Neurocomputing 191 (2016) 263–272

Contents lists available at ScienceDirect

Neurocomputing

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

Canonical sparse cross-view correlation analysis

Chen Zu, Daoqiang Zhang*

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form
8 January 2016
Accepted 19 January 2016
Communicated by Yue Gao
Available online 23 February 2016

Keywords:
Canonical correlation analysis
Feature extraction
Sparse representation
Cross correlation
Dimensionality reduction

ABSTRACT

Recently, multi-view feature extraction has attracted great interest and Canonical Correlation Analysis (CCA) is a powerful technique for finding the linear correlation between two view variable sets. However, CCA does not consider the structure and cross view information in feature extraction, which is very important for subsequence tasks. In this paper, a new approach called Canonical Sparse Cross-view Correlation Analysis (CSCCA) is proposed to address this problem. We first construct similarity matrices by performing sparse representation between within-class samples. Then local manifold information and cross-view correlations are incorporated into CCA. Furthermore, a kernel version of CSCCA (KCSCCA) is proposed to reveal the nonlinear correlation relationship between two sets of features. We compare CSCCA and KCSCCA with existing multi-view feature extraction methods and perform experiments on both artificial data set and real world databases including multiple features and face data sets. The experimental results demonstrate the merits of our proposed method.

1. Introduction

Multi-view learning [1,2] which learns patterns or features from instances with multiple representations has been one of the hotspots in machine learning community. It has been shown that learning from multiple representations of data often achieves better performance than traditional single view learning methods. Recently, multi-view learning techniques have been extended to multi-view regression [3] and multi-view clustering [4].

Canonical correlation analysis (CCA) [5] is a learning method to find linear relationship between two groups of multidimensional variables. The goal of CCA is to seek two bases which would maximize the correlation of data by projecting two-view data obtained from various information sources, e.g. sound and image. In the past decades, CCA and its variants have been successfully applied to many fields such as image processing [6], pattern recognition [7,8], medical image analysis [9,10] and data regression analysis [11].

Standard CCA is an unsupervised linear dimensionality reduction method. It cannot preserve local structure in canonical subspaces and either cannot reveal nonlinear correlation relationship. In order to extract features with discriminant information, variants of CCA called discriminant CCA (DCCA) [12], random correlation ensemble (RCE) [13] and discriminative extended CCA (DECCA) [14] are proposed. For example, DCCA not only considers the correlation between two corresponding views of a sample, but also uses all the cross-view correlation between within-class examples. Several works have also discussed the relationships between CCA and LDA, especially when the data features are used in one view and the class labels are used in the other view [15,16]. It is shown that CCA and LDA have some equivalent relations [17]. In order to deal with nonlinear circumstance, some nonlinear CCA algorithms have been proposed in the literature [18]. Kernel methods [19,20] are widely used to reveal nonlinear structure in the original input space, and have been introduced into CCA (e.g., Kernel CCA (KCCA)) [21]. KCCA first maps the data into high dimensional feature space by implicit nonlinear mappings, and then traditional CCA is performed in the feature space in which the nonlinear problem in the original space is converted into a linear one. However, like many other kernel methods, one disadvantage of KCCA is the choice of appropriate kernel and kernel parameters. Neural networks based nonlinear CCA suffers from some intrinsic limitations such as long-time training, slow convergence and local minima [18].

In recent years, locality preserving methods such as locally linear embedding (LLE) [22], Isomap [23] and locality preserving projections (LPP) [24] have achieved a remarkable flourish in single-view dimensionality reduction. These methods preserve the neighborhood information so as to discover the low dimensional manifold structure embedded in the original high dimensional space. Inspired by the similar idea, Sun and Chen proposed a locality preserving CCA (LPCCA) [25]. LPCCA incorporates the local structure information into CCA and decomposes the global nonlinear problem into many local linear ones, consequently, in each

* Corresponding author.
E-mail address: dqzhang@nuaa.edu.cn (D. Zhang).

http://dx.doi.org/10.1016/j.neucom.2016.01.053
0925-2312/

small neighborhood field the problem can be treated as linear CCA and the global problem can be solved by optimizing the combination or integration of these local sub-problems. It has been shown that LPCCA performs better than CCA in discovering intrinsic structure of data for some applications, e.g., data visualization and pose estimation. Nevertheless, LPCCA only concerns the correlation between sample pairs and the discrimination of the extracted features which is important in subsequent classification task, while LPCCA is dependent on the parameter k which is manually chosen through experience.

Several supervised multi-view feature extraction methods have been proposed in recent researches. For example, Diethe et al. extended the convex formulation for Kernel Fisher Discriminant Analysis to multiple views [26]. Chen et al. proposed Hierarchical Multi-view Fisher Discriminant Analysis to improve the performance in classification and dimensionality reduction of multi-view task [27]. Sharma proposed a general multi-view feature extraction approach called Generalized Multiview Analysis (GMA) [28]. Although these works can do well in supervised learning situations, but they do not consider the intrinsic structures of the data, such as manifold structure. Local discrimination CCA (LDCCA) [29] and discriminative locality preserving CCA (DLPCCA) [30] can be seemed as extensions of CCA which use label and neighborhood information. Specifically, LDCCA not only considers the correlations between sample pairs but also the correlations between samples and their local neighborhoods. DLPCCA based on LDCCA can use label discriminative information to improve classification performance and preserve the geometric structure of data to enhance the smoothness of the extracted features. It worth noting that LPCCA, LDCCA and DLPCCA directly use the standard Euclidean distance to measure the similarity between data points which may be affected by outliers for the deficiency of robustness of Euclidean distance. In LPCCA, the locality means that the global nonlinear problem is decomposed into local linear ones. So the local structure information can be preserved in the canonical subspace. In LDCCA, the locality means the local neighborhood sample pairs are used to compute the correlations while DLPCCA considers the local structure information in two views separately.

In this paper, we propose a novel learning method for multi-view data called canonical sparse cross-view correlation analysis (CSCCA). We first construct similarity matrices by performing ℓ_1 norm sparse representation on within-class samples. Then local manifold structure information and cross-view correlation are incorporated into CCA. Here, we use sparse reconstruction because ℓ_1 norm is more robust to noises than Euclidean distance. The proposed method not only preserves the local structure information in two views separately, but also the structure information in the cross view. It is worth noting that many works have investigated sparse CCA [31,32]. In those papers, the projective vectors obtained by CCA need to be sparse that means there are a lot of zero values in it, while the sparse in our method means sparse reconstruction. The sparse representation achieved by minimizing a ℓ_1 regularization related objective function chooses its neighborhood automatically and does not have to encounter model parameters. The sparse in our method is totally different from those sparse CCA. The proposed method can be efficiently solved via generalized eigenvalue decomposition. Although the solution is similar to that of [33], they are derived from different motivation. Moreover, we extend CSCCA to kernel version (KCSCCA) to find nonlinear correlation. Experimental results on both synthetic and real world data sets including multiple features data set and face databases validate the effectiveness of the proposed method.

The rest of the paper is organized as follows: In Section 2, CCA is briefly reviewed, the proposed CSCCA and KCSCCA are then introduced in detail. The experiments and results on various data sets are given in Section 3. Finally, we conclude this paper in Section 4.

2. Proposed method

2.1. Canonical correlation analysis

Given a set of pair-wise data $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}^q$, where $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are samples from different views. Define that $\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$ and $\boldsymbol{Y}=[\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n] \in \mathbb{R}^{q\times n}$. We assume that the data have been preprocessed with zero mean. CCA seeks to find two basis or projection vectors $\boldsymbol{w}_x \in \mathbb{R}^p$ and $\boldsymbol{w}_y \in \mathbb{R}^q$, such that the canonical variables $x=\boldsymbol{w}_x^T\boldsymbol{x}_i$ and $y=\boldsymbol{w}_y^T\boldsymbol{y}_i$ would be maximally correlated. The objective function of CCA can be formulated as

$$\max_{\boldsymbol{w}_x,\boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T\boldsymbol{C}_{xy}\boldsymbol{w}_y}{\sqrt{(\boldsymbol{w}_x^T\boldsymbol{C}_{xx}\boldsymbol{w}_x)(\boldsymbol{w}_y^T\boldsymbol{C}_{yy}\boldsymbol{w}_y)}} \tag{1}$$

where $\boldsymbol{C}_{xx}=\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T=\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{C}_{yy}=\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^T=\boldsymbol{Y}\boldsymbol{Y}^T$ are within-sets covariance matrices and $\boldsymbol{C}_{xy}=\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{y}_i^T=\boldsymbol{X}\boldsymbol{Y}^T$ is between-sets covariance matrix. Since the two basis vectors are scale independent, $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ can be obtained by solving the following optimization problem with constraints:

$$\begin{aligned} \max_{\boldsymbol{w}_x,\boldsymbol{w}_y} \quad & \boldsymbol{w}_x^T\boldsymbol{C}_{xy}\boldsymbol{w}_y \\ \text{s.t.} \quad & \boldsymbol{w}_x^T\boldsymbol{C}_{xx}\boldsymbol{w}_x=1, \quad \boldsymbol{w}_y^T\boldsymbol{C}_{yy}\boldsymbol{w}_y=1 \end{aligned} \tag{2}$$

The optimization problem of CCA can be solved by applying Lagrangian equation to Eq. (2) and we can obtain the following generalized eigenvalue decomposition problem:

$$\begin{bmatrix} & \boldsymbol{C}_{xy} \\ \boldsymbol{C}_{xy}^T & \end{bmatrix}\begin{bmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\boldsymbol{C}_{xx} & \\ & \boldsymbol{C}_{yy}\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{bmatrix} \tag{3}$$

2.2. Canonical sparse cross-view correlation analysis

In this section, in order to cope with the nonlinear problems and improve the performance of CCA in subsequent classification task, we propose a novel feature extraction method called canonical sparse cross-view correlation analysis (CSCCA). In CSCCA, the local structure information is incorporated and the cross correlations between two views from within-class samples are used by sparse representation.

The optimization problem of CCA can be written in the equivalent form [25] as:

$$\begin{aligned} \max_{\boldsymbol{w}_x,\boldsymbol{w}_y} \quad & \boldsymbol{w}_x^T\cdot\sum_{i=1}^n\sum_{j=1}^n(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^T\cdot\boldsymbol{w}_y \\ \text{s.t.} \quad & \boldsymbol{w}_x^T\cdot\sum_{i=1}^n\sum_{j=1}^n(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\cdot\boldsymbol{w}_x=1 \\ & \boldsymbol{w}_y^T\cdot\sum_{i=1}^n\sum_{j=1}^n(\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^T\cdot\boldsymbol{w}_y=1 \end{aligned} \tag{4}$$

We incorporate the local structure information and the within-class cross correlations into Eq. (4). The objective function of CSCCA can be formulated as follows:

$$\begin{aligned} \max_{\boldsymbol{w}_x,\boldsymbol{w}_y} \quad & \boldsymbol{w}_x^T\cdot\left(\sum_{i=1}^n\sum_{j=1}^n S_{ij}^x(\boldsymbol{x}_i-\boldsymbol{x}_j)S_{ij}^y(\boldsymbol{y}_i-\boldsymbol{y}_j)^T+\sum_{i=1}^n\sum_{j=1}^n(S_{ij}^x+S_{ij}^y)\boldsymbol{x}_i\boldsymbol{y}_j^T\right)\cdot\boldsymbol{w}_y \\ \text{s.t.} \quad & \boldsymbol{w}_x^T\cdot\sum_{i=1}^n\sum_{j=1}^n S_{ij}^x(\boldsymbol{x}_i-\boldsymbol{x}_j)S_{ij}^x(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\cdot\boldsymbol{w}_x=1 \\ & \boldsymbol{w}_y^T\cdot\sum_{i=1}^n\sum_{j=1}^n S_{ij}^y(\boldsymbol{y}_i-\boldsymbol{y}_j)S_{ij}^y(\boldsymbol{y}_i-\boldsymbol{y}_j)^T\cdot\boldsymbol{w}_y=1 \end{aligned} \tag{5}$$

where S_{ij}^x and S_{ij}^y are elements of similarity matrices $\boldsymbol{S}_x$ and $\boldsymbol{S}_y$. As we can see that there are some differences between Eqs. (4) and (5). First the similarity matrices $\boldsymbol{S}_x$ and $\boldsymbol{S}_y$ are obtained by sparse reconstruction instead of the standard Euclidean distance. The sparse weight matrix can reflect the intrinsic geometric properties of the data to some extent and naturally preserve potential discriminant information [34]. We therefore expect that the characterization in the original high-dimensional embedding space can be preserved in the low-dimensional embedding space. Second we emphasize the correlations between within-class by adding the correlation term $\sum_{i=1}^{n}\sum_{j=1}^{n}(S_{ij}^x+S_{ij}^y)\boldsymbol{x}_i\boldsymbol{y}_j^T$.

The similarity matrix $\boldsymbol{S}_x$ is defined as follows. Given c classes samples:

$$\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]=[\boldsymbol{x}_1^{(1)},\ldots,\boldsymbol{x}_{n_1}^{(1)},\ldots,\boldsymbol{x}_1^{(c)},\ldots,\boldsymbol{x}_{n_c}^{(c)}]$$

where $\boldsymbol{x}_i^{(j)}\in\mathbb{R}^d$ denotes the i-th sample in the j-th class. For sample $\boldsymbol{x}_i^{(j)}$ from the j-th class, we find a sparse reconstructive weight vector $\boldsymbol{S}_i^w$ by the following minimization problem:

$$\min_{\boldsymbol{S}_i^w\geq 0}\frac{1}{2}\|\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{S}_i^w\|+\eta\|\boldsymbol{S}_i^w\|_1 \tag{6}$$

where η is a balancing factor which controls the sparsity of $\boldsymbol{S}_i^w$ and $\boldsymbol{S}_i^w$ is an vector where the element in the position of $\boldsymbol{x}_i^{(j)}$ is zero to ignore degenerated solution.

$$\boldsymbol{S}_i^w=[\overbrace{0,\ldots,0}^{\sum_{k=1}^{j-1}n_k},S_{i1},\ldots,S_{i,i-1},0,S_{i,i+1},\ldots,S_{in_j},\overbrace{0,\ldots,0}^{n-\sum_{k=1}^{j}n_k}]^T$$

The optimization problem of Eq. (6) can be solved by SLEP toolbox [35]. It is worth noting that we only use the samples with the same label to reconstruct $\boldsymbol{x}_i$ for preserving the class information. When we get $\tilde{\boldsymbol{S}}_i^w$ which is the optimal solution of Eq. (6), the sparse reconstructive weight matrix $\boldsymbol{S}_x$ can be denoted as follows:

$$\boldsymbol{S}^w=[\tilde{\boldsymbol{S}}_1^w,\ldots,\tilde{\boldsymbol{S}}_n^w]$$
$$\boldsymbol{S}_x=\boldsymbol{S}^w+(\boldsymbol{S}^w)^T$$

The reconstructive weight matrix $\boldsymbol{S}_x$ is a $n\times n$ symmetric matrix, of which the elements $\tilde{S}_{ij}$ represents the contribution of each $\boldsymbol{x}_j$ to reconstruct $\boldsymbol{x}_i$. Generally speaking, if the element $\tilde{S}_{ij}$ is bigger, the sample $\boldsymbol{x}_j$ is more important to reconstruct $\boldsymbol{x}_i$.

The optimization prob lem (5) can be rewritten after some algebraic manipulations as

$$\begin{aligned}\max_{\boldsymbol{w}_x,\boldsymbol{w}_y}\quad & \boldsymbol{w}_x^T\boldsymbol{X}\boldsymbol{R}\boldsymbol{Y}^T\boldsymbol{w}_y\\ \text{s.t.}\quad & \boldsymbol{w}_x^T\boldsymbol{X}\boldsymbol{S}_{xx}\boldsymbol{X}^T\boldsymbol{w}_x=1\\ & \boldsymbol{w}_y^T\boldsymbol{Y}\boldsymbol{S}_{yy}\boldsymbol{Y}^T\boldsymbol{w}_y=1\end{aligned} \tag{7}$$

where $\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]$, $\boldsymbol{Y}=[\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n]$, $\boldsymbol{R}=2\boldsymbol{S}_{xy}+\boldsymbol{S}_x+\boldsymbol{S}_y$, $\boldsymbol{S}_{xx}=\boldsymbol{D}_{xx}-\boldsymbol{S}_x\circ\boldsymbol{S}_x$, $\boldsymbol{S}_{yy}=\boldsymbol{D}_{yy}-\boldsymbol{S}_y\circ\boldsymbol{S}_y$, $\boldsymbol{S}_{xy}=\boldsymbol{D}_{xy}-\boldsymbol{S}_x\circ\boldsymbol{S}_y$, the symbol $\circ$ denotes an operator such that $(\boldsymbol{A}\circ\boldsymbol{B})_{ij}=\boldsymbol{A}_{ij}\boldsymbol{B}_{ij}$ for matrices $\boldsymbol{A}$, $\boldsymbol{B}$ with the same size and $\boldsymbol{A}_{ij}$ denotes the ij-th entry of $\boldsymbol{A}$, $\boldsymbol{D}_{xx}(\boldsymbol{D}_{yy},\boldsymbol{D}_{xy})$ is a diagonal matrix of size n by n, and its i-th diagonal entry equal to the sum of the entries in the i-th row of the matrix $\boldsymbol{S}_x\circ\boldsymbol{S}_x(\boldsymbol{S}_y\circ\boldsymbol{S}_y,\boldsymbol{S}_x\circ\boldsymbol{S}_y)$. More details can be found in the Appendix.

The reason why we use sparse reconstruction to build similarity matrices is that the sparsity is not only an important way to encode the domain knowledge, but also beneficial to computational tractability. The sparest representation has natural discriminating power: considering face images, the most compact expression of a certain face image is generally given by the face images from the same class. Furthermore, the sparse representation achieved by minimizing a ℓ_1 regularization related objective function chooses its neighborhood automatically and does not have to encounter model parameters such as the neighborhood size and heat kernel width, which are generally difficult to set in practice. It is worth noting that the sparsity in our method means the sparse representation instead of controlling the zero values in projective vectors, which indicates that our proposed approach is different from those sparse CCA.

According to Eq. (3), the optimization problem of CSCCA can be solved by following generalized eigenvalue decomposition:

$$\begin{bmatrix}&\boldsymbol{X}\boldsymbol{R}\boldsymbol{Y}^T\\ \boldsymbol{Y}\boldsymbol{R}\boldsymbol{X}^T&\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{bmatrix}=\lambda\begin{bmatrix}\boldsymbol{X}\boldsymbol{S}_{xx}\boldsymbol{X}^T&\\ &\boldsymbol{Y}\boldsymbol{S}_{yy}\boldsymbol{Y}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{bmatrix} \tag{8}$$

Following the idea in [36], Eq. (8) can be efficiently computed via singular value decomposition (SVD) technique. The details of the solution are presented in Algori thm 1. After obtaining d eigenvectors for each view corresponding to d generalized eigenvalues λ_i, $i=1,\ldots,d$, we denote them as $\boldsymbol{W}_x=[\boldsymbol{w}_x^1,\ldots,\boldsymbol{w}_x^d]$ and $\boldsymbol{W}_y=[\boldsymbol{w}_y^1,\ldots,\boldsymbol{w}_y^d]$. For samples $\boldsymbol{x}$, $\boldsymbol{y}$, the features can be extracted as follows [37]

$$\begin{aligned}&\boldsymbol{W}_x^T\boldsymbol{x}+\boldsymbol{W}_y^T\boldsymbol{y}\\ &\begin{bmatrix}\boldsymbol{W}_x^T\boldsymbol{x}\\ \boldsymbol{W}_y^T\boldsymbol{y}\end{bmatrix}\end{aligned} \tag{9}$$

The two feature combination methods in Eq. (9) are referred to as parallel combination and serial combination, denoted as FFS1 and FFS2, respectively. With the fused features, we can implement classification tasks through any classifier. In this paper, we use the nearest neighbor classifier.

Algorithm 1. CSCCA.

Input: Training sets $\boldsymbol{X}\in\mathbb{R}^{p\times n}$, $\boldsymbol{Y}\in\mathbb{R}^{q\times n}$
Output: Projection matrices $\boldsymbol{W}_x,\boldsymbol{W}_y$

1. Construct $\boldsymbol{S}_x,\boldsymbol{S}_y$.
2. Define $\boldsymbol{R}=\boldsymbol{S}_{xy}+\boldsymbol{S}_x+\boldsymbol{S}_y$;
3. Compute matrices $\tilde{\boldsymbol{C}}_{xy}=\boldsymbol{X}\boldsymbol{R}\boldsymbol{Y}^T$, $\tilde{\boldsymbol{C}}_{xx}=\boldsymbol{X}\boldsymbol{S}_{xx}\boldsymbol{X}^T$, $\tilde{\boldsymbol{C}}_{yy}=\boldsymbol{Y}\boldsymbol{S}_{yy}\boldsymbol{Y}^T$.
4. Compute matrix $\boldsymbol{H}=\tilde{\boldsymbol{C}}_{xx}^{-\frac{1}{2}}\tilde{\boldsymbol{C}}_{xy}\tilde{\boldsymbol{C}}_{yy}^{-\frac{1}{2}}$;
5. Perform SVD decomposition on $\boldsymbol{H}$: $\boldsymbol{H}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$;
6. Choose $\boldsymbol{U}=[U_1\ldots U_d]$, $\boldsymbol{V}=[V_1\ldots V_d]$, $d<n$;
7. Obtain $\boldsymbol{W}_x=\tilde{\boldsymbol{C}}_{xx}^{-\frac{1}{2}}\boldsymbol{U}$, $\boldsymbol{W}_y=\tilde{\boldsymbol{C}}_{yy}^{-\frac{1}{2}}\boldsymbol{V}$;

2.3. Kernelized CSCCA

Kernel methods have achieved a remarkable flourish in dimensionality reduction research and amount of kernel based learning algorithms such as KPCA [38], KICA [39] and KCCA [36] have been proposed. According to pattern separability theorem [40], after mapping into a high dimension feature space a complicated pattern classification problem will be more linear separable than in the original low dimension input space. With the help of kernel trick, we extend CSCCA to its kernel version (KCSCCA) to enhance the classification ability for nonlinear problems.

The solution of CSCCA can be expressed as the linear combination of samples, $\boldsymbol{w}_x=\boldsymbol{X}\boldsymbol{\alpha}$ and $\boldsymbol{w}_y=\boldsymbol{Y}\boldsymbol{\beta}$, where $\boldsymbol{\alpha}\in\mathbb{R}^n$ and $\boldsymbol{\beta}\in\mathbb{R}^n$ are the combination coefficients. Assume the nonlinear mapping $\phi:\boldsymbol{x}\mapsto\phi(\boldsymbol{x})$ and $\psi:\boldsymbol{y}\mapsto\psi(\boldsymbol{y})$, which maps samples into feature space. Let $\phi(\boldsymbol{X})=[\phi(\boldsymbol{x}_1),\ldots,\phi(\boldsymbol{x}_n)]$ and $\psi(\boldsymbol{Y})=[\psi(\boldsymbol{y}_1),\ldots,\psi(\boldsymbol{y}_n)]$ denote the sample matrices in feature space. Because KCSCCA is performing CSCCA in feature space, the solution of KCSCCA $\boldsymbol{w}_\phi$ and $\boldsymbol{w}_\psi$ can be expressed as the linear combination of samples $\{\phi(\boldsymbol{x}_i)\}_{i=1}^n$ and $\{\psi(\boldsymbol{y}_i)\}_{i=1}^n$, i.e. $\boldsymbol{w}_\phi=\phi(\boldsymbol{X})\boldsymbol{\alpha}$ and $\boldsymbol{w}_\psi=\psi(\boldsymbol{Y})\boldsymbol{\beta}$. The objective function of KCSCCA is to optimize

$$\begin{aligned}\max_{\boldsymbol{\alpha}\boldsymbol{\beta}}\quad & \boldsymbol{\alpha}^T\boldsymbol{K}_x\boldsymbol{R}\boldsymbol{K}_y\boldsymbol{\beta}\\ \text{s.t.}\quad & \boldsymbol{\alpha}^T\boldsymbol{K}_x\boldsymbol{S}_{xx}\boldsymbol{K}_x\boldsymbol{\alpha}=1\end{aligned}$$

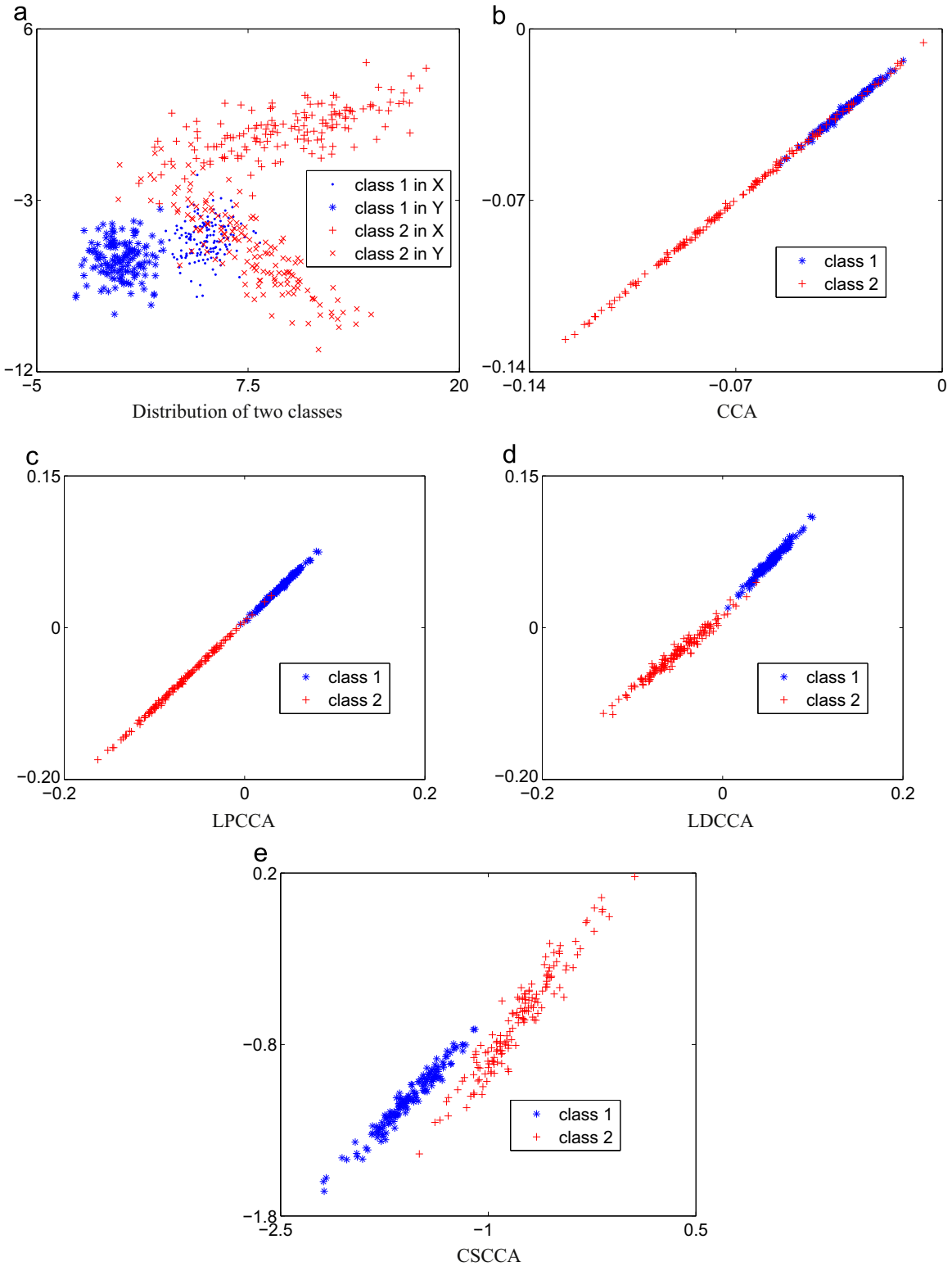


Fig. 1. Distribution of the first pair of features extracted by CCA, LPCCA, LDCCA and CSCCA.

$$\boldsymbol{\beta}^T \boldsymbol{K}_y \boldsymbol{S}_{yy} \boldsymbol{K}_y \boldsymbol{\beta} = 1 \tag{10}$$

where $\boldsymbol{K}_x = \phi(\boldsymbol{X})^T \phi(\boldsymbol{X})$ and $\boldsymbol{K}_y = \psi(\boldsymbol{Y})^T \psi(\boldsymbol{Y})$. It is worth nothing that we still compute the reconstruction matrices $\boldsymbol{S}_x$ and $\boldsymbol{S}_y$ in the original space rather than in the feature space. $K_x(\cdot,\cdot)$ and $K_y(\cdot,\cdot)$ are kernel functions by which we can obtain the inner product of samples in the feature space without knowing the explicit form of ϕ and ψ. The solution of Eq. (10) is similar to that in CSCCA.

When we get the generalized eigenvectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$, $\boldsymbol{W}_\phi$ and $\boldsymbol{W}_\psi$ can be denoted as follows:

$$\begin{aligned} \boldsymbol{W}_\phi &= \phi(\boldsymbol{X})[\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d] = [\boldsymbol{w}_{\phi 1}, \ldots, \boldsymbol{w}_{\phi d}] \\ \boldsymbol{W}_\psi &= \psi(\boldsymbol{Y})[\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d] = [\boldsymbol{w}_{\psi 1}, \ldots, \boldsymbol{w}_{\psi d}] \end{aligned} \tag{11}$$

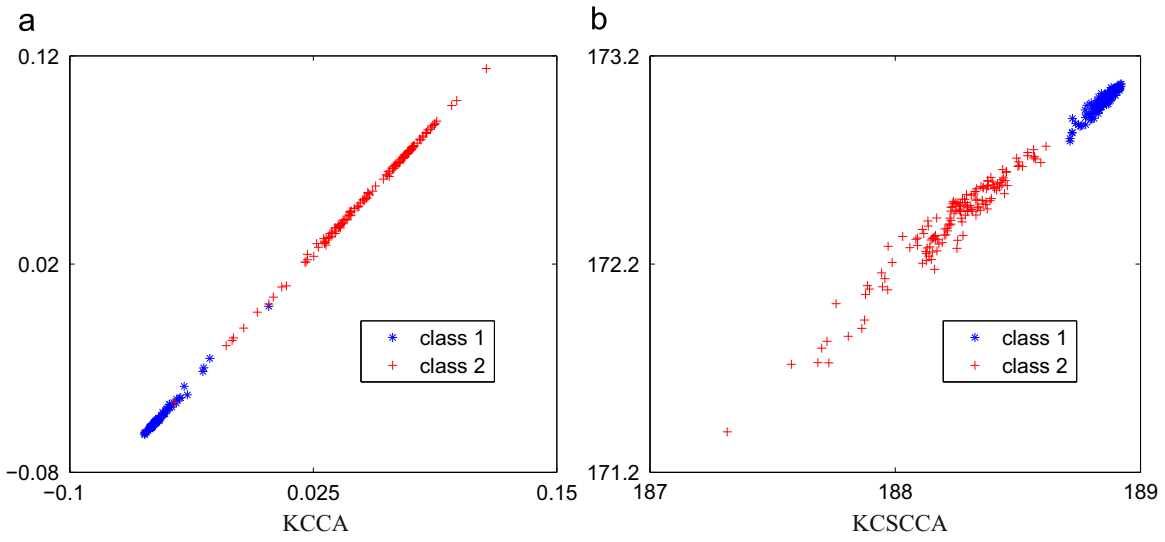


Fig. 2. Distribution of the first pair of features extracted by KCCA and KCSCCA.

Table 1
Recognition accuracies on multiple features database (FFS1).

X	Y	CCA	LPCCA	DCCA	KCCA	LDCCA	CSCCA	KCSCCA
fac	fou	0.8683	0.9398	0.9512	0.8691	0.9515	**0.9828**$^+$	**0.9706***
fac	kar	0.9632	0.9687	0.9739	0.9476	0.9777	**0.9841**$^+$	**0.9801***
fac	mor	0.7584	0.8011	0.9076	0.6867	0.8701	**0.9625**$^+$	**0.9713***
fac	pix	0.9471	0.9611	0.9733	0.9440	0.9740	**0.9836**$^+$	**0.9832***
fac	zer	0.8529	0.9275	0.9603	0.8243	0.9659	**0.9834**$^+$	**0.9748***
fou	kar	0.8970	0.9483	0.9380	0.8651	0.9374	**0.9744**$^+$	**0.9884***
fou	mor	0.7581	0.7027	0.8102	0.6593	0.8046	**0.8179**	**0.8419***
fou	pix	0.8295	0.8547	0.9318	0.8662	0.9306	**0.9745**$^+$	**0.9886***
fou	zer	0.8249	0.8424	0.8382	0.8206	0.8353	**0.8571**$^+$	**0.8801***
kar	mor	0.7872	0.8245	0.8902	0.6754	0.8612	**0.9318**$^+$	**0.9784***
kar	pix	0.9627	0.9739	0.9459	0.9545	0.9447	**0.9781**$^+$	**0.9823***
kar	zer	0.9127	0.9613	0.9368	0.8309	0.9474	**0.9719**$^+$	**0.9806***
mor	pix	0.7246	0.7178	0.8784	0.6716	0.8530	**0.9426**$^+$	**0.9830***
mor	zer	0.7194	0.6922	0.7939	0.6662	**0.7967**	0.7816	**0.8272***
pix	zer	0.8232	0.9310	0.9307	0.8269	0.9373	**0.9719**$^+$	**0.9827***

Table 2
Recognition accuracies on multiple features database (FFS2).

X	Y	CCA	LPCCA	DCCA	KCCA	LDCCA	CSCCA	KCSCCA
fac	fou	0.8776	0.9613	0.9787	0.8691	**0.9796**	**0.9866**$^+$	0.9595
fac	kar	0.9638	0.9697	0.9786	0.9476	**0.9804**	**0.9841**$^+$	0.9785
fac	mor	0.7700	0.8418	0.9284	0.6937	0.8950	**0.9726**$^+$	**0.9707***
fac	pix	0.9468	0.9620	0.9750	0.9440	0.9762	**0.9834**$^+$	**0.9831***
fac	zer	0.8601	0.9501	0.9768	0.8243	**0.9787**	**0.9825**$^+$	0.9750
fou	kar	0.9227	0.9656	0.9684	0.8651	0.9665	**0.9816**$^+$	**0.9900***
fou	mor	0.7636	0.7805	**0.8252**	0.6565	0.8215	0.8236	**0.8418***
fou	pix	0.8476	0.8912	0.9630	0.8662	0.9587	**0.9832**$^+$	**0.9902***
fou	zer	0.8350	0.8529	0.8464	0.8209	0.8499	**0.8647**$^+$	**0.8703***
kar	mor	0.8156	0.8828	0.9224	0.6788	0.9001	**0.9411**$^+$	**0.9784***
kar	pix	0.9630	0.9737	0.9411	0.9545	0.9579	**0.9776**	**0.9824***
kar	zer	0.9203	0.9697	0.9587	0.8311	0.9611	**0.9746**$^+$	**0.9796***
mor	pix	0.7500	0.7479	0.9103	0.6726	0.8831	**0.9535**$^+$	**0.9829***
mor	zer	0.7482	0.7293	**0.8120**	0.6709	0.8096	0.7989	**0.8273***
pix	zer	0.8404	0.9548	0.9525	0.8269	0.9561	**0.9757**$^+$	**0.9829***

For any sample $(\boldsymbol{x},\boldsymbol{y})$, we can extract features as follows:

$$\boldsymbol{W}_\phi^T\phi(\boldsymbol{x}) = [\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_d]^T\phi(\boldsymbol{X})^T\phi(\boldsymbol{x}) = [\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_d]^T[K_x(\boldsymbol{x}_1,\boldsymbol{x}),\ldots,K_x(\boldsymbol{x}_n,\boldsymbol{x})]^T \in \mathbb{R}^n$$

$$\boldsymbol{W}_\psi^T\psi(\boldsymbol{y}) = [\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d]^T\psi(\boldsymbol{Y})^T\psi(\boldsymbol{y}) = [\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d]^T[K_y(\boldsymbol{y}_1,\boldsymbol{y}),\ldots,K_y(\boldsymbol{y}_n,\boldsymbol{y})]^T \in \mathbb{R}^n \tag{12}$$

With the extracted features by KCSCCA, we can obtain the combined representations through feature fusion strategies FFS1 and FFS2 in Eq. (9), respectively.

3. Experiments and results

In this section, we evaluate our methods CSCCA and KCSCCA on both artificial and real world data sets including multiple features data set[1] and two face databases,[2] ORL and PIE. We first use CSCCA and KCSCCA to extract features from multi-view data as well as other five related methods, CCA, LPCCA, DCCA, KCCA and LDCCA. Then the nearest neighborhood classifier is employed to evaluate the classification performance of different methods.

The type of kernel and its related parameters are important parameters to be determined in kernel based methods. In our experiments, we adopt the Gaussian kernel, i.e. $K_x(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma_x^2)$ and $K_y(\boldsymbol{y}_i,\boldsymbol{y}_j) = \exp(-\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2/2\sigma_y^2)$. The optimal values of kernel widths ${\sigma_x}^2$ and ${\sigma_y}^2$ are found by searching the following parameter spaces

$$[2^{-2},2^{-1},2^0,2^1,2^2]\times\sigma_{x_0}^2 \quad \text{and} \quad [2^{-2},2^{-1},2^0,2^1,2^2]\times\sigma_{y_0}^2$$

$$\sigma_{x_0}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2 \quad \text{and} \quad \sigma_{y_0}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2 \tag{13}$$

where $\sigma_{x_0}^2$ and $\sigma_{y_0}^2$ are the mean square distances of samples $\boldsymbol{X}$ and $\boldsymbol{Y}$. In order to control the sparsity of weight matrices in CSCCA and KCSCCA, a balancing factor η is introduced, which is optimized by searching in the range $\eta \in [0.001, 0.01, 0.1, 1, 10, 100]$. Furthermore, the effect of η will be demonstrated in practice in Section 3.4. There is also a parameter k in LPCCA and LDCCA, i.e. the number of the nearest neighbors, which will be searched in the range from one to the number of training data. All the parameters are chosen

[1] http://archive.ics.uci.edu/ml
[2] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

Table 3
Recognition accuracies on ORL database.

	X	Y	CCA	LPCCA	DCCA	KCCA	LDCCA	CSCCA	KCSCCA
FFS1	64×64	32×32	0.8735	0.8585	0.9080	0.8705	0.9250	**0.9490**$^{+}$	**0.9325***
	64×64	Wav	0.8920	0.8555	0.9120	0.8895	0.9270	**0.9495**$^{+}$	**0.9405***
	64×64	Lbp	0.9175	0.8880	**0.9820**	0.9095	**0.9825**	0.9790	0.9785
	32×32	Wav	0.8860	0.8710	0.9320	0.8855	0.9315	**0.9565**$^{+}$	**0.9440**
	32×32	Lbp	0.9210	0.8835	**0.9805**	0.9135	**0.9810**	0.9785	0.9780
	Wav	Lbp	0.9220	0.8905	**0.9860**	0.9165	**0.9830**	0.9795	0.9820
FFS2	64×64	32×32	0.8735	0.8585	0.9075	0.8705	0.92220	**0.9495**$^{+}$	**0.9350***
	64×64	Wav	0.8920	0.8555	0.9115	0.8895	0.9290	**0.9490**$^{+}$	**0.9410***
	64×64	Lbp	0.9200	0.8890	**0.9845**	0.9095	0.9845	**0.9850**	0.9820
	32×32	Wav	0.8865	0.8715	0.9320	0.8855	0.9270	**0.9565**$^{+}$	**0.9440**
	32×32	Lbp	0.9200	0.8810	**0.9850**	0.9135	0.9830	**0.9830**	0.9795
	Wav	Lbp	0.9210	0.8925	**0.9845**	0.9165	0.9830	**0.9865**	0.9820

Table 4
Classification accuracies on PIE database.

	X	Y	CCA	LPCCA	DCCA	KCCA	LDCCA	CSCCA	KCSCCA
FFS1	64×64	32×32	0.9314	0.9345	0.9462	0.9240	0.9594	**0.9672**$^{+}$	**0.9750***
	64×64	Wav	0.9331	0.9365	0.9470	0.9330	0.9613	**0.9682**$^{+}$	**0.9755***
	64×64	Lbp	0.9503	0.9524	**0.9858**	0.9349	0.9822	0.9834	**0.9844**
	32×32	Wav	0.9518	0.9547	0.9668	0.9452	0.9607	**0.9688**	**0.9762***
	32×32	Lbp	0.9495	0.9512	**0.9849**	0.9328	0.9812	0.9840	**0.9845**
	Wav	Lbp	0.9508	0.9526	0.9656	0.9349	0.9820	**0.9841**$^{+}$	**0.9842***
FFS2	64×64	32×32	0.9313	0.9345	0.9457	0.9240	0.9587	**0.9674**$^{+}$	**0.9748***
	64×64	Wav	0.9331	0.9365	0.9470	0.9330	0.9612	**0.9682**$^{+}$	**0.9757***
	64×64	Lbp	0.9552	0.9543	**0.9870**	0.9352	0.9842	**0.9853**	0.9842
	32×32	Wav	0.9521	0.9547	0.9665	0.9453	0.9601	**0.9689**	**0.9761***
	32×32	Lbp	0.9538	0.9530	**0.9861**	0.9332	0.9826	**0.9856**	0.9835
	Wav	Lbp	0.9548	0.9550	0.9672	0.9353	0.9801	**0.9854**$^{+}$	**0.9840***

in their respective ranges through 10-fold cross validation on training data. For each data set, the experiments are repeated 10 times and the averaged accuracies are presented. For all methods, the subspace dimensionality is set from 1 to 50, such that we choose the best result of different algorithms with different subspace dimensionality. If the maximum dimension of the two views is less than 50, the subspace dimensionality is set to the maximum dimension of the two views.

3.1. Toy problem

We consider a two-class classification problem containing 150 two-dimensional samples with two views denoted as $\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2]$ and $\boldsymbol{Y}=[\boldsymbol{Y}_1,\boldsymbol{Y}_2]$, where $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ represent the i $(i=1,2)$ class samples respectively. $\boldsymbol{X}_i$ satisfies the Gaussian distribution $N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_1=[10.18,0.66]^T$, $\boldsymbol{\Sigma}_1=[14,3.75;3.75,15]$, $\boldsymbol{\mu}_2=[5,-5]^T$, $\boldsymbol{\Sigma}_2=[1,0;0,1]$. We get $\boldsymbol{Y}$ from the following transformation:

$$\boldsymbol{y}_i=\boldsymbol{W}^T\boldsymbol{x}_i+\boldsymbol{\varepsilon},\quad i=1,\ldots,150. \tag{14}$$

where $\boldsymbol{W}=[0.6,-\sqrt{1/2};0.8,\sqrt{1/2}]$ and $\boldsymbol{\varepsilon}$ is additional Gaussian noise which follows $N(\boldsymbol{\mu}_\varepsilon,\boldsymbol{\Sigma}_\varepsilon)$, in which $\boldsymbol{\mu}_\varepsilon=[1,1]^T$, $\boldsymbol{\Sigma}_\varepsilon=[0.01,0;0,0.01]$.

The distribution of the data is shown in Fig. 1(a). And Fig. 1(b)–(e) plots the distribution of the first pair of features $(\boldsymbol{w}_{x1}^T\boldsymbol{x},\boldsymbol{w}_{y1}^T\boldsymbol{y})$ extracted by CCA, LPCCA, LDCCA and CSCCA, respectively. Fig. 2 compares the result of KCCA with that of KCSCCA. From the experimental results, we can see that CCA can reveal the linear relationship of the original features, but the two classes are severely overlapped. The overlapped phenomena still exist in the result of LPCCA and LDCCA. In CSCCA, two classes are separated better than CCA, LPCCA and LDCCA and the data features extracted by KCSCCA are separated completely while there is still some overlap in KCCA. Figs. 1 and 2 indicate that the features extracted by CSCCA and KCSCCA can preserve much discriminate information.

3.2. Multiple feature recognition

Multiple Feature data set which consists 2000 samples of handwritten digits 0–9 with six different feature sets is picked out from UCI machine learning repository. Each class contains 200 examples. The six feature sets which describe the data set from different views include profile correlations, flourier coefficient of the character shapes, Karhunen–Loève expansion coefficient, pixel average, Zernike moments and morphological characteristics. The name and the dimension of those features are (fac, 216), (fou, 76), (kar, 64), (pix, 240), (zer, 47) and (mor, 6).

We choose two sets of these six features as $\boldsymbol{X}$ view and $\boldsymbol{Y}$ view, so there are 15 ($C_6^2=15$) view settings in total. For each class, we randomly choose 100 samples in each class for training, the other samples for testing. Thus there are 1000 training samples and 1000 testing samples. We extract features by CSCCA and KCSCCA as well as CCA, LPCCA, DCCA, KCCA and LDCCA. Then we perform classification based on the extracted features using nearest neighbor classifier. Average results over 10 independent runs of two feature fusion strategies are reported in Tables 1 and 2. We have also performed a statistical test (paired t test at 95% significance level)

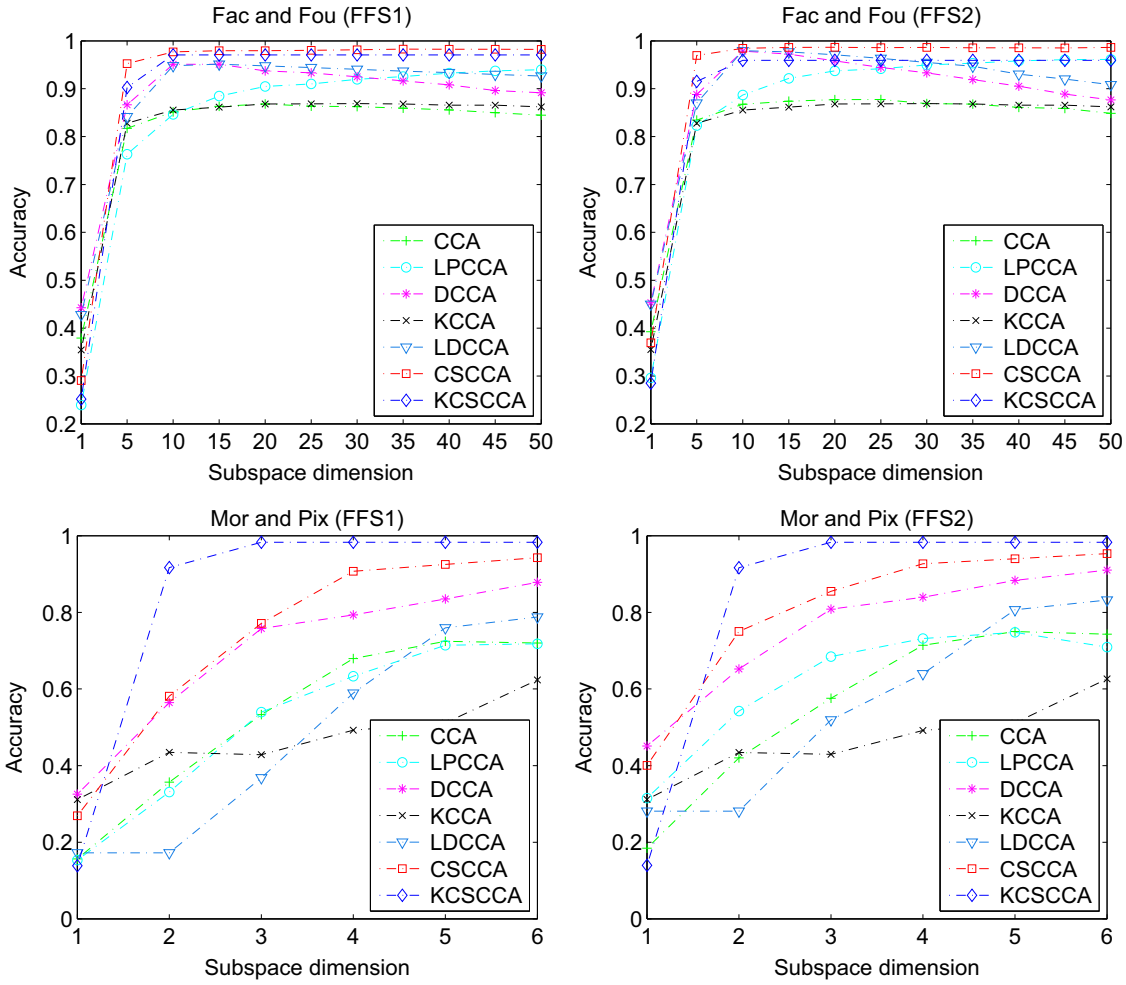


Fig. 3. Recognition accuracies on multiple features database with a different dimension.

between our methods and the best performing method excluding CSCCA and KCSCCA. In Tables 1 and 2, we use bold type and other three symbols to point out the results. In the same classification task, the highest accuracy of the results is represented in bold type with underline and the second-highest one is shown in bold type. The plus symbol + denotes that the classification result obtained by CSCCA is statistically significant. Analogously, the asterisk notation ∗ indicates that the accuracy of KCSCCA is statistically significant. The bold type and the three symbols in Tables 3 and 4 express the same meaning in Tables 1 and 2. It can be seen from Table 1, the accuracy of CSCCA and KCSCCA is mostly better than other methods. In contrast, the LDCCA provides the second best result only once. A similar phenomenon can be found in Table 2. DCCA provides the second best result only two times and LDCCA achieves the second best performance only three times. It also shows that in most cases our methods perform significantly better than other methods. Fig. 3 shows the effect of subspace dimensionality d on performances of the algorithms. Two typical view settings are picked. It indicates that in most cases our methods achieve better accuracies than other methods at various dimensions.

3.3. Face recognition

In this subsection, we carry out experiments on two databases including ORL and PIE to verify the performances of CSCCA and KCSCCA for face recognition. Following the preprocessing steps in [41], face areas are cropped and the size of each cropped image is 64×64 pixels, which is used as the first view. Actually images with different resolutions can provide information at different levels and can be regarded as different views of images, thus each image is resized to 32×32 pixels to generate the second view. Double Daubechies wavelet transform is performed on all images and the low frequency components are chosen as the third view. We use the local binary pattern (LBP) histograms which have been proved to be efficient patterns for representing face images [42], as the fourth view. So there are 6 view combinations in our experiments. Our methods are compared with CCA, LPCCA, DCCA, KCCA and LDCCA in each combination and the results of 10 independent runs are reported.

3.3.1. ORL database

ORL data set contains 40 persons' images and everyone has 10 images which are taken at different times, varying the lighting, facial expressions (open or closed eyes, smiling or not smiling) and facial details (glasses or no glasses). All images are all sized 112×92 pixels with 256-level gray scale. We randomly choose five images of one person for training and the rest images for testing, so that here are 200 training images and 200 testing images in total. Table 3 presents the results of different algorithms on ORL database. The table shows that CSCCA outperforms other methods in FFS2 and achieves the highest accuracy on ORL database. KCSCCA also gets good performance in several view settings. Fig. 4 validates the effectiveness of the proposed methods with different subspace dimensions.

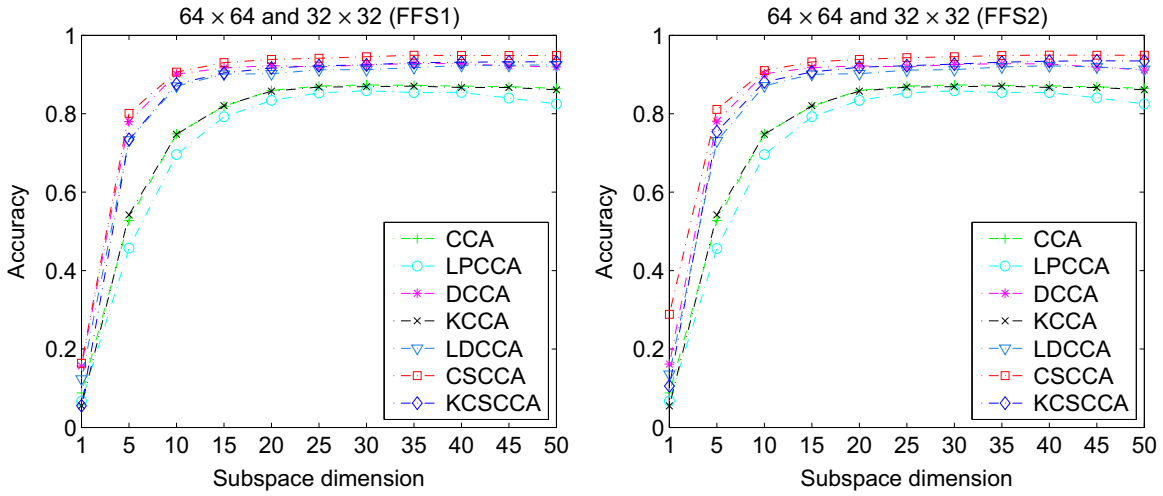


Fig. 4. Recognition accuracies on ORL database with a different dimension.

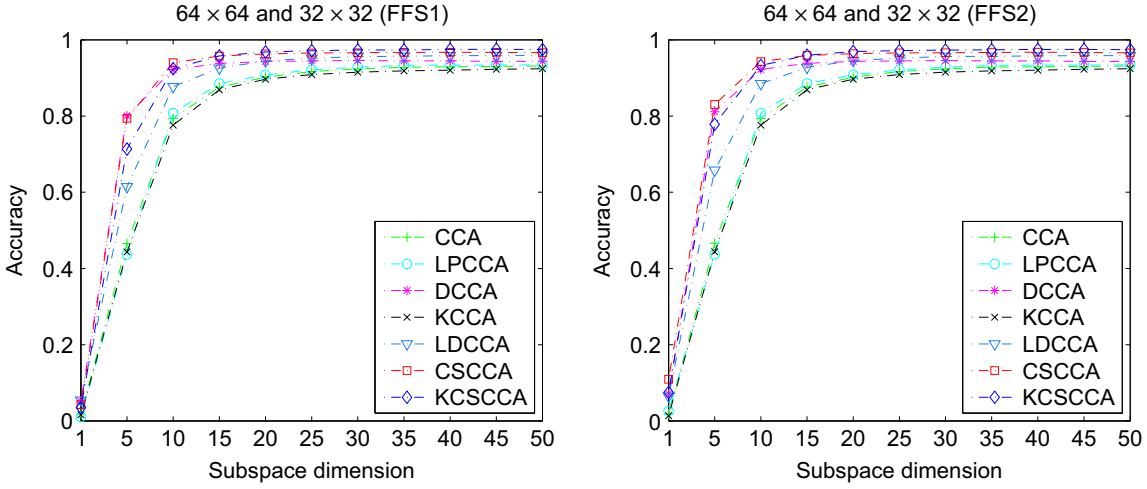


Fig. 5. Recognition accuracies on PIE database with a different dimension.

3.3.2. *PIE database*

PIE data set contains 3329 images of 68 individuals, in which subject of number 36 has only 46 images and the rest have 49 images. Because PIE is much larger than ORL database, 20 images of one person are chosen randomly as training examples and the remaining images are for classification. Table 4 gives the classification accuracies of different algorithms on PIE data set. From Table 4, we can see that CSCCA and KCSCCA achieve the mostly highest accuracy in a different view combination. Other methods only DCCA gets best performance four times. Fig. 5 also shows that our methods are better than other methods with a different dimension.

3.4. *Effect of parameter* η

The η in constructing the weight matrices is an important parameter in CSCCA. However, to the best of our knowledge, it is challenging to determine the appropriate value of η in advance. In our experiments, we choose the optimal value of the parameter by searching in the range $\eta \in [0.001, 0.01, 0.1, 1, 10, 100]$ via 10-fold cross validation on the training set. In order to see how different values of η will affect the final performance, we plot the curves of CSCCA method when different values of parameters are used in experiments. Fig. 6 shows the results of three different view settings using feature fusion strategy one (FFS1) on multiple features data set, ORL database and PIE database, respectively.

As can be seen from Fig. 6, parameter η affects the final performance. Specifically, when η is in [0.001 1], CSCCA would reach the preferable accuracy. When η is larger than 10, the classification accuracy drops down dramatically. The reason is that when η increases, the ℓ_1 regularization term of Eq. (6) becomes less important and there will be many nonzero values in the reconstructive weight vector, which will overlook the intrinsic geometric structure of data. In contrast, when η is between 10^{-3} and 1, the weight matrices obtained via Eq. (6) will catch the intrinsic local structure of data and the perfect performances will be achieved.

4. Conclusion

In this paper, we have presented a novel CCA model called CSCCA for multi-view feature extraction. CSCCA can preserve the within-class local structure information. Besides, CSCCA considers not only the correlations between the two views of the same sample but also the cross-view correlations between within-class samples. The local structure and cross-view correlation are automatically obtained in similarity matrices, which are constructed by performing sparse representation between within-class samples.

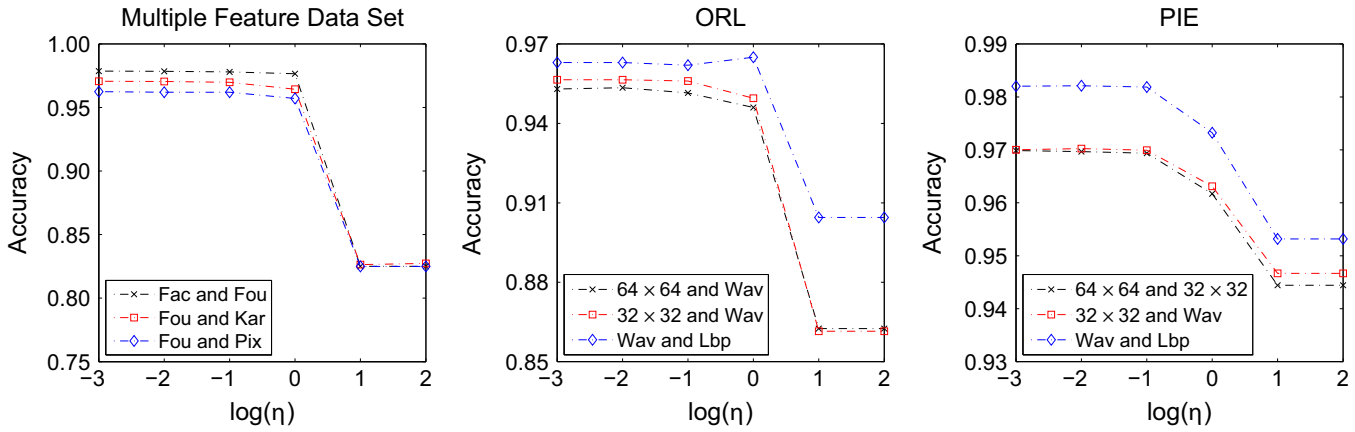


Fig. 6. Recognition accuracies on multiple features data set, ORL and PIE database with a different η in FFS1.

The sparest representation has natural discriminating power and is also beneficial to computational tractability. Furthermore, we propose a kernel generalization of CSCCA (KCSCCA). The experimental results on a series of databases show the effective performance of CSCCA and KCSCCA compared with related methods.

In KCSCCA, we use the weight matrices calculated in original space rather than in the feature space. We will investigate whether the performances can be improved with the weight matrices in feature space. In current work, we use CSCCA and KCSCCA only in classification task. However, it is interesting to apply the proposed methods to other tasks such as data visualization. There are usually more than two view feature sets in real work applications, so how to handle multiple views is very important in our following research. In our future work, we will extend our proposed approach to semi-supervised situation. Nonnegative matrix factorization [43,44] instead of sparse representation will be adopted to obtain weight matrices.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 61422204, 61473149), the Jiangsu Natural Science Foundation for Distinguished Young Scholar (No. BK20130034), the Specialized Research Fund for the Doctoral Program of Higher Education under grant (No. 20123218110009), and the NUAA Fundamental Research Funds under grant (No. NE2013105).

Appendix A. The derivation of Eq. (7)

The objective function of Eq. (7) contains two terms and the second term can be easily rewritten into compact form:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(S_{ij}^{x}+S_{ij}^{y})\boldsymbol{x}_i\boldsymbol{y}_j^T=\boldsymbol{X}(\boldsymbol{S}_x+\boldsymbol{S}_y)\boldsymbol{Y}^T \tag{15}$$

Expand the first term:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}(\boldsymbol{x}_i-\boldsymbol{x}_j)S_{ij}^{y}(\boldsymbol{y}_i-\boldsymbol{y}_j)^T=\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}S_{ij}^{y}(\boldsymbol{x}_i\boldsymbol{y}_i^T+\boldsymbol{x}_j\boldsymbol{y}_j^T-\boldsymbol{x}_i\boldsymbol{y}_j^T-\boldsymbol{x}_j\boldsymbol{y}_i^T) \tag{16}$$

The quadratic term

$$\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}S_{ij}^{y}\boldsymbol{x}_i\boldsymbol{y}_i^T=\sum_{j=1}^{n}\sum_{i=1}^{n}S_{ij}^{x}S_{ij}^{y}\boldsymbol{x}_i\boldsymbol{y}_i^T$$

$$=\sum_{j=1}^{n}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\begin{bmatrix}S_{1j}^{x}S_{1j}^{y} & & \\ & \ddots & \\ & & S_{nj}^{x}S_{nj}^{y}\end{bmatrix}\begin{bmatrix}\boldsymbol{y}_1^T\\ \vdots\\ \boldsymbol{y}_n^T\end{bmatrix}$$

$$=\boldsymbol{X}\begin{bmatrix}\sum_{j=1}^{n}S_{1j}^{x}S_{1j}^{y} & & \\ & \ddots & \\ & & \sum_{j=1}^{n}S_{nj}^{x}S_{nj}^{y}\end{bmatrix}\boldsymbol{Y}^T=\boldsymbol{X}\cdot diag(\boldsymbol{S}_x\circ\boldsymbol{S}_y)\cdot\boldsymbol{Y}^T \tag{17}$$

Similarly,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}S_{ij}^{y}\boldsymbol{x}_i\boldsymbol{y}_j^T=\boldsymbol{X}\begin{bmatrix}\sum_{i=1}^{n}S_{i1}^{x}S_{i1}^{y} & & \\ & \ddots & \\ & & \sum_{i=1}^{n}S_{in}^{x}S_{in}^{y}\end{bmatrix}\boldsymbol{Y}^T=\boldsymbol{X}\cdot diag(\boldsymbol{S}_x\circ\boldsymbol{S}_y)\cdot\boldsymbol{Y}^T \tag{18}$$

The cross term $\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}S_{ij}^{y}\boldsymbol{x}_i\boldsymbol{y}_j^T=\sum_{i=1}^{n}\boldsymbol{x}_i\sum_{j=1}^{n}S_{ij}^{x}S_{ij}^{y}\boldsymbol{y}_j^T=\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{b}_i^T\boldsymbol{Y}^T$, where $\boldsymbol{b}_i^T=(S_{i1}^{x}S_{i1}^{y},\ldots,S_{in}^{x}S_{in}^{y})$ is the i-th row of matrix $\boldsymbol{S}_x\circ\boldsymbol{S}_y$, then Eq. (18) can be written as $\boldsymbol{X}(\boldsymbol{S}_x\circ\boldsymbol{S}_y)\boldsymbol{Y}^T$. For the symmetry of matrix $\boldsymbol{S}_x$ and $\boldsymbol{S}_y$, we can also get,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}S_{ij}^{y}\boldsymbol{x}_j\boldsymbol{y}_i^T=\boldsymbol{X}(\boldsymbol{S}_x\circ\boldsymbol{S}_y)^T\boldsymbol{Y}^T=\boldsymbol{X}(\boldsymbol{S}_x\circ\boldsymbol{S}_y)\boldsymbol{Y}^T \tag{19}$$

Arrange these equations, we can get the compact form of the first term:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}(\boldsymbol{x}_i-\boldsymbol{x}_j)S_{ij}^{y}(\boldsymbol{y}_i-\boldsymbol{y}_j)^T=2\boldsymbol{X}\,diag(\boldsymbol{S}_x\circ\boldsymbol{S}_y)\boldsymbol{Y}^T-2\boldsymbol{X}(\boldsymbol{S}_x\circ\boldsymbol{S}_y)\boldsymbol{Y}^T=2\boldsymbol{X}\boldsymbol{S}_{xy}\boldsymbol{Y}^T \tag{20}$$

Finally, the compact form of the objective function can be written as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}^{x}(\boldsymbol{x}_i-\boldsymbol{x}_j)S_{ij}^{y}(\boldsymbol{y}_i-\boldsymbol{y}_j)^T+\sum_{i=1}^{n}\sum_{j=1}^{n}(S_{ij}^{x}+S_{ij}^{y})\boldsymbol{x}_i\boldsymbol{y}_j^T=\boldsymbol{X}(2\boldsymbol{S}_{xy}+\boldsymbol{S}_x+\boldsymbol{S}_y)\boldsymbol{Y}^T \tag{21}$$

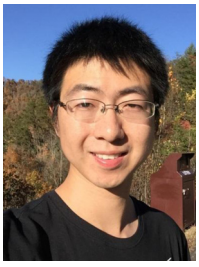
References

[1] M. Liu, D. Zhang, E. Adeli-Mosabbeb, D. Shen, Inherent structure based multi-view learning with multi-template feature representation for alzheimer's disease diagnosis, IEEE Trans. Biomed. Eng., http://dx.doi.org/10.1109/TBME.2015.2496233, in press.

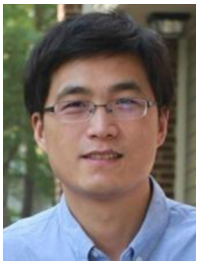
[2] V. Sindhwani, D.S. Rosenberg, An rkhs for multi-view learning and manifold co-regularization, in: Proceedings of the International Conference on Machine Learning (ICML'08), Helsinki, Finland, 2008, pp. 976–983.

[3] S.M. Kakade, D.P. Foster, Multi-view regression via canonical correlation analysis, in: Proceedings of the Conference on Learning Theory (COLT'07), San Diego, CA, USA, 2007, pp. 82–96.

[4] D.R. Hardoon, K. Pelcksman, Pair-wise Cluster Analysis, arXiv preprint arXiv:1009.3601.
[5] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3) (1936) 321–377.
[6] Y. Hel-Or, The Canonical Correlations of Color Images and Their Use for Demosaicing, Technical Report, HP Laboratories Israel, 2004.
[7] F. Wang, D. Zhang, A new locality-preserving canonical correlation analysis algorithm for multi-view dimensionality reduction, Neural Process. Lett. 37 (2) (2013) 135–146.
[8] Y.-H. Yuan, Q.-S. Sun, H.-W. Ge, Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition, Pattern Recognit. 47 (3) (2014) 1411–1424.
[9] M. Liu, D. Zhang, D. Shen, View-centralized multi-atlas classification for Alzheimer's disease diagnosis, Hum. Brain Mapp. 36 (5) (2015) 1847–1865.
[10] C. Zu, B. Jie, M. Liu, S. Chen, D. Shen, D. Zhang, A.D.N. Initiative, et al., Label-aligned multi-task feature learning for multimodal classification of alzheimer's disease and mild cognitive impairment, Brain Imaging Behav. (2015) 1–12.
[11] Y. Lu, D.P. Foster, large scale canonical correlation analysis with iterative least squares, in: Advances in Neural Information Processing Systems (NIPS'14), 2014, pp. 91–99.
[12] T. Sun, S. Chen, J. Yang, P. Shi, A novel method of combined feature extraction for recognition, in: Proceedings of the IEEE International Conference on Data Mining (ICDM'08), Pisa, Italy, 2008, pp. 1043–1048.
[13] J. Zhang, D. Zhang, A novel ensemble construction method for multi-view data using random cross-view correlation between within-class examples, Pattern Recognit. 44 (6) (2011) 1162–1171.
[14] O. Arandjelović, Discriminative extended canonical correlation analysis for pattern set matching, Mach. Learn. 94 (3) (2014) 353–370.
[15] T. Sun, S. Chen, Class label versus sample label-based cca, Appl. Math. Comput. 185 (1) (2007) 272–283.
[16] Y. Shin, C. Park, Analysis of correlation based dimension reduction methods, Int. J. Appl. Math. Comput. Sci. 21 (3) (2011) 549–558.
[17] O. Kursun, E. Alpaydin, O.V. Favorov, Canonical correlation analysis using within-class coupling, Pattern Recognit. Lett. 32 (2) (2011) 134–144.
[18] W.W. Hsieh, Nonlinear canonical correlation analysis by neural networks, Neural Netw. 13 (10) (2000) 1095–1105.
[19] W. Wu, J. He, J. Zhang, A kernelized discriminant analysis algorithm based on modified generalized singular value decomposition, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08), Las Vegas, Nevada, 2008, pp. 1353–1356.
[20] D. Zhang, S. Chen, Clustering incomplete data using kernel-based fuzzy c-means algorithm, Neural Process. Lett. 18 (3) (2003) 155–162.
[21] P.L. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, Int. J. Neural Syst. 10 (5) (2000) 365–377.
[22] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[23] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[24] X. He, P. Niyogi, Locality preserving projections, in: Advances in Neural Information Processing Systems (NIPS'04), vol. 16, Whistler, Canada, 2004, p. 153.
[25] T. Sun, S. Chen, Locality preserving cca with applications to data visualization and pose estimation, Image Vis. Comput. 25 (5) (2007) 531–543.
[26] T. Diethe, D.R. Hardoon, J. Shawe-Taylor, Constructing nonlinear discriminants from multiple data views, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20–24, 2010, Proceedings, Part I Constructing Nonlinear Discriminants from Multiple Data Views, Springer, Berlin, Heidelberg, 2010, pp. 328–343.
[27] Q. Chen, S. Sun, Hierarchical multi-view Fisher discriminant analysis, In: C. Leung, M. Lee, J. Chan (Eds.), Neural Information Processing, Lecture Notes in Computer Science, vol. 5864, Springer, Berlin, Heidelberg, 2009, pp. 289–298.
[28] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12), IEEE, Rhode Island, USA, 2012, pp. 2160–2167.
[29] Y. Peng, D. Zhang, J. Zhang, A new canonical correlation analysis algorithm with local discrimination, Neural Process. Lett. 31 (1) (2010) 1–15.
[30] X. Zhang, N. Guan, Z. Luo, L. Lan, Discriminative locality preserving canonical correlation analysis, in: Pattern Recognition: Chinese Conference, CCPR 2012, Beijing, China, September 24–26, 2012. Proceedings Discriminative Locality Preserving Canonical Correlation Analysis Springer, Berlin, Heidelberg, 2012, pp. 341-349.
[31] D.R. Hardoon, J. Shawe-Taylor, Sparse canonical correlation analysis, Mach. Learn. 83 (3) (2011) 331–353.
[32] A. Lykou, J. Whittaker, Sparse cca using a lasso with positivity constraints, Comput. Stat. Data Anal. 54 (12) (2010) 3144–3157.
[33] A. Kimura, M. Sugiyama, H. Sakano, H. Kameoka, Designing various multivariate analysis at will via generalized pairwise expression, IPSJ Online Trans. 6 (0) (2013) 45–54.
[34] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341.
[35] J. Liu, S. Ji, J. Ye, SLEP: Sparse Learning with Efficient Projections, Arizona State University, 2009. URL ⟨http://www.public.asu.edu/~jye02/Software/SLEP⟩.
[36] T. Melzer, M. Reiter, H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognit. 36 (9) (2003) 1961–1971.
[37] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, D.-S. Xia, A new method of feature fusion and its application in image recognition, Pattern Recognit. 38 (12) (2005) 2437–2448.
[38] J. Shawe-Taylor, C.K. Williams, N. Cristianini, J. Kandola, On the eigenspectrum of the gram matrix and the generalization error of kernel-pca, IEEE Trans. Inf. Theory 51 (7) (2005) 2510–2522.
[39] F.R. Bach, M.I. Jordan, Kernel independent component analysis, J. Mach. Learn. Res. 3 (2002) 1–48.
[40] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electron. Comput. 14 (3) (1965) 326–334.
[41] D. Cai, X. He, Y. Hu, J. Han, T. Huang, Learning a spatially smooth subspace for face recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, USA, 2007, pp. 1–7.
[42] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, In: Proceedings of the European Conference on Computer Vision Computer Vision (ECCV'04), Prague, Czech Republic, 2004, pp. 469–481.
[43] D. Zhang, W. Liu, An efficient nonnegative matrix factorization approach in flexible kernel space, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'09), Pasadena, California, 2009, pp. 1345–1350.
[44] D. Zhang, Z.-H. Zhou, S. Chen, Non-negative matrix factorization on kernels, In: Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI'06), Guilin, China, 2006, pp. 404–412.



Chen Zu received the B.S. and M.S. degrees from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2010 and 2013, respectively. He is currently pursuing a Ph.D. degree at Nanjing University of Aeronautics and Astronautics. His current research interests include neuroimaging analysis, machine learning, pattern recognition, and data mining.



Daoqiang Zhang received the B.S. and Ph.D. degrees in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1999 and 2004, respectively. In 2004, he joined the Department of Computer Science and Engineering, NUAA, as a Lecturer, where he is currently a Professor. His current research interests include machine learning, pattern recognition, data mining, and medical image analysis. He has published over 100 scientific articles in refereed international journals such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, Neuroimage, Human Brain Mapping, and conference proceedings such as International Joint Conferences on Artificial Intelligence, IEEE International Conference on Data Mining, and International Conference on Medical Image Computing and Computer Assisted Interventions.

Dr. Zhang is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence and the Artificial Intelligence and Pattern Recognition Society of the China Computer Federation.