,j} a_{ij}^2}. \tag{7}$$

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 estimator of HSIC, written as HSIC$(\mathcal{Z}, \mathcal{F}, \mathcal{G})$, is given by

$$\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) = (n-1)^{-2} tr(\mathbf{K}_1 \mathbf{H} \mathbf{K}_2 \mathbf{H}), \tag{8}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are the Gram matrices with $k_{1,ij} = k_1(\mathbf{x}_i, \mathbf{x}_j)$, $k_{2,ij} = k_2(\mathbf{y}_i, \mathbf{y}_j)$. $h_{ij} = \delta_{ij} - 1/N$ centers the Gram matrix to have zero mean in the feature space. For more details of HSIC, please refer to the paper [36].

C. Flexible Multi-View Dimensionality Co-Reduction

To explore the correlations among multiple views, we introduce the kernel matching based on HSIC as a co-regularization to encourage the new representations of different views to be of sufficient dependence in Eq. (3). Accordingly, the objective function is formulated as

$$\max_{\mathbf{Z}^{(v)} \in \mathbb{R}^{K^{(v)} \times N}} \sum_{v=1}^{V} tr(\mathbf{Z}^{(v)} \mathbf{L}^{(v)} \mathbf{Z}^{(v)^{\mathsf{T}}}) + \lambda \sum_{v \neq u} \text{HSIC}(\mathbf{Z}^{(v)}, \mathbf{Z}^{(u)}), \tag{9}$$

where $\lambda > 0$ is the tradeoff parameter. The first term guarantees the smoothness within each view independently, and the second term enforces that the leaned representations should jointly depend on each other. With the projections $\mathbf{P}^{(v)}$'s, the objective function is derived as

$$\begin{aligned} \max_{\mathbf{P}^{(v)} \in \mathbb{R}^{K^{(v)} \times D^{(v)}}} & \sum_{v=1}^{V} tr(\mathbf{P}^{(v)} \mathbf{X}^{(v)} \mathbf{L}^{(v)} \mathbf{X}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)^{\mathsf{T}}}) \\ & + \lambda \sum_{v \neq u} \text{HSIC}(\mathbf{P}^{(v)} \mathbf{X}^{(v)}, \mathbf{P}^{(u)} \mathbf{X}^{(u)}), \\ \text{s.t. } & \mathbf{P}^{(v)} \mathbf{P}^{(v)^{\mathsf{T}}} = \mathbf{I}, \quad v = 1, \ldots, V. \end{aligned} \tag{10}$$

To solve the problem, we adopt alternating maximization strategy for our objective function. Since we use inner product kernel for HSIC, i.e., $\mathbf{K}^{(v)} = \mathbf{Z}^{(v)^{\mathsf{T}}} \mathbf{Z}^{(v)} = \mathbf{X}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)} \mathbf{X}^{(v)}$, HSIC can be rewritten in the form of matrix trace as

$$\begin{aligned} & \text{HSIC}(\mathbf{P}^{(v)} \mathbf{X}^{(v)}, \mathbf{P}^{(u)} \mathbf{X}^{(u)}) \\ & \quad = tr(\mathbf{K}^{(v)} \mathbf{H} \mathbf{K}^{(u)} \mathbf{H}) \\ & \quad = tr(\mathbf{X}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)} \mathbf{X}^{(v)} \mathbf{H} \mathbf{K}^{(u)} \mathbf{H}) \\ & \quad = tr(\mathbf{P}^{(v)} \mathbf{X}^{(v)} \mathbf{H} \mathbf{K}^{(u)} \mathbf{H} \mathbf{X}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)^{\mathsf{T}}}). \end{aligned} \tag{11}$$

With the alternating maximizing strategy, we can approximately solve Eq. (10) by maximizing with respect to one view once at a time while fixing the other views. With all but one fixed, we maximize the following objective function

$$\begin{aligned} \mathcal{O}(\mathbf{P}^{(v)}) &= tr(\mathbf{P}^{(v)} \mathbf{X}^{(v)} \mathbf{L}^{(v)} \mathbf{X}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)^{\mathsf{T}}}) \\ &\quad + \lambda \sum_{u \neq v} \text{HSIC}(\mathbf{P}^{(v)} \mathbf{X}^{(v)}, \mathbf{P}^{(u)} \mathbf{X}^{(u)}) \\ &= tr(\mathbf{P}^{(v)} \mathbf{X}^{(v)} \mathbf{L}^{(v)} \mathbf{X}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)^{\mathsf{T}}}) \\ &\quad + \lambda \sum_{u \neq v} \text{tr}(\mathbf{P}^{(v)} \mathbf{X}^{(v)} \mathbf{H} \mathbf{K}^{(u)} \mathbf{H} \mathbf{X}^{(v)^{\mathsf{T}}} \mathbf{P}^{(v)^{\mathsf{T}}}) \\ &= tr\big(\mathbf{P}^{(v)} (\mathbf{A} + \lambda \mathbf{B}) \mathbf{P}^{(v)^{\mathsf{T}}}\big) \end{aligned} \tag{12}$$

with

$$\begin{aligned} \mathbf{A} &= \mathbf{X}^{(v)} \mathbf{L}^{(v)} \mathbf{X}^{(v)^{\mathsf{T}}}, \\ \mathbf{B} &= \sum_{u \neq v} \mathbf{X}^{(v)} \mathbf{H} \mathbf{K}^{(u)} \mathbf{H} \mathbf{X}^{(v)^{\mathsf{T}}}. \end{aligned} \tag{13}$$

Note that, under the condition $\mathbf{P}^{(v)} \mathbf{P}^{(v)^{\mathsf{T}}} = \mathbf{I}$, the above problem is an eigenvalue decomposition task which can be efficiently solved.

Remarks: With the inner product kernel, HSIC maximization turns out to be equivalent to maximizing the trace of the product of the projected representation $\mathbf{P}^{(v)} \mathbf{X}^{(v)}$ as in Eq. (11). Then, our objective can be optimized with eigenvalue decomposition as in Eq. (12). Hence, it is simple to implement.

Connection With Disagreement Penalty: For our kernel matching constraint term, here we give an explanation from the other point of view. Specifically, our proposed method is a more general way to penalize the disagreement across multiple views. For any two affinity matrices corresponding to $\mathbf{Z}^{(v)}$ and $\mathbf{Z}^{(u)}$, the measure of disagreement between them can be defined as [4]

$$D(\mathbf{Z}^{(v)}, \mathbf{Z}^{(u)}) = \left\| \frac{\mathbf{K}_{\mathbf{Z}^{(v)}}}{||\mathbf{K}_{\mathbf{Z}^{(v)}}||_F^2} - \frac{\mathbf{K}_{\mathbf{Z}^{(u)}}}{||\mathbf{K}_{\mathbf{Z}^{(u)}}||_F^2} \right\|_F^2, \qquad (14)$$

where $\mathbf{K}_{\mathbf{Z}^{(v)}}$ is the affinity matrix for $\mathbf{Z}^{(v)}$, and $|| \cdot ||_F$ denotes the Frobenius norm of a matrix. The affinity matrices are normalized by their Frobenius norms, which makes them to be comparable across different affinity matrices. By using the linear kernel $k(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\mathsf{T}\mathbf{z}_j$ as the similarity measure in Eq. (14), we have $\mathbf{K}_{\mathbf{Z}^{(v)}} = \mathbf{Z}^{(v)^\mathsf{T}}\mathbf{Z}^{(v)}$. Under the condition $\mathbf{Z}^{(v)}\mathbf{Z}^{(v)^\mathsf{T}} = \mathbf{I}$, we have $||\mathbf{K}_{\mathbf{Z}^{(v)}}||_F^2 = tr(\mathbf{K}_{\mathbf{Z}^{(v)}}\mathbf{K}_{\mathbf{Z}^{(v)}}^T) = tr(\mathbf{Z}^{(v)^\mathsf{T}}\mathbf{Z}^{(v)}\mathbf{Z}^{(v)^\mathsf{T}}\mathbf{Z}^{(v)}) = K^{(v)}$, where $K^{(v)}$ is the dimensionality of the v^{th} subspace. Given the condition $K^{(v)} = K^{(u)}$, Eq. (14) can be rewritten as the following by ignoring the constant additive and scaling terms

$$\mathrm{D}(\mathbf{Z}^{(v)}, \mathbf{Z}^{(u)}) = -tr(\mathbf{Z}^{(v)^\mathsf{T}}\mathbf{Z}^{(v)}\mathbf{Z}^{(u)^\mathsf{T}}\mathbf{Z}^{(u)}). \qquad (15)$$

Note that, with the regularization by HSIC in Eq. (8), our method is more general than the method in [4], and the advantages of our method over the method in [4] are summarized as follows: 1) The method in [4] is only suitable for the case that the reduced dimensionalities are equal for different views, since Eq. (14) holds only under the condition $K^{(v)} = K^{(u)}$. This limitation makes it not such flexible as the kernel matching in our method. 2) The Eq. (14) holds only when the affinity matrices are constructed with the linear kernel $k(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\mathsf{T}\mathbf{z}_j$. By contrast, other nonlinear kernels [38] could be adopted in HSIC used in our method. 3) Unlike the method [4], our method explicitly learns the projections from high dimensional space to low dimensional space, which could be use for new data in the out-of-sample manner. 4) Moreover, experimental results further demonstrate the advantage of our method over the method in [4] as shown in Fig. 3.

D. Complexity and Convergence Analysis

The major computation of MDcR is composed of three parts, i.e., updating $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{P}^{(v)}$ for each iteration. For simplicity, we suppose the dimensionality of each view is D. For updating $\mathbf{A}$ and $\mathbf{B}$, the main complexity is the matrix multiplication, and the complexity of updating $\mathbf{A}$ and $\mathbf{B}$ are $O(DN^2)$ and $O(VDN^2)$, respectively, where V is the number of views and N is the sample number. Updating $\mathbf{P}^{(v)}$ is an eigendecomposition problem, and the computation complexity of eigendecomposition is $O(N^3)$. Hence the complexity of our method is $O(T(VDN^2 + N^3))$, where T is the number of iterations.

The alternating maximization is carried out until convergence. To prevent the algorithm being stuck in a local minimum [42], we initialize the projections of each view using NaMDR in Eq. (4), which directly reduces dimensionalities of multiple views without any correlation constraint. Since the objective is non-decreasing with the iterations, the algorithm is guaranteed to convergence. The algorithm of our MDcR is shown in Alg. 1. In practice, we monitor that the convergence is reached within less than five iterations in our experiments.

Algorithm 1 The Optimization Algorithm for MDcR

```
Input: Multi-view data $\mathcal{D} = \{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$,
       dimensionalities of learned subspaces
       $\{K^{(1)}, ..., K^{(V)}\}$, parameters $\lambda$
for each $v \in V$ do
    Initialize $\mathbf{P}^{(v)}$ by single view solution in Eq. (4).
end
while not converged do
    for each $v \in V$ do
        Compute $\mathbf{P}^{(v)}$ by objective function (12).
    end
end
Output: Projections $\mathcal{P} = \{\mathbf{P}^{(1)}, ..., \mathbf{P}^{(V)}\}$.
```

IV. Experiments

In this section, we test our method on two tasks: clustering and recognition. Five benchmark datasets are employed. Beyond the regular experiments, we use UCI Multiple Features dataset to validate the flexibility and effectiveness of our method in handling the unbalanced dimensionality case, while the ORL dataset is used to test the model (learned projections) transfer ability.

Yale[1]: This dataset contains 165 grayscale images of 15 individuals with 11 images per subject. We resize the images into 64 × 64 and extract 3 types of features: intensity (4096 dimensions), LBP (3304 dimensions) and Gabor (6750 dimensions). The features of the following face datasets, i.e., ORL and Notting-Hill, are extracted with the same manner.

ORL[2]: ORL dataset contains 10 different images of each of 40 distinct subjects, which were taken at different times, varying the lighting, facial expressions and facial details. The dataset acts as a testing data to test the learned low-dimensional projection from Yale Face dataset. The 3 types of features are the same to Yale.

Notting-Hill: The dataset *Notting-Hill* [43] is derived from the movie "Notting-Hill". Faces of 5 main casts are used, including 550 faces in 76 tracks. The original dataset consists of the facial images of the size of 120×150. To reduce the computational cost and the memory requirements, we downsample each facial image to 40×50. The 3 types of features are the same to Yale.

UCI Multiple Features (UCI-MF)[3]: This dataset consists of handwritten numbers ('0'-'9') with 200 patterns per class. Three types of features are extracted: 47 Zernike moments, 240 pixel averages in 2 × 3 windows and 6 morphological features.

Still DB: Still DB is a still images dataset used for action recognition. It contains 467 images and six classes. We extract Sift Bow (200 dimensions), Color Sift Bow (200 dimensions) and Shape context Bow (200 dimensions) as their features.

MSRC [44]: The image dataset contains 7 classes. Three types of features are extracted, i.e., CENT (1302 dimensions) [45], COLOR (48 dimensions) [46] and GIST

[1] http://vision.ucsd.edu/content/yale-face-database

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[3] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

TABLE II

Clustering Performances of Dimensionality Reduction Methods on **Yale Face** Dataset

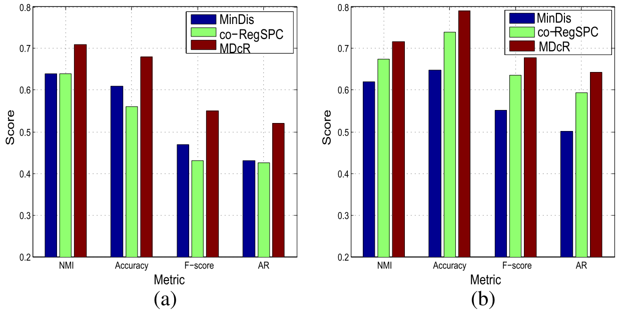
Feature	Method	NMI	Accuracy	F-score	AR
Single	PCA	0.588 ± 0.008	0.577 ± 0.014	0.390 ± 0.015	0.348 ± 0.013
	CCA	0.608 ± 0.006	0.581 ± 0.021	0.398 ± 0.017	0.360 ± 0.008
	NaMDR	0.608 ± 0.015	0.579 ± 0.016	0.419 ± 0.021	0.380 ± 0.013
	MDcR	**0.657 ± 0.013**	**0.628± 0.019**	**0.487± 0.018**	**0.441 ± 0.018**
Multiple	$\text{PCA}_{\text{FC}}^{1}$	0.642 ± 0.004	0.623 ± 0.001	0.453 ± 0.000	0.416 ± 0.000
	$\text{PCA}_{\text{FC}}^{2}$	0.656 ± 0.003	0.656 ± 0.009	0.468 ± 0.003	0.432 ± 0.003
	LPP_{FC}	0.633 ± 0.026	0.552 ± 0.014	0.433± 0.026	0.394 ± 0.027
	CCA*	0.652 ± 0.021	0.642 ± 0.015	0.445 ± 0.009	0.407 ± 0.007
	PLS*	0.697 ± 0.032	0.678 ± 0.042	0.543 ± 0.032	0.513 ± 0.035
	DSE*	0.662 ± 0.004	0.648 ± 0.010	0.511 ± 0.008	0.478 ± 0.009
	NaMDR	0.668 ± 0.020	0.650 ± 0.030	0.490 ± 0.025	0.465 ± 0.026
	MDcR*	**0.719 ± 0.010**	**0.688 ± 0.018**	**0.569 ± 0.016**	**0.540 ± 0.017**

TABLE III

Clustering Performances of Dimensionality Reduction Methods on **Notting-Hill** Dataset

Feature	Method	NMI	Accuracy	F-score	AR
Single	PCA	0.495 ± 0.000	0.633 ± 0.000	0.541 ± 0.000	0.418 ± 0.000
	CCA	0.372 ± 0.000	0.601 ± 0.000	0.483 ± 0.000	0.343 ± 0.000
	NaMDR	0.543 ± 0.000	0.645 ± 0.000	0.582 ± 0.000	0.469 ± 0.000
	MDcR	**0.588 ± 0.000**	**0.654 ± 0.000**	**0.614 ± 0.000**	**0.508 ± 0.000**
Multiple	$\text{PCA}_{\text{FC}}^{1}$	0.395 ± 0.000	0.547 ± 0.000	0.435 ± 0.000	0.282 ± 0.000
	$\text{PCA}_{\text{FC}}^{2}$	0.405 ± 0.000	0.561 ± 0.000	0.453 ± 0.000	0.290 ± 0.000
	LPP_{FC}	0.443 ± 0.021	0.516 ± 0.029	0.491 ± 0.023	0.343 ± 0.035
	CCA*	0.422 ± 0.000	0.629 ± 0.000	0.629 ± 0.000	0.393 ± 0.000
	PLS*	0.524 ± 0.000	0.698 ± 0.000	0.611 ± 0.000	0.500 ± 0.000
	DSE*	0.645 ± 0.000	**0.703 ± 0.000**	0.667 ± 0.000	0.577 ± 0.000
	NaMDR	0.603 ± 0.000	0.661 ± 0.000	0.608 ± 0.000	0.503 ± 0.000
	MDcR*	**0.709 ± 0.000**	0.694 ± 0.000	**0.695 ± 0.000**	**0.611 ± 0.000**

Fig. 2. Example images of different datasets.



(a) (b)

Fig. 3. Comparison between the multi-view clustering methods [4], [51] and ours. Note that, since the methods in [4] and [51] are two multi-view clustering methods (not used for dimensionality reduction), we only compared the clustering performance. (a) Yale Face dataset. (b) UCI Multiple Features dataset.

(512 dimensions) [47]. Example images of these datasets are shown in Fig. 2.

We compare our method with the following baselines:

- **PCA** [9]: We provide two ways of utilizing PCA for multi-view setting. The first is $\text{PCA}_{\text{FC}}^{1}$, which employs PCA to reduce dimensionality for each view independently, and then concatenates all the low-dimensional views together. The other is $\text{PCA}_{\text{FC}}^{2}$, which concatenates all the views firstly and then employs PCA to reduce dimensionality to a low-dimensional space.
- **CCA*** [10]: Multi-view dimensionality reduction method based on CCA constraint learns a common subspace to compare different views.
- **TCCA*** [11]: The method generalizes CCA for arbitrary number of views by analyzing the covariance tensor of the different views. We conduct this method on three out of five datasets due to its rather high computational cost.
- **PLS*** [13]: The Partial Least Squares (PLS) maps different views to a common linear subspace. We use the best two views since the method can only deal with the 2-view case.
- **DSE*** [15]: The Distributed Spectral Embedding maps different views to a common linear subspace.
- **NaMDR:** We also provide a naive multi-view dimensionality reduction method in Eq. (4) as a baseline, which

TABLE IV

Clustering Performances of Dimensionality Reduction Methods on **UCI Multiple Features** Dataset

Feature	Method	NMI	Accuracy	F-score	AR
Single	PCA	0.506 ± 0.002	0.563 ± 0.001	0.439 ± 0.001	0.393 ± 0.002
	CCA	0.386 ± 0.015	0.445 ± 0.026	0.346 ± 0.017	0.268 ± 0.020
	NaMDR	0.532 ± 0.016	0.570 ± 0.007	0.463 ± 0.001	0.416 ± 0.001
	MDcR	**0.601 ± 0.006**	**0.651 ± 0.001**	**0.519 ± 0.005**	**0.553 ± 0.001**
Multiple	$\text{PCA}_{\text{FC}}^{1}$	0.565 ± 0.001	0.647 ± 0.001	0.513 ± 0.002	0.485 ± 0.001
	$\text{PCA}_{\text{FC}}^{2}$	0.569 ± 0.001	0.648 ± 0.001	0.516 ± 0.000	0.462 ± 0.000
	LPP_{FC}	0.590 ± 0.020	0.576 ± 0.037	0.518 ± 0.026	0.463± 0.029
	CCA^*	0.427 ± 0.002	0.494 ± 0.021	0.389 ± 0.004	0.320 ± 0.020
	TCCA^*	0.613 ± 0.020	0.666 ± 0.040	0.559 ± 0.028	0.510 ± 0.031
	PLS^*	0.588 ± 0.002	0.707 ± 0.005	0.558 ± 0.003	0.509 ± 0.004
	DSE^*	0.518 ± 0.001	0.575 ± 0.001	0.485 ± 0.001	0.428 ± 0.000
	NaMDR	0.582 ± 0.001	0.632 ± 0.002	0.524 ± 0.021	0.471 ± 0.002
	MDcR^*	**0.643 ± 0.001**	**0.732 ± 0.001**	**0.612 ± 0.003**	**0.559 ± 0.001**

TABLE V

Clustering Performances of DR Methods on **Still DB** Dataset

Feature	Method	NMI	Accuracy	F-score	AR
Single	PCA	0.045 ± 0.000	0.257 ± 0.000	0.194 ± 0.000	0.024 ± 0.000
	CCA	0.026 ± 0.015	0.245 ± 0.026	0.175 ± 0.017	0.008 ± 0.020
	NaMDR	0.073 ± 0.000	**0.289 ± 0.007**	0.210 ± 0.000	**0.046 ± 0.000**
	MDcR	**0.082 ± 0.000**	0.279 ± 0.000	**0.220 ± 0.001**	0.043 ± 0.000
Multiple	$\text{PCA}_{\text{FC}}^{1}$	0.069 ± 0.000	**0.316 ± 0.000**	0.206 ± 0.000	0.043 ± 0.000
	$\text{PCA}_{\text{FC}}^{2}$	0.070 ± 0.000	0.290 ± 0.000	0.198 ± 0.000	0.0411 ± 0.000
	LPP_{FC}	0.218 ± 0.000	0.290 ± 0.009	0.215 ± 0.004	0.001 ± 0.000
	CCA^*	0.039 ± 0.002	0.248 ± 0.021	0.185 ± 0.004	0.015 ± 0.020
	TCCA^*	0.055 ± 0.002	0.265 ± 0.003	0.207 ± 0.005	0.037 ± 0.004
	PLS^*	0.058 ± 0.000	0.280 ± 0.000	0.210 ± 0.000	0.039 ± 0.000
	DSE^*	0.076 ± 0.000	0.265 ± 0.000	0.209 ± 0.000	0.038 ± 0.000
	NaMDR	0.089 ± 0.000	0.312 ± 0.000	0.221 ± 0.000	0.059 ± 0.000
	MDcR^*	**0.122 ± 0.000**	0.301 ± 0.000	**0.233 ± 0.000**	**0.074 ± 0.000**

TABLE VI

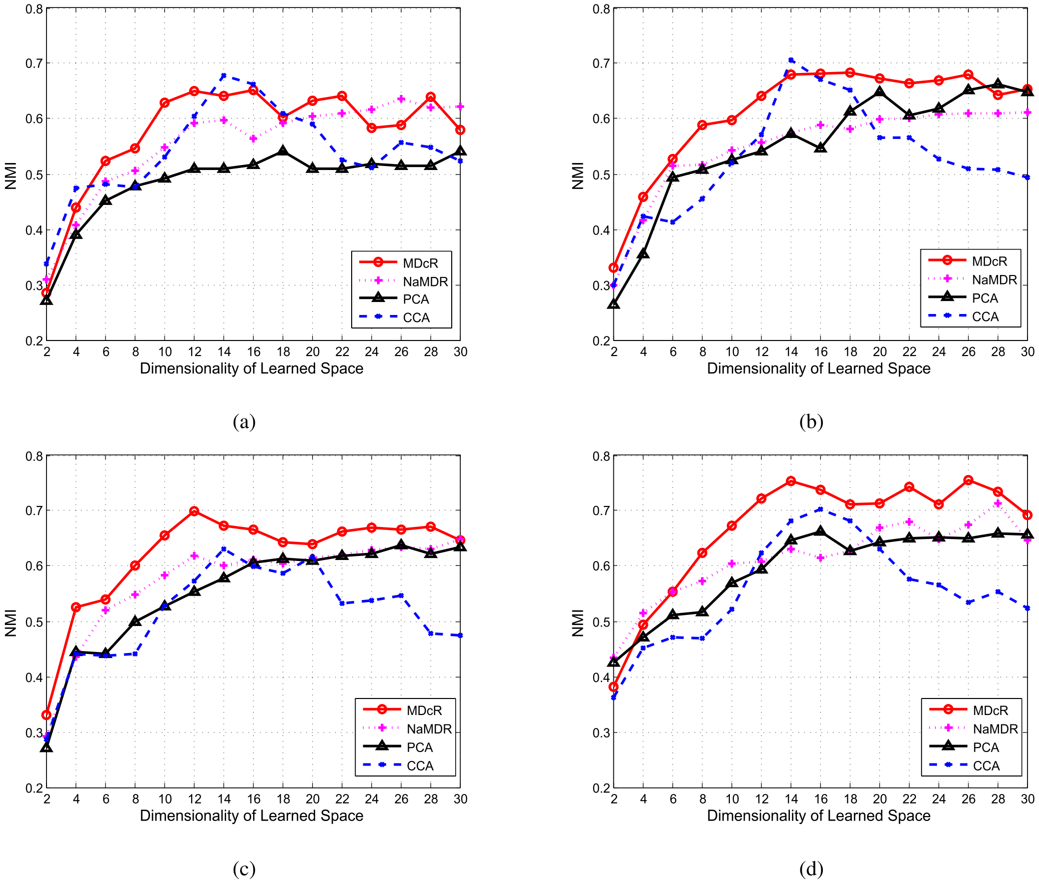
Clustering Performances of Dimensionality Reduction Methods on MSRC Dataset

Feature	Method	NMI	Accuracy	F-score	AR
Single	PCA	0.274 ± 0.001	0.418 ± 0.060	0.282 ± 0.001	0.162 ± 0.001
	CCA	0.154 ± 0.006	0.327 ± 0.002	0.204 ± 0.001	0.007 ± 0.001
	NaMDR	0.413 ± 0.001	0.488 ± 0.007	0.390 ± 0.000	0.286 ± 0.000
	MDcR	**0.488 ± 0.003**	**0.563 ± 0.000**	**0.457 ± 0.001**	**0.364 ± 0.000**
Multiple	$\text{PCA}_{\text{FC}}^{1}$	0.586 ± 0.027	0.738 ± 0.044	0.588 ± 0.030	0.522 ± 0.035
	$\text{PCA}_{\text{FC}}^{2}$	0.569 ± 0.033	0.720 ± 0.054	0.570 ± 0.037	0.509 ± 0.044
	LPP_{FC}	0.619 ± 0.022	0.750 ± 0.020	0.593 ± 0.022	0.526 ± 0.025
	CCA^*	0.126 ± 0.002	0.295 ± 0.021	0.185 ± 0.004	0.053 ± 0.020
	TCCA^*	0.135 ± 0.017	0.370 ± 0.023	0.198 ± 0.011	0.055 ± 0.012
	PLS^*	0.445 ± 0.004	0.559 ± 0.023	0.442 ± 0.014	0.349 ± 0.015
	DSE^*	0.614 ± 0.000	0.681 ± 0.000	0.598 ± 0.000	0.533 ± 0.000
	NaMDR	0.732 ± 0.009	0.823 ± 0.007	0.707 ± 0.001	0.660 ± 0.006
	MDcR^*	**0.736 ± 0.002**	**0.852 ± 0.001**	**0.735 ± 0.008**	**0.692 ± 0.002**

independently reduces the dimensionality of each view without any correlation constraint between different views.

Note that, PCA and NaMDR only explore the correlation within each view independently, while the methods marked with $*$ explore the correlations of different views.

For Yale, ORL and Notting-Hill, the dimensionalies of the three views are 4096, 3304 and 6750, respectively. The dimensionalites of three views for Still DB are all 200. We reduce the dimensionality of each view to 20 for all the methods. For UCI Multiple Features dataset, since the dimensionalities of its three views are significantly unbalanced (e.g., morphological features only with 6 dimensions), we reduce the other two views to 20, while preserve morphological features with 6 dimensions for all methods except CCA. CCA learns a common space with dimensionality of 6 on UCI Multiple Features dataset because the dimensionality



(a)

(b)

(c)

(d)

Fig. 4. Clustering performance with respect to varying dimensionality. (a) 1st View. (b) 2nd View. (c) 3rd View. (d) Concatenation.

of the common subspace must be equal or less than the lowest dimensionality of the multiple views. After obtaining the low-dimensional representations, we perform standard spectral clustering algorithm [48] on them to obtain the final clustering results. With the low-dimensional representations after dimensionality reduction, we report performances of two different settings: the average performance of three single views (**SingleView**) and the performance with concatenated views (**MultipleView**). Specifically, after dimensionality reduction, we performed clustering/classification for each single view, and reported the average performance of different views. The deviations reported for single view are also in the same manner (i.e., average of these single view deviations).

A. Clustering Experiment

For clustering, four evaluation metrics are employed: normalized mutual information (NMI), accuracy (ACC), adjusted rand index (AR), and F-score [49], [50]. The higher value indicates better clustering quality for all the four metrics. We report results on these diverse measures to perform a comprehensive evaluation. The final step of the standard spectral clustering [48] is k-means, hence we run it 30 times to report the average value and standard deviation.

The clustering performances are shown in the Tables II-VI, where the multiple view performances (**MultipleView**) are better than the independent view (**SingleView**). This demonstrates that multiple views can improve the results of clustering, even without any correlation constraint (e.g., NaMDR). Our MDcR method achieves the best results in terms different metrics on all the datasets. Furthermore, MDcR improves NaMDR obviously in terms of all the evaluation metrics, which validates the effectiveness of exploring correlations among different views via kernel matching constraint. Furthermore, we give insight investigation for the reason of promising performance of our method. As shown in Fig. 4, in a big picture, our MDcR can achieve promising results on different views with different reduced dimensions and thus induce the superior multi-view results.

B. Recognition Experiment

For recognition task, we employ accuracy as our evaluation metric, which is widely used in face and image recognition. The kNN (where k=1) classifier is used for recognition experiment, which is similar to most existing works [52]. The accuracy is simply calculated by the ratio of the corrected recognized testing samples. For each dataset, we partition it into the gallery and probe sets with different numbers. For ease of representation, G_m/P_n means m images per class are randomly selected for training and the remaining n images are for testing. For each G_m/P_n, we repeat the experiments over 30 random splits and report the average score as well as the standard deviation. Note that, for Notting-Hill and Still DB, the sample numbers of different classes are not equal, thus given the same size of training samples for different classes,

TABLE VII

Recognition Performances of Dimensionality Reduction Methods on **Yale Face** Dataset

Feature	Method	G3/P8	G4/P7	G5/P6	G6/P5
Single	PCA	0.592 ± 0.019	0.637 ± 0.022	0.623 ± 0.029	0.633 ± 0.035
	CCA	0.675 ± 0.037	0.681 ± 0.044	0.708 ± 0.017	0.699 ± 0.040
	NaMDR	0.620 ± 0.015	0.669 ± 0.045	0.668 ± 0.036	0.689 ± 0.037
	MDcR	**0.703 ± 0.023**	**0.695 ± 0.014**	**0.701 ± 0.019**	**0.717± 0.031**
Multiple	PCA$_{\text{FC}}^{1}$	0.689 ± 0.018	0.725 ± 0.021	0.712 ± 0.024	0.734 ± 0.002
	PCA$_{\text{FC}}^{2}$	0.668 ± 0.012	0.716 ± 0.029	0.732 ± 0.021	0.741 ± 0.002
	LPP$_{\text{FC}}$	0.620 ± 0.026	0.647 ± 0.030	0.655 ± 0.037	0.700 ± 0.046
	CCA*	0.689 ± 0.012	0.716 ± 0.055	0.720 ± 0.026	0.727 ± 0.034
	PLS*	0.653 ± 0.027	0.693 ± 0.043	0.728± 0.026	0.696 ± 0.028
	DSE*	0.661 ± 0.025	0.655 ± 0.035	0.666 ± 0.028	0.672 ± 0.026
	NaMDR	0.695 ± 0.021	0.730 ± 0.028	0.733 ± 0.038	0.741 ± 0.021
	MDcR*	**0.721 ± 0.022**	**0.753 ± 0.022**	**0.763 ± 0.009**	**0.776 ± 0.009**

TABLE VIII

Recognition Performances of Dimensionality Reduction Methods on **Notting-Hill** Dataset

Feature	Method	G10	G20	G30	G40
Single	PCA	0.904 ± 0.023	0.940 ± 0.014	0.962 ± 0.005	0.963 ± 0.012
	CCA	0.749 ± 0.048	0.833 ± 0.015	0.862 ± 0.025	0.881 ± 0.023
	NaMDR	0.932 ± 0.014	0.950 ± 0.017	0.971 ± 0.008	0.974 ± 0.004
	MDcR	**0.941 ± 0.017**	**0.973 ± 0.005**	**0.982 ± 0.003**	**0.981 ± 0.002**
Muliple	PCA$_{\text{FC}}^{1}$	0.950 ± 0.024	0.973 ± 0.016	0.991 ± 0.005	0.985 ± 0.010
	PCA$_{\text{FC}}^{2}$	0.951 ± 0.025	0.970 ± 0.020	0.989 ± 0.007	0.983 ± 0.009
	LPP$_{\text{FC}}$	0.769 ± 0.037	0.870 ± 0.029	0.913 ± 0.019	0.932 ± 0.021
	CCA*	0.765 ± 0.045	0.846 ± 0.021	0.879 ± 0.025	0.892 ± 0.024
	PLS*	0.930 ± 0.028	0.970 ± 0.007	0.975 ± 0.008	0.981 ± 0.004
	DSE*	0.919 ± 0.042	0.930 ± 0.025	0.951 ± 0.007	0.950 ± 0.011
	NaMDR	0.957 ± 0.013	0.976 ± 0.019	0.985 ± 0.010	0.989 ± 0.006
	MDcR*	**0.972 ± 0.020**	**0.989 ± 0.009**	**0.994 ± 0.004**	**0.996 ± 0.003**

TABLE IX

Recognition Performances of Dimensionality Reduction Methods on **UCI Multiple Features** Dataset

Feature	Method	G10/P190	G20/P180	G30/P170	G50/P150
Single	PCA	0.557 ± 0.015	0.570 ± 0.030	0.577 ± 0.022	0.592 ± 0.032
	CCA	0.198 ± 0.028	0.201 ± 0.041	0.198 ± 0.017	0.189 ± 0.027
	NaMDR	0.667 ± 0.022	0.706 ± 0.024	0.713 ± 0.013	0.718 ± 0.026
	MDcR	**0.712± 0.001**	**0.749 ± 0.007**	**0.762 ± 0.036**	**0.789 ± 0.021**
Muliple	PCA$_{\text{FC}}^{1}$	0.867 ± 0.017	0.915 ± 0.019	0.918 ± 0.011	0.919 ± 0.030
	PCA$_{\text{FC}}^{2}$	0.883 ± 0.013	0.918 ± 0.003	0.926± 0.006	0.938 ± 0.002
	LPP$_{\text{FC}}$	0.844 ± 0.013	0.888 ± 0.007	0.905 ± 0.007	0.929 ± 0.006
	CCA*	0.261 ± 0.003	0.282 ± 0.029	0.257 ± 0.038	0.264 ± 0.032
	TCCA*	0.724 ± 0.023	0.745 ± 0.009	0.752 ± 0.009	0.760 ± 0.014
	PLS*	0.709 ± 0.040	0.759 ± 0.010	0.775 ± 0.017	0.798 ± 0.006
	DSE*	0.523 ± 0.092	0.450 ± 0.089	0.419 ± 0.103	0.346 ± 0.064
	NaMDR	0.885 ± 0.030	0.898 ± 0.025	0.891 ± 0.031	0.918 ± 0.017
	MDcR*	**0.912 ± 0.012**	**0.939 ± 0.021**	**0.946 ± 0.009**	**0.954 ± 0.030**

the test sample numbers are not the same. The results are shown in Tables VII-XI. Our MDcR steadily outperforms the other methods with a significant margin. Moreover, without exploring the relationships within each view itself, the performances of multi-view methods are unstable. For example, CCA achieves promising results on the Yale Face dataset while the performance degrades sharply on the UCI Multiple Features dataset.

Overall, according to the experimental results, we have the following observations. First, directly concatenating all the views is not reasonable since the performance of PCA$_{\text{FC}}^{2}$ is not such promising. Second, reducing each view independently is not sufficient, since the performances of PCA$_{\text{FC}}^{1}$ and NaMDR are obviously lower than ours. Third, although CCA and PLS explore the correlations between different views, they ignore exploring the relationship within individual view. Our method simultaneously explores the relationship (i.e., graph regularization for smoothness) within each view and correlations (i.e., HSIC for dependence) across views to produces more promising results. Finally, for the views with unbalanced dimensionalities, it is unreasonable to project all views to the common low-dimensional space. For example, CCA failed on

TABLE X

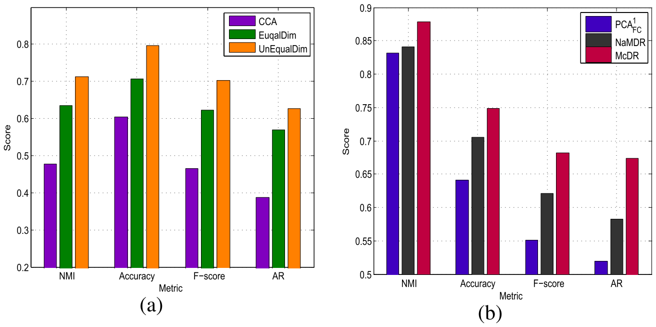
Recognition Performances of Dimensionality Reduction Methods on **Still DB** Dataset

Feature	Method	G10	G20	G30	G40
Single	PCA	0.255 ± 0.021	0.283 ± 0.031	0.280 ± 0.030	0.298 ± 0.022
	CCA	0.216 ± 0.023	0.229 ± 0.030	0.198 ± 0.017	0.249 ± 0.022
	NaMDR	0.269 ± 0.016	0.270 ± 0.015	0.283 ± 0.027	0.298 ± 0.027
	MDcR	**0.274** ± **0.027**	**0.283** ± **0.023**	**0.287** ± **0.034**	**0.293** ± **0.016**
Muliple	PCA$_{\text{FC}}^{1}$	0.305 ± 0.034	0.317 ± 0.034	0.322 ± 0.009	0.347 ± 0.027
	PCA$_{\text{FC}}^{2}$	0.302 ± 0.032	0.304 ± 0.035	0.316 ± 0.019	0.335 ± 0.030
	LPP$_{\text{FC}}$	0.305 ± 0.033	0.319 ± 0.019	0.335± 0.024	0.343 ± 0.031
	CCA*	0.239 ± 0.031	0.242± 0.028	0.257 ± 0.038	0.260 ± 0.032
	TCCA*	0.188 ± 0.016	0.190 ± 0.022	0.192 ± 0.023	0.202 ± 0.027
	PLS*	0.245 ± 0.033	0.242 ± 0.041	0.243 ± 0.029	0.273 ± 0.033
	DSE*	0.268 ± 0.025	0.253 ± 0.025	0.266 ± 0.048	0.265 ± 0.022
	NaMDR	0.303 ± 0.0188	0.313 ± 0.030	0.314 ± 0.027	0.336 ± 0.024
	MDcR*	**0.322** ± **0.031**	**0.325** ± **0.025**	**0.342** ± **0.029**	**0.360** ± **0.030**

TABLE XI

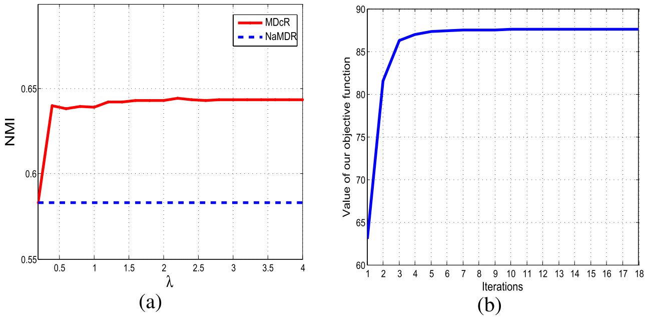
Recognition Performances of Dimensionality Reduction Methods on MSRC Dataset

Feature	Method	G4	G8	G12	G16
Single	PCA	0.522 ± 0.016	0.530 ± 0.027	0.540 ± 0.019	0.563 ± 0.038
	CCA	0.385 ± 0.037	0.430 ± 0.026	0.441 ± 0.044	0.457 ± 0.039
	NaMDR	0.528 ± 0.027	0.569 ± 0.034	0.596 ± 0.034	0.608 ± 0.027
	MDcR	**0.536** ± **0.037**	**0.568** ± **0.034**	**0.601** ± **0.035**	**0.618** ± **0.016**
Muliple	PCA$_{\text{FC}}^{1}$	0.732 ± 0.042	0.713 ± 0.029	0.738 ± 0.051	0.740 ± 0.003
	PCA$_{\text{FC}}^{2}$	0.724 ± 0.035	0.710 ± 0.031	0.732 ± 0.041	0.736 ± 0.003
	LPP$_{\text{FC}}$	0.506 ± 0.034	0.581 ± 0.052	0.637 ± 0.029	0.654 ± 0.057
	CCA*	0.472 ± 0.031	0.537 ± 0.032	0.559 ± 0.041	0.613 ± 0.014
	TCCA*	0.443 ± 0.044	0.498 ± 0.033	0.510 ± 0.04	0.512 ± 0.040
	PLS*	0.502 ± 0.090	0.489 ± 0.051	0.530 ± 0.062	0.511 ± 0.050
	DSE*	0.729 ± 0.038	0.763 ± 0.039	0.746 ± 0.030	0.762 ± 0.026
	NaMDR	0.754 ± 0.030	0.813 ± 0.027	0.838 ± 0.018	0.838 ± 0.022
	MDcR*	**0.779** ± **0.025**	**0.827** ± **0.031**	**0.854** ± **0.027**	**0.851** ± **0.034**



(a) (b)

Fig. 5. Results on multi-view data with unbalanced dimensionalies (a) and results of learned projections for new data (b).



(a) (b)

Fig. 6. Parameter tuning in terms of NMI performance (a); Objective function value convergence curve (b).

the UCI Multiple Features dataset due to its limitation of common space for all views.

C. Model Flexibility and Transferring Ability

For the case that the dimensionalities of different views are unbalanced (e.g., UCI Multiple Features dataset), we compare three settings, i.e., CCA with the best subspace dimensionality ($k = 4$), MDcR with equal dimensionality ($k = 6$) of subspaces (*EqualDim*) and MDcR with different subspace dimensionalities (i.e., $k^{(1)} = k^{(2)} = 12$ and $k^{(3)} = 6$) (*UnEqualDim*). The clustering results, as shown in Fig. 5(a), indicate that our method with the flexible setting are more reasonable.

Generally, the projection-based dimensionality reduction methods could apply the learned low-dimensional projections for new data. To evaluate this issue, we employ the *ORL* dataset as the new data to test the learned low-dimensional projections from Yale Face dataset. As shown in Fig. 5(b) for clustering, our method obtains an obvious improvement on NaMDR about 4% and 5% in terms of NMI and accuracy, respectively. This further demonstrates the importance of exploring correlations among different views.

D. Parameter Tuning and Convergence

Fig. 6 shows the parameter tuning and convergence curves of our objective function on the UCI Multiple Features dataset. $\lambda \geq 0$ is an essential parameter in our MDcR approach, which controls the strength of the dependencies across different views. Fig. 6(a) shows the clustering results of our method with different values of parameter λ. The performance is basically stable and the significant improvement could be expected compared to NaMDR while $\lambda > 0.5$. We also test the convergence rate of our method. Empirically, our algorithm converges fast in less than 3 iterations, as shown in Fig. 6(b).

V. Conclusions

We have proposed a multi-view dimensionality reduction method, Multi-view Dimensionality co-Reduction (MDcR), which employs the HSIC to explicitly enforce the dependence across different views. We have shown that our method is more general and effective than the existing co-regularized method. Extensively empirical study suggests that the proposed approach can effectively explore the underlying complementary information of the given multi-view data, and outperforms the compared dimensionality reduction methods. By incorporating nonlinear universal kernel, the optimization is not the eigenvalue decomposition problem. One solution is applying the gradient descent method to update each view in each iteration, which is computationally expensive. Nevertheless, adopting nonlinear universal kernels is quite interesting for addressing more general correlations and we will consider it in our future work.

References

[1] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1150–1157.

[2] M. Lades *et al.*, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.*, vol. 42, no. 3, pp. 300–311, Mar. 1993.

[3] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[4] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.

[5] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2392–2398.

[6] Y. Yang, H.-J. Ye, D.-C. Zhan, and Y. Jiang, "Auxiliary information regularized machine for multiple modality feature learning," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 1033–1039.

[7] J. Yu, M. Wang, and D. Tao, "Semisupervised multiview distance metric learning for cartoon synthesis," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4636–4648, Nov. 2012.

[8] H. Zhao and Y. Fu, "Dual-regularized multi-view outlier detection," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 4077–4083.

[9] I. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2002.

[10] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[11] 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.

[12] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang, "Sparse unsupervised dimensionality reduction for multiple view data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 10, pp. 1485–1496, Oct. 2012.

[13] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 593–600.

[14] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.

[15] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.

[16] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[17] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[18] T. Zhou and D. Tao, "Double shrinking sparse dimension reduction," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 244–257, Jan. 2013.

[19] C. Xu, D. Tao, and C. Xu. (2013). "A survey on multi-view learning." [Online]. Available: https://arxiv.org/abs/1304.5634

[20] A. Kumar and H. Daumé, III, "A co-training approach for multi-view spectral clustering," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 393–400.

[21] M. White, X. Zhang, D. Schuurmans, and Y.-L. Yu, "Convex multi-view subspace learning," in *Proc. 25th Adv. Neural Inf. Process. Syst.*, 2012, pp. 1673–1681.

[22] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.

[23] D. P. Foster, S. M. Kakade, and T. Zhang, "Multi-view dimensionality reduction via canonical correlation analysis," Dept. Statist., Pennsylvania Univ., Philadelphia, PA, USA, Tech. Rep. TTI-TR-2008-4, 2008.

[24] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, vol. 11. Newbury Park, CA, USA: Sage, 1978.

[25] D. Engel, L. Hüttenberger, and B. Hamann, "A survey of dimension reduction methods for high-dimensional data analysis and visualization," in *OASIcs—OpenAccess Series in Informatics*, vol. 27. Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[26] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[27] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 47–55.

[28] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 414–427, Jan. 2016.

[29] N. Chen, J. Zhu, and E. P. Xing, "Predictive subspace learning for multi-view data: A large margin approach," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 361–369.

[30] N. Quadrianto and C. H. Lampert, "Learning multi-view neighborhood preserving projections," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 425–432.

[31] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. Int. Conf. Data Mining*, 2004, pp. 19–26.

[32] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace Clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 586–594.

[33] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 1582–1590.

[34] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Proc. Subspace, Latent Struct. Feature Selection*, 2006, pp. 34–51.

[35] M. Xiao and Y. Guo, "Feature space independent semi-supervised domain adaptation via kernel matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 54–66, Jan. 2015.

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

[37] N. Quadrianto, L. Song, and A. J. Smola, "Kernelized sorting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1289–1296.

[38] D. Niu, J. G. Dy, and M. I. Jordan, "Iterative discovery of multiple alternativeclustering views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1340–1353, Jul. 2014.

[39] 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.

[40] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, “Graph regularized transductive classification on heterogeneous information networks,” in *Proc. ECMLPKDD*, 2010, pp. 570–586.

[41] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, “Learning with local and global consistency,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 321–328.

[42] D. Niu, J. G. Dy, and M. I. Jordan, “Multiple non-redundant spectral clustering views,” in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 831–838.

[43] B. Wu, Y. Zhang, B. G. Hu, and Q. Ji, “Constrained clustering and its application to face clustering in videos,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3507–3514.

[44] J. Winn and N. Jojic, “LOCUS: Learning object classes with unsupervised segmentation,” in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1. Oct. 2005, pp. 756–763.

[45] J. Wu and J. M. Rehg, “Where am I: Place instance and category recognition using spatial PACT,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[46] H. Yu, M. Li, H.-J. Zhang, and J. Feng, “Color texture moments for content-based image retrieval,” in *Proc. Int. Conf. Image Process.*, vol. 3. Jun. 2002, pp. 929–932.

[47] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[48] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2. 2001, pp. 849–856.

[49] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 2008.

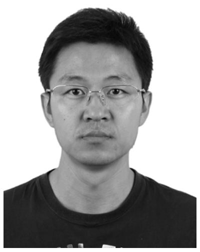
[50] D. Cai, X. He, and J. Han, “Document clustering using locality preserving indexing,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[51] V. R. de Sa, “Spectral clustering with two views,” in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 20–27.

[52] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, “Learning a spatially smooth subspace for face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

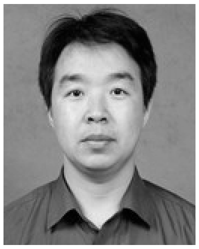
[53] P. Huang, Y. Wang, and M. Shao, “A new method for multi-view face clustering in video sequence,” in *Proc. Int. Conf. Data Mining Workshops*, 2008, pp. 869–873.

[54] J. Lei, C. Zhang, Y. Fang, Z. Gu, N. Ling, and C. Hou, “Depth sensation enhancement for multiple virtual view rendering,” *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 457–469, 2015.



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.

Huazhu Fu received the B.S. degree in mathematical sciences from Nankai University in 2006, the M.E. degree in mechatronics engineering from the Tianjin University of Technology in 2010, and the Ph.D. degree in computer science from Tianjin University, China, in 2013. He was a Research Fellow with Nanyang Technological University, Singapore. He is currently a Research Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His research interests include computer vision, image processing, and medical image analysis.



Qinghua Hu (SM’13) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively. He was a Post-Doctoral Fellow with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, from 2009 to 2011, and now he is a Full Professor with School of Computer Science and Technology, Tianjin University. He has authored over 150 journal and conference papers in the areas of granular computing-based machine learning, reasoning with uncertainty, pattern recognition, and fault diagnosis. His current research interests include multimodality learning, metric learning, and uncertainty modeling and reasoning with fuzzy sets, rough sets and probability theory. He was the Program Committee Co-Chair of the International Conference on Rough Sets and Current Trends in Computing in 2010, the Chinese Rough Set and Soft Computing Society in 2012 and 2014, and the International Conference on Rough Sets and Knowledge Technology and the International Conference on Machine Learning and Cybernetics in 2014, and the General Co-Chair of IJCRS 2015. He is the PC Co-Chair of CCML 2017 and CCCV 2017.

Pengfei Zhu (M’15) received the B.S. and M.S. degrees from the Harbin Institute of Technology, Harbin, China, in 2009 and 2011, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2015. He is currently an Associate Professor with the School of Computer Science and Technology, Tianjin University. His research interests are focused on machine learning and computer vision.

Xiaochun Cao (SM’14) received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, since 2012. He spent about three years with ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and co-authored over 120 journal and conference papers. He is a fellow of the IET. He is on the Editorial Board of the IEEE Transactions of Image Processing. His dissertation was nominated for the University of Central Floridas university-level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.