Multiview Latent Space Learning With Feature Redundancy Minimization

Tao Zhou, *Member, IEEE*, Changqing Zhang, Chen Gong, *Member, IEEE*, Harish Bhaskar, and Jie Yang

***Abstract*—Multiview learning has received extensive research interest and has demonstrated promising results in recent years. Despite the progress made, there are two significant challenges within multiview learning. First, some of the existing methods directly use original features to reconstruct data points without considering the issue of feature redundancy. Second, existing methods cannot fully exploit the complementary information across multiple views and meanwhile preserve the view-specific properties; therefore, the degraded learning performance will be generated. To address the above issues, we propose a novel multiview latent space learning framework with feature redundancy minimization. We aim to learn a latent space to mitigate the feature redundancy and use the learned representation to reconstruct every original data point. More specifically, we first project the original features from multiple views onto a latent space, and then learn a shared dictionary and view-specific dictionaries to, respectively, exploit the correlations across multiple views as well as preserve the view-specific properties. Furthermore, the Hilbert–Schmidt independence criterion is adopted as a diversity constraint to explore the complementarity of multiview representations, which further ensures the diversity from multiple views and preserves the local structure of the data in each view. Experimental results on six public datasets have demonstrated the effectiveness of our multiview learning approach against other state-of-the-art methods.**

***Index Terms*—Complementary information, Hilbert–Schmidt independence criterion (HSIC), latent space, multiview learning, redundancy minimization.**

Manuscript received July 22, 2018; revised November 3, 2018; accepted November 18, 2018. Date of publication December 14, 2018; date of current version February 25, 2020. This work was supported in part by the China Post-Doctoral Science Foundation under Grant 2016M601597, in part by NSFC of China under Grant 61876107, Grant 6151101179, Grant 61602337, and Grant 61602246, in part by the 973 Plan of China under Grant 2015CB856004, in part by NSF of Jiangsu Province under Grant BK20171430, in part by the Fundamental Research Funds for the Central Universities under Grant 30918011319, in part by the Open Project of State Key Laboratory of Integrated Services Networks (Xidian University, ID: ISN19-03), in part by the "Summit of the Six Top Talents" Program under Grant DZXX-027, in part by the "Lift Program for Young Talents" of Jiangsu Province, and in part by the "CAST Lift Program for Young Talents." This paper was recommended by Associate Editor D. Tao. *(Corresponding authors: Jie Yang; Chen Gong.)*

T. Zhou is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Inception Institute of Artificial Intelligence, Abu Dhabi 51133, UAE (e-mail: taozhou17@gmail.com).

C. Zhang is with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China (e-mail: zhangchangqing@tju.edu.cn).

C. Gong is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China (e-mail: chen.gong@njust.edu.cn).

H. Bhaskar is with Zero One Infinity Consulting Service Ltd., Mississauga, ON L5N6G9, Canada (e-mail: harishbhasky@gmail.com).

J. Yang is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jieyang@sjtu.edu.cn).

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

Digital Object Identifier 10.1109/TCYB.2018.2883673

I. Introduction

RECENTLY, multiview (or multimodal) learning has attracted significant attention as it is able to characterize an object via harnessing the diverse information from multiple sources in many real-world applications [1]–[7]. For example, an image can be described by using different features, such as SIFT, LBP, HOG, etc. The news on the Internet usually consists of texts, images, and videos. Sufficient research results have demonstrated that the model performance can be substantially improved by combining multiple views of data. This is mainly because different views depict different perspectives of the data which further provide complementary information for data description and model training. Thus, one key challenge for multiview learning is how to effectively integrate the information of multiple views and exploit the underlying structures within data to obtain the improved performance.

A number of multiview learning approaches have been proposed in the last decade. Co-training [8] is one of the earliest multiview learning schemes, which are alternately trained based on the unlabeled data of two distinct views to maximize their agreement. After that, many of its variants have also been developed and obtained promising results [9], [10]. Next, some multiview methods are proposed such that the data of multiple views can be projected into a common space. For example, canonical correlation analysis (CCA) [11] and its related variants (e.g., multiview CCA [12]) learn multiple projections to map multiview data into a common space. Distributed spectral embedding [13] learns a low-dimensional and sufficiently smooth embedding over all views simultaneously. Partial least squares (PLS) [14] projects different views into a common linear subspace by using a standard regression methodology. Multiple kernel learning (MKL)-based approaches [15], [16] learn a low-dimensional common representation across multiple views. Besides, non-negative matrix factorization (NMF)-based multiview learning methods have also attracted wide attention. For example, multi-NMF [17] formulates a joint multiview NMF learning framework, which encourages the representations of all views to be compromised to a common result.

Additionally, as an important branch of multiview learning, the clustering methods for multiview data have also been widely developed. Such approaches can be divided into three main types. First, graph-based methods [18], [19] exploit the correlations across different views via using the multiple graph fusion strategies. The methods of second type are usually based on co-training and co-regularization [20], [21]. Finally, the third category relies on subspace learning [22]–[25], which aims to find an underlying low-dimensional subspace for representing data points rather than assuming that they are distributed uniformly across the entire feature space. Specifically, the work of [22] proposes an iterative strategy to achieve multiview spectral clustering by minimizing the divergence of the latent data-clustered representation for each view. The work of [23] presents a Markov chain method for multiview clustering, which recovers a shared transition probability matrix via low-rank and sparse decomposition. Gao *et al*. [24] performed subspace clustering on each individual modality and then integrates these modalities into a common indicator matrix. Besides, the work [25] assumes that multiple views originate from one underlying latent representation, and then clustering is performed on such latent data representation.

In general, most of the existing approaches suffer from the following challenges. First, in the real-world applications, the original high-dimensional data often contain feature redundancy, which makes the relationships among different examples not be accurately depicted in the original feature space. As a result, the performances of existing multiview learning algorithms still need further improvements. Second, it also remains a challenging problem on how to simultaneously capture the correlation across multiple views and exploit the diversity within each individual view to achieve better multiview learning performance.

To address the challenges above, in this paper, we present a novel multiview latent space learning framework with feature redundancy minimization (FRM). The proposed method minimizes redundancy by learning a latent space to render new data representations that accurately depict the relationships among different views. Within the latent space, our approach employs shared and specific dictionaries to capture both the consensus and particular information of different views. Moreover, the Hilbert–Schmidt independence criterion (HSIC) is introduced as a diversity constraint to enhance the complementarity of multiview representations, which further ensures the diversity from multiple views and preserves the local structure of data in each view. The basic flow of our proposed framework is depicted in Fig. 1. For algorithm validation, we compare the performances of our FRM methodology and other state-of-the-art multiview approaches on six benchmark datasets. Experimental results demonstrate the effectiveness and the feasibility of the proposed approach. The main contributions of this paper are summarized below.

1) We formulate a unified framework for multiview subspace learning by constructing a latent space with FRM.
2) Within the latent space, our approach simultaneously captures the correlations across multiple views as well as exploiting view-specific information from each view.

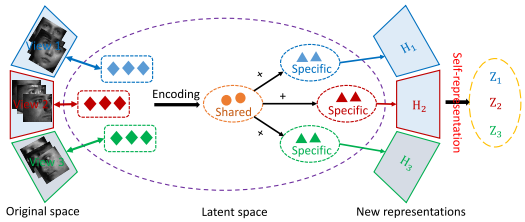


Fig. 1. Illustration of the proposed multiview latent space learning framework, where "+" denotes to cascade the shared and view-specific dictionaries together. First, we project the original features from multiview data into a latent space through redundancy minimization, and then the low-dimensional features are represented using shared and view-specific dictionaries. The newly encoded representations is then used for reconstructing the data points. Within the latent space, we learn both shared and specific dictionaries to exploit both the correlations across multiple views and preserve view-specific property, respectively. Note that, different colors in the cell mean different views, dotted rectangles represent the learned low-dimensional features, dotted ellipses represent the learned dictionaries, and double arrows denote feature projection and reconstruction.

3) To ensure that the new representations within latent space from different views can provide enhanced complementary information, we use HSIC to penalize the information redundancy among these new representations.
4) Our approach can reduce the feature redundancy by discovering the shared information and strengthening the representation diversity of various views.

The rest of this paper is organized as follows. Section II briefly reviews some related works. Section III presents the proposed method, optimization solution, and computation complexity analysis. Extensive experiments on benchmark datasets are conducted in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

In this section, we briefly review three related topics with this paper: 1) dictionary learning; 2) multiview learning; and 3) low-rank subspace learning.

A. Dictionary Learning

Given a data matrix $\mathbf{X} \in \mathbb{R}^{L\times N}$ with N data points and L features. Existing dictionary learning framework is defined as

$$\min_{\mathbf{D},\mathbf{H}} \|\mathbf{X}-\mathbf{D}\mathbf{H}\|_F^2 + \lambda\Psi(\mathbf{D},\mathbf{H}) \tag{1}$$

where λ is a tradeoff parameter, $\mathbf{D}$ and $\mathbf{H}$ denote a dictionary and the encoding coefficient matrix, respectively. $\Psi(\mathbf{D},\mathbf{H})$ is used to enforce characteristic properties on $\mathbf{D}$ and $\mathbf{H}$. Recently, dictionary learning has been extensively used in face recognition [26], [27], object classification [28], visual tracking [29], [30], and so on, which have been shown good performance. Besides, some studies have been developed to learn class-specific dictionary and shared dictionary to characterize the class-specific property and exploit the correlation among different classes, respectively [31]. Further, some classification techniques [26], [32] simultaneously consider to train a classifier by extending the above framework as follows:

$$\min_{\mathbf{D},\mathbf{H},\mathbf{W}} \|\mathbf{X}-\mathbf{D}\mathbf{H}\|_F^2 + \lambda_1\|\mathbf{Y}-\mathbf{W}\mathbf{H}\|_F^2 + \lambda_2\Psi(\mathbf{D},\mathbf{H}) \tag{2}$$

where λ_1 and λ_2 are tradeoff parameters, $\mathbf{Y}$ is the label matrix, and $\mathbf{W}$ is a linear projection matrix. In this paper, we integrate dictionary learning into multiview learning framework, and the encoding coefficient matrix can be regarded as new representations which are used to reconstruct the data points.

B. Multiview Learning

Multiview learning aims to exploit the relationships between different views to improve the performance. Since real-world data are often collected from multiple views, multiview learning has attracted widespread attention over the last decades and has been successfully applied to different applications, such as classification, clustering, dimensionality reduction, and so on. Existing multiview learning approaches can be divided into three main categories [33]: 1) co-training; 2) MKL; and 3) subspace learning. Specifically, co-training approaches aim to maximize the agreement between two distinct views from unlabeled data in a semisupervised manner. Following that, many research works have been developed by following the basic idea of co-training [9], [10]. MKL-based approaches [15], [16] aim to seek different kernels for different views and then combine them to process the training data. Subspace learning-based approaches aim to learn a latent subspace across different views, which assume that the input views are generated from the learned latent subspace. Currently, subspace learning-based multiview learning [25], [34] is quite popular as the learned subspace can be used to conduct both classification and clustering tasks.

C. Low-Rank Subspace Learning

Low-rank representation (LRR)-based methods [35] have become well known for its robustness to the noise/corrupted data, and have been demonstrated to be effective for tackling many machine learning tasks. The general model of LRR can be formulated as follows:

$$\min_{\mathbf{Z},\mathbf{E}} \ \text{rank}(\mathbf{Z}) + \lambda\|\mathbf{E}\|_p, \quad \text{s.t.} \ \ \mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E} \tag{3}$$

where $\mathbf{X}$ is a data matrix, $\mathbf{A}$ is a dictionary matrix that can linearly span the data space, $\mathbf{E}$ is a sparse additive error matrix, $\|\mathbf{E}\|_p$ denotes certain regularization strategy (such as the ℓ_1 and $\ell_{2,1}$ cases) to model the noise, and λ is a regularization parameter. Besides, rank($\mathbf{Z}$) denotes the rank of coefficients matrix $\mathbf{Z}$, however, rank minimization problem is in general NP hard, thus the trace norm $\|\mathbf{Z}\|_*$ can be adopted to be a surrogate of the rank of $\mathbf{Z}$. Further, the data matrix itself $\mathbf{X}$ is directly used as the dictionary, therefore, (3) can be reformulated as

$$\min_{\mathbf{Z},\mathbf{E}} \ \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_p, \quad \text{s.t.} \ \ \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}. \tag{4}$$

Based on LRR, a lot of studies [36]–[38] have been proposed to find a more robust subspace with low-rank constraint.

III. Proposed Method

In this section, we first present our novel multiview learning framework. Then, we design an optimization solution, and provide complexity analysis on our approach.

A. Formulation

Given a data set $\mathbf{X}_v \in \mathbb{R}^{L_v \times N}$, where $\mathbf{X}_v$ denotes the features matrix of the vth view ($v = 1, \ldots, V$), with L_v and N being the dimensionality of features from the vth view and the number of examples, respectively. Therefore, $\mathbf{X}_v$ can be represented by a learned dictionary, which is

$$\min_{\mathbf{D}_0,\mathbf{D}_v,\mathbf{H}_v} \sum_v \left\| \Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v \right\|_F^2 \tag{5}$$

where $\Theta_v \in \mathbb{R}^{L_v \times L_0}$ (L_0 denotes the dimensionality of latent space), is a linear transformation matrix that is used to link the original input space and the learned latent space. $\mathbf{D}_0 \in \mathbb{R}^{L_0 \times K_0}$ and $\mathbf{D}_v \in \mathbb{R}^{L_0 \times K_v}$ represent the shared dictionary and view-specific dictionary, respectively, and K_0 and K_v are the corresponding numbers of atoms in the two dictionaries $\mathbf{D}_0$ and $\mathbf{D}_v$, respectively. In addition, $\mathbf{H}_v$ is the new representations for the vth view in latent space that can be used for the subsequent clustering and classification.

Further, to enhance the complementarity of multiview representations, the encoded representations of different views are encouraged to be of sufficient diversity. Next, we briefly introduce a diversity regularization term. Let us define two kernel spaces $\mathcal{F}$ and $\mathcal{G}$, the inner product between vectors in the two spaces can then be given by the kernel functions, i.e., $k_1(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ and $k_2(\mathbf{y}_i, \mathbf{y}_j) = \langle \phi(\mathbf{y}_i), \phi(\mathbf{y}_j) \rangle$, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ are two different variable sets, and $\phi(\cdot)$ maps the original features to a kernel space. Following the work in [39], to avoid the direct estimation of an unknown joint distribution $p_{\mathbf{xy}}$ over the spaces $\mathcal{F}$ and $\mathcal{G}$, we utilize an empirical version of HSIC as a diversity term, of which the definition is below.

Definition 1: Given a set of N independent observations collected from the joint distribution p_{xy}, i.e., $\mathbf{Z} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\} \subseteq \mathcal{X} \times \mathcal{Y}$, then we can define an estimator of HSIC($\mathbf{Z}, \mathcal{F}, \mathcal{G}$) as

$$\text{HSIC}(\mathbf{Z}, \mathcal{F}, \mathcal{G}) = (N-1)^{-2}\text{tr}(\mathbf{K}_1\mathbf{H}\mathbf{K}_2\mathbf{H}) \tag{6}$$

where $k_{1,ij} \coloneqq k_1(\mathbf{x}_i, \mathbf{x}_j)$ and $k_{2,ij} \coloneqq k_2(\mathbf{y}_i, \mathbf{y}_j)$. $h_{ij} \coloneqq \delta_{ij} - 1/N$ centralizes the two Gram matrices $\mathbf{K}_1$ and $\mathbf{K}_2$, which makes them have zero mean. The more details of HSIC can be found in [39] and [40].

To encourage the encoded representations (i.e., $\mathbf{H}_v, v = 1, \ldots, V$) of different views to be sufficiently diverse, the objective function in (5) is expressed as

$$\min_{\mathbf{D}_0,\mathbf{D}_v,\mathbf{H}_v} \ \sum_v \left\| \Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v \right\|_F^2 + \beta \sum_v \sum_{w \neq v} \text{HSIC}(\mathbf{H}_v, \mathbf{H}_w) \tag{7}$$

where β is a tradeoff parameter. After obtaining the new data representation $\mathbf{H}_v$ for vth view, the objective function for self-representation-based subspace learning is given by

$$\min_{\mathbf{Z}_v,\mathbf{E}_v} \|\mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v\|_F^2 + \lambda\|\mathbf{E}_v\|_{2,1} \tag{8}$$

where λ is a non-negative tradeoff parameter, and $||.||_{2,1}$ denotes $\ell_{2,1}$-norm which encourages the columns of a matrix to be zero [35], i.e., $\|\mathbf{E}\|_{2,1} = \sum_{j=1}^{N}\sqrt{\sum_{i=1}^{M}[\mathbf{E}_{ij}]^2}$, where $\mathbf{E} \in \mathbb{R}^{M\times N}$.

Finally, the objective function of multiview latent space learning framework with FRM can be formulated as

$$\begin{aligned}\min_{\Theta_v,\mathbf{P}_v,\mathbf{Z}_v,\mathbf{E}_v,\mathbf{H}_v,\mathbf{D}_v,\mathbf{D}_0} & \sum_v \|\mathbf{Z}_v\|_* + \lambda\sum_v \|\mathbf{E}_v\|_{2,1} \\ & + \beta\sum_v \|\Theta_v\|_F^2 \\ & + \underbrace{\gamma\sum_v\sum_{w\neq v}\mathrm{HSIC}(\mathbf{H}_v,\mathbf{H}_w)}_{\text{diversity term}} \\ \text{s.t.} \ & \underbrace{\Theta_v^\top\mathbf{X}_v = [\mathbf{D}_0,\mathbf{D}_v]\mathbf{H}_v + \mathbf{E}_v^1}_{\text{latent space learning term}} \\ & \underbrace{\mathbf{X}_v = \mathbf{P}_v\Theta_v^\top\mathbf{X}_v + \mathbf{E}_v^2}_{\text{reconstruction term}} \\ & \underbrace{\mathbf{H}_v = \mathbf{H}_v\mathbf{Z}_v + \mathbf{E}_v^3}_{\text{self-representation term}}, \mathbf{E}_v = \left[\mathbf{E}_v^1;\mathbf{E}_v^2;\mathbf{E}_v^3\right] \\ & \mathbf{D}_0^\top\mathbf{D}_0 = \mathbf{I}, \mathbf{D}_v^\top\mathbf{D}_v = \mathbf{I}, \mathbf{P}_v^\top\mathbf{P}_v = \mathbf{I} \end{aligned} \quad (9)$$

where $\|.\|_*$ is the matrix nuclear norm, which enforces the subspace representation to be low rank. To clearly show the proposed formulation, we list the main notations in Table I. In detail, the critical properties of our proposed formulation are explained below.

1) The latent space learning term (i.e., $\Theta_v^\top\mathbf{X}_v = [\mathbf{D}_0,\mathbf{D}_v]\mathbf{H}_v + \mathbf{E}_v^1$) is utilized to learn the shared dictionary and view-specific dictionaries. First, we learn a projection matrix Θ_v for the vth view to project the original features into a subspace spanned latent space. Within the latent space, we assume that different views are composed of a shared dictionary and view-specific dictionaries. As a consequence, our model can exploit the correlation across multiple views by leveraging the shared dictionary. Besides, orthogonal constraints are imposed on $\mathbf{D}_0$ and $\mathbf{D}_v$ (i.e., $\mathbf{D}_0^\top\mathbf{D}_0 = \mathbf{I}$ and $\mathbf{D}_v^\top\mathbf{D}_v = \mathbf{I}$) to prevent the trivial solution. If this constraint is missing, the values in $\mathbf{H}_v$ could be arbitrarily large.
2) The reconstruction term (i.e., $\mathbf{X}_v = \mathbf{P}_v\Theta_v^\top\mathbf{X}_v + \mathbf{E}_v^2$) is used to ensure that a good reconstruction of the original data can be obtained by using the learned low-dimensional features. Its main advantage is to reduce the redundancy as well as preserve critical and useful information in data. Besides, an orthogonal constraint is also introduced, i.e., $\mathbf{P}_v^\top\mathbf{P}_v = \mathbf{I}$.
3) The self-representation term (i.e., $\mathbf{H}_v = \mathbf{H}_v\mathbf{Z}_v + \mathbf{E}_v^3$) is used to reconstruct the data points by using the learned new representations (i.e., $\mathbf{H}_v$). In addition, we utilize the HSIC as a diversity term [i.e., $\mathrm{HSIC}(\mathbf{H}_v,\mathbf{H}_w)$, $w \neq v$] to explore the complementarity of multiview representations. Since each view of the data could contain some knowledge that other views do not have, this information can strengthen the ability to exploit the diversity within each view for improving the multiview learning performance. In addition, there are two different understandings of $\mathbf{H}_v$: one is the matrix $\mathbf{H}_v$ records the coding coefficients that linearly represent the input data by the dictionary atoms, and the other is that the $\mathbf{H}_v$ can be regarded as new feature representations of data points, which plays an important role in representing the data structure from each view.
4) The constraint term $\|\mathbf{Z}_v\|_*$ prevents the trivial solution by enforcing the self-representation to be low rank. Additionally, for the constraint $\mathbf{E}_v = [\mathbf{E}_v^1;\mathbf{E}_v^2;\mathbf{E}_v^3]$, we vertically concatenate the column of errors, which enforces the columns of $\mathbf{E}_v^1$, $\mathbf{E}_v^2$, and $\mathbf{E}_v^3$ to jointly have consistent magnitude values. The effectiveness has been previously investigated in [25]. $\ell_{2,1}$-norm encourages the columns of $\mathbf{E}_v$ to be zero. Thus, an underlying assumption here is that the corruptions in data are example specific.

TABLE I
MAIN NOTATIONS USED IN THE PROPOSED FORMULATION

Notation	Description
$\mathbf{X}_v$	Feature matrix for the v-th view
Θ_v	Linear transformation matrix for the v-th view
$\mathbf{P}_v$	Reconstruction matrix for the v-th view
$\mathbf{H}_v$	Low-dimension feature matrix for the v-th view
$\mathbf{Z}_v$	View-specific self-representation matrix for the v-th view
$\mathbf{D}_v$	View-specific dictionary for the v-th view
$\mathbf{D}_0$	Shared dictionary
$\mathbf{E}_v$	Error terms
L_v	Dimension of original features for the v-th view
L_0	Dimension in latent space
K_v	Dimension of learned representations in view-specific dictionary
K_0	Dimension of learned representations in shared dictionary
V	Number of multiple views
v	View index
λ, β, γ	Regularization parameters

B. Optimization

The objective function in (9) is not jointly convex with respect to all variables. Hence, we can utilize the ADMM [41] algorithm to efficiently and effectively solve our problem. To adopt ADMM strategy to our problem, we introduce auxiliary variables $\mathbf{J}_v$ to replace $\mathbf{Z}_v$ to make our problem separable. Then, we have the following equivalent problem:

$$\begin{aligned}\min_{\Theta_v,\mathbf{P}_v,\mathbf{Z}_v,\mathbf{E}_v,\mathbf{H}_v,\mathbf{D}_v,\mathbf{D}_0} & \sum_v \|\mathbf{Z}_v\|_* + \lambda\sum_v \|\mathbf{E}_v\|_{2,1} \\ & + \beta\sum_v \|\Theta_v\|_F^2 \\ & + \gamma\sum_v\sum_{w\neq v}\mathrm{HSIC}(\mathbf{H}_v,\mathbf{H}_w) \\ \text{s.t.} \ & \Theta_v^\top\mathbf{X}_v = [\mathbf{D}_0,\mathbf{D}_v]\mathbf{H}_v + \mathbf{E}_v^1, \\ & \mathbf{X}_v = \mathbf{P}_v\Theta_v^\top\mathbf{X}_v + \mathbf{E}_v^2 \\ & \mathbf{H}_v = \mathbf{H}_v\mathbf{Z}_v + \mathbf{E}_v^3, \mathbf{E}_v = \left[\mathbf{E}_v^1;\mathbf{E}_v^2;\mathbf{E}_v^3\right] \\ & \mathbf{D}_0^\top\mathbf{D}_0 = \mathbf{I}, \mathbf{D}_v^\top\mathbf{D}_v = \mathbf{I}, \mathbf{P}_v^\top\mathbf{P}_v = \mathbf{I} \\ & \mathbf{Z}_v = \mathbf{J}_v. \end{aligned} \quad (10)$$

Thus, the augmented Lagrangian function can be given by

$$\begin{aligned}
&\mathcal{L}\Big(\Theta_v, \mathbf{P}_v, \mathbf{J}_v, \mathbf{Z}_v, \mathbf{E}_v, \mathbf{H}_v, \mathbf{D}_v, \mathbf{Y}_v^1, \mathbf{Y}_v^2, \mathbf{Y}_v^3, \mathbf{Y}_v^4, \mathbf{D}_0\Big) \\
&= \sum_v \|\mathbf{J}_v\|_* + \lambda \sum_v \|\mathbf{E}_v\|_{2,1} \\
&\quad + \beta \sum_v \|\Theta_v\|_F^2 + \gamma \sum_v \sum_{w \neq v} \mathrm{HSIC}(\mathbf{H}_v, \mathbf{H}_w) \\
&\quad + \sum_v \Phi\Big(\mathbf{Y}_v^1, \Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\Big) \\
&\quad + \sum_v \Phi\Big(\mathbf{Y}_v^2, \mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v - \mathbf{E}_v^2\Big) \\
&\quad + \sum_v \Phi\Big(\mathbf{Y}_v^3, \mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3\Big) \\
&\quad + \sum_v \Phi\Big(\mathbf{Y}_v^4, \mathbf{Z}_v - \mathbf{J}_v\Big) \\
&\text{s.t. } \mathbf{E}_v = \Big[\mathbf{E}_v^1; \mathbf{E}_v^2; \mathbf{E}_v^3\Big], \mathbf{D}_0^\top \mathbf{D}_0 = \mathbf{I}, \mathbf{D}_v^\top \mathbf{D}_v = \mathbf{I}, \mathbf{P}_v^\top \mathbf{P}_v = \mathbf{I}
\end{aligned} \tag{11}$$

where $\Phi(\mathbf{Y}, \mathbf{Q}) = (\mu/2)\|\mathbf{Y}\|_F^2 + \langle \mathbf{Y}, \mathbf{Q}\rangle$, with $\langle \cdot, \cdot \rangle$ denoting the matrix inner product, μ is a positive penalty scalar, and $\mathbf{Y}_v^1$, $\mathbf{Y}_v^2$, $\mathbf{Y}_v^3$, and $\mathbf{Y}_v^4$ are Lagrangian multipliers. Next, we detail the subproblems regarding each of the variables $\mathbf{J}_v$, $\mathbf{Z}_v$, $\mathbf{E}_v$, $\mathbf{H}_v$, $\mathbf{D}_v$, Θ_v, $\mathbf{P}_v$, and $\mathbf{D}_0$.

J_v-Subproblem: The associated optimization problem with respect to $\mathbf{J}_v$ can be written as

$$\begin{aligned}
&\min_{\mathbf{J}_v} \ \|\mathbf{J}_v\|_* + \Phi\Big(\mathbf{Y}_v^4, \mathbf{Z}_v - \mathbf{J}_v\Big) \\
\Leftrightarrow &\min_{\mathbf{J}_v} \ \frac{1}{\mu}\|\mathbf{J}_v\|_* + \frac{1}{2}\|\mathbf{J}_v - \Big(\mathbf{Z}_v + \mathbf{Y}_v^4/\mu\Big)\|_F^2
\end{aligned} \tag{12}$$

which can be solved by using a singular value thresholding operator [42], namely $\mathbf{J}_v = \mathbf{U}\mathbf{S}_{(1/\mu)}(\mathbf{\Sigma})\mathbf{V}^\top$, where $\mathbf{S}_{(1/\mu)}(\Sigma_{ii}) = \mathrm{sign}(\Sigma_{ii})\max(\Sigma_{ii} - 1/\mu, 0)$ is a soft-thresholding operator, and $\mathbf{Z}_v + \mathbf{Y}_v^4/\mu = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the SVD of $\mathbf{Z}_v + \mathbf{Y}_v^4/\mu$.

Z_v-Subproblem: By fixing every other variables to constant, we update $\mathbf{Z}_v$ by solving

$$\begin{aligned}
&\min_{\mathbf{Z}_v} \ \Phi\Big(\mathbf{Y}_v^3, \mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3\Big) + \Phi\Big(\mathbf{Y}_v^4, \mathbf{Z}_v - \mathbf{J}_v\Big) \\
\Leftrightarrow &\min_{\mathbf{Z}_v} \ \frac{\mu}{2}\|\mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3 + \mathbf{Y}_v^3/\mu\|_F^2 \\
&\qquad + \frac{\mu}{2}\|\mathbf{Z}_v - \mathbf{J}_v + \mathbf{Y}_v^4/\mu\|_F^2.
\end{aligned} \tag{13}$$

Taking the derivative with respect to $\mathbf{Z}_v$ and setting it to zero, we obtain

$$\mathbf{Z}_v = \Big(\mathbf{H}_v^\top \mathbf{H}_v + \mathbf{I}\Big)^{-1}\Big(\mathbf{H}_v^\top\Big(\mathbf{H}_v - \mathbf{E}_v^3 + \mathbf{Y}_v^3/\mu\Big) + \mathbf{J}_v - \mathbf{Y}_v^4/\mu\Big). \tag{14}$$

E_v-Subproblem: The error term $\mathbf{E}_v$ can be updated by solving

$$\begin{aligned}
\min_{\mathbf{E}_v} \ &\lambda\|\mathbf{E}_v\|_{2,1} + \Phi\Big(\mathbf{Y}_v^1, \Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\Big) \\
&+ \Phi\Big(\mathbf{Y}_v^2, \mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v - \mathbf{E}_v^2\Big) \\
&+ \Phi\Big(\mathbf{Y}_v^3, \mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3\Big).
\end{aligned} \tag{15}$$

It is equivalently to solving the following problem:

$$\min_{\mathbf{E}_v} \frac{\lambda}{\mu}\|\mathbf{E}_v\|_{2,1} + \frac{1}{2}\|\mathbf{E}_v - \mathbf{G}_v\|_F^2 \tag{16}$$

where $\mathbf{G}_v$ can be formed by vertically concatenating the matrices $\Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v + \mathbf{Y}_v^1/\mu$, $\mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v + \mathbf{Y}_v^2/\mu$, and $\mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v + \mathbf{Y}_v^3/\mu$. Then, an $\ell_{2,1}$ minimization operator as in [35] can be used to obtain the optimal $\mathbf{E}_v$.

H_v-Subproblem: Dropping all unrelated terms with respect to $\mathbf{H}_v$ yields

$$\begin{aligned}
\min_{\mathbf{H}_v} \ &\gamma \sum_{w \neq v} \mathrm{HSIC}(\mathbf{H}_v, \mathbf{H}_w) \\
&+ \Phi\Big(\mathbf{Y}_v^1, \Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\Big) \\
&+ \Phi\Big(\mathbf{Y}_v^3, \mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3\Big).
\end{aligned} \tag{17}$$

By following [40], we utilize the inner product kernel for HSIC constraint, i.e., $\mathbf{K}_v = \mathbf{H}_v^\top \mathbf{H}_v$. Then, we have the following equation:

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{H}_v, \mathbf{H}_w) &= \mathrm{tr}\Big(\mathbf{H}_v\mathbf{K}\mathbf{H}_v^\top\Big) \\
\text{with } \ \mathbf{K} &= \sum_{w \neq v} \mathbf{M}\mathbf{K}_w\mathbf{M}
\end{aligned} \tag{18}$$

where $\mathbf{M} = [\mathbf{m}_{ij}]$ with $\mathbf{m}_{ij} = \delta_{ij} - 1/n$. The details can be found in [39] and [40]. By plugging (18) into (17), and setting the derivative of (17) to $\mathbf{H}_v$ to zero, we get the following closed-form solution:

$$\begin{aligned}
\mathbf{H}_v = &\Big([\mathbf{D}_0, \mathbf{D}_v]^\top\Big(\Theta_v^\top \mathbf{X}_v + \mathbf{Y}_v^1/\mu - \mathbf{E}_v^1\Big) \\
&+ \Big(\mathbf{E}_v^3 - \mathbf{Y}_v^3/\mu\Big)\mathbf{Z}_v'^\top\Big)\Big(\mathbf{I} + \mathbf{Z}_v'\mathbf{Z}_v'^\top + \frac{2\gamma}{\mu}\mathbf{K}\Big)^{-1}
\end{aligned} \tag{19}$$

where $\mathbf{Z}_v' = \mathbf{I} - \mathbf{Z}_v$ and $\mathbf{I}$ is an identity matrix.

D_v-Subproblem: The associated optimization problem with respect to $\mathbf{D}_v$ can be written as

$$\min_{\mathbf{D}_v^\top \mathbf{D}_v = \mathbf{I}} \ \Phi\Big(\mathbf{Y}_v^1, \Theta_v^\top \mathbf{X}_v - \mathbf{D}_0\mathbf{H}_v^s - \mathbf{D}_v\mathbf{H}_v^c - \mathbf{E}_v^1\Big) \tag{20}$$

where $\mathbf{H}_v = [\mathbf{H}_v^s; \mathbf{H}_v^c]$, and $\mathbf{H}_v^s$ and $\mathbf{H}_v^c$ are corresponding to $\mathbf{D}_0$ and $\mathbf{D}_v$, respectively. In other words, $\mathbf{H}_v^s$ is the new representation learned from the shared dictionary $\mathbf{D}_0$, while $\mathbf{H}_v^c$ is the new representation learned from view-specific dictionary $\mathbf{D}_v$. Equation (20) contains a matrix orthogonality constraint which has been used in [25] and [43]. The detailed optimization steps for solving $\mathbf{D}_v$ can be found in [25] and [43].

Θ_v-Subproblem: The associated optimization problem with respect to Θ_v can be written as

$$\begin{aligned}
\min_{\Theta_v} \ &\beta\|\Theta_v\|_F^2 + \Phi\Big(\mathbf{Y}_v^1, \Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\Big) \\
&+ \Phi\Big(\mathbf{Y}_v^2, \mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v - \mathbf{E}_v^2\Big).
\end{aligned} \tag{21}$$

It is equivalent to optimizing the following problem:

$$\begin{aligned}
\min_{\Theta_v} \ &\beta\|\Theta_v\|_F^2 + \frac{\mu}{2}\|\Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v + \mathbf{Y}_v^1/\mu - \mathbf{E}_v^1\|_F^2 \\
&+ \frac{\mu}{2}\|\mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v + \mathbf{Y}_v^2/\mu - \mathbf{E}_v^2\|_F^2.
\end{aligned} \tag{22}$$

Algorithm 1: Solving Problem (11) via ADMM

Input: Multi-view matrices: $\{X_1, \ldots, X_V\}$, hyper-parameters λ, β and γ, and L_0, K_0 and K_v.
Initialize: $\mathbf{Y}_v^1 = \mathbf{Y}_v^2 = \mathbf{Y}_v^3 = \mathbf{Y}_v^4 = \mathbf{0}$ $(v = 1, \ldots, V)$, $\varepsilon = 10^{-6}$, $\rho = 1.5$, $\mu = 10^{-4}$, $max_\mu = 10^6$.
Output: $\mathbf{Z}_1, \ldots, \mathbf{Z}_V$.
while *not converged* **do**
  Update $\mathbf{J}_v$, $\mathbf{Z}_v$, $\mathbf{E}_v$, $\mathbf{H}_v$, $\mathbf{D}_v$, Θ_v, $\mathbf{P}_v$, and $\mathbf{D}_0$ via Eq. (12), Eq. (14), Eq. (16), Eq. (19), Eq. (20), Eq. (23), Eq. (24), and Eq. (25), respectively;
  Update multipliers $\mathbf{Y}_v^1$, $\mathbf{Y}_v^2$, $\mathbf{Y}_v^3$, and $\mathbf{Y}_v^4$ via Eq. (26);
  Update the parameter μ via $\mu := \min(\rho\mu, max_\mu)$;
  Check convergence conditions:
  $\|\Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\|_\infty < \varepsilon$,
  $\|\mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v - \mathbf{E}_v^2\|_\infty < \varepsilon$,
  $\|\mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3\|_\infty < \varepsilon$, and
  $\|\mathbf{Z}_v - \mathbf{J}_v\|_\infty < \varepsilon$.
end

Taking the derivative of the above objective with respect to Θ_v and setting it to zero, we have the following closed-form solution:

$$\Theta_v = \left(\frac{2\beta}{\mu}\mathbf{I} + 2\mathbf{X}_v\mathbf{X}_v^\top\right)^{-1} \times \left\{\mathbf{X}_v\left(\left(\mathbf{X}_v + \mathbf{Y}_v^2/\mu - \mathbf{E}_v^2\right)^\top \mathbf{P}_v + \left([\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{Y}_v^1/\mu + \mathbf{E}_v^1\right)^\top\right)\right\}. \quad (23)$$

P_v-Subproblem: The associated optimization problem with respect to $\mathbf{P}_v$ can be written as

$$\min_{\mathbf{P}_v^\top \mathbf{P}_v = \mathbf{I}} \Phi\left(\mathbf{Y}_v^2, \mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v - \mathbf{E}_v^2\right). \quad (24)$$

We also use the same method from [25] and [43] to obtain an optimal $\mathbf{P}_v$.

D_0-Subproblem: By dropping all other unrelated terms, we optimize $\mathbf{D}_0$ by

$$\begin{aligned} &\min_{\mathbf{D}_0^\top \mathbf{D}_0 = \mathbf{I}} \sum_v \Phi\left(\mathbf{Y}_v^1, \Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\right) \\ \Leftrightarrow &\min_{\mathbf{D}_0^\top \mathbf{D}_0 = \mathbf{I}} \sum_v \frac{\mu}{2}\|\mathbf{X}_v - \mathbf{D}_0\mathbf{H}_v^s - \mathbf{D}_v\mathbf{H}_v^c + \mathbf{Y}_v^1/\mu - \mathbf{E}_v^1\|_F^2 \\ \Leftrightarrow &\min_{\mathbf{D}_0^\top \mathbf{D}_0 = \mathbf{I}} \left\|\left[\mathbf{X}_1', \ldots, \mathbf{X}_V'\right] - \mathbf{D}_0\left[\mathbf{H}_1^s; \ldots; \mathbf{H}_V^s\right]\right\|_F^2 \end{aligned} \quad (25)$$

where $\mathbf{X}_v' = \Theta_v^\top \mathbf{X}_v - \mathbf{D}_v\mathbf{H}_v^c + \mathbf{Y}_v^1/\mu - \mathbf{E}_v^1$. Then, we can obtain the optimal $\mathbf{D}_0$ by using the same optimization strategy as solving $\mathbf{D}_v$ in (20).

Multipliers Updating: The multipliers $\mathbf{Y}_v^1$, $\mathbf{Y}_v^2$, $\mathbf{Y}_v^3$, and $\mathbf{Y}_v^4$ $(v = 1, \ldots, V)$ can be updated with the following rule:

$$\begin{cases} \mathbf{Y}_v^1 := \mathbf{Y}_v^1 + \mu\left(\Theta_v^\top \mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\right) \\ \mathbf{Y}_v^2 := \mathbf{Y}_v^2 + \mu\left(\mathbf{X}_v - \mathbf{P}_v\Theta_v^\top \mathbf{X}_v - \mathbf{E}_v^2\right) \\ \mathbf{Y}_v^3 := \mathbf{Y}_v^3 + \mu\left(\mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3\right) \\ \mathbf{Y}_v^4 := \mathbf{Y}_v^4 + \mu(\mathbf{Z}_v - \mathbf{J}_v). \end{cases} \quad (26)$$

The variables are updated iteratively until convergence. The details of optimizing (11) via ADMM algorithm are summarized in Algorithm 1.

TABLE II
DETAILS OF SIX BENCHMARK DATASETS

Datasets	# Size	# Views	# Clusters
Notting-Hill	4660	3	5
CMU-PIE	5440	3	68
Caltech101	441	6	7
MSRCV1	210	6	7
Oxford Flowers	1360	3	17
Still DB	467	3	6

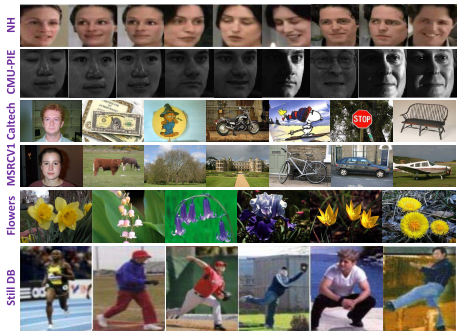


Fig. 2. Sampling images of six benchmark datasets in our experiments ("NH" is short for "Notting-Hill Video Face").

C. Complexity Analysis

The major computational burden of our Algorithm 1 lies in updating $\mathbf{J}_v$, $\mathbf{Z}_v$, $\mathbf{E}_v$, $\mathbf{H}_v$, $\mathbf{D}_v$, Θ_v, $\mathbf{P}_v$, and $\mathbf{D}_0$. Specifically, updating $\mathbf{J}_v$ with SVD operation takes $\mathcal{O}(N^3)$ complexity and updating $\mathbf{Z}_v$ requires $\mathcal{O}(N^3)$ complexity due to matrix inversion. The complexity of optimizing $\mathbf{E}_v$ is $\mathcal{O}(N^2)$. The complexities of updating $\mathbf{H}_v$ and Θ_v are $\mathcal{O}(N^3)$. Moreover, the complexities of updating $\mathbf{D}_v$, $\mathbf{P}_v$, and $\mathbf{D}_0$ are $\mathcal{O}(K_v^2 N)$, $\mathcal{O}(L_0^2 N)$, and $\mathcal{O}(K_0^2 N)$, respectively. Thus, the total complexity of our algorithm is deduced as $\mathcal{O}(T(V(K_0^2+K_v^2+L_0^2)N+VN^2+VN^3))$, where T is the total iteration number, and V is the number of views. Further, considering $N \gg V$ for multiview data setting, the main computational complexity of the proposed approach is $\mathcal{O}(T(V(K_0^2 + K_v^2 + L_0^2)N + VN^3))$.

IV. EXPERIMENTAL RESULTS

In this section, we first explain the experimental setup, and then compare our method with other state-of-the-art multiview clustering and classification approaches. Subsequently, we investigate some critical properties of our proposed multiview learning approach.

A. Experimental Setup

Datasets: We evaluate the effectiveness of the proposed approach by using the following six popular benchmark datasets (the statistics of the employed datasets are summarized in Table II, some example images of these six datasets are shown in Fig. 2, and the value in bracket denotes the dimension of feature).

1) *Notting-Hill Video Face [40]:* The examples in this dataset are captured from the movie "Notting-Hill," where the faces of five main casts are used which lead to totally 4660 face examples. For this dataset, we resize

the images into 48×48 and extract three types of features, including gray-level intensity (2000), LBP (3304), and Gabor (6750).

2) *CMU-PIE*[1]*:* It consists of 68 subjects in total, with large variances within the same subject but in different poses. We randomly select 80 examples from each subject to construct 5440 facial images in the evaluation subset, where all face images are cropped to the size of 64×64. Also, three types of features, namely gray intensity (1024), LBP (256), and HOG (496), are used for this paper.
3) *Caltech101*[2]*:* This image data set consists of 101 categories of images for object recognition problem. We follow previous work [44] and select the images from seven widely used classes, i.e., Dolla-bill, Face, Garfield, Motorbike, Snoopy, Stop-sign, and Windsor-chair. Specifically, six types of features are extracted, including CENTRIST (1302), CMT (48), GIST (512), HOG (100), LBP (256), and SIFT (441).
4) *MSRCV1*[3]*:* This dataset consists of 240 images and 8 object classes. Similar to [25], we select the examples of seven classes, i.e., Cow, Tree, Building, Airplane, Face, Car, and Bicycle. We also extract six types of features which are CENTRIST (1302), CMT (48), GIST (512), HOG (100), LBP (256), and SIFT (210).
5) *Oxford Flowers*[4]*:* This dataset is composed of 1360 examples with 17 flower categories. In this dataset, χ^2 distance matrices for three different visual features [45], i.e., color (1360), texture (1360), and shape (1360), are used to form three views.
6) *Still DB [46]:* The dataset is a still image dataset which is made up of 467 images with six classes of actions. Three types of features are extracted, i.e., Sift Bow (200), Color Sift Bow (200), and Shape Context Bow (200).

Parametric Settings: We set $L_0 = 100$ (L_0 denotes the dimensionality of latent space) for all datasets, and set the number of atoms from the view-specific dictionary as $K_v = 60$ ($v = 1, \ldots, V$). Besides, the number of atoms within the shared dictionary K_0 is tuned in the range of $\{10, 20, \ldots, 60\}$. The parameters λ, β, and γ are selected within the range of $\{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$.

B. Comparison on Clustering Tasks

Compared Methods: To verify the effectiveness of our proposed framework, we first implement clustering and compare our results to some recent state-of-the-art single-view and multiview clustering methods, which include the following.

1) $Single_{best}$: This method performs standard spectral clustering algorithm [47] by selectively using the most informative view.
2) LRR_{best}: This method performs LRR [35] by selectively using the most informative view. We tune the involved parameter λ in the range of $\{0.001, 0.01, 0.02, 0.05, 0.1, 0.2, 1, 2, 5\}$.
3) $S3C_{best}$ *[48]:* This method carries out the clustering on every single view and then outputs the best performance. The parameter spaces are $\lambda \in \{10^{-5}, 10^{-4}, \ldots, 10^{5}\}$ and $\alpha \in [0.03, 0.3]$, respectively.
4) *FeatConcate:* This method concatenates the features of all views and then conducts the standard spectral clustering.
5) *ConcatePCA:* This method concatenates the features from all views and then applies PCA [51] to obtain a low-dimensional subspace representation. Further, it conducts the standard spectral clustering on the low-dimensional representation. The optimal dimensionality is searched in the range of $\{100, 200, \ldots, 500\}$.
6) *Co-Reg SPC [21]:* This method co-regularizes the clustering hypotheses to enforce corresponding data point in each view to have the same cluster membership. Its parameter λ is searched in the range of $\{0, 0.02, \ldots, 0.1\}$.
7) *Min-Dis [50]:* This method creates a bipartite graph and tries to minimize the disagreement among various views. The final result is obtained through spectral clustering.
8) *MVSC [24]:* This method performs subspace clustering on individual views and then fuses their outputs to obtain the final result. Its parameters λ_1 and λ_2 are searched in the range of $\{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$.
9) *RMSC [23]:* This method recovers a shared low-rank transition probability matrix for clustering. Its parameter λ is searched from 0.005 to 100.
10) *Multi-NMF [17]:* This method searches a compatible clustering solution across multiple views by minimizing the differences between data representations of each view and the consensus matrix.
11) *DiMSC [40]:* This method enforces the diversity of different views using the HSIC criterion. Its parameter λ_s is searched in the range of $[0.01, 0.03]$, and the parameter λ_v is searched in the range of $\{20, 40, \ldots, 180\}$.
12) *LMSC [25]:* This method assumes that each view is originated from an underlying latent representation. Its parameter λ is decided within $\{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$.

For clustering, after obtaining the optimal $\mathbf{Z}_v$ for the vth view, we adopt the existing spectral clustering algorithm [47] on a similarity matrix $\mathbf{S}$, i.e., $\mathbf{S} = (1/V)\sum_v(|\mathbf{Z}_v| + |\mathbf{Z}_v^\top|)$. To achieve a fair comparison for all the compared methods, we directly use the source codes provided by the authors to obtain the best results. Finally, we report the mean values and standard deviations for all methods over 30 independent trials. In addition, to evaluate the clustering performance, six popular metrics, including normalized mutual information, accuracy (ACC), adjusted Rand index, F-score, precision, and recall are utilized in this paper. Each metric penalizes or favors different properties in the clustering task, and hence we report the results on these different measures to achieve a comprehensive evaluation of our method against all state-of-the-art methods. Note that a higher value indicates better clustering performance.

[1] http://vasc.ri.cmu.edu/idb/html/face/
[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[3] http://research.microsoft.com/en-us/projects/objectclassrecognition/
[4] http://www.robots.ox.ac.uk/vgg/data/flowers/

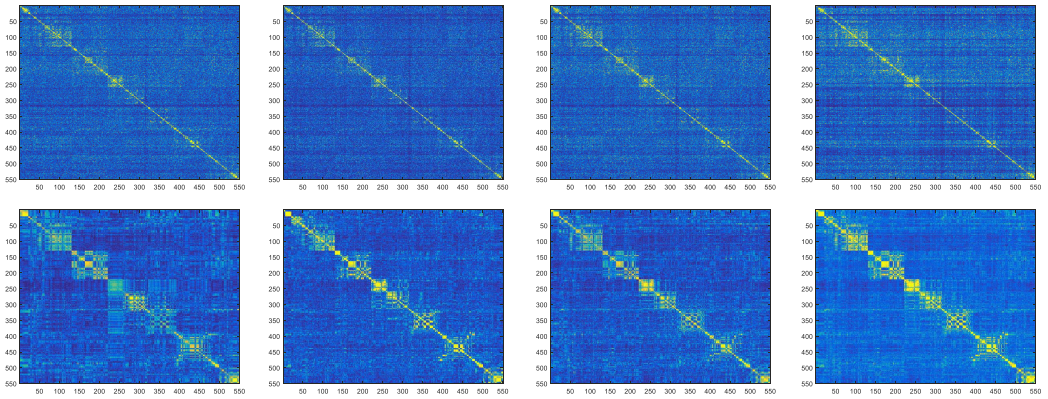


Fig. 3. Comparison of similarity matrices of LRR (top row) and our FRM (bottom row) on different views and their combinations.

TABLE III
RESULTS (MEAN ± STANDARD DEVIATION) ON NOTTING-HILL DATASET

Feature	Method	NMI	ACC	AR	F-score	Precision	Recall
Single	$\text{Single}_{\text{best}}$ [47]	0.723±0.008	0.813±0.000	0.712±0.020	0.775±0.015	0.774±0.018	0.776±0.013
	LRR_{best} [35]	0.650±0.002	0.754±0.004	0.655±0.006	0.713±0.005	0.714±0.004	0.712±0.006
	S3C_{best} [48]	0.645±0.002	0.746±0.000	0.613±0.003	0.697±0.003	0.702±0.002	0.692±0.003
Multiple	FeatConcate [47]	0.628±0.028	0.673±0.033	0.612±0.041	0.696±0.032	0.699±0.032	0.693±0.031
	ConcatePCA [49]	0.632±0.009	0.733±0.008	0.598±0.015	0.685±0.012	0.691±0.010	0.680±0.014
	Co-Reg SPC [21]	0.660±0.003	0.758±0.000	0.616±0.004	0.699±0.000	0.705±0.003	0.694±0.003
	Min-Dis [50]	0.707±0.003	0.791±0.000	0.689±0.002	0.758±0.002	0.750±0.002	0.765±0.003
	MVSC [24]	0.748±0.007	0.821±0.000	0.745±0.003	0.805±0.002	0.748±0.003	0.792±0.013
	RMSC [23]	0.685±0.011	0.782±0.017	0.688±0.028	0.755±0.021	0.760±0.025	0.751±0.021
	MultiNMF [17]	0.724±0.020	0.820±0.014	0.727±0.017	0.788±0.013	0.774±0.015	0.801±0.018
	DiMSC [40]	0.799±0.001	0.843±0.021	0.787±0.001	0.834±0.001	0.822±0.005	0.836±0.009
	LMSC [25]	0.786±0.011	0.840±0.012	0.781±0.009	0.811±0.012	0.825±0.008	0.817±0.019
	FRM	**0.788±0.012**	**0.901±0.008**	**0.820±0.009**	**0.859±0.010**	**0.864±0.011**	**0.855±0.009**

TABLE IV
RESULTS (MEAN ± STANDARD DEVIATION) ON CMU-PIE DATASET

Feature	Method	NMI	ACC	AR	F-score	Precision	Recall
Single	$\text{Single}_{\text{best}}$ [47]	0.611±0.009	0.396±0.013	0.274±0.013	0.286±0.013	0.268±0.012	0.306±0.019
	LRR_{best} [35]	0.673±0.011	0.527±0.009	0.330±0.014	0.341±0.017	0.290±0.016	0.415±0.011
	S3C_{best} [48]	0.628±0.017	0.436±0.014	0.283±0.011	0.285±0.008	0.277±0.011	0.316±0.012
Multiple	FeatConcate [47]	0.608±0.013	0.416±0.021	0.298±0.016	0.309±0.016	0.291±0.014	0.330±0.019
	ConcatePCA [49]	0.645±0.010	0.478±0.023	0.339±0.028	0.350±0.021	0.324±0.021	0.379±0.023
	Co-Reg SPC [21]	0.597±0.012	0.384±0.013	0.271±0.004	0.265±0.013	0.263±0.017	0.291±0.013
	Min-Dis [50]	0.587±0.010	0.374±0.021	0.273±0.017	0.282±0.017	0.293±0.022	0.307±0.019
	MVSC [24]	0.611±0.017	0.516±0.021	0.324±0.014	0.377±0.019	0.331±0.016	0.329±0.015
	RMSC [23]	0.604±0.027	0.413±0.017	0.317±0.013	0.297±0.016	0.304±0.018	0.347±0.023
	MultiNMF [17]	0.607±0.018	0.498±0.021	0.322±0.015	0.337±0.015	0.293±0.021	0.386±0.021
	DiMSC [40]	0.652±0.014	0.521±0.012	0.357±0.012	0.401±0.019	0.311±0.013	0.384±0.011
	LMSC [25]	0.692±0.006	0.542±0.018	0.413±0.019	0.422±0.015	0.331±0.018	0.437±0.014
	FRM	**0.710±0.006**	**0.582±0.018**	**0.429±0.019**	**0.438±0.015**	**0.382±0.018**	**0.471±0.013**

Fig. 3 shows the comparison of affinity matrices between LRR and our FRM method on Still DB dataset, which constructs the affinity matrices of different views and also their combinations. From Fig. 3, it is evident that our FRM reveals the underlying clustering structures more clearly than using LRR.

In addition, we report the performances of all methods on six benchmark datasets by using six evaluation metrics. The detailed results are presented in Tables III−VIII. In each table, the values in **bold** indicate the best performance among all methods. Specifically, Tables III and IV show the clustering results on two face datasets, Notting-Hill and CMU-PIE, respectively. Video face clustering in Notting-Hill dataset is very challenging because the appearances of faces often vary significantly due to the lighting conditions. From Table III, it can be seen that our approach outperforms all other single-view and multiview clustering methods. From Table IV, it can be observed that our approach performs best when compared with all other methods. More importantly, our multiview subspace clustering approach performs better than all the single-view methods, which indicate that our approach can effectively explore distinct information from multiple views to improve the clustering performance. This is because our approach utilizes the learned representations in latent space

TABLE V
RESULTS (MEAN ± STANDARD DEVIATION) ON CALTECH101 DATASET

Feature	Method	NMI	ACC	AR	F-score	Precision	Recall
Single	$\text{Single}_{\text{best}}$ [47]	0.589±0.009	0.629±0.007	0.523±0.012	0.576±0.009	0.586±0.014	0.566±0.003
	LRR_{best} [35]	0.639±0.002	0.646±0.003	0.580±0.001	0.649±0.002	0.631±0.001	0.623±0.003
	S3C_{best} [48]	0.578±0.000	0.611±0.007	0.504±0.009	0.559±0.007	0.568±0.010	0.551±0.006
Multiple	FeatConcate [47]	0.603±0.017	0.641±0.020	0.526±0.034	0.601±0.023	0.624±0.021	0.579±0.024
	ConcatePCA [49]	0.651±0.012	0.672±0.017	0.558±0.009	0.634±0.013	0.639±0.014	0.621±0.010
	Co-Reg SPC [21]	0.623±0.003	0.590±0.005	0.549±0.005	0.620±0.004	0.645±0.005	0.598±0.003
	Min-Dis [50]	0.624±0.004	0.701±0.023	0.552±0.007	0.623±0.006	0.645±0.006	0.603±0.007
	MVSC [24]	**0.683±0.002**	0.712±0.003	0.596±0.004	0.675±0.004	0.566±0.003	0.667±0.003
	RMSC [23]	0.625±0.026	0.708±0.023	0.614±0.018	0.675±0.027	0.699±0.018	0.654±0.023
	MultiNMF [17]	0.646±0.021	0.676±0.012	0.589±0.019	0.655±0.019	0.672±0.020	0.640±0.017
	DiMSC [40]	0.642±0.011	0.727±0.003	0.615±0.008	0.675±0.008	0.704±0.008	0.648±0.010
	LMSC [25]	0.652±0.013	0.710±0.012	0.593±0.010	0.664±0.009	0.656±0.014	0.661±0.012
	FRM	0.664±0.009	**0.766±0.010**	**0.651±0.009**	**0.710±0.011**	**0.687±0.007**	**0.712±0.012**

TABLE VI
RESULTS (MEAN ± STANDARD DEVIATION) ON MSRCV1 DATASET

Feature	Method	NMI	ACC	AR	F-score	Precision	Recall
Single	$\text{Single}_{\text{best}}$ [47]	0.574±0.032	0.668±0.051	0.536±0.010	0.535±0.043	0.571±0.009	0.612±0.009
	LRR_{best} [35]	0.569±0.008	0.676±0.009	0.502±0.010	0.524±0.009	0.543±0.009	0.587±0.007
	S3C_{best} [48]	0.612±0.005	0.688±0.009	0.514±0.006	0.583±0.006	0.572±0.009	0.594±0.010
Multiple	FeatConcate [47]	0.613±0.042	0.672±0.031	0.505±0.032	0.575±0.024	0.566±0.021	0.586±0.027
	ConcatePCA [49]	0.621±0.022	0.702±0.015	0.541±0.009	0.607±0.014	0.595±0.011	0.617±0.015
	Co-Reg SPC [21]	0.569±0.013	0.653±0.017	0.512±0.010	0.587±0.018	0.543±0.010	0.583±0.011
	Min-Dis [50]	0.657±0.017	0.745±0.044	0.567±0.008	0.628±0.007	0.615±0.015	0.643±0.010
	MVSC [24]	0.615±0.012	0.695±0.008	0.506±0.014	0.573±0.015	0.525±0.013	0.616±0.012
	RMSC [23]	0.656±0.026	0.747±0.017	0.577±0.033	0.637±0.035	0.623±0.023	0.653±0.025
	MultiNMF [17]	0.647±0.021	0.752±0.013	0.560±0.018	0.622±0.022	0.615±0.015	0.629±0.018
	DiMSC [40]	0.665±0.015	0.795±0.011	0.596±0.016	0.655±0.014	0.646±0.013	0.665±0.016
	LMSC [25]	0.653±0.011	0.806±0.013	0.599±0.017	0.652±0.017	0.612±0.012	0.663±0.011
	FRM	**0.759±0.012**	**0.871±0.011**	**0.715±0.013**	**0.755±0.012**	**0.750±0.013**	**0.761±0.014**

TABLE VII
RESULTS (MEAN ± STANDARD DEVIATION) ON OXFORD FLOWERS DATASET

Feature	Method	NMI	ACC	AR	F-score	Precision	Recall
Single	$\text{Single}_{\text{best}}$ [47]	0.416±0.005	0.392±0.007	0.236±0.002	0.281±0.002	0.276±0.002	0.287±0.002
	LRR_{best} [35]	0.419±0.006	0.396±0.009	0.239±0.003	0.284±0.003	0.279±0.004	0.290±0.003
	S3C_{best} [48]	0.417±0.002	0.392±0.004	0.236±0.003	0.281±0.004	0.276±0.003	0.286±0.002
Multiple	FeatConcate [47]	0.425±0.017	0.410±0.013	0.242±0.010	0.287±0.011	0.281±0.014	0.293±0.015
	ConcatePCA [49]	0.430±0.012	0.426±0.019	0.239±0.013	0.285±0.012	0.274±0.012	0.297±0.012
	Co-Reg SPC [21]	0.443±0.002	0.429±0.003	0.280±0.004	0.322±0.004	0.315±0.004	0.329±0.003
	Min-Dis [50]	0.375±0.002	0.401±0.039	0.258±0.005	0.216±0.006	0.213±0.005	0.221±0.005
	MVSC [24]	0.438±0.017	0.432±0.013	0.261±0.012	0.304±0.011	0.295±0.014	0.314±0.013
	RMSC [23]	0.396±0.014	0.385±0.016	0.231±0.019	0.249±0.011	0.234±0.012	0.256±0.010
	MultiNMF [17]	0.434±0.013	0.415±0.015	0.251±0.018	0.277±0.012	0.283±0.015	0.295±0.011
	DiMSC [40]	0.442±0.011	0.434±0.014	0.266±0.009	0.310±0.008	0.302±0.007	0.318±0.010
	LMSC [25]	0.444±0.009	0.442±0.009	0.275±0.007	0.318±0.012	0.312±0.011	0.325±0.011
	FRM	**0.485±0.008**	**0.478±0.013**	**0.316±0.007**	**0.357±0.009**	**0.346±0.011**	**0.367±0.010**

to reconstruct the data points, which can effectively reduce the feature redundancy to boost the clustering performance. Besides, within the latent space, our approach simultaneously exploits the correlations across multiple views and preserves view-specific property. Further, HSIC-based diversity term is utilized to enforce the complementary information, which could serve as a valuable complement to multiview clustering.

On the Caltech101 dataset, as shown in Table V, some approaches achieve good performance, while our method still outperforms all other compared methods. Table VI shows the clustering result on MSRCV1 dataset, and it can be observed that our approach reports significantly better performance than the competing baselines. Tables VII and VIII show the clustering results on Oxford Flowers and Still DB datasets, respectively. Compared with other baseline methods, LMSC and DiMSC achieve relatively better performances. Overall, our approach obtains much better clustering performance than the state-of-the-art methods.

C. Comparison on Classification Tasks

To further verify the effectiveness of the proposed framework, we test its ability on processing classification tasks based on the learned representations (i.e., $\mathbf{H}_v$). The k-NN ($k = 1$ in our experiment) classifier is a classic technique for classification, which has been applied in some existing works [52], [53].

TABLE VIII
RESULTS (MEAN ± STANDARD DEVIATION) ON STILL DB DATASET

Feature	Method	NMI	ACC	AR	F-score	Precision	Recall
Single	$\text{Single}_{\text{best}}$ [47]	0.105±0.008	0.294±0.009	0.061±0.000	0.219±0.006	0.222±0.001	0.216±0.002
	LRR_{best} [35]	0.109±0.003	0.306±0.004	0.061±0.004	0.219±0.000	0.223±0.003	0.212±0.004
	S3C_{best} [48]	0.101±0.002	0.288±0.002	0.062±0.002	0.220±0.003	0.224±0.003	0.217±0.002
Multiple	FeatConcate [47]	0.122±0.004	0.296±0.011	0.073±0.003	0.232±0.003	0.233±0.003	0.231±0.003
	ConcatePCA [49]	0.114±0.009	0.319±0.018	0.088±0.010	0.254±0.012	0.240±0.006	0.253±0.024
	Co-Reg SPC [21]	0.115±0.002	0.274±0.002	0.076±0.000	0.234±0.002	0.234±0.005	0.245±0.008
	Min-Dis [50]	0.097±0.000	0.336±0.014	0.082±0.008	0.233±0.004	0.232±0.003	0.237±0.011
	MVSC [24]	0.123±0.016	0.312±0.021	0.086±0.008	0.271±0.007	0.227±0.005	0.248±0.011
	RMSC [23]	0.089±0.009	0.305±0.010	0.073±0.011	0.221±0.002	0.231±0.004	0.219±0.002
	MultiNMF [17]	0.113±0.014	0.314±0.011	0.073±0.009	0.236±0.017	0.229±0.015	0.245±0.011
	DiMSC [40]	0.122±0.008	0.323±0.002	0.083±0.001	0.249±0.000	0.235±0.004	0.256±0.002
	LMSC [25]	0.136±0.003	0.327±0.003	0.084±0.011	**0.269±0.005**	0.235±0.007	0.247±0.012
	FRM	**0.151±0.005**	**0.358±0.008**	**0.101±0.005**	0.268±0.005	**0.256±0.004**	**0.271±0.006**

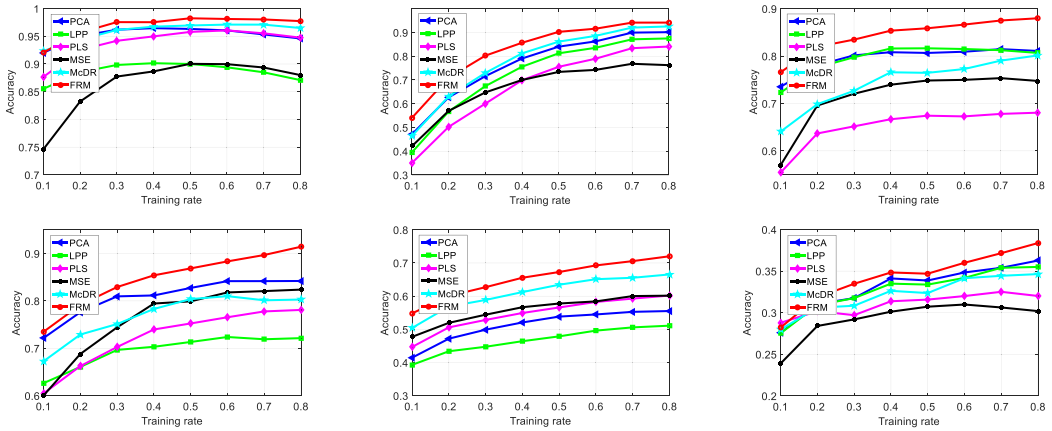


Fig. 4. Classification performance with respect to varying training rate (i.e., r) on six benchmark datasets (from left to right and top to bottom: Notting-Hill, CMU-PIE, Caltech101, MSRCV1, Oxford Flowers, and Still DB datasets, respectively).

The classification accuracy can be calculated by the ratio of the corrected classified against all considered test examples.

For each dataset, we split the entire dataset into the training and test sets. Specifically, $r \times N$ $(0 < r < 1)$ examples are randomly selected to build the training set and the remaining examples are for testing, where N is the size of the whole dataset and r denotes the ratio of training data to N. In the training set, the number of examples belonging to different classes is kept identical, while the number of test examples for different classes may not the same. Above dataset splitting is randomly repeated 30 times, and thus all the compared methods should independently run 30 times to generate the averaged classification accuracy. In this experiment, the compared methods are PCA [51], LPP [54], PLS [55], MSE [13], and McDR [53].

Fig. 4 shows the classification performances of various methods with respect to the varied ratio (i.e., $r = 0.1, 0.2, \ldots, 0.8$) on six benchmark datasets. From the classification results shown in Fig. 4, we have the following observations. First, directly concatenating the features of all views (i.e., PCA and LPP methods) is not reasonable, for example, LPP obtains relatively better performance on CMU-PIE, Caltech101, and Still DB datasets, while it obtains much worse performance on the remaining datasets. Second, although MSE and PLS methods can explore the correlations between different views, they ignore the relationship within individual view. Overall, our proposed approach simultaneously learns a shared dictionary to exploit the correlations across multiple views and learns the view-specific dictionary to preserve the property of individual view, which effectively learns the compact representations for improving the classification accuracy.

D. Model Property Evaluation

1) Evaluation of Redundancy Rate: To verify that our proposed approach reduces the redundancy feature information across multiview representations, we define a redundancy rate (RR) evaluation metric as

$$\text{RR} = \frac{\sum_{i=1}^{N} \sum_{v=1, w \neq v}^{V} |\text{Corr}(\mathbf{h}_v^i, \mathbf{h}_w^i)|}{V(V-1)N} \tag{27}$$

where $\text{Corr}(\cdot, \cdot)$ denotes the Pearson correlation coefficient between the two vectors, which can be used to measure the linear correlation between two variables. Similar to [56], the average sum of similarity of all N data samples in all pairs of views can be obtained in a range of $\{0, \ldots, 1\}$, where 0 means a completely complementary result, and 1 vice versa. We compare the RR of the proposed approach with PCA, multi-NMF, and McDR, which first obtain the representation of each view and then get the final multiview representation by averaging them. Table IX shows the comparison results. From Table IX,

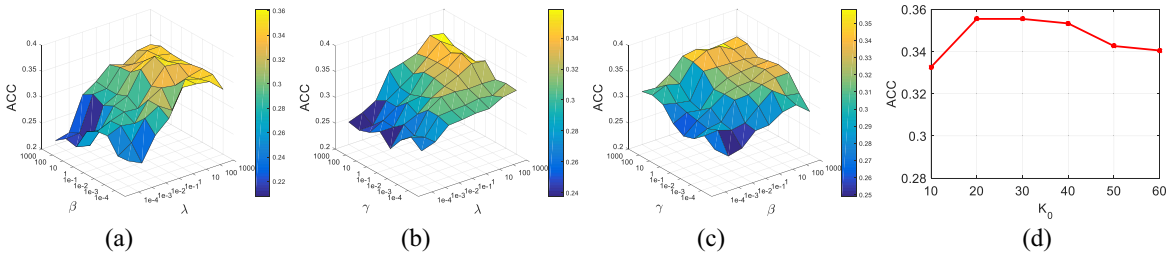


Fig. 5. Clustering performance with respect to different parameters: (a) λ and β; (b) λ and γ; (c) β and γ; and (d) K_0 on Still DB dataset.

TABLE IX
COMPARISON OF VARIOUS METHODS ON RR. THE BEST RESULT FOR EACH DATASET IS MARKED IN **BOLD**

Datasets	PCA	MultiNMF	McDR	FRM
Notting-Hill	0.4148	0.8106	0.3715	**0.1141**
CMU-PIE	0.3927	0.5416	0.3175	**0.1022**
Caltech101	0.3299	0.6870	0.4213	**0.0748**
MSRCV1	0.2572	0.5078	0.3107	**0.1159**
Oxford Flowers	0.5022	0.5778	0.4236	**0.1249**
Still DB	0.4786	0.6209	0.5313	**0.0951**

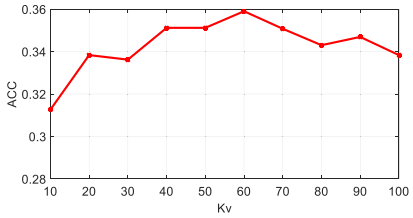


Fig. 6. Performance of the proposed approach with respect to ACC measure when varying the parameter K_v on Still DB dataset.

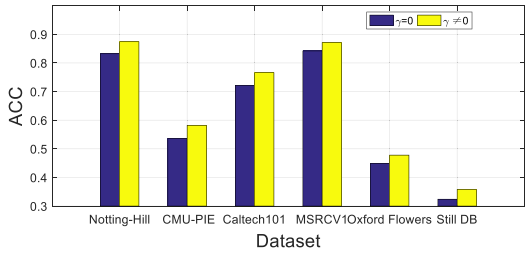


Fig. 7. Performance of the proposed approach with respect to ACC measure when using the HSIC term ($\gamma \neq 0$) and without using it ($\gamma = 0$).

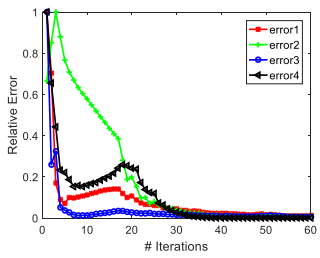


Fig. 8. Convergence curves of our proposed approach on Still DB dataset in terms of various relative errors.

it can be seen that our proposed approach can enforce the complementarity across multiple views and meanwhile effectively reducing the feature redundancy both.

2) Effects of Key Parameters: In our approach, there are three regularization parameters, i.e., λ, β, and γ in (9), which are used to balance different terms. We show the effect of these different parameters on our algorithm by using the Still DB dataset in Fig. 5. Fig. 5 illustrates the clustering performance when one parameter is fixed while the remaining two parameters are changed. It is observed that the clustering performance of our proposed approach is generally satisfactory (i.e., ACC ≥ 0.32) when $\lambda \geq 1$, $\beta \geq 1$, and $\gamma \geq 0.1$. Besides, we also notice that the parameter K_0, which denotes the number of atoms in the shared dictionary $\mathbf{D}_0$, can also influence the performance of our proposed approach. To show the effect of K_0, we vary K_0 within a range $\{10, \ldots, 60\}$ and study the model output. The produced ACC under different selections of K_0 is shown in Fig. 5(d). From Fig. 5(d), we see that good performance can be obtained when $K_0 \in [20, 40]$. Overall, the proposed algorithm can obtain promising clustering performance when $K_0 \in [20, 40]$. Besides, we tested the effects of K_v on Still DB dataset, when fixed all other parameters, to tune K_v in a range of $[10, 20, \ldots, 100]$. Fig. 6 shows the performance of the proposed approach with respect to ACC measure when varying the parameter K_v on Still DB dataset. From Fig. 6, it can be clearly seen that our method obtains better clustering performance (i.e., ACC > 0.34) when K_v in the range of $[40, 50, \ldots, 90]$, and it obtains the best performance when $K_v = 60$. Thus, we set $K_v = 60$ for all datasets.

Aiming to promote the diversity of the new learned representations in the proposed formulation, we employ the HSIC to penalize for dependence between data in different views. Accordingly, the redundancy among different views (especially the view-specific information) could be also reduced. Further, to verify the effectiveness of the HSIC in the proposed framework, we report the comparison clustering results (i.e., ACC) between using the diversity term and without using it in Fig. 7. From Fig. 7, it can be seen that the HSIC term helps to improve the multiview clustering performance.

3) Convergence Study: In this part, we provide empirical evidence to demonstrate the convergence of our approach on real-world data. Specifically, we investigate how the relative errors (i.e., $\|\Theta_v^T\mathbf{X}_v - [\mathbf{D}_0, \mathbf{D}_v]\mathbf{H}_v - \mathbf{E}_v^1\|_F^2$ as "error1," $\|\mathbf{X}_v - \mathbf{P}_v\Theta_v^T\mathbf{X}_v - \mathbf{E}_v^2\|_F^2$ as "error2," $\|\mathbf{H}_v - \mathbf{H}_v\mathbf{Z}_v - \mathbf{E}_v^3\|_F^2$ as "error3," and $\|\mathbf{Z}_v - \mathbf{J}_v\|_F^2$ as "error4") vanish when the iteration proceeds. The convergence curves on Still DB dataset are presented in Fig. 8. From Fig. 8, we can see that the convergence conditions are all reached within less than 60 iterations.

E. Discussion and Extension

Our method learns a latent space with redundancy minimization, which can reduce the effects of noise and outliers to learn more informative representations. The new representations can effectively depict the underlying relationships among different samples to improve the clustering and classification performance. In addition, some related studies [57]–[59] have indicated that redundant features can have significant adverse the effect on learning performance, thus, it is necessary to address this limitation for feature selection (or feature representation learning). These studies have also verified the effectiveness of redundancy minimization to learn more informative features. This is consistent with this paper, as it is expected to learn more informative features (representations) to reconstruct data points.

Linear projection employed in our framework (i.e., latent space learning) is a simple but effective technique for high-dimensional data, and it is easy to resolve in practice. In order to capture more complex correlations, some nonlinear methods (e.g., kernel technique [60] and deep networks [61]) will be introduced in our model in the future work. Besides, it often takes much time to project high-dimensional features into a latent space, especially, the feature reconstruction will increase the computation complexity. Thus, hashing technique [62], [63] can be also introduced in our model to accelerate the multiview learning speed.

In addition, the proposed framework can be easily extended to some related applications, e.g., visual tracking [64]–[67], classification tasks [68]–[72], etc. As in visual tracking, fusion of multiple features is an effective approach to improve tracking performance, thus it is also critical to reduce the redundancy of high-dimensional multiple features and exploit the correlations across multiple features. Therefore, we can consider some extensions based on the current framework in the future work.

V. Conclusion

In this paper, we have presented a novel multiview subspace learning framework by minimizing the feature redundancy in a learned latent space. Different from most existing multiview learning approaches that directly deploy the original redundant data, our approach can effectively improve its performances by utilizing the learned compact representations in a latent space. Importantly, within this latent space, our approach simultaneously captures the underlying correlations cross multiple views and takes advantage of the information embedded in each view to preserve the view-specific property. We have shown that the enhanced complementary information is helpful to multiview subspace learning. Extensive experimental results have demonstrated the superiority of the proposed multiview subspace learning approach when compared with the other state-of-the-art methods in terms of both clustering and classification tasks.

References

[1] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2531–2544, Dec. 2015.

[2] Y. Xie, W. Zhang, Y. Qu, L. Dai, and D. Tao, "Hyper-Laplacian regularized multilinear multiview self-representations for clustering and semisupervised learning," *IEEE Trans. Cybern.*, to be published.

[3] C. Zhang *et al.*, "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[4] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3516–3529, Nov. 2017.

[5] C. Gong, "Exploring commonality and individuality for multi-modal curriculum learning," in *Proc. AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 1926–1933.

[6] Z. Ding and Y. Fu, "Robust multiview data analysis through collective low-rank subspace," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1986–1997, May 2018.

[7] C. Gong *et al.*, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3249–3260, Jul. 2016.

[8] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, Madison, WI, USA, 1998, pp. 92–100.

[9] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in *Proc. Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 1135–1142.

[10] S. Yu *et al.*, "Bayesian co-training," *J. Mach. Learn. Res.*, vol. 12, pp. 2649–2680, Sep. 2011.

[11] J. Sun and S. Keates, "Canonical correlation analysis on data with censoring and error information," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 1909–1919, Dec. 2013.

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

[13] 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.

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

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

[16] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. Int. Conf. Mach. Learn.*, Banff, AB, Canada, 2004, p. 6.

[17] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. SIAM Int. Conf. Data Min.*, Austin, TX, USA, 2013, pp. 252–260.

[18] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. Int. Joint Conf. Artif. Intell.*, Austin, TX, USA, 2015, pp. 2750–2756.

[19] W. Tang, Z. Lu, and I. S. Dhillon, "Clustering with multiple graphs," in *Proc. IEEE Int. Conf. Data Min.*, 2009, pp. 1016–1021.

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

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

[22] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018.

[23] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proc. Int. Joint Conf. Artif. Intell.*, Quebec City, QC, Canada, 2014, pp. 2149–2155.

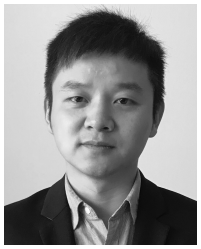
[24] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 4238–4246.

[25] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Latent multi-view subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 4279–4287.

[26] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

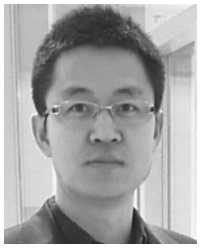
[27] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2586–2593.

[28] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, 2014.
[29] T. Zhou *et al.*, "Online discriminative dictionary learning for robust object tracking," *Neurocomputing*, vol. 275, pp. 1801–1812, Jan. 2018.
[30] T. Zhou, F. Liu, H. Bhaskar, and J. Yang, "Robust visual tracking via online discriminative and low-rank dictionary learning," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2643–2655, Sep. 2018.
[31] S. Gao, I. W.-H. Tsang, and Y. Ma, "Learning category-specific dictionary and shared dictionary for fine-grained image categorization," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 623–634, Feb. 2014.
[32] J. Zheng, Z. Jiang, and R. Chellappa, "Cross-view action recognition via transferable dictionary learning," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2542–2556, Jun. 2016.
[33] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.
[34] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 129–136.
[35] G. Liu *et al.*, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[36] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4322–4334, Nov. 2015.
[37] L. Zhang *et al.*, "Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding," *Pattern Recognit.*, vol. 48, no. 10, pp. 3102–3112, 2015.
[38] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1582–1590.
[39] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf, "Measuring statistical dependence with Hilbert–Schmidt norms," in *Proc. Int. Conf. Algorithm Learn. Theory*, vol. 16. Singapore, 2005, pp. 63–78.
[40] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 586–594.
[41] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, 2011, pp. 612–620.
[42] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[43] J. Huang, F. Nie, and H. Huang, "Spectral rotation versus K-means in spectral clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, Bellevue, WA, USA, 2013, pp. 431–437.
[44] D. Dueck and B. J. Frey, "Non-metric affinity propagation for unsupervised image categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.
[45] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. New York, NY, USA, 2006, pp. 1447–1454.
[46] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu, "Recognizing actions from still images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Tampa, FL, USA, 2008, pp. 1–4.
[47] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2002, pp. 849–856.
[48] C.-G. Li and R. Vidal, "Structured sparse subspace clustering: A unified optimization framework," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 277–286.
[49] H. Abdi and L. J. Williams, "Principal component analysis," *Interdiscipl. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.
[50] V. R. De Sa, "Spectral clustering with two views," in *Proc. Int. Conf. Mach. Learn. Workshop Learn. Multiple Views*, Lille, France, 2005, pp. 20–27.
[51] I. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2002.
[52] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, "Learning a spatially smooth subspace for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, 2007, pp. 1–7.
[53] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao, "Flexible multi-view dimensionality co-reduction," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 648–659, Feb. 2017.
[54] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2004, pp. 153–160.
[55] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 593–600.
[56] J. Wang *et al.*, "Diverse non-negative matrix factorization for multiview data representation," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2620–2632, Sep. 2017.
[57] Z. Zhao *et al.*, "Efficient spectral feature selection with minimum redundancy," in *Proc. AAAI Conf. Artif. Intell.*, Atlanta, GA, USA, 2010, pp. 673–678.
[58] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015.
[59] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, "Feature selection with ensembles, artificial variables, and redundancy elimination," *J. Mach. Learn. Res.*, vol. 10, pp. 1341–1366, Jul. 2009.
[60] J. Shawe-Taylor *et al.*, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[61] J. T. Zhou *et al.*, "SC2Net: Sparse LSTMs for sparse coding," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 4588–4595.
[62] J. Wang *et al.*, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.
[63] J. T. Zhou *et al.*, "Transfer hashing: From shallow to deep," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6191–6201, Dec. 2018.
[64] X. Lan *et al.*, "Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2022–2037, Apr. 2018.
[65] T. Zhou, H. Bhaskar, F. Liu, and J. Yang, "Graph regularized and locality-constrained coding for robust visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2153–2164, Oct. 2017.
[66] X. Lan *et al.*, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5826–5841, Dec. 2015.
[67] X. Lan *et al.*, "Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1194–1201.
[68] J. T. Zhou, I. W. Tsang, S. S. Ho, and K. R. Muller, "N-ary decomposition for multi-class classification," *Mach. Learn. J.*, 2018.
[69] C. Gong, X. Chang, M. Fang, and J. Yang, "Teaching semi-supervised classifier via generalized distillation," in *Proc. Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, 2018, pp. 2156–2162.
[70] Y. Zhang *et al.*, "Temporally constrained sparse group spatial patterns for motor imagery BCI," *IEEE Trans. Cybern.*, to be published.
[71] Y. Zhang *et al.*, "Sparse Bayesian classification of EEG for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2256–2267, Nov. 2016.
[72] C. Gong *et al.*, "A regularization approach for instance-based superset label learning," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 967–978, Mar. 2018.



Tao Zhou (M'17) received the M.S. degree in computer application technology from Jiangnan University, Wuxi, China, in 2012 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2016.

His current research interests include machine learning, computer vision, and medical image analysis.



Changqing Zhang received the B.S. and M.S. degrees in computer science from the College of Computer, Sichuan University, Chengdu, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2016.

He is an Assistant Professor with the College of Intelligence and Computing, Tianjin University. He has been a Postdoctoral Research Fellow with the Department of Radiology and BRIC, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. His current research interests include machine learning, computer vision, and medical image analysis.

Chen Gong (M'16) received the B.E. degree in automation from the East China University of Science and Technology, Shanghai, China, in 2010, the first Doctoral degree in pattern recognition and intelligent systems from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2016, and the second Doctoral degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2017.

He is currently a Full Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. He has published over 50 technical papers at prominent journals and conferences, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, CVPR, AAAI, IJCAI, and ICDM. His current research interests include machine learning, data mining, and learning-based vision problems.

Dr. Gong was a recipient of the "Excellent Doctoral Dissertation" awarded by SJTU and the Chinese Association for Artificial Intelligence. He was also enrolled by the "Summit of the Six Top Talents" Program of Jiangsu Province, China, and the "Lift Program for Young Talents" of the China Association for Science and Technology. He also serves as the Reviewer for over 20 international journals, such as *Artificial Intelligence Journal*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is also the PC Member of several top-tier conferences, such as ICML, AAAI, IJCAI, ICDM, and AISTATS.

Harish Bhaskar received the Ph.D. degree in computer science from Loughborough University, Loughborough, U.K., in 2007.

He was a Research Associate with the University of Manchester, Manchester, U.K., and Lancaster University, Lancaster, U.K. He is currently an Assistant Professor with the Visual Signal Analysis and Processing Research Center, Khalifa University, Abu Dhabi, UAE. His current research interests include computer vision, image processing, visual data mining, medical imaging, quantum information, and machine vision.

Jie Yang received the Ph.D. degree in computer science from Hamburg University, Hamburg, Germany, in 1994.

He is currently a Professor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. He has led many research projects, had one book published in Germany, and authored over 200 journal papers. His current research interests include object detection and recognition, data fusion and data mining, and medical image processing.