Multi-View Missing Data Completion

Lei Zhang, *Member, IEEE*, Yao Zhao, *Senior Member, IEEE*, Zhenfeng Zhu, Dinggang Shen, *Fellow, IEEE*, and Shuiwang Ji, *Senior Member, IEEE*

Abstract—A growing number of multi-view data arises naturally in many scenarios, including medical diagnosis, webpage classification, and multimedia analysis. A challenge in learning from multi-view data is that not all instances are fully represented in all views, resulting in missing view data. In this paper, we focus on feature-level completion for missing view of multi-view data. Aiming at capturing both semantic complementarity and identical distribution among different views, an Isomorphic Linear Correlation Analysis (ILCA) method is proposed to linearly map multi-view data to a feature-isomorphic subspace through learning a set of excellent isomorphic features, thereby unfolding the shared information from different views. Meanwhile, we assume that missing view obeys normal distribution. Then, the missing view data matrix can be modeled as a low-rank component plus a sparse contribution. Thus, to accomplish missing view completion, an Identical Distribution Pursuit Completion (IDPC) model based on the learned features is proposed, in which the identical distribution constraint of missing view to the other available one in the feature-isomorphic subspace is fully exploited. Comprehensive experiments on several multi-view datasets demonstrate that our proposed framework yields promising results.

Index Terms—Multi-view learning, missing view, feature-level completion, sparse learning, trace norm, optimization

1 Introduction

With the increase of data modality in representing real-world objects, more and more multi-view data become available in various fields, including medical diagnosis, webpage classification, and multimedia analysis. These data have multiple views that generally correspond to distinct sets of feature representations for the same set of underlying objects. A challenge in learning from multi-view data is that not all instances are fully represented in all views, resulting in missing view data. The missing view problem in multi-view learning is different from the missing data problem in mono-view learning, as the missing of a view results in the missing of all attributes in the same view. For example, in the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1] database, many data only have Magnetic Resonance Imaging (MRI) measurement, yet lack Positron Emission Tomography (PET) scan, resulting in a scenario shown in Fig. 1.

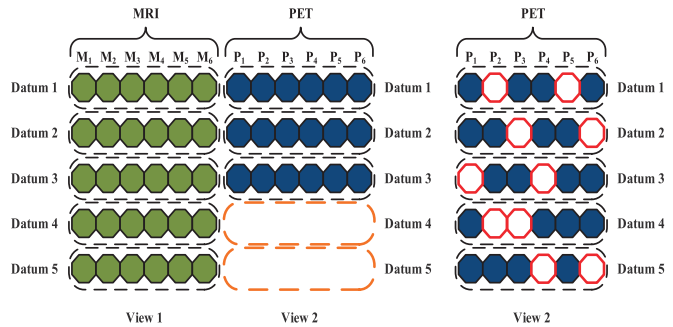
- *L. Zhang is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China. E-mail: zhanglei1@iie.ac.cn.*
- *Y. Zhao and Z. Zhu are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China. E-mail: {yzhao, zhfzhu}@bjtu.edu.cn.*
- *D. Shen is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, and the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-701, Korea. E-mail: dgshen@med.unc.edu.*
- *S. Ji is with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752. E-mail: sji@eecs.wsu.edu.*

Manuscript received 6 Aug. 2017; revised 2 Jan. 2018; accepted 6 Jan. 2018. Date of publication 10 Jan. 2018; date of current version 1 June 2018. (Corresponding authors: Lei Zhang and Shuiwang Ji.) Recommended for acceptance by J. M. Phillips. 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.2018.2791607

More notably, since each view of multi-view data may contain some common and consistent information, multi-view learning can be employed to reduce the noise, as well as to learn the correlations between different views to obtain higher-level information [2], [3], [4], [5], [6], [7]. Nevertheless, missing view data are directly discarded in general, resulting in a severe loss of available information. Furthermore, to the best of our knowledge, little efforts have focused on recovering missing view of multi-view data. Consequently, the above-mentioned applications face great challenge in the real world. To bridge this gap, our work aims to develop an effective feature-level completion method for missing view of multi-view data.

Nevertheless, missing view completion of multi-view data is highly challenging. First of all, since different views (forms, modalities, or sources) span heterogeneous low-level feature spaces, there is no explicit correspondence among the heterogeneous representations from different views. For example, as shown in Fig. 2, the co-occurring image and text in a web page convey the same semantic concept from the perspectives of vision and writing, respectively, so it is not straightforward to directly measure the relationship between heterogeneous representations. Thus, there is a need to build a feature-isomorphic subspace to capture the semantic complementarity among different views. Note that the feature-isomorphic subspace refers to the mappings of heterogeneous representations from different views into a common feature space, in which the same dimension and attributes are used to represent the same semantic concept.

Meanwhile, for the multi-view data in the feature-isomorphic subspace, it can be assumed as illustrated in Fig. 3 that they are under both semantic complementarity and identical distribution constraints. The complementarity constraint refers to the semantic complementarity among



(a) missing view (b) incomplete data

Fig. 1. Missing view and incomplete data. Large chunks of missing view data are marked by the orange dotted line in Fig. 1a. The hollow red solid-line wireframes represent the location of the missing values in the incomplete data in the mono-view setting in Fig. 1b.

(a) The digit 7 in different forms.

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

(c) The MRI measurement and PET scan of brain.

Fig. 2. The cases of multi-view data.

different views that makes much more the consistent information from different views fully contained in the isomorphic representations of multi-view data. Note that the consistent information is the commonality among the heterogeneous representations from different views. Unlike the semantic complementarity constraint, the identical distribution constraint takes high distributive similarity among different views which can group the samples of the same class from the same view together while keeping the instances from different categories away from each other simultaneously. Hence, another issue we need further to deal with for completing missing view of multi-view data is to recover missing view under both semantic complementarity and identical distribution constraints.

1.1 Main Contributions

In this work, we develop a set of methods and algorithms to address the above challenges. The key contributions of this work are highlighted as follows:

- A general feature-level framework for completing missing view to obtain the complex representations for multi-view data is proposed. In this framework, a feature-isomorphic subspace is learned to build a bridge between multiple heterogeneous low-level feature spaces.
- To build a feature-isomorphic subspace to capture both semantic complementarity and identical distribution among different views, we propose a novel Isomorphic Linear Correlation Analysis (ILCA) model with maximum neighbourhood criterion and orthogonal constraints, unfolding the shared information from different views. The maximum neighbourhood criterion in ILCA takes charge of highly correlating the learned features with the class, and the correlations among the features can be removed by the orthogonal constraints. Thus, multiple heterogeneous low-level feature spaces are linearly projected into a feature-isomorphic subspace through a set of learned excellent isomorphic features.
- A new Identical Distribution Pursuit Completion (IDPC) method based on the learned features is proposed to recover missing view of multi-view data, in which the identical distribution constraint of missing view to the other available one in the feature-isomorphic subspace is fully exploited. Consequently, the feature-level completion of missing view is accomplished while noisy information is repressed in the recovered missing view representations of multi-view data.

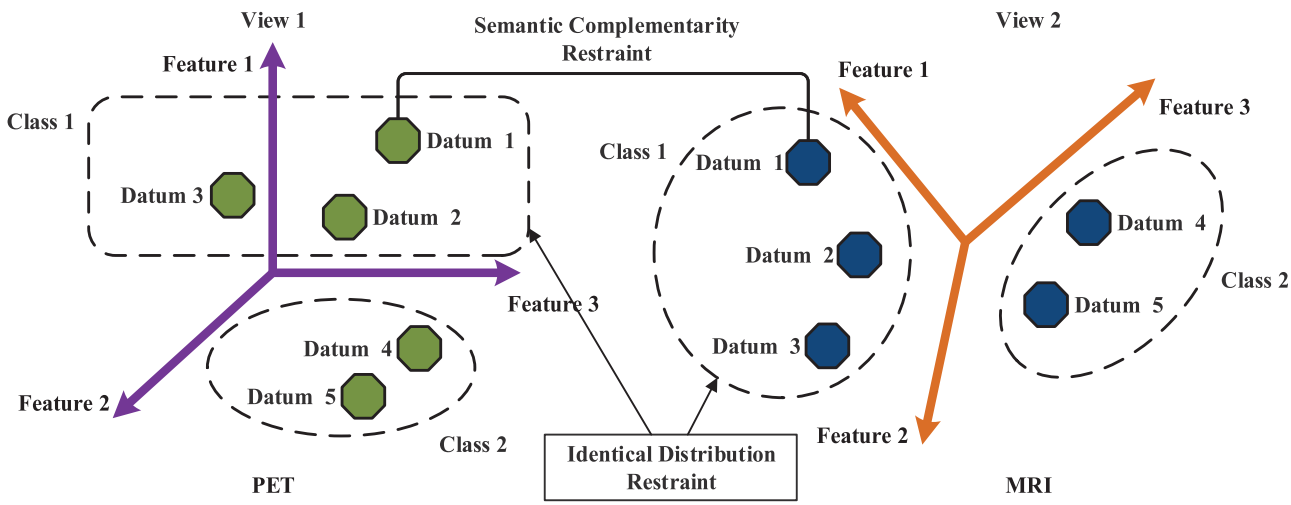


Fig. 3. Semantic complementarity and identical distribution restraints on multi-view data.

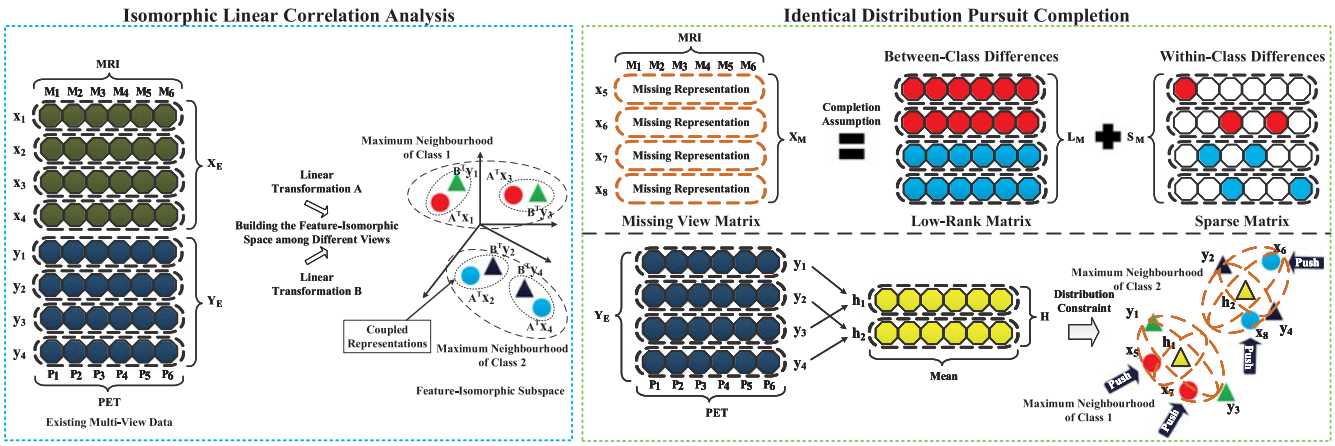


Fig. 4. The proposed framework for completing missing view of multi-view data.

- Extensive experiments on four multi-view datasets are conducted to demonstrate the effectiveness of the proposed framework.

1.2 Organization

The remainder of this paper is organized as follows: We present a general feature-level framework for completing missing view to obtain the integrated representations for multi-view data in Section 2.1. In Section 2.2, a novel Isomorphic Linear Correlation Analysis model is developed for correlating different views through learning a set of excellent isomorphic features. We build a new Identical Distribution Pursuit Completion model to recover missing view of multi-view data under both semantic complementarity and identical distribution restraints in Section 2.3. Furthermore, Section 3 provides an efficient algorithm to solve the proposed framework and analyzes the computational complexities and convergence rates of the proposed algorithms. Section 4 gives a broad overview of some related work. Experimental results and analyses are reported in Section 5. Section 6 concludes this paper.

1.3 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 $X_E = [x_1, \ldots, x_{n_1}]^T \in \mathbb{R}^{n_1 \times d_x}$ and $Y_E = [y_1, \ldots, y_{n_1}]^T \in \mathbb{R}^{n_1 \times d_y}$ be two sets of existing 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_1 is the number of available samples, and d_x and d_y are the dimensionalities of the heterogeneous low-level feature spaces V_x and V_y. Note that for $i = 1, \ldots, n_1$, (x_i, y_i) represents the ith couple of heterogeneous representations. We assume that both $\{x_i\}_{i=1}^{n_1}$ and $\{y_i\}_{i=1}^{n_1}$ are centered, i.e., $\sum_{i=1}^{n_1} x_i = 0$ and $\sum_{i=1}^{n_1} y_i = 0$. Let the data matrix $X_M = [x_{n_1+1}, \ldots, x_{n_1+n_2}]^T \in \mathbb{R}^{n_2 \times d_x}$ be a set of missing representations from the V_x and the data matrix $Y_M = [y_{n_1+1}, \ldots, y_{n_1+n_2}]^T \in \mathbb{R}^{n_2 \times d_y}$ be a set of existing heterogeneous representations from the V_y corresponding to the missing representations X_M.

We use $||A||_* = \sum_{i=1}^{r} \sigma_i$ to denote the trace (nuclear) norm of a matrix $A = [a_{ij}] \in \mathbb{R}^{p \times q}$, 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. $||A||_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}^2}$ is the Frobenius norm of A. If A is a square matrix, then 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.

Additionally, let $|H|$ be the number of elements in the set H; $\triangledown f(C)$ denotes the gradient of any smooth function $f(\cdot)$ at the point C; for $w \in \mathbb{R}^p$, we denote by $diag(w)$ the diagonal matrix having the components of the vector w on the diagonal; let D be a set of representations, $mean(D)$ denotes the average value of D. $I_k \in \mathbb{R}^k$ is an identity matrix.

2 The Proposed Formulation

We propose a general feature-level framework to complete missing view of multi-view data. A graphical illustration of the proposed formulation is given in Fig. 4 to facilitate the understanding the proposed formulations and algorithms significantly.

2.1 Overview of the Proposed Formulations

We provide an overview of the proposed formulations by using the example in Fig. 4. In this example, a set of multi-view data consists of the views MRI and PET. However, the MRI view is missing, such as all attributes in the representations x_5, x_6, x_7, and x_8 are totally absent.

To recover missing view of multi-view data, a feature-isomorphic subspace is learned by ILCA model to build a bridge between multiple heterogeneous low-level feature spaces in the proposed framework, in which the same dimension and attributes are used to represent the same semantic concept. Specifically, to fully exploit both semantic complementarity and similar distributions among different views as shown in Fig. 3, multiple linear transformations A and B are learned using the existing multi-view data X_E and Y_E to eliminate the heterogeneity across them. Thus, a feature-isomorphic subspace is obtained by a set of learned excellent isomorphic features, in which the correlated representations from different views are coupled together to capture the commonality among the heterogeneous representations from different views. Consequently, some maximum neighbourhoods are established among different categories, such as the maximum neighbourhoods of Class 1 and Class 2 in Fig. 4. We can measure the correlation among the multi-view data in the feature-isomorphic

subspace directly. For example, the second co-occurring samples x_2 and y_2 are projected to the feature-isomorphic subspace to eliminate the heterogeneity across them through the linear transformations A and B. In addition, 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 in the feature-isomorphic subspace. For instance, the heterogeneous representations of the co-occurring samples (x_1, y_1) and (x_3, y_3) take high distributive similarity.

Furthermore, we assume that missing view representations obey normal distribution. Then, the expectation naturally corresponds to between-class differences, and the variance represents within-class differences. The rank is used to capture the between-class differences, and the sparsity to mine the within-class differences. Consequently, the missing view matrix X_M is composed of a low-rank matrix L_M and a sparse matrix S_M. According to this completion assumption, the missing view of multi-view data X_M is recovered by IDPC model through exploiting both semantic complementarity and similar distributions among different views learned by ILCA model.

Moreover, some noisy information is inevitably involved in the recovered missing view representations in the process of completion. These factors may seriously affect the performance of the recovered representations. To eliminate the noises effectively, a data distribution constraint induced by a mean matrix H is introduced to push the recovered representations into the neighbourhood centered on the mean of the samples of the same class. The i-the row vector of H is the mean values of the existing samples Y_E with the same class label. For instance, the mean of the representations y_1 and y_3 forms the row vector h_1 because y_1 and y_3 belong to Class 1. Meanwhile, the recovered representation x_5 and x_7 from the missing view MRI are pushed into the neighbourhood centered on the mean h_1 of the samples of the same class from the view PET, and coupled together with the corresponding representation y_5 and y_7 from the view PET in the feature-isomorphic subspace.

With the complementary information from the feature-isomorphic subspace, the recovered representations of different classes as displayed in Fig. 4 will be more likely to be linearly separable in the feature-isomorphic subspace.

2.2 Isomorphic Linear Correlation Analysis

In the following, a novel ILCA model is developed for capturing both semantic complementarity and identical distribution among different views through learning a set of excellent isomorphic features. Our work is motivated by a few prior studies. Recently, Hall [8] have pointed out that discriminative feature set contains features that are highly correlated with the class, yet uncorrelated to each other. Furthermore, Jin et al [9] have shown that the orthogonal constraints on a matrix can be used to effectively remove the correlations among different features. Following the above-mentioned theoretical results [8], [9], we propose a novel ILCA model with maximum neighbourhood criterion and orthogonal constraints to linearly map multiple heterogeneous low-level feature spaces to a feature-isomorphic subspace. Meanwhile, the correlated representations from different views are coupled together to capture both semantic complementarity and identical distribution among different views.

Specifically, let $\mathcal{S}_X$ and $\mathcal{S}_Y$ be the sets of sample pairs from the same class in views V_x and V_y, respectively, and $\mathcal{D}_X$ and $\mathcal{D}_Y$ are the sets of sample pairs from different categories in views V_x and V_y, respectively. Then the within-class scatter matrices can be expressed as follows:

$$J_{\mathcal{S}} = \sum_{\forall (x_i, x_j) \in \mathcal{S}_X} (x_i - x_j)(x_i - x_j)^T, \tag{1}$$

$$R_{\mathcal{S}} = \sum_{\forall (y_i, y_j) \in \mathcal{S}_Y} (y_i - y_j)(y_i - y_j)^T. \tag{2}$$

Meanwhile, the between-class scatter matrices are defined as follows:

$$J_{\mathcal{D}} = \sum_{\forall (x_i, x_j) \in \mathcal{D}_X} (x_i - x_j)(x_i - x_j)^T, \tag{3}$$

$$R_{\mathcal{D}} = \sum_{\forall (y_i, y_j) \in \mathcal{D}_Y} (y_i - y_j)(y_i - y_j)^T. \tag{4}$$

Based on the above definitions, we propose the following optimization problem:

$$\begin{aligned} \boldsymbol{\Psi_1}: \; & \min_{A,B} \parallel X_E A - Y_E B \parallel_F^2 - \alpha(tr(A^T J_{\mathcal{D}} A) + tr(B^T R_{\mathcal{D}} B)) \\ & + \beta(tr(A^T J_{\mathcal{S}} A) + tr(B^T R_{\mathcal{S}} B)) \\ & s.t. \quad A^T A = I_k \quad and \quad B^T B = I_k, \end{aligned} \tag{5}$$

where $A \in \mathbb{R}^{d_x \times k}$, $B \in \mathbb{R}^{d_y \times k}$, $k \in \{1, \ldots, min(d_x, d_y)\}$ is the dimensionality of the feature-isomorphic subspace, $tr(A^T J_{\mathcal{D}} A) + tr(B^T R_{\mathcal{D}} B)$ is a between-class penalty, and $tr(A^T J_{\mathcal{S}} A) + tr(B^T R_{\mathcal{S}} B)$ is a within-class compactness, and α and β are two trade-off parameters. The motivation of introducing the orthogonal constraints in Eq. (5) is to effectively remove the correlations among different features in the same view. Additionally, a maximum neighbourhood criterion is added into the model Ψ_1 to learn the identical distribution among different views. The maximum neighbourhood criterion refers to the trace difference consisting of within-class compactness and between-class penalty. It can be used to group the samples of the same class from the same view together while keeping the instances from different categories away from each other simultaneously. Consequently, a maximum neighbourhood is established between different categories.

Furthermore, Yang et al. have pointed out in [10] through extensive experiments that the classification accuracy is increased significantly by the complex vectors induced by Parallel Feature Fusion Strategy (PFFS) and also demonstrate that the complex vectors are more effective than the union vectors induced by classical Serial Feature Fusion Strategy (SFFS). On the other hand, since the increase of dimension is avoided in PFFS, much computational time is saved. Moreover, it has been proved in [11] that the recognition rate of feature fusion representations is far higher than that of each single feature representations. Therefore, based on above-mentioned theoretical supports [10], [11], the PFFS is utilized to establish the common representations. The details are as follows: for the ith pair of heterogeneous

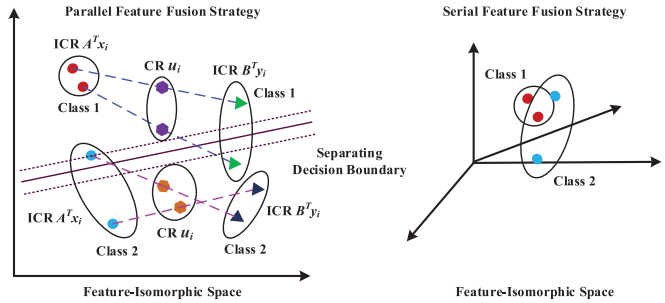


Fig. 5. Parallel and serial feature fusion strategy.

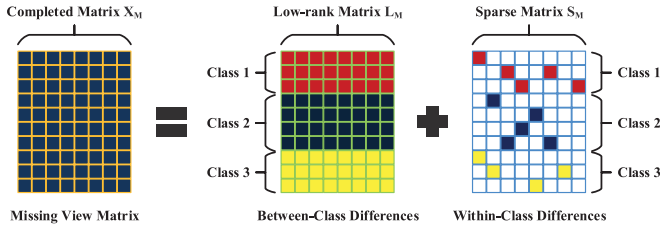


Fig. 6. The composition of missing view data.

representations (x_i, y_i), we can obtain their own Isomorphic Correlated Representations (ICR) with the optimal A^* and B^* by

$$\mu_{x_i} = A^{*T} x_i \quad and \quad \mu_{y_i} = B^{*T} y_i. \tag{6}$$

Consequently, we can obtain a Complex Representations (CR) μ_i in the feature-isomorphic subspace based on μ_{x_i} and μ_{y_i}:

$$\mu_i = (\mu_{x_i} + \mu_{y_i})/2. \tag{7}$$

As shown in Fig. 5, the union vectors are always high dimensional and contain much redundant information and some conflicting information which are unfavorable for recognition. However, the favorable discriminatory information is retained and at the same time the unfavorable redundant or conflicting information is eliminated in the CR.

In Section 3.1, an efficient algorithm is proposed to solve the problem Ψ_1.

2.3 Identical Distribution Pursuit Completion

In this section, we propose a new feature-level missing view completion method, known as IDPC model, to recover missing view of multi-view data. Our method is built on the basis of both semantic complementarity and identical distribution among different views learned in the proposed model Ψ_1. Some previous studies inspire our work. Following the idea behind the robust PCA, we assume that missing view obeys normal distribution. Then, the expectation naturally corresponds to between-class differences, and the variance represents within-class differences. In [12], [13], it has been justified that the rank is a powerful tool to capture between-class differences information in the matrix case. In addition, it has been proved in [14] that the sparse representations can effectively uncover the within-class differences of data. Therefore, we suppose that the missing view data can be represented as in Fig. 6, where the matrix X_M in the view V_x can be modeled as a low-rank part L_M plus a sparse contribution S_M. Then the low-rank component L_M naturally corresponds to the between-class differences, and the sparse component S_M captures the within-class differences. Thus, to recover missing view of multi-view data, it is essential to impose the low-rank and sparse constraints on the recovered missing view representations.

Recently, Candès and Recht [15], Recht et al [16], and Candès and Tao [17] have shown that the trace norm of a matrix can be used to approximate the rank of the matrix. In addition, Wright et al. [14] have shown that the sparse representations computed by ℓ_1-minimization can effectively uncover the identifying characteristics of data. Moreover, some noisy information is usually included in the recovered missing view representations when capturing underlying information. These factors may seriously affect the performance of the representations. Therefore, it is necessary to eliminate the noisy components in the recovered missing view representations. Recently, Weinberger et al [18] have shown that the data distribution induced by distance metric can eliminate noise to a large extent.

In view of the above-mentioned theoretical results [12], [13], [14], [15], [16], [17], [18], we propose a novel IDPC method that incorporates trace norm constraint, ℓ_1-norm regularization, and data distribution constraint. By exploiting both semantic complementarity and identical distribution among different views, missing view of multi-view data are recovered by the proposed IDPC method while noisy information is effectively suppressed.

Specifically, we have built a feature-isomorphic subspace to capture both semantic complementarity and identical distribution among different views through learning a set of excellent isomorphic features. Let (A^*, B^*) be the optimal solutions of the problem Ψ_1. Then the proposed approach can be formulated as follows:

$$\mathbf{\Omega_1}: \begin{array}{ll} \underset{L_M, S_M}{min} & \| (L_M + S_M)A^* - Y_M B^* \|_F^2 + \gamma \|S_M\|_1 \\ s.t. & \|(L_M + S_M)A^* - HB^*\|_F^2 \le \pi \quad and \\ & \|L_M\|_* \le \varepsilon, \end{array} \tag{8}$$

where L_M and S_M denote the between-class and within-class differences of the missing view representations X_M, respectively, γ is a trade off, π is a distance metric parameter, and ε is a pre-specified positive parameter to control the amount of information carried by the recovered missing view representations. The matrix $H = [h_1, \ldots, h_{n_2}]^T \in \mathbb{R}^{n_2 \times d_y}$ imposes a data distribution constraint in order to ensure the recovered missing view representations $L_M + S_M = X_M$ having identical distribution with the existing view representations Y_E, thereby eliminating noise to a large extent. Let the nearest neighbor y_{NN}^i of the ith sample from the Y_M be contained in the Y_E. Each row h_i^T in the matrix H consists of the mean values of the existing samples Y_E with the same class label corresponding to the y_{NN}^i. Specifically, let $\mathcal{C}_X^t$ and $\mathcal{C}_Y^t$ be the sample sets of tth class from the view V_x and V_y, respectively. We define

$$D_t = \{y_j | y_j \in Y_E \wedge y_j \in C_Y^t \wedge y_{NN}^i \in C_Y^t\}, \tag{9}$$

$$D^i = \bigcup\nolimits_t D_t, \tag{10}$$

$$h_i = mean(D^i), \tag{11}$$

where D^i is the sample sets of the same class as the y^i_{NN} in the Y_E.

The motivation of introducing the first item in the objective function in Eq. (8) is to make the recovered representations carrying much more semantically complementary information by the learned linear transformations by ILCA model. In addition, the trace norm constraint imposed on the between-class differences L_M will bring about the recovered representations linear separable as much as possible in the feature-isomorphic subspace. Moreover, much more identifying characteristics are involved in the recovered representations by adding the ℓ_1-norm regularization on the within-class differences S_M. Furthermore, to eliminate the noises at the greatest extent involved in the recovered representations, the data distribution constraint (the first constraint in Eq. (8)) composed of the maximum neighbourhood criterion learned by ILCA is introduced to pull the recovered representations into the neighbourhood of the mean value of the samples of the same class as the center. Thereby, the distribution of the recovered representations is identical with the instances of the same category in the feature-isomorphic subspace.

Thus, the proposed IDPC model is different from the existing matrix completion methods because of the full consideration of both semantic complementarity and identical distribution among different views. To the best of our knowledge, no existing efforts have focused on this type of completion.

Note that solving the problem Ω_1 in Eq. (8) directly is a challenging task for two main reasons. First, it is difficult to seek the solution that satisfies the data distribution constraint. Second, the trace norm constraints are not smooth, which makes it even more difficult to compute the optimum. Thus, we propose to use Lagrangian duality to augment the objective function with a weighted sum of the data distribution constraint to obtain a solvable convex problem Ω_2 as follows:

$$\mathbf{\Omega_2}: \begin{array}{ll} \underset{L_M,S_M}{min} & \| (L_M+S_M)A^* - Y_M B^* \|_F^2 + \gamma \|S_M\|_1 \\ & + \eta(\|(L_M+S_M)A^* - HB^*\|_F^2 - \pi) \\ s.t. & \|L_M\|_* \le \varepsilon. \end{array} \tag{12}$$

Section 3.2 presents an efficient algorithm to compute the optimum for the problem Ω_2.

3 Efficient Algorithms for the Proposed Formulations

In this section, we develop efficient algorithms to solve the proposed formulations. Specifically, an iterative algorithm for solving the ILCA model Ψ_1 in Section 2.2 is presented in the Section 3.1. Additionally, we show in Section 3.2 how to solve the IDPC model Ω_2 proposed in Section 2.3. Furthermore, the computational complexities of the proposed algorithms are analyzed in Section 3.3.

3.1 An Efficient Solver for Ψ_1

For notational simplicity, we denote the optimization problem Ψ_1 by

$$\underset{Z \in \mathcal{C}}{min} \quad f(Z), \tag{13}$$

where $f(\cdot) = \|\cdot\|_F^2 - \alpha tr(\cdot) + \beta tr(\cdot)$ is a smooth objective function, $Z = [A \quad B]$ represents the optimization variables, and $\mathcal{C}$ is a closed domain set with respect to each variable A and B

$$\mathcal{C} = \{Z | A^T A = I_k, B^T B = I_k\}. \tag{14}$$

Obviously, the non-convex optimization problem in Eq. (13) is generally difficult to optimize due to the orthogonal constraints. However, Guo and Xiao have pointed out in [19] that Gradient Descent Method with Curvilinear Search (GDMCS) in [20] can effectively solve non-convex optimization problem for a local optimal solution as long as the Armijo-Wolfe conditions are satisfied.

Furthermore, since the objective function in Eq. (13) is smooth, the gradient of the objective function with respect to A, B can be easily computed, respectively. Accordingly, it is appropriate to use the gradient descent method to solve the problem Ψ_1 in Eq. (13).

Algorithm 1. Isomorphic Linear Correlation Analysis (**ILCA**)

Input: $f(\cdot)$, $Z = [A \ B]$, $\epsilon > 0, 0 < \mu < 1, 0 < \rho_1 < \rho_2 < 1$.
Output: Z^*.

Compute $[A] = \textbf{\textit{Schmidt}}(A)$.
Compute $[B] = \textbf{\textit{Schmidt}}(B)$.
for $i = 1$ to m
  Compute G_1 and G_2 according to Eqs. (15), (16).
  if $\|G_1\|_F^2 + \|G_2\|_F^2 \le \epsilon$ then stop and exit.
  Compute F_1 and F_2 using Eqs. (17), (18).
  Compute $f'_\tau(Q_1(0), Q_2(0))$ via Eq. (24).
  Set $\tau = 1$.
  for $step = 1$ to $max - step$
    Compute $Q_1(\tau)$ and $Q_2(\tau)$ using Eqs. (19), (20).
    Compute $f'_\tau(Q_1(\tau), Q_2(\tau))$ via Eq. (23).
    if Armijo-Wolfe conditions in Eqs. (21), (22) are satisfied then break.
    Set $\tau = \mu\tau$.
  end-for
  if $step > max - step$ then stop and exit.
  Update $A = Q_1(\tau)$ and $B = Q_2(\tau)$.
end-for
Set $Z^* = [A \ B]$.

In each iteration of the gradient descent procedure, given the current feasible point (A, B), the gradients can be computed as follows:

$$\begin{aligned} G_1 = \nabla_A f(A,B) &= 2X_E^T X_E - 2X_E^T Y_E B \\ &- \alpha(J_{\mathcal{D}} + J_{\mathcal{D}}^T)A + \beta(J_{\mathcal{S}} + J_{\mathcal{S}}^T)A, \end{aligned} \tag{15}$$

$$\begin{aligned} G_2 = \nabla_B f(A,B) &= 2Y_E^T Y_E - 2Y_E^T X_E A \\ &- \alpha(R_{\mathcal{D}} + R_{\mathcal{D}}^T)B + \beta(R_{\mathcal{S}} + R_{\mathcal{S}}^T)B. \end{aligned} \tag{16}$$

We then compute two skew-symmetric matrices

$$F_1 = G_1 A^T - A G_1^T, \tag{17}$$

$$F_2 = G_2 B^T - B G_2^T. \tag{18}$$

It is easy to see $F_1^T = -F_1$ and $F_2^T = -F_2$. The next new point can be searched as a curvilinear function of a step size variable τ, such that

$$Q_1(\tau) = (I + \tau F_1/2)^{-1}(1 - \tau F_1/2)A, \tag{19}$$

$$Q_2(\tau) = (I + \tau F_2/2)^{-1}(1 - \tau F_2/2)B. \tag{20}$$

It is easy to verify that $Q_1(\tau)^T Q_1(\tau) = I$ and $Q_2(\tau)^T Q_2(\tau) = I$ for all $\tau \in \mathbb{R}$. Thus we can stay in the feasible region along the curve defined by τ. Moreover, $dQ_1(0)/d\tau$ and $dQ_2(0)/d\tau$ are equal to the projections of $(-G_1)$ and $(-G_2)$ onto the tangent space $\mathcal{C}$ at the current point (A, B). Hence $\{Q_1(\tau), Q_2(\tau)\}_{(\tau \geq 0)}$ is a descent path in the close neighborhood of the current point. We thus apply a similar strategy as the standard backtracking line search to find a proper step size τ using curvilinear search, while guaranteeing the iterations to converge to a stationary point. We determine a proper step size τ as one satisfying the following Armijo-Wolfe conditions [20]

$$f(Q_1(\tau), Q_2(\tau)) \leq f(Q_1(0), Q_2(0)) + \rho_1 \tau f_\tau^{'}(Q_1(0), Q_2(0)), \tag{21}$$

$$f_\tau^{'}(Q_1(\tau), Q_2(\tau)) \geq \rho_2 f_\tau^{'}(Q_1(0), Q_2(0)). \tag{22}$$

Here $f_\tau^{'}(Q_1(\tau), Q_2(\tau))$ is the derivative of f with respect to τ,

$$\begin{aligned} &f_\tau^{'}(Q_1(\tau), Q_2(\tau)) = \\ &- tr((\triangledown_A f(Q_1(\tau), Q_2(\tau)))^T \Big(I + \frac{\tau}{2} F_1\Big)^{-1} F_1 \Big(\frac{A + Q_1(\tau)}{2}\Big) \\ &- tr((\triangledown_B f(Q_1(\tau), Q_2(\tau)))^T \Big(I + \frac{\tau}{2} F_2\Big)^{-1} F_2 \Big(\frac{B + Q_2(\tau)}{2}\Big). \end{aligned} \tag{23}$$

Therefore,

$$\begin{aligned} f_\tau^{'}(Q_1(0), Q_2(0)) &= -tr(G_1^T (G_1 A^T - A G_1^T) A) \\ &\quad - tr(G_2^T (G_2 B^T - B G_2^T) B) \\ &= -\frac{\|F_1\|_F^2}{2} - \frac{\|F_2\|_F^2}{2}. \end{aligned} \tag{24}$$

The overall algorithm is given in Algorithm 1, where the function $\boldsymbol{Schmidt}(\cdot)$ [21] denotes the GramSchmidt process.

3.2 An Efficient Solver for Ω_2

This section provides an efficient algorithm to solve the model Ω_2 proposed in Section 2.3. Similarly, the optimization problem Ω_2 can be simplified as

$$\min_{\Theta \in \mathcal{Q}} \quad F(\Theta) = w(\Theta) + \gamma g(\Theta), \tag{25}$$

where $w(\cdot) = \|\cdot\|_F^2$ is a smooth function, $g(\cdot) = \|\cdot\|_1$ is an undifferentiable function, $\Theta = [L_\Theta \;\; S_\Theta]$ represents the optimization variables, and $\mathcal{Q}$ is a closed and convex domain set defined as

$$\mathcal{Q} = \{\Theta | \|L_\Theta\|_* \leq \varepsilon\}. \tag{26}$$

Obviously, the optimization problem in Eq. (25) is non-convex. However, Ando and Zhang have testified in [22] that the alternating optimization method can effectively solve non-convex problem. They have also pointed out that this method usually did not lead to serious problems since given the local optimal solution of one variable, the solution of other variables would still be globally optimal.

Additionally, the problem in Eq. (25) is separately convex with respect to L_Θ and S_Θ. Furthermore, as $w(\cdot)$ is continuously differentiable with Lipschitz continuous gradient [23] with respect to L_Θ and S_Θ, respectively. Thus, through combining Accelerated Projected Gradient (APG) [23] method and alternating optimization approach [22], the problem in Eq. (25) can be effectively solved.

Algorithm 2. Identical Distribution Pursuit Completion (**IDPC**)

Input: $F(\cdot)$, $w(\cdot)$, $g(\cdot)$, $\Theta_0 = [L_{\Theta_0} \;\; S_{\Theta_0}]$, γ, $\varepsilon > 0$, $\mu > 0$, $\rho > 0$, $\tau_1 > 0$, $\eta_1 > 0$, $t = 1$, $q = 1$.
Output: Θ^*.

Set $L_{\Theta_1} = L_{\Theta_0}$ and $S_{\Theta_1} = S_{\Theta_0}$.
for $i = 1, 2, \cdots, m$ do
  Fix S and approximately solve for L.
  Define $F_{\tau, L_P}(L_\Theta) = w(L_P) + \langle \triangledown w(L_P), L_\Theta - L_P \rangle + \tau \|L_\Theta - L_P\|_F^2/2 + \gamma g(L_\Theta)$.
  for $j = 1, 2, \cdots, h_1$ do
    Set $\alpha_j = (t-1)/t$.
    Compute $L_P = (1 + \alpha_j) L_{\Theta_i} - \alpha_j L_{\Theta_{i-1}}$.
    Compute $\triangledown_{L_P} w(L_P)$.
    While (true)
      Compute $\widehat{L_P} = L_P - \triangledown_{L_P} w(L_P)/\tau_i$.
      Compute $[L_{\Theta_{i+1}}] = \boldsymbol{EPTNC}(\widehat{L_P}, \varepsilon)$.
      if $F(L_{\Theta_{i+1}}) \leq F_{\tau_i, L_P}(L_{\Theta_{i+1}})$, then break;
      else Update $\tau_i = \tau_i \times 2$.
      end-if
    end-while
    Update $t = \big(1 + \sqrt{1 + 4t^2}\big)/2$, $\tau_{i+1} = \tau_i$.
  end-for
  Fix L and approximately solve for S.
  Define $F_{\tau, S_P}(S_\Theta) = w(S_P) + \langle \triangledown w(S_P), S_\Theta - S_P \rangle + \eta \|S_\Theta - S_P\|_F^2/2 + \gamma g(S_\Theta)$.
  for $k = 1, 2, \cdots, h_2$ do
    Set $\beta_k = (q-1)/q$.
    Compute $S_P = (1 + \beta_k) L_{\Theta_i} - \beta_k S_{\Theta_{i-1}}$.
    Compute $\triangledown_{S_P} w(S_P)$.
    While (true)
      Compute $\widehat{S_P} = S_P - \triangledown_{S_P} w(S_P)/\eta_i$.
      Compute $[S_{\Theta_{i+1}}] = \boldsymbol{STO}(\widehat{S_P}, \mu, \rho)$.
      if $F(S_{\Theta_{i+1}}) \leq F_{\eta_i, S_P}(S_{\Theta_{i+1}})$, then break;
      else Update $\eta_i = \eta_i \times 2$.
      end-if
    end-while
    Update $q = \big(1 + \sqrt{1 + 4q^2}\big)/2$, $\eta_{i+1} = \eta_i$.
  end-for
end-for
Set $\Theta^* = [L_{\Theta_{i+1}} \;\; S_{\Theta_{i+1}}]$.

APG belongs to the first-order gradient schemes and its global convergence rate is optimal among all first-order methods [24], [25], which will construct a searching point sequence $\{S_i\}$ to update a solution point sequence $\{Z_i\}$. Note that, in the APG algorithm, the Euclidean projection of a given point p onto the convex set $\mathcal{G} = \{\theta | \|\theta\|_* \leq m\}$ can be defined by:

$$proj_{\mathcal{G}}(s) = \arg \min_{\theta \in \mathcal{G}} \|\theta - s\|_F^2/2, \tag{27}$$

where m is a pre-specified positive constant. The projection procedure can be solved efficiently via Efficient Projection on Trace Norm Constraints (EPTNC) [26].

TABLE 1
Computational Complexity

Method	Computational Complexity
ILCA	$O((4k^2\sum_{i=1}^{v} d_i) * m)$
IDPC	$O((d_x n_2^2 h_1 + d_x n_2^2 h_2) * m)$

EPTNC is an efficient gradient-related projection method which can formulate the problem in Eq. (27) as a simple singular optimization by projecting a vector onto a simplex. It is widely applied in approaching a sparse solution in sparse feature learning when the number of examples and the dimension are large. Then we can use the EPTNC algorithm to solve Eq. (27). The details of this procedure are given in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2018.2791607.

In the APG algorithm, the unconstrained optimization associated with the non-differentiable function $g(\cdot)$ can be defined as

$$\theta_* = arg \ \min_{\theta} \quad \mu\|\theta\|_1 + \rho\|\theta - s\|_F^2, \tag{28}$$

where μ and ρ are two pre-specified positive parameters.

Soft thresholding Operator (STO) [27] is a very popular tool to solve a non-smooth regular convex function. This operator is a proximal mapping of the ℓ_1-norm to yield sparse representations. Since the ℓ_1-norm is separable, the operator exerts influence on each element separately in a vector or matrix when it is used in a vector or matrix. Because of the widespread application of ℓ_1 penalties, the property of soft thresholding operator to efficiently find the sparse solution to the problem in Eq. (28) becomes very useful. Then we can use the STO algorithm to solve Eq. (28), and the details are given in the supplementary material, available online.

Finally, when applying the APG method for solving the problem in Eq. (25), the Euclidean projection $\Theta = [L_\Theta \ \ S_\Theta]$ of a given point $P = [L_P \ \ S_P]$ onto the set $\mathcal{Q}$ is defined by

$$proj_{\mathcal{Q}}(P) = \arg \ \min_{\Theta \in \mathcal{Q}} \|\Theta - P\|_F^2/2. \tag{29}$$

By combining APG, EPTNC, and STO, we can solve the problem in Eq. (25), and the details are given in Algorithm 2.

3.3 Analysis of Computational Complexity

In this section, we will discuss the Computational Complexities (CC) of the proposed ILCA and IDPC algorithms.

Obviously, because $k \in \{1, \ldots, min(d_1, \ldots, d_i)\}$ (d_i denotes the dimensionalities of the view V_i), the computational cost $O((1+k)k/2)$ of the function $Schmidt(\cdot)$ [21] is relatively small. Consequently, the computational cost of ILCA algorithm depends mostly on the cost for computing the value of the algorithm GDMCS. Wen and Yin have proved in [20] the computational complexity of GDMCS as $O(4dk^2)$ where d and k are the dimensionalities of the low-level feature space and feature-isomorphic subspace, respectively. Thus, the dominating computational complexity of ILCA algorithm is $O((4k^2\sum_{i=1}^{v} d_i) * m)$ (v is the number of views, and m is the number of iteration).

Additionally, as shown in Algorithm 2, the main computational complexity for IDPC is involved with the solving of the functions $EPTNC(\cdot)$ [26] and $STO(\cdot)$ [27], respectively. These two functions will consume most computational time to calculate the Singular Value Decomposition (SVD) [28] of approximate solution. Accordingly, the computational complexities of EPTNC and STO are $O(d_x n_2^2)$. Therefore, IDPC has the computational complexity of $O((d_x n_2^2 h_1 + d_x n_2^2 h_2) * m)$ (h_1 and h_2 are the number of iteration).

The dominating CCs of the proposed ILCA and IDPC algorithms are listed in the Table 1.

We can see from Table 1 that the computational cost of ILCA largely depends the dimensionalities of the low-level feature spaces, since the dimensionality k of the feature-isomorphic subspace is usually small. Besides, with the increase in the number of missing multi-view data, the computational time of IDPC will rise continuously.

4 Related Work

Our proposed work is related to some prior methods for mining the correlation between different views in multi-view learning and some matrix completion methods for mono-view data completion.

4.1 Existing Isomorphic Methods

To eliminate the heterogeneity across different views, many techniques have been proposed recently, including dimensionality reduction [29], [30], kernel methods [31], [32], and subspace learning [19], [33], [34].

4.1.1 Dimensionality Reduction

CCA (Canonical Correlation Analysis) [29], [35], [36] and OPLS (Orthogonal Partial Least Squares) [30] are two classical statistical analysis techniques for modeling correlation between sets of observed variables. 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 OPLS maximizes their covariance. When one of views is the predictors induced from class label, it has been shown that CCA is equivalent to Linear Discriminant Analysis (LDA) [29]. Additionally, Multi-view CCA (MCCA) (>2 views) [37] is also a classical algorithm, which the label representation is used as the third view. However, Bach and Jordan [38] also have proved that that LDA is only equivalent to CCA in the two-variable case on the condition that their own generalized eigenvalue problems are equivalent. Therefore, three-view CCA or MCCA (>2 views) is not equivalent to LDA.

4.1.2 Kernel Methods

Kernel CCA (KCCA) offers an alternative solution for CCA by implicitly mapping multi-view data into a feature-isomorphic subspace. Recently, Hardoon et al. [31] proposed a general method using KCCA to learn a semantic representation of web images and their associated texts. Moreover, Andrew et al. presented a Deep CCA (DCCA) in [32] to learn complex nonlinear transformations of two associated views. Unlike KCCA, DCCA does not require an inner

product, which provides a flexible nonlinear alternative to KCCA.

4.1.3 Subspace Learning

Recently, some subspace learning methods have been proposed for multi-view classification. Guo [33] developed a convex subspace representation learning approach for general multi-view clustering. In [34], a large margin classifier was constructed by integrating the nature of the multi-view setting into the transfer learning framework and imposing the consistencies among multiple views. A subspace co-regularized multi-view learning method was presented in [19] to project input data into a low-dimensional subspace.

4.2 Matrix Completion Methods

The completion of missing data in the mono-view setting can generally be formulated as a matrix completion problem. A number of studies have introduced various Matrix Completion (MC) algorithms to complete the matrix with incomplete data. The existing MC methods involve nuclear norm [27], [39], [40], [41], [42], [43], [44], statistical analysis [45], K-nearest neighbor [46], and Singular Value Decomposition [28], which have gained promising performance in some applications.

4.2.1 Nuclear Norm

Most of the existing MC approaches are based on the nuclear norm. Cai et al. [27] introduces a Singular Value Thresholding (SVT) algorithm to approximate the incomplete data matrix with a matrix with minimum nuclear norm among all matrices obeying a set of convex constraints. In [39], Toh and Yun proposed a Nuclear Norm-regularized Least Squares (NNLS) method to solve an unconstrained nonsmooth convex optimization problem. A MC algorithm, called Truncated Nuclear Norm Regularization (TNNR) was developed in [40] by minimizing the truncated nuclear norm. Marjanovic and Solo constructed a ℓ_p penalized least squares problem for MC in [41]. They used the l_p Accelerated Projected Gradient (ℓ_pAPG) [23] algorithm to solve the problem. Robust Principal Component Analysis (RPCA) [42] completed matrix by minimizing a weighted combination of the nuclear norm and the ℓ_1 norm. Xiao and Guo [43] proposed a cross-language MC method to produce a complete parallel document-term matrix for all documents in two languages. A robust transfer PCA method was presented in [44] for recovering low-rank matrix from a heavily corrupted observation matrix by leveraging related uncorrupted auxiliary data.

4.2.2 Other Related Approaches

In [45], Schneider proposed a statistical analysis algorithm, called Expectation Maximization (EM), to impute missing values by estimating the mean and the covariance matrix of an incomplete dataset. Based on the Euclidean distance between samples, K-Nearest Neighbor [46] method can be used to impute missing value by replacing the missing value in the data matrix with the corresponding value from the nearest column. SVD [28] is a standard MC method based on low-rank approximation. It first provide some initial guesses (such as 0) to the missing data values, and then decompose the filled-in matrix to obtain a low-rank approximation. Next, it update the missing values as their corresponding values in the low-rank estimation until convergence. In [47], [48], Xiang et al. proposed a incomplete Source and Feature Selection (iSFS) model to complete the block-wise missing data in multi-source problem. Srivastava and Salakhutdinov [49] proposed a Multimodal Deep Boltzmann Machines (MDBM) model to fill in missing modalities by sampling from the conditional distributions over them given the observed ones.

5 Experimental Study

In this section, we evaluate and analyze the effectiveness of the proposed formulations and algorithms for missing view completion of multi-view data.

5.1 Datasets

Our experiments are conducted on five publicly available multi-view datasets, namely, UCI Multiple Features (UCI MFeat) [50], Alzheimer's Disease Neuroimaging Initiative [1], Wikipedia [3], Corel 5K [51], and MIR Flickr [52]. Due to limited space, the details of the datasets are provided in the supplementary material, available online.

5.2 Experimental Setup

Note that all the data are normalized to unit length. Each dataset is randomly separated into a training set and a test set. The training samples account for 80 percent of each original dataset, and the remaining ones act as the test data. Such a partition of each dataset is repeated five times and the average performance is reported. In the training set and test set, 10 percent of multi-view data have missing view.

Some key parameters of all the methods in our experiments are tuned using the 5-fold cross-validation based on the AUC (area under the receiver operating characteristic curve) on the training set. For each time, one-fold data set is used for testing while the other folds are used for training. The training set can be split further into training part and validation part for parameter tuning. The final classification accuracy is the average of the accuracies across all 5 cross-validation folds. Particularly, the LIBSVM classifier serves as the benchmark for the tasks of classification in the experiments.

We use the AUC and Mean Reconstruction Error (MRE) score to evaluate the proposed framework. The MRE is defined as $\Sigma_i \|x_i - x_i'\|/n$ (x' is the recovered multi-view instance), which is a commonly used criterion in matrix completion.

5.3 Evaluation on Single and Integrated View

To evaluate the ability of the proposed ILCA model for capturing both semantic complementarity and identical distribution among different views, we compare the complex representations μ as given in Eq. (7) with the original expressions of either single view.

For the proposed ILCA model, the dimensionality k of the feature-isomorphic subspace is specified by $min(d_x, d_y)$ and the trade-off parameters α and β are tuned on the sets $\{10^i | i = -2, -1, 0, 1, 2\}$.

Clearly, it can be observed from Table 2 that the CR μ as given in Eq. (7) outperforms the original expressions of

TABLE 2
Classification Performance of Single and Integrated Views in Terms of AUC

Dataset	Representations		
UCI MFeat	fou	zer	CR
	0.9015	0.9285	**0.9536**
ADNI	MRI	PET	CR
	0.5492	0.6803	**0.7213**
Wikipedia	Image	Text	CR
	0.5886	0.7822	**0.8254**

either single view. This observation verifies the effectiveness of ILCA for capturing the semantic complementarity among different views.

5.4 Comparison of CCA, OPLS, LDA, and ILCA

The purpose of comparing the proposed ILCA model and CCA [29], OPLS [30], and LDA [53] is to show the importance of mining the identical distribution among different views. Here, the dimensionality k of the feature-isomorphic subspace is specified by $min(d_x, d_y)$ for both OPLS and CCA. For LDA, we set the dimensionality k of the low-dimensional subspace to the number of class labels.

Due to their inherent limitations, OPLS and CCA can only project the multi-view data into a low-dimensional space according to Eq. (7) without the full consideration of identical distribution among different views. Therefore, the feature-isomorphic spaces learned by OPLS and CCA may contain much more noise, which groups the instances from different categories together while keeping the samples of the same

class away from each other simultaneously. Additionally, since LDA is originally developed for handling mono-view problems, it can only learn some limited distributional information among different views.

The proposed ILCA model linearly maps multiple heterogeneous low-level feature spaces to a feature-isomorphic one using orthogonal constraints and maximum neighbourhood criterion. As shown in Table 3, the superiority of ILCA over CCA, OPLS, and LDA in the classification performance is quite clear. For example, nearly 20 percent gain is achieved for the ADNI dataset. It means that ILCA can learn the identical distribution among different views more effectively than CCA, OPLS, and LDA.

5.5 Analysis of Explicit and Implicit Projection

Here we analyze the explicit and implicit projections. As mentioned above, CCA may not extract useful descriptors

TABLE 3
Classification Performance of CCA, OPLS, LDA, and ILCA in Terms of AUC

Dataset	Method			
	CCA	OPLS	LDA	ILCA
UCI MFeat	0.9314	0.9026	0.9229	**0.9536**
ADNI	0.4590	0.5574	0.6380	**0.7213**
Wikipedia	0.7158	0.7144	0.7646	**0.8339**

TABLE 4
Classification Performance of MCCA, KCCA, DCCA, and ILCA in Terms of AUC

Dataset	Method			
	MCCA	KCCA	DCCA	ILCA
UCI MFeat	0.6187	0.6371	0.8494	**0.9536**
ADNI	0.5451	0.5738	0.5393	**0.7213**
Wikipedia	0.6482	0.8096	0.8196	**0.8339**

of data due to its inherent limitation [31]. KCCA [31] and DCCA [32] offer an alternative solution by nonlinearly mapping the multi-view data into a feature-isomorphic subspace. However, unlike KCCA and DCCA, our proposed ILCA model adopts explicit projecting method with orthogonal constraints and maximum neighbourhood criterion. Thus, ILCA could potentially learn better linear feature set than KCCA and DCCA. Although MCCA [37] uses label information as the third view, it computes low-dimensional embedding ($p \leq min(d_x, d_y, q)$, q is the number of class label.) of sets of variables simultaneously in the same way as CCA. Obviously, because q is relatively small, i.e., $q \ll d_x, d_y$, the dimensionality of the shared feature space obtained by MCCA is much smaller than one of the feature-homogeneous space learned by CCA, leading to the loss of a great deal of information.

To confirm this viewpoint, ILCA, MCCA, KCCA, and DCCA are compared in classification performance. For MCCA, $p = min(d_x, d_y, q)$ (q is the number of class label). For KCCA [31] and DCCA [32], we tune the dimensionality k of the feature-isomorphic subspace on the candidate set $\{i \times 200 | i = 1, 2, 3, \ldots, 10\}$, and Gaussian kernel is used in KCCA.

We can see from Table 4 that it is very difficult for KCCA and DCCA to capture much complementary information without orthogonal constraints and maximum neighbourhood criterion as in ILCA, although they all can map the multi-view data into a feature-isomorphic subspace. This observation indicates that the linear features learned by ILCA are superior to the nonlinear features obtained by KCCA and DCCA.

Moreover, as shown in Table 4, it is very difficult for MCCA to capture much complementary information, although the label representations are used as the third view. This observation indicates that it is not very helpful to using directly the label information as a single view in MCCA.

5.6 Comparison of Completion Algorithms

Similar to SVT [27], NNLS [39], TNNR [40], ℓ_pAPG [41], RPCA [42], the proposed IDPC model is also a completion method based on the trace norm. But the major difference of IDPC with the other models lies in that it fully takes into account the identical distribution among different views. In addition, though kNN [46] and EM [45] use the mean value to replace missing value, some complementary information will be lost due to the lack of consideration of the semantic complementarity among different views. Moreover, though iSFS [47], [48] handles both feature-level and source-level analysis and MDBM [49] can be used to fill-in missing modalities given the observed ones, they still does not

TABLE 5
Classification Performance of Completion Algorithms in Terms of AUC

Dataset	Method									
	SVT	RPCA	kNN	EM	TNNR	NNLS	ℓ_pAPG	iSFS	MDBM	IDPC
UCI MFeat	0.9067	0.9161	0.9147	0.9044	0.8896	0.9161	0.8998	0.9234	0.9128	**0.9456**
ADNI	0.6721	0.6823	0.6393	0.6885	0.6557	0.6691	0.7049	0.6829	0.7259	**0.7526**
Wikipedia	0.7675	0.7148	0.7916	0.7905	0.7148	0.7984	0.7246	0.8019	0.7998	**0.8218**
MIR Flickr	0.7538	0.7412	0.7765	0.8005	0.7368	0.7927	0.7385	0.8187	0.8360	**0.8558**

address the identical distribution among different views. So the proposed IDPC model might be more favorable to complete missing view than the compared methods.

To validate this point, we first use the existing multi-view data to construct an incomplete matrix, in which the missing values refer to missing view of multi-view data and then apply SVT, NNLS, TNNR, ℓ_pAPG, RPCA, iSFS, MDBM, kNN, and EM to complete missing view. Then DCCA is applied to project the recovered multi-view data into a feature-isomorphic subspace to obtain the CR μ according to Eq. (7).

For our proposed framework, ILCA is performed first before IDPC is carried out. For the proposed IDPC model, the distance metric parameter π is set to the number of missing data, the trade off γ and the nonnegative constraint parameter ε are selected from the set $\{10^i | i = -2, -1, 0, 1, 2\}$. The parameter settings in the compared methods are the same as in their original literatures. The dimension k of the feature-isomorphic subspace in DCCA and ILCA are specified by the best values selected out by the experiment in Section 5.5.

It can be observed from Table 5 that IDPC shows an obvious advantage over the other methods. This comparison shows that, in contrast to the compared approaches, IDPC is highly effective on recovering missing view of multi-view data because it fully exploits both semantic complementarity and identical distribution among different views.

5.7 Comparison in Different Missing Rates

To test the performance of the proposed IDPC in different missing rates, we further compare the classification performances and reconstruction errors of IDPC with other completion methods such as iSFS [47], [48], RPCA [42], kNN [46], and MDBM [49] in the larger MIR Flickr dataset. We tune the missing rates on the set $\{10\%, 15\%, 20\%, 25\%\}$.

We can see from Fig. 7a that IDPC is superior to other completion methods in classification performance. This observation further confirms that IDPC can effectively recover the missing view of multi-view data. Nevertheless, with the increasing of missing rate, the performance of IDPC will degrade. Thus, IDPC also has some limitations that it need a certain number of existing samples to complete missing view.

Moreover, it can be observed from Fig. 7b that the reconstruction effect of IDPC is better than other completion methods. This is a strong proof that some complementary information will be lost for other completion methods due to the lack of consideration of the semantic complementarity among different views. However, with the increasing in missing rate, IDPC takes less obvious advantage over other completion methods. This once again shows that IDPC is based on a certain number of existing samples.

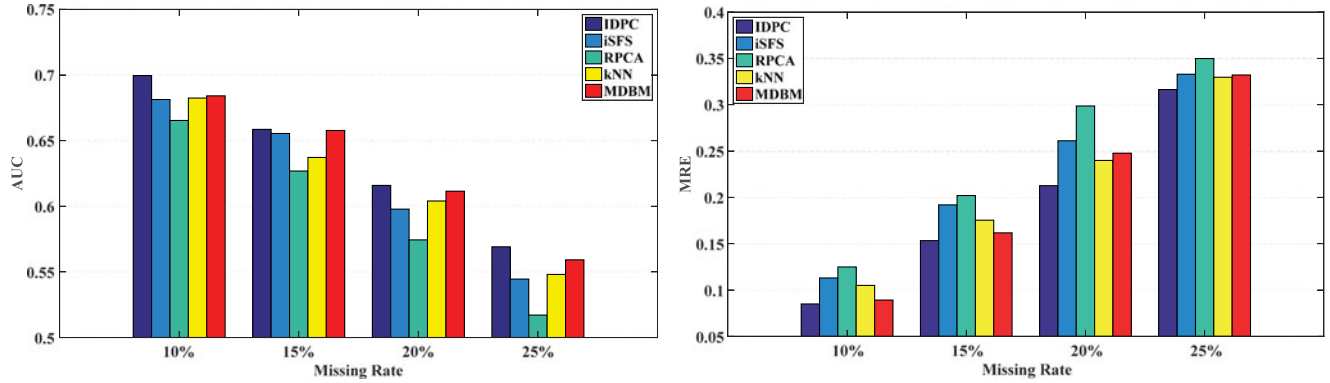
5.8 Analysis of Convergence Rate

In order to investigate the convergence behaviors of the proposed ILCA and IDPC, we plot the objective values of these two methods in different iterations on the UCI MFeat, ADNI, and Wikipedia datasets in Fig. 8.

We can observe that ILCA and IDPC converge very fast, especially at early iterations. This is consistent with our theoretical results in Section 3.3 and confirms that the proposed methods in Algorithms 1 and 2 can reach the local optimal objective value rapidly.

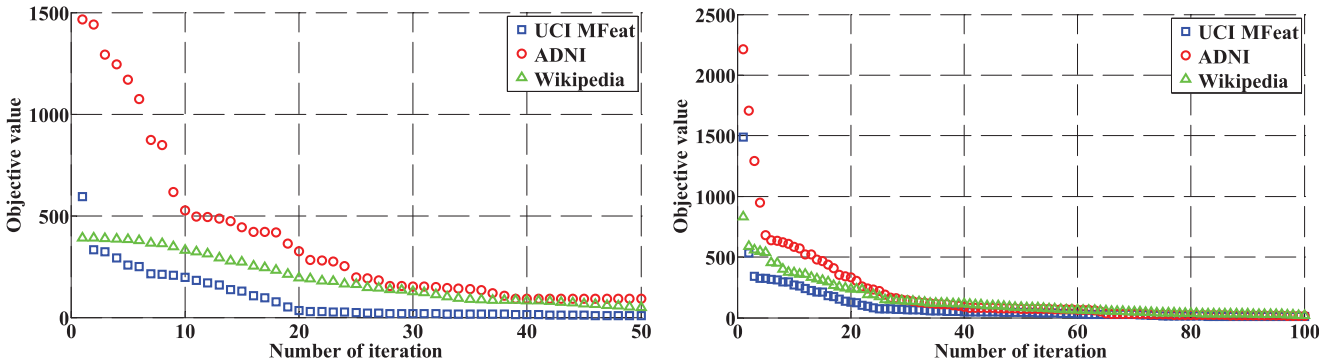
5.9 Evaluation of Multi-Pass Performance

The so-called multi-pass performance refers to the repetitive and alternate performing of ILCA and IDPC to improve the performance of multi-view learning. To verify the effect of multi-pass ILCA+IDPC, we compare the classification performance of ILCA, IDPC, and CCA in different times of



(a) Classification Performance (b) Reconstruction Error

Fig. 7. Comparison in different missing rates.



(a) Convergence Rate of ILCA (b) Convergence Rate of IDPC

Fig. 8. Analysis of convergence rate.

repetition. The experiment is performed on the COREL 5K dataset, in which the dataset is separated into a training set (50 percent of COREL 5K data), a validation set (30 percent of COREL 5K data), and a test set (20 percent of COREL 5K data). All the multi-view data in the validation set completely lack the representations from the DenseHue view, and 10 percent of multi-view data have missing DenseHue view in the test set.

The training samples are used by the ILCA model in Eq. (5) first to learn two optimal linear transformations A^* and B^*. Then the missing DenseHue view in the validation set is recovered by IDPC model in Eq. (8) based the learned linear transformations A^* and B^*. Furthermore, the training set and validation set are incorporated into a bigger set to train the ILCA model once again, and then the learned results are used to validate the classification performances of ILCA and IDPC in the test set. This process will be repeated six times. For CCA, the classification performance is verified in the completed test set recovered by IDPC in each repetition.

It can be shown from Fig. 9 that the classification performance of CCA rises up constantly with the increase of the number of repetition. This indicates that the recovered missing view in turn indeed improve the performance of multi-view learning, since the proposed framework enhances the qualities multi-view representations and increases the number of multi-view samples. Additionally, we also see from Fig. 9 that ILCA and IDPC can benefit from each other through multiple repetitive and alternate learning, and the procedure converges to a good level. Therefore, this procedure can be utilized to obtain better recovered results.

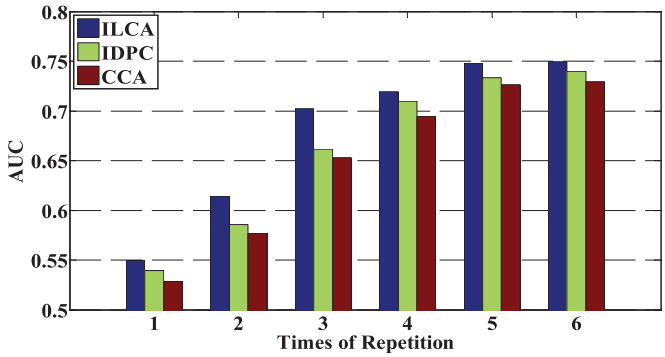


Fig. 9. Comparison of multi-pass performance.

5.10 Parameter Sensitivity of ILCA and IDPC

Due to limited space, the details are given in the supplementary material, available online.

6 Conclusion

In this paper, we have investigated the missing view problem in multi-view data. We developed a general feature-level framework to recover missing view to obtain CR for multi-view data. Within this framework, a feature-isomorphic subspace is learned by the proposed ILCA model to unfold the shared information from different views. We assume that missing view obeys normal distribution. Then, the expectation naturally corresponds to between-class differences, and the variance represents within-class differences. Therefore, the missing view data matrix can be modeled as a low-rank component plus a sparse contribution. Furthermore, we also proposed a IDPC model to recover missing view of multi-view data on the basis of the identical distribution constraint of missing view to the other available one in the feature-isomorphic subspace. Practically, the proposed ILCA and IDPC 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 missing view problem in multi-view data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 61601458, No. 61532005, No. 61572068), US National Science Foundation (IIS-1615035, IIS-1633359), and National Key Research and Development of China (No. 2016YFB0800404).

References

[1] M. Liu, D. Zhang, and D. Shen, "Ensemble sparse classification of alzheimer's disease," *NeuroImage*, vol. 60, no. 2, pp. 1106–1116, 2012.
[2] S. M. Landau, et al., "Associations between cognitive, functional, and FDG-pet measures of decline in ad and MCI," *Neurobiol. Aging*, vol. 32, no. 7, pp. 1207–1218, 2011.
[3] N. Rasiwasia, et al., "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 251–260.
[4] 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.
[5] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2160–2167.

[6] Z. Zhang, M. Zhao, and T. W. S. 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.
[7] B. Qian, X. Wang, J. Ye, and I. Davidson, "A reconstruction error based framework for multi-label and multi-view learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 594–607, Mar. 2015.
[8] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, Hillcrest, Hamilton, New Zealand, 1999.
[9] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.
[10] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu, "Feature fusion: Parallel strategy versus serial strategy," *Pattern Recognit.*, vol. 36, no. 6, pp. 1369–1381, 2003.
[11] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognit.*, vol. 38, no. 12, pp. 2437–2448, 2005.
[12] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[13] 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.
[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[15] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[16] 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.
[17] 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.
[18] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1473–1480.
[19] Y. Guo and M. Xiao, "Cross language text classification via subspace co-regularized multi-view learning," in *Proc. ACM Int. Conf. Mach. Learn.*, 2012, pp. 915–922.
[20] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1–2, pp. 397–434, 2013.
[21] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA, USA: Siam, 2000.
[22] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, no. 3, pp. 1817–1853, 2005.
[23] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Berlin, Germany: Springer Science & Business Media, 2004, vol. 87.
[24] A. Nemirovski, *Efficient methods in convex programming*. Lecture Notes, 2005.
[25] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.
[26] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. ACM Int. Conf. Mach. Learn.*, 2008, pp. 272–279.
[27] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[28] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: JHU Press, 2012, vol. 3.
[29] 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.
[30] H. Wold, "Partial least squares," *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: Wiley-Interscience, 2006.
[31] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computat.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[32] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ACM Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[33] Y. Guo, "Convex subspace representation learning from multi-view data," in *Proc. Nat. Conf. Artif. Intell.*, 2013, vol. 1, Art. no. 2.
[34] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence, "Multi-view transfer learning with a large margin approach," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1208–1216.
[35] L. Sun, S. Ji, and J. Ye, *Multi-Label Dimensionality Reduction*. Boca Raton, FL, USA: CRC Press, 2013.
[36] X. Zhu, H.-I. Suk, and D. Shen, "Multi-modality canonical feature selection for alzheimers disease diagnosis," in *Proc. Springer Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2014, pp. 162–169.
[37] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2002.
[38] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Department of Statistics, University of California, Berkeley, CA, USA, Rep. no. 688, 2005.
[39] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, no. 615–640, 2010, Art. no. 15.
[40] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2117–2130, Sep. 2013.
[41] G. Marjanovic and V. Solo, "On ℓ_p optimization and matrix completion," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5714–5724, Nov. 2012.
[42] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2011, Art. no. 11.
[43] M. Xiao and Y. Guo, "A novel two-step method for cross language representation learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1259–1267.
[44] Y. Guo, "Robust transfer principal component analysis with rank constraints," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1151–1159.
[45] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *J. Climate*, vol. 14, no. 5, pp. 853–871, 2001.
[46] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[47] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Multi-source learning with block-wise missing data for alzheimer's disease prediction," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 185–193.
[48] S. Xiang, et al.,"Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, 2014.
[49] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2949–2980, 2014.
[50] M. Van Breukelen, R. P. W. Duin, D. M. J. Tax, and J. E. Den Hartog, "Handwritten digit recognition by combined classifiers," Kybernetika, vol. 34, no. 4, pp. 381–386, 1998.
[51] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 902–909.
[52] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inform. Retrieval*, 2008, pp. 39–43.
[53] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Cambridge, MA, USA: Academic press, 2013.

Lei Zhang received the PhD degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2015. He is currently an assistant research fellow of the Institute of Information Engineering, Chinese Academy of Science (CAS), Beijing, China. He was a visiting scholar in the Department of Computer Science, College of Sciences, Old Dominion University, Norfolk, Virginia, USA, during 2014. His research interests include data mining, computer vision, machine learning, and multimedia content analysis.

Yao Zhao received the PhD degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an associate professor with 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 visited the Swiss Federal Institute of Technology in Lausanne (EPFL) in October 2015, and the University of Southern California from December 2017 to March 2018. 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 leads several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. He serves on the editorial boards of several international journals, including as an associate editor of *IEEE TCYB, IEEE SPL,* an area editor of *Signal Processing: Image Communication* (Elsevier), and an associate editor of *Circuits, System, and Signal Processing* (Springer). 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 fellow of IET.

Zhenfeng Zhu received the PhD degree in pattern recognition and intelligence system from the Institute of Automation, Chinese Academy of Sciences, in 2005. He was a visiting scholar with the Department of Computer Science and Engineering, Arizona State University, AZ, USA, in 2010. He is currently a professor with the Institute of Information Science, Beijing Jiaotong University. His general research interests include computer vision, machine learning, and data mining.

Dinggang Shen is the Jeffrey Houpt Distinguished Investigator, and a professor of radiology, Biomedical Research Imaging Center (BRIC), computer science, and biomedical engineering in the University of North Carolina at Chapel Hill (UNC-CH). He is currently directing the Center for Image Analysis and Informatics, the Image Display, Enhancement, and Analysis (IDEA) Lab in the Department of Radiology, and also the medical image analysis core in the BRIC. He was a tenure-track assistant professor at the University of Pennsylvanian (UPenn), and a faculty member at Johns Hopkins University. Dr. Shen's research interests include medical image analysis, computer vision, and pattern recognition. He has published more than 800 papers in the international journals and conference proceedings. He serves as an editorial board member for eight international journals. He has also served on the Board of Directors, The Medical Image Computing and Computer Assisted Intervention (MICCAI) Society, in 2012-2015. He is Fellow of IEEE and The American Institute for Medical and Biological Engineering (AIMBE).

Shuiwang Ji received the PhD degree in computer science from Arizona State University, Tempe, Arizona, in 2010. Currently, he is an associate professor in the School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington. His research interests include machine learning, data mining, computational neuroscience, and bioinformatics. He received the National Science Foundation CAREER Award in 2014. He is currently an associate editor for the *ACM Transactions on Knowledge Discovery from Data, IEEE Transactions on Neural Networks and Learning Systems,* and *BMC Bioinformatics*. He is an elected editorial board member of Data Mining and Knowledge Discovery and a senior member of IEEE.