Mining Semantically Consistent Patterns for Cross-View Data

Lei Zhang, Yao Zhao, *Senior Member, IEEE*, Zhenfeng Zhu, Shikui Wei, and Xindong Wu, *Fellow, IEEE*

Abstract—In some real world applications, like information retrieval and data classification, we often are confronted with the situation that the same semantic concept can be expressed using different views with similar information. Thus, how to obtain a certain Semantically Consistent Patterns (SCP) for cross-view data, which embeds the complementary information from different views, is of great importance for those applications. However, the heterogeneity among cross-view representations brings a significant challenge on mining the SCP. In this paper, we propose a general framework to discover the SCP for cross-view data. Specifically, aiming at building a feature-isomorphic space among different views, a novel Isomorphic Relevant Redundant Transformation (IRRT) is first proposed. The IRRT linearly maps multiple heterogeneous low-level feature spaces to a high-dimensional redundant feature-isomorphic one, which we name as mid-level space. Thus, much more complementary information from different views can be captured. Furthermore, to mine the semantic consistency among the isomorphic representations in the mid-level space, we propose a new Correlation-based Joint Feature Learning (CJFL) model to extract a unique high-level semantic subspace shared across the feature-isomorphic data. Consequently, the SCP for cross-view data can be obtained. Comprehensive experiments on three data sets demonstrate the advantages of our framework in classification and retrieval.

Index Terms—Cross-view, cross-media, shared subspace learning, heterogeneous data, dimensionality reduction

1 Introduction

With the rapid development of information technology, cross-view data have been widely available in the real world. The so-called cross-view data refer to information items with similar underlying contents, which arrive in different forms, backgrounds or modalities, and so on. For example, in handwritten digit recognition [1], the same digit is written in different forms by different persons; for semantic scene classification [2], different natural scenes (backgrounds) may contain some similar objects; in the case of multimedia retrieval [3], the co-occurring text and image of different modalities, carrying similar semantic information, generally exist in a webpage also known as a Multimedia Document (MMD) [4]. These examples can be well illustrated by three publicly available cross-view data sets, namely, UCI Multiple Features (UCI MFeat) [5], COREL 5K [6], and Wikipedia [7], as shown in Fig. 1.

Naturally, if the representations of different views can be integrated into a certain Semantically Consistent Patterns (SCP) covering the overall complementary information from all views, then the resulting consistent representation will be more favorable for fully exploiting the complementarity among different views.

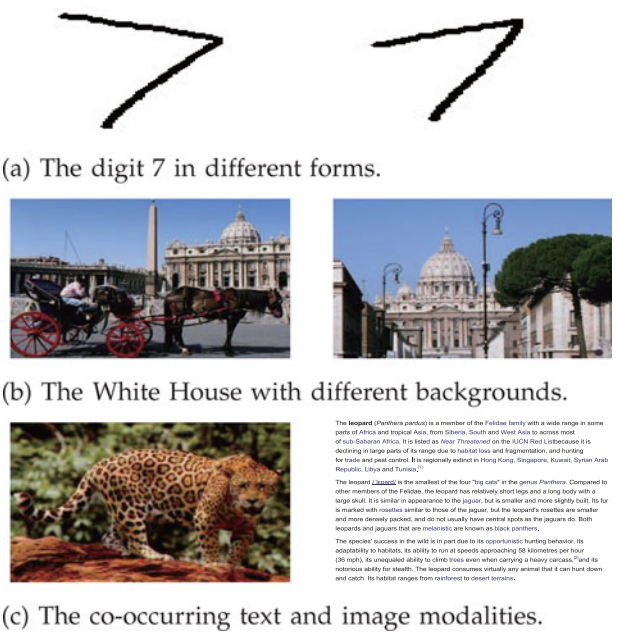
However, it is a challenging task to mine the SCP for cross-view data. First of all, since different views span heterogeneous low-level feature spaces, there is no explicit correspondence among the cross-view representations. For example, the co-occurring image and text in a webpage convey the same semantic concept from the perspectives of vision and writing, respectively, so it is not easy to directly measure the relationship between them based on their own heterogeneous representations. Therefore, to correlate different views, an issue to be first addressed is to build a mid-level feature-isomorphic space, in which the complementary information from different views will be fully embedded.

Meanwhile, for the isomorphic representation in the mid-level space, it can be assumed as illustrated in Fig. 2 that it is mainly composed of requisite, redundant, and noisy components, respectively [8]. The requisite component refers to the complimentary information among isomorphic representations that is requisite for building the Semantically Consistent Patterns with prior knowledge. Unlike the requisite component, the later two refer to non-requisite information. But the difference between them lies in that the redundant component takes high relativity with the requisite component, whereas the noisy one takes no relativity with both the requisite and redundant components. Hence, another issue we need further to deal with for mining the SCP is to extract a unique high-level semantic subspace shared across the feature-isomorphic data. Thus, the above requisite component can be well preserved without the redundant and noisy components being remained.

- *Y. Zhao is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing 100044, China, and the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China. E-mail: yzhao@bjtu.edu.cn.*
- *L. Zhang, Z. Zhu, and S. Wei are with the Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China. E-mail: {10112061, zhfzhu, shkwei}@bjtu.edu.cn.*
- *X. Wu is with the Department of Computer Science, University of Vermont, 33 Colchester Avenue, Burlington, VT 05405. E-mail: xwu@uvm.edu.*

Manuscript received 6 Oct. 2013; revised 17 Mar. 2014; accepted 18 Mar. 2014. Date of publication 25 Mar. 2014; date of current version 26 Sept. 2014. Recommended for acceptance by H. Xiong.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2014.2313866

1041-4347 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.



(a) The digit 7 in different forms.

(b) The White House with different backgrounds.

(c) The co-occurring text and image modalities.

Fig. 1. The cases of cross-view data. Examples from UCI MFeat (top row), COREL 5K (mid row), and Wikipedia (bottom row) data sets.

1.1 Main Contributions

To address the issues mentioned above, the key contributions of this paper are highlighted as follows:

- A general framework for mining the Semantically Consistent Patterns for cross-view data is proposed. In this framework, a mid-level redundant feature-isomorphic space is learned to build a bridge between multiple heterogeneous low-level feature spaces and a unique high-level semantically shared one.
- We propose a novel Isomorphic Relevant Redundant Transformation (IRRT) with low rank constraints, which linearly maps multiple heterogeneous low-level feature spaces to a mid-level redundant feature-isomorphic one, to capture complementary information from different views. To the best of our knowledge, no existing efforts have focused on this type of mapping.
- A new trace ratio based shared subspace learning algorithm, called Correlation-based Joint Feature Learning (CJFL) model, is proposed to extract a unique high-level semantic subspace shared across the feature-isomorphic data. By exploiting the correlation across isomorphic representations, the requisite component is maintained to a large extent while eliminating the redundant and noisy information.
- Extensive experiments on three publicly available cross-view data sets are conducted to demonstrate the effectiveness of the proposed framework.

1.2 Organization

The remainder of this paper is organized as follows: Section 2 gives a broad overview of some related works and defines the notations to be used throughout this paper. We present a general framework for mining the Semantically Consistent Patterns for cross-view data in Section 3.1. In Section 3.2, a novel Isomorphic Relevant Redundant Transformation is developed for correlating different views. We build a new Correlation-based Joint Feature Learning model to mine the semantic consistency among isomorphic representations in Section 3.3. Experimental results and analyses are reported in Section 4. Section 5 concludes this paper.

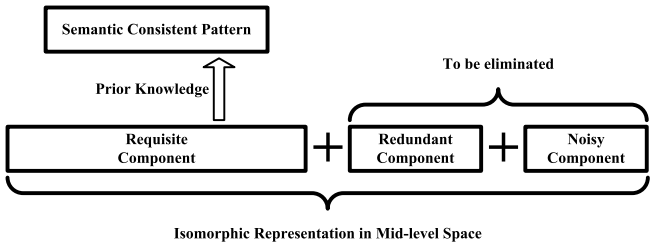


Fig. 2. Components of Isomorphic representation in mid-level space.

2 Related Works and Notations

This section reviews some related works and sets up some notations.

2.1 Related Works

To eliminate the heterogeneity across different views, in the past decades, some classical statistical analysis techniques for modeling correlation between sets of observed variables have been proposed, such as Canonical Correlation Analysis (CCA) [9] and Partial Least Squares (PLS) [10]. They both compute low-dimensional embedding of sets of variables simultaneously. The main difference of them is that CCA maximizes the correlation between variables in the embedded space, while PLS maximizes their covariance.

In particular, CCA is a classical and generally accepted method for feature extraction in the multi-view problem [11]. It has a wide range of applications such as computational biology, financial analysis, and information retrieval, and so on. Furthermore, some state-of-the-art algorithms proposed recently in the multi-view problem have indicated that CCA can be applied to cross-media retrieval [7], [12], clustering [13], and classification [14] of multi-view data. In addition, when one of the views is the predictors induced from the class label, it has been proven that CCA is equivalent to LDA [9].

In [7], a cross-view retrieval method based on CCA has been proposed to obtain the common representations among different views. Hardoon et al. [12] presented a general method using kernel CCA to learn a semantic representation to web images and their associated text. In addition, a multi-view clustering approach via CCA has been proposed by Chaudhuri et al. [13] to project multiple views of the data into a low-dimensional subspace. Furthermore, Sharma et al. [14] presented a general CCA-based multi-view feature extraction approach, which can be used for cross-view classification and retrieval.

Fig. 3 gives an overall illustration of CCA. As shown in Fig. 3, since the features between the heterogeneous presentations from different views are usually complementary, the extracted features from the heterogeneous presentations by CCA are beneficial for the retrieval, clustering, and classification of multi-view data.

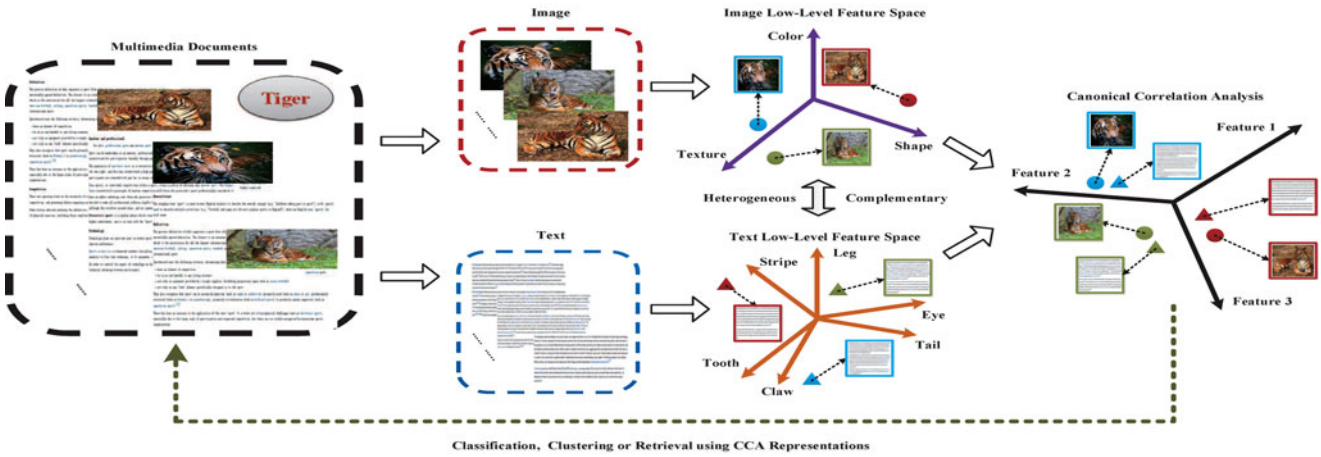


Fig. 3. Canonical correlation analysis.

However, CCA may not extract useful descriptors of data due to its inherent limitation [12]. Kernel CCA offers an alternative solution by implicitly nonlinearly mapping the data into a high-dimensional feature space. Recently, Hardoon et al. [12] proposed a general method using KCCA to learn a semantic representation to web images and their associated text.

In addition, some transformational methods have also been recently proposed for multi-view classification. Kusakunniran et al. [15] presented a View Transformation Model (VTM) from a different point of view using Support Vector Regression (SVR). In [16], Koterba et al. studied a Multi-view Active Appearance Model (MAAM) for fitting and construction. Huang et al. [17] proposed a Vector Boosting (VB) algorithm, called Multi-View Face Detector (MVFD), to divide the entire face space into smaller subspaces.

Meanwhile, although data can be represented by a great deal of features, they are known to be noisy, ambiguous, incomplete, and subjective. These factors can seriously affect the performance of data representations. However, Chen et al. [18] have recently pointed out that the shared structures among different representations generally embody the consistence and coherence of features that co-occur at different representations. Thus, such shared structures tend to characterize beneficially the semantic concept while eliminating noise and redundancy.

Therefore, a number of researchers have recently introduced various shared subspace learning algorithms to mine the semantic consistency among isomorphic representations. The so-called shared subspace learning aims to capture the common features (semantic consistency).

Up to now, the existing shared subspace learning methods involve multi-task learning [19], [20], [21], [22], multi-label classification [23], [24], multi-class classification [25], [26], and matrix factorization [27], [28], [29], [30], which have gained promising performances in some real applications.

In multi-task learning, each task is generally provided with different training samples, but all tasks share the same set of features. Ando and Zhang [19] introduced an Alternating Structure Optimization (ASO) formulation for learning a shared predictive structure from multiple related tasks. Moreover, a framework of Convex Multi-Task Feature Learning (CMTFL) was proposed in [20] for learning a shared feature subspace on which all tasks performed well.

When all tasks share the same set of training data and features, multi-task learning is equivalent to multi-label learning in which each sample can be associated with multiple labels. Specifically, as all the labels share the same input space in multi-label classification, the semantics conveyed by different labels are usually correlated. Thus, Ji et al. [23] proposed a shared-subspace learning framework based on the least squares loss, called Shared-Subspace for Multi-Label Classification (SSMLC), to exploit the correlation information in multi-label learning.

Multi-class learning deals with the learning scenario where each sample is associated with a single label. From this viewpoint, multi-class learning can be seen as a special case of multi-label learning. In multi-class learning, different classes may be built on some underlying common characteristics and thus related to each other. Hence, for extracting a low-dimensional shared subspace in the multi-class problem, a formulation called Shared Structures in Multi-Class Classification (SSMCC) has been proposed by Amit et al. [25], in which a low-rank transformation was computed.

Additionally, different from the above-mentioned supervised mode, an unsupervised Joint Shared Nonnegative Matrix Factorization (JSNMF) method has been recently proposed in [27] to capture the shared base vectors between two data sets with their individual bases corresponding to the discriminant subspace. This approach was formulated under the framework of Nonnegative Matrix Factorization (NMF) [31], which uses an auxiliary source to improve the performance from a primary data set on the basis of nonnegativity constraints on all matrices involved.

2.2 Notations

Here we establish some notations to be used throughout this paper. Assume V_x and V_y are two different views. Let the data matrices $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d_x}$ and $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times d_y}$ be two sets of heterogeneous representations from the V_x and V_y, respectively, where $x_i \in \mathbb{R}^{d_x}$ is the ith sample from V_x, $y_i \in \mathbb{R}^{d_y}$ is the ith sample from V_y, n is the number of training samples, d_x and d_y are the

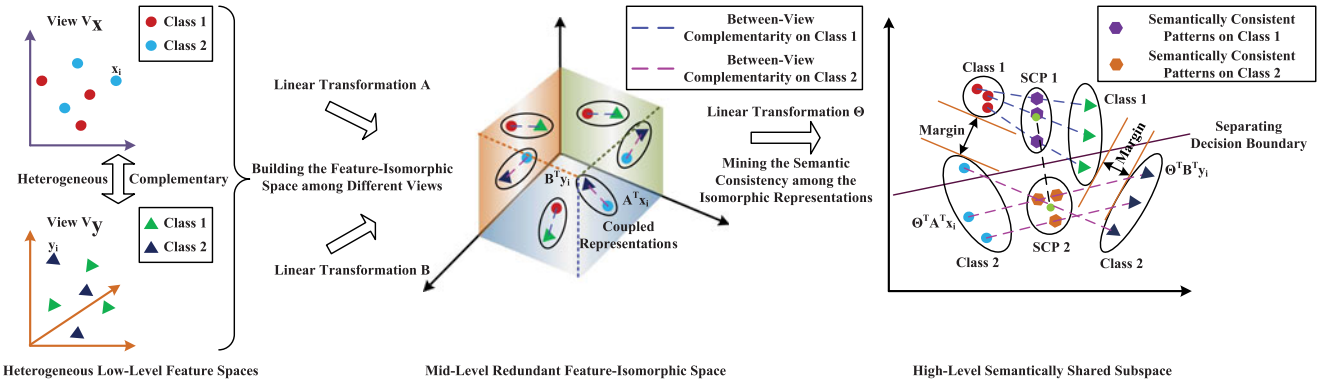


Fig. 4. Framework for mining semantically consistent patterns for cross-view data.

dimensionalities of the heterogeneous low-level feature spaces V_x and V_y. Note that for $i = 1, \ldots, n$, (x_i, y_i) represents the ith couple of heterogeneous representations. We assume that both $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are centered, i.e., $\sum_{i=1}^n x_i = 0$ and $\sum_{i=1}^n y_i = 0$.

We use $\|A\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^q a_{ij}^2}$ to denote the Frobenius norm of a matrix $A = [a_{ij}] \in \mathbb{R}^{p \times q}$, and $\|A\|_* = \sum_{i=1}^r \sigma_i$ is the trace norm of A, where $r = rank(A)$ denotes the rank of A and $\{\sigma_i\}_{i=1}^r$ is the set of singular values of A in a non-increasing order. Let $tr(A) = \sum_{i=1}^p a_{ii}$ be the trace of A. For two matrices A and B, $\langle A, B \rangle = tr(A^T B)$ denotes the matrix inner product. For a vector $b \in \mathbb{R}^p$, let $\|b\|_2 = \sqrt{\sum_{i=1}^p b_i^2}$ be the ℓ_2-norm of b.

Besides, $\nabla f(C)$ denotes the gradient of any smooth function $f(\bullet)$ at the point C; let $|H|$ be the number of elements in the set H; for $w \in \mathbb{R}^p$, we denote by $diag(w)$ the diagonal matrix having the components of the vector w on the diagonal; I is the identity matrix.

3 Discovering Semantically Consistent Patterns for Cross-View Data

Here we propose a general framework to mine the SCP for cross-view data. To facilitate the understanding of our proposed framework, Fig. 4 gives an overall illustration of the proposed framework. More details are presented in the following sections.

3.1 The Proposed Framework

As shown in Fig. 4, a mid-level high-dimensional redundant feature-isomorphic space is learned to build a bridge between multiple heterogeneous low-level feature spaces and a unique high-level semantically shared one in the proposed framework.

Specifically, to fully exploit the complementarity among different views, multiple linear transformations are learned to eliminate the heterogeneity across them. Thus, a mid-level redundant feature-isomorphic space is obtained, in which the correlated representations from different views are coupled together to capture much more complementary information from different views. Accordingly, we can directly measure the correlation among the cross-view data in the mid-level high-dimensional space. For example, the ith co-occurring samples x_i and y_i are projected to the redundant feature-isomorphic space to eliminate the heterogeneity across them.

Furthermore, to mine the semantic consistency among the isomorphic representations, a unique high-level semantic subspace shared across the feature-isomorphic data is extracted with prior knowledge. In the semantically shared subspace, the samples of the same class from the same view can be grouped together while keeping the instances from different categories away from each other simultaneously. Thus, the redundant and noisy information in the mid-level space is eliminated effectively. For instance, the ith coupled representation is mapped into a semantically shared subspace for keeping their complementary.

With the requisite complementary information from the mid-level space, the resulting SCP as displayed in Fig. 4 will be more likely to be linearly separable, compared sharply with any single view.

3.2 Building a Feature-Isomorphic Space among Different Views

A novel Isomorphic Relevant Redundant Transformation is developed for correlating different views and an efficient algorithm for solving the IRRT model is presented in this section.

3.2.1 The Proposed IRRT Model

As a classical correlation analysis method, CCA [9] can obtain a low-dimensional embedding of the sets of heterogeneous representations from different views in a feature-isomorphic space. It can be formulated as follows:

$$\begin{aligned} \min_{A,B} \quad & \|XA - YB\|_F^2 \\ s.t. \quad & A^T X^T XA = I \quad and \quad B^T Y^T YB = I, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{d_x \times p}$, $B \in \mathbb{R}^{d_y \times p}$, and $p \in \{1, \ldots, min(d_x, d_y)\}$. Then for the ith couple of heterogeneous representations (x_i, y_i), we can obtain their own isomorphic correlated representations with the optimal A^* and B^* by:

$$\mu_{x_i} = A^{*T} x_i \quad and \quad \mu_{y_i} = B^{*T} y_i. \tag{2}$$

Furthermore, we can get an integrated representation μ_i in the feature-isomorphic space based on μ_{x_i} and μ_{y_i}:

$$\mu_i = (\mu_{x_i} + \mu_{y_i})/2. \tag{3}$$

Although based on a good mathematical formulation, the reduced dimensionality p must not be larger than $min(d_x, d_y)$ due to the inherent limitation of CCA. It means that some requisite information may be lost in the dimension-reduced space. In practice, as illustrated in Fig. 2, the underlying complementary information (requisite component) from different views tends to be hidden in a high-dimensional space mingled with the redundant and noisy components [8].

Recently, Liu et al. [32] have pointed out that the rank is a powerful tool to capture some type of underlying information in the matrix case. Nevertheless, "$rank(\bullet)$" is not a convex function, which leads to the difficulty in finding the optimal solution. Fortunately, Candès and Recht [33], Recht et al. [34], and Candès and Tao [35] have theoretically justified that the trace norm of a matrix can be used to approximate the rank of the matrix. Therefore, based on the above-mentioned strong theoretical supports [8], [32], [33], [34], [35], we propose a novel Isomorphic Relevant Redundant Transformation with trace norm constraints to linearly map multiple heterogeneous low-level feature spaces to a high-dimensional redundant feature-isomorphic one. Consequently, the correlated representations from different views are coupled together to capture much more complementary information. We name this redundant feature-isomorphic space as a mid-level space. Hence, we have the following optimization problem:

$$\boldsymbol{\Psi_1}: \begin{array}{ll} \min\limits_{A,B} & \|XA - YB\|_F^2 \\ s.t. & \|XA\|_* \le \varepsilon \quad and \quad \|YB\|_* \le \gamma, \end{array} \tag{4}$$

where ε and γ are pre-specified positive parameters to control the amount of information carried by the transformed data. The motivation of introducing the trace norm constraints in Eq. (4) is to capture much more underlying complementary information from different views in the feature-isomorphic space.

Seemingly, our proposed IRRT model looks like an extension of CCA. However, in fact, IRRT is completely different from CCA in terms of its theoretical basis.

Unlike the reduced dimensionality in CCA, i.e., $p \le min(d_x, d_y)$, our proposed IRRT model can linearly project the cross-view data into a feature-isomorphic space of even higher dimensions. That is to say, p may be greater than both d_x and d_y, i.e., $p \gg max(d_x, d_y)$. It is worth to note that no similar numerical method has been yet proposed and IRRT is greatly different from well-known kernel methods without an explicit high-dimensional projection.

In addition, CCA imposes the orthogonal constraints on the canonical variables so as to project two views of the same set of objects onto a lower-dimensional space in which they are maximally correlated. Different from the orthogonal constraints in CCA, the low rank constraints are enforced in IRRT on the transformed data with the aim of linearly mapping multi-view data to a high-dimensional redundant feature-isomorphic space.

Finally, as a classical statistical analysis technology, CCA is generally converted into an eigenvalue problem in which the close-form solutions can be obtained. However, we will prove below that the proposed IRRT method can be converted to a relaxed convex optimization problem for which there is an iterative algorithm which converges to an optimal solution.

Thus, in essence, our IRRT method and CCA do not have much in common. It is not a simple extension of CCA.

Note that to solve the problem Ψ_1 in Eq. (4) directly is not a trivial task for two main reasons. First, it is a non-convex problem, although it is separately convex with respect to each variable A or B. Second, the trace norm constraints are not smooth, which makes it even more difficult to find the optimum. However, Lemma 1 shows that the trace norm constraints on the transformed data in Ψ_1 can be relaxedly converted into the trace norm constraints on the projection matrices in Ψ_2.

Lemma 1. *For a positive number δ and any two conformable matrices C and D, if*

$$\|C\|_*\|D\|_* \le \delta,$$

then

$$\|CD\|_* \le \delta.$$

Proof. As the trace norm is a matrix norm, it satisfies the compatibility for any two conformable matrices [36, Chapter 5.2, page 280]. So we can get

$$\|CD\|_* \le \|C\|_*\|D\|_*.$$

Thus if $\|C\|_*\|D\|_* \le \delta$, then $\|CD\|_* \le \delta$. This completes the proof of the lemma. □

According to Lemma 1, if the pre-specified positive parameters ε and γ in Ψ_1 satisfy

$$\|X\|_*\|A\|_* \le \varepsilon \quad and \quad \|Y\|_*\|B\|_* \le \gamma, \tag{5}$$

then we can obviously obtain $\|XA\|_* \le \varepsilon$ and $\|YB\|_* \le \gamma$. Thus the trace norm constraints in Ψ_1 can be converted into

$$\|A\|_* \le \varepsilon/\|X\|_* \quad and \quad \|B\|_* \le \gamma/\|Y\|_*. \tag{6}$$

Consequently, with the relaxed constraints in Eq. (6), the formulation Ψ_1 can be reformulated as follows:

$$\boldsymbol{\Psi_2}: \begin{array}{ll} \min\limits_{A,B} & \|XA - YB\|_F^2 \\ s.t. & \|A\|_* \le \varepsilon/\|X\|_* \; and \; \|B\|_* \le \gamma/\|Y\|_*. \end{array} \tag{7}$$

3.2.2 An Efficient Solver for Ψ_2

For notational simplicity, we denote the optimization problem Ψ_2 by:

$$\min_{Z \in \mathcal{C}} \quad f(Z), \tag{8}$$

where $f(\bullet) = \| \bullet \|_F^2$ is a smooth objective function, $Z = [A_Z \quad B_Z]$ symbolically represents the optimization variables, and $\mathcal{C}$ is the closed and convex domain set defined by:

$$\mathcal{C} = \{Z | \|A_Z\|_* \le \varepsilon/\|X\|_*, \|B_Z\|_* \le \gamma/\|Y\|_*\}. \tag{9}$$

As $f(\bullet)$ is continuously differentiable with Lipschitz continuous gradient L [37]:

$$\|\triangledown f(Z_x) - \triangledown f(Z_y)\|_F \le L\|Z_x - Z_y\|_F, \forall Z_x, Z_y \in \mathcal{C}, \tag{10}$$

it is appropriate to adopt the Accelerated Projected Gradient (APG) [37] method to solve the problem in Eq. (8). APG has been successfully applied in the following minimization problem:

$$\min_{z\in\mathcal{G}} \quad g(z), \tag{11}$$

where $g(\bullet)$ is a smooth objective function, z is the optimization variable, and $\mathcal{G}$ is the feasible domain of the optimization problem.

Note that, in the APG algorithm, the euclidean projection of a given point s onto the convex set $\mathcal{G}=\{z|\|z\|_* \le m\}$ can be defined by:

$$proj_{\mathcal{G}}(s) = arg \quad \min_{z\in\mathcal{G}} \|z-s\|_F^2/2, \tag{12}$$

where m is a pre-specified positive constant. Then we can use the algorithm proposed in [38] to solve Eq. (12). The details are given in Algorithm 1.

Algorithm 1: Efficient Projection on Trace Norm Constraints (**EPTNC**) [38]

Input: s, $\mathcal{G}$.
Output: z^*.
1: Compute singular value decomposition of s as $s=U_s\Sigma_s V_s^T$.
2: Set $\rho=|H|$, $H=\{i\in 1,\cdots,n|\sigma_i>0,\sigma_i$ is the i-th singular value of $s\}$.
3: Define $\theta=\left(\sum_{i=1}^{\rho}\sigma_i - m\right)/\rho$.
4: Set $\tau_i = max\{\sigma_i-\theta,0\}$.
5: Define $\Sigma_{z^*}=diag(\sigma_1,\cdots,\sigma_\rho,0)$.
6: Compute $z^*=U_s\Sigma_{z^*}V_s^T$.

Algorithm 2: Isomorphic Relevant Redundant Transformation (**IRRT**)

Input: $f(\bullet)$, $Z_0=[A_{Z_0}\ B_{Z_0}]$, γ_1, $\mathcal{C}$, $t_0=1$, and $max\text{–}iter$.
Output: Z^*.
1: Define $f_{\gamma,S}(Z)=f(S)+\langle\nabla f(S),Z-S\rangle+\gamma\|Z-S\|_F^2/2$.
2: Set $A_{Z_1}=A_{Z_0}$ and $B_{Z_1}=B_{Z_0}$.
3: for i =1,2,$\cdots$,$max\text{–}iter$ do
4: Set $a_i=(t_{i-1}-1)/t_{i-1}$.
5: Compute $A_{S_i}=(1+\alpha_i)A_{Z_i}-\alpha_i A_{Z_{i-1}}$.
6: Compute $B_{S_i}=(1+\alpha_i)B_{Z_i}-\alpha_i B_{Z_{i-1}}$.
7: Set $S_i=[A_{S_i}\quad B_{S_i}]$.
8: Compute $\nabla_{A_S}f(A_{S_i})$ and $\nabla_{B_S}f(B_{S_i})$.
9: while (true)
10: Compute $\widehat{A_S}=A_{S_i}-\nabla_{A_S}f(A_{S_i})/\gamma_i$.
11: Compute $\widehat{B_S}=B_{S_i}-\nabla_{B_S}f(B_{S_i})/\gamma_i$.
12: Compute $[A_{Z_{i+1}}]=\boldsymbol{EPTNC}(\widehat{A_S},\mathcal{C})$.
13: Compute $[B_{Z_{i+1}}]=\boldsymbol{EPTNC}(\widehat{B_S},\mathcal{C})$.
14: Set $Z_{i+1}=[A_{Z_{i+1}}\quad B_{Z_{i+1}}]$.
15: if $f(Z_{i+1})\le f_{\gamma_i,S_i}(Z_{i+1})$, then break;
16: else Update $\gamma_i=\gamma_i\times 2$.
17: end-if
18: end-while
19: Update $t_i=\left(1+\sqrt{1+4t_{i-1}^2}\right)/2$ and $\gamma_{i+1}=\gamma_i$.
20: end-for
21: Set $Z^*=Z_{i+1}$.

When applying the APG method for solving the problem in Eq. (8), the euclidean projection $Z=[A_Z\quad B_Z]$ of a given point $S=[A_S\quad B_S]$ onto the set $\mathcal{C}$ is defined by:

$$proj_{\mathcal{C}}(S) = arg \quad \min_{Z\in\mathcal{C}} \|Z-S\|_F^2/2. \tag{13}$$

By combining APG and Algorithm 1, we can solve the problem in Eq. (8). The details are given in Algorithm 2.

3.3 Mining the Semantic Consistency among Isomorphic Representations

This section presents a new shared subspace learning algorithm, called Correlation-based Joint Feature Learning model, to mine the semantic consistency among isomorphic representations, and shows how to solve the CJFL model.

3.3.1 The Proposed CJFL Model

In Section 3.2, we have built a mid-level high-dimensional redundant feature-isomorphic space for correlating different views, in which the embedded requisite component tends to be exact, clear, complete, and objective. However, as shown in Fig. 2, some redundant and noisy components inevitably co-exist with the requisite one in the space. These factors can seriously affect the performance of the mid-level data representations.

Therefore, it is essential to extract a unique high-level low-dimensional semantic subspace shared across the feature-isomorphic data to eliminate both the redundant and noisy information in the mid-level high-dimensional redundant space.

Recently, some trace ratio algorithms such as Linear Discriminant Analysis (LDA) [39], Semantic Subspace Projection (SSP) [40], and Trace Ratio Optimization Problem (TROP) [41] have been proven to be effective in redundancy and noise reduction. For the purpose of finding a projection matrix W to simultaneously minimize the within-class distance while maximizing the between-class distance, a trace ratio optimization problem is formulated as follows [41]:

$$\max_{W^TW=I} \quad \frac{tr(W^THW)}{tr(W^TGW)}, \tag{14}$$

where H and G denote the between-class and within-class scatter matrices, respectively. However, since these methods were originally developed for handling single view problems, they do not fully take into account the correlation across isomorphic representations.

Thereby, unlike some previous supervised shared subspace learning methods based on least squares [19], [20], [23], [25] and matrix factorization [27] techniques, we propose a new trace ratio based shared subspace learning algorithm, called Correlation-based Joint Feature Learning model. By exploiting the correlations across isomorphic representations, CJFL could extract a unique high-level semantically shared subspace. In this subspace, the requisite component will be maintained to a large extent without the redundant and noisy information being remained. Correspondingly, the SCP for cross-view data can be obtained.

Specifically, let (A^*,B^*) be the optimal solutions of the problem Ψ_2. Then we have the sets of isomorphic relevant redundant representations $J=\{a_i=A^{*T}x_i\}_{i=1}^n$ and $R=\{b_i=B^{*T}y_i\}_{i=1}^n$. Let $\mathcal{C}_X^t$ and $\mathcal{C}_Y^t$ be the sample sets of tth class

from J and R, respectively. We define

$$\mathcal{S}_X^t = \{(a_i, a_j) \,|\, a_i, a_j \in C_X^t, i \neq j\}, \tag{15}$$

$$\mathcal{S}_Y^t = \{(b_i, b_j) \,|\, b_i, b_j \in C_Y^t, i \neq j\}, \tag{16}$$

$$\mathcal{D}_X^{tk} = \{(a_i, a_j) \,|\, a_i \in C_X^t \wedge a_j \in C_X^k, i \neq j, t \neq k\}, \tag{17}$$

$$\mathcal{D}_Y^{tk} = \{(b_i, b_j) \,|\, b_i \in C_Y^t \wedge b_j \in C_Y^k, i \neq j, t \neq k\}. \tag{18}$$

Let

$$\mathcal{S}_X = \bigcup\nolimits_t \mathcal{S}_X^t \quad and \quad \mathcal{S}_Y = \bigcup\nolimits_t \mathcal{S}_Y^t, \tag{19}$$

$$\mathcal{D}_X = \bigcup\nolimits_t \bigcup\nolimits_k \mathcal{D}_X^{tk} \quad and \quad \mathcal{D}_Y = \bigcup\nolimits_t \bigcup\nolimits_k \mathcal{D}_Y^{tk}. \tag{20}$$

Obviously, each pair of data from $\mathcal{S}_X$ or $\mathcal{S}_Y$ is semantically similar to each other and the one from $\mathcal{D}_X$ or $\mathcal{D}_Y$ is semantically dissimilar to each other.

To eliminate the redundant and noisy information in the mid-level high-dimensional space, we need to learn a linear transformation $\Theta \in \mathbb{R}^{p\times k}$ with prior knowledge (class information in our case) to parameterize the semantically shared subspace, where k is the dimensionality of the subspace. Mathematically, we would like to minimize the within-class distance as follows:

$$\begin{aligned}
&\sum_{\forall(a_i,a_j)\in\mathcal{S}_X} \left(\Theta^T a_i - \Theta^T a_j\right)^T \left(\Theta^T a_i - \Theta^T a_j\right) \\
&\quad + \sum_{\forall(b_i,b_j)\in\mathcal{S}_Y} \left(\Theta^T b_i - \Theta^T b_j\right)^T \left(\Theta^T b_i - \Theta^T b_j\right) \\
&= \sum_{\forall(a_i,a_j)\in\mathcal{S}_X} tr\left(\Theta^T (a_i - a_j)(a_i - a_j)^T \Theta\right) \\
&\quad + \sum_{\forall(b_i,b_j)\in\mathcal{S}_Y} tr\left(\Theta^T (b_i - b_j)(b_i - b_j)^T \Theta\right) \\
&= tr(\Theta^T J_{\mathcal{S}} \Theta) + tr(\Theta^T R_{\mathcal{S}} \Theta) \\
&= tr(\Theta^T (J_{\mathcal{S}} + R_{\mathcal{S}}) \Theta),
\end{aligned} \tag{21}$$

where

$$J_{\mathcal{S}} = \sum_{\forall(a_i,a_j)\in\mathcal{S}_X} (a_i - a_j)(a_i - a_j)^T, \tag{22}$$

$$R_{\mathcal{S}} = \sum_{\forall(b_i,b_j)\in\mathcal{S}_Y} (b_i - b_j)(b_i - b_j)^T, \tag{23}$$

and $J_{\mathcal{S}} + R_{\mathcal{S}}$ is a joint within-class scatter matrix from both J and R. Meanwhile, we also expect to maximize the between-class distance as follows:

$$\begin{aligned}
&\sum_{\forall(a_i,a_j)\in\mathcal{D}_X} \left(\Theta^T a_i - \Theta^T a_j\right)^T \left(\Theta^T a_i - \Theta^T a_j\right) \\
&\quad + \sum_{\forall(b_i,b_j)\in\mathcal{D}_Y} \left(\Theta^T b_i - \Theta^T b_j\right)^T \left(\Theta^T b_i - \Theta^T b_j\right) \\
&= \sum_{\forall(a_i,a_j)\in\mathcal{D}_X} tr\left(\Theta^T (a_i - a_j)(a_i - a_j)^T \Theta\right) \\
&\quad + \sum_{\forall(b_i,b_j)\in\mathcal{D}_Y} tr\left(\Theta^T (b_i - b_j)(b_i - b_j)^T \Theta\right) \\
&= tr(\Theta^T J_{\mathcal{D}} \Theta) + tr(\Theta^T R_{\mathcal{D}} \Theta) \\
&= tr(\Theta^T (J_{\mathcal{D}} + R_{\mathcal{D}}) \Theta),
\end{aligned} \tag{24}$$

where

$$J_{\mathcal{D}} = \sum_{\forall(a_i,a_j)\in\mathcal{D}_X} (a_i - a_j)(a_i - a_j)^T, \tag{25}$$

$$R_{\mathcal{D}} = \sum_{\forall(b_i,b_j)\in\mathcal{D}_Y} (b_i - b_j)(b_i - b_j)^T, \tag{26}$$

and $J_{\mathcal{D}} + R_{\mathcal{D}}$ is a joint between-class scatter matrix from both J and R. To simultaneously minimize the within-class distance while maximizing the between-class distance, it is straightforward to formulate the above problem as a trace ratio optimization problem:

$$\mathbf{\Omega_1}: \min_{\Theta^T\Theta=I} \frac{tr(\Theta^T (J_{\mathcal{D}} + R_{\mathcal{D}})\Theta)}{tr(\Theta^T (J_{\mathcal{S}} + R_{\mathcal{S}})\Theta)}, \tag{27}$$

where the orthogonal constraint for Θ is used to eliminate the redundant information in the mid-level space, which takes high relativity with the requisite component. Unlike the scatter matrices in LAD [39], SSP [40], and TROP [41], both the joint within-class and between-class scatter matrices $J_{\mathcal{S}} + R_{\mathcal{S}}$ and $J_{\mathcal{D}} + R_{\mathcal{D}}$ make a full use of the identity of sample distributions from different views in the mid-level feature-isomorphic space.

On the other hand, the complementarity across isomorphic representations should be well preserved. Thus, we can redefine the formulation Ω_1 by:

$$\mathbf{\Omega_2}: \min_{\Theta^T\Theta=I} \frac{tr(\Theta^T (J_{\mathcal{D}} + R_{\mathcal{D}})\Theta)}{tr(\Theta^T (J_{\mathcal{S}} + R_{\mathcal{S}})\Theta) + \alpha\|J\Theta - R\Theta\|_F^2 + \beta\|\Theta\|_F^2}, \tag{28}$$

where the term $\|J\Theta - R\Theta\|_F^2$ denotes the correlation-based residual to avoid violating the intrinsic structure of the coupled representations, the regularization term $\|\Theta\|_F^2$ controls the complexity of the model, and α, $\beta > 0$ are the regularization parameters.

3.3.2 An Efficient Solver for Ω_2

The optimal Θ^* for the problem in Eq. (28) can be obtained by maximizing the following trace difference problem:

$$\begin{aligned}
\Theta^* &= arg \max_{\Theta^T\Theta=I} tr(\Theta^T (J_{\mathcal{D}} + R_{\mathcal{D}})\Theta) - \eta_t tr(\Theta^T (J_{\mathcal{S}} + R_{\mathcal{S}})\Theta) \\
&\quad - \eta_t \alpha tr((J\Theta - R\Theta)^T (J\Theta - R\Theta)) - \eta_t \beta tr(\Theta^T\Theta) \\
&= arg \max_{\Theta^T\Theta=I} tr(\Theta^T (J_{\mathcal{D}} + R_{\mathcal{D}} - \eta_t (J_{\mathcal{S}} + R_{\mathcal{S}}))\Theta) \\
&\quad - \eta_t tr(\alpha(\Theta^T J^T J\Theta - 2\Theta^T J^T R\Theta + \Theta^T R^T R\Theta) + \beta\Theta^T\Theta) \\
&= arg \max_{\Theta^T\Theta=I} tr(\Theta^T (J_{\mathcal{D}} + R_{\mathcal{D}} - \eta_t (J_{\mathcal{S}} + R_{\mathcal{S}} \\
&\quad + \alpha(J^T J - 2J^T R + R^T R) + \beta I))\Theta),
\end{aligned} \tag{29}$$

where η_t [see Eq. (32)] is the trace ratio value of the tth iteration. Hence, Θ^* is composed of the eigenvectors corresponding to the k largest eigenvalues of the matrix $J_{\mathcal{D}} + R_{\mathcal{D}} - \eta_t(J_{\mathcal{S}} + R_{\mathcal{S}} + \alpha(J^T J - 2J^T R + R^T R) + \beta I)$. We can use the iterative algorithm proposed in [41] to solve the problem in Eq. (29). The details are given in Algorithm 3.

Algorithm 3: Correlation-based Joint Feature Learning (**CJFL**)

Input: an arbitrary columnly orthogonal matrix Θ_0, the matrices J, R, $J_{\mathcal{D}}$, $R_{\mathcal{D}}$, $J_{\mathcal{S}}$, and $R_{\mathcal{S}}$, a positive integer h, two positive numbers α and β, and $max\text{–}iter$.

Output: Θ^*.

1: for t =0,1,2,$\cdots$,$max\text{–}iter$ do

2: Compute

$$\eta_t = \frac{tr(\Theta_t^T(J_{\mathcal{D}}+R_{\mathcal{D}})\Theta_t)}{tr(\Theta_t^T(J_{\mathcal{S}}+R_{\mathcal{S}})\Theta_t)+\alpha\|J\Theta_t-R\Theta_t\|_F^2+\beta\|\Theta_t\|_F^2}. \quad (30)$$

4: Perform eigen-decomposition of the matrix $J_{\mathcal{D}}+R_{\mathcal{D}}-\eta_t(J_{\mathcal{S}}+R_{\mathcal{S}}+\alpha(J^TJ-2J^TR+R^TR)+\beta I)$ as $P\Lambda P^T$.

5: Θ_{t+1} is given by the column vectors of the matrix P corresponding to the h largest eigenvalue.

6: end-for

7: Set $\Theta^* = \Theta_{t+1}$.

3.3.3 Semantically Consistent Patterns

Let (A^*, B^*) be the optimal solution of the problem Ψ_2 and Θ^* be the optimal one of the problem Ω_2. Then, for the ith couple of heterogeneous representations (x_i, y_i), we can obtain their own isomorphic relevant representations with the optimal A^*, B^*, and Θ^* as follows:

$$\tau_{x_i} = \Theta^{*T}A^{*T}x_i \quad and \quad \tau_{y_i} = \Theta^{*T}B^{*T}y_i. \quad (31)$$

In addition, we can exploit the consistent representation τ_i of different views, i.e., the Semantically Consistent Patterns for the cross-view data in the semantically shared subspace based on τ_{x_i} and τ_{y_i}:

$$\tau_i = (\tau_{x_i} + \tau_{y_i})/2. \quad (32)$$

4 Experimental Results and Analyses

In this section, we evaluate and analyze the effectiveness of the learned SCP by the proposed framework for cross-view data.

4.1 Data Sets

Our experiments are conducted on three publicly available cross-view data sets, namely, UCI Multiple Features (UCI MFeat) [5], COREL 5K [6], and Wikipedia [7]. The statistics of the data sets are given in Table 1.

• UCI MFeat data set

It consists of features of handwritten numerals ('0'-'9'), which are represented in terms of six different feature sets. Each number represents a class in which there are 200 different written forms corresponding to the same original character. The fou and zer feature sets were randomly picked up to test the performance of our proposed models.

• Corel 5K data set

It contains 260 categories of images of various contents ranging from animals to vehicles, which are represented in terms of 15 different feature sets. Each category includes a number of pictures under different natural scenarios corresponding to the same semantic class. Similarly, we selected the DenseHue and HarrisHue feature sets at random. DenseHue and HarrisHue are two visual features provided by Guillaumin et al. [42]. They use local SIFT features [43] and local hue histograms [44]. DenseHue is computed on a dense multiscale grid, and HarrisHue is computed on regions found with a Harris interest-point detector.

• Wikipedia data set

It is composed of 2,866 MMDs [4] selected from the Wikipedia's featured article collection, in which each MMD includes a single image and at least 70 words of text corresponding to the same semantic concept. Instead of directly using an original representation as in [7], the TF-IDF encoding method is used to form the text representation.

Brief descriptions of the chosen feature sets in the above-mentioned data sets are listed in Table 2.

4.2 Experimental Setup

Note that all the data are normalized to unit length. Each data set is randomly separated into a training set and a test set. The training samples account for 80 percent of each original data set, and the remaining ones act as the test data. Such a partition of each data set is repeated five times and the average performance is reported.

In real-world applications such as classification and retrieval, the parameters in the proposed framework can be alternately set by five-fold cross-validation based on the AUC and MAP, respectively. Specifically, for IRRT, the nonnegative constraint parameters ε and γ are first set to certain pre-specified values. Then the dimensionality p of the feature-isomorphic relevant redundant space is tuned on the pre-specified candidate set. When an appropriate dimensionality is determined, it is necessary to select ε and γ on the pre-specified candidate set with the fixed dimensionality. Similarly, for the dimensionality k of the shared subspace and the regularization parameters α and β in CJFL, we can set them in the same way as with IRRT.

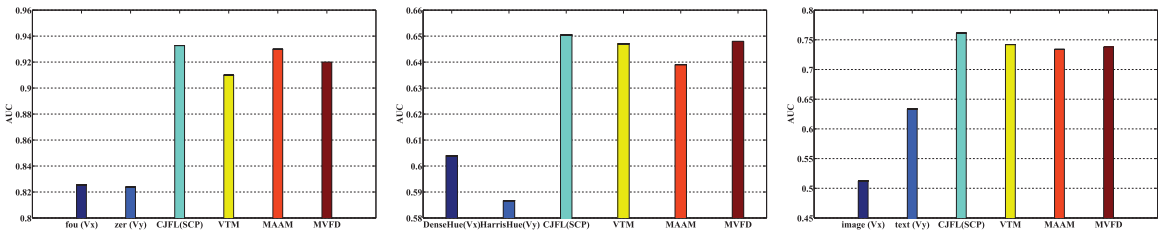
Particularly, the k-nearest ($k = 5$) neighbor classifier serves as the benchmark for the tasks of classification. In the case of retrieval, we use the euclidean distance and the Mean Average Precision (MAP) score to evaluate the retrieval performance.

TABLE 1
Statistics of the Cross-View Data Sets

Dataset	Total Attributes	Total Classes	Total Samples
UCI MFeat	649	10	2000
COREL 5K	37152	260	4999
Wikipedia	258	10	2866

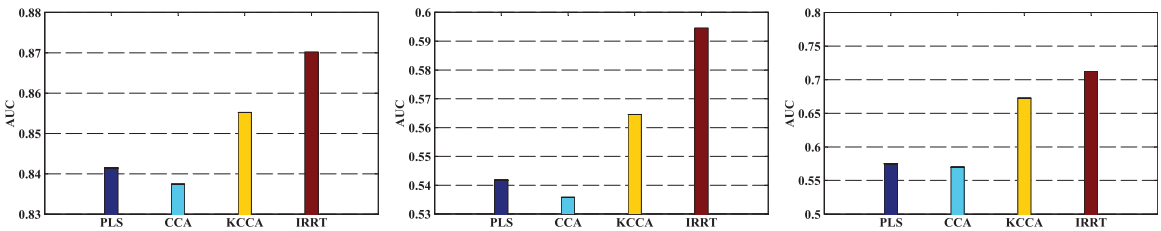
TABLE 2
Brief Descriptions of the Feature Sets

Dataset	Feature Set	Total Attributes	Total Labels	Total Instances
UCI MFeat	fou (V_x)	76	10	2000
	zer (V_y)	47	10	2000
Corel 5K	DenseHue (V_x)	100	260	4999
	HarrisHue (V_y)	100	260	4999
Wikipedia	image (V_x)	128	10	2866
	text (V_y)	130	10	2866



(a) UCI MFeat (b) COREL 5K (c) Wikipedia

Fig. 5. Comparisons of classification performance of single and integrated views.



(a) UCI MFeat (b) COREL 5K (c) Wikipedia

Fig. 6. Comparisons of classification performance of PLS, CCA, KCCA, and IRRT.

4.3 Comparison of Single and Integrated Views

To show the advantages of an integrated view over single views, the classification performances of them are illustrated in Fig. 5.

For the proposed IRRT model, we tune the dimensionality p of the feature-isomorphic relevant redundant space on the set $\{2^i \times 100 | i = 1, 2, 3, 4, 5\}$ and the nonnegative constraint parameters ε and γ on the sets $\{10^i | i = -4, -3, -2, -1, 0, 1, 2, 3, 4\}$. For the proposed CJFL model, the dimensionality k of the shared subspace is selected from the set $\{i \times 10 | i = 1, 2, 3, \ldots, 30\}$ and the regularization parameters α and β on the candidate sets $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The parameter h in Step 5 of Algorithm 3 is the number of positive singular values of the matrix P in Step 4 of Algorithm 3 and $\Theta \in \mathbb{R}^{h \times k}$. The parameter settings in VTM [15], MAAM [16], and MVFD [17] are the same as in their original references.

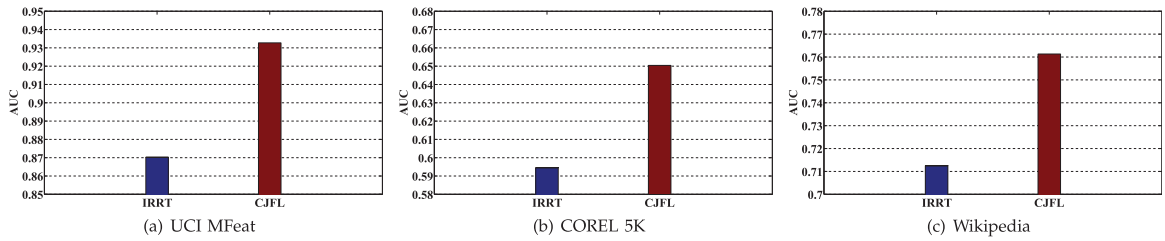
Clearly, it can be observed from Fig. 5 that the SCP τ as given in Eq. (32) greatly outperforms the original expressions of either single view. This observation confirms our previous hypothesis that the SCP will be more favorable for fully exploiting the complementarity among different views. In addition, because of implementing low-rank constraints and eliminating the redundant and noisy information, our proposed framework is not worse than other state-of-the-art multi-view classification methods in classification performance.

4.4 Evaluation on PLS, CCA, KCCA, and IRRT

To evaluate the potentiality of capturing complementary information of the proposed IRRT model, we further make a comparison between IRRT and the three classical correlation analysis methods CCA [9], PLS [10], and KCCA [12]. Here, the dimensionality p of the feature-isomorphic space is specified by $min(d_x, d_y)$ for both PLS and CCA. For KCCA, Gaussian kernel is used and p is identical to the dimensionality in IRRT.

As mentioned above, due to its inherent limitations, PLS and CCA can only project the cross-view data into a low-dimensional space according to Eq. (3). However, as illustrated in Fig. 2, the underlying complementary information from different views tends to be hidden in a high-dimensional space. In addition, we can see from Fig. 6 that it is very difficult for KCCA to capture much more complementary information without low-rank constraints as in IRRT, although both KCCA and IRRT can map the cross-view data into a high-dimensional space.

Just to pursue such a purpose, the proposed IRRT model linearly maps multiple heterogeneous low-level feature spaces to a high-dimensional redundant feature-isomorphic one with low-rank constraints. As shown in Fig. 6, the superiority of IRRT over CCA, PLS, and KCCA in the classification performance is quite remarkable. For example, nearly 14 percent gain is achieved for the Wikipedia data set. It



(a) UCI MFeat (b) COREL 5K (c) Wikipedia

Fig. 7. Comparisons of classification performance of IRRT and CJFL.

means that IRRT can capture much more complementary information than CCA, PLS, and KCCA.

4.5 Analysis of IRRT and CJFL

The propose of comparing the proposed IRRT and CJFL models is twofold. One is to confirm the above-mentioned assumption that the data representation in the mid-level feature-isomorphic space is composed all of requisite, redundant, and noisy components; the other is to verify the competence of CJFL for eliminating the redundant and noisy information.

Based on IRRT, we can build a mid-level high-dimensional redundant feature-isomorphic space by correlating heterogeneous low-level representations from different views. However, as shown in Fig. 2, some redundant and noisy components inevitably co-exist with the requisite component in the mid-level high-dimensional space. In order to eliminate the redundant and noisy information in the mid-level space, according to Eq. (32), CJFL will extract a unique high-level low-dimensional semantic subspace shared across the feature-isomorphic data obtained by IRRT.

From Fig. 7, we can see that CJFL achieves much better classification performance than IRRT without the involvement of semantic information. This result is consistent with our previous assumption that some redundant and noisy components are unavoidably contained in the mid-level high-dimensional space. It also indicates that CJFL indeed has a powerful capacity of eliminating the redundant and noisy information.

4.6 Comparison of Trace Ratio Algorithms

In essence, like LDA [39], SSP [40], and TROP [41], the proposed CJFL model is also a dimensionality reduction method based on the trace ratio. But the explicit difference of CJFL from the former models lies in that it fully takes into account the correlation across isomorphic representations. So the latter will be more favorable to mine the semantic consistency.

To validate this point, we first use IRRT to project the cross-view data into a mid-level high-dimensional redundant feature-isomorphic space and then apply LDA, SSP, TROP, and CJFL to extract a low-dimensional semantic subspace according to Eq. (32). For LDA, SSP, and TROP, we set k, the dimensionality of the low-dimensional subspace, to the number of class labels.

It can be observed from Fig. 8 that CJFL shows an obvious advantage over the other methods based on the trace ratio. This fact shows that, in contrast to the compared approaches, CJFL is effective on maintaining the requisite component covering the complimentary information from different views in the mid-level space.

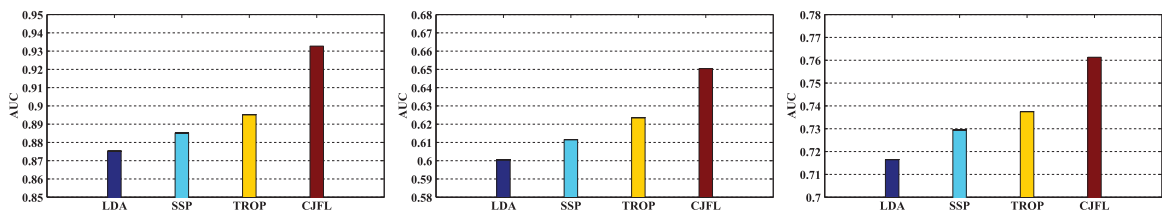
4.7 Evaluation on Shared Subspace Learning Algorithms

As illustrated in Fig. 4, the proposed framework mines a unique high-level semantically shared subspace in which the SCP for cross-view data can be obtained. Here we compare our framework with some other recently proposed shared subspace learning methods, such as ASO [19], CMTFL [20], SSMLC [23], SSMCC [25], and JSNMF [27].

We use six different methods which are IRRT+CJFL, CCA+ASO, CCA+CMTFL, CCA+SSMLC, CCA+SSMCC, and CCA+JSNMF to produce different SCPs according to Eq. (32). Note that, different from our framework, CCA [9] is first implemented to build a feature-isomorphic space among different views prior to the compared approaches. Brief descriptions of the six methods are given in Table 3.

Specifically, both the regularization parameters in ASO and SSMLC are set to 10^{-3} as in [19] and [23]. For JSNMF, the threshold is set to 10^{-2} as in [27]. The tradeoff parameter in SSMCC is tuned from the set $\{10^i | i = -5, -8, \ldots, 4, 5\}$. We select the regularization parameter in CMTFL from the set $\{10^i | i = -6, -5, \ldots, 2, 3\}$.

It comes to our notice from Fig. 9 that in comparison to other existing shared subspace learning algorithms, the proposed IRRT+CJFL framework can achieve the best classification performance. It implies that our framework can mine

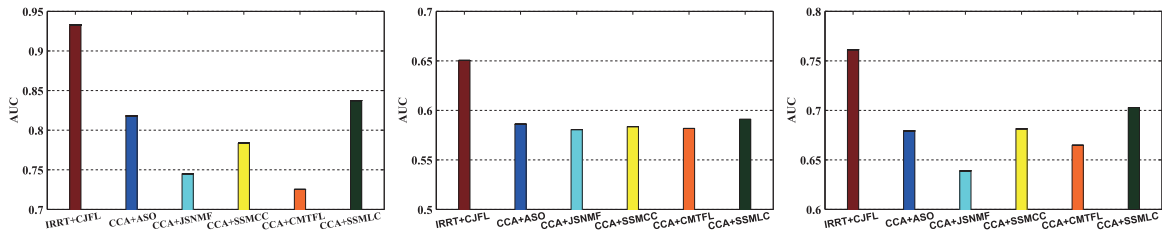


(a) UCI MFeat (b) COREL 5K (c) Wikipedia

Fig. 8. Comparisons of classification performance of trace ratio algorithms.

TABLE 3
Brief Descriptions of the Six Methods

Method	Description
IRRT+CJFL	IRRT is performed first before CJFL is carried out.
CCA+ASO	CCA is performed first before ASO is carried out.
CCA+CMTFL	CCA is performed first before CMTFL is carried out.
CCA+SSMLC	CCA is performed first before SSMLC is carried out.
CCA+SSMCC	CCA is performed first before SSMCC is carried out.
CCA+JSNMF	CCA is performed first before JSNMF is carried out.



(a) UCI MFeat (b) COREL 5K (c) Wikipedia

Fig. 9. Comparisons of classification performance of shared subspace learning algorithms.

TABLE 4
MAP Scores on the Cross-View Data Sets

Dataset	Query	Supervised		Semi-supervised		Unsupervised		
		CJFL	GMLDA	LRGARF	SCM	IRRT	CCA	PLS
UCI MFeat	fou (V_x)	**0.5422**	0.3998	0.4749	0.2545	**0.3314**	0.1830	0.1960
	zer (V_y)	**0.5031**	0.3619	0.4370	0.2208	**0.2971**	0.1571	0.1680
	Average	**0.5227**	0.3809	0.4559	0.2377	**0.3142**	0.1700	0.1820
Corel 5K	DenseHue (V_x)	**0.3191**	0.2387	0.2731	0.1593	**0.2191**	0.1344	0.1504
	HarrisHue (V_y)	**0.2840**	0.2146	0.2443	0.1384	**0.1876**	0.1145	0.1265
	Average	**0.3015**	0.2267	0.2587	0.1488	**0.2033**	0.1245	0.1385
Wikipedia	image (V_x)	**0.4632**	0.3646	0.4121	0.2314	**0.2848**	0.1669	0.1819
	text (V_y)	**0.4252**	0.3304	0.3796	0.2106	**0.2508**	0.1274	0.1414
	Average	**0.4442**	0.3475	0.3959	0.2210	**0.2678**	0.1472	0.1617

the semantic consistency among isomorphic representations more accurately than the other shared subspace learning algorithms.

4.8 Analysis of Cross-View Retrieval Performance

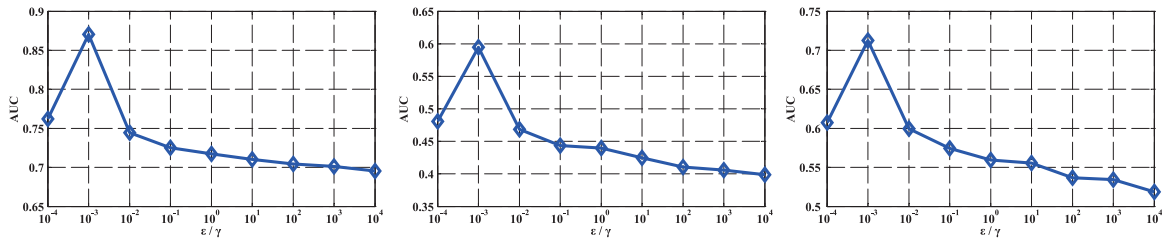
Now we analyze the cross-view retrieval performance of the proposed IRRT+CJFL framework. It consists of two independent models IRRT and CJFL. Practicably, both of them can be applied to cross-view retrieval. The so-called cross-view retrieval refers to the retrieval of a view in response to another query view. It is central to many applications of practical interest, such as finding on the web the picture that best matches a given text or searching the text that best illustrates a given picture.

Particularly, some recent works for cross-view retrieval such as Local Regression and Global Alignment and Relevance Feedback (LRGARF) [3], Semantic Correlation Matching (SCM) [7], and Generalized Multi-view Linear Discriminant Analysis (GMLDA) [14] are used to make retrieval performance comparisons with our models. Depending on their supervised mode, CJFL and GMLDA can be categorized into supervised methods, LRGARF and SCM belong to semi-supervised approaches, and IRRT, CCA and PLS are unsupervised models.

For GMLDA, the positive parameters are set to 10 and 1, and the constraint parameter is set to the trace ratio of the square symmetric defined matrices. We specify the regularization parameter of LRGARF by 1,000, and set the dimensionality k of the low-dimensional subspace to the number of class labels. For SCM, the semantic spaces produced by logistic regression are 10-dimensional probability simplexes.

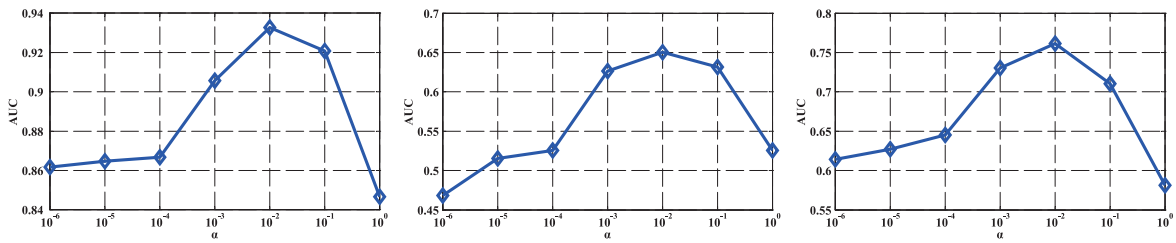
For CCA and IRRT, the cross-view retrieval is carried out using the μ_x and μ_y (see Eq. (2)) as their query view, respectively. Similarly, we use the τ_x and τ_y (see Eq. (31)) as the query view for CJFL. The MAP scores of all the methods for cross-view retrieval are listed in Table 4.

It can be seen from Table 4 that, as unsupervised approaches for cross-view retrieval, IRRT is superior to CCA and PLS and even better than the semi-supervised algorithm SCM. It implies that the feature-isomorphic space by IRRT can capture much more complimentary information from different views than CCA. In addition, compared with other supervised and semi-supervised methods with prior knowledge (class information), the CJFL achieves the



(a) UCI MFeat (b) COREL 5K (c) Wikipedia

Fig. 10. Sensitivity of Parameters ε and γ for IRRT.



(a) UCI MFeat (b) COREL 5K (c) Wikipedia

Fig. 11. Sensitivity of Parameter α for CJFL.

best performance. The comparisons with IRRT, CCA and PLS show that the cross-view retrieval performance can be greatly improved with the involvement of prior knowledge.

4.9 Parameter Sensitivity of IRRT and CJFL

We further evaluate the parameter sensitivity of the proposed models IRRT and CJFL with respect to their own important parameters. Since the parameters ε and γ in Eq. (4) control the amount of information carried by the transformed data, here we focus on the evaluation of performance variation of IRRT with respect to ε and γ. Fig. 10 shows the performance variation of IRRT with different parameter values of ε and γ, which suggests the optimal value for both ε and γ is 10^{-3}.

In addition, since the parameter α in Eq. (28) is applied to avoid violating the intrinsic structure of the coupled representations, it is crucial to observe performance variation of CJFL with respect to α. It can be found from Fig. 11 that CJFL achieves the best performance on all three data sets when α is set to 10^{-2}.

5 Conclusion

In this paper, we have investigated the problem of consistently representing the objects from different views on the basis of their correlations or complementarities. We developed a general framework to project cross-view data into a unique high-level low-dimensional semantically shared subspace to mine the SCP for cross-view data. Within this framework, the proposed IRRT model builds a bridge between multiple heterogeneous low-level feature spaces and a unique high-level semantically shared space. Furthermore, we also proposed a CJFL model to mine the semantic consistency among isomorphic representations.

Practically, the proposed IRRT and CJFL in our framework can be easily extended to multi-view cases. In addition, they are so flexible that either algorithm combined with other existing algorithms can be applied to solve the cross-view problem.

Acknowledgments

This work was supported in part by National Basic Research Program of China (No. 2012CB316400), National Natural Science Foundation of China (No. 61025013, No. 61172129, No. 61210006), Program for Changjiang Scholars and Innovative Research Team in University (No. IRT201206), Program for New Century Excellent Talents in University (No. 13-0661), Fundamental Research Funds for the Central Universities (No. 2012JBZ012), and the State Key Laboratory of Rail Traffic Control and Safety (No. RCS2014ZT13).

References

[1] Z. Lu, Z. Chi, and W. C. Siu, "Extraction and optimization of B-spline PBD templates for recognition of connected handwritten digit strings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 132–139, Jan. 2002.

[2] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, 2004.

[3] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[4] Y. Zhuang, Y. Yang, F. Wu, and Y. Pan, "Manifold learning based cross-media retrieval: A solution to media object complementary nature," *J. VLSI Signal Process.*, vol. 46, no. 2, pp. 153–164, 2007.

[5] R. P. W. Duin. UCI repository of machine learning databases. (1998). [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Multiple+Features

[6] M. Guillaumin, J. Verbeek, and C. Schmid, "Multiple instance metric learning from automatically labeled bags of faces," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 634–647.

[7] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM. Int. Conf. Multimedia*, 2010, pp. 251–260.

[8] Q. Cheng, H. Zhou, and J. Cheng, "The Fisher-Markov selector fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1217–1233, Jun. 2011.
[9] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, Jan. 2011.
[10] H. Wold, "Partial least squares, and " in *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: Wiley, 2006.
[11] B. Thompson, "Canonical correlation analysis, and " in *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ, USA: Wiley, 2005.
[12] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[13] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. ACM Int. Conf. Mach. Learn.*, 2009, pp. 129–136.
[14] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2160–2167.
[15] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 974–981.
[16] S. Koterba, S. Baker, I. Matthews, C. Hu, J. Xiao, J. Cohn, and T. Kanade, "Multi-view AAM fitting and camera calibration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 511–518.
[17] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 446–453.
[18] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning a shared predictive structure from multiple tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1025–1038, May 2013.
[19] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.
[20] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.
[21] T. N. Huy, H. Shao, B. Tong, and E. Suzuki, "A feature-free and parameter-light multi-task clustering framework," *Knowl. Inf. Syst.*, vol. 32, no. 1, pp. 251–276, 2013.
[22] H. Fei and J. Huan, "Structured feature selection and task relationship inference for multi-task learning," *Knowl. Inf. Syst.*, vol. 35, no. 2, pp. 345–364, 2013.
[23] S. Ji, L. Tang, S. Yu, and J. Ye, "Extracting shared subspace for multi-label classification," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 381–389.
[24] X. Kong, M. Ng, and Z. Zhou, "Transductive multi-label learning via label set propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 704–719, Mar. 2013.
[25] Y. Amit, M. Fink, and N. Srebro, S. Ullman, "Uncovering shared structures in multiclass classification," in *Proc. Int. Conf. Mach. Learn.*, 2007, vol. 24, pp. 17–24.
[26] Z. Zhang, M. Zhao, and T. Chow, "Binary- and multi-class group sparse canonical correlation analysis for feature extraction and classification," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2192–2205, Oct. 2013.
[27] S. K. Gupta, D. Phung, B. Adams, T. Tran, and S. Venkatesh, "Nonnegative shared subspace learning and its application to social media retrieval," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1169–1178.
[28] H. Ma, W. Zhao, and Z. Shi, "A nonnegative matrix factorization framework for semi-supervised document clustering with dual constraints," *Knowl. Inf. Syst.*, vol. 36, no. 3, pp. 629–651, 2013.
[29] Z.-Y. Zhang, T. Li, and C. Ding, "Non-negative Tri-factor tensor decomposition with applications," *Knowl. Inf. Syst.*, vol. 34, no. 2, pp. 243–265, 2013.
[30] Y. Wang and Y. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.
[31] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Processing Syst.*, 2001, vol. 13, pp. 556–562.
[32] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.
[33] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[34] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.
[35] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.
[36] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA, USA: SIAM Publishers, 2000.
[37] Y. Nesterov, *Introductory Lectures on Convex Programming*. Norwell, MA, USA: Kluwer, 2004.
[38] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 272–279.
[39] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, NY, USA: Academic, 1991.
[40] J. Yu and Q. Tian, "Learning image manifolds by semantic subspace projection," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 297–306.
[41] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Comput. Vis. Pattern Recog.*, 2007, pp. 457–464.
[42] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 902–909.
[43] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[44] J. Van De Weijer and C. Schmid, "Coloring local feature extraction," in *Proc. Eur. Conf. Comput. Vis.*, 2006, vol. 60, no. 2, pp. 334–348.

Lei Zhang received the ME degree from the School of Computer and Communication Engineering from Tianjin University of Technology, Tianjin, China, in 2010. He is currently working toward the PhD degree in the Institute of Information Science, Beijing Jiaotong University, China. His research interests include data mining, pattern recognition, and multimedia content analysis.

Yao Zhao (M'06-SM'12) received the BS degree from Fuzhou University, China, in 1989, and the ME degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), China, in 1996. He became an associate professor at BJTU in 1998 and became a professor in 2001. From 2001 to 2002, he was a senior research fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the editorial boards of several international journals, including as associate editors of *IEEE Transactions on Cybernetics*, *IEEE Signal Processing Letters*, and an area editor of *Signal Processing: Image Communication* (Elsevier), etc. He was named a distinguished young scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a senior member of the IEEE.

Zhenfeng Zhu received the PhD degree in pattern recognition and intelligence system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005. He is currently an associate professor of the Institute of Information Science, Beijing Jiaotong University. He has been a visiting scholar at the Department of Computer Science and Engineering, Arizona State University, during 2010. His research interests include image and video understanding, computer vision, and machine learning.

Shikui Wei received PhD degree in signal and information processing from Beijing Jiaotong University (BJTU), China, in 2010. From 2010 to 2011, he was a research fellow in the School of Computer Engineering, Nanyang Technological University, Singapore. He is currently an associate professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include computer vision, image/video analysis and retrieval, and copy detection.

Xindong Wu received the bachelor's and master's degrees in computer science from the Hefei University of Technology, China, and the PhD degree in artificial intelligence from the University of Edinburgh, United Kingdom. He is a professor of computer science at the University of Vermont, and a Yangtze River scholar in the School of Computer Science and Information Engineering at the Hefei University of Technology, China. His research interests include data mining, knowledge-based systems, and web information exploration. He is the Steering Committee chair of the IEEE International Conference on Data Mining (ICDM), the editor-in-chief of *Knowledge and Information Systems* (KAIS, by Springer), and a series editor of the Springer Book Series on Advanced Information and Knowledge Processing (AIKP). He was the editor-in-chief of the *IEEE Transactions on Knowledge and Data Engineering* (*TKDE*, by the IEEE Computer Society) between 2005 and 2008. He served as a program committee chair/cochair for ICDM 2003 (the 2003 IEEE International Conference on Data Mining), KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), and CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management). He is a fellow of the IEEE and AAAS.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.