Pattern Recognition 45 (2012) 2005–2018

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data

Xiaohong Chen [a,b], Songcan Chen [b,c,*], Hui Xue [d], Xudong Zhou [b]

[a] Department of Mathematics, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
[b] Department of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
[c] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
[d] School of Computer Science and Engineering, Southeast University, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 16 August 2011
Received in revised form
12 November 2011
Accepted 15 November 2011
Available online 25 November 2011

Keywords:
Multi-view data
Correlation analysis
Semi-supervised learning
Semi-paired learning
Dimensionality reduction

ABSTRACT

Canonical correlation analysis (CCA) is a popular and powerful dimensionality reduction method to analyze paired multi-view data. However, when facing semi-paired and semi-supervised multi-view data which widely exist in real-world problems, CCA usually performs poorly due to its requirement of data pairing between different views and un-supervision in nature. Recently, several extensions of CCA have been proposed, however, they just handle the semi-paired scenario by utilizing structure information in each view or just deal with semi-supervised scenario by incorporating the discriminant information. In this paper, we present a general dimensionality reduction framework for semi-paired and semi-supervised multi-view data which naturally generalizes existing related works by using different kinds of prior information. Based on the framework, we develop a novel dimensionality reduction method, termed as semi-paired and semi-supervised generalized correlation analysis (S^2GCA). S^2GCA exploits a small amount of paired data to perform CCA and at the same time, utilizes both the global structural information captured from the unlabeled data and the local discriminative information captured from the limited labeled data to compensate the limited pairedness. Consequently, S^2GCA can find the directions which make not only maximal correlation between the paired data but also maximal separability of the labeled data. Experimental results on artificial and four real-world datasets show its effectiveness compared to the existing related dimensionality reduction methods.

1. Introduction

In real world, we often meet with such a case that one object is represented by two or more types of features, e.g., gene can be represented by the genetic activity feature and text information feature [1], the same person has visual and audio features [2], each webpage can be represented by the text in the page and the hyperlinks jointly [3], CAD-catalogs are represented by some kind of 3D model like Bezier curves or polygon meshes and additional textual information like descriptions of technical [4]. This kind of data is usually called multimodal or multi-modality [1,2,5–8], multiple outlooks [9], multi-represented objection [4] or multi-view [3,10–14] data (for convenience, we will uniformly call them multi-view data hereafter). Analyzing such multi-view data to acquire useful information and knowledge has attracted more and more attentions recently. These works include dimensionality reduction (DR) [7,8,14–21], regression [22] and clustering [1,4,11]. In this paper, we focus on DR for multi-view data with the aim to avoid the curse of dimensionality [23] and overfitting brought by high dimensionality for good generalization [15], i.e., learning the appropriate low-dimensional representations for high dimensional data for subsequent task.

In recent years, a number of efficient algorithms [7,8,14–21] emerged to address this problem for discovering inherent structures and relations among different views. Among all the methods, canonical correlation analysis (CCA) [16–18] is the most widely used one. It works with two sets of related variables ($\mathbf{x}$, $\mathbf{y}$), and aims to find the directions that maximize the correlation between the two sets of projected representations in the low-dimensional space. In its implementation, CCA requires the data be rigorously paired or one-to-one correspondence among different views due to its correlation definition. However, such requirement is usually not satisfied in real life due to various reasons, e.g., (1) different sampling frequencies of sensors acquiring data or sensor faulty in an audio-video system, which result in

* Corresponding author at: Department of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China. Tel.: +86 25 84892452.
E-mail address: s.chen@nuaa.edu.cn (S. Chen).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.11.008

non-synchronicity between signals from different channels and even the missing of samples of certain views so that the multi-view data cannot keep one-to-one correspondence any more [7]. (2) Even having sufficient individual-view data, pairing them is still difficult, time consuming, even expensive since needing the efforts from experienced human annotators. Meanwhile, unpaired multi-view data are relatively easier to be collected. So we are often given only a few paired and a lot of unpaired multi-view data. We refer such data as semi-paired multi-view data. In literature, it is also named as weakly-paired multi-view data [6] or partially-paired multi-view data [19]. The common approaches to analyze such type of data include: (1) directly discarding unpaired data and performing correlation analysis just on the paired data, which usually results in overfitting on the given data and poor generalization for unseen samples especially when the paired data is scarce. (2) Creating synthetic samples in terms of certain criterion with the aim to generate paired multi-view data for correlation analysis [24]. These methods cannot achieve reasonable improvement due to not incorporating the prior information (such as clustering hypothesis and manifold hypothesis) of the data. Now, the key point to address this problem is how to utilize the meaningful prior information hidden in additional unpaired data. Most recently, some improved algorithms of CCA that can effectively deal with semi-paired multi-view data have emerged. Typically, Blaschko et al. [20] proposed semi-supervised Laplacian regularization of kernel canonical correlation (SemiLRKCCA) to find a set of highly correlated directions by exploiting the intrinsic manifold geometry structure of all data (paired and unpaired). Another paradigm is SemiCCA [15]. It essentially resembles the manifold regularization [25], i.e., using the global structure of the whole training data including both paired and unpaired samples to regularize CCA. Consequently, SemiCCA seamlessly bridges CCA and principal component analysis (PCA) [26,27], and inherits some characteristics of both PCA and CCA. It is necessary to mention that the actual meaning of "semi-" in SemiCCA and SemiLRKCCA is "semi-paired" rather than "semi-supervised" in popular semi-supervised learning literature [28,29]. Most recently, Gu et al. [19] proposed partially paired locality correlation analysis (PPLCA), which effectively deals with the semi-paired scenario of wireless sensor network localization by virtue of the combination of the neighborhood structure information in data. SemiCCA, SemiLRKCCA and PPLCA all cater well for semi-paired multi-view scenario and thus achieve better empirical results than CCA through preserving original paired information and deeply utilizing the structure information simultaneously.

As we have known, discriminative information is quite important for DR serving the classification task. However, SemiCCA, SemiLRKCCA and PPLCA are unsupervised DR methods, thus only concerning the between-view correlation embedded the structure information of each view is generally not enough for better classification accuracy. Concretely, SemiLRKCCA utilizes the graph Laplacians constructed through the unsupervised within-view k-nearest neighbors with regardless of the labeled or unlabeled data. SemiCCA employs unsupervised PCA as within-view regularization terms to do semi-paired learning. PPLCA replaces total mean with the neighborhood means into the formulation of CCA in each view such that PPLCA can incorporate the unpaired data information. Due to not exploiting the class information, the above three methods unavoidably result in the limitation of recognition performance. To overcome the limitation, Sun et al. [8] proposed the discriminative canonical correlation analysis (DCCA) for supervised multi-view data. DCCA aims to obtain DR with discrimination by maximizing the within-class correlation while minimizing the between-class correlation. Next, Sun et al. [7] further extended DCCA to the fully supervised and semi-paired scenario and developed the DCCA with Missing Samples (DCCAM). Borrowing the idea of DCCA, Peng et al. [21] proposed the local discrimination CCA(LDCCA) by incorporating the idea of local discriminant analysis [30] into CCA. Specifically, LDCCA takes local discriminant information of each view data into account for defining the local between-class covariance and local within-class covariance matrices and thus attempts to achieve effective between-class separation by maximizing local within-class correlations and minimizing local between-class correlations simultaneously. Essentially, the common key of above three methods is to construct the within-class and between-class correlation matrices. However, such a construction is only suit for the case that the class label-aligned discriminant information is given for each view data, hence their performance will degrade greatly when just few labeled data can be available.

Although DCCA, DCCAM and LDCCA can work reasonably well in fully supervised case, in many real-world applications such as image classification, web page classification and protein function prediction, labeled samples are harder to be collected than unlabeled samples since the labeling process is relatively expensive and time consuming. Thus, a semi-supervised(SSL) scenario occurred [28,29,31]. Recently, the multi-view DR in semi-supervised scenario has received increasing attention as a learning paradigm. For example, Foster et al. [22] performed CCA first for unlabeled data and then least squares regression for given labeled data in the CCA-generated lower dimensional subspace. Kursun and Alpaydin [32] proposed a Semi-supervised CCA(SCCA). In its implementation, a key ingredient is to rebuild two-view data and then perform correlation analysis, i.e., first for the one view, SCCA keeps the other view when class label is absent, otherwise replaces the samples by the corresponding class-centers, and then performs semi-supervised DR for this view data, the same process is repeated for the other view. Most recently, Hou et al. [14] developed a multiple view semi-supervised dimensionality reduction (MVSSDR) method with the discriminative information from given within-view pairwise must-link and cannot-link constraints (similar to SSDR [33]). Here a pair of "must-link" samples implies that they belong to the same classes of the same view and a pair of "cannot-link" samples implies that they belong to different classes of the same view. MVSSDR exploits the disparate structures and different statistical properties of different views to achieve better performance than SSDR which is only fit for all the concatenated representations of all the views. The above two methods [14,32] deal with a fully paired and semi-supervised multi-view case. Undoubtedly, such a strict pairing requirement among views naturally limits their applications in real world.

With the successive emergence of new application problems and the rapid development of data collection and processing techniques, multi-view data is more complex and diverse, i.e., between-view data may be paired or unpaired, and within-view data may be labeled or unlabeled simultaneously. According to whether the multi-view data under study is fully paired or not, the existing corresponding DR methods can be roughly categorized into paired ones (CCA, SCCA, MVSSDR, DCCA and LDCCA) and semi-paired ones (SemiCCA, SemiLRCCA, PPLCA and DCCAM). The former can further be divided into unsupervised, semi-supervised and supervised ones. The latter is subdivided into supervised and unsupervised ones. Table 1 summarizes the characteristics of the above related methods in terms of pairing information, discriminative information and structural information used.

From the "paired information" and "discriminative information" columns of Table 1, we observe that, there is no DR method to deal with semi-paired and semi-supervised multi-view data. Furthermore, we find that besides the paired information, both discriminative information and structural information are

Table 1
Comparison of CCA, SemiCCA, SemiLRCCA, DCCA, LDCCA, DCCAM, MVSSDR, SCCA and PPLCA.

	Paired information		Discriminative information			Structural information	
	Paired	Semi-paired	Unsupervised	Semi-supervised	Supervised	Local[a]	Global
CCA [16–18]	✓		✓				
SemiCCA [15]		✓	✓				✓
SemiLRCCA [20]		✓	✓			✓	
DCCA [8]	✓				✓		
LDCCA [21]	✓				✓	✓	
DCCAM [7]		✓			✓		
MVSSDR [14]	✓			✓			
SCCA [32]	✓			✓			
PPLCA [19]		✓	✓			✓	

[a] "Local" means to use the data neighborhood information (e.g., manifold information) to construct scatter matrix.

meaningful for DR. Consequently, in this paper, we try to design a general framework called semi-paired and semi-supervised dimensionality reduction(S^2DR), especially for multi-view data by combining the semi-paired correlation analysis and the semi-supervised DR into a unified framework, which takes not only the discriminant information but also the within-view structural (local and global) information into account.

Based on our S^2DR framework, we put forward a novel multi-view DR algorithm, and refer it as semi-paired and semi-supervised generalized correlation analysis (S^2GCA). S^2GCA makes as maximal correlation as possible by performing CCA on given paired data, while preserves geometric structure of unlabeled data as sufficiently as possible and separates labeled data from different classes as far as possible. Consequently, S^2GCA can seek the desirable directions which not only have maximal correlation for paired data but also reflect the separability for the labeled data. Experimental results on a toy dataset and four publicly-available datasets including semi-supervised learning data (SSL) [34,35], Multiple Feature Database(MFD) [36], WebKB dataset [37] and advertisement dataset (Ads) [38] show its effectiveness compared to the related DR methods.

Finally, it is worthwhile to highlight several advantages of our S^2GCA as follows:

(1) To the best of our knowledge, S^2GCA is the first DR method to deal with the semi-paired and semi-supervised multi-view data. A general framework is further constructed in such scenario including SemiCCA and SemiLRCCA as its special cases.
(2) Different from unsupervised SemiLRCCA and SemiCCA which just utilize global or local (manifold) structure of each view data, S^2GCA fuses not only the global and local structural information but also the discriminative information into a single objective function, consequently, making it more effective and flexible in modeling the given data since not limited to whether paired and/or unpaired data should have labels.
(3) Compared with the traditional semi-supervised DR methods which can only be applicable in single-view data, S^2GCA can perform semi-supervised learning on two or more views data simultaneously and thus can capture the latent knowledge in data more sufficiently. Compared to existing multi-view semi-supervised methods such as SCCA and MVSSDR which work on semi-supervised and fully paired multi-view data, S^2GCA is free of the limitation of the correspondence between different views to great extent.
(4) Compared with the works on supervised multi-view data, such as DCCA, DCCAM and LDCCA, S^2GCA copes with semi-supervised multi-view data, which is more general and more applicable.
(5) S^2GCA characterizes the optimization objective as a generalized eigenvalue problem, which can be solved simply and efficiently as CCA, SCCA, DCCA, DCCAM, LDCCA, PPLCA, SemiCCA and SemiLRCCA.

The rest of the paper is organized as follows. Section 2 gives a brief review of the related works. In Section 3, we put forward a general DR framework for multi-view data, semi-paired and semi-supervised dimensionality reduction (S^2DR). We then utilize the S^2DR framework as a general platform to design S^2GCA algorithm, including the motivation, formulation and solution in Section 4. Then we present the experimental results and analysis both on toy data and real-world datasets including SSL, MFD, WebKB and Ads databases in Section 5. The conclusions and future works are listed in Section 6.

2. Related works

2.1. CCA: canonical correlation analysis

Given n pairs of pairwise samples $\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)\}$ centralized by subtracting the total samples means from each sample. Let $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]\in R^{p\times n}$ and $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_n]\in R^{q\times n}$. CCA [16–18] attempts to find a set of projections (or directions) $\mathbf{w}_x$ and $\mathbf{w}_y$ for each view such that the correlation between $\mathbf{w}_x^T\mathbf{x}$ and $\mathbf{w}_y^T\mathbf{y}$ is maximized. The corresponding objective can be described as follows:

$$\max_{\mathbf{w}_x,\mathbf{w}_y}\frac{\mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x}\sqrt{\mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y}} \tag{1}$$

Evidently, it can be expressed by the following equality constrained optimization problem [18]:

$$\begin{aligned}\max_{\mathbf{w}_x,\mathbf{w}_y}\ & \mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{w}_y\\ \text{s.t.}\ & \mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x=1\\ & \mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y=1\end{aligned} \tag{2}$$

By the Lagrange technique [18], the optimization of (2) boils down to solving a generalized eigenvalue problem

$$\begin{bmatrix}0 & \mathbf{X}\mathbf{Y}^T\\ \mathbf{Y}\mathbf{X}^T & 0\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\mathbf{X}\mathbf{X}^T & 0\\ 0 & \mathbf{Y}\mathbf{Y}^T\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix} \tag{3}$$

Further, we can jointly get two projection matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ consisting of the top r ($\le\min(p,q)$) generalized eigenvectors of (3). In this way, a common dimensionality reduced subspace maximizing the between-view correlation is established.

In fact, CCA is difficult to work effectively for nonlinearly-correlated data due to its linearity in nature. Consequently, kernel

canonical correlation analysis (KCCA) [39] is developed by kernelizing CCA to effectively compensate for this drawback.

2.2. SemiLRKCCA: semi-supervised Laplacian regularization of KCCA

Given a set of data $\{\mathbf{x}_1,\ldots,\mathbf{x}_{n_p},\mathbf{x}_{n_p+1},\ldots,x_{n_x}\}$ from $\mathbf{X}$-view and the other set of data $\{\mathbf{y}_1,\ldots,\mathbf{y}_{n_p},\mathbf{y}_{n_p+1},\ldots,\mathbf{y}_{n_y}\}$ from $\mathbf{Y}$-view, respectively, where $(\mathbf{x}_i,\mathbf{y}_i)$, $i=1,2,\ldots,n_p$ are paired ones and the rest are unknown whether to be paired. $n_x(n_y)$ is the total number of samples in $\mathbf{X}$-view ($\mathbf{Y}$-view). For $\mathbf{X}$-view, we denote the paired data matrix $\tilde{\mathbf{X}}=[\mathbf{x}_1,\ldots,\mathbf{x}_{n_p}]\in R^{p\times n_p}$ and the matrix including all data with and without correspondences $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_{n_p},\mathbf{x}_{n_p+1},\ldots,\mathbf{x}_{n_x}]\in R^{p\times n_x}$. Similarly for $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$. According to its definition, CCA (KCCA) is not suitable for such a semi-paired or partially-paired scenario. In order to overcome this shortcoming, Blaschko et al. [20] applied the manifold regularization technique [25] to KCCA (CCA) and consequently developed a semi-supervised Laplacian regularization of KCCA (SemiLRKCCA) to tackle such scenario. SemiLRKCCA can be built by optimizing the following problem:

$$\max_{\alpha,\beta}\frac{\alpha^T\mathbf{K}_{X\tilde{X}}\mathbf{K}_{\tilde{Y}Y}\beta}{\sqrt{\alpha^T(\mathbf{K}_{X\tilde{X}}\mathbf{K}_{\tilde{X}X}+\mathbf{R}_X)\alpha\beta^T(\mathbf{K}_{Y\tilde{Y}}\mathbf{K}_{\tilde{Y}Y}+\mathbf{R}_Y)\beta}} \tag{4}$$

where $\mathbf{R}_X=\varepsilon_X\mathbf{K}_{XX}+(\gamma_X/n_x^2)\mathbf{K}_{XX}\mathbf{L}_X\mathbf{K}_{XX}$, $\mathbf{L}_x$ is the empirical graph Laplacian as defined in manifold learning [25], which is constructed from the n_x samples both of paired and unpaired. The involved kernel matrices in (4) for $\mathbf{X}$-view are defined as $\mathbf{K}_{XX}=\phi_x(\mathbf{X})^T\phi_x(\mathbf{X})$, $\mathbf{K}_{X\tilde{X}}=\phi_x(\mathbf{X})^T\phi_x(\tilde{\mathbf{X}})$, $\mathbf{K}_{\tilde{X}X}=\phi_x(\tilde{\mathbf{X}})^T\phi_x(\mathbf{X})$ and $\mathbf{K}_{\tilde{X}\tilde{X}}=\phi_x(\tilde{\mathbf{X}})^T\phi_x(\tilde{\mathbf{X}})$, where $\phi_x(\cdot):R^p\to R$ is the kernel function especially defined for $\mathbf{X}$-view. Kernel matrices for $\mathbf{Y}$-view are defined analogously.

It should be pointed that SemiLRKCCA invloves 13 parameters in total to be tuned in the learning process, consequently, resulting in high learning cost and even inapplicable in actual applications. In order to keep the consistency with our proposal later and discover the intrinsic characteristic of SemiLRKCCA, we specially reduce nonlinear SemiLRKCCA to its linear version by virtue of the following equations and rename it as SemiLRCCA

$$\alpha^T\mathbf{K}_{X\tilde{X}}\mathbf{K}_{\tilde{Y}Y}\beta=\alpha^T\mathbf{X}^T\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\mathbf{Y}\beta=\mathbf{w}_x^T\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\mathbf{w}_y \tag{5}$$

$$\alpha^T\mathbf{K}_{X\tilde{X}}\mathbf{K}_{\tilde{X}X}\beta=\alpha^T\mathbf{X}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{X}\beta=\mathbf{w}_x^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{w}_x \tag{6}$$

$$\alpha^T\mathbf{R}_X\alpha=\alpha^T\left(\varepsilon_X\mathbf{K}_{XX}+\frac{\gamma_X}{n_x^2}\mathbf{K}_{XX}\mathbf{L}_X\mathbf{K}_{XX}\right)\alpha=\mathbf{w}_x^T\left[\varepsilon_X\mathbf{I}+\frac{\gamma_X}{n_x^2}\mathbf{X}\mathbf{L}_x\mathbf{X}^T\right]\mathbf{w}_x \tag{7}$$

As a result, we get the following optimization problem corresponding to (4):

$$\max_{\mathbf{w}_x,\mathbf{w}_y}\frac{\mathbf{w}_x^T\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T+\varepsilon_X\mathbf{I}+(\gamma_X/n_x^2)\mathbf{X}\mathbf{L}_x\mathbf{X}^T)\mathbf{w}_x}\sqrt{\mathbf{w}_y^T(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T+\varepsilon_Y\mathbf{I}+(\gamma_Y/n_y^2)\mathbf{Y}\mathbf{L}_y\mathbf{Y}^T)\mathbf{w}_y}} \tag{8}$$

From the denominator of (8), we can conveniently observe its embedding way for the local structural information and reduce its motivation of regularization clearer. Furthermore, SemiLRCCA shows the conversion from the original CCA (paired correlation analysis) to semi-paired correlation analysis. Likewise, through introducing the equality constraint, above optimization problem (8) can be reformulated as

$$\begin{aligned}\max\ & \mathbf{w}_x^T\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\mathbf{w}_y\\ \text{s.t.}\ & \mathbf{w}_x^T\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T+\varepsilon_X\mathbf{I}+\frac{\gamma_X}{n_x^2}\mathbf{X}\mathbf{L}_x\mathbf{X}^T\right)\mathbf{w}_x=1\\ & \mathbf{w}_y^T\left(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T+\varepsilon_Y\mathbf{I}+\frac{\gamma_Y}{n_y^2}\mathbf{Y}\mathbf{L}_y\mathbf{Y}^T\right)\mathbf{w}_y=1\end{aligned} \tag{9}$$

Again with the Lagrange technique, we obtain the following generalized eigenvalue problem

$$\begin{bmatrix}0 & \tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\\ \tilde{\mathbf{Y}}\tilde{\mathbf{X}}^T & 0\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T+\varepsilon_X\mathbf{I}+\frac{\gamma_X}{n_x^2}\mathbf{X}\mathbf{L}_x\mathbf{X}^T & 0\\ 0 & \mathbf{w}_y^T\left(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T+\varepsilon_Y\mathbf{I}+\frac{\gamma_Y}{n_y^2}\mathbf{Y}\mathbf{L}_y\mathbf{Y}^T\right)\mathbf{w}_y\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix} \tag{10}$$

2.3. SemiCCA: semi-supervised learning of canonical correlation analysis

With the aim to avoid overfitting resulted from CCA when paired data is scarce, SemiCCA [15] gives the following direct eigenvalue problem with no concrete optimization objective:

$$\begin{bmatrix}(1-\mu)\mathbf{C}_{xx} & \mu\tilde{\mathbf{C}}_{xy}\\ \mu\tilde{\mathbf{C}}_{yx} & (1-\mu)\mathbf{C}_{yy}\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\mu\tilde{\mathbf{C}}_{xx}+(1-\mu)I_p & 0\\ 0 & \mu\tilde{\mathbf{C}}_{yy}+(1-\mu)I_q\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix} \tag{11}$$

Likewise, for consistency and contrast with our work later, from (11), we also deduce its corresponding objective function as follows:

$$\begin{aligned}\max_{\mathbf{w}_x,\mathbf{w}_y}\ & 2\mu\mathbf{w}_x^T\tilde{C}_{xy}\mathbf{w}_y+(1-\mu)(\mathbf{w}_x^TC_{xx}\mathbf{w}_x+\mathbf{w}_y^TC_{yy}\mathbf{w}_y)\\ \text{s.t.}\ & \mu(\mathbf{w}_x^T\tilde{C}_{xx}\mathbf{w}_x+\mathbf{w}_y^T\tilde{C}_{yy}\mathbf{w}_y)+(1-\mu)(\mathbf{w}_x^T\mathbf{w}_x+\mathbf{w}_y^T\mathbf{w}_y)=1\end{aligned} \tag{12}$$

where $\mathbf{C}_{xx}=(1/n_x)\mathbf{X}\mathbf{X}^T$, $\mathbf{C}_{yy}=(1/n_y)\mathbf{Y}\mathbf{Y}^T$, $\tilde{\mathbf{C}}_{xx}=(1/n_p)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$, $\tilde{\mathbf{C}}_{yy}=(1/n_p)\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$ and $\tilde{\mathbf{C}}_{xy}=(1/n_p)\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T$. The first term of (12) ensures the correlation between the paired data to be maximized and the second term ensures the covariances of $\mathbf{X}$ and $\mathbf{Y}$ to be maximized, respectively. Evidently, SemiCCA combines CCA just applicable in the paired data and PCA [26,27] of all the data with a tradeoff parameter μ. Incorporating the global structure of the data into CCA has been shown better than just relying on the paired information provided by a small amount of paired samples [15].

2.4. DCCA: discriminative canonical correlation analysis

Given n pairs of mean-normalized paired samples $\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)\}$ coming from c classes, DCCA [8] aims to seek a set of projection vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ such that the within-class correlation is maximized and the between-class correlation is minimized. It has been formulated as the following optimization problem:

$$\begin{aligned}\max_{\mathbf{w}_x,\mathbf{w}_y}\ & \mathbf{w}_x^T(\mathbf{C}_w-\eta\mathbf{C}_b)\mathbf{w}_y\\ \text{s.t.}\ & \mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x=\mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y=1\end{aligned} \tag{13}$$

where $\mathbf{C}_w=\sum_{i=1}^{c}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\mathbf{x}_k^{(i)}\mathbf{y}_l^{(i)T}$ is the within-class correlation matrix and $\mathbf{C}_b=\sum_{i=1}^{c}\sum_{\substack{j=1\\ j\neq i}}^{c}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}\mathbf{x}_k^{(i)}\mathbf{y}_l^{(j)T}$ is the between-class correlation matrix, η is a balance factor which trades-off $\mathbf{C}_w$ and $\mathbf{C}_b$. Due to the fact that $\mathbf{C}_b=-\mathbf{C}_w$ in supervised case, DCCA can be shortly expressed as

$$\begin{aligned}\max_{\mathbf{w}_x,\mathbf{w}_y}\ & \mathbf{w}_x^T\mathbf{C}_w\mathbf{w}_y\\ \text{s.t.}\ & \mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x=\mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y=1\end{aligned} \tag{14}$$

Using the Lagrange multiplier technique, (14) is easily turned into the following generalized eigenvalue problem:

$$\begin{bmatrix}0 & \mathbf{C}_w\\ \mathbf{C}_w^T & 0\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\mathbf{X}\mathbf{X}^T & 0\\ 0 & \mathbf{Y}\mathbf{Y}^T\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix} \tag{15}$$

As a result, DCCA can be established by solving (15).

2.5. SCCA: semi-supervised canonical correlation analysis

Given a fully paired two-view dataset $\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_{n_p},\mathbf{y}_{n_p})\}$. Each view contains both labeled and unlabeled samples. For realizing its semi-supervised DR, SCCA [32] pre-processes the dataset according to class information partially available: for either view, it keeps the samples without class-label of the other view unchanged and otherwise replaces them by the corresponding class centers, then performs CCA on the pre-processed dataset. Similar procedure is repeated for the second view. SCCA is finally devised by such two CCAs together.

3. Semi-paired and semi-supervised dimensionality reduction (S^2DR): a general dimensionality reduction framework for multi-view data

Up to date, a number of DR approaches for multi-view data have been proposed. Although with different motivations and different objectives, all these approaches adopt the between-view correlation as a common measure and seek the low-dimensional representations for original high-dimensional data by optimizing the measure with aiming to preserve the maximal between -view correlation along with other prior knowledge. So, in this section, we attempt to establish a unified DR framework for semi-paired and semi-supervised multi-view data and thus provide a common perspective in understanding the relationship between these algorithms and in the next section, further develop a new DR method from the framework.

Given semi-paired and semi-supervised multi-view data, in addition to the paired information and the discriminant information, the (local and global) structural information implicitly in each view is also important for DR. For semi-supervised multi-view data, existing DR methods [14,32] make use of the discriminant information (class label or pairwise constraints) to improve between-class separability in the low-dimensional space. For semi-paired multi-view data, existing DR methods [7,15,19,20] usually embed the latent structural information of data in the form of regularization into the classical CCA's objective function. Their experimental results showed that both discriminant information and structural information are quite important for DR. Encouraged by their successes, we develop a unified DR framework in the form of a common objective function for such a semi-paired and semi-supervised scenario and term it as semi-paired and semi-supervised dimensionality reduction(S^2DR). Specifically, we define the following objective $J(\mathbf{w}_x,\mathbf{w}_y)$

$$\max J(\mathbf{w}_x,\mathbf{w}_y)=J_{paired}(\mathbf{w}_x,\mathbf{w}_y)+\eta_1 J_{supervised}(\mathbf{w}_x,\mathbf{w}_y)+\eta_2 J_{structured}(\mathbf{w}_x,\mathbf{w}_y) \quad (16)$$

Then we can characterize all of the above correlation-based methods in a unified form. In (16), $J_{paired}(\mathbf{w}_x,\mathbf{w}_y)$, $J_{supervised}(\mathbf{w}_x,\mathbf{w}_y)$ and $J_{structured}(\mathbf{w}_x,\mathbf{w}_y)$ measure individual gains, respectively, for the between-view paired information, the within-view discriminant information and the within-view structural information and η_1,η_2 are parameters to tune the balance among the three kinds of prior knowledge. In fact, both $J_{supervised}(\mathbf{w}_x,\mathbf{w}_y)$ and $J_{structured}(\mathbf{w}_x,\mathbf{w}_y)$ can be treated as regularized terms and represent different prior information.

From (16), we can find that all the above algorithms can be subsumed in S^2DR framework. Concretely,

(1) SemiCCA, SemiLRCCA and PPLCA share a common objective consisting of the first and third terms of (16), however, their major difference lies in that SemiCCA focuses more on global structure of each view, while SemiLRCCA and PPLCA emphasize more local structure of each view.

(2) For DCCA, DCCAM, LDCCA, MVSSDR and SCCA, though their objective functions are respectively formulated as a single term, in fact, the term can accordingly be decomposed into the first and the second terms of (16). It needs to mention that the first three methods deal with fully supervised and fully paired multi-view data while the last two methods cope with semi-supervised and fully paired scenario.

4. Semi-paired and semi-supervised generalized correlation analysis (S^2GCA)

Based on the S^2DR framework, we further develop a new algorithm, which incorporates both the discriminative information and the structural information into CCA (objective function), to cater for such a new semi-paired and semi-supervised case, called semi-paired and semi-supervised generalized correlation analysis (S^2GCA).

4.1. Motivation

For given semi-paired and semi-supervised multi-view data $\{\mathbf{x}_1,\ldots,\mathbf{x}_{n_p},\mathbf{x}_{n_p+1},\ldots,x_{n_x}\}$ and $\{\mathbf{y}_1,\ldots,\mathbf{y}_{n_p},\mathbf{y}_{n_p+1},\ldots,\mathbf{y}_{n_y}\}$, as in Semi-LRCCA. Each view only contains a few labeled samples coming from c classes and abundant unlabeled samples, as in semi-supervised learning. Our goal is to seek projection vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ which make not only the between-view correlation as maximal as possible but also the within-view separability among different classes as maximal as possible. Towards this end, we manage to mine the prior knowledge hidden in the data to obtain relatively reasonable projections by introducing the idea of semi-supervised LFDA (SELF) [34] to the correlation analysis. SELF is the semi-supervised extension of LFDA [35] by adding the PCA objective into the objective of LFDA. It constructs several crucial Laplacian matrices to reflect both the structural information and the discriminative information according to k-nearest neighbors. We first perform similar semi-supervised learning on **X**-view and **Y**-view, as in SELF, then embed the Laplicians matrices constructed from each view to CCA's objective as the regularization terms, and finally form an optimization problem subject to certain specific constraints. In the following subsection, we detail S^2GCA in the following problem formulation and its solving.

4.2. Formulation

Owing to the formulation involves several crucial matrices describing structural and discriminant information of given data, we will construct and introduce them in the next subsections by virtue of the idea of SELF [34].

4.2.1. Construct local within-class matrix and local between-class matrix

For avoiding notational confusion, we denote $\mathbf{X}_l=[\mathbf{x}_1,\ldots,\mathbf{x}_l]$ to be the labeled data coming from c different classes. By the graph embedding [40], we define the local within-class matrix $\mathbf{S}_{lw}^X$ and the local between-class matrix $\mathbf{S}_{lb}^X$ to reflect the local discriminative information

$$\mathbf{S}_{lw}^X=\frac{1}{2}\sum_{i,j=1}^{l}(\mathbf{S}_w^X)_{ij}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T \quad (17)$$

$$\mathbf{S}_{lb}^X=\frac{1}{2}\sum_{i,j=1}^{l}(\mathbf{S}_b^X)_{ij}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T \quad (18)$$

where $\mathbf{S}_w^X$ and $\mathbf{S}_b^X$ are matrices respectively having the (i,j)th element defined by

$$(\mathbf{S}_w^X)_{ij}=\begin{cases}\mathbf{A}_{ij}/n_t & \text{if } \mathbf{x}_i,\ \mathbf{x}_j \text{ belong to the class } t\\ 0 & \text{otherwise}\end{cases} \tag{19}$$

$$(\mathbf{S}_b^X)_{ij}=\begin{cases}\mathbf{A}_{ij}(1/l-1/n_t) & \text{if } \mathbf{x}_i,\ \mathbf{x}_j \text{ belong to the class } t\\ 1/l & \text{otherwise}\end{cases} \tag{20}$$

Here n_t is the number of labeled samples in class $t(\sum_{t=1}^{c} n_t=l)$ and $\mathbf{A}_{ij}$ is the affinity between $\mathbf{x}_i$ and $\mathbf{x}_j$ based on local scaling heuristic [41] and defined as

$$\mathbf{A}_{ij}=\begin{cases}\exp(-\|\mathbf{x}_j-\mathbf{x}_i\|^2/\sigma_i\sigma_j) & \text{if } \mathbf{x}_i,\ \mathbf{x}_j \text{ belong to the same class}\\ 0 & \text{otherwise}\end{cases} \tag{21}$$

Evidently, $\mathbf{A}_{ij}$ is large if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" and small if "far apart". The parameter $\sigma_i=\|\mathbf{x}_i^{(k)}-\mathbf{x}_i\|$ represents the local scaling around $\mathbf{x}_i$, where $\mathbf{x}_i^{(k)}$ is the kth nearest neighbor of the labeled sample $\mathbf{x}_i$ among all the samples of **X**-view.

4.2.2. Construct regularized local within-class matrix and regularized local between-class matrix

One the one hand, considering the likely instability of $\mathbf{S}_{lb}^X$ in case of a few labeled data, we introduce the total scatter matrix $\mathbf{S}_T^X$ to stabilize $\mathbf{S}_{lb}^X$ as a regularization term, thus forming regularized local between-class scatter matrix $\mathbf{S}_{rlb}^X$ (22) as defined in SELF

$$\mathbf{S}_{rlb}^X=(1-\delta)\mathbf{S}_{lb}^X+\delta\mathbf{S}_T^X \tag{22}$$

where $\mathbf{S}_T^X=\mathbf{X}\mathbf{X}^T-n_x\overline{\mathbf{x}}\overline{\mathbf{x}}^T$ and the sample mean $\overline{\mathbf{x}}=(1/n_x)\sum_{i=1}^{n_x}\mathbf{x}_i$ are, calculated from both label and unlabeled data of **X**-view, and $\delta\in[0,1]$ is a trading-off parameter between the local discriminant structure and the global structure. Now maximizing $\mathbf{S}_{rlb}^X$ implies that unlabeled data separate from each other to preserve the global structure, and meanwhile the sample pairs in different classes farther apart.

On the other hand, the identity matrix $\mathbf{I}_p$ is added to $\mathbf{S}_{lw}^X$ as a regularization term for avoiding its instability which may suffer from ill-conditioned, as a result, forming regularized local within-class scatter matrix $\mathbf{S}_{rlw}^X$

$$\mathbf{S}_{rlw}^X=(1-\delta)\mathbf{S}_{lw}^X+\delta\mathbf{I}_p \tag{23}$$

For **Y**-view, the regularized local between-class scatter matrix $\mathbf{S}_{rlb}^Y$ and regularized local within-class scatter matrix $\mathbf{S}_{rlw}^Y$ can be defined analogously.

4.2.3. Embed into CCA's objective

Now we attempt to introduce an objective for inventing our new algorithm. Specifically, we give the optimization problem (24) by defining the following objective function, which embodies our intuition:

$$\max_{\mathbf{w}_x,\mathbf{w}_y}\ \mathbf{w}_x^T\tilde{C}_{xy}\mathbf{w}_y+\frac{\eta}{2}\left[\mathbf{w}_x^T(S_{rlb}^X-S_{rlw}^X)\mathbf{w}_x+\mathbf{w}_y^T(S_{rlb}^Y-S_{rlw}^Y)\mathbf{w}_y\right]$$
$$\text{s.t.}\quad \mathbf{w}_x^T\tilde{C}_{xx}\mathbf{w}_x+\mathbf{w}_y^T\tilde{C}_{yy}\mathbf{w}_y=1 \tag{24}$$

where $\tilde{\mathbf{C}}_{xx}=(1/n_p)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$, $\tilde{\mathbf{C}}_{yy}=(1/n_p)\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$ and $\tilde{\mathbf{C}}_{xy}=(1/n_p)\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T$, η is the regularization parameter which controls the balance between the between-view correlation and the within-view semi-supervised DR objectives. From problem (24), we can find that the first term ensures as maximal correlation between the paired data as possible and the second term tries to maximally separate samples from different classes and maximally preserve the global structure for **X**-view and **Y**-view respectively. In fact, our embedding manner is flexible since any existing similar definitions [40,42] can be used to substitute those corresponding matrices. More importantly, formulation (24) suits for almost all semi-paired and semi-supervised scenario. Thus we term such a DR method as semi-paired and semi-supervised generalized correlation analysis (S^2GCA).

Remark. Although the above derivation is just for linear case, but in fact, it can be easily generalized to the nonlinear version via the powerful kernel trick [43].

4.3. Solving

In order to design S^2GCA technically, we need to solve problem (24). By the Lagrangian technique, we define the following function:

$$L(\mathbf{w}_x,\mathbf{w}_y,\lambda)=\mathbf{w}_x^T\tilde{C}_{xy}\mathbf{w}_y+\frac{\eta}{2}\left[\mathbf{w}_x^T(S_{rlb}^X-S_{rlw}^X)\mathbf{w}_x+\mathbf{w}_y^T(S_{rlb}^Y-S_{rlw}^Y)\mathbf{w}_y\right]$$
$$-\frac{\lambda}{2}\left[\mathbf{w}_x^T\tilde{C}_{xx}\mathbf{w}_x+\mathbf{w}_y^T\tilde{C}_{yy}\mathbf{w}_y-1\right] \tag{25}$$

where λ is the Lagrangian multiplier. Differentiating (25) with respect to $\mathbf{w}_x$, $\mathbf{w}_y$ and zeroing their derivatives, we have

$$\frac{\partial L}{\partial\mathbf{w}_x}=\tilde{C}_{xy}\mathbf{w}_y+\eta(S_{rlb}^X-S_{rlw}^X)\mathbf{w}_x-\lambda\tilde{C}_{xx}\mathbf{w}_x=0 \tag{26}$$

$$\frac{\partial L}{\partial\mathbf{w}_y}=\tilde{C}_{xy}^T\mathbf{w}_x+\eta(S_{rlb}^Y-S_{rlw}^Y)\mathbf{w}_y-\lambda\tilde{C}_{yy}\mathbf{w}_y=0 \tag{27}$$

Then Eqs. (26) and (27) can be expressed, respectively

$$\eta(S_{rlb}^X-S_{rlw}^X)\mathbf{w}_x+\tilde{C}_{xy}\mathbf{w}_y=\lambda\tilde{C}_{xx}\mathbf{w}_x \tag{28}$$

$$\tilde{C}_{xy}^T\mathbf{w}_x+\eta(S_{rlb}^Y-S_{rlw}^Y)\mathbf{w}_y=\lambda\tilde{C}_{yy}\mathbf{w}_y \tag{29}$$

With some algebraic operations, Eqs. (28) and (29) can be boiled down to the following equation:

$$\begin{bmatrix}\eta(\mathbf{S}_{rlb}^X-\mathbf{S}_{rlw}^X) & \tilde{C}_{xy}\\ \tilde{C}_{xy}^T & \eta(S_{rlb}^Y-S_{rlw}^Y)\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix}=\begin{bmatrix}\lambda\tilde{C}_{xx}\mathbf{w}_x\\ \lambda\tilde{C}_{yy}\mathbf{w}_y\end{bmatrix} \tag{30}$$

Owing to the fact that

$$\begin{bmatrix}\lambda\tilde{C}_{xx}\mathbf{w}_x\\ \lambda\tilde{C}_{yy}\mathbf{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\tilde{C}_{xx} & 0\\ 0 & \tilde{C}_{yy}\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix} \tag{31}$$

So we get the following generalized eigenvalue equation:

$$\begin{bmatrix}\eta(\mathbf{S}_{rlb}^X-\mathbf{S}_{rlw}^X) & \tilde{C}_{xy}\\ \tilde{C}_{xy}^T & \eta(S_{rlb}^Y-S_{rlw}^Y)\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\tilde{C}_{xx} & 0\\ 0 & \tilde{C}_{yy}\end{bmatrix}\begin{bmatrix}\mathbf{w}_x\\ \mathbf{w}_y\end{bmatrix} \tag{32}$$

Then we select a set of eigenvectors $(\mathbf{w}_{xi},\mathbf{w}_{yi})s$ corresponding to the top d largest non-negative eigenvalues of (32). Thus we obtain two projection matrices $\mathbf{W}_x=[\mathbf{w}_{x1},\mathbf{w}_{x2},\ldots,\mathbf{w}_{xd}]$ and $\mathbf{W}_y=[\mathbf{w}_{y1},\mathbf{w}_{y2},\ldots,\mathbf{w}_{yd}]$ for **X**-view and **Y**-view, respectively (the reason refers to the Appendix). Then we in turn use these matrices to project the high-dimensional data of each view and produce the low-dimensional representations $\mathbf{W}_\mathbf{x}^T\mathbf{x}$ and $\mathbf{W}_y^T\mathbf{y}$ for $\mathbf{x}$ and $\mathbf{y}$. As a result, S^2GCA implements DR for semi-paired and semi-supervised multi-view data. The pseudo-code of S^2GCA is summarized in Table 2.

From (32), we can find that S^2GCA will degenerate to (1) CCA when $\eta=0$ and $n_p=n_x=n_y$; (2) SemiCCA when $\delta=1$, $\eta=2$; (3) similar to SemiLRCCA when $\delta=0$, $\eta=2$. Therefore, DR algorithm S^2GCA based on the framework S^2DR is general and flexible in modeling multi-view data.

Table 2
Pseudo-code for S^2GCA.

Input:	Semi-paired and semi-supervised multi-view data: $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_{n_p},\mathbf{x}_{n_p+1},\ldots,\mathbf{x}_{n_x}]\in R^{p\times n_x}$ and $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_{n_p},\mathbf{y}_{n_p+1},\ldots,\mathbf{y}_{n_y}]\in R^{q\times n_x}$ where $\tilde{\mathbf{X}}=[\mathbf{x}_1,\ldots,\mathbf{x}_{n_p}]$ and $\tilde{\mathbf{Y}}=[\mathbf{y}_1,\ldots,\mathbf{y}_{n_p}]$ are paired data. The number k of the nearest neighbors of labeled sample $\mathbf{x}_i$; the parameters δ,η
Output:	Projection matrices: $\mathbf{W}_x=[\mathbf{w}_{x1},\mathbf{w}_{x2},\ldots,\mathbf{w}_{xd}]$, $\mathbf{W}_y=[\mathbf{w}_{y1},\mathbf{w}_{y2},\ldots,\mathbf{w}_{yd}]$
Step1:	Compute corresponding matrices for paired data $\tilde{\mathbf{C}}_{xx}=(1/n_p)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$, $\tilde{\mathbf{C}}_{yy}=(1/n_p)\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$, $\tilde{\mathbf{C}}_{xy}=(1/n_p)\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T$
Step2:	Compute local within-class and between-class scatter matrices: $\mathbf{S}_{lb}^X$, $\mathbf{S}_{lw}^X$, $\mathbf{S}_{lb}^Y$ and $\mathbf{S}_{lw}^Y$
Step3:	Compute regularized local within-class and between-class matrices: $\mathbf{S}_{rlb}^X$, $\mathbf{S}_{rlw}^X$, $\mathbf{S}_{rlb}^Y$ and $\mathbf{S}_{rlw}^Y$
Step4:	Solve problem (32) to obtain a set of projection vectors $(\mathbf{w}_{xi},\mathbf{w}_{yi})\quad i=1,2,\ldots,d$.
Step5:	Form projection matrices $\mathbf{W}_x=[\mathbf{w}_{x1},\mathbf{w}_{x2},\ldots,\mathbf{w}_{xd}]$ and $\mathbf{W}_y=[\mathbf{w}_{y1},\mathbf{w}_{y2},\ldots,\mathbf{w}_{yd}]$

5. Experiments and analyses

To evaluate the proposed DR algorithm S^2GCA, we systematically compare it with the related algorithms,[1] including CCA [16], DCCA [8], SemiCCA [15], SemiLRCCA [20] and SCCA [32] on both toy and real-world datasets. Firstly, we present an experiment on a synthetic dataset for performance comparison by illustrating their optimal directions in both correlation and separability when the labeled data is scarce. Secondly, we perform DR on four standard benchmark datasets including SSL database [34,35], MFD dataset [36], WebKB dataset [37] and Ads dataset [38] and then using the nearest neighbor classifier to perform classification.

In all experiments, the regularization factor η of S^2GCA is selected from $\{2^{-5},2^{-4},\ldots,2^4,2^5\}$, the balance parameter δ from $\{0.1,0.2,\ldots,1\}$. In addition, the number k, of the nearest neighbors of each labeled data is taken in $\{(c+1),2(c+1),\ldots,m(c+1)\}$ with $m(c+1)\le l$ where $\boldsymbol{c}$ is the class number, l is the number of labeled samples of training dataset. We carry out three-fold cross-validation for each dataset to select the appropriate parameters with optimal test performances.

5.1. Toy problem

Here we directly use the two-view toy dataset with two classes [8] in which each class consists of 100 two-dimensional samples. Let $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2]$ and $\mathbf{Y}=[\mathbf{Y}_1,\mathbf{Y}_2]$, where $\mathbf{X}_i$, $i=1,2$, denotes a matrix composed by the ith class data. They are randomly generated from the Gaussian distributions $N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, $i=1,2$, where $\boldsymbol{\mu}_1=\begin{pmatrix}10.18\\0.66\end{pmatrix}$, $\boldsymbol{\Sigma}_1=\begin{pmatrix}15 & 3.75\\3.75 & 15\end{pmatrix}$, $\boldsymbol{\mu}_2=\begin{pmatrix}5\\-5\end{pmatrix}$, $\boldsymbol{\Sigma}_2=\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$. Furthermore, we define $\mathbf{y}_i=\mathbf{W}^T\mathbf{x}_i+\varepsilon_i$, where ε_i follows the Gaussian noise with mean $\boldsymbol{\mu}=\begin{pmatrix}1\\1\end{pmatrix}$ and variance $\boldsymbol{\Sigma}=\begin{pmatrix}0.01 & 0\\0 & 0.01\end{pmatrix}$. Consequently, $\mathbf{x}_i$ and $\mathbf{y}_i$ satisfy linear correlation relation to a certain degree. Then we randomly select half of samples per class for training and the rest for testing. For the training set, half samples of each class are paired while randomly-selected three samples of each class are labeled and respectively illustrated as the filled circle "●" and diamond "◆" in Fig. 1, which shows the distribution of **X**-view and **Y**-view training data, respectively. And the testing samples are shown in Fig. 3.

In this toy experiment, we do not perform SCCA owing to the fact unsuitful for visualization.

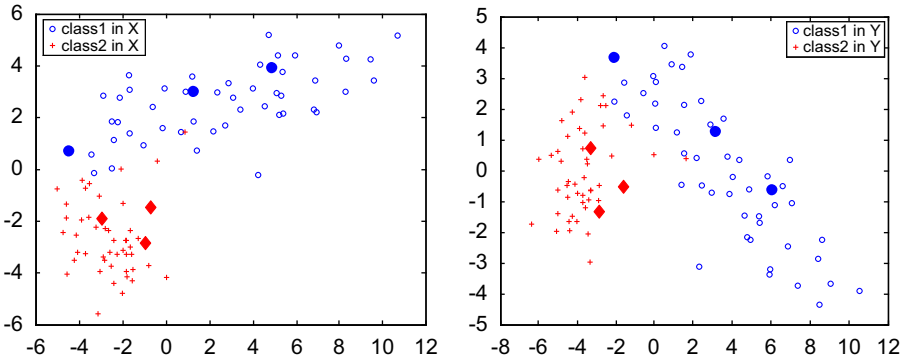

Fig. 1. The distributions of training samples.

Fig. 2(a), (b), (c), (d) and (e) shows all the first pair of features $(\mathbf{w}_x^T\mathbf{x},\mathbf{w}_y^T\mathbf{y})$ for training samples extracted, respectively, by CCA, DCCA, SemiCCA, SemiLRCCA and S^2GCA and Fig. 4 (a), (b), (c), (d) and (e) shows the first pair of extracted features for testing samples. And the horizontal and vertical coordinates correspond to **x** and **y** components. Jointly from Figs. 2 and 4, we can observe that

(1) Though indeed well discovering linear correlation between the first pair of canonical components, CCA results in relatively large overlapping between classes both for training and testing sets, meaning poor separability, due to its unsupervised nature. Although incorporating the class label information, DCCA still yields the overlapping between classes for both training samples and testing samples due to the scarcity of labeled samples, the classification accuracies of CCA and DCCA in two-dimensional projected space are both 0.95.
(2) Compared to CCA and DCCA, SemiLRCCA produces relatively less overlapping as shown in Figs. 2d and 4d, and achieves the classification accuracy of 0.98 in the two-dimensional

[1] We do not compare MVSSDR due to that its discriminative information is given in the form of the pairwise cannot-link and must-link constraints.

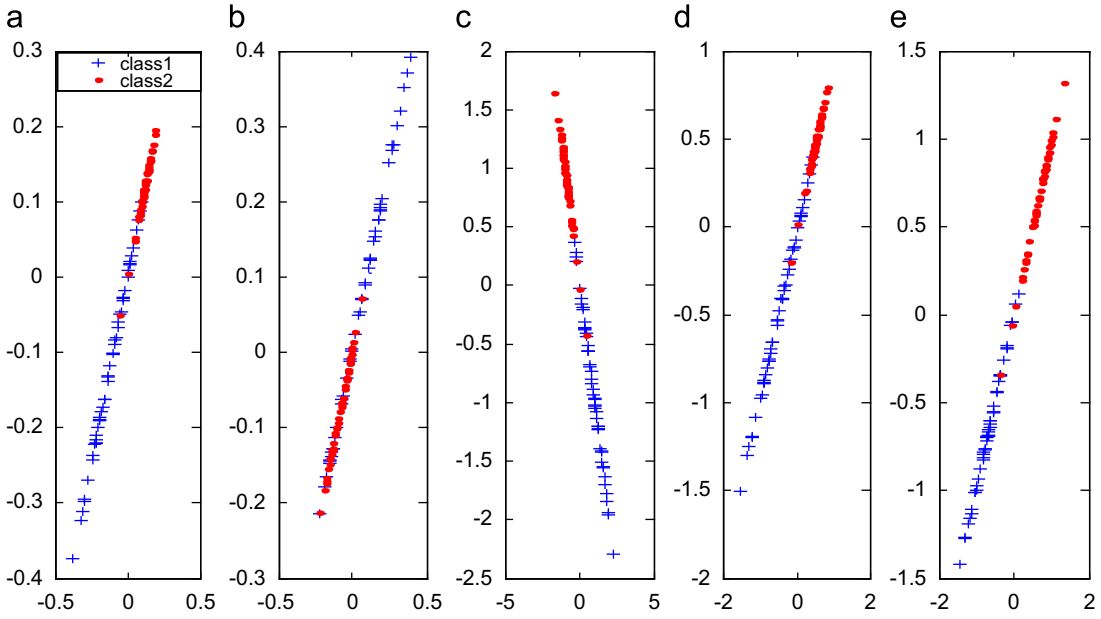

Fig. 2. The illustration of the first pair of features extracted by compared methods for training samples.

projected sapce, superior to both CCA and DCCA. SemiCCA outperforms SemiLRCCA on both training and testing sets. Although having comparable training performance to S^{2}GCA, SemiCCA still performs relatively poorly compared to S^{2}GCA on testing set by 2% as shown in Fig. 4e and Table 3.

(3) Compared with all the other methods, S^{2}GCA not only inherits the merits of preserving correlation, but also well separates different classes, thus performs better for both training and testing sets, and achieves accuracies of 0.99 in two-dimensional projected space. Moreover, by comparing Fig. 2 with Fig. 4, we can also find that S^{2}GCA has better generalization ability for unseen samples.

Table 3
Accuracies of CCA, DCCA, SemiCCA, SemiLRCCA and S^{2}GCA for testing samples.

	CCA	DCCA	SemiCCA	SemiLRCCA	S^{2}GCA
X	0.95	0.95	0.97	0.98	0.99
Y	0.95	0.95	0.97	0.98	0.99

5.2. Experiments on semi-supervised learning database

To further investigate the effectiveness of S^{2}GCA, we also perform experiments in benchmark SSL datasets [34,35]. SSL datasets[2] consist of nine semi-supervised learning datasets. Here, we just select six binary class datasets from them, i.e., SSL1, SSL2, SSL3, SSL4, SSL5 and SSL7. Tables 4 and 5 give their detailed descriptions.

In this paper, experimental settings such as training and testing samples, as well as the labeled and unlabeled data follow [34,35]. In addition, for creating two-view data for single view data which acts as one view, the PCA-reduced data are used as the other view. Then we perform the experimental comparison among CCA, DCCA, SCCA, SemiCCA, SemiLRCCA and S^{2}GCA on original data and PCA-reduced data with 95% energy preservation. We randomly select 10%, 20%, 30%, 40% and 50% samples from training sets as the paired data and the rest as the unpaired data. The average recognition accuracies of the five methods over 12 repetitions are shown in Fig. 5 for 10 labeled case and Fig. 6 for 100 labeled cases, respectively.

From Figs. 5 and 6, we can obtain several insights as follows:

(1) S^{2}GCA outperforms the other five methods on most cases, especially on the ten-labeled data cases. For example, on SSL1_10 and SSL4_10, the improvement of S^{2}GCA is remarkable. The experimental results show that preserving global structure of unlabeled data as maximally as possible and simultaneously separating labeled data in different classes as maximally as possible is helpful for seeking projections favorite subsequent classification.

(2) With the increase of paired data (proportion), DCCA keeps the invariable recognition accuracy due to its irrelevance with the paired information. The performance of other five methods all get improvement to different extents on most cases, especially CCA which just bases on paired data. S^{2}GCA utilizes not only paired information but also structural (global and local) and discriminative information in data. Consequently, its performance rises relatively smoothly as the number of paired data increases. Similar results for SCCA, SemiCCA and SemiLRCCA.

(3) When the number of labeled data reaches 100, all the methods achieve better recognition accuracy relative to the case of the 10 labeled data on most datasets.

5.3. Experiments on multiple feature handwritten digit database (MFD)

The Multiple Feature (handwritten) digit data set (MFD)[3] is selected from UCI machine learning repository [36]. It involves six sets of features of handwritten digits form 0 to 9. Each class contains 200 samples and the total sample size is 2000. The six feature sets are flourier coefficients (Fou,76), contour correlation characteristics (Fac,216), Karhunen-Loève expansion coefficients (Kar,64), pixel average in 2×3 windows (Pix,240), morphological characteristics (Mor,6) and Zernike moments (Zer,47). And the dimension of each feature is listed after the feature abbreviation in the bracket.

[2] The data sets are available from 'http://www.kyb.tuebingen.mpg.de/ssl-book/'.

[3] The data sets are from 'http://www.ics.uci.edu/~mlearning/MLSummary.html'.

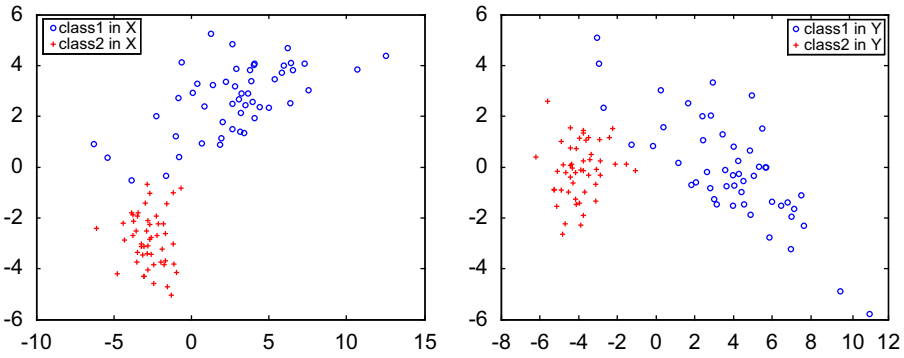

Fig. 3. The distributions of testing samples.

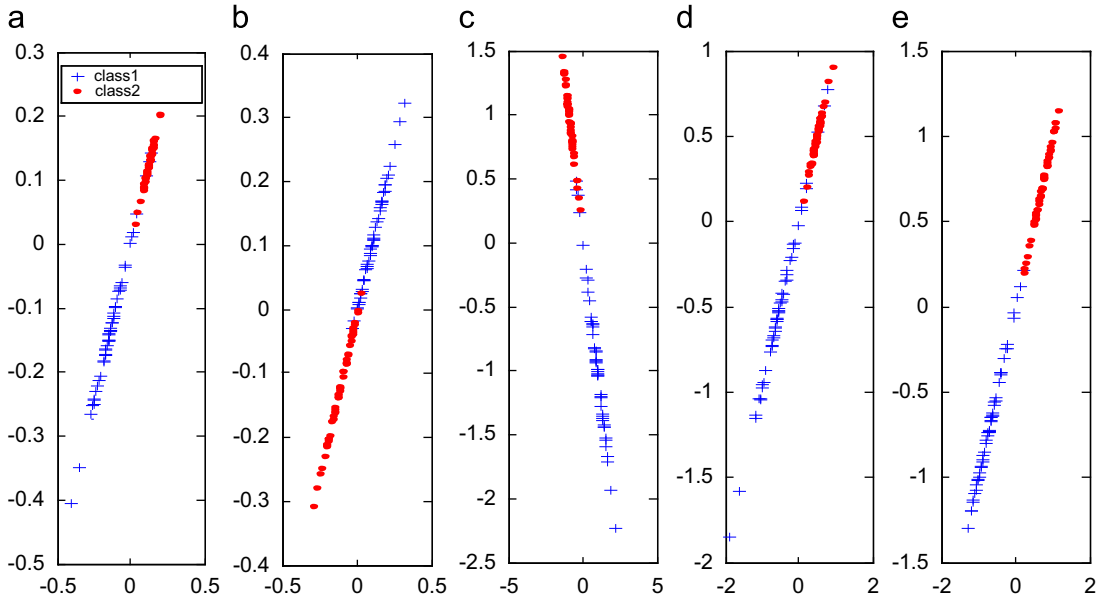

Fig. 4. The illustration of the first pair of features extracted by compared methods for testing samples.

Table 4
The attributes of the six datasets with 10 labeled samples in the SSL database.

Dataset	SSL1_10	SSL2_10	SSL3_10	SSL4_10	SSL5_10	SSL7_10
Number of dimension	241	241	241	117	241	241
Number of labeled data	10	10	10	10	10	10
Number unlabeled data	1490	1490	1490	390	1490	1490

Table 5
The attributes of the six datasets with 100 labeled samples in the SSL database.

Dataset	**SSL1_100**	**SSL2_100**	**SSL3_100**	**SSL4_100**	**SSL5_100**	**SSL7_100**
Number of dimension	241	241	241	117	241	241
Number of labeled data	100	100	100	100	100	100
Number unlabeled data	1400	1400	1400	300	1400	1400

Firstly, we choose any two sets of features as **X**-view and **Y**-view, thus there are 15 combinations of the six features in total. Among 200 samples per-class, we randomly select half of each class for training and the remaining for testing. Secondly, the training portion is further split into paired and unpaired ones where the ratio of the paired to the unpaired is 50:50 of the training samples per-class. In addition, we randomly select 10% of the training samples as labeled data used for semi-supervised learning and the rest leaves unlabeled. Owing to lacking enough paired information in the testing samples, we just give the classification accuracies on the dimension-reduced data of individual view. The parameters of S^2GCA are searched by cross-validation for optimizing performance. And the parameters corresponding to the best results in the validation is finally used in testing. We repeat the experiments ten times and report their average results in Table 6 for **X**-view and Table 7 for **Y**-view where the best performances are highlighted in bold.

From Tables 6 and 7, we can obtain several attractive observations as follows:

(1) It is obvious that S^2GCA prominently outperforms CCA on all cases, both of $\mathbf{W}_x^T\mathbf{X}$ and $\mathbf{W}_y^T\mathbf{Y}$. Although SCCA incorporates the discriminative information into DR and is superior to CCA on

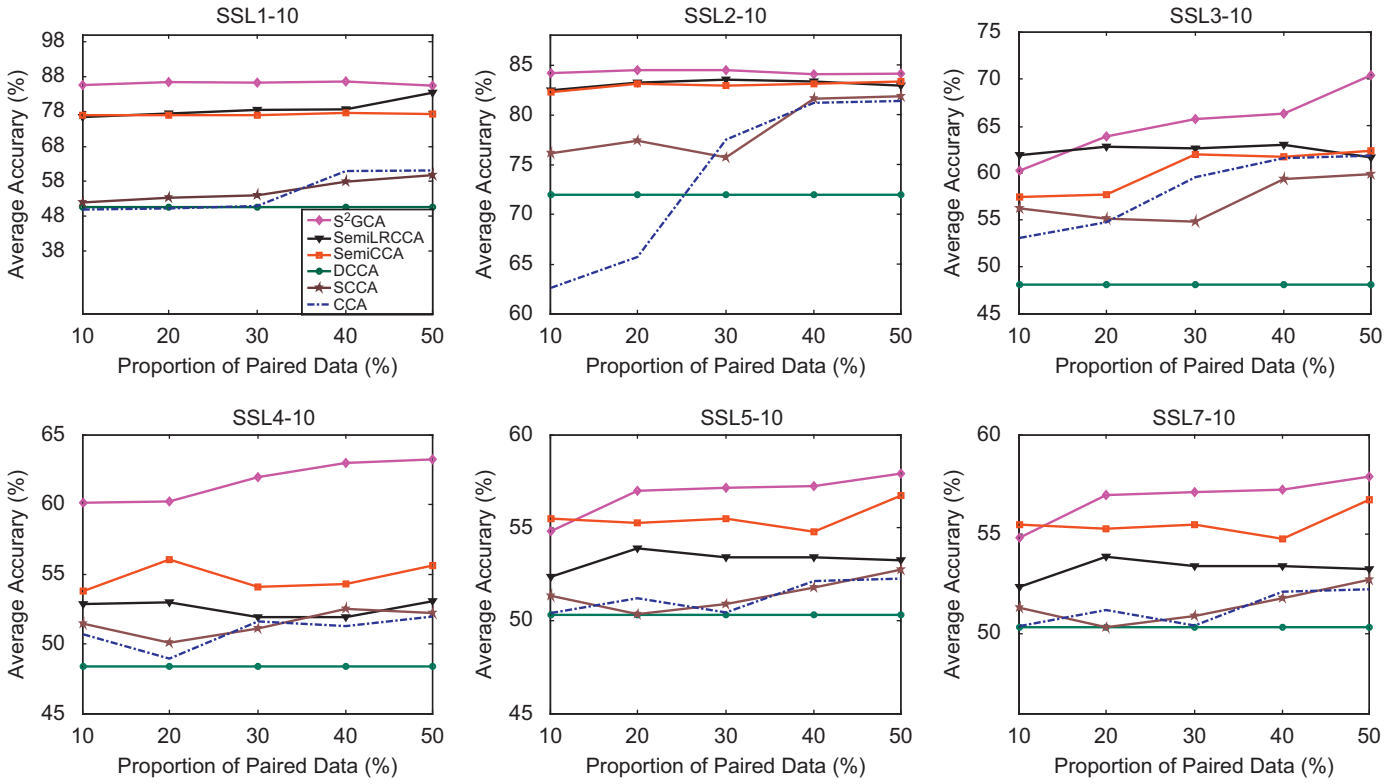

Fig. 5. Comparisons of the performance of the six methods on six different SSL datasets with 10 labeled data.

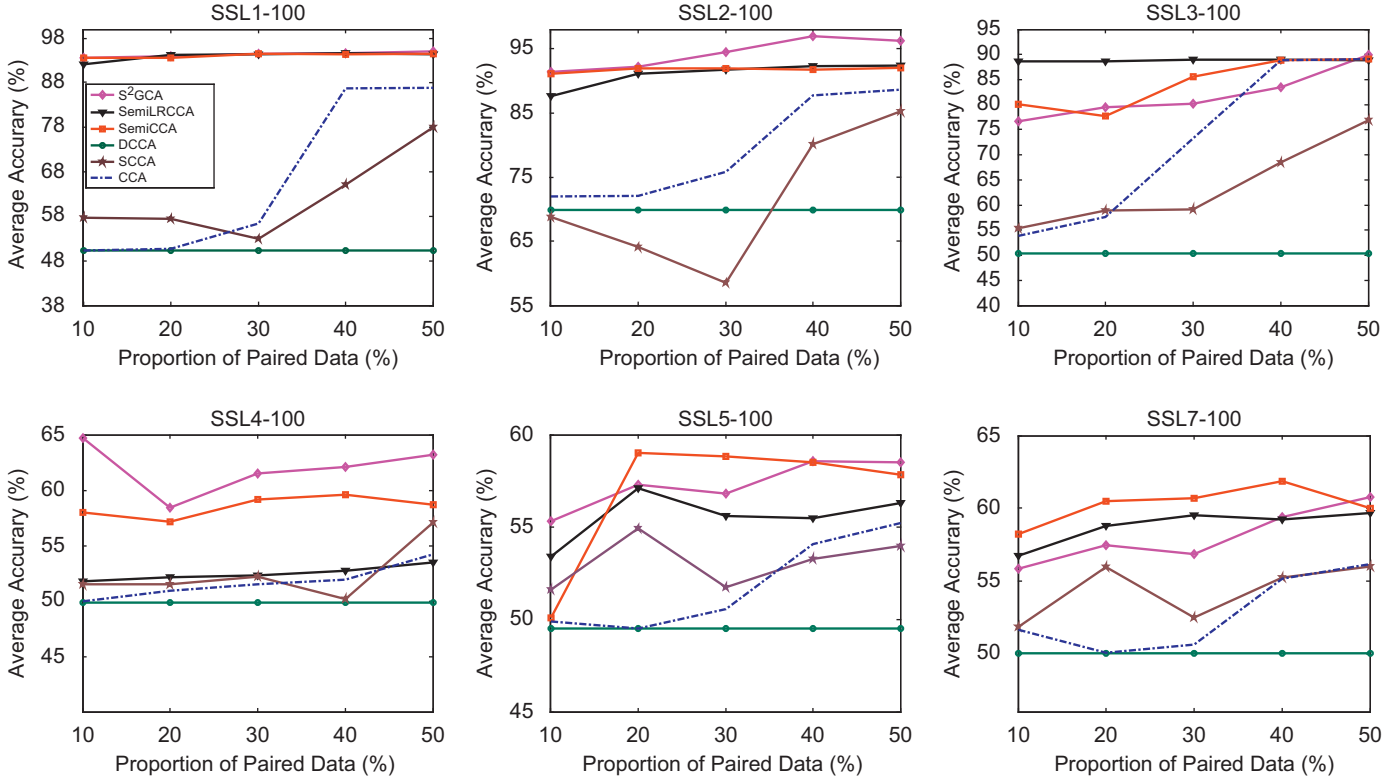

Fig. 6. Comparisons of the performance of the six methods on six different SSL datasets with 100 labeled data.

most case of $\mathbf{W}_x^T\mathbf{X}$ and $\mathbf{W}_y^T\mathbf{Y}$, S^{2}GCA still excels SCCA significantly on all cases. The result validates that only emphasizing on the paired information or class information is NOT enough for DR, especially when paired data and labeled data are few.

(2) Compared with SemiCCA, S^{2}GCA achieves better recognition accuracy on 13 out of the 15 feature combinations for $\mathbf{W}_x^T\mathbf{X}$, especially achieving the maximum improvement of 29% on Fac and Zer combination, of 24% on Fac and Mor combination,

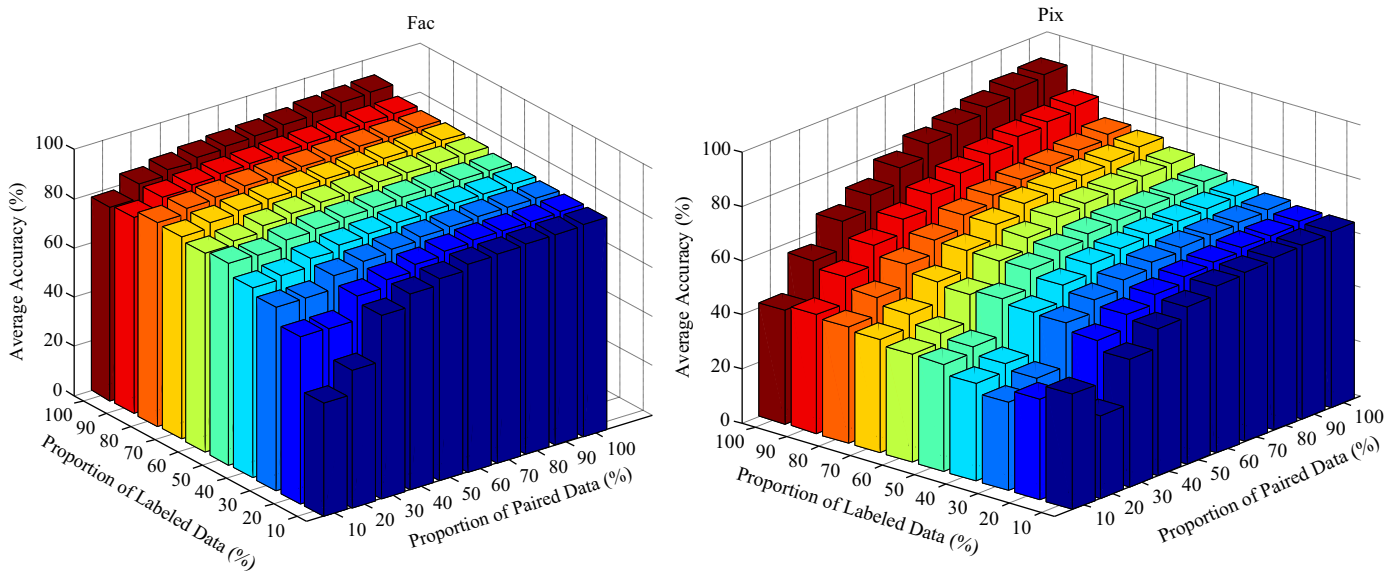

Fig. 7. Comparisons of the performance of S^{2}GCA on Far and Pix.

of 21% on Fac and Pix combination. S^{2}GCA improves more than 5% on the 8 cases, and obtains comparable results on the other 4 cases. S^{2}GCA achieves better recognition accuracy on 12 out of the 15 feature combinations for $\mathbf{W}_y^T\mathbf{Y}$.

(3) Compared with SemiLRCCA, S^{2}GCA provides better results on 11 out of the 15 feature combinations for $\mathbf{W}_x^T\mathbf{X}$. S^{2}GCA exceeds SemiLRCCA to different degrees from 2% to 22%. Especially on the combination of cases 6,7,9,10,13,14, S^{2}GCA improves more than 7% in classification accuracies and still obtains a small quantity of improvement on the other 2 cases. S^{2}GCA outperforms SemiLRCCA on 11 out of the 15 feature combinations for $\mathbf{W}_y^T\mathbf{Y}$, and its improvement is also prominent. The superiority of S^{2}GCA further validates the reasonability of the semi-supervised learning incorporated with semi-paired correlation analysis.

(4) Although taking the class label into account in constructing the between-class and within-class correlation matrices, DCCA performs still poorly when the labeled samples are scarce, which attributes to the fact that it just concerns discriminant information rather than the intrinsic structural information in data.

Secondly, for illustrating performance change of our method with both the paired data ratio and the labeled data ratio, we simultaneously vary the numbers of the labeled data and the paired data in the training set to evaluate the contribution of the different prior information. Here we select Far and Pix combination. Fig. 7 shows that the performance of S^{2}GCA increases monotonously with the increase of the paired ratio and the labeled ratio of the training samples.

5.4. Experiments on WebKB

The WebKB dataset[4] used in [37] consists of web pages collected from web sites of computer science departments of various universities. The dataset contains two descriptions: (1) *fulltext*—description on the web pages. (2) *inlinks*—the hyperlinks pointing to the page. It is natural to take these two descriptions as two views, i.e., the *fulltext* view and the *inlinks* view. There are 1051 pages in total, which have been manually classified into two classes: course (230) and non-course (821). The original *fulltext* and *inlinks* documents are processed to 3000-dimensional and 1840-dimensional vectors, respectively. For our experiments, we

[4] The data sets are from 'http://www.cs.cmu.edu/afs/cs/project/theo-11/www/wwkb/'.

Table 6
Classification accuracy of the six methods on MFD database (**X**-view).

	X	Y	CCA	DCCA	SCCA	SemiCCA	SemiLRCCA	S^{2}GCA
1	Fac	Fou	0.4625	0.1088	0.5476	0.7238	**0.8872**	0.8527
2	Fac	Kar	0.6179	0.1119	0.5431	0.7594	0.8831	**0.8882**
3	Fac	Mor	0.5762	0.1037	0.3671	0.5928	0.8178	**0.8368**
4	Fac	Pix	0.4433	0.0993	0.7599	0.5886	**0.8948**	0.7911
5	Fac	Zer	0.5463	0.1026	0.5117	0.61	0.8844	**0.9037**
6	Fou	Kar	0.5013	0.4721	0.6631	0.7727	0.6989	**0.7848**
7	Fou	Mor	0.5405	0.4291	0.4423	0.7467	0.5636	**0.784**
8	Fou	Pix	0.4891	0.4721	0.6755	0.7746	0.7199	**0.7845**
9	Fou	Zer	0.528	0.4721	0.6387	0.7663	0.6875	**0.7899**
10	Kar	Mor	0.5918	0.5859	0.5471	**0.8422**	0.7295	0.8186
11	Kar	Pix	0.6683	0.7578	0.8575	0.8422	**0.9031**	0.8492
12	Kar	Zer	0.6639	0.7578	0.8284	0.8422	0.7982	**0.8946**
13	Mor	Pix	0.6544	0.6521	0.6773	0.6586	0.6381	**0.7036**
14	Mor	Zer	0.6544	0.6521	0.6773	0.6588	0.6384	**0.7017**
15	Pix	Zer	0.3687	0.1038	0.3035	0.7308	**0.8258**	0.6728

Table 7
Classification accuracy of the six methods on MFD database (**Y**-view).

	X	Y	CCA	DCCA	SCCA	SemiCCA	SemiLRCCA	S^{2}GCA
1	Fac	Fou	0.4891	0.4212	0.6754	0.7736	0.739	**0.784**
2	Fac	Kar	0.6683	0.7032	0.8575	0.8486	0.8903	**0.9035**
3	Fac	Mor	0.6544	0.6521	0.6773	0.6561	0.6428	**0.7068**
4	Fac	Pix	0.3309	0.1088	0.4987	0.7308	**0.8886**	0.586
5	Fac	Zer	0.5772	0.5938	0.7300	0.6746	0.743	**0.8173**
6	Fou	Kar	0.6683	0.7578	0.8575	0.8422	0.8149	**0.8685**
7	Fou	Mor	0.6544	0.6521	0.6773	0.6583	0.6329	**0.7021**
8	Fou	Pix	0.3101	0.1094	0.3169	0.7308	**0.8235**	0.573
9	Fou	Zer	0.5772	0.6513	0.7300	0.6746	0.6382	**0.8024**
10	Kar	Mor	0.6544	0.6521	0.6773	0.6585	0.6419	**0.7008**
11	Kar	Pix	0.6171	0.1339	0.3110	0.7308	**0.8906**	0.7966
12	Kar	Zer	0.5772	0.6513	0.7300	0.6746	0.6757	**0.8037**
13	Mor	Pix	0.3881	0.1019	0.2761	**0.7309**	0.6902	0.5421
14	Mor	Zer	0.5808	0.5896	0.4707	0.6746	0.5429	**0.7466**
15	Pix	Zer	0.5772	0.5976	0.7300	0.6746	0.7094	**0.8067**

Table 8
Classification accuracy of the six methods on WebKB.

Supervised ratio (%)	CCA	DCCA	SCCA	SemiCCA	SemiLRCCA	S^2GCA
Fullext						
5	0.7411	0.9032	0.7135	0.9141	0.8821	**0.9288**
10	0.7857	0.8796	0.7670	**0.9461**	0.9408	**0.9461**
20	0.8170	0.8057	0.7735	0.9531	0.9146	**0.9556**
Inlinks						
5	0.8109	0.9099	0.7800	0.9078	0.8585	**0.9360**
10	0.8091	0.8703	0.7914	0.9023	0.8558	**0.9267**
20	0.8347	0.8324	0.8827	0.9269	0.8758	**0.9352**

Table 9
Description of the five features of advertisement data.

Abbreviation	Description of the feature	Dimensions
Alt	Information of the alt terms	111
Cap	Information of the words occurring near the anchor text	19
url	Information of phrases occurring in the URL	457
Origurl	Information of the image's URL	495
Ancurl	Information of the anchor text	472

first perform PCA to reduce the dimensionality to 100 both for *fulltext* and *inlinks*. As usual, we randomly select half of each class for training and the rest for testing. The training samples are further split into paired and unpaired portions with the ratio of 1:1 per-class. We further randomly choose 5%, 10% and 20% of the training samples as labeled samples and the rest as unlabeled ones for semi-supervised learning. We report the 10-run average results and highlight the best performances in bold in Table 8.

From Table 8, it is observed that:

(1) S^2GCA is overall better than or comparable to the other five compared algorithms in recognition performance. More importantly, S^2GCA achieves satisfied accuracy both for the *fulltext* view and the *inlinks* view, even with a small number of labeled samples. Its superiority validates the reasonability of appropriately fusing the class information and the structural information in DR process, which accords with the well-known "No Free Lunch" Theorem [44], i.e., making sufficient use of prior knowledge can promote the leaning performance.
(2) With the increase of labeled data, the performance on the *fulltext* view of all the compared methods except DCCA is consistently improved to different extent, however, on the *inlinks* view, consistently increased only for both SCCA and SemiLRCCA, decreased for DCCA and fluctuated for the rest three methods. Finally, it is worth to point out that as validated in [28], the increase of labeled data is not always beneficial to semi-supervised learning.

5.5. Experiments on advertisement data (Ads)

The internet advertisements data set[5] is selected from UCI machine learning repository. It contains 3279 samples among which 458 (roughly 14%) are advertisements. Each sample is treated as a binary vector with quite large sparsity. The task is to predict whether the web page is an advertisement ("ad") or not ("non-ad"). Details of data creation and the feature design are described in Kushmerick [38]. Here like [14], we select five feature sets listed in Table 9.

[5] The data sets are from 'http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements'.

Table 10
Classification accuracy of the six methods on Ads (**X**-view).

	X	**Y**	CCA	DCCA	SCCA	SemiCCA	SemiLRCCA	S^2GCA
1	Alt	Cap	0.3285	0.2749	0.3298	0.3333	0.3434	**0.3772**
2	Alt	Ancurl	0.2921	0.2749	0.2564	0.3190	0.3428	**0.3628**
3	Alt	Origurl	0.2813	0.2749	0.2564	0.3223	0.3348	**0.3763**
4	Alt	url	0.2830	0.2749	0.2564	0.3264	0.3367	**0.3771**
5	Cap	Ancurl	0.1557	0.1559	0.1534	0.1628	0.1647	**0.1674**
6	Cap	Origurl	0.1565	0.1559	0.1534	0.1629	0.1648	**0.1665**
7	Cap	url	0.1563	0.1559	0.1534	0.1641	0.1648	**0.1669**
8	Ancurl	Origurl	0.5554	0.5601	0.4434	0.6189	0.6592	**0.7243**
9	Ancurl	url	0.5553	0.5601	0.4388	0.6212	0.6672	**0.7239**
10	Origurl	url	0.6206	0.5767	0.4995	0.6278	0.6910	**0.7358**

Table 11
Classification accuracy of the six methods on Ads (**Y**-view).

	X	**Y**	CCA	DCCA	SCCA	SemiCCA	SemiLRCCA	S^2GCA
1	Alt	Cap	0.1595	0.1559	0.1534	0.1640	0.1636	**0.1688**
2	Alt	Ancurl	0.6037	0.5601	0.5511	0.6470	0.6639	**0.7223**
3	Alt	Origurl	0.6579	0.5767	0.6278	0.6437	0.6921	**0.7399**
4	Alt	url	0.7246	0.6651	0.7172	0.7167	0.7630	**0.8148**
5	Cap	Ancurl	0.6195	0.5601	0.5549	0.6421	0.6612	**0.7295**
6	Cap	Origurl	0.6647	0.5767	0.6663	0.6434	0.6937	**0.7330**
7	Cap	url	0.7251	0.6651	0.6677	0.7092	0.7539	**0.8056**
8	Ancurl	Origurl	0.6642	0.5767	0.4995	0.6415	0.6984	**0.740**4
9	Ancurl	url	0.6739	0.6651	0.5148	0.7244	0.7667	**0.8062**
10	Origurl	url	0.6391	0.6651	0.5477	0.7170	0.7559	**0.8032**

As usually, we choose any two sets of features as **X-**view and **Y-**view respectively, thus there are 10 combinations of the five features in total. Then we randomly select half of each class for training and the rest for testing. The training samples are further split in 1:1 manner into paired and unpaired portions for each class. Further for semi-supervised learning, we randomly choose 10% of the training samples labeled and the rest unlabeled. The classification accuracies averaged over 10 independent trials are summarized in Tables 10 and 11, and the best performances are highlighted in bold.

From Tables 10 and 11, we can make several interesting observations:

(1) For **X**-view, S^2GCA achieves the highest accuracies in all cases and excels the best result among the other methods. Specifically, it improves more than 6% on case 8, 9 and10, 4% on case 3 and 4, 2% on case 1 and 2 and a slight progress on the last three cases. For **Y**-view, S^2GCA shows overall better performance than all the other algorithms. Specifically, it achieves improvements of 6% on the three combinations 2, 5 and 10, 4% on the two combinations 4, 7 and 8 and more than 4% on the three cases. Totally, compared with the related works, the low-dimensional features extracted by S^2GCA can embody the intrinsic discriminative structure to greater extent and thus can be favorable for subsequent classification.
(2) For this dataset, the six methods including ours do not exhibit sufficient prominent performance due to its sparsity of features. To achieve better performance, we can adopt some favorable preprocessing techniques for the dataset, such as latent semantic analysis (LSA) [45].

In general, a series of experimental results show superiority of S^2GCA in almost all cases with different paired ratios and different labeled ratios.

6. Conclusions and future works

Encouraged by the success of semi-paired correlation analysis and semi-supervised learning, we present a unified DR framework for semi-paired and semi-supervised multi-view data, which appears more frequently in real world. Based on the framework, we further propose a new linear DR algorithm, namely S^{2}GCA. Different from the existing correlation-based DR methods, S^{2}GCA can not only preserve the global structure of unlabeled data but also achieve the maximal separability of different classes. Different from the existing (semi-)supervised DR methods, S^{2}GCA relaxes the fully paired and fully labeled requirement for dataset. Consequently, it is general and flexible. The experimental results on both toy and benchmark datasets show its encouraging performance.

There are several directions deserved future study:

(1) Modeling design: because the solving for both S^{2}GCA and SemiCCA is finally boiled down to solving a general eigenvalue problem, the decoupled property in the form of the Eqs. (3), (10) and (15) is lost and thus makes the involved solving relatively complicated. Therefore, exploring capably-decoupled modeling is important for more efficient solving, even performance improvement.
(2) Classifier design: due to the scarcity of labeled data, using the nearest neighbor strategy to classify new data in reduced space will not be quite reasonable, thus one more reasonable strategy is to design a semi-supervised classifier directly, which will facilitate performance improvement since in the DRed space, the total number of the samples is now the sum of the samples from two views and thus is enlarged to relatively more sufficient extent.

Acknowledgments

This work was supported by National Natural Science Foundations of China (NSFC) under Grant no. 61170151 and NUAA Research Funding under Grant no. NP2011030, partially supported by NSFC under Grant nos. 60905002, 11001128.

Appendix

Let $\begin{bmatrix}\mathbf{w}_{x_i}\\ \mathbf{w}_{y_i}\end{bmatrix}$ is the eigenvector corresponding to the ith eigenvalue of (32), that is

$$\begin{bmatrix}\eta(\mathbf{S}^X_{rlb}-\mathbf{S}^X_{rlw}) & \tilde{C}_{xy}\\ \tilde{C}^T_{xy} & \eta(S^Y_{rlb}-S^Y_{rlw})\end{bmatrix}\begin{bmatrix}\mathbf{w}_{x_i}\\ \mathbf{w}_{y_i}\end{bmatrix}=\lambda_i\begin{bmatrix}\tilde{C}_{xx} & 0\\ 0 & \tilde{C}_{yy}\end{bmatrix}\begin{bmatrix}\mathbf{w}_{x_i}\\ \mathbf{w}_{y_i}\end{bmatrix}$$

Next considering the following objective value of (25) for the d eigenvectors:

$$\sum_{i=1}^{d}\left[\mathbf{w}^T_{x_i}\tilde{C}_{xy}\mathbf{w}_{y_i}+\frac{\eta}{2}\mathbf{w}^T_{x_i}(\mathbf{S}^X_{rlb}-\mathbf{S}^X_{rlw})\mathbf{w}_{x_i}+\frac{\eta}{2}\mathbf{w}^T_{y_i}\eta(S^Y_{rlb}-S^Y_{rlw})\mathbf{w}_{y_i}\right]$$

$$=\frac{1}{2}\sum_{i=1}^{d}[\eta\mathbf{w}^T_{x_i}(\mathbf{S}^X_{rlb}-\mathbf{S}^X_{rlw})\mathbf{w}_{x_i}+\mathbf{w}^T_{x_i}\tilde{C}_{xy}\mathbf{w}_{y_i}]$$

$$+\frac{1}{2}\sum_{i=1}^{d}[\mathbf{w}^T_{y_i}\tilde{C}^T_{xy}\mathbf{w}_{x_i}+\eta\mathbf{w}^T_{y_i}(\mathbf{S}^Y_{rlb}-\mathbf{S}^Y_{rlw})\mathbf{w}_{y_i}]$$

$$=\frac{1}{2}\sum_{i=1}^{d}[\lambda_i\mathbf{w}^T_{x_i}\tilde{C}_{xx}\mathbf{w}_{x_i}+\lambda_i\mathbf{w}^T_{y_i}\tilde{C}_{yy}\mathbf{w}_{x_i}]$$

$$=\frac{1}{2}\sum_{i=1}^{d}\lambda_i[\mathbf{w}^T_{x_i}\tilde{C}_{xx}\mathbf{w}_{x_i}+\mathbf{w}^T_{y_i}\tilde{C}_{yy}\mathbf{w}_{x_i}]=\frac{1}{2}\sum_{i=1}^{d}\lambda_i$$

Consequently, in order to obtain the optimal value of (25), we choose the d eigenvectors corresponding to the top d largest non-negative eigenvalues to form our final projection matrices.

References

[1] P. Glenisson, J. Mathys, B.D. Moor, Meta-clustering of gene expression data and literature-based information, Proceedings of the SIGKDD Explore Newsletter 5 (2) (2003) 101–112.
[2] Y. Wu, E.Y. Chang, K.C. Chang, J.R. Smith, Optimal multimodal fusion for multimedia data analysis, in: Proceedings of the 12th Annual ACM International Conference on Multimedia, 2004, pp. 572–579.
[3] Q. Chen, S. Sun, Hierarchical multi-view Fisher discriminant analysis, Proceedings of the Advances in Neural Information Processing Systems (2009) 289–298.
[4] K. Kailing, H. Kriegel, A. Pryakhin, M. Schubert, Clustering multi-represented objects with noise, in: Proceedings of the Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Sydney, Australia, 2004, pp. 394–403.
[5] H. Tong, J. He, M. Li, C. Zhang, W. Ma, Graph based multi-modality learning, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, Singapore, 2005, pp. 862–871.
[6] C.H. Lampert, O. Kromer, Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning, in: Proceedings of the 11th European Conference on Computer Vision, Hersonissos, Greece, 2010, pp. 566–579.
[7] T. Sun, S. Chen, J. Yang, X. Hu, P. Shi, Discriminative canonical correlation analysis with missing samples, in: Proceedings of the World Congress on Computer Science and Information Engineering, 2009, pp. 95–99.
[8] T. Sun, S. Chen, J. Yang, P. Shi, A novel method of combined feature extraction for recognition, in: Proceedings of the IEEE Conferences on Data Mining, Pisa, Italy 2008, pp. 1043–1048.
[9] M. Harel, S. Mannor, Learning from multiple outlooks, in: Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.
[10] M. Sun, H. Su, S. Savarese, F.F. Li, A multi-view probabilistic model for 3D object classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[11] K. Chaudhuri, S.M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009.
[12] L. Hoegaerts, J.A.K. Suykens, J. Vandewalle, B. De Moor, Subset based least squares subspace regression in RKHS, Neurocomputing 63 (2005) 293–323.
[13] S. Rüping, T. Scheffer, Learning with multiple views, in: ICML Workshop on Learning with Multiple View, 2005.
[14] C. Hou, C. Zhang, Y. Wu, F. Nie, Multiple view semi-supervised dimensionality reduction, Pattern Recognition 43 (2010) 720–730.
[15] A. Kimura, H. Kameoka, M. Sugiyama, SemiCCA: efficient semi-supervised learning of canonical correlations, in: Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010.
[16] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–372.
[17] F.R. Banch, M.I. Jordan, A Probability Interpretation of Canonical Correlation Analysis. 2005, Technical Report 688. Department of Statistics, University of California, Berkeley.
[18] D.R. Hardoon, S. Szedmak, J.S. Tayloy, Canonical correlation analysis: an overview with application to learning method, Neural Computation 16 (2004) 2639–2664.
[19] J. Gu, S. Chen, Y. Zhuang, Localization with incompletely paired data in complex wireless sensor network, IEEE Transaction on Wireless Communication 10 (9) (2011) 2841–2849.
[20] M.B. Blaschko, C.H. Lampert, A. Gretton, Semi-supervised laplacian regularization of kernel canonical correlation analysis, Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science 5211 (2008) 133–145.
[21] Y. Peng, D. Zhao, J. Zhang, A new canonical correlation analysis algorithm with local discrimination, Neural Processing Letters 31 (2010) 1–15.
[22] D.P. Foster, R. Johnson, S.M. Kakade, T. Zhang, Multi-View Dimensionality Reduction via Canonical Correlation Analysis. Technical Report TR-2009-5, TTI-Chicago, 2009.
[23] T.E. Bellman, Adaptive Control Processes, Princeton University Press, Princeton, 1961.
[24] M. Slaney, M. Covell, FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks, in: Proceedings of the Advances in Neural Information Processing Systems, Colorado, US, 2000.
[25] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from examples, The Journal of Machine Learning 7 (2006) 2399–2434.
[26] A. Maćkiewicz, W. Ratajczak, Principal components analysis, Computers & Geosciences 19 (3) (1993) 303–342.

[27] K. Pearson, On lines and planes of closest fit to systems of points in space, The London Edinburgh and Dublin Philosophical Magazine and Journal of Science 6 (1901) 559–572.
[28] X. Zhu, Semi-Supervised Learning Literature Survey, Department of Computer Sciences, University of Wisconsin, Madison, 2008.
[29] O. Chapelle, B. Scholkopf, A. Zien, Semi-Supervised Learning, MIT press, Cambridge, MA, 2006.
[30] M. Loog, D.d. Ridder, Local discriminant analysis, in: Proceedings of the 18th International Conference on Pattern Recognition, vol. 3, Hong Kong, China, 2006, pp. 328–331.
[31] Y. Song, F. Nie, C. Zhang, S. Xiang, A unified framework for semi-supervised dimensionality reduction, Pattern Recognition 41 (2008) 2789–2799.
[32] O. Kursun, E. Alpaydin, Canonical correlation analysis for multiview semi-supervised feature extraction, in: Proceedings of the ICANISC 2010, Part I, LNAI 6113: 2010, pp. 430–436.
[33] D. Zhang, Z. Zhou, S. Chen, Semi-supervised dimensionality reduction, in: Proceedings of the SIAM Conference on Data Mining, 2007.
[34] M. Sugiyama, T. Idé, S. Nakajima, J. Sese, Semi-supervised local Fisher discriminant analysis for dimensionality reduction, Machine Learning 78 (2010) 35–61.
[35] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, Journal of Machine Learning Research 8 (2007) 1027–1061.
[36] M.V. Breukelen, R.P.W. Duin, D.M.J. Tax, J.E.D. Hartog, Handwritten digit recognition by combined classifiers, Kybernetika 34 (4) (1998) 381–386.
[37] K. Nigam, T. Ghani, Analyzing the effectiveness and applicability of co-training, in: Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, Virginia, US, 2000, pp. 86–93.
[38] N. Kushmerick, Learning to remove internet advertisements, in: Proceedings of the Third Annual Conference on Autonomous Agents, California: Stanford, 1999, 175–181.
[39] T. Melzer, M. Reiter, H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognition 36 (9) (2003) 1961–1971.
[40] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions:a general framework for dimensionality reduction, IEEE Transaction on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.
[41] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, Advances in Neural Information Processing Systems 17 (2005) 1601–1608.
[42] R. Chatpatanasiri, B. Kijsirikul, A unified semi-supervised dimensionality reduction framework for manifold learning, Neurocomputing 73 (2010) 1631–1640.
[43] J.S. Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, England, 2004.
[44] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, 2001.
[45] A. Trivedi, P. Rai, S. DuVall, Exploiting tag and word correlation for improved webpage clustering, in: Proceedings of the SMUC, Toronto, Canada, 2010.

Xiaohong Chen received the B.S. degree in mathematics from Qufu Normal University, in 1998. In 2001, she received her M.S. degree in mathematics from Nanjing University of Aeronautics & Astronautic (NUAA) and then worked at NUAA as an assistant lecturer. Now she is currently a Ph.D. student in computer application technology at NUAA. Her research interests include dimensionality reduction and semi-supervised learning.

Songcan Chen received the B.S. degree in mathematics from Hangzhou University (now merged into Zhejiang University) in 1983. In December 1985, he completed the M.Sc. degree in computer applications at Shanghai Jiaotong University and then worked at Nanjing University of Aeronautics and Astronautics (NUAA) in January 1986 as an assistant lecturer. There he received a Ph.D. degree in communication and information systems in 1997. Since 1998, as a full professor, he has been with the Department of Computer Science and Engineering at NUAA. His research interests include pattern recognition, machine learning and neural computing. In these fields, he has authored or coauthored over 130 scientific journal papers.

Hui Xue received her B.S. degree in mathematics from Nanjing Normal University in 2002. In 2005, she received her M.S. degree in mathematics from Nanjing University of Aeronautics & Astronautics (NUAA). And she also received her Ph.D. degree in computer application technology at NUAA in 2008. Since 2009, as a university instructor, she has been with the School of Computer Science & Engineering at Southeast University. Her research interests include pattern recognition, machine learning andneural computing.

Xudong Zhou received the B.S. and M.S. degree in computer science from Taiyuan University of Technology in 2001 and 2004, respectively. And then he worked in College of Information Engineering of Yanzhou University as an assistant lecturer. Now he is currently a Ph.D. student in computer application technology at NUAA. His research interests include dimensionality reduction and semi-supervised learning.