Bi-Level Multi-View Latent Space Learning

Zhe Xue, Guorong Li, Shuhui Wang, Weigang Zhang, and Qingming Huang, *Senior Member*, *IEEE*

***Abstract*—Different kinds of features describe different aspects of image data, and each feature can be treated as a view when we take it as a particular understanding of images. Leveraging multiple views provides a richer and comprehensive description than only using single view. However, multi-view data are often represented by high-dimensional heterogeneous features, so it is meaningful to find a low-dimensional consensus representation from multiple views. In this paper, we propose an unsupervised multi-view dimensionality reduction method for images based on bi-level latent space learning. As different views have different physical meanings and statistical properties, they are not directly comparable. Therefore, we learn the comparable representation for each view in the first level. The shared and private nature of multi-view data are exploited to accurately preserve the information of each view. Then, we fuse different views into a low-dimensional representation by conducting joint matrix factorization in the second level. To guarantee the low-dimensional representation to be compact and discriminative, the intrinsic geometric structure of data are utilized. Besides, our method considers to resist the outliers and noise contained in multi-view data which may influence the learned representation and deteriorate its semantic consistency. We design appropriate optimization objectives to learn the latent spaces in different levels. Compared with the existing methods, our method could provide a more flexible multi-view learning strategy that not only accurately captures the information of each view but also is robust to outliers and noise, which can obtain a more discriminative and compact low-dimensional representation. Experiments on two real-world image datasets demonstrate the advantages of our method over the existing multi-view dimensionality reduction methods.**

***Index Terms*—Multi-view, latent space, matrix factorization, image and video classification.**

I. Introduction

MANY real-world problems involve multi-view data where instances are characterized by multiple representations. For example, images can be represented by different visual features describing color, shape, texture and other visual information. A webpage has different contents such as text, image, video, etc. Different views generate different descriptions about the same instance and they can complement with each other. Compared with only using a single view, a more accurate and robust representation can be obtained by leveraging multiple views. However, directly using high-dimensional multi-view features will lead to the degraded performance of the model due to the "curse of dimensionality". A practical solution is to find a low-dimensional consensus representation which could effectively characterize the multi-view data and generate better learning performance.

To address the low-dimensional representation learning problem in multi-view learning, many methods have been presented [1], [2], [3], [4], [5], [6]. Some of them [1], [2], [3] use graphs to model multi-view data, and then fuse all the obtained graphs into a unified graph or learn a unified embedding by enforcing the embedding of each view to be consistent. A general spectral embedding framework is first proposed for multi-view dimensionality reduction [1]. Patches are introduced to represent the multi-view data in [2], and then the low-dimensional representation is learned by patch alignment. Besides, matrix factorization techniques are adopted in [4], [5], [6] to learn a new representation from multiple views. Structured sparse PCA [7] is adopted in [5] to encode the information of each view into the low-dimensional representation. The common representation in [6] is obtained by jointly factorizing the data matrix of each view into a new common subspace with group sparse regularization.

Nevertheless, most of the existing methods only partially solved the key problems in multi-view learning as follows. First, different views describe different aspects of the same instance, and some information is shared among all the views while some independent information only exists in certain view. Taking account for the shared and private nature can represent the information of each view more accurately and better exploit the complementary nature of multi-view data. Second, real data may be distributed in a low-dimensional manifold and such nonlinearity should not be ignored. By preserving the manifold structure of data in the low-dimensional representation, a more compact and discriminative representation can be obtained. Third, multi-view data often contain outliers and noise which may greatly influence the learning performance. So the proposed model should be capable of accommodating these unreliable information.

In light of all the above mentioned factors, we propose an unsupervised multi-view dimensionality reduction method for image data called bi-level multi-view latent space learning (BLMV), which aims to achieve better image classification and annotation performance in the learned low-dimensional subspace. Our bi-level learning strategy not only accurately captures the independent information of each view but also effectively resists the noise during the low-dimensional representation learning. In addition, the discriminating power of the learned representation can be improved by preserving the

Z. Xue, G. Li and Q. Huang are with the School of Computer and Control Engineering, University of Chinese Academy of Sciences (UCAS), and the Key Laboratory of Big Data Mining and Knowledge Management, UCAS, Beijing, 101408, China (e-mail: xuezhe10@mails.ucas.ac.cn; liguorong@ucas.ac.cn; qmhuang@ucas.ac.cn)

S. Wang and Q. Huang are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190 China (e-mail: wangshuhui@ict.ac.cn).

W. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai, 264209, China, and is also with the University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing, 100049, China (e-mail: wgzhang@hit.edu.cn).

Corresponding author: Guorong Li, Qingming Huang.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology

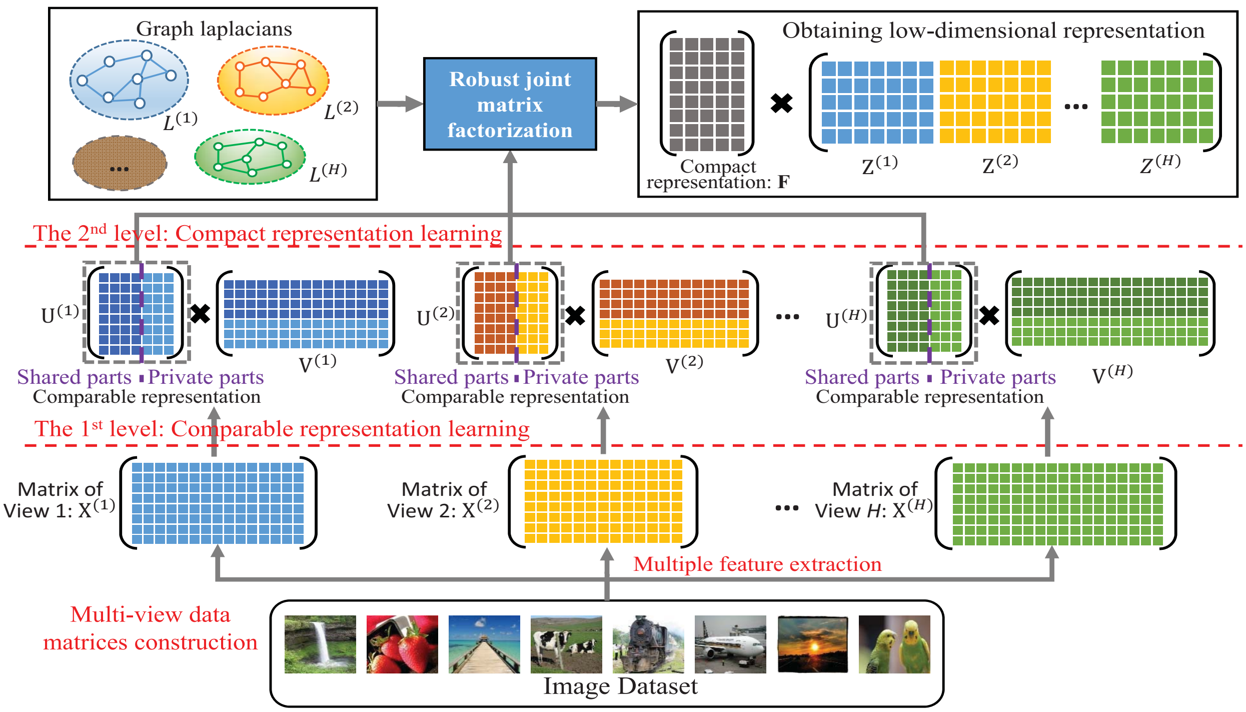


Fig. 1. The framework of BLMV. We extract H types of visual features from data to construct the multi-view matrix $\{\mathbf{X}^{(i)}\}_{i=1}^{H}$ which are represented by different colors of rectangles. In the first level, the comparable representation $\{\mathbf{U}^{(i)}\}_{i=1}^{H}$ are obtained, both shared and private parts are explored during learning. The second level learns the compact representation $\mathbf{F}$ from $\{\mathbf{U}^{(i)}\}_{i=1}^{H}$ by joint matrix factorization and the intrinsic geometric structure $\{L^{(i)}\}_{i=1}^{H}$ of multi-view data are preserved in $\mathbf{F}$.

local geometric structure of multi-view data. The framework of BLMV is shown in Fig. 1. Since different views with different physical meanings are not directly comparable, the objective of the first level is to learn *comparable representation* for each view (matrix $\mathbf{U}^{(i)}$ in Fig. 1). This representation is obtained by non-negative matrix factorization (NMF) [8] which provides naturally interpretable latent spaces. Considering the shared and private information of views, the latent spaces are divided into shared and private parts. The shared parts are assumed to be consistent with each other to leverage the complementary nature of multi-view data, and the private parts are learned adaptively by utilizing group sparse regularization. Consequently, the comparable representation accurately encodes the information of each view. In the second level, we conduct dimensionality reduction for the comparable representation and denote it as *compact representation* (matrix $\mathbf{F}$ in Fig. 1). By jointly factorizing the comparable representation towards a unified representation, the information of each view can be effectively fused into the learned latent space. Meanwhile, the nonlinearity structure of multi-view data which can be modeled by nearest neighbor graphs is preserved into the learned compact representation. Taking account for the different description abilities of different views and the noise contained in multi-view data, we estimate the weight for each view and introduce the $\ell_{2,1}$-norm into the loss function, which endows our model with resistancy to the unreliable views and robustness to noise and outliers.

We incorporate the learning of *comparable representation* and *compact representation* into a joint optimization framework, which learns the optimal consensus low-dimensional representation from multiple views. To solve the proposed objective function, an efficient iterative algorithm is derived and the convergence can be rigorously guaranteed. We conduct several image classification and annotation experiments on two real-world image datasets to demonstrate the effectiveness of the learned low-dimensional representation. Experimental results show the promising performance of BLMV over the existing methods. The main contributions of this work can be summarized as follows:

- A flexible multi-view dimensionality reduction method based on bi-level learning strategy is proposed. By designing appropriate optimization objectives in different levels, the information of each view can be more effectively captured and encoded into the low-dimensional representation.
- Both shared and private latent factors of multiple views are explored during the comparable representation learning. Through imposing reasonable regularization on different types of latent factors, the proposed method better takes advantage of the complementary properties of multi-view data.
- The proposed method exploits the manifold structure of each view to reveal the nonlinearity inherent in data, which makes the learned low-dimensional representation more compact and discriminative. Furthermore, the robustness of the method is improved by estimating the different importance of views and adopting $\ell_{2,1}$-norm loss function for multi-view fusion.

The rest of the paper is organized as follows. We introduce the related work in Section II. Then we elaborate the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology

proposed model, optimization scheme and the convergence proof in Section III and Section IV. Section V presents the experimental results on image classification and annotation. Finally, we conclude this paper in Section VI.

II. Related Work

Multi-view learning deals with data described in multiple views and exploits the multiple descriptions to improve the learning performance. Many multi-view works have been proposed for supervised learning [9], [10], [11], [12] and unsupervised learning [1], [2], [13], [14], [15]. This paper focuses on the latter. Most of the existing multi-view methods adopt two principles for learning tasks [16]. The first one is to make different views consistent by minimizing the disagreement of each view. The second principle assumes that each view contains some knowledge that the others do not contain, thus different views can complement with each other. By adopting the two properties, multi-view learning methods exploit knowledge from multiple sources and achieve better performance than single-view based methods.

For the first principle, some works try to make the learning results of different views agree with each other. Canonical correlation analysis (CCA) extracts a common subspace from two views by maximizing the correlation among them. Chaudhuri *et al.* [17] adopt CCA to project multi-view data onto a common low-dimensional subspace for clustering. By constructing a graph for each view, some methods [18], [19], [20] adopt spectral clustering techniques to learn a consensus representation from multiple views. Cai *et al.* [18] minimize the distance of the low-dimensional embedding of each view to obtain the unified clustering results. Considering the different importance of views, Huang *et al.* [19] assign a weight to each view which reflects its importance. This strategy makes the learning results more accurate and robust. Patch is used in [2], [21] to model the local structure of a sample on a view. Then the unified representation is obtained by global coordinate alignment, which makes all the low-dimensional embedding of each view consistent with each other.

The second principle tries to exploit the complementary nature of multi-view data. Co-training method [22] is originally proposed in semi-supervised learning and then it is used for multi-view learning. It assumes samples labeled by predictors on one view is useful for training predictors on other views. The two predictors can exchange the complementary information during co-training process [3], [23]. By constructing a kernel for each view, multiple kernel learning (MKL) aims to combine multiple kernels together to obtain an optimal kernel. This kernel contains complementary information of different views and provides a more comprehensive measurement of similarity. Lin *et al.* [24] generalize the framework of MKL for dimensionality reduction, which provides convenience of using multiple image features.

Furthermore, some multi-view works [4], [5], [25], [26] consider the shared and independent parts of each view to more accurately preserve the multi-view information. Structured sparsity is used in [4] to separate the latent space into shared and independent parts for human pose estimation. The independent information of each view is first encoded into a matrix, and then structured sparse PCA [7] is used to learn the low-dimensional representation in [5]. A semi-supervised latent factor learning method is proposed in [25], where the shared parts are made consistent with each other to exploit the complementary information.

Many works have shown that image data are more likely to reside on a low-dimensional submanifold of the ambient space [27], [28], [29]. The traditional dimensionality reduction methods such as PCA and NMF ignore the possible nonlinearity inherent in data, while such nonlinearity can be preserved by manifold learning methods. So some researches such as embedding learning [28], [30], [31], feature extraction [29], [32], [33], [34] and clustering [35], [36], [37] are developed based on manifold learning. These works adopt the locally invariant property, *i.e.*, if two samples are close in the intrinsic geometry of data manifold, then they should have similar embeddings. Both single-view [38], [39] and multi-view methods [40] demonstrate that exploiting the intrinsic manifold structure of data can enhance the discriminating power of the learned latent space and further improve the learning performance. It should be noted that [38] may suffer from trivial solutions when directly adding the manifold regularization to the objective function [41]. Another constraints should be imposed to make the problem well-defined and obtain more reliable solutions [41], [42], [43], [44].

The real data may contain undesirable noises and outliers and the commonly adopted least square error function is vulnerable to the unreliable data. To improve the robustness of the model, $\ell_{2,1}$-norm is introduced in [45], [46], [47] to cope with the noise and outliers. Compared with adopting least square function, the errors generated by outliers cannot dominate the objective function because they are not squared with $\ell_{2,1}$-norm, so that more robust solutions can be obtained. To solve $\ell_{2,1}$-norm minimization problem, Nie *et al.* [45] first develop an efficient algorithm whose convergence can be guaranteed and then extent it to solve other general $\ell_{2,1}$-norm minimization problems [48].

III. Bi-level Multi-view Latent Space Learning

A. Preliminary

For an arbitrary matrix $\mathbf{A}$, the (i,j)-th entry, the i-th row and the j-th column are denoted by $\mathbf{A}_{ij}$, $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$ respectively. $Tr[\mathbf{A}]$ is the trace of $\mathbf{A}$. $\mathbf{A} \odot \mathbf{B}$ and $\mathbf{A} \oslash \mathbf{B}$ represent element-wise multiplication and division of matrices $\mathbf{A}$ and $\mathbf{B}$, respectively. $(\mathbf{A}, \mathbf{B})$ represents the horizontal concatenation of $\mathbf{A}$ and $\mathbf{B}$, and $(\mathbf{A}; \mathbf{B})$ is the vertical concatenation of them. $\mathbf{1}_M \in \mathbb{R}^{M \times 1}$ denotes a vector of ones. For $\mathbf{A} \in \mathbb{R}^{N \times M}$, the Frobenius norm is $||\mathbf{A}||_F$, and $\ell_{2,1}$-norm is defined as

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{M} \mathbf{A}_{ij}^2} \tag{1}$$

Non-negative Matrix Factorization (NMF) imposes non-negative constraints on the learned latent space and provides a more interpretable and meaningful representation. Given a non-negative matrix $\mathbf{X} = [x_1, x_2, ..., x_n]^T \in \mathbb{R}^{N \times M}$, each row of $\mathbf{X}$ represents a sample. NMF aims to factorize $\mathbf{X}$ into two non-negative matrices,

$$\mathbf{X} \approx \mathbf{UV} \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{N\times K}$ and $\mathbf{V} \in \mathbb{R}^{K\times M}$. We usually have $K \ll N$ and $K \ll M$ in practice. $\mathbf{V}$ can be regarded as the learned basis representing the original data $\mathbf{X}$. Each row of $\mathbf{U}$ is the low-dimensional representation of each sample with respect to basis $\mathbf{V}$.

Given multi-view data set consisting of N image samples with H views, they can be denoted by a set of matrices $\mathcal{X} = \{\mathbf{X}^{(i)} \in \mathbb{R}^{N\times M_i}\}_{i=1}^{H}$, where M_i is the dimensionality of the i-th view. Our objective is to learn the compact representation of images $\mathbf{F} \in \mathbb{R}^{N\times R}$ where $R < M_i, \forall i = 1, ..., H$.

B. Comparable Representation Learning

In the first level, we want to learn the comparable representation from heterogeneous image features. Since NMF can provide naturally interpretable and meaningful latent spaces, we adopted it here to make multiple views with different properties comparable in the learned non-negative latent spaces. The data matrix is factorized as $\mathbf{X}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}$, where $\mathbf{U}^{(i)} \in \mathbb{R}^{N\times K}$ is the learned comparable representation for the i-th view, $\mathbf{V}^{(i)} \in \mathbb{R}^{K\times M_i}$ is the learned basis matrix. Meanwhile, the comparable representation should be capable of accurately representing the information of each view, so the shared and private property of multi-view data is exploited. We assume that the latent factors of each view are composed of shared and private parts. Shared parts referred to the latent factors that are used for generating all views, while private parts are only used for generating certain view. Taking account of this, the basis matrix is constructed as $\mathbf{V}^{(i)} = \left(\mathbf{V}_S^{(i)}; \mathbf{V}_P^{(i)}\right)$, where $\mathbf{V}_S^{(i)} \in \mathbb{R}^{K_S\times M_i}$ represents the shared part of latent factors and $\mathbf{V}_P^{(i)} \in \mathbb{R}^{K_P\times M_i}$ is the private part, $K = K_S + K_P$. The coefficient matrix is also constructed as $\mathbf{U}^{(i)} = \left(\mathbf{U}_S^{(i)}, \mathbf{U}_P^{(i)}\right)$, where the shared and private parts are $\mathbf{U}_S^{(i)} \in \mathbb{R}^{N\times K_S}$ and $\mathbf{U}_P^{(i)} \in \mathbb{R}^{N\times K_P}$, respectively. To control the proportion of the latent factors of shared parts and private parts, we introduce parameter $\theta = K_S/K$, which reflects the importance of shared parts in comparable representation learning. Thus, we have $K_S = ROUND(\theta \cdot K)$ and $K_P = K - K_S$.

Next, how to accurately learn the shared and private parts according to the property of each view is the critical problem. For private parts learning, we adopt group sparsity regularization, which is an effective technique to discover meaningful latent factors in latent space learning [4], [6]. To learn the private latent factors of each view according to its property, we impose $\ell_{2,1}$-norm on private parts of latent factors $\mathbf{V}_P^{(i)}$ which encourages some rows to be zeroed out. The intuition behind this is that we expect the private part of each view to depend on a subset of the latent dimensions, i.e. $\widetilde{K} \leq K_P$, where $\widetilde{K}$ is the number of the learned latent factors in $\mathbf{V}_P^{(i)}$. By using this regularization, the private latent factors of each view can be learned adaptively. For shared parts learning, as the shared parts of different views refer to the same latent factors and they are directly comparable, we make the coefficients of shared parts consistent to leverage complementary information of multiple views. Although several regularization functions can be used to make them consistent [3], [49], considering the complexity of the proposed model, we adopt a simple and efficient regularization manner and define the objective function as follows

$$\begin{aligned}
&\mathcal{O}_{cp}(\mathbf{U}^{(i)}, \mathbf{V}^{(i)}) = \min \sum_{i=1}^{H} \Big[\big\|\mathbf{X}^{(i)} - \big(\mathbf{U}_S^{(i)}, \mathbf{U}_P^{(i)}\big)\big(\mathbf{V}_S^{(i)}; \mathbf{V}_P^{(i)}\big)\big\|_F^2 \\
&+ \eta \big\|\mathbf{V}_P^{(i)}\big\|_{2,1} \Big] + \lambda \sum_{i=1}^{H-1} \sum_{j=i+1}^{H} \big\|\mathbf{U}_S^{(i)} - \mathbf{U}_S^{(j)}\big\|_F^2 \\
&s.t.\ \mathbf{U}^{(i)} \geq 0,\ \mathbf{V}^{(i)} \geq 0,\ \forall i = 1, 2, ..., H,
\end{aligned} \tag{3}$$

where η and λ are two parameters to control the strength of group sparsity and consistent regularization.

C. Compact Representation Learning

The objective of the second level is to learn the compact representation $\mathbf{F} \in \mathbb{R}^{N\times R}$ from the comparable representation $\{\mathbf{U}^{(i)}\}_{i=1}^{H}$. We expect to encode the multi-view information into the compact representation and another matrix factorization process is conducted. However, there are three issues that should be noticed. First, the description ability of different views are different. Some views generate more reliable description than the others and it is preferable to selecting these views for representation. Second, for the learned comparable representation $\{\mathbf{U}^{(i)}\}_{i=1}^{H}$, some views may contain inaccurate description of multi-view data and noise, which makes the learned compact representation deviate from the true value. Third, the intrinsic manifold structure of multi-view data should be preserved in compact representation to improve its discriminative power.

A direct solution is concatenating $\{\mathbf{U}^{(i)}\}_{i=1}^{H}$ into a new matrix $\mathbf{U} = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(H)})$ which contains the information of all views. By factorizing $\mathbf{U}$, the multi-view information can be encoded into the compact representation. However, this strategy treats all views equally important during fusion. The contribution of different views on the final result $\mathbf{F}$ should be different. Therefore, a more reasonable method is first let $\|\mathbf{U} - \mathbf{FZ}\|_F = \sum_{i=1}^{H} \|\mathbf{U}^{(i)} - \mathbf{FZ}^{(i)}\|_F$, where $\mathbf{Z} = (\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, ..., \mathbf{Z}^{(H)}), \mathbf{Z}^{(i)} \in \mathbb{R}^{R\times K}$. Then we assign a weight to each view to represent its importance. The view that achieves smaller factorization residual is considered to be more reliable. In addition, we need a robust loss function to cope with the outliers and noise contained in each view. By encouraging the row-wise sparsity of the residual matrix, $\ell_{2,1}$-norm accommodates outliers and noise in a better way than standard one, and it has been adopted in several robust learning tasks [46], [47]. So we introduce $\ell_{2,1}$-norm penalty to improve the robustness and the compact representation can be learned by solving the following objective

$$\begin{aligned}
&\mathcal{O}_{pt}(\mathbf{U}^{(i)}, \mathbf{F}, \mathbf{Z}^{(i)}, \gamma_i) = \min \sum_{i=1}^{H} \gamma_i^P \big(\|\mathbf{U}^{(i)} - \mathbf{FZ}^{(i)}\|_{2,1}\big) \\
&s.t.\ \mathbf{F} \geq 0, \mathbf{Z}^{(i)} \geq 0, \gamma_i \geq 0, \forall i = 1, ..., H, \sum\nolimits_{i=1}^{H} \gamma_i = 1,
\end{aligned} \tag{4}$$

where γ_i is the weight assigned to the i-th view indicating the importance of that view. The exponent $P \geq 1$ is a parameter

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology

to control the smoothness of weights γ. $P = 1$ can lead to completely sparse weights which only a single view is selected. For $P > 1$, the weights will become smoother as the increase of the value of P.

Furthermore, we expect to capture the intrinsic manifold structure of data which can be effectively modeled by a nearest neighbor graph. Given N samples $\{x_i\}_{i=1}^N$ with H views, we construct a p-nearest neighbor graph $\mathbf{G}^{(h)}$ with affinity matrix $\mathbf{W}^{(h)} \in \mathbb{R}^{N\times N}$ for the h-th view. Then each graph encodes the local geometric structure information of the corresponding view. Gaussian kernel is one of the most commonly used similarity and we have $\mathbf{W}_{ij}^{(h)} = exp(-||x_i - x_j||^2/2\sigma^2)$. Then the graph Laplacian is defined as $\mathbf{L}^{(h)} = \mathbf{D}^{(h)} - \mathbf{W}^{(h)}$, where $\mathbf{D}^{(h)}$ is a diagonal matrix and $\mathbf{D}_{ii}^{(h)} = \sum_l \mathbf{W}_{il}^{(h)}$. However, it is infeasible to directly encode multiple Laplacian matrices $\{\mathbf{L}^{(h)}\}_{h=1}^H$ into the compact representation $\mathbf{F}$. Considering the different importance of views and make the reliable views contribute more, we also impose weights on each graph Laplacian matrix when encoding the structure information. Thus the objective function for learning compact representation is revised as

$$\begin{aligned} &\mathcal{O}_{pt}(\mathbf{U}^{(i)}, \mathbf{F}, \mathbf{Z}^{(i)}, \gamma_i) = \\ &\min \sum_{i=1}^{H} \gamma_i^P \Big[\big\| \mathbf{U}^{(i)} - \mathbf{F}\mathbf{Z}^{(i)} \big\|_{2,1} + \beta Tr(\mathbf{F}^T \mathbf{L}^{(i)} \mathbf{F}) \Big] \\ &s.t.\ \mathbf{Z}^{(i)} \geq 0, \gamma_i \geq 0, \forall i = 1, ..., H, \sum\nolimits_{i=1}^{H} \gamma_i = 1, \\ &\quad \mathbf{F} \geq 0, \mathbf{F}\mathbf{1}_R = \mathbf{1}_N, \end{aligned} \tag{5}$$

where β controls the strength of preserving structure information. We impose the ℓ_1 normalization constraints on the rows of $\mathbf{F}$ to handle the trivial solution problem [41], so that more reliable solutions can be obtained.

D. The Unified Objective Function

To learn the optimal comparable representation and compact representation simultaneously, the ultimate objective of BLMV is formulated by integrating the two subproblems into a unified function

$$\begin{aligned} &\mathcal{O}_U(\mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{F}, \mathbf{Z}^{(i)}, \gamma_i) = \mathcal{O}_{cp} + \mathcal{O}_{pt} = \\ &\min \sum_{i=1}^{H} \Big[\big\| \mathbf{X}^{(i)} - \big(\mathbf{U}_S^{(i)}, \mathbf{U}_P^{(i)}\big)\big(\mathbf{V}_S^{(i)}; \mathbf{V}_P^{(i)}\big) \big\|_F^2 + \eta \big\| \mathbf{V}_P^{(i)} \big\|_{2,1} \Big] \\ &+ \lambda \sum_{i=1}^{H-1} \sum_{j=i+1}^{H} \| \mathbf{U}_S^{(i)} - \mathbf{U}_S^{(j)} \|_F^2 + \sum_{i=1}^{H} \gamma_i^P \Big[\big\| \mathbf{U}^{(i)} - \mathbf{F}\mathbf{Z}^{(i)} \big\|_{2,1} \\ &+ \beta Tr(\mathbf{F}^T \mathbf{L}^{(i)} \mathbf{F}) \Big], \\ &s.t.\ \mathbf{U}^{(i)} \geq 0, \mathbf{V}^{(i)} \geq 0, \mathbf{Z}^{(i)} \geq 0, \gamma_i \geq 0, \forall i = 1, ..., H, \\ &\quad \mathbf{F} \geq 0, \mathbf{F}\mathbf{1}_R = \mathbf{1}_N, \sum\nolimits_{i=1}^{H} \gamma_i = 1. \end{aligned} \tag{6}$$

By solving problem (6), we can obtain the compact representation $\mathbf{F}$.

IV. Optimization

Apparently, problem (6) is not convex over all variables $\mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{F}, \mathbf{Z}^{(i)}$ and γ_i simultaneously, so we derive an iteration optimization algorithm to solve it. In each iteration, only one variable is updated while the others remain unchanged. We adopt the multiplicative iteration method to solve $\mathbf{U}^{(i)}, \mathbf{V}^{(i)}$ and $\mathbf{Z}^{(i)}$. The detailed derivation of updating rule for $\mathbf{U}^{(i)}$ is provided in Section IV-A. Since the derivation of updating rules of $\mathbf{V}^{(i)}$ and $\mathbf{Z}^{(i)}$ are similar to $\mathbf{U}^{(i)}$, we directly present the updating rules. Since $\mathbf{F}$ is imposed by an ℓ_1 normalization constraint, we introduce the method developed in [42], [43] to solve $\mathbf{F}$ (SectionIV-B). γ_i can be solved by using Lagrange multiplier method (Section IV-C). Next, the convergence proof of our algorithm is provided, and then we analyse the computational complexity of the proposed algorithm. The optimization process is summarized in Algorithm 2.

A. Update $\mathbf{U}^{(i)}, \mathbf{V}^{(i)} and\ \mathbf{Z}^{(i)}$

We derive the updating rules for the i-th view. The Lagrange multipliers ϕ_{jk} are introduced for constraint $\mathbf{U}_{jk}^{(i)} \geq 0$, and we denote them by a matrix $\Phi = [\phi_{jk}]$. Since $\mathbf{U}^{(i)}$ is made up by two parts $\big(\mathbf{U}_S^{(i)}, \mathbf{U}_P^{(i)}\big)$, we also separate $\Phi = (\Phi_S, \Phi_P)$ and $\mathbf{Z}^{(i)} = (\mathbf{Z}_S^{(i)}, \mathbf{Z}_P^{(i)})$ where $\Phi_S \in \mathbb{R}^{N\times K_S}$, $\Phi_P \in \mathbb{R}^{N\times K_P}$, $\mathbf{Z}_S^{(i)} \in \mathbb{R}^{N\times K_S}$ and $\mathbf{Z}_P^{(i)} \in \mathbb{R}^{N\times K_P}$. In order to solve $\mathbf{U}^{(i)}$, we keep the parts which are related to $\mathbf{U}^{(i)}$ from $\mathcal{O}_U$, and the Lagrange is

$$\begin{aligned} &\mathcal{L}(\mathbf{U}^{(i)}) = \big\| \mathbf{X}^{(i)} - \big(\mathbf{U}_S^{(i)}, \mathbf{U}_P^{(i)}\big)\big(\mathbf{V}_S^{(i)}; \mathbf{V}_P^{(i)}\big) \big\|_F^2 \\ &+ \lambda \sum_{j=1}^{H} \big\| \mathbf{U}_S^{(i)} - \mathbf{U}_S^{(j)} \big\|_F^2 + \gamma_i^P \big\| \big(\mathbf{U}_S^{(i)}, \mathbf{U}_P^{(i)}\big) - \mathbf{F}\big(\mathbf{Z}_S^{(i)}, \mathbf{Z}_P^{(i)}\big) \big\|_{2,1} \\ &- Tr[\mathbf{U}_S^{(i)} \Phi_S^T] - Tr[\mathbf{U}_P^{(i)} \Phi_P^T] \end{aligned} \tag{7}$$

The partial derivative of $\mathcal{L}(\mathbf{U}^{(i)})$ with respect to $\mathbf{U}_S^{(i)}$ is

$$\begin{aligned} &\frac{\partial \mathcal{L}(\mathbf{U}^{(i)})}{\partial \mathbf{U}_S^{(i)}} = -2\mathbf{X}^{(i)} \mathbf{V}_S^{(i)^T} + 2\mathbf{U}_S^{(i)} \mathbf{V}_S^{(i)} \mathbf{V}_S^{(i)^T} + 2\mathbf{U}_P^{(i)} \mathbf{V}_P^{(i)} \mathbf{V}_S^{(i)^T} \\ &+ 2\lambda \sum_{j=1}^{H} \big(\mathbf{U}_S^{(i)} - \mathbf{U}_S^{(j)}\big) + 2\gamma_i^P \big(\mathbf{\Lambda}_1^{(i)} \mathbf{U}_S^{(i)} - \mathbf{\Lambda}_1^{(i)} \mathbf{F}\mathbf{Z}_S^{(i)}\big) - \Phi_S \end{aligned} \tag{8}$$

where we use $\mathbf{\Lambda}_1^{(i)}$ to represent a diagonal matrix with elements $\big(\mathbf{\Lambda}_1^{(i)}\big)_{ll} = \frac{1}{2\left\| \left(\mathbf{U}_S^{(i)} - \mathbf{F}\mathbf{Z}_S^{(i)}\right)_{l.} \right\|}$, and the diagonal matrices in the following parts are also expressed in this way.

Using the Karush-Kuhn-Tuckre (KKT) condition $(\Phi_S)_{jk} (\mathbf{U}_S^{(i)})_{jk} = 0$, we obtain the following updating rule

$$\begin{aligned} &\big(\mathbf{U}_S^{(i)}\big)_{jk} \leftarrow \big(\mathbf{U}_S^{(i)}\big)_{jk} \nabla_{\mathbf{U}_S^{(i)}}, \text{ and } \nabla_{\mathbf{U}_S^{(i)}} = \\ &\frac{\big(\mathbf{X}^{(i)} \mathbf{V}_S^{(i)^T} + \lambda \sum_{j=1}^{H} \mathbf{U}_S^{(j)} + \gamma_i^P \mathbf{\Lambda}_1^{(i)} \mathbf{F}\mathbf{Z}_S^{(i)}\big)_{jk}}{\big(\mathbf{U}_S^{(i)} \mathbf{V}_S^{(i)} \mathbf{V}_S^{(i)^T} + \mathbf{U}_P^{(i)} \mathbf{V}_P^{(i)} \mathbf{V}_S^{(i)^T} + \lambda H \mathbf{U}_S^{(i)} + \gamma_i^P \mathbf{\Lambda}_1^{(i)} \mathbf{U}_S^{(i)}\big)_{jk}}. \end{aligned} \tag{9}$$

Similarly, the updating rules for $\mathbf{U}_P^{(i)}, \mathbf{V}^{(i)}$ and $\mathbf{Z}^{(i)}$ can be obtained as follows

$$\begin{aligned} &(\mathbf{U}_P^{(i)})_{jk} \leftarrow (\mathbf{U}_P^{(i)})_{jk} \nabla_{\mathbf{U}_P^{(i)}}, \text{ and } \nabla_{\mathbf{U}_P^{(i)}} = \\ &\frac{(\mathbf{X}^{(i)} \mathbf{V}_P^{(i)^T} + \gamma_i^P \mathbf{\Lambda}_2^{(i)} \mathbf{F}\mathbf{Z}_P^{(i)})_{jk}}{(\mathbf{U}_S^{(i)} \mathbf{V}_S^{(i)} \mathbf{V}_P^{(i)^T} + \mathbf{U}_P^{(i)} \mathbf{V}_P^{(i)} \mathbf{V}_P^{(i)^T} + \gamma_i^P \mathbf{\Lambda}_2^{(i)} \mathbf{U}_P^{(i)})_{jk}}, \end{aligned} \tag{10}$$

$$(\mathbf{V}_S^{(i)})_{jk} \leftarrow (\mathbf{V}_S^{(i)})_{jk} \frac{(\mathbf{U}_S^{(i)^T} \mathbf{X}^{(i)})_{jk}}{(\mathbf{U}_S^{(i)^T} \mathbf{U}_S^{(i)} \mathbf{V}_S^{(i)} + \mathbf{U}_S^{(i)^T} \mathbf{U}_P^{(i)} \mathbf{V}_P^{(i)})_{jk}}, \tag{11}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology

$\left(\mathbf{V}_P^{(i)}\right)_{jk} \leftarrow \left(\mathbf{V}_P^{(i)}\right)_{jk} \nabla_{\mathbf{V}_P^{(i)}}$, and $\nabla_{\mathbf{V}_P^{(i)}} =$

$$\frac{\left(\mathbf{U}_P^{(i)^T}\mathbf{X}^{(i)}\right)_{jk}}{\left(\mathbf{U}_P^{(i)^T}\mathbf{U}_S^{(i)}\mathbf{V}_S^{(i)}+\mathbf{U}_P^{(i)^T}\mathbf{U}_P^{(i)}\mathbf{V}_P^{(i)}+\eta\mathbf{\Lambda}_3^{(i)}\mathbf{V}_P^{(i)}\right)_{jk}}, \quad (12)$$

$$\mathbf{Z}_{jk}^{(i)} \leftarrow \mathbf{Z}_{jk}^{(i)} \frac{\left(\mathbf{F}^T\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)}\right)_{jk}}{\left(\mathbf{F}^T\mathbf{\Lambda}_4^{(i)}\mathbf{F}\mathbf{Z}^{(i)}\right)_{jk}}, \quad (13)$$

where $\left(\mathbf{\Lambda}_2^{(i)}\right)_{ll} = \frac{1}{2\left\|\left(\mathbf{U}_P^{(i)}-\mathbf{F}\mathbf{Z}_P^{(i)}\right)_{l\cdot}\right\|}$, $\left(\mathbf{\Lambda}_3^{(i)}\right)_{ll} = \frac{1}{2\left\|\left(\mathbf{V}_P^{(i)}\right)_{l\cdot}\right\|}$ and $\left(\mathbf{\Lambda}_4^{(i)}\right)_{ll} = \frac{1}{2\left\|\left(\mathbf{U}^{(i)}-\mathbf{F}\mathbf{Z}^{(i)}\right)_{l\cdot}\right\|}$.

B. Update F

By keeping the related parts of $\mathbf{F}$ in $\mathcal{O}_U$, we obtain the following minimization problem

$$\begin{aligned} &\min S(\mathbf{F}) \\ &s.t.\ \mathbf{F} \geq 0, \mathbf{F}\mathbf{1}_R = \mathbf{1}_N, \end{aligned} \quad (14)$$

where

$$S(\mathbf{F}) = \sum_{i=1}^{H} \gamma_i^P \left[\left\|\mathbf{U}^{(i)} - \mathbf{F}\mathbf{Z}^{(i)}\right\|_{2,1} + \beta Tr(\mathbf{F}^T\mathbf{L}^{(i)}\mathbf{F})\right]. \quad (15)$$

To effectively solve $\mathbf{F}$ with $\ell_{2,1}$-norm loss function, we introduce the following equation

$$\begin{aligned} J_1(\mathbf{F}) = &\sum_{i=1}^{H} \gamma_i^P Tr\big[\mathbf{U}^{(i)^T}\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)} - 2\mathbf{U}^{(i)^T}\mathbf{\Lambda}_4^{(i)}\mathbf{F}\mathbf{Z}^{(i)} \\ &+ \mathbf{Z}^{(i)^T}\mathbf{F}^T\mathbf{\Lambda}_4^{(i)}\mathbf{F}\mathbf{Z}^{(i)} + \beta\mathbf{F}^T(\mathbf{D}^{(i)} - \mathbf{W}^{(i)})\mathbf{F}\big]. \end{aligned} \quad (16)$$

where $\left(\mathbf{\Lambda}_4^{(i)}\right)_{ll} = \frac{1}{2\left\|\left(\mathbf{U}^{(i)}-\mathbf{F}\mathbf{Z}^{(i)}\right)_{l\cdot}\right\|}$.

In order to cope with the ℓ_1 normalization constraint of $\mathbf{F}$, we introduce another variable $\widetilde{\mathbf{F}}$ and define $\mathbf{F}_{jk} = \frac{\widetilde{\mathbf{F}}_{jk}}{\sum_s \widetilde{\mathbf{F}}_{js}}$. Then we can optimize $J_1(\mathbf{F})$ by minimizing $J_2(\widetilde{\mathbf{F}})$, where

$$\begin{aligned} J_2(\widetilde{\mathbf{F}}) = &\sum_{i=1}^{H} \gamma_i^P \big[Tr(\mathbf{U}^{(i)^T}\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)}) - 2\sum_{jk}(\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)}\mathbf{Z}^{(i)^T})_{jk}\frac{\widetilde{\mathbf{F}}_{jk}}{\sum_s \widetilde{\mathbf{F}}_{js}} \\ &+ \sum_{jk}[\mathbf{\Lambda}_4^{(i)}(\widetilde{\mathbf{F}} \oslash (\widetilde{\mathbf{F}}\mathbf{1}_R\mathbf{1}_R^T))\mathbf{Z}^{(i)}\mathbf{Z}^{(i)^T}]_{jk}\frac{\widetilde{\mathbf{F}}_{jk}}{\sum_s \widetilde{\mathbf{F}}_{js}} \\ &+ \beta\sum_{jkl}\mathbf{D}_{jl}^{(i)}\frac{\widetilde{\mathbf{F}}_{jk}}{\sum_s \widetilde{\mathbf{F}}_{js}}\frac{\widetilde{\mathbf{F}}_{lk}}{\sum_s \widetilde{\mathbf{F}}_{ls}} - \beta\sum_{jkl}\mathbf{W}_{jl}^{(i)}\frac{\widetilde{\mathbf{F}}_{jk}}{\sum_s \widetilde{\mathbf{F}}_{js}}\frac{\widetilde{\mathbf{F}}_{lk}}{\sum_s \widetilde{\mathbf{F}}_{ls}}\big]. \end{aligned} \quad (17)$$

The auxiliary function [50] of $J_2(\widetilde{\mathbf{F}})$ can be constructed as

$$\begin{aligned} Z(\widetilde{\mathbf{F}}, \widetilde{\mathbf{F}}') = &\sum_{i=1}^{H} \gamma_i^P \big[Tr(\mathbf{U}^{(i)^T}\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)}) \\ &- 2\sum_{jk}(\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)}\mathbf{Z}^{(i)^T})_{jk}\frac{\widetilde{\mathbf{F}}'_{jk}}{\sum_s \widetilde{\mathbf{F}}'_{js}}(1 + \log\frac{\widetilde{\mathbf{F}}_{jk}/\sum_s \widetilde{\mathbf{F}}_{js}}{\widetilde{\mathbf{F}}'_{jk}/\sum_s \widetilde{\mathbf{F}}'_{js}}) \\ &+ \sum_{jk}[\mathbf{\Lambda}_4^{(i)}(\widetilde{\mathbf{F}}' \oslash (\widetilde{\mathbf{F}}'\mathbf{1}_R\mathbf{1}_R^T))\mathbf{Z}^{(i)}\mathbf{Z}^{(i)^T}]_{jk}\frac{(\widetilde{\mathbf{F}}_{jk}/\sum_s \widetilde{\mathbf{F}}_{js})^2}{\widetilde{\mathbf{F}}'_{jk}/\sum_s \widetilde{\mathbf{F}}'_{js}} \\ &+ \beta\sum_{jk}\frac{[\mathbf{D}^{(i)}(\widetilde{\mathbf{F}}' \oslash (\widetilde{\mathbf{F}}'\mathbf{1}_R\mathbf{1}_R^T))]_{jk}(\widetilde{\mathbf{F}}_{jk}/\sum_s \widetilde{\mathbf{F}}_{js})^2}{\widetilde{\mathbf{F}}'_{jk}/\sum_s \widetilde{\mathbf{F}}'_{js}} \\ &- \beta\sum_{jkl}\mathbf{W}_{jl}^{(i)}\frac{\widetilde{\mathbf{F}}'_{jk}}{\sum_s \widetilde{\mathbf{F}}'_{js}}\frac{\widetilde{\mathbf{F}}'_{lk}}{\sum_s \widetilde{\mathbf{F}}'_{ls}}(1 + \log\frac{(\widetilde{\mathbf{F}}_{jk}/\sum_s \widetilde{\mathbf{F}}_{js})(\widetilde{\mathbf{F}}_{lk}/\sum_s \widetilde{\mathbf{F}}_{ls})}{(\widetilde{\mathbf{F}}'_{jk}/\sum_s \widetilde{\mathbf{F}}'_{js})(\widetilde{\mathbf{F}}'_{lk}/\sum_s \widetilde{\mathbf{F}}'_{ls})})\big]. \end{aligned} \quad (18)$$

Setting the partial derivative of $\widetilde{\mathbf{F}}_{jk}$ to zero, we have

$$\frac{\partial Z(\widetilde{\mathbf{F}}, \widetilde{\mathbf{F}}')}{\widetilde{\mathbf{F}}_{jk}} = \sum_t \frac{\partial Z(\widetilde{\mathbf{F}}, \widetilde{\mathbf{F}}')}{\partial(\frac{\widetilde{\mathbf{F}}_{jt}}{\sum_s \widetilde{\mathbf{F}}_{js}})}\frac{\partial(\frac{\widetilde{\mathbf{F}}_{jt}}{\sum_s \widetilde{\mathbf{F}}_{js}})}{\partial\widetilde{\mathbf{F}}_{jk}} = 0. \quad (19)$$

By replacing $\frac{\widetilde{\mathbf{F}}_{jt}}{\sum_s \widetilde{\mathbf{F}}_{js}} = \mathbf{F}_{jt}$, we can obtain

$$\begin{aligned} \frac{\partial Z(\widetilde{\mathbf{F}}, \widetilde{\mathbf{F}}')}{\partial(\frac{\widetilde{\mathbf{F}}_{jt}}{\sum_s \widetilde{\mathbf{F}}_{js}})} = &\sum_{i=1}^{H} \gamma_i^P \big[-2(\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)}\mathbf{Z}^{(i)^T})_{jt}\frac{\mathbf{F}'_{jt}}{\mathbf{F}_{jt}} \\ &+ 2(\mathbf{\Lambda}_4^{(i)}\mathbf{F}'\mathbf{Z}^{(i)}\mathbf{Z}^{(i)^T})_{jt}\frac{\mathbf{F}_{jt}}{\mathbf{F}'_{jt}} + \frac{2\beta(\mathbf{D}^{(i)}\mathbf{F}')_{jt}\mathbf{F}_{jt}}{\mathbf{F}'_{jt}} \\ &- \frac{2\beta(\mathbf{W}^{(i)}\mathbf{F}')_{jt}\mathbf{F}'_{jt}}{\mathbf{F}_{jt}}\big]. \end{aligned} \quad (20)$$

In addition, we have

$$\frac{\partial(\frac{\widetilde{\mathbf{F}}_{jt}}{\sum_s \widetilde{\mathbf{F}}_{js}})}{\partial\widetilde{\mathbf{F}}_{jk}} = \frac{\delta_{kt} - \mathbf{F}_{jt}}{\sum_s \widetilde{\mathbf{F}}_{js}}, \quad \delta_{tk} = \begin{cases} 1 & if\ t = k \\ 0 & otherwise \end{cases} \quad (21)$$

Let $\mathbf{\Gamma} = \left[\sum_{i=1}^{H} \gamma_i^P(\mathbf{\Lambda}_4^{(i)}\mathbf{U}^{(i)}\mathbf{Z}^{(i)^T} + \beta\mathbf{W}^{(i)}\mathbf{F}')\right] \odot \mathbf{F}'$ and $\mathbf{\Theta} = \left[\sum_{i=1}^{H} \gamma_i^P(\mathbf{\Lambda}_4^{(i)}\mathbf{F}'\mathbf{Z}^{(i)}\mathbf{Z}^{(i)^T} + \beta\mathbf{D}^{(i)}\mathbf{F}')\right] \oslash \mathbf{F}'$, then equation(19) can be written as

$$\mathbf{\Theta}_{jk}\mathbf{F}_{jk}^2 + \sum_s[\mathbf{\Gamma}_{js} - \mathbf{\Theta}_{js}\mathbf{F}_{js}^2]\mathbf{F}_{jk} - \mathbf{\Gamma}_{jk} = 0 \quad (22)$$

By summing equation (22) over k, we obtain

$$\begin{aligned} &\sum_k \mathbf{\Theta}_{jk}\mathbf{F}_{jk}^2 + \sum_s[\mathbf{\Gamma}_{js} - \mathbf{\Theta}_{js}\mathbf{F}_{js}^2]\sum_k \mathbf{F}_{jk} - \sum_k \mathbf{\Gamma}_{jk} \\ &= \sum_s[\mathbf{\Gamma}_{js} - \mathbf{\Theta}_{js}\mathbf{F}_{js}^2](\sum_k \mathbf{F}_{jk} - 1) = 0, \end{aligned} \quad (23)$$

where we can observe that the solution of equation (23) always satisfies ℓ_1 normalization for almost all β. To obtain $\mathbf{F}$, we compute the fixed point of equation (22) by Algorithm 1 which is developed in [42], [43]. The convergence of Algorithm 1 has been proved in [42] and only a few iterations are needed to converge (always less than 10 iterations in our experiments).

Algorithm 1: Computing the fixed point of equation (22)

```
Input: Γ, Θ, F, MaxIter
Output: F
while not converged do
    Ψ = (Γ − Θ ⊙ F²) 1_R 1_R^T;
    F = (√(Ψ² + 4ΘΓ) − Ψ) ⊘ (2Θ);
    I = {i | Ψ_i1 < 0};
    if I is not empty then
        F_I· ← diag(F_I· 1_R)^(−1) F_I·;
    end
end
```

C. Update γ_i

Using Lagrange multiplier method, we can derive the following updating rule for γ_i.

If $P > 1$,

$$\gamma_i = 1 \Big/ \sum_{j=1}^{H} \left(\mathbf{\Delta}^{(i)} \Big/ \mathbf{\Delta}^{(j)} \right)^{1/(P-1)} \tag{24}$$

For $P = 1$ case, we have

$$\gamma_i = \begin{cases} 1, & i = \text{argmin}_i \mathbf{\Delta}^{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

where $\mathbf{\Delta}^{(i)} = \left\| \mathbf{U}^{(i)} - \mathbf{F}\mathbf{Z}^{(i)} \right\|_{2,1} + \beta Tr(\mathbf{F}^T \mathbf{L}^{(i)} \mathbf{F})$.

D. Convergence Analysis

In each iteration, BLMV needs to update all the variables $\gamma_i, \mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{F}$ and $\mathbf{Z}^{(i)}$. It is obviously that the lower bound of objective function $\mathcal{O}_U$ in (6) is zero. To prove the convergence of the updating rules, we need to show that $\mathcal{O}_U$ is non-increasing under each updating step. Since the solution for γ_i is obtained by Lagrange multiplier method, the objective is non-increasing after updating and it is convergent. For the other variables $\mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{F}$ and $\mathbf{Z}^{(i)}$, we make use of auxiliary function method [8] to prove the convergence. The detailed proof of the convergence of updating $\mathbf{F}$ is provided, and we skip the proofs with respect to $\mathbf{U}^{(i)}, \mathbf{V}^{(i)}$ and $\mathbf{Z}^{(i)}$ since they are similar in spirit to the proof with respect to $\mathbf{F}$.

First, we introduce the following lemma, and then the convergence is proved in Theorem 1.

Lemma 1: In each update of $\mathbf{F}$, the following inequation holds

$$\begin{aligned} & S(\mathbf{F}^{t+1}) - S(\mathbf{F}^t) \leq \\ & \sum_{i=1}^{H} \gamma_i^P \{ Tr[(\mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)})^T \mathbf{\Lambda}_4^{(i)} (\mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)})] \\ & + \beta Tr(\mathbf{F}^{t+1^T} \mathbf{L}^{(i)} \mathbf{F}^{t+1}) - Tr[(\mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)})^T \mathbf{\Lambda}_4^{(i)} (\mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)})] \\ & - \beta Tr(\mathbf{F}^{t^T} \mathbf{L}^{(i)} \mathbf{F}^{t}) \}, \end{aligned} \tag{26}$$

where $\left(\mathbf{\Lambda}_4^{(i)}\right)_{ll} = \frac{1}{2\left\| \left(\mathbf{U}^{(i)} - \mathbf{F}^t \mathbf{Z}^{(i)} \right)_{l.} \right\|}$.

Proof:

$$\begin{aligned} & S(\mathbf{F}^{t+1}) - S(\mathbf{F}^t) \\ & = \sum_{i=1}^{H} \gamma_i^P \Big[\left\| \mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)} \right\|_{2,1} + \beta Tr(\mathbf{F}^{t+1^T} \mathbf{L}^{(i)} \mathbf{F}^{t+1}) \\ & - \left\| \mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)} \right\|_{2,1} - \beta Tr(\mathbf{F}^{t^T} \mathbf{L}^{(i)} \mathbf{F}^{t}) \Big]. \end{aligned} \tag{27}$$

Comparing (27) with the right hand side of (26), we only need to prove the following inequality holds

$$\begin{aligned} & \sum_{i=1}^{H} \gamma_i^P \Big[\left\| \mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)} \right\|_{2,1} - \left\| \mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)} \right\|_{2,1} \Big] \leq \\ & \sum_{i=1}^{H} \gamma_i^P \Big\{ Tr[(\mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)})^T \mathbf{\Lambda}_4^{(i)} (\mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)})] \\ & - Tr[(\mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)})^T \mathbf{\Lambda}_4^{(i)} (\mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)})] \Big\}. \end{aligned} \tag{28}$$

It is obvious to prove the above inequation by a similar inequation which is given by Lemma 3 in [46]. □

Theorem 1: The objective function $\mathcal{O}_U$ in equation (6) is non-increasing after updating $\mathbf{F}$.

Proof: Based on the properties of the auxiliary function, we have $J_1(\mathbf{F}^{t+1}) \leq Z(\mathbf{F}^{t+1}, \mathbf{F}^t) \leq Z(\mathbf{F}^t, \mathbf{F}^t) = J_1(\mathbf{F}^t)$. Thus $J_1(F)$ will monotonically decrease under the update of $\mathbf{F}$, so we have $J_1(\mathbf{F}^{t+1}) \leq J_1(\mathbf{F}^t)$, that is,

$$\begin{aligned} & \sum_{i=1}^{H} \gamma_i^P \{ Tr[(\mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)})^T \mathbf{\Lambda}_4^{(i)} (\mathbf{U}^{(i)} - \mathbf{F}^{t+1}\mathbf{Z}^{(i)})] \\ & + \beta Tr(\mathbf{F}^{t+1^T} \mathbf{L}^{(i)} \mathbf{F}^{t+1}) \} \leq \\ & \sum_{i=1}^{H} \gamma_i^P \{ Tr[(\mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)})^T \mathbf{\Lambda}_4^{(i)} (\mathbf{U}^{(i)} - \mathbf{F}^{t}\mathbf{Z}^{(i)})] \\ & + \beta Tr(\mathbf{F}^{t^T} \mathbf{L}^{(i)} \mathbf{F}^{t}) \}, \end{aligned} \tag{29}$$

where $\left(\mathbf{\Lambda}_4^{(i)}\right)_{ll} = \frac{1}{2\left\| \left(\mathbf{U}^{(i)} - \mathbf{F}^t \mathbf{Z}^{(i)} \right)_{l.} \right\|}$.

According to Lemma 1, we have $S(\mathbf{F}^{t+1}) - S(\mathbf{F}^t) \leq 0$. This proves that the objective function $\mathcal{O}_U$ in equation (6) is non-increasing after updating $\mathbf{F}$ in each iteration. □

E. Computational Complexity Analysis

Before updating each variable of BLMV, the nearest neighbour graphs are constructed for different views. We need $O(N^2 M_i)$ to construct these graphs. During the updating process of BLMV, we calculate the computational cost in each iteration based on updating rules. It should be noted that the p-nearest neighbor graphs $\{\mathbf{W}^{(i)}\}_{i=1}^H$ are highly sparse matrices. Therefore, the computational cost to calculate $\mathbf{W}^{(i)}\mathbf{F}$ is only $O(pNR)$. The overall computational complexities for each variable are: $O(NM_iK)$ for updating $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$, and $O(NKR)$ for updating $\mathbf{F}$ and $\mathbf{Z}^{(i)}$, where N is the number of samples, M_i, K and R are the dimensionality of the i-th view, the comparable representation and the compact representation, respectively. The computational cost of the updating process is comparable with that of standard NMF, therefore our method is efficient.

Algorithm 2: The algorithm of BLMV

Input: Multi-view data matrices $\mathcal{X} = \{\mathbf{X}^{(i)}\}_{i=1}^H$, graph Laplacians $\mathbf{L}^{(h)} = \mathbf{D}^{(h)} - \mathbf{W}^{(h)}, \forall i = 1, ..., H$ and parameters: λ, η, β, P, θ, K, R

Output: The compact representation $\mathbf{F}$

```
Initialize {U^(i), V^(i), Z^(i)}_{i=1}^H, F with non-negative
values, and let γ_i = 1/H, ∀i = 1, ..., H
while not converged do
    for i = 1 to H do
        Update U_S^(i) and U_P^(i) by equation (9) and (10);
        Update V_S^(i) and V_P^(i) by equation (11) and (12);
    end
    for i = 1 to H do
        Update Z^(i) by equation (13);
    end
    Compute Γ and Θ;
    Update F by Algorithm 1;
    Update γ_i, ∀i = 1, ..., H by equation (24) and (25);
end
```

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology

V. Experimental Results

To verify the effectiveness of the proposed method BLMV, we investigate its performance in the tasks of image classification and annotation on two benchmark real-world image datasets. Our method is compared to several representative multi-view learning methods to demonstrate its effectiveness. First, we introduce the datasets, the compared methods and experimental settings. Then, we present the image classification and annotation performance of the proposed method and the other methods. Finally, parameter sensitivity analysis are provided to illustrate the properties of BLMV.

A. Datasets

Two real-world image datasets NUS-WIDE (NUS) [51] and PASCAL VOC'07 (VOC) [52] are adopted in our experiments. NUS is a web image dataset and it includes $269,648$ images and associated tags from Flickr. There are 81 ground truth labels can be used for evaluation. VOC dataset contains around $10,000$ images obtained from Flickr web site and all of them are annotated for 20 categories. Some example images from the two datasets are shown in Fig. 2 and Fig. 3, respectively.

From each of the two datasets, we randomly sample 6000 images as our datasets. Five different visual features are extracted as five views of image data. These visual features and their dimensionalities are block-wise color moments (729) [53], bag of visual words based on SIFT (1000) [54], HOG (3100) [55], GIST (512) [56] and LBP (928) [57].

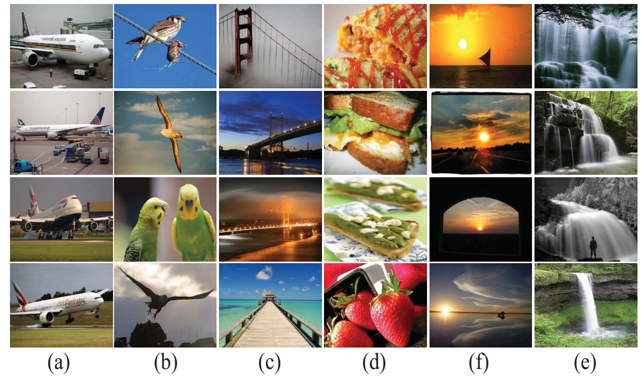


Fig. 2. Some example images from NUS-WIDE dataset. (a) Airport. (b) Bird. (c) Bridge. (d) Food. (e) Sunset. (f) Waterfall.

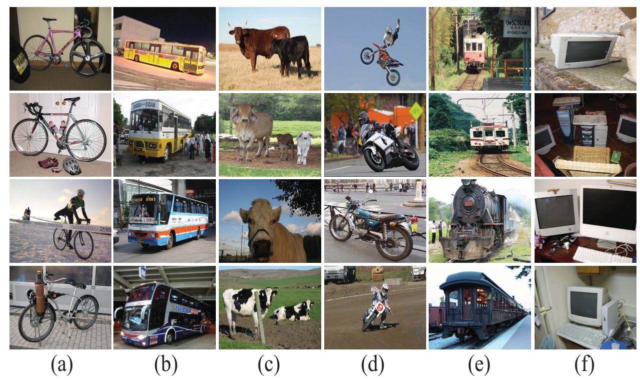


Fig. 3. Some example images from PASCAL VOC dataset. (a) Bicycle. (b) Bus. (c) Cow. (d) Motorbike. (e) Train. (f) Tvmonitor.

B. Compared Methods

To demonstrate the effectiveness of BLMV, we learn the low-dimensional representation by several compared methods, including two traditional PCA-based methods (1-2), six recently proposed multi-view learning methods (3-8) and two versions of our model with different settings (9, 10).

1) *Single view PCA* (sPCA): Perform principle component analysis (PCA) for each view to obtain the low-dimensional representation and report the results with best performance.
2) *Multi-view PCA* (mPCA): Concatenate each view together and then perform PCA to obtain the low-dimensional representation.
3) *Multi-view Spectral Embedding* (MVSE): A spectral embedding method for multi-view dimensionality reduction proposed in [1].
4) *Multi-feature Spectral Clustering with Minimax Optimization* (MSCMO): A multi-view spectral clustering method [58] and the learned feature embedding is used as the low-dimensional representation.
5) *Group Sparse Multi-view Patch Alignment Framework* (GSMVPA): A joint multi-view feature extraction and feature selection method [21] based on patch alignment.
6) *Sparse Dimensionality Reduction for Multi-view Data* (SSMVD): A multi-view dimensionality reduction method [5] based on structured sparse PCA which improves flexibility in sharing information across different views.
7) *Ensemble Manifold Regularized Sparse Low-rank Approximation* (EMRSLRA): A multi-view feature embedding method [40] that based on least-squares component analysis and ensemble manifold learning.
8) *Co-regularized Non-negative Matrix Factorization* (CoNMF): An NMF based multi-view clustering method [49]. The learned coefficient matrix is used as the low-dimensional representation. Both pair-wise and cluster-wise CoNMF are performed and best results are reported.
9) *BLMV without the first level* (BLMV_SL): Let $\mathbf{X}^{(i)}$ instead of $\mathbf{U}^{(i)}$ to feed into the second level of BLMV. This method is to validate that the proposed bi-level learning strategy can more effectively leverage the multi-view complementary information.
10) *BLMV using Frobenius Norm* (BLMV_F): Replace $\ell_{2,1}$-norm with F-norm in the second level of BLMV. This baseline is used to test how our method (with $\ell_{2,1}$-norm) is resistent to noise and outliers.

C. Evaluation Metrics

To evaluate the image classification and annotation performance, we adopt the three criteria as adopted in [5]: accuracy (ACC), the area under the receiver operating characteristics curve (AUC) and F1 (F-measure) scores. They are typical evaluation metrics for the tasks of image classification and annotation.

ACC is defined as ACC=N_r/N_t, where N_t is the total number of test images to be labeled and N_r is the number of images that are annotated with right labels or correctly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology

classified. The detailed definition of AUC and F1 scores can be referred to [59].

D. Experimental Setting

We evaluate the effectiveness of low-dimensional representation learned by BLMV in the tasks of image classification and annotation. Specifically, all the multi-view data are first embedded into a low-dimensional space by using both the compared methods and BLMV. Then, we conduct image classification in this space based on linear SVM classifiers. The one-against-the-rest scheme is adopted where one classifier is trained for each category. For all the methods, we learn the low-dimensional representation with different dimensionalities $R = \{50, 100, 150, 200, 250, 300, 350, 400\}$, and then image classification is conducted for each R. Higher classification performance indicates that more effective low-dimensional representation is obtained. We randomly partition the dataset into training set and test set ten times and the averaged performance are reported.

The parameters of the compared methods are set as suggested in their works. There are several parameters in our method BLMV. To model the geometric structure of image data, we construct a 7-NN graph for each view and the corresponding graph Laplacian matrix. The Gaussian kernel width parameter σ is estimated by the average of the pair-wise distances. For the rest of the parameters, we determine them by cross-validation. In the first random partition of training and test set, the training set are split into five folds. Then we tune the parameters by five-fold cross-validation and choose the parameter settings with the best average performance. These parameter settings are used for training and testing. Thus we set $\lambda = 1$, $\eta = 1$, $P = 2$, $\theta = 0.75$, $K = 400$ and $\beta = 1e-4$ for NUS and VOC dataset in our experiments. To further illustrate the properties of these parameters, we will present the detailed parameter sensitivity analysis in Section V-F.

E. Results Analysis

We first test image classification and annotation performance with dimensionality $R = 100$, and we randomly select 1000 images as training set and the rest images are used for testing. The mean and standard error of image classification and annotation performance on NUS and VOC datasets are listed in Table I. It is obvious that BLMV achieves the best image classification and annotation performance compared with the other methods on both datasets, which demonstrates the effectiveness of our method in multi-view low-dimensional representation learning. Compared with the best results obtained by the other methods, our method achieves improvement of 2.1% in ACC, 1.0% in AUC and 1.5% in F1 score for NUS dataset and 2.7% in ACC, 1.0% in AUC and 2.9% in F1 score for VOC dataset.

Moreover, we conduct experiments with different size of training set and fix $R = 100$. The size of training set is set to be $\{500, 1000, 1500, 2000, 2500\}$ and the detailed performance are shown in Fig. 4. It can be clearly observed that as we use more data for training classifiers, the performance of each method improves stably. Compared with the other methods, BLMV achieves the best performance with different number of training data, which indicates that the low-dimensional representation obtained by BLMV is more discriminative and effective than that obtained by the other methods.

TABLE I
IMAGE CLASSIFICATION AND ANNOTATION COMPARISON ON DIFFERENT DATASETS.

(a) Performance comparison on NUS dataset.

Method	ACC Score	AUC Score	F1 Score
sPCA	0.669 ± 0.004	0.744 ± 0.005	0.303 ± 0.004
mPCA	0.707 ± 0.007	0.791 ± 0.003	0.345 ± 0.003
MVSE	0.701 ± 0.012	0.749 ± 0.003	0.317 ± 0.006
MSCMO	0.716 ± 0.010	0.758 ± 0.004	0.330 ± 0.003
GSMVPA	0.696 ± 0.006	0.780 ± 0.003	0.336 ± 0.004
SSMVD	0.710 ± 0.005	0.793 ± 0.003	0.348 ± 0.004
EMRSLRA	0.700 ± 0.006	0.792 ± 0.002	0.343 ± 0.004
CoNMF	0.699 ± 0.014	0.793 ± 0.008	0.343 ± 0.005
BLMV_SL	0.724 ± 0.013	0.788 ± 0.008	0.349 ± 0.007
BLMV_F	0.717 ± 0.010	0.797 ± 0.006	0.354 ± 0.005
BLMV	$\mathbf{0.737 \pm 0.007}$	$\mathbf{0.803 \pm 0.006}$	$\mathbf{0.363 \pm 0.006}$

(b) Performance comparison on VOC dataset.

Method	ACC Score	AUC Score	F1 Score
sPCA	0.816 ± 0.011	0.754 ± 0.007	0.224 ± 0.005
mPCA	0.827 ± 0.006	0.786 ± 0.003	0.247 ± 0.004
MVSE	0.823 ± 0.005	0.748 ± 0.004	0.223 ± 0.004
MSCMO	0.830 ± 0.009	0.756 ± 0.006	0.232 ± 0.004
GSMVPA	0.845 ± 0.010	0.777 ± 0.007	0.251 ± 0.003
SSMVD	0.849 ± 0.006	0.794 ± 0.003	0.268 ± 0.002
EMRSLRA	0.840 ± 0.005	0.793 ± 0.004	0.263 ± 0.005
CoNMF	0.859 ± 0.007	0.777 ± 0.003	0.254 ± 0.003
BLMV_SL	0.871 ± 0.008	0.778 ± 0.005	0.269 ± 0.005
BLMV_F	0.873 ± 0.006	0.786 ± 0.003	0.277 ± 0.006
BLMV	$\mathbf{0.886 \pm 0.005}$	$\mathbf{0.804 \pm 0.004}$	$\mathbf{0.297 \pm 0.006}$

To further demonstrate the performance improvement of our method compared with the other methods in image classification and annotation, we conduct image classification and annotation experiments at different dimensionalities $R = \{50, 100, ..., 400\}$. 1000 images are used for training and the rest are used for testing. Fig. 5 shows the detailed experimental results. We can observe that the two spectral embedding methods MVSE and MSCMO usually achieve their best performance at lower dimensions, and the performance degraded as the dimension increasing. This is because the learned basis by these two methods are orthogonal and the first few dimensions may have high variance and they are quite discriminative. However, the variance of basis becomes lower and lower as the dimension increases, so the representation becomes ambiguous and meaningless since it is dominated by these low discriminative dimensions. On the other hand, some latent space based methods such as SSMVD, CONMF and BLMV achieve better performance at higher dimensions. This is reasonable because more information can be preserved with more latent factors. Compared with the other methods, BLMV generally achieves better performance at different dimensions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology

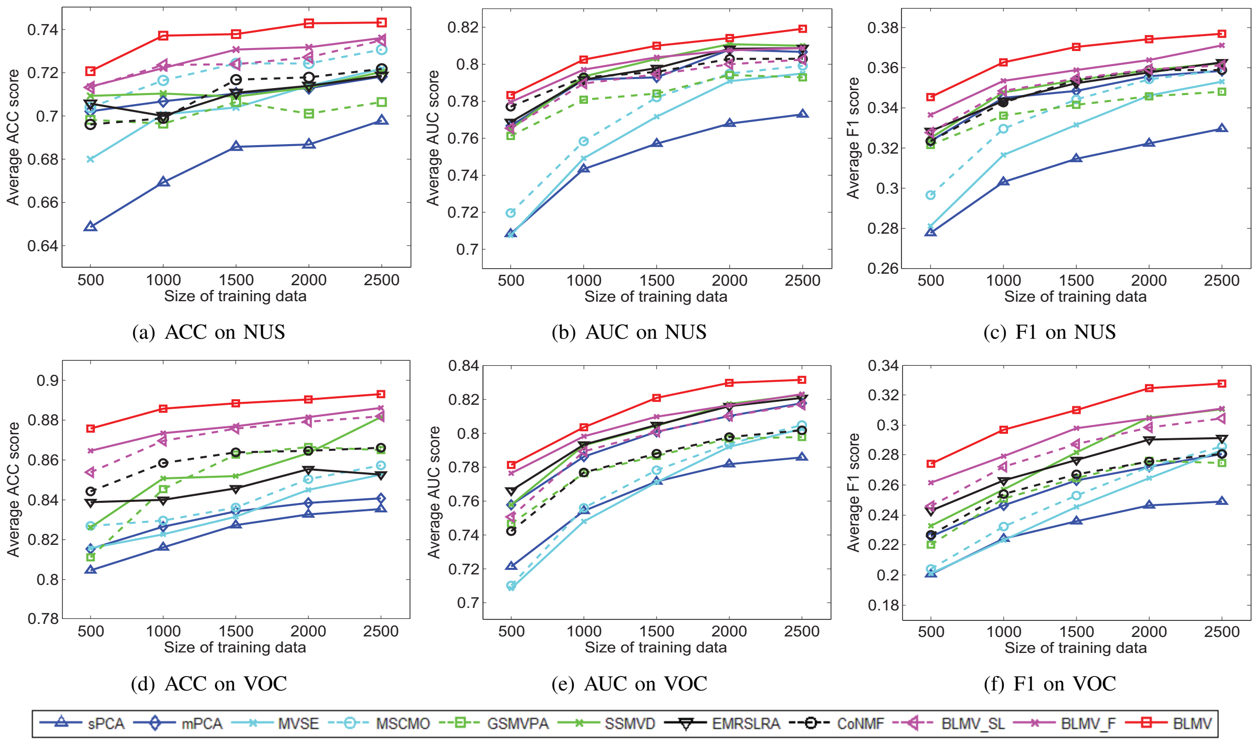


Fig. 4. The performance comparison of image classification and annotation with different numbers of training data on NUS and VOC datasets. The size of training set varies in the range of $\{500, 1000, 1500, 2000, 2500\}$ and R is fixed to 100.

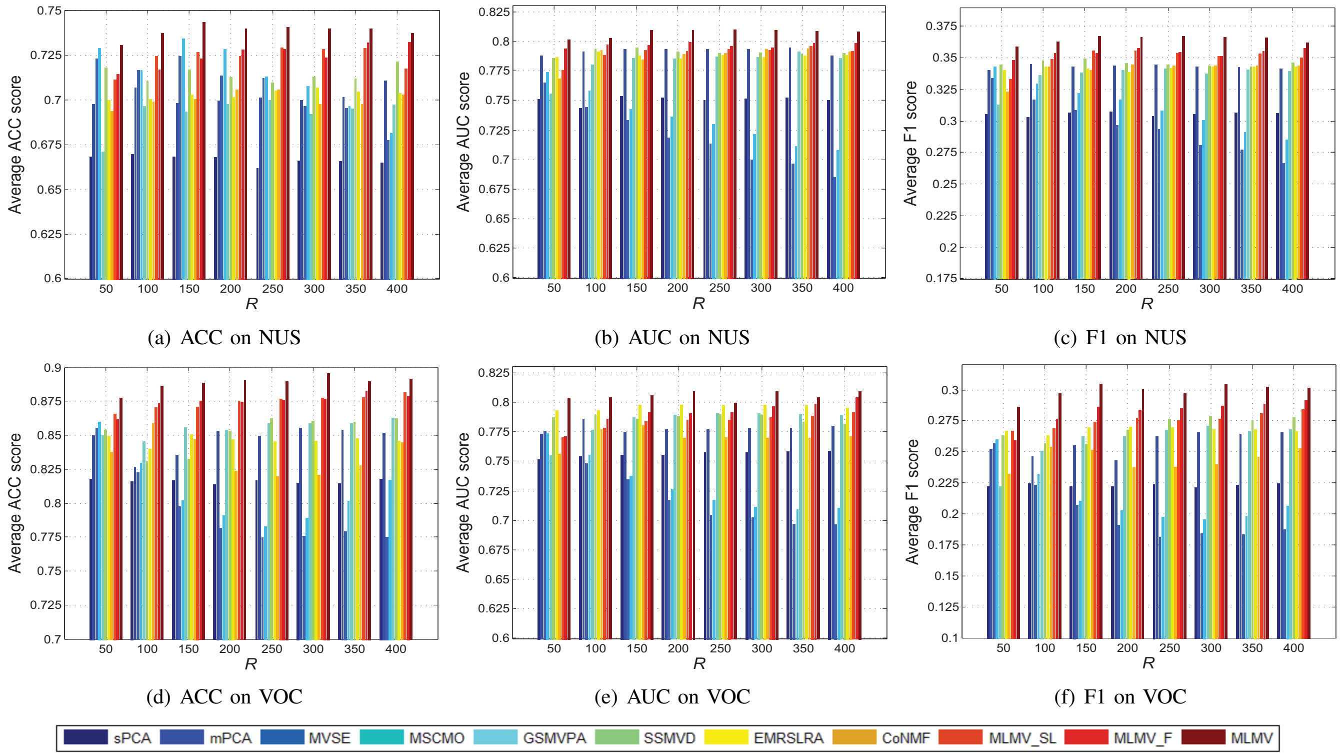
In particular, BLMV obtains relatively stable performance when $R > 50$, which indicates that its performance is not sensitive to R and BLMV can effectively encode the multi-view information at a wide range of dimension settings.

Throughout the experiments, we reveal several interesting points as follows:

1) All the multi-view dimensionality reduction methods are superior to the single view based method sPCA. This indicates that exploiting multiple views can generate a more powerful description than only using a single view.
2) From the performance comparison between two multi-view spectral embedding methods MVSE and MSCMO, we can find that by considering the different importance of views and minimizing pairwise disagreement between any two views, MSCMO learns a more harmonic consensus representation. Therefore, it fuses multi-view information more effectively and achieves better performance than MVSE.
3) Compared with EMRSLRA, BLMV can more effectively conduct multi-view fusion. It learns non-negative latent spaces which are naturally interpretable so that different views with different physical meanings are comparable. Besides, by assigning different weights to multiple views during multi-view matrix factorization, BLMV can accommodate unreliable views and obtain more accurate results than EMRSLRA.
4) BLMV outperforms another non-negative latent space learning method CoNMF. CoNMF only regularizes the coefficient matrices of different views towards a consensus, whereas BLMV simultaneously utilizes the shared and private nature as well as the intrinsic geometric structure information of multi-view data, which helps to obtain a more compact and discriminative low-dimensional representation.
5) BLMV_SL performs unsatisfactorily because it directly learns the low-dimensional representation from heterogeneous multi-view features which ignores the incomparability of different views. In addition, neither shared nor private information of multi-view data can be well captured by BLMV_SL so that its overall performance is limited. The experimental results of BLMV_SL illustrates that comparable representation learning helps to accurately capture the information of each view and further enhances the effectiveness of compact representation. The proposed bi-level learning strategy is both reasonable and efficient for multi-view representation learning.
6) BLMV further improves the image classification performance compared with BLMV_F, which demonstrates that the adopted $\ell_{2,1}$-norm in the second level is capable of resisting the outliers and noise contained in the multi-view data and provides more accurate and reliable learning results. Specifically, compared with BLMV_F from Table I, BLMV gains 2.0%, 0.6%, and 0.9% improvement in terms of ACC, AUC, and F1 scores on NUS dataset and it also gains 1.3%, 1.8%, and 2.0% improvement in terms of ACC, AUC, and F1 scores on VOC dataset.

Next, we illustrate the weight of each view γ_i learned by BLMV and analyze the different importance of views. The detailed weights are shown in Fig. 6. We can observe that the weights learned from the two datasets share similar distribution properties, i.e., HOG, SIFT and LBP obtain higher weights than color moments and GIST. This indicates that the images from the two real-world datasets have similar visual appearance, and using the local features such as HOG,

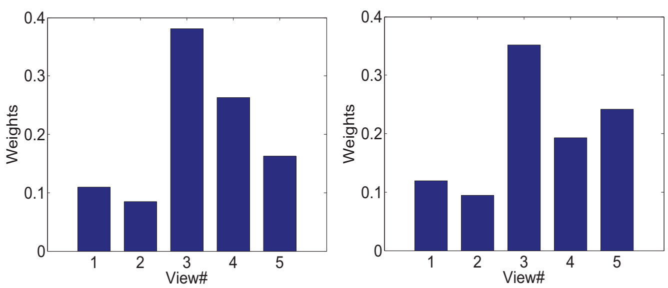
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology



(a) ACC on NUS (b) AUC on NUS (c) F1 on NUS

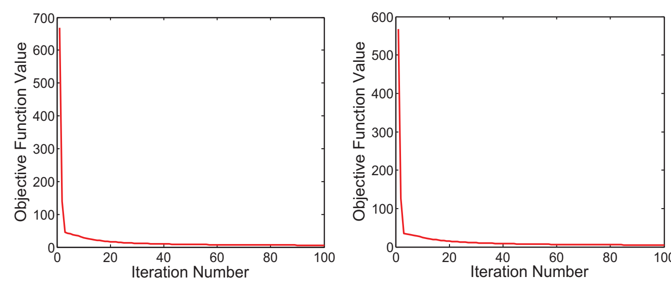
(d) ACC on VOC (e) AUC on VOC (f) F1 on VOC

Fig. 5. The performance comparison of image classification and annotation with different dimensionalities $R = \{50, 100, ..., 400\}$. The size of training data is set to 1000.



(a) NUS (b) VOC

Fig. 6. The learned weights γ_i on each dataset. View 1 to view 5 represent color moments, GIST, HOG, SIFT and LBP features, repectively.



(a) Convergence curve on NUS (b) Convergence curve on VOC

Fig. 7. Convergence curves of BLMV over NUS and VOC datasets.

SIFT and LBP can achieve better discrimination results. Color moments and GIST are relatively unreliable since they obtain lower weights. HOG, SIFT and LBP generate more accurate descriptions so that they are more important for multi-view dimensionality reduction.

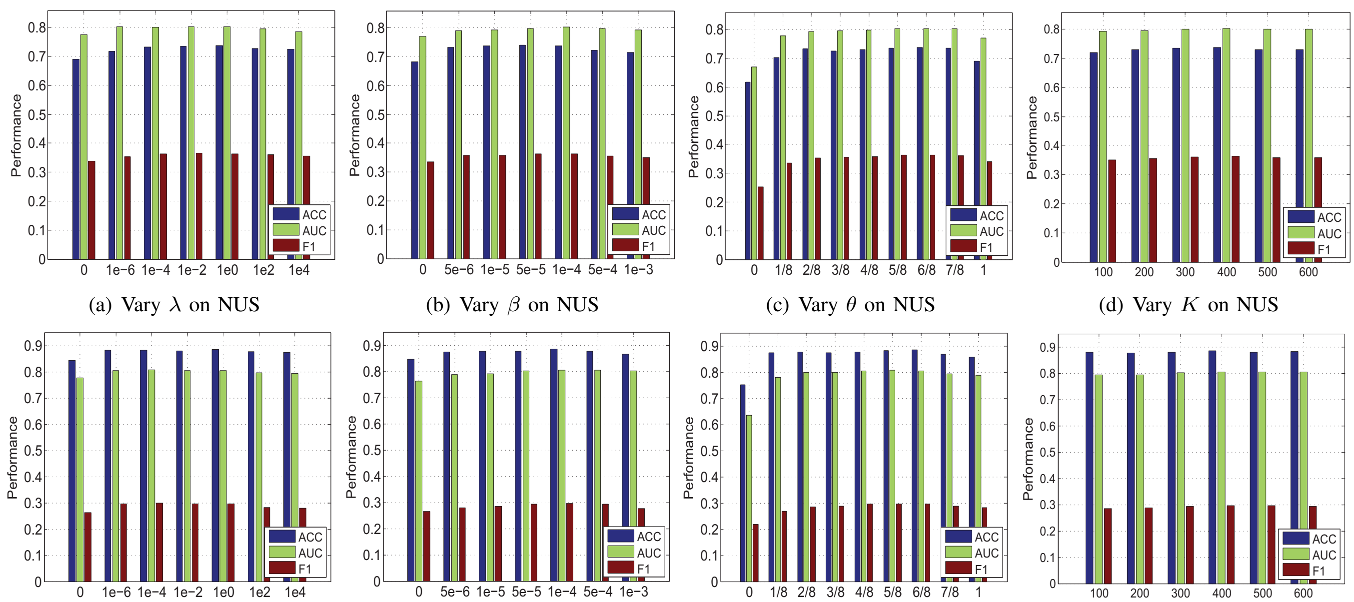
In the end, the convergence curves of the two datasets are shown in Fig. 7, where we can see that the objective function value is non-increasing after each iteration. Fig. 7 indicates that the proposed optimization algorithm can effectively solve the objective function and converges quickly.

F. Parameter Sensitivity Analysis

There are four critical parameters λ, β, θ and K in BLMV. λ controls the consistency of the shared parts of each view. β measures the strength of encoding the intrinsic geometric structure into the low-dimensional representation. θ controls the proportion of shared parts and private parts in comparable representation learning. K is the dimensionality of comparable representation. To analysis the property of each parameter, we use the same parameter settings that are obtained by cross-validation as introduced in Section V-D, and then vary one at a time while fixing the others to see the performance variations. The number of training data is set to 1000 and the dimensionality of compact representation R is set to 100. The averaged performance on NUS and VOC datasets are reported in Fig. 8.

First, we vary λ in the range of {0, 1e-6, 1e-4,..., 1e4} and the results are shown in Fig. 8 (a) and (e). The performance of BLMV is not good when $\lambda = 0$. Because no consistent constraints are imposed on the shared parts of each view, the views cannot complement with each other well. Then, as λ becomes larger, different views can share the information with each other and BLMV achieves good performance in the range of [1e-4, 1e2] on both datasets. Nevertheless, the performance of BLMV degrades slightly when λ becomes too large ($\lambda = $1e4). This is mainly due to the shared parts are highly consistent with each other under this condition so that the original properties of each view cannot be well preserved. In general, the performance of BLMV is not sensitive to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2607842, IEEE Transactions on Circuits and Systems for Video Technology



(a) Vary λ on NUS (b) Vary β on NUS (c) Vary θ on NUS (d) Vary K on NUS

(e) Vary λ on VOC (f) Vary β on VOC (g) Vary θ on VOC (h) Vary K on VOC

Fig. 8. Parameter sensitivity analysis on NUS and VOC datasets. The averaged performance of BLMV with different values of the four parameters λ, β, θ and K on NUS and VOC datasets are illustrated.

parameter λ in a wide range of [1e-4, 1e2], which can provide satisfied results.

Then, we investigate the influence of parameter β in BLMV. From Fig. 8 (b) and (f), we observe that good image classification and annotation results can be obtained when $\beta \in$ [5e-6, 1e-4] for NUS dataset and $\beta \in$ [5e-5, 5e-4] for VOC dataset. When $\beta = 0$, no structure information is encoded in the low-dimensional representation so that the performance of BLMV is unsatisfied. As β becomes larger, more and more structure information are preserved and the performance is improved accordingly. This variation tendency demonstrates that the nonlinearity inherent in multi-view data is important and considering such intrinsic structure can make the learned low-dimensional representation more compact and discriminative. In addition, when β is set to a large value of 1e-3, the performance is degraded slightly since the learned representation is affected by such excessive regularization. Similar to λ, the overall performance will not be greatly influenced by varying β in a wide range of [5e-6, 5e-4].

Fig. 8 (c) and (g) illustrate how the performance varies with parameter θ. For NUS and VOC dataset, the satisfied performance can be achieved when $\theta \in$ [3/8, 6/8] and [2/8, 6/8], respectively. The shared information cannot be well encoded when $\theta \leq 1/8$ and the private information of each view is not fully captured for $\theta > 7/8$. BLMV cannot generate satisfied performance in both cases. It indicates that exploiting the shared and private nature of multi-view data will help BLMV accurately capturing the information of each view and better utilizing the complementary information, which improves the performance of learning tasks.

Finally, we investigate the performance of BLMV with different K. From Fig. 8 (d) and (h), we observe that good performance is achieved on both NUS and VOC datasets by our method when K is around 400. The performance is limited when $K < 200$ since there are not enough latent factors to fully encode the multi-view information. The appropriate results can be obtained for $K > 200$, where BLMV is not sensitive to the variation of K because the number of latent factors is capable of sufficiently preserving the information of each view.

VI. Conclusion

In this paper, we propose an unsupervised multi-view dimensionality reduction method for image data, which is based on bi-level latent space learning. The first level aims to learn the comparable representation from multiple views with different physical meanings. Both shared and independent parts of each view are explored to accurately represent the information of each view. In the second level, we adopt a robust loss function and preserve the intrinsic manifold structure of images which guarantees the learned representation to be more reliable and discriminative. By introducing the bi-level learning strategy, our method possesses more powerful abilities to leverage the complementary nature of multi-view data and generates promising results. An efficient iterative algorithm is also developed to solve the objective function. Image classification and annotation experiments conducted on two real-world datasets demonstrate the effectiveness of the proposed method. In our future work, we will find a smarter way to determine the parameters in our method. Moreover, we expect to exploit the category information and generalize the method to supervised learning tasks, so that a more discriminative and semantically consistent low-dimensional representation can be obtained.

Acknowledgment

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, in part by National Natural Science Foundation of China: 61303153, 61332016, 61620106009, 61572488 and 61672497, in part by 863 program of China:

2014AA015202, and in part by Bureau of Frontier Sciences and Education (CAS): QYZDJ-SSW-SYS013.

REFERENCES

[1] B. Long, S. Y. Philip, and Z. M. Zhang, "A general model for multiple view unsupervised learning." in *SDM*. SIAM, 2008, pp. 822–833.
[2] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 6, pp. 1438–1446, 2010.
[3] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.
[4] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Advances in Neural Information Processing Systems*, 2010, pp. 982–990.
[5] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang, "Sparse unsupervised dimensionality reduction for multiple view data," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 10, pp. 1485–1496, 2012.
[6] J. Kim, R. Monteiro, and H. Park, "Group sparsity in nonnegative matrix factorization." in *SDM*. SIAM, 2012, pp. 851–862.
[7] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," *arXiv preprint arXiv:0909.1440*, 2009.
[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.
[9] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 5, pp. 733–746, 2009.
[10] H. Wang, F. Nie, H. Huang, and C. Ding, "Heterogeneous visual features fusion via sparse multimodal machine," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3097–3102.
[11] Y. Han, Y. Yang, F. Wu, and R. Hong, "Compact and discriminative descriptor inference using multi-cues," *Image Processing, IEEE Transactions on*, vol. 24, no. 12, pp. 5114–5126, 2015.
[12] Z. Xue, G. Li, and Q. Huang, "Joint multi-view representation learning and image tagging," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1366–1372.
[13] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 352–360.
[14] Z. Xue, G. Li, S. Wang, C. Zhang, W. Zhang, and Q. Huang, "Gomes: A group-aware multi-view fusion approach towards real-world image clustering," in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.
[15] L. Chu, Y. Zhang, G. Li, S. Wang, W. Zhang, and Q. Huang, "Effective multimodality fusion framework for cross-media topic detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 556–569, 2016.
[16] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.
[17] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 129–136.
[18] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1977–1984.
[19] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Affinity aggregation for spectral clustering," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 773–780.
[20] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1737–1744.
[21] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, and Y. Y. Tang, "Group sparse multiview patch alignment framework with view consistency for image classification," *Image Processing, IEEE Transactions on*, vol. 23, no. 7, pp. 3126–3137, 2014.
[22] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.
[23] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 393–400.
[24] Y. Y. Lin, T. L. Liu, and C. S. Fuh, "Multiple kernel learning for dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 6, pp. 1147–1160, 2011.
[25] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *Neural Networks and Learning Systems, IEEE Transactions on*, 2014.
[26] C. Zhang, C. Liang, L. Li, J. Liu, Q. Huang, and Q. Tian, "Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–10, 2016.
[27] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268–2269, 2000.
[28] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[29] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16. MIT, 2004, p. 153.
[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[31] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *Image Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 1921–1932, 2010.
[32] F. Nie, S. Xiang, Y. Song, and C. Zhang, "Orthogonal locality minimizing globality maximizing projections for feature extraction," *Optical Engineering*, vol. 48, no. 1, pp. 017 202–017 202, 2009.
[33] Y. Liu, B. Liao, and Y. Han, "Discriminative multi-view feature selection and fusion," in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.
[34] C. Zhang, Q. Huang, and Q. Tian, "Contextual exemplar classifier based image representation for classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2016.
[35] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering." in *NIPS*, vol. 14, 2001, pp. 585–591.
[36] F. Nie, C. Ding, D. Luo, and H. Huang, "Improved minmax cut graph clustering with nonnegative relaxation," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 451–466.
[37] L. Chu, S. Wang, S. Liu, Q. Huang, and J. Pei, "Alid: scalable dominant cluster detection," *Proceedings of the VLDB Endowment*, vol. 8, no. 8, pp. 826–837, 2015.
[38] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1548–1560, 2011.
[39] Z. Zhang and K. Zhao, "Low-rank matrix approximation with manifold regularization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 7, pp. 1717–1729, 2013.
[40] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding," *Pattern Recognition*, 2015.
[41] Q. Gu, C. Ding, and J. Han, "On trivial solution and scale transfer problems in graph regularized nmf," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 1288.
[42] M. Sun and H. V. hamme, "Large scale graph regularized non-negative matrix factorization with l1 normalization based on kullback–leibler divergence," *Signal Processing, IEEE Transactions on*, vol. 60, no. 7, pp. 3876–3880, 2012.
[43] L. Du and Y.-D. Shen, "Towards robust co-clustering." in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2013.
[44] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 3, p. 11, 2014.
[45] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l21-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.
[46] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 673–682.
[47] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l21-norm regularized discriminative feature selection for unsupervised learning," in *IJCAI*

Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, no. 1. Citeseer, 2011, p. 1589.

[48] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1062–1070.

[49] X. He, M.-Y. Kan, P. Xie, and X. Chen, "Comment-based multi-view clustering of web 2.0 items," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 771–782.

[50] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 359–368.

[51] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, p. 48.

[52] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[53] M. A. Stricker and M. Orengo, "Similarity of color images," in *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*. International Society for Optics and Photonics, 1995, pp. 381–392.

[54] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[56] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[57] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[58] H. Wang, C. Weng, and J. Yuan, "Multi-feature spectral clustering with minimax optimization," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 4106–4113.

[59] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

Zhe Xue received the B.S. degree in Electronic Engineering from Civil Aviation University of China, Tianjin, China, in 2010. He is currently pursuing the Ph.D. degree in University of Chinese Academy of Sciences, Beijing, China. His research interests include machine learning, pattern recognition, computer vision and multimedia data mining.

Guorong Li received her B.S. degree in computer science from Renmin University of China, in 2006 and Ph.D. degree in computer science from the Graduate University of the Chinese Academy of Sciences in 2012. Now, she is an associate professor within the University of Chinese Academy of Sciences. Her research interests include object tracking, pattern recognition, cross-media analysis and multi-label learning.

Shuhui Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006 and the Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2012. He is a researcher with Institute of Computing Technology, Chinese Academy of Sciences, and Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include semantic image analysis, image and video retrieval, and large-scale web multimedia data mining.

Weigang Zhang received the B.S., M.S. and Ph.D. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2003, 2005 and 2016, respectively. He is an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China, and is also a postdoctoral researcher at the University of Chinese Academy of Sciences, CAS, Beijing, China. His research interests include multimedia analysis, video processing and cross-meida computing.

Qingming Huang is currently a professor and deputy dean in the School of Computer and Control Engineering, University of Chinese Academy of Sciences (UCAS). His research interests include multimedia computing, image/video processing, pattern recognition and computer vision. He has published more than 300 academic papers in international journals such as IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on CSVT, and top level international conferences including ACM Multimedia, ICCV, CVPR, VLDB, IJCAI, etc.