Multi-View Data Fusion Oriented Clustering via Nuclear Norm Minimization

Aiping Huang, *Member, IEEE*, Tiesong Zhao, *Senior Member, IEEE*, and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*— Image clustering remains challenging when handling image data from heterogeneous sources. Fusing the independent and complementary information existing in heterogeneous sources together facilitates to improve the image clustering performance. To this end, we propose a joint learning framework of multi-view image data fusion and clustering based on nuclear norm minimization. Specifically, we first formulate the problem as matrix factorization to a shared clustering indicator matrix and a representative coefficient matrix. The former is constrained with orthogonality and nonnegativity, which ensures the validation of clustering assignments. The latter is imposed with nuclear norm minimization to achieve compression of principal components for performance improvement. Then, an alternating minimization strategy is employed to efficiently decompose the multi-variable optimization problem into several small solvable sub-problems with closed-form solutions. Extensive experimental results on real-world image and video datasets demonstrate the superiority of proposed method over other state-of-the-art methods.**

***Index Terms*— Unsupervised learning, data fusion, multi-view clustering, matrix factorization, nuclear norm.**

I. Introduction

A VARIETY of image datasets in real world comprise various multi-view representations. As illustrated in Fig. 1, an image can be represented by different features, an animation can be encoded as a video or as an audio, and a scenario can be described either in images or in texts. Given specific learning tasks, discovering hidden patterns and latent semantics from these views refers to multi-view learning. Extensive studies [1]–[4] have revealed that multi-view learning is more effective, robust, promising and general than single-view learning, because it considers the diversity of different views and exploits the joint merits of these views. Therefore, multi-view technology is of great potential for benefiting many applications such as image recognition [5], image segmentation [6], natural language processing [7] and multimedia understanding [8].

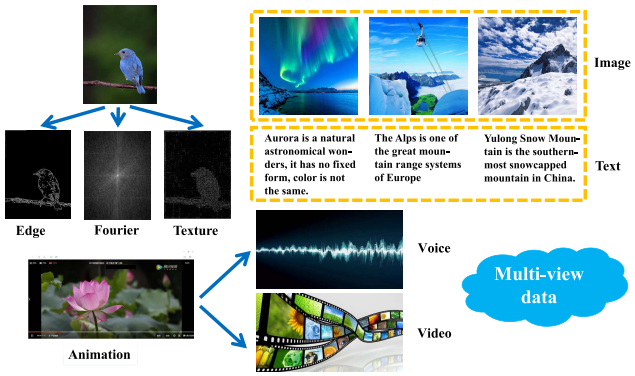


Fig. 1. Examples of multi-view data.

Multi-view clustering aims at grouping data points into a certain number of patterns by exploiting compatible and complementary information of multi-view data. The past decade has witnessed a booming of multi-view clustering methods based on Canonical Correlation Analysis (CCA), matrix factorization, subspace and spectral. CCA is an important technique for identifying relationships among different views and simultaneously constructing a common space. It has been widely used in multi-view scenario for addressing the clustering problem of paired data [9]–[11] and incomplete view [12], [13]. Matrix factorization-based methods [14], [15] are equivalent to the relaxed K-means by decomposing the feature matrix into a centroid matrix and the cluster assignment. Subspace clustering [16], [17] aims at learning a latent low-dimensional representation to correctly cluster data points based on the learned features. Recently, many subspace-based methods have been proposed, including diversity-induced [18], multi-manifold regularized [19], deep low-rank ensemble [20], correlation consensus based [21], and tensor based [22], [23] multi-view clustering. Spectral-based clustering involves low-dimensional embedding of the affinity matrix between samples, followed by K-means clustering. It is a representative model of unsupervised learning with several variants developed for multi-view clustering, such as co-regularized [24], co-training [25], robust [26], low-rank [27], [28], one-step [29], kernel based [30] and essential tensor [3] multi-view spectral clustering.

Manuscript received December 3, 2019; revised May 22, 2020, August 2, 2020, and September 21, 2020; accepted October 1, 2020. Date of publication October 15, 2020; date of current version October 21, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61671152 and Grant 61901119. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiangqian Wu. *(Corresponding author: Tiesong Zhao.)*

Aiping Huang is with the Fujian Key Laboratory for Intelligent Processing and Wireless Transmission of Media Information, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China (e-mail: sxxhap@163.com).

Tiesong Zhao is with the Fujian Key Laboratory for Intelligent Processing and Wireless Transmission of Media Information, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China, and also with the Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou 350108, China (e-mail: t.zhao@fzu.edu.cn).

Chia-Wen Lin is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan, and also with the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Digital Object Identifier 10.1109/TIP.2020.3029883

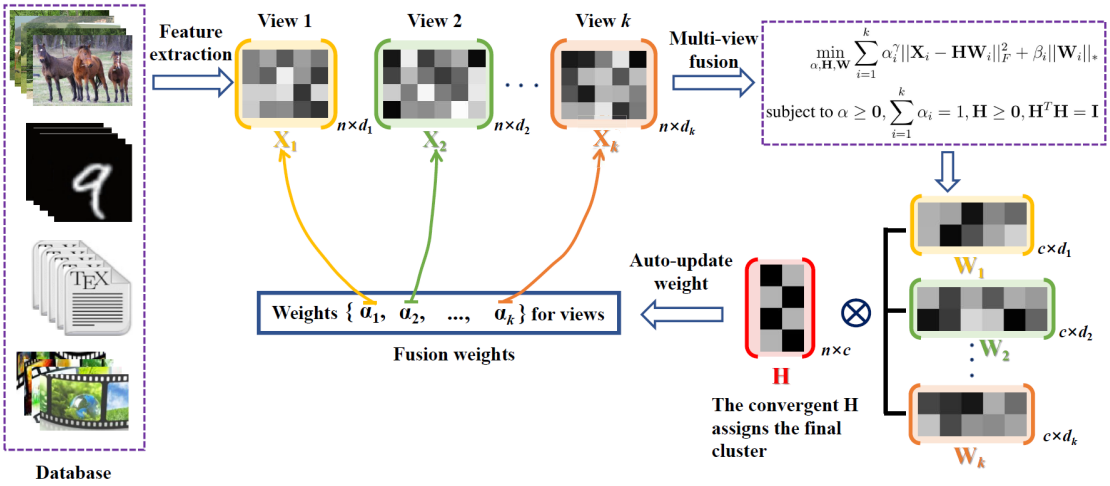


Fig. 2. Framework for the proposed method. In our method, features of images from k views are extracted, then the proposed MVDFC problem is formulated as a canonical form of matrix factorization of a shared indicator matrix and respective representative coefficient matrices.

Certainly, there are still plenty of substantial research results that are not mentioned above, such as graph learning based [31], [32], consensus guided [21], [33] and incomplete [34] multi-view clustering. However, most of these multi-view clustering methods mainly focused on finding a reduced subspace structure, and then employing a two-step pipeline to obtain features and cluster sequentially, which often results in a suboptimal clustering performance. Moreover, it is crucial to scoop out favorable geometric structures from the given data that may overlap. To this end, we propose a joint learning framework for both multi-view data fusion and clustering in this paper. The former task is to search a compact feature fusion with regularized structured learning, while the latter aims to achieve accurate clustering with a tailored low-dimensional representation.

As demonstrated in Fig. 2, the proposed framework first extracts the features of images from k views, and then formulates the proposed multi-view data fusion oriented clustering (MVDFC) problem as a canonical form of matrix factorization of a shared indicator matrix and respective representative coefficient matrices. Through iteratively minimizing the canonical form and updating view weight, we ultimately obtain a shared indicator matrix for assigning the cluster labels. Note that the shared matrix is imposed with orthogonality and non-negativity constraints to ensure the validation of the clustering indicator matrix. Simultaneously, the coefficient matrices are constrained with nuclear norm minimization, aiming at capturing principal components of different views for performance improvement. The main contributions of this paper are summarized as follows:

- Proposing a joint model that integrates multiple views into a comprehensive representation to simultaneously learn multi-view fusion weights and make clustering assignments.
- Factorizing the proposed problem formulation into several small sub-problems, and then employing an alternative optimization method to solve the problems by deriving respective closed-form solutions.
- Proposing an efficient iterative optimization algorithm to properly make clustering assignments and update fusion weights.
- Conducting extensive experiments on several real-world image and video datasets to demonstrate the superiority of our method over the compared state-of-the-art methods.

The rest of this paper is arranged as follows. Section II introduces some related work to this paper. In Section III, a joint learning framework of multi-view data fusion and clustering is proposed and the corresponding iterative algorithm is developed. Section IV presents the experimental results and analyses. We explore the extension of the proposed framework to general multi-view learning in Section V. Finally, this paper is concluded in Section VI.

II. Related Work

Multi-view data captures rich compatible and complementary information among diverse data sources, which can be beneficial to clustering tasks.

Early multi-view clustering studies mainly focused on handling two-view data. For example, [35] extended K-means to settle the clustering problem of two conditionally independent views. Reference [24] proposed a co-regularized multi-view spectral clustering method in which the eigenvectors of graph Laplacian are adopted as predictor functions. To incorporate more views, [36] developed a multiple graph framework to automatically learn optimal weights of different views. Reference [37] presented a robust large-scale multi-view clustering method based on K-means, in which $L_{2,1}$-norm was introduced to improve the robustness when handling outliers. In [26], a shared low-rank transition probability matrix from different views was recovered and then fed to a standard Markov chain model for clustering. Reference [21] employed correlation consensus among multi-view data to formulate a subspace clustering framework. Reference [27] designed an iterative low-rank based structured optimization algorithm to preserve the local manifold structure of heterogeneous views. Reference [38] proposed a group-aware multi-view

TABLE I
SUMMARY OF KEY NOTATIONS USED IN THE PAPER

Notations	Explanations
$\{\mathbf{X}_i\}_{i=1}^k$	multi-view data of k views with $\mathbf{X}_i \in \mathbb{R}^{n\times d_i}$
$\mathbf{H} \in \mathbb{R}^{n\times c}$	the shared clustering indicator matrix
$\widehat{\mathbf{H}} \in \mathbb{R}^{n\times c}$	the optimal solution of $\mathbf{H}$
$\mathbf{W}_i \in \mathbb{R}^{c\times d_i}$	clustering coefficient matrix for the i-th view
$\widehat{\mathbf{W}}_i \in \mathbb{R}^{c\times d_i}$	the optimal solution of $\mathbf{W}_i$
$\mathbf{X}_i^{test}$	test data for the i-th view in supervised scene
$\mathbf{Y}^{test}$	the test class label in supervised scene
$\mathbf{Y} \in \mathbb{R}^l$	the column vector of class labels
$\mathbf{S} \in \mathbb{R}^{n\times n}$	predefined matrix for prior knowledge
$\alpha \in \mathbb{R}^k$	the weight vector for different views
$\widehat{\alpha} \in \mathbb{R}^k$	the optimal value of α
t	the number of iterations
t_{max}	the maximum number of iterations
n, k	the numbers of samples and views
d_i	the dimension of the i-th view data matrix
c	the number of clusters / classes
N, l	the numbers of training and labeled samples

fusion approach with pair-wise fusion and center-wise fusion to improve fusion accuracy while reducing computational complexity.

The aforementioned methods perform clustering in a separate procedure, *i.e.*, first learning a shared representation for multi-view features, then employing a traditional method such as K-means to produce the final clusters. Similar approaches also include CCA-based multi-view clustering [11]–[13]. They project the multi-view high-dimensional data onto one common low-dimensional space, where existing clustering methods are subsequently conducted. However, the separate clustering steps usually result in suboptimal performance because the relationship between multi-view feature learning and clustering is not well exploited.

Recently, several advanced methods have been proposed. For example, in [4], Nie *et al*. designed a multi-view model to perform clustering and local structure learning simultaneously; In [39], they proposed a self-weighted method to directly assign the cluster label for each data point without any postprocessing; In [40], they presented an adaptively weighted procrustes technique to learn an indicator matrix for multi-view clustering. In addition, Huang *et al*. [41] used kernel spaces to learn similarity relationships among different views and perform multi-view clustering task simultaneously. Different from the above methods, we propose a joint framework for multi-view data fusion and clustering assignments, where orthogonality and nonnegativity are imposed to constrain the clustering indicator matrix, and nuclear norm minimization is utilized to restrict the coefficient matrices.

III. JOINT MULTI-VIEW DATA FUSION AND CLUSTERING VIA NUCLEAR NORM MINIMIZATION

In this section, we propose a joint learning framework of multi-view data fusion and clustering with an effective optimization algorithm. The key notations used in the paper are summarized in Table I.

A. Problem Formulation

For an input of multi-view data sources $\{\mathbf{X}_i\}_{i=1}^k$ with the view number k and $\mathbf{X}_i \in \mathbb{R}^{n\times d_i}$, each view is regarded as a consolidated description of data points. MVDFC is to fuse all collected features into a unified representation to learn an efficient low-dimensional embedding for the clustering task. It is noted that all individual views share the same clustering indicator matrix but possess diverse representation coefficient matrices. Consequently, the objective function of MVDFC can be abstracted as the following canonical form

$$\min_{\mathbf{H},\mathbf{W}} \sum_{i=1}^{k} \ell(\mathbf{X}_i, \mathbf{H}\mathbf{W}_i) + \phi(\mathbf{H}) + \psi(\mathbf{W}_i). \tag{1}$$

Here, $\mathbf{H} \in \mathbb{R}^{n\times c}$ is a shared clustering indicator matrix to be learned, $\mathbf{W} = [\mathbf{W}_1, \cdots, \mathbf{W}_k] \in \mathbb{R}^{c\times \sum_{i=1}^{k} d_i}$ with $\mathbf{W}_i$ being the coefficient matrix associated with the i-th view, $\ell(\circ, \circ)$ is a specific loss function, and $\phi(\mathbf{H})$, $\psi(\mathbf{W}_i)$ are respective constraints for $\mathbf{H}$, $\mathbf{W}_i$. Intuitively, data from various views are inclined to capture different properties, which motivates us to utilize an adaptive weight vector, instead of equal importance, to highlight significant views. Therefore, the above optimization problem is further represented as

$$\min_{\alpha,\mathbf{H},\mathbf{W}} \sum_{i=1}^{k} \alpha_i^{\gamma} \ell(\mathbf{X}_i, \mathbf{H}\mathbf{W}_i) + \phi(\mathbf{H}) + \psi(\mathbf{W}_i). \tag{2}$$

where $\alpha = [\alpha_1, \cdots, \alpha_k]^T \in \mathbb{R}^k$ is the vector of weights for the k views and $\gamma \geq 0$ is a predefined constant to enhance the metric learning of loss function.

Next, we will explore the specific forms of the loss function and constraints. We aim to directly derive an indicator matrix $\mathbf{H}$ for clustering assignment from a joint framework without any postprocessing, which induces two necessary but not sufficient conditions, *i.e.*, $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ and $\mathbf{H} \geq \mathbf{0}$. Meanwhile, the coefficient matrix $\mathbf{W}_i$ is restricted with a compact compression and further represented as nuclear norm minimization to capture the principal components of different views. When specifying the loss function based on F-norm, the objective function of MVDFC is written as

$$\begin{aligned} &\min_{\alpha,\mathbf{H},\mathbf{W}} \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{H}\mathbf{W}_i||_F^2 + \beta_i ||\mathbf{W}_i||_* \\ &\text{subject to } \alpha \geq \mathbf{0}, \sum_{i=1}^{k} \alpha_i = 1, \mathbf{H} \geq \mathbf{0}, \mathbf{H}^T\mathbf{H} = \mathbf{I}, \end{aligned} \tag{3}$$

where the conditions $\alpha \geq \mathbf{0}$ and $\sum_{i=1}^{k} \alpha_i = 1$ jointly guarantee the regularization of α.

Due to the orthogonality constraint $\mathbf{H}^T\mathbf{H} = \mathbf{I}$, the optimization problem becomes a high-order (fourth-order) matrix factorization problem, which makes it difficult to solve. To address the problem, a relaxation trick is proposed. Let $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_c] \in \mathbb{R}^{n\times c}$ where $\mathbf{h}_i \in \mathbb{R}^n$ is a column vector for $i \in \{1, 2, \cdots, c\}$. It is clear that

$$\mathbf{h}_i^T\mathbf{h}_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \tag{4}$$

implying $\sum_{j\neq i}\mathbf{h}_i^T\mathbf{h}_j = 0$ which is equivalently represented as

$$Tr(\mathbf{H}\mathbf{\Theta}\mathbf{H}^T) = 0 \tag{5}$$

where $\mathbf{\Theta} \in \mathbb{R}^{c\times c}$ is a constant matrix, defined as

$$\mathbf{\Theta} = \begin{bmatrix} 0 & 1 & \cdots & 1 & 1 \\ 1 & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}. \tag{6}$$

Consequently, the objective function of MVDFC is then rewritten as

$$\min_{\alpha,\mathbf{H},\mathbf{W}} \sum_{i=1}^{k} \alpha_i^{\gamma} \|\mathbf{X}_i - \mathbf{H}\mathbf{W}_i\|_F^2 + \beta_i \|\mathbf{W}_i\|_* + \lambda Tr(\mathbf{H}\mathbf{\Theta}\mathbf{H}^T)$$
$$\text{subject to } \alpha \geq \mathbf{0}, \mathbf{1}^T\alpha = 1, \mathbf{H} \geq \mathbf{0}, \tag{7}$$

where $\mathbf{1} \in \mathbb{R}^k$ is a column vector whose entries are equal to one, and $\lambda > 0$ is a parameter to weigh the importance of the orthogonality of $\mathbf{H}$. Notice that $\mathbf{\Theta}$ can be extended to be any graph Laplacian matrix.

B. Alternating Optimization Algorithm

In order to settle the aforementioned optimization problem, two intermediate optimization matrix variables $\mathbf{G} \in \mathbb{R}^{n\times c}$ and $\mathbf{V} = [\mathbf{V}_1, \cdots, \mathbf{V}_k] \in \mathbb{R}^{c\times \sum_{i=1}^{k} d_i}$ are introduced to separate $\mathbf{H}$ and $\mathbf{W}$. Consequently, Equation (7) can be rewritten as

$$\min_{\alpha,\mathbf{H},\mathbf{G},\mathbf{W},\mathbf{V}} \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i\|_F^2 + \beta_i \|\mathbf{W}_i\|_* + \lambda Tr(\mathbf{G}\mathbf{\Theta}\mathbf{G}^T)$$
$$\text{subject to } \alpha \geq \mathbf{0}, \mathbf{1}^T\alpha = 1, \mathbf{H} \geq \mathbf{0}, \mathbf{H}-\mathbf{G}=\mathbf{0}, \mathbf{W}-\mathbf{V}=\mathbf{0}. \tag{8}$$

With the alternating direction method of multipliers (ADMM), the optimization problem can be equivalently transformed into minimizing

$$\mathcal{L}(\alpha, \mathbf{H}, \mathbf{G}, \mathbf{W}, \mathbf{V}, \mathbf{Y}_1, \mathbf{Y}_2) = \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2$$
$$+ \beta_i ||\mathbf{W}_i||_* + \lambda Tr(\mathbf{G}\mathbf{\Theta}\mathbf{G}^T) + Tr(\mathbf{Y}_1^T(\mathbf{W}-\mathbf{V}))$$
$$+ \frac{\rho_1}{2}||\mathbf{W}-\mathbf{V}||_F^2 + Tr(\mathbf{Y}_2^T(\mathbf{H}-\mathbf{G})) + \frac{\rho_2}{2}||\mathbf{H}-\mathbf{G}||_F^2$$
$$\text{subject to } \alpha \geq \mathbf{0}, \mathbf{1}^T\alpha = 1, \mathbf{H} \geq \mathbf{0}, \tag{9}$$

where $\mathbf{Y}_1 \in \mathbb{R}^{c\times \sum_{i=1}^{k} d_i}$ and $\mathbf{Y}_2 \in \mathbb{R}^{n\times c}$ can be regarded as two penalty matrices for narrowing $\mathbf{W}-\mathbf{V}$ and $\mathbf{H}-\mathbf{G}$ to zero matrix $\mathbf{0}$.

The optimization problem (9) can then be solved by the following iterative updating rules, with initialization settings α^0, $\mathbf{H}^0$, $\mathbf{G}^0$, $\mathbf{W}^0$, $\mathbf{V}^0$, $\mathbf{Y}_1^0$, $\mathbf{Y}_2^0$:

$$\alpha^{t+1} = \arg\min_{\alpha} \mathcal{L}(\alpha, \mathbf{H}^t, \mathbf{G}^t, \mathbf{W}^t, \mathbf{V}^t, \mathbf{Y}_1^t, \mathbf{Y}_2^t), \tag{10}$$
$$\mathbf{H}^{t+1} = \arg\min_{\mathbf{H}} \mathcal{L}(\alpha^{t+1}, \mathbf{H}, \mathbf{G}^t, \mathbf{W}^t, \mathbf{V}^t, \mathbf{Y}_1^t, \mathbf{Y}_2^t), \tag{11}$$
$$\mathbf{G}^{t+1} = \arg\min_{\mathbf{G}} \mathcal{L}(\alpha^{t+1}, \mathbf{H}^{t+1}, \mathbf{G}, \mathbf{W}^t, \mathbf{V}^t, \mathbf{Y}_1^t, \mathbf{Y}_2^t), \tag{12}$$
$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}} \mathcal{L}(\alpha^{t+1}, \mathbf{H}^{t+1}, \mathbf{G}^{t+1}, \mathbf{W}, \mathbf{V}^t, \mathbf{Y}_1^t, \mathbf{Y}_2^t), \tag{13}$$
$$\mathbf{V}^{t+1} = \arg\min_{\mathbf{V}} \mathcal{L}(\alpha^{t+1}, \mathbf{H}^{t+1}, \mathbf{G}^{t+1}, \mathbf{W}^{t+1}, \mathbf{V}, \mathbf{Y}_1^t, \mathbf{Y}_2^t), \tag{14}$$
$$\mathbf{Y}_1^{t+1} = \mathbf{Y}_1^t + \rho_1(\mathbf{W}^{t+1} - \mathbf{V}^{t+1}), \tag{15}$$
$$\mathbf{Y}_2^{t+1} = \mathbf{Y}_2^t + \rho_2(\mathbf{H}^{t+1} - \mathbf{G}^{t+1}). \tag{16}$$

Next, we will find the optimal solutions to all the above subproblems with respect to α, $\mathbf{H}$, $\mathbf{G}$, $\mathbf{W}$ and $\mathbf{V}$, respectively.

Updating α when keeping H, G, W and V. Eliminating constant terms of $\mathcal{L}(\alpha, \mathbf{H}, \mathbf{G}, \mathbf{W}, \mathbf{V}, \mathbf{Y}_1, \mathbf{Y}_2)$, the subproblem of updating α is written as

$$\min_{\alpha} \mathcal{J} = \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2 \quad \text{subject to } \alpha \geq \mathbf{0}, \mathbf{1}^T\alpha = 1. \tag{17}$$

Using the Lagrange multiplier method, we construct a Lagrangian function as

$$\mathcal{L}(\alpha, \mu) = \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2 - \mu(\mathbf{1}^T\alpha - 1). \tag{18}$$

Taking the derivative of $\mathcal{L}(\alpha, \mu)$ with respect to α and μ,

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \gamma \alpha_i^{\gamma-1} ||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2 - \mu, \quad \frac{\partial \mathcal{L}}{\partial \mu} = \mathbf{1}^T\alpha - 1. \tag{19}$$

Setting $\frac{\partial \mathcal{L}}{\partial \alpha} = \mathbf{0}$ and $\frac{\partial \mathcal{L}}{\partial \mu} = 0$, we have

$$\alpha_i = \frac{\left(||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2\right)^{1/(1-\gamma)}}{\sum_{i=1}^{k}\left(||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2\right)^{1/(1-\gamma)}}. \tag{20}$$

Updating H when keeping α, G, W and V. The subproblem of updating $\mathbf{H}$ aims to solve the minimization problem

$$\min_{\mathbf{H}\geq\mathbf{0}} \frac{\rho_2}{2}||\mathbf{H}-\mathbf{G}||_F^2 + Tr(\mathbf{Y}_2^T(\mathbf{H}-\mathbf{G})) \tag{21}$$
$$= \min_{\mathbf{H}\geq\mathbf{0}} \frac{\rho_2}{2}||\mathbf{H}-\mathbf{G}+\frac{1}{\rho_2}\mathbf{Y}_2||_F^2 - \frac{1}{2\rho_2}||\mathbf{Y}_2||_F^2 \tag{22}$$

which refers to the following closed-form solution:

$$\mathbf{H}^* = [\mathbf{G} - \frac{1}{\rho_2}\mathbf{Y}_2]_+ \tag{23}$$

where $[a]_+$ is the positive part of a, *i.e.*, $[a]_+ = \max\{a, 0\}$. Once the clustering indicator matrix $\mathbf{H}^*$ is achieved, the clustering label y_i for input sample x_i can be calculated by the following decision function:

$$y_i = \arg\max_j \mathbf{H}_{ij}^*.$$
$$\forall i = 1, 2, \cdots, n. \ \forall j = 1, 2, \cdots, c. \tag{24}$$

Updating G when keeping α, H, W and V. The subproblem of updating $\mathbf{G}$ is formulated as

$$\min_{\mathbf{G}} \mathcal{J} = \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2 + \lambda Tr(\mathbf{G}\mathbf{\Theta}\mathbf{G}^T) + Tr(\mathbf{Y}_2^T(\mathbf{H}-\mathbf{G})) + \frac{\rho_2}{2}||\mathbf{H}-\mathbf{G}||_F^2. \quad (25)$$

Notice that the above optimization problem is unconstrained, hence its optimal solution is attained at $\frac{\partial \mathcal{J}}{\partial \mathbf{G}} = \mathbf{0}$. Setting the derivative of $\mathcal{J}$ with respect to $\mathbf{G}$ to zero, we have

$$\frac{\partial \mathcal{J}}{\partial \mathbf{G}} = \sum_{i=1}^{k} \alpha_i^{\gamma} (-2\mathbf{X}_i\mathbf{V}_i^T + 2\mathbf{G}\mathbf{V}_i\mathbf{V}_i^T) + 2\lambda\mathbf{G}\mathbf{\Theta} - \mathbf{Y}_2 + \rho_2(\mathbf{G}-\mathbf{H}) \triangleq \mathbf{0} \quad (26)$$

which leads to the optimal value

$$\mathbf{G}^* = \left(\sum_{i=1}^{k} \alpha_i^{\gamma} \mathbf{X}_i\mathbf{V}_i^T + \frac{\mathbf{Y}_2}{2} + \frac{\rho_2\mathbf{H}}{2}\right)\left(\sum_{i=1}^{k} \alpha_i^{\gamma} \mathbf{V}_i\mathbf{V}_i^T + \lambda\mathbf{\Theta} + \frac{\rho_2\mathbf{I}}{2}\right)^{-1}. \quad (27)$$

Updating W when keeping α, H, G, and V. The subproblem of updating $\mathbf{W}$ is equivalent to

$$\min_{\mathbf{W}} \mathcal{J} = \sum_{i=1}^{k} \frac{\rho_1}{2}||\mathbf{W}_i - \mathbf{V}_i||_F^2 + Tr(\mathbf{Y}_{1,i}^T(\mathbf{W}_i - \mathbf{V}_i)) + \beta_i||\mathbf{W}_i||_* = \min_{\mathbf{W}} \sum_{i=1}^{k} \frac{\rho_1}{2}||\mathbf{W}_i - \mathbf{V}_i + \frac{1}{\rho_1}\mathbf{Y}_{1,i}||_F^2 + \beta_i||\mathbf{W}_i||_* - \frac{1}{2\rho_1}||\mathbf{Y}_{1,i}||_F^2 \quad (28)$$

where $\mathbf{Y}_1 = [\mathbf{Y}_{1,1}, \cdots, \mathbf{Y}_{1,k}]$ with $\mathbf{Y}_{1,i} \in \mathbb{R}^{c\times d_i}$. The above equation follows a closed-form solution with

$$\mathbf{W}_i^* = \mathbf{U}\mathcal{D}_{\beta_i/\rho_1}(\mathbf{\Sigma})\mathbf{V}^T \quad (29)$$

in which $\mathbf{V}_i - \frac{1}{\rho_1}\mathbf{Y}_{1,i} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition (SVD) and $\mathcal{D}_{\tau}(\mathbf{\Sigma})$ is the soft-thresholding operator [42], *i.e.*,

$$\mathcal{D}_{\tau}(\mathbf{\Sigma}) = diag(\left[[\mathbf{\Sigma}_{11} - \tau]_+, \cdots, [\mathbf{\Sigma}_{cc} - \tau]_+\right]). \quad (30)$$

Updating V when keeping α, H, G, and W. The subproblem of updating $\mathbf{V}$ is transformed into

$$\min_{\mathbf{V}} \mathcal{J} = \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{G}\mathbf{V}_i||_F^2 + Tr(\mathbf{Y}_1^T(\mathbf{W}-\mathbf{V})) + \frac{\rho_1}{2}||\mathbf{W}-\mathbf{V}||_F^2. \quad (31)$$

Similar to Problem (25), the above optimization issue is unconstrained, hence its optimal value is attained by setting $\frac{\partial \mathcal{J}}{\partial \mathbf{V}} = \mathbf{0}$. Taking the derivative of $\mathcal{J}$ with respect to $\mathbf{V}_i$, we obtain

$$\frac{\partial \mathcal{J}}{\partial \mathbf{V}_i} = \alpha_i^{\gamma}(-2\mathbf{G}^T\mathbf{X}_i + 2\mathbf{G}^T\mathbf{G}\mathbf{V}_i) - \mathbf{Y}_{1,i} + \rho_1(\mathbf{V}_i - \mathbf{W}_i). \quad (32)$$

Setting $\frac{\partial \mathcal{J}}{\partial \mathbf{V}_i} = \mathbf{0}$, we can compute the optimal solution as

$$\mathbf{V}_i^* = \left(2\alpha_i^{\gamma}\mathbf{G}^T\mathbf{G} + \rho_1\mathbf{I}\right)^{-1}\left(2\alpha_i^{\gamma}\mathbf{G}^T\mathbf{X}_i + \mathbf{Y}_{1,i} + \rho_1\mathbf{W}_i\right). \quad (33)$$

Algorithm 1 Algorithm for MVDFC

Input:
1: The data matrices: $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_k$.
2: Number of clusters: c.
3: The parameters: nonlinear factor γ, regularization parameter β_i, weighted coefficient λ and $\{\rho_i\}_{i=1}^2$.

Output:
1: The clustering label for input data.
2: The weight vector $\alpha = [\alpha_1, \cdots, \alpha_k]^T$.

Initialization:
1: Initialize $\alpha = [\frac{1}{k}, \cdots, \frac{1}{k}]^T$.
2: Initialize matrix variables $\mathbf{H}$, $\mathbf{G}$, $\mathbf{W}$, $\mathbf{V}$ and $\{\mathbf{Y}_i\}_{i=1}^2$.

Procedure:
1: **while** Not converged and $t < t_{max}$ **do**
2: Update α^{t+1} with Equation (20) // Lagrange multiplier method;
3: Update $\mathbf{H}^{t+1}$ with Equation (23) // Closed-form solution;
4: Update $\mathbf{G}^{t+1}$ with Equation (27) // Unconstrained minimization;
5: Update $\mathbf{W}^{t+1}$ with Equation (29) // Soft-thresholding operator;
6: Update $\mathbf{V}^{t+1}$ with Equation (33) // Unconstrained minimization;
7: $\mathbf{Y}_1^{t+1} = \mathbf{Y}_1^t + \rho_1(\mathbf{W}^{t+1} - \mathbf{V}^{t+1})$ // Gradient decent method;
8: $\mathbf{Y}_2^{t+1} = \mathbf{Y}_2^t + \rho_2(\mathbf{H}^{t+1} - \mathbf{G}^{t+1})$ // Gradient decent method;
9: $t \leftarrow t+1$;
10: **end while**
11: Obtain the clustering labels by using (24);
12: **return** The clustering label for input data and the weight vector.

Summarizing the aforementioned analyses of respective optimal solutions to all subproblems, an algorithm for MVDFC is presented in Algorithm 1.

The proposed algorithm employs the ADMM strategy to decompose the multi-variable optimization problem into several small solvable sub-problems. At each iteration, we obtain the closed-form solution of $\mathbf{H}^{t+1}$ and $\mathbf{W}^{t+1}$. In [43], the convergence of ADMM has already been proven. Accordingly, the proposed algorithm converges as well. Meanwhile, experimental evidence on real data also validates the good convergence behavior.

As to the computational complexity of the proposed method, computing $\mathbf{H}$ and $\mathbf{G}$ consumes $O(nc)$ and $O(n^2c + ndc)$. Simultaneously, updating $\mathbf{W}$ and $\mathbf{V}$ respectively requires $O(c^3 + dc^2)$ and $O(ndc)$. Considering that the number c of clusters is much smaller than d and n, the overall computational complexity is $O(n^2 + nd)$.

IV. Experimental Results

In this section, comprehensive experiments on real-world image and video datasets are conducted for performance evaluation of the proposed method.

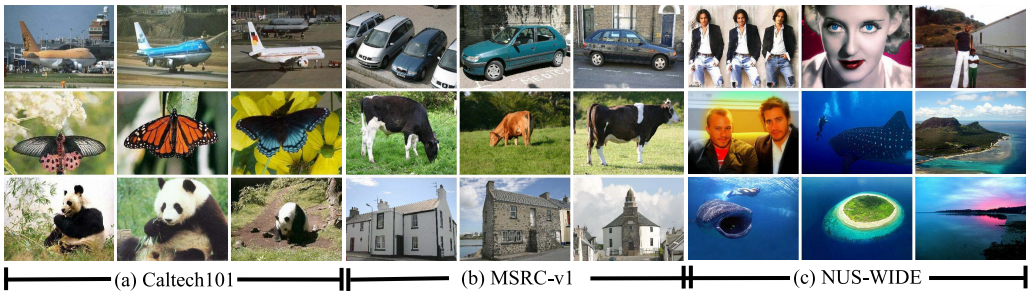


Fig. 3. An illustration of samples images from three selected datasets.

TABLE II
SUMMARY OF THE EVALUATION DATASETS

Datasets	# Samples	# Views	# Total features	# Classes	Data types
ALOI	1,079	4	218	10	Object image
Caltech101	2,386	6	3,766	20	Object image
MNIST	10,000	3	69	10	Digit image
MSRC-v1	210	5	1,892	7	Object image
NUS-WIDE	1,600	6	235	10	Web image
Youtube	2,000	6	1,589	10	Video data

A. Dataset

Six publicly available datasets are employed to conduct fair evaluations for all the compared methods. These datasets are derived from real-world applications and are appropriate for exploring multi-view learning issues.

ALOI is a collection of color images with small foreground objects, which are taken with various viewing angles, illumination directions, and object orientations. Its features are extracted from four views, including RGB color histograms, HSV color histograms, color similarity and Haralick texture features[1].

Caltech101 is an image database with 101 categories, from which we select the most widely used 20 classes, a total of 2,386 samples, each with 3,766 features from six views, including 48-D Gabor feature, 40-D wavelet moments features, 254-D CENTRIST features, 1,984-D HOG features, 512-D GIST features, and 928-D local binary pattern (LBP) features[2].

MNIST is a collection of handwritten digits, from which 10,000 samples are used for testing, along with three views of features produced by IsoProjection with 30 dimensions, linear discriminant analysis with 9 dimensions, and neighborhood preserving embedding with 30 dimensions[3].

MSRC-v1 is a set of 210 images with 7 classes, each class containing 30 images[4]. We extract visual features from several views, namely 24-D CMT features, 576-D HOG features, 512-D GIST features, 254-D CENTRIST features, and 256-D LBP features.

NUS-WIDE is a web image database for object annotation and retrieval[5]. For better fitting multi-view learning problems, we select 1,600 images from 10 categories, each with 160 images. The extracted features consist of 64-D color histogram, 144-D color correlogram, 75-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments, and 500-D bag-of-words.

Youtube is a dataset of multi-view video games[6]. 2,000 samples are selected, each being described from six views consisting of both audio features (mfcc, volume stream, and spectrogram stream) and visual features (cuboids histogram, hist motion estimate, and hog features).

Several sample images are demonstrated in Fig. 3 and a brief description of the aforementioned datasets is presented in Table II.

B. Compared Algorithms and Parameter Selection

In order to fairly evaluate the effectiveness and efficiency of the proposed algorithm, we compare our method with the following six representative and state-of-the-art clustering approaches.

K-means is a classical data clustering method and usually serves as a benchmark of clustering tasks. It is inclined to cluster data into spherical distributions and is sensitive to initial values.

MVCC (multi-view concept clustering) [44] exploits complementary knowledge from multi-view data, rather than solely relying on an individual view. It casts multi-view clustering as the problem of concept decomposition with local manifold structure regularization.

AMGL (auto-weighted graph learning) [36] is founded on multi-graph spectral clustering and searches the optimal weight of each graph. Moreover, it further extends the multi-view clustering problem to semi-supervised learning, and recasts it as a convex optimization problem with a global solution.

MLAN (multi-view learning with adaptive neighbors) [45] is a framework to address multi-view learning and local

[1]http://aloi.science.uva.nl/
[2]http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[3]http://yann.lecun.com/exdb/mnist/
[4]https://www.microsoft.com/en-us/research/project/image-understanding/
[5]https://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
[6]http://archive.ics.uci.edu/ml/datasets/YouTube+Multiview+Video+Games+Dataset

structure embedding simultaneously, that yields an optimal graph with a certain number of connected components, corresponding to exact clustering assignments.

MVKSC (multi-view kernel spectral clustering) [30] utilizes a weighted kernel canonical correlation method to exploit complementary information from multiple views so as to improve performance.

MSC-IAS [46] first utilizes multi-view information to learn an intact space, then constructs the intactness-aware similarity matrix in the space by HSIC, finally employs the spectral clustering on the obtained similarity matrix to perform clustering.

There are several algorithmic parameters need to be set in advance. For the proposed MVDFC, we treat the regularization parameter for each view equally, and its parameters are fixed as follows: the nonlinear factor $\gamma = 2$, regularization parameter $\beta_i = 1$, and weighted coefficient $\lambda = 10$. For all the algorithms for comparison, we implement them through the source codes provided by authors and adopt their default settings if feasible. Due to the initialization sensitivities of clustering methods, all the compared methods are repeated 20 times with random initializations and the corresponding mean values and standard deviations are reported. For single-view clustering method K-means, we concatenate feature vectors of different views together for the all-view clustering setting. As to MVCC, we set the regularization parameter $\alpha = 100$, and two trade-off coefficients $\beta = 100$ and $\gamma = 10$. For AMGL, the maximal number of iterations is fixed as 100, and the number of nearest neighbors for similarity matrix construction is tuned as 5. Regarding MLAN, the number of adaptive neighbors is set as 9, and the maximal number of iterations is set as 30. For MVKSC, the regularization parameter and kernel parameter are set as 1 and 0.1, respectively. For MSC-IAS, its parameters are fixed as follows: $\lambda_2 = 0.1$, the dimension of intact space $d = 500$ and the nearest neighbor number $k = 3$.

C. Evaluation Metrics

In order to assess the clustering performances of all compared methods, we adopt four widely-used objective quality metrics including clustering accuracy (ACC), normalized mutual information (NMI), purity (Purity), and adjusted rand index (ARI).

Given data points $\{x_i\}_{i=1}^n$, let p_i and q_i be the predicted clustering label and the ground truth, respectively. ACC is defined as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(p_i, map(q_i))}{n} \tag{34}$$

with $\delta(a, b) = 1$ if $a = b$, and $\delta(a, b) = 0$ otherwise. Herein, $map(\circ)$ is the best permutation mapping from the predicted clustering label onto an equivalent label of the dataset, which is frequently represented as a maximum matching problem [47].

Assuming the predicted clustering results $\widetilde{\mathbb{C}} = \{\widetilde{C}_i\}_{i=1}^{\widetilde{c}}$ and the ground-truth labels $\mathbb{C} = \{C_j\}_{j=1}^{c}$, NMI is defined as

$$\mathrm{NMI}(\mathbb{C}, \widetilde{\mathbb{C}}) = \frac{\sum_{i=1}^{\widetilde{c}} \sum_{j=1}^{c} |\widetilde{C}_i \cap C_j| log \frac{n|\widetilde{C}_i \cap C_j|}{|\widetilde{C}_i||C_j|}}{\sqrt{(\sum_{i=1}^{\widetilde{c}} |\widetilde{C}_i| log \frac{|\widetilde{C}_i|}{n})(\sum_{j=1}^{c} |C_j| log \frac{|C_j|}{n})}}. \tag{35}$$

A merit of NMI is that it does not necessarily increase with the number of clusters, which makes it a popular metic of clustering quality.

Purity is computed by counting the maximal number of correctly clustered data. Formally, it is represented by

$$\mathrm{Purity}(\mathbb{C}, \widetilde{\mathbb{C}}) = \frac{1}{n} \sum_{i=1}^{\widetilde{c}} \max_j |\widetilde{C}_i \cap C_j|. \tag{36}$$

Here, it is assumed that each predicted cluster is assigned with the class that is the most frequent in the cluster.

As a widely used metric in clustering performance validation, ARI evaluates the similarity between two data clusterings. Let $n_{ij} = |\widetilde{C}_i \cap C_j|$, $a_i = \sum_{j=1}^{c} n_{ij}$ and $b_j = \sum_{i=1}^{\widetilde{c}} n_{ij}$ for all $i \in \{1, \cdots, \widetilde{c}\}$ and $j \in \{1, \cdots, c\}$, from which it follows

$$\mathrm{ARI}(\mathbb{C}, \widetilde{\mathbb{C}}) = \frac{\Sigma_{ij} \binom{n_{ij}}{2} - [\Sigma_i \binom{a_i}{2} \Sigma_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2}[\Sigma_i \binom{a_i}{2} + \Sigma_j \binom{b_j}{2}] - [\Sigma_i \binom{a_i}{2} \Sigma_j \binom{b_j}{2}] / \binom{n}{2}}. \tag{37}$$

It is worth mentioning that the aforementioned $\widetilde{c}$ and c are not necessarily equal. Moreover, ACC, NMI, Purity, and ARI range between 0 to 1, and are positively correlated to the clustering performance.

D. Results and Analyses

We compare the performances of the proposed MVDFC algorithm against that of the other six state-of-the-art algorithms in terms of the four widely used metrics on ALOI, Caltech101, MNIST, MSRC-v1, NUS-WIDE, and Youtube datasets, as reported in Table III, where the numbers in the parentheses are the standard deviations. From the table, we have the following observations.

First, MVDFC outperforms the other state-of-the-art multi-view clustering methods in most cases. Specifically, MVDFC significantly outperforms the other methods on ALOI, MNIST, MSRC-v1 and Youtube, though it just achieves comparable performances on Caltech101 and NUS-WIDE. The superiority of MVDFC lies in the fact that the proposed joint learning framework facilitates to learn a compact low-dimensional representation and make an accurate clustering assignment simultaneously. Meanwhile, the nuclear norm minimization step effectively adjusts the low-dimensional embedding towards enhancing the robustness of the learned data fusion subspace.

Second, the compared methods also exhibit respective strengths in various scenarios. MSC-IAS is a subspace clustering model which can avoid the information loss for insufficient views. By recovering an intact space from multi-view data, MSC-IAS exceeds all the compared algorithms in terms of NMI and Purity on NUS-WIDE. MLAN and AGML are parameter-free methods. They can automatically allocate the weight for each view without additional parameters. They perform well on Caltech101 and MSRC-v1, but achieve limited performances on NUS-WIDE. This instability may due to the constructed graphs are easily affected by the type and quality of data.

Third, multi-view clustering is evidently more promising than single-view one, as can be evidenced from the fact that each multi-view clustering method outperforms single-view K-means. This also indicates that the complementary knowledge

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT MULTI-VIEW CLUSTERING ALGORITHMS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD (THE HIGHER THE BETTER)

Method	ACC (%)						NMI (%)					
	ALOI	Caltech101	MNIST	MSRC-v1	NUS-WIDE	Youtube	ALOI	Caltech101	MNIST	MSRC-v1	NUS-WIDE	Youtube
K-means	48.7 (3.1)	33.0 (2.5)	51.6 (2.5)	46.8 (3.3)	17.2 (0.9)	23.8 (1.2)	48.2 (1.9)	34.4 (1.2)	50.6 (2.2)	40.1 (1.2)	3.90 (0.3)	14.8 (0.5)
MVCC	60.4 (3.9)	43.4 (4.2)	76.6 (4.9)	59.7 (5.1)	-	21.7 (0.3)	59.8 (3.6)	52.3 (2.6)	72.1 (1.5)	55.4 (4.5)	-	11.8 (0.1)
AMGL	51.7 (5.0)	52.0 (1.8)	84.5 (11.6)	65.7 (9.9)	10.4 (0.9)	24.2 (1.4)	54.2 (2.6)	**54.6 (2.2)**	**80.8 (5.3)**	62.1 (6.0)	0.33 (0.0)	14.8 (0.3)
MLAN	60.1 (6.4)	**52.8 (1.0)**	62.4 (1.2)	68.1 (0.0)	13.8 (0.0)	16.7 (0.6)	60.6 (5.1)	47.7 (0.4)	69.2 (0.1)	63.0 (0.0)	3.33 (0.0)	6.54 (0.2)
MVKSC	60.4 (0.0)	28.0 (0.0)	21.9 (0.0)	65.7 (0.0)	16.9 (2.6)	25.1 (0.0)	58.4 (0.0)	32.9 (0.0)	24.1 (0.0)	56.7 (0.0)	2.48 (0.2)	13.9 (0.0)
MSC-IAS	58.3 (5.8)	43.8 (2.5)	68.1 (0.9)	53.5 (5.4)	30.6 (0.6)	27.6 (0.7)	69.5 (1.6)	38.0 (2.5)	72.1 (0.5)	51.2 (2.7)	**20.6 (0.6)**	16.0 (0.6)
MVDFC	**65.6 (4.9)**	47.2 (2.5)	**87.7 (0.4)**	**70.2 (4.8)**	**31.5 (0.8)**	**27.8 (0.7)**	**70.7 (2.5)**	54.4 (1.2)	75.4 (0.5)	**70.2 (4.8)**	14.5 (1.0)	**27.6 (0.7)**

Method	Purity (%)						ARI (%)					
	ALOI	Caltech101	MNIST	MSRC-v1	NUS-WIDE	Youtube	ALOI	Caltech101	MNIST	MSRC-v1	NUS-WIDE	Youtube
K-means	49.9 (2.3)	59.1 (1.4)	57.3 (3.0)	48.8 (3.2)	17.7 (1.0)	27.1 (0.9)	33.7 (2.4)	20.1 (2.0)	36.4 (3.0)	26.5 (2.8)	2.27 (0.2)	7.99 (0.8)
MVCC	**73.9 (4.1)**	73.6 (2.3)	77.6 (4.0)	63.3 (5.1)	-	25.4 (0.3)	43.6 (2.8)	32.8 (3.6)	63.9 (3.1)	41.7 (5.9)	-	5.55 (0.2)
AMGL	54.9 (3.3)	69.0 (0.6)	86.3 (8.8)	69.2 (6.9)	10.1 (0.1)	28.8 (0.0)	33.0 (3.7)	27.9 (1.9)	75.1 (11.4)	48.3 (9.6)	0.50 (0.1)	7.20 (0.5)
MLAN	61.7 (5.7)	66.6 (0.0)	67.6 (0.2)	73.3 (0.0)	14.6 (0.0)	17.7 (0.3)	35.9 (6.5)	19.9 (1.0)	53.6 (0.7)	50.4 (0.0)	0.19 (0.0)	2.22 (0.2)
MVKSC	64.4 (0.0)	61.4 (0.0)	22.1 (0.0)	70.5 (0.0)	17.1 (2.6)	28.9 (0.0)	43.8 (0.0)	15.1 (0.0)	16.6 (0.0)	45.8 (0.0)	0.61 (0.1)	7.01 (0.0)
MSC-IAS	66.8 (3.3)	57.9 (2.1)	73.8 (1.0)	59.1 (3.4)	**36.2 (1.3)**	31.0 (0.4)	52.8 (5.1)	19.3 (2.4)	60.3 (1.2)	33.5 (3.7)	10.6 (0.4)	9.53 (0.5)
MVDFC	71.5 (3.0)	**77.1 (0.9)**	**87.7 (0.4)**	**74.3 (4.9)**	32.3 (0.7)	**31.2 (1.0)**	**52.9 (3.2)**	**39.2 (2.9)**	**76.3 (0.8)**	**54.6 (4.9)**	**11.6 (0.6)**	**11.1 (0.4)**

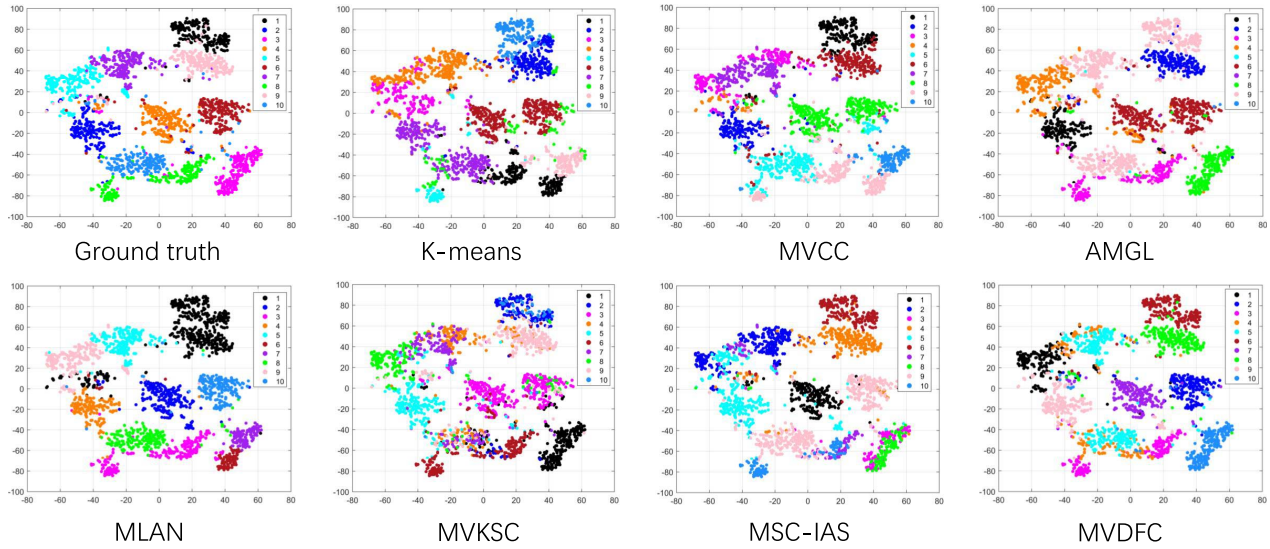


Fig. 4. Visualizations of different multi-view clustering methods on the selected sample dataset.

gained from multiple data sources can provide a beneficial potential for a more accurate unsupervised learning.

Fig. 4 also illustrates the learning performances for different multi-view clustering approaches on HW which consists of handwritten numerals from 0 to 9 digit classes, with total 2,000 patterns from 6 heterogeneous views, labeled with 10 classes. We stack the feature vectors of 6 views together. Then the high-dimensional feature is mapped onto a 2D subspace by a well-known dimension reduction method t-SNE [48]. Finally, we color the mapped 2D data with the cluster results obtained by different methods to visualize the performances of different multi-view clustering methods. Evidently, a better performance should be closer to the ground truths. In this sense, the visualized results shown in Fig. 4 reflect that the proposed MVDFC can effectively group similar samples into the same clusters.

E. Model Discussions

1) Runtime Analyses: The run-time of each compared multi-view clustering method is reported in Table IV. Note that we record the averaged time consumption of each algorithm by repeating 20 times with random initializations. From the table, K-means is the fastest, which is consequently adopted as an unsupervised initialization technique in many cases. On the contrary, AMGL consumes the longest time on almost all datasets. MVDFC performs reasonably well in terms of run-time.

2) Convergence Analyses: To further validate the convergence property of our method, we respectively compute the objective function values on ALOI, Caltech101, MNIST, MSRC-v1, NUS-WIDE and Youtube datasets. The corresponding curves are presented in Fig. 5. From the figure, we can see the objective function values decrease with the iteration number and converge stably and quickly. It suggests that the proposed algorithm has fast and stable convergence behavior.

Fig. 6 shows the evolution of view weights on different test datasets from the 1st iteration. At the beginning, we treat each view equally and set $\alpha_i = \frac{1}{k}$ as initialization, where α_i is the i-th view weight and k is the number of views.

TABLE IV
TIME CONSUMPTION OF THE PROPOSED METHOD AND SIX COMPARISON ALGORITHMS (IN SECONDS)

Datasets	K-means	MVCC	AMGL	MLAN	MVKSC	MSC-IAS	MVDFC
ALOI	0.15	15.23	25.17	4.94	4.34	17.22	0.44
Caltech101	3.69	199.04	92.26	65.99	112.63	94.66	25.24
MNIST	0.56	886.91	5,942.90	2,802.40	2,433.60	1,369.4	8.01
MSRC-v1	0.05	6.80	0.23	0.27	0.45	5.69	0.54
NUS-WIDE	3.69	-	2,000.00	1,029.37	29.74	35.14	12.19
Youtube	2.50	285.55	43.70	44.71	54.59	65.35	19.74

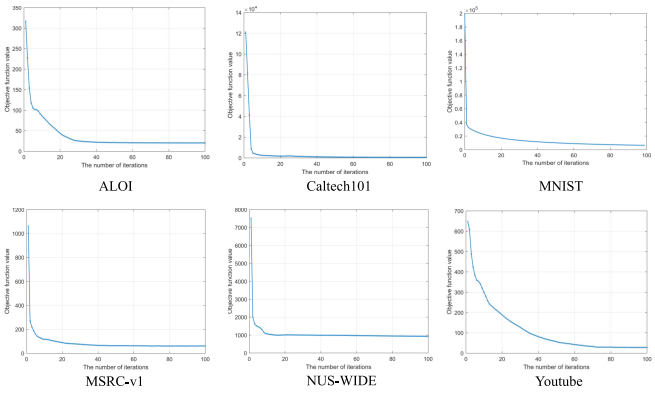


Fig. 5. Convergence curves of MVDFC on all datasets.

From Fig. 6, we can clearly find that the weight curves rapidly converge after a limited number of iterations. It can also be seen from Fig. 6 that there are dominating views on ALOI, Caltech101 and MSRC-v1 datasets after learning the weights for individual views. To investigate its impact, we conduct a comparison between MVDFC and the method utilizing the features of individual views only. From the results in Fig. 7, integrating different views together can significantly improve the performance compared with single-view clustering running on any specific view. It further validates the effectiveness of MVDFC, proving that MVDFC can adequately fuse different views and assign appropriate weights to explore the hidden complementary information among multi-view data.

3) Parameter Sensitivity: In order to verify the robustness of the proposed method with respect to different settings, parameter sensitivity analyses are conducted, as shown in Figs. 8, 9 and 10. We employ a grid searching strategy to find the best choices for all parameters on each test dataset, and then select an appropriate uniform value for each parameter to implement multi-view clustering in all benchmark datasets.

While keeping $\beta_i = 1$ and $\gamma = 2$, the clustering performance of the proposed method is reported in Fig. 8, where the weighting coefficient λ ranges in $\{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$. This figure shows that the best clustering results of MVDFC are attained at $\lambda \in [0.1, 10]$ with high probability. Specifically, MVDFC achieves the best performances on ALOI and Caltech101 with $\lambda = 1$, and with $\lambda = 10$ on MNIST, MSRC-v1, and Youtube. Notice that λ is positively correlated to the clustering performance on NUS-WIDE, and $\lambda = 1 \times 10^4$ leads to the best result.

With fixed $\lambda = 10$ and $\gamma = 2$, the experimental results of the proposed method are presented in Fig. 9, in which the regularization parameter β_i varies in $\{10^{-4}, 10^{-3}, \cdots, 10^{4}\}$. It shows that $\beta_i = 1$ brings out the best performance of the proposed method on ALOI, Caltech101, MNIST, and Youtube. Interestingly, this best parameter setting changes to $\lambda = 0.1$ on MSRC-v1 and $\lambda = 10$ on NUS-WIDE. Overall, the proposed method performs well when β_i ranges in $[0.1, 10]$.

Setting $\lambda = 10$ and $\beta_i = 1$, the empirical studies of the proposed method are demonstrated in Fig. 10. Herein, the nonlinear factor γ ranges in $\{1, 2, \cdots, 9\}$. The figure shows that we can mostly reach the best clustering performances when $\gamma = 2$ or $\gamma = 3$ on ALOI, Caltech101, MSRC-v1, NUS-WIDE and Youtube, then the performances sharply decrease when $\gamma \in [5, 9]$. It may be attributed to the following reason. As we know, $\alpha_i \in [0, 1]$ and the first term $\alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{H}\mathbf{W}_i||_F^2$ in our model aims to learn a shared clustering indicator matrix with a minimal fitting error. When γ is large, the coefficient α_i^{γ} is small, which indicates that a larger fitting error is allowed. Therefore, it is reasonable that the clustering performance curves dramatically rise to a maximum value then start to drop afterwards.

Overall, the above experiments indicate that the performance gains from the sufficiency of information but would be hurt by the redundancy.

V. EXTENSION TO SUPERVISED AND SEMI-SUPERVISED MULTI-VIEW LEARNING

In this section, we show the extensions of the proposed unsupervised multi-view clustering framework to semi-supervised and supervised multi-view classification.

A. Semi-Supervised Multi-View Classification

Traditional clustering methods refer to unsupervised ones. However, there are often a small amount of labeled training samples available in many computer vision and machine learning tasks. Semi-supervised learning can leverage the limited supervised information to guide the learning. In semi-supervised multi-view scenarios, recent works such as hypergraph-based [49] and parameter-free [50] semi-supervised multi-view classification have been developed. In fact, the framework represented in Equation (3) can be extended to tackle semi-supervised classification problems.

Denote the given training data as $\{\{\mathbf{X}_i\}_{i=1}^k, \mathbf{Y}\}$ with the i-th view data matrix $\mathbf{X}_i \in \mathbb{R}^{n \times d_i}$ and class label $\mathbf{Y} \in \mathbb{R}^l$ where $l \ll n$ implies that the number of labeled samples is much smaller than that of total samples. Certainly, we aim to make full use of the labeled samples to improve the learning performance. Naturally, the predicted classification results for such labeled samples should be as close as possible

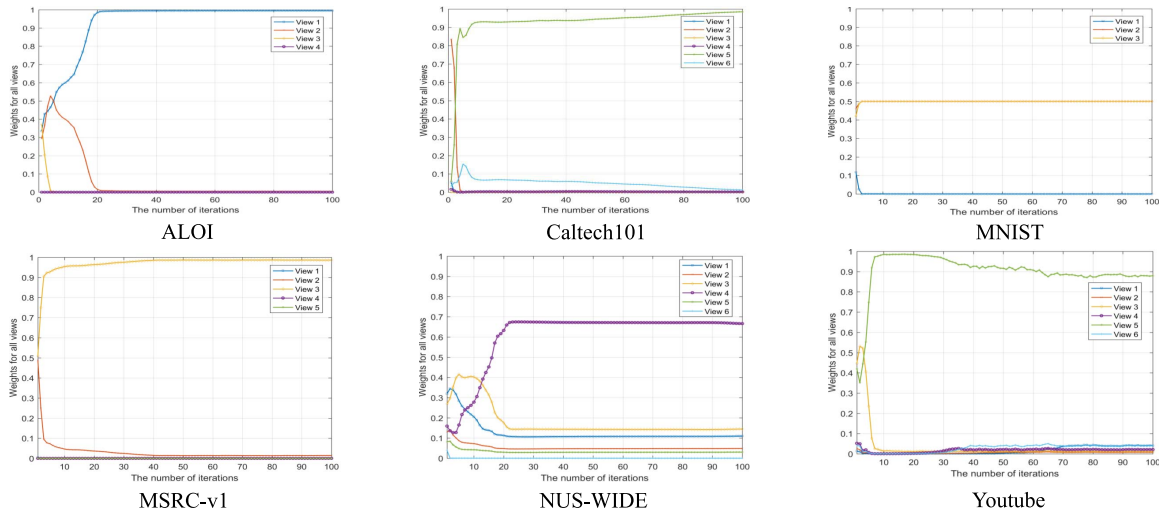


Fig. 6. The evolution of view weights on different test datasets.

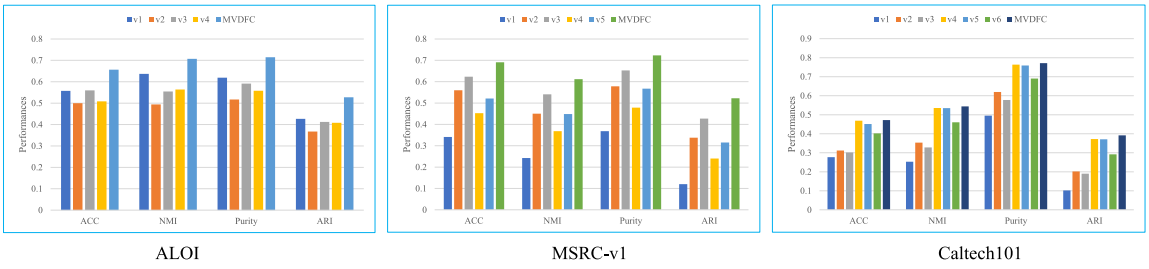


Fig. 7. Comparison between MVDFC and such method utilizing only one specific views feature.

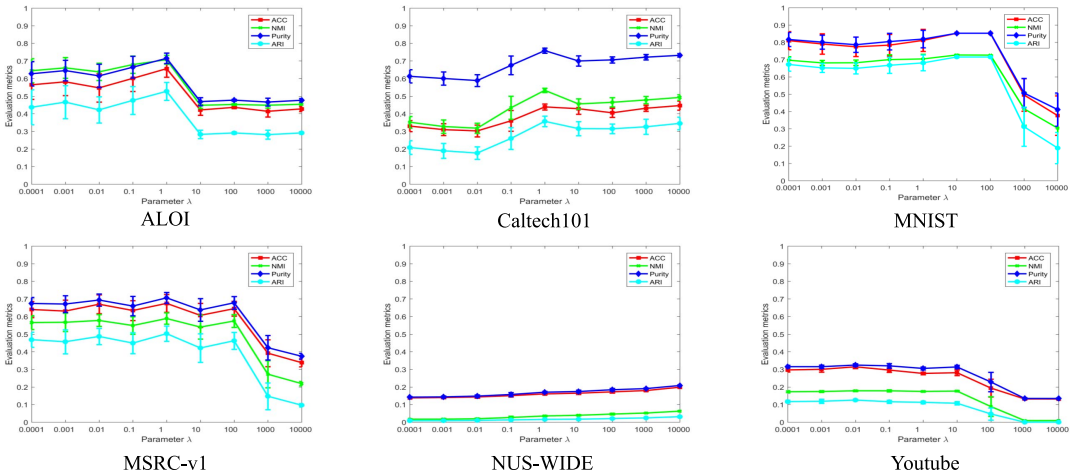


Fig. 8. Performances of MVDFC with various weighting coefficient λ ranging in $\{10^{-4}, 10^3, \cdots, 10^4\}$ while fixing $\beta_i = 1$ and $\gamma = 2$.

to the given labels. Towards this end, the small amount of labeled samples serve as prior knowledge, which can then be embedded into the formulated classification optimization problem, defined by

$$\min_{\alpha,\mathbf{H},\mathbf{W}} \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{S}\mathbf{H}\mathbf{W}_i||_F^2 + \beta_i ||\mathbf{W}_i||_* \quad \text{subject to } \alpha \geq \mathbf{0}, \sum_{i=1}^{k} \alpha_i = 1, \mathbf{H} \geq \mathbf{0}, \mathbf{H}^T\mathbf{H} = \mathbf{I} \tag{38}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ is a predefined matrix to preserve the prior knowledge from data distribution. Without loss of generality, it is assumed that the first l data points are labeled and the number of classes is c. The embedded matrix $\mathbf{S} = [\mathbf{S}_{ij}]_{n\times n}$ is then defined as the following block matrix

$$\mathbf{S} = \begin{bmatrix} \mathbf{L}_{l\times c} & \mathbf{O}_{l\times(n-c)} \\ \mathbf{O}_{(n-l)\times c} & \mathbf{I}_{(n-l)\times(n-c)} \end{bmatrix}. \tag{39}$$

Herein, $\mathbf{O}_{l\times(n-c)}$ and $\mathbf{O}_{(n-l)\times c}$ are respectively $l \times (n - c)$ and $(n - l) \times c$ zero matrices, $\mathbf{I}_{(n-l)\times(n-c)}$ is a

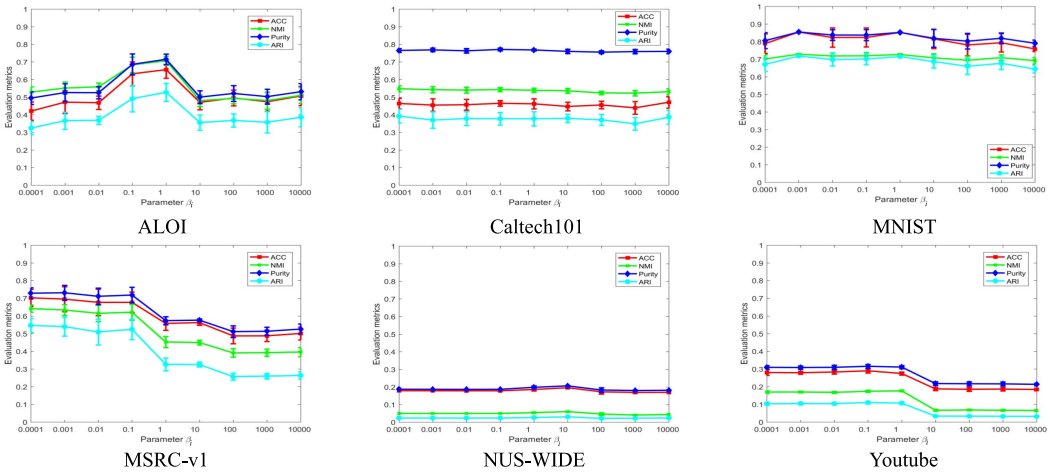


Fig. 9. Performance of MVDFC with various regularization parameter β_i ranging in $\{10^{-4}, 10^{3}, \cdots, 10^{4}\}$ while fixing $\lambda = 10$ and $\gamma = 2$.

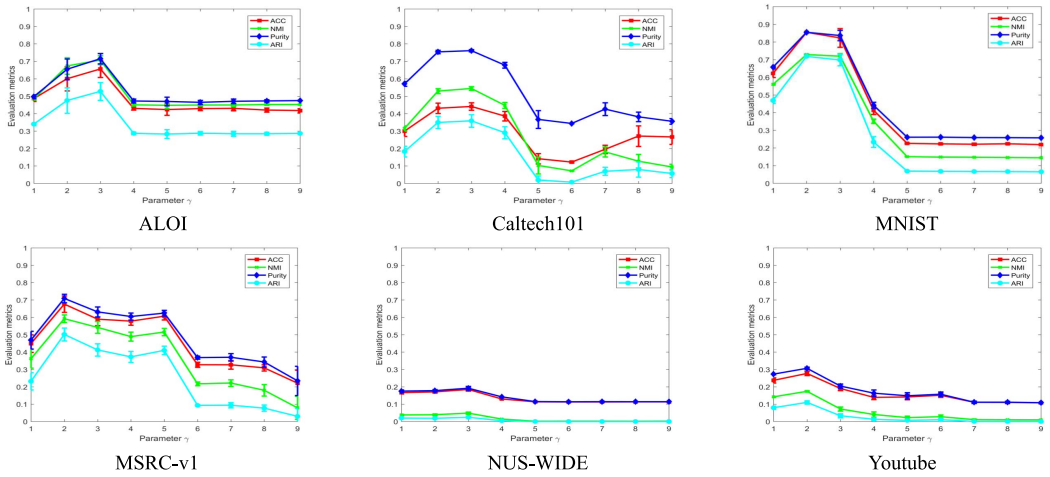


Fig. 10. Performances of MVDFC with various nonlinear factor γ ranging in $\{1, 2, \cdots, 9\}$ while fixing $\lambda = 10$ and $\beta_i = 1$.

$(n-l) \times (n-c)$ identity matrix, and $\mathbf{L}_{l\times c} \in \{0, 1\}$ contains the labeled information, where the (i, j)-th entry of $\mathbf{L}_{l\times c}$ is equal to one if and only if the i-th data point belongs to the j-th class. Consequently, the embedded matrix ensures that the optimization problem above tends to learn a consistent clustering indicator $\mathbf{H}$ with the given $\mathbf{S}$. Notice that the above semi-supervised multi-view data fusion oriented classification (semi-MVDFC) optimization problem can be solved by MVDFC framework.

Fig. 11 presents an intuitive illustration of semi-MVDFC on the HW dataset. Evidently, a better performance should be closer to the ground truths. From the figure, semi-MVDFC can well boost the utilization of label information to obtain more accurate class labels compared with MVDFC, which indicates the good scalability of MVDFC.

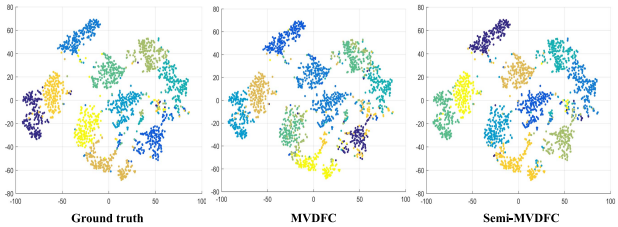


Fig. 11. Visualization of semi-MVDFC on HW with t-SNE.

B. Supervised Multi-View Classification

Benefiting from the label information, supervised multi-view classification has attracted increasing attention in recent years. For example, [51] proposed a new discriminative regression to address the multi-view feature learning problem, which was further enhanced to be more discriminative for subsequent classification. Huang *et al.* [52] combined multi-view canonical correlation analysis (MCCA) and multi-view spectral embedding (MSE) for supervised PoISAR image classification. We also can extend MVDFC to address supervised multi-view classification problems.

Under a supervised scenario, we denote the given multi-view training data as $\{\{\mathbf{X}_i\}_{i=1}^{k}, \mathbf{Y}\}$ with the i-th view data matrix $\mathbf{X}_i \in \mathbb{R}^{n\times d_i}$ and class label $\mathbf{Y} \in \mathbb{R}^{n}$. In this sense, the class label information is available for guiding the searching directions of learning algorithms.

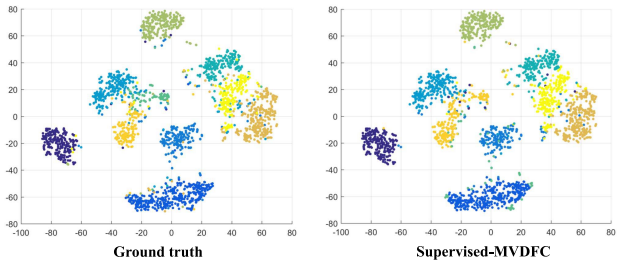


Fig. 12. Visualization of supervised-MVDFC on MNIST with t-SNE.

By contrast, $\mathbf{H} = [\mathbf{H}_{ij}]_{n\times c}$ denotes a given matrix from the training data, representing an indicator for the class label assignment, defined by

$$\mathbf{H}_{ij} = \begin{cases} 1, & \text{the sample } \mathbf{x}_i \text{ belongs to the } j\text{-th class;} \\ 0, & \text{otherwise.} \end{cases} \tag{40}$$

As a consequence, Equation (3) for the supervised multi-view classification problem can be rewritten as

$$\min_{\alpha,\mathbf{W}} \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i - \mathbf{H}\mathbf{W}_i||_F^2 + \beta_i ||\mathbf{W}_i||_* \quad \text{subject to } \alpha \geq \mathbf{0}, \sum_{i=1}^{k} \alpha_i = 1 \tag{41}$$

where α is a weighted vector for all views, and $\mathbf{W} = [\mathbf{W}_1, \cdots, \mathbf{W}_k] \in \mathbb{R}^{c\times \sum_{j=1}^{k} d_j}$ is a coefficient matrix for the multi-view data.

To obtain an explicit solution to the test data, the optimization problem above is rewritten as

$$\min_{\alpha,\mathbf{W}} \sum_{i=1}^{k} \alpha_i^{\gamma} ||\mathbf{X}_i\mathbf{W}_i^T - \mathbf{H}||_F^2 + \beta_i ||\mathbf{W}_i||_* \quad \text{subject to } \alpha \geq \mathbf{0}, \sum_{i=1}^{k} \alpha_i = 1 \tag{42}$$

where α and $\mathbf{W}$ are parameters to be learned from the training data. The optimization problem above can also be solved by MVDFC framework. We denote the learned optimal α and $\mathbf{W}$ as $\widehat{\alpha}$ and $\widehat{\mathbf{W}}$, respectively. With given multi-view test data $\{\mathbf{X}_i^{test}\}_{i=1}^{k}$ with $\mathbf{X}_i \in \mathbb{R}^{N\times d_i}$ and N being the number of test samples, the classification task is to provide the predicted class label $\mathbf{Y}^{test} \in \{0,1\}^N$ of $\{\mathbf{X}_i^{test}\}_{i=1}^{k}$.

As mentioned before, the class labels for all test data can be predicted by solving the following optimization problem

$$\min_{\mathbf{H}} J(\mathbf{H}) = \sum_{i=1}^{k} \widehat{\alpha}_i^{\gamma} ||\mathbf{X}_i^{test}\widehat{\mathbf{W}}_i^T - \mathbf{H}||_F^2 + \beta_i ||\mathbf{W}_i||_*. \tag{43}$$

Since $J(\mathbf{H})$ is an unconstrained objective function, the optimal solution of (43) is attained at $\partial J(\mathbf{H})/\partial \mathbf{H} = \mathbf{0}$, that is

$$\frac{\partial J(\mathbf{H})}{\partial \mathbf{H}} = \sum_{i=1}^{k} \widehat{\alpha}_i^{\gamma} (\mathbf{H} - \mathbf{X}_i^{test}\widehat{\mathbf{W}}_i^T) = \mathbf{0}. \tag{44}$$

As a result, the optimal solution $\widehat{\mathbf{H}}$ becomes

$$\widehat{\mathbf{H}} = \frac{\sum_{i=1}^{k} \widehat{\alpha}_i^{\gamma} \mathbf{X}_i^{test}\widehat{\mathbf{W}}_i^T}{\sum_{i=1}^{k} \widehat{\alpha}_i^{\gamma}}, \tag{45}$$

which suggests the class label assignments $\mathbf{Y}^{test}$. Fig. 12 illustrates the performances of the extended supervised multi-view data fusion oriented classification (supervised-MVDFC) method on MNIST. Herein, we randomly produce 2,000 samples from MNIST for better visualization. A better performance should be closer to the ground truths. The visualized results shown in Fig. 12 reflect the fine classification ability of the supervised-MVDFC, which further indicates MVDFC can be effectively generalized to supervised multi-view learning.

VI. Conclusion

In this paper, we proposed an efficient multi-view clustering method by nuclear norm minimization, which achieves compression of principal components for the learned feature representation. In the proposed formulation, multi-view data fusion and clustering are integrated into a unified framework. Besides, we further proposed an alternating optimization algorithm based on dual ADMM framework with a closed-form solution to each subproblem. Comprehensive experimental results validate the effectiveness and efficiency of the proposed method. Furthermore, we have also shown the extension of the proposed framework to semi-supervised and supervised multi-view learning tasks. Recently, deep learning models exhibit strong ability in dealing with various kinds of supervised learning tasks. In the further work, we will devote more efforts to connecting with neural networks to address multi-view semi/fully supervised classification problems.

References

[1] J. Li, B. Zhang, G. Lu, and D. Zhang, "Generative multi-view and multi-feature learning for classification," *Inf. Fusion*, vol. 45, pp. 215–226, Jan. 2019.
[2] Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1774–1782, Jul. 2019.
[3] J. Wu, Z. Lin, and H. Zha, "Essential tensor learning for multi-view spectral clustering," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5910–5922, Dec. 2019.
[4] F. Nie, G. Cai, J. Li, and X. Li, "Auto-weighted multi-view learning for image clustering and semi-supervised classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1501–1511, Mar. 2018.
[5] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. T. Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Inf. Sci.*, vol. 480, pp. 287–304, Apr. 2019.
[6] Z. Tao, H. Liu, H. Fu, and Y. Fu, "Multi-view saliency-guided clustering for image cosegmentation," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4634–4645, Sep. 2019.
[7] G. Li, K. Chang, and S. C. H. Hoi, "Multiview semi-supervised learning with consensus," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 11, pp. 2040–2051, Nov. 2012.
[8] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1180–1197, May 2017.
[9] 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. (ICML)*, 2009, pp. 129–136.
[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] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proc. 17th Int. Conf. Artif. Intell. Statist.*, 2014, pp. 823–831.

[12] S. Li, Y. Jiang, and Z. Zhou, "Partial multi-view clustering," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1968–1974.

[13] Q. Yin, S. Wu, and L. Wang, "Incomplete multi-view clustering via subspace learning," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2015, pp. 383–392.

[14] J. Wen, Z. Zhang, Y. Xu, and Z. Zhong, "Incomplete multi-view clustering via graph regularized matrix factorization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 593–608.

[15] F. Nie, S. Shi, and X. Li, "Auto-weighted multi-view co-clustering via fast matrix factorization," *Pattern Recognit.*, vol. 102, pp. 107–207, 2020.

[16] M. Yin, W. Huang, and J. Gao, "Shared generative latent representation learning for multi-view clustering," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 6688–6695.

[17] M. Chen, L. Huang, C. Wang, and D. Huang, "Multi-view clustering in latent embedding space," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 3513–3520.

[18] 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. (CVPR)*, Jun. 2015, pp. 586–594.

[19] L. Zong, X. Zhang, L. Zhao, H. Yu, and Q. Zhao, "Multi-view clustering via multi-manifold regularized non-negative matrix factorization," *Neural Netw.*, vol. 88, pp. 74–89, Apr. 2017.

[20] Z. Xue, J. Du, D. Du, and S. Lyu, "Deep low-rank subspace ensemble for multi-view clustering," *Inf. Sci.*, vol. 482, pp. 210–227, May 2019.

[21] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3939–3949, Nov. 2015.

[22] M. Cheng, L. Jing, and M. K. Ng, "Tensor-based low-dimensional representation learning for multi-view clustering," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2399–2414, May 2019.

[23] M. Yin, J. Gao, S. Xie, and Y. Guo, "Multiview subspace clustering via tensorial t-Product representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 851–864, Mar. 2019.

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

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

[26] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2149–2155.

[27] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan, "Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2153–2159.

[28] Y. Wang and L. Wu, "Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering," *Neural Netw.*, vol. 103, pp. 1–8, Jul. 2018.

[29] X. Zhu, S. Zhang, W. He, R. Hu, C. Lei, and P. Zhu, "One-step multi-view spectral clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 2022–2034, Oct. 2019.

[30] L. Houthuys, R. Langone, and J. A. K. Suykens, "Multi-view kernel spectral clustering," *Inf. Fusion*, vol. 44, pp. 46–56, Nov. 2018.

[31] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "Comic: Multi-view clustering without parameter selection," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 5092–5101.

[32] Q. Gao, W. Xia, Z. Wan, D.-Y. Xie, and P. Zhang, "Tensor-SVD based graph learning for multi-view subspace clustering," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 3930–3937.

[33] H. Liu and Y. Fu, "Consensus guided multi-view clustering," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 4, pp. 1–21, Jul. 2018.

[34] X. Liu *et al.*, "Late fusion incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2410–2423, Oct. 2019.

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

[36] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1881–1887.

[37] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.

[38] Z. Xue, G. Li, S. Wang, J. Huang, W. Zhang, and Q. Huang, "Beyond global fusion: A group-aware fusion approach for multi-view image clustering," *Inf. Sci.*, vol. 493, pp. 176–191, Aug. 2019.

[39] F. Nie, J. Li, and X. Li, "Self-weighted multiview clustering with multiple graphs," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2564–2570.

[40] F. Nie, L. Tian, and X. Li, "Multiview clustering via adaptively weighted procrustes," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2022–2030.

[41] S. Huang, Z. Kang, I. W. Tsang, and Z. Xu, "Auto-weighted multi-view clustering via kernelized graph learning," *Pattern Recognit.*, vol. 88, pp. 174–184, Apr. 2019.

[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, Jan. 2010.

[43] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[44] H. Wang, Y. Yang, and T. Li, "Multi-view clustering via concept factorization with local manifold regularization," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 1245–1250.

[45] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2408–2414.

[46] X. Wang, Z. Lei, X. Guo, C. Zhang, H. Shi, and S. Z. Li, "Multi-view subspace clustering with intactness-aware similarity," *Pattern Recognit.*, vol. 88, pp. 50–63, Apr. 2019.

[47] L. Lovasz and M. D. Plummer, *Matching Theory*. Providence, RI, USA: American Mathematical Society, 2009.

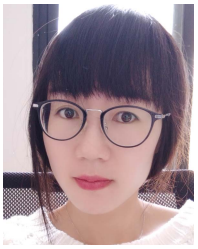
[48] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[49] J. J. Whang *et al.*, "MEGA: Multi-view semi-supervised clustering of hypergraphs," *Proc. VLDB Endowment*, vol. 13, no. 5, pp. 698–711, Jan. 2020.

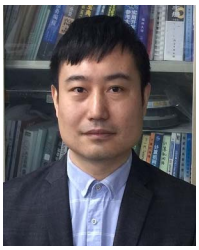
[50] R. Zhang, F. Nie, and X. Li, "Semisupervised learning with parameter-free similarity of label and side information," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 405–414, Feb. 2019.

[51] M. Yang, C. Deng, and F. Nie, "Adaptive-weighting discriminative regression for multi-view classification," *Pattern Recognit.*, vol. 88, pp. 236–245, Apr. 2019.

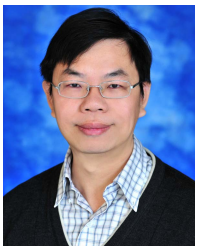
[52] X. Huang, X. Nie, H. Qiao, and B. Zhang, "Supervised polsar image classification by combining multiple features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 22–25.



Aiping Huang (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from Putian University, Putian, China, in 2011, and the M.S. degree in basic mathematics from Minnan Normal University, Zhangzhou, China, in 2014. She is currently pursuing the Ph.D. degree with the Collage of Physics and Information Engineering, Fuzhou University. She served as a Lecturer for the School of Information Science and Technology, Xiamen University Tan Kah Kee College, from 2014 to 2018. Her research interests include computer vision, image processing, machine learning, and data mining.



Tiesong Zhao (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2006, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2011. He served as a Research Associate for the Department of Computer Science, City University of Hong Kong, from 2011 to 2012, a Postdoctoral Fellow for the Department of Electrical and Computer Engineering, University of Waterloo, from 2012 to 2013, and a Research Scientist for the Ubiquitous Multimedia Laboratory, The State University of New York at Buffalo, from 2014 to 2015. He is currently a Minjiang Distinguished Professor with the College of Physics and Information Engineering, Fuzhou University, China. His research interests include multimedia signal processing, coding, quality assessment and transmission. Due to his contributions in video coding and transmission, he received the Fujian Science and Technology Award for Young Scholars in 2017. He has also been serving as an Associate Editor for *Electronics Letters* (IET) since 2019.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is also the Deputy Director of the AI Research Center of NTHU. He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, from 2000 to 2007. Prior to joining academia, he worked with the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, from 1992 to 2000. His research interests include image and video processing, computer vision, and video networking. He served as a Distinguished Lecturer for the IEEE Circuits and Systems Society from 2018 to 2019, a Steering Committee Member for the IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2015, and the Chair for the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. His articles received the Best Paper Award of the IEEE VCIP 2015, the Top 10% Paper Awards of the IEEE MMSP 2013, and the Young Investigator Award of VCIP 2005. He received the Outstanding Electrical Professor Award presented by Chinese Institute of Electrical Engineering in 2019, and the Young Investigator Award presented by the Ministry of Science and Technology, Taiwan, in 2006. He is currently the Chair of the Steering Committee of the IEEE ICME. He has been serving as the President of the Chinese Image Processing and Pattern Recognition Association, Taiwan, since 2019. He has served as the Technical Program Co-Chair for the IEEE ICME 2010, and the General Co-Chair for the IEEE VCIP 2018, and the Technical Program Co-Chair for the IEEE ICIP 2019. He has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, *IEEE MultiMedia*, and *Journal of Visual Communication and Image Representation*.