Front. Comput. Sci., 2016, 10(5): 856–869
DOI 10.1007/s11704-015-4538-7

RESEARCH ARTICLE

Multi-view dimensionality reduction via canonical random correlation analysis

Yanyan ZHANG[1*], Jianchun ZHANG[2*], Zhisong PAN (✉)[1], Daoqiang ZHANG (✉)[2]

1 College of Command Information Systems, PLA University of Science and Technology, Nanjing 210007, China
2 Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract Canonical correlation analysis (CCA) is one of the most well-known methods to extract features from multi-view data and has attracted much attention in recent years. However, classical CCA is unsupervised and does not take discriminant information into account. In this paper, we add discriminant information into CCA by using random cross-view correlations between within-class samples and propose a new method for multi-view dimensionality reduction called canonical random correlation analysis (RCA). In RCA, two approaches for randomly generating cross-view correlation samples are developed on the basis of bootstrap technique. Furthermore, kernel RCA (KRCA) is proposed to extract nonlinear correlations between different views. Experiments on several multi-view data sets show the effectiveness of the proposed methods.

Keywords canonical correlation analysis, discriminant, multi-view, dimensionality reduction

1 Introduction

Objects in the real world can be described by several sets of features or views in some cases, and multiple representations of objects can be easily obtained in many applications, such as images and their semantic descriptions, morphological features of handwritten characters and their pixel information. Earlier researches have shown that using complementary information contained in multiple views instead of simply combining them into a big view could increase classification accuracy [1, 2]. In multi-view setting, each view can be viewed as a gross description of particular aspects of observations, so the correlations between different views will contain fine features for representing objects [3]. In recent years, many extensions of multi-view learning techniques have been extended to different kinds of application fields, such as multi-view distance metric learning methods for speaker identification [4] and image processing [5, 6], multiple view clustering and multiple spectral dimensionality reduction [7], multi-view sparse unsupervised dimensionality reduction [8]. In Ref. [3], Xia et al. developed a multi-view spectral embedding (MSE) algorithm to explore the complementary property of different views. Furthermore, in Ref. [9], Xie et al. proposed a multi-view stochastic neighbor embedding (m-SNE) method to integrate heterogeneous features into a unified representation based on probabilistic framework.

Intuitively, complementary information and correlation information between different views can be used for multi-view learning. However, in traditional multi-view learning methods, correlation information have not been well explored. Canonical correlation analysis (CCA), developed by Hotelling [10], the most well-known two-view based method,

Received November 26, 2014; accepted September 25, 2015

E-mail: hotpzs@hotmail.com; dqzhang@nuaa.edu.cn

* These authors contributed equally to this work

is often used to reveal correlation relationships between two sets of features (or views). CCA can be seen as a two-view extension of PCA [11]. Classical CCA works with only two sets of variables and can only find linear relationships between them. So in the past years, many generalizations of CCA are suggested to cope with problems emerging in different fields [12, 13]. The generalization of extending two-set CCA to multi-set CCA is encouraging. Vía et al. [13] reformulated this generalization as a set of coupled least square problems to develop a neural model for CCA. Kernel extension of CCA (KCCA) is first proposed by Akaho [12]. In Ref. [14], Hardoon et al. associated the images with its semantic descriptions through KCCA and retrieved images based on text queries. In Ref. [15], Yang et al. proposed a centerized version of kernel feature with the kernels-as-features idea for CCA. Sun et al. introduced local neighborhood information into CCA and decomposed global nonlinear problem into a set of local linear sub problems [16]. Blaschko et al. proposed semi-supervised kernel canonical correlation analysis [17] and Laplacian regularized KCCA that can find the directions for representing the structure of the data and increasing class separation [18].

CCA is inherently an unsupervised method and the label information can not be utilized in CCA, which limits its classification performance in practice. Intuitively, correlations within the same classes should be superior to the correlations between different classes. In order to make up this shortcoming of classic CCA, label information is taken into account in some supervised or semi-supervised extensions of CCA methods, such as supervised regularized canonical correlation analysis [19], supervised penalized canonical correlation analysis [20], Intra-View and Inter-View Supervised Correlation Analysis [21], 3CCA using for face recognition [22], spectral supervised CCA for facial expression recognition [23], semi-supervised subspace learning for brain resonance imaging data [24]. Recently, Sun et al. proposed an supervised extension of CCA called discriminant CCA (DCCA) [25] to maximize within-class correlations and minimize the between-class correlations at the same time. In DCCA, not only the correlation between two views of a sample but also all the cross-view correlations between within-class samples are used to increase class separation. Indeed, it is not necessary to consider all the cross-view correlations between within-class samples into DCCA, because there exist redundancy. In this paper, following DCCA we attempt to incorporate discriminant information into CCA and propose a simple feature extraction method called canonical random correlation analysis (RCA). However, different from DCCA, in RCA we use partial random cross-view correlations between within-class examples. Also, it is worth noting that our proposed RCA algorithm is different from recent random projections (RP) for dimensionality reduction [26]. Although both methods use randomness in dimensionality reduction, RP focuses on single-view dimensionality reduction by generating a random projection matrix independently on data, while RCA is for multi-view dimensionality reduction by randomly generating the within-class cross correlation samples from data. It is important in RCA to determine which within-class cross correlation to be used. In this paper, two approaches are developed to produce within-class cross correlations, called RCA-I and RCA-II respectively. Both approaches generate cross correlation randomly except that there are some constraints imposed on RCA-II. Comparative experiments with other multi-view dimensionality reduction methods including standard CCA, locality-preserving CCA (LPCCA) [16], DCCA and partial least square (PLS) [27] on several multi-view databases, validate the effectiveness of RCA.

The rest of the paper is organized as follows. In Section 2, we give the basic theory of CCA and some discussions about CCA. We extend the standard cross correlation to sample cross correlation in Section 3, where the two approaches RCA-I and RCA-II are developed, and the algorithm of RCA is presented. The nonlinear extension of RCA based on kernel methods is introduced in Section 4. Then experimental results and relevant analysis are shown in Section 5. Finally, we conclude this paper in Section 6.

2 Preliminaries

Canonical correlation analysis (CCA) deals with two sets of input variables, each for one data view. CCA could provide useful information about the implicit semantics of data samples by maximizing the correlations between the two variable spaces. In this section, we will give a short review of CCA and discuss the pros and cons of CCA.

2.1 Canonical correlation analysis

Suppose that the training data are described by two views $S = \{(x_i, y_i)\}_{i=1}^{n}$, corresponding to two random vectors with zero means $x \in R^p$ and $y \in R^q$ respectively, where n is the size of data set. CCA attempts to find two sets of directions, one set for each view, such that the two views would be maximally correlated when being projected onto the two sets of directions respectively. The projections are called canonical variables. Assume w_x and w_y denote a pair of directions for

the respective view, the problem of CCA can be formulated as

$$\begin{aligned} &arg\max_{w_x,w_y} \frac{E[(w_x^{\mathrm{T}}x)(w_y^{\mathrm{T}}y)^{\mathrm{T}}]}{\sqrt{(E[(w_x^{\mathrm{T}}x)(w_x^{\mathrm{T}}x)^{\mathrm{T}}])(E[(w_y^{\mathrm{T}}y)(w_y^{\mathrm{T}}y)^{\mathrm{T}}])}} \\ &= arg\max_{w_x,w_y} \frac{w_x^{\mathrm{T}}E[xy^{\mathrm{T}}]w_y}{\sqrt{(w_x^{\mathrm{T}}E[xx^{\mathrm{T}}]w_x)(w_y^{\mathrm{T}}E[yy^{\mathrm{T}}]w_y)}} \\ &= arg\max_{w_x,w_y} \frac{w_x^{\mathrm{T}}C_{xy}w_y}{\sqrt{(w_x^{\mathrm{T}}C_{xx}w_x)(w_y^{\mathrm{T}}C_{yy}w_y)}}, \end{aligned} \tag{1}$$

where $E[\cdot]$ denotes empirical expectation, C_{xy} denotes between-sets covariance matrix and C_{xx}, C_{yy} denote within-sets covariance matrices. Because both w_x and w_y are scale independent, Eq. (1) is equivalent to

$$\begin{aligned} &arg\max_{w_x,w_y} w_x^{\mathrm{T}}C_{xy}w_y \\ &s.t. \begin{cases} w_x^{\mathrm{T}}C_{xx}w_x = 1, \\ w_y^{\mathrm{T}}C_{yy}w_y = 1. \end{cases} \end{aligned} \tag{2}$$

Through applying Lagrangian equation to Eq. (2), the optimization problem of CCA can be converted to generalized eigenvalue decomposition problem, see Eq. (3).

$$\begin{bmatrix} & C_{xy} \\ C_{xy}^{\mathrm{T}} & \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}. \tag{3}$$

w_x and w_y can be solved by Eq. (3). The directions w_x and w_y corresponding to the largest eigenvalue are called as the first pair of canonical basis, then the second pair and so on. Correlated features are extracted through projecting the two sets of samples onto the two sets of directions respectively. The dimension of canonical correlation subspace is equal to the number of the pairs of canonical bases.

2.2 Discussion of CCA

Classical CCA works well to reveal the correlation relationship between two sets of input variables. Correlated features extracted by CCA characterize the implicit semantics of original data. As illustrated in Section 2.1, CCA does not take advantages of label information during dimensionality reduction. Figure 1 shows the classification results on a subset of the Multiple Feature data set with two views **Fac** and **Fou** (see Section 5.2). From Fig. 1(a) we can see the linear relationships between the first pair of canonical variables out of 207 pairs extracted by CCA. Figure 1(b) shows the two-dimensional PCA representation of all of the canonical variables fused according to FFS-II (Eq. (12)). Although CCA can reveal the hidden relationships between different views, it does not retain any discriminative information after reducing dimensionality. Actually, CCA is an unsupervised method and it can be regarded as a two-view extension of PCA [11]. Therefore, classification performance of CCA is limited. Assume that $X = [x_1x_2\cdots x_n]$, $Y = [y_1y_2\cdots y_n]$ are data matrices for two views. CCA optimization can be rewritten as Eq. (4), based on which the sample cross correlation will be derived in Section 3.

$$\begin{aligned} &arg\max_{w_x,w_y} w_x^{\mathrm{T}}XY^{\mathrm{T}}w_y = arg\max_{w_x,w_y} w_x^{\mathrm{T}}(\sum_{i=1}^{n} x_iy_i^{\mathrm{T}})w_y \\ &s.t. \begin{cases} w_x^{\mathrm{T}}XX^{\mathrm{T}}w_x = w_x^{\mathrm{T}}(\sum_{i=1}^{n} x_ix_i^{\mathrm{T}})w_x = 1, \\ w_y^{\mathrm{T}}YY^{\mathrm{T}}w_y = w_y^{\mathrm{T}}(\sum_{j=1}^{n} y_jy_j^{\mathrm{T}})w_y = 1. \end{cases} \end{aligned} \tag{4}$$

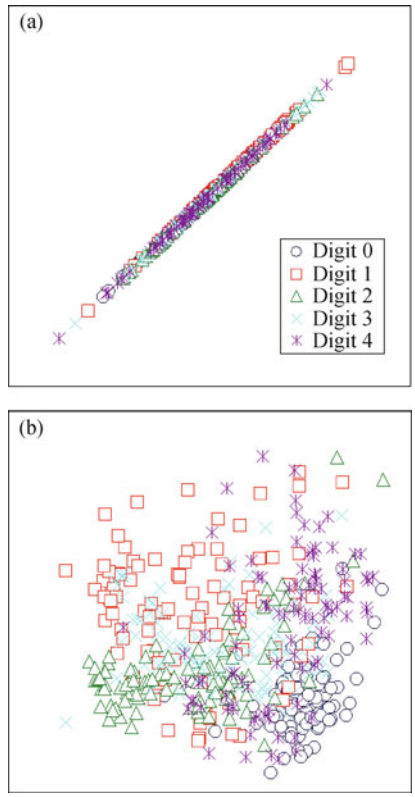

Fig. 1 Demonstration of CCA on a subset of the Multiple Feature data set (digit 0–4). (a) The linear relationships between the first pair of canonical variables out of 207 pairs extracted by CCA; (b) two-dimensional PCA representation of all of the canonical variables fused according to FFS-II (Eq. (12))

Since training samples may come from multiple categories, it is evident that the correlation relationship for one category must be different from another category to some extent, which is ignored by classical CCA (see Eq. (4)). Canonical correlation analysis uses correlations of each sample pair from two views to estimate correlations of two sets of views. Moreover, the differences of samples from different classes are not noticed in CCA. Intuitively, correlations within the

same class should be superior to the correlations between different classes. In order to make up this shortcoming of classic CCA, label information is taken into account in some supervised extensions of CCA methods. In classification tasks, it is intended that discriminative information among various classes should be retained when being projected into canonical correlation subspaces. In Section 3.2, we show that class separation can be increased through introducing cross correlations to CCA. Classical CCA can only identify the linear relationships between the two data views. Nonlinear correlation relationships may be common in many real world applications, where classical CCA performs poorly. In the past, some nonlinear methods such as kernel extension of CCA (KCCA) [12] and local structure preserving CCA (LPCCA) had been developed [16]. However, extra discussions about them are beyond the scope of this paper.

3 Canonical random correlation analysis

In this section, we present our method canonical random correlation analysis (RCA). As noted in the previous section, correlated features extracted by CCA help discover the correlation relationships hidden behind the two views of training data, but important discriminative information for classification might not be preserved. CCA performs dimensionality reduction without considering the label information. RCA explores to introduce cross correlation to classical CCA to retain as much useful discriminative information as possible, see Fig. 2. From Fig. 2(a) we can see the linear relationships between the first pair of canonical variables out of 20 pairs extracted by RCA, and Fig. 2(b) shows the two-dimensional PCA representation of all of the canonical variables fused according to FFS-II (Eq. (12)).

3.1 Sample cross correlations

In other fields such as signal processing, cross correlation is a standard method of measuring similarity of two time series [28]. It is also used to match pattern templates in pattern recognition [29]. Suppose there are two series $x(i)$ and $y(i)$, where $i = 1, 2, \ldots, N$, the standard cross correlation at delay d is defined as

$$r(d) = \frac{\sum_{i=1}^{N}(x(i) - m_x)(y(i-d) - m_y)}{\sqrt{\sum_{i=1}^{N}(x(i) - m_x)^2}\sqrt{\sum_{i=1}^{N}(y(i) - m_y)^2}}, \tag{5}$$

where d is the time delay, m_x and m_y are the means of respective series.

Borrowing the idea from the standard cross correlation, not rigorously, we define sample cross correlation as

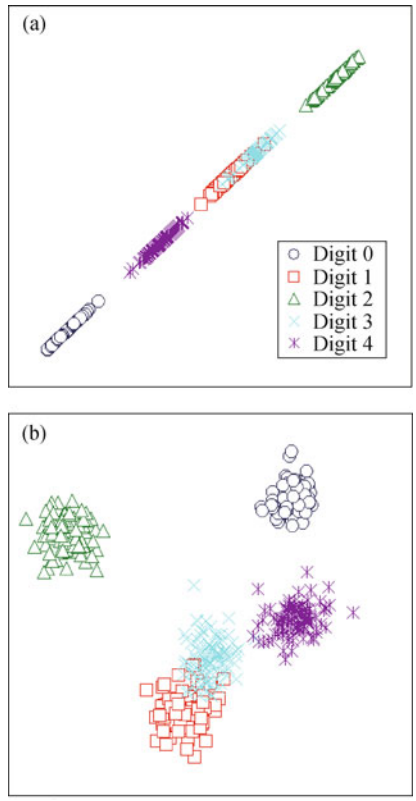

Fig. 2 Demonstration of RCA on the same data as in Fig. 1. (a) The linear relationships between the first pair of canonical variables out of 20 pairs extracted by RCA; (b) two-dimensional PCA representation of all of the canonical variables fused according to FFS-II (Eq. (12))

$$R = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j^{\mathrm{T}}}{\sqrt{\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}}\sqrt{\sum_{j=1}^{n} y_j y_j^{\mathrm{T}}}}, \tag{6}$$

where $(x_i, y_i) \in S$ is centered observations, and each sum term $x_i y_j^{\mathrm{T}}$ is referred to as a cross correlation term, or correlation term for short.

From Eqs. (4) and (6), it can be found that CCA is a special case of application of sample cross correlation. Now we can extend the optimization of classical CCA based on the notion of sample cross correlation to the new canonical cross-correlation analysis as

$$\begin{aligned} &arg\max_{w_x, w_y} w_x^{\mathrm{T}}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j^{\mathrm{T}}\Big) w_y \\ &s.t. \begin{cases} w_x^{\mathrm{T}}\big(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\big) w_x = 1, \\ w_y^{\mathrm{T}}\big(\sum_{j=1}^{n} y_j y_j^{\mathrm{T}}\big) w_y = 1. \end{cases} \end{aligned} \tag{7}$$

The optimization problem can be solved similarly as CCA. Considering that the samples may come from multiple classes, if x_i and y_j are with the same class label, then $x_i y_j^{\mathrm{T}}$ is called within-class correlation item, otherwise, it is called between-class correlation term. To preserve effective

discriminant information, only within-class correlation terms are considered in our method.

3.2 Random sampling

In this section, we present our method RCA. In RCA, a particular set of weighted cross correlation terms are chosen within every class. However, it is challenging to determine the proper correlation term set. Different from DCCA where all the cross-view correlations between within-class examples are used, in RCA we use partial random cross-view correlations between within-class examples because there exists redundancy in the cross-view correlations of within-class examples and not all those cross-view correlations are required. In RCA, we adopt a rather simple fashion to construct the correlation term set, i.e., sampling examples with replacements randomly within every class. Here, two approaches are proposed to implement RCA. We show that through extra constraints, the number of correlation terms needed can be decreased further.

Suppose all samples in S fall into c classes $\{w_k\}_{k=1}^{c}$, and $\mathcal{X}_k$, $\mathcal{Y}_k$ are the jth subset of respective view of training data S. And we assume that $\tilde{\mathcal{X}}_k$ and $\tilde{\mathcal{Y}}_k$ are the jth subset in final correlation term set. Now two approaches, called RCA-I and RCA-II respectively, are developed to determine the correlation term sets $\tilde{\mathcal{X}}_k$ and $\tilde{\mathcal{Y}}_k$.

In RCA-I, we sample $\mathcal{X}_k$ and $\mathcal{Y}_k$ with replacement to form corresponding $\tilde{\mathcal{X}}_k$ and $\tilde{\mathcal{Y}}_k$, and let $\tilde{\mathcal{X}}_k$ and $\tilde{\mathcal{Y}}_k$ have the same size as $\mathcal{X}_k$ and $\mathcal{Y}_k$. In fact, $\tilde{\mathcal{X}}_k$ and $\tilde{\mathcal{Y}}_k$ can be seen as special bootstrap samples of $\mathcal{X}_k$ and $\mathcal{Y}_k$, because the action is taken over every class instead of the entire training set. In practice, the process will be repeated multiple times, say t times. As a result, t sets of bootstrap samples are generated, i.e., $\tilde{\mathcal{X}}_k^{(l)}$, $\tilde{\mathcal{Y}}_k^{(l)}$, $l = 1, 2, \ldots, t$, where the superscript l in parentheses represents the lth bootstrap sample and

$$\begin{cases} \tilde{\mathcal{X}}_k^{(l)} = \{\tilde{x}_i^{(l)}\}_{i=1}^{n_k}, \tilde{x}_i^{(l)} \in \mathcal{X}_k, \\ \tilde{\mathcal{Y}}_k^{(l)} = \{\tilde{y}_j^{(l)}\}_{j=1}^{n_k}, \tilde{y}_j^{(l)} \in \mathcal{Y}_k, \end{cases}$$

where $k = 1, 2, \ldots, c$, and n_k denotes the size of $\mathcal{X}_k$ and $\mathcal{Y}_k$. Therefore, the size of final correlation term set would be $\sum_{k=1}^{c} tn_k = tn$. Some elements may occur several times and the frequencies of correlation terms are defined as their weights. For those correlation terms that have not been in the final correlation term set, their weights are set to zero. RCA-I can be formulated as

$$arg \max_{w_x, w_y} w_x^{\mathrm{T}} \Big(\sum_{k=1}^{c} \sum_{l=1}^{t} \sum_{i=1}^{n_k} \tilde{x}_i^{(l)} \tilde{y}_i^{(l)\mathrm{T}}\Big) w_y. \tag{8}$$

In theory, Discriminant CCA can be viewed as a special version of RCA. In DCCA, all the within-class correlation terms are used and the weights are set to 1, while only partial random with-class correlation terms are used in RCA.

RCA-II looks a bit like RCA-I except for some extra constraints. In RCA-II, only the second view $\mathcal{Y}$ is considered to be sampled and the first view is kept unchanged. RCA-II can be formulated as

$$arg \max_{w_x, w_y} w_x^{\mathrm{T}} \Big(\sum_{k=1}^{c} \sum_{l=1}^{t} \sum_{i=1}^{n_k} x_i \tilde{y}_i^{(l)\mathrm{T}}\Big) w_y. \tag{9}$$

In RCA-I, samples in $\mathcal{X}_k$ and $\mathcal{Y}_k$ may occur in $\tilde{\mathcal{X}}_k$ and $\tilde{\mathcal{Y}}_k$ multiple times or not. It is clear that prior information about the given sample data could not be fully utilized, so more cross correlation terms are needed in RCA-I to achieve class separation. RCA-II is designed to remain unchanged in the first view and to make sampling only on the second view. Thus prior information about the data can be made better use of, and less cross correlation terms are required to achieve good class separation. The subsequent experiments illustrate the point.

3.3 The RCA algorithm

Now, we present the RCA algorithm (both I and II). Let $\mathcal{X} = \bigcup_{k=1}^{c} \mathcal{X}_k$, $\mathcal{Y} = \bigcup_{k=1}^{c} \mathcal{Y}_k$, and let X, Y still denote data matrices of $\mathcal{X}$ and $\mathcal{Y}$. A $n \times n$ weight matrix $R_w \in R^{n \times n}$ is constructed to store weight values of cross correlations, where n is the size of $\mathcal{X}$ (or $\mathcal{Y}$). The (i, j) entry of R_w corresponds to the weight of the cross correlation term $x_i y_j^{\mathrm{T}}$. So R_w can be represented as a block diagonal matrix, and the kth block R_{w_k} corresponds to the kth class subset. Two examples of R_{w_k} for RCA-I and RCA-II may look as follows:

$$R_{w_k}^{(\mathrm{I})} = \begin{bmatrix} 2 & 0 & 0 & 0 & 1 \\ 2 & 0 & 0 & 1 & 2 \\ 0 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix},$$

$$R_{w_k}^{(\mathrm{II})} = \begin{bmatrix} 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix},$$

where we assume that there are five samples in the kth class subset and t is set to three. The superscripts I and II represent the two approaches respectively. Sums of all elements in $R_{w_k}^{(\mathrm{I})}$

and $R_{w_k}^{(\mathrm{II})}$ are equal to fifteen. According to RCA-II, the row sums of $R_{w_k}^{(\mathrm{II})}$ is equal to the value of t, i.e., three. Figure 3 shows the differences between CCA and RCA from perspective of correlation terms. The points in 3-dimensional space and the points in the bottom 2-d plane represent two views respectively. Different marks (circle and triangle) indicate different classes (two classes). The dashed lines represent correlation terms used in respective method and its widths denote weights. CCA (Fig. 3(a)) only considers pair-wise correlation terms. RCA randomly chooses within-class correlation terms, so there may be points who are not contained by any correlation term (the arrow pointing).

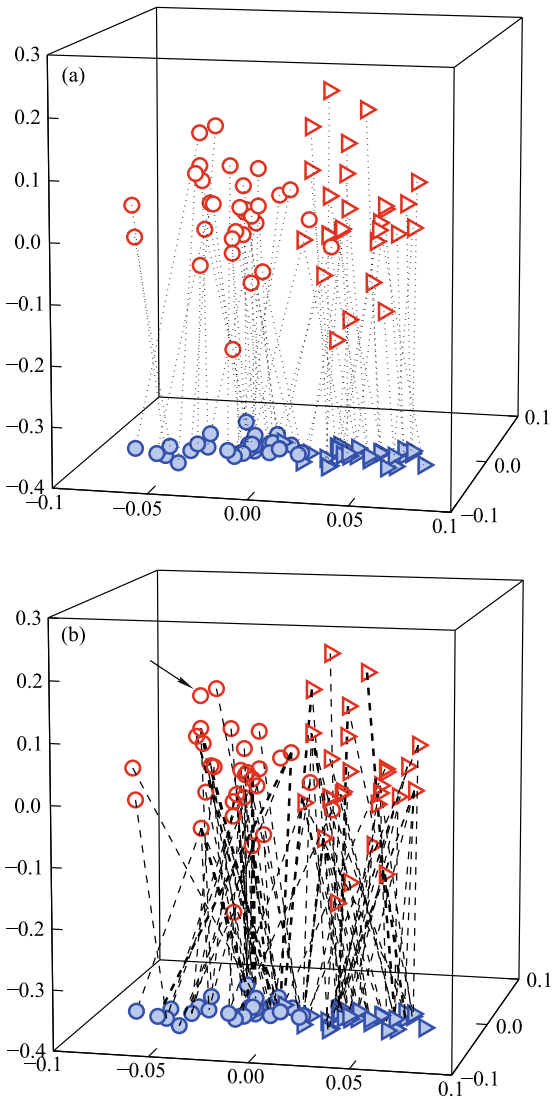

Fig. 3 Differences between (a) CCA and (b) RCA from perspective of correlation terms

Now, RCA (both I and II) can be reformulated as

$$arg \max_{w_x, w_y} w_x^{\mathrm{T}} X R Y^{\mathrm{T}} w_y, \tag{10}$$

where $R = R_w + R_w^{\mathrm{T}}$ is set to guarantee symmetry of the correlation relationships, which implies that all symmetric terms, e.g., $x_j y_i^{\mathrm{T}}$ with respect to $x_i y_j^{\mathrm{T}}$, are taken into account automatically to reinforce correlated relation further. The above optimization problem can be solved by the following generalized eigenvalue decomposition according to Eq. (3):

$$\begin{bmatrix} & XRY^{\mathrm{T}} \\ YRX^{\mathrm{T}} & \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} XX^{\mathrm{T}} & \\ & YY^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}. \tag{11}$$

Similar to CCA, a sequence of canonical variable pairs can be solved, and the dimension of canonical correlation subspace equals to the number of pair of canonical bases. Let $W_x = [w_{x_1} w_{x_2} \cdots w_{x_d}]$ and $W_y = [w_{y_1} w_{y_2} \cdots w_{y_d}]$, where the subscripts indicate the sequence of canonical variable pairs and d is the number of canonical variable pairs. W_x and W_y are called canonical projection matrixs. In Ref. [30], the authors proposed to fuse features extracted by CCA via two feature fusion strategies (FFS). FFS-I adds the two groups of canonical variables, and FFS-II combines them into a large canonical vector. For any example in training set $(x_i, y_i) \in S$, we have

$$\begin{aligned} &\text{FFS-I}: W_x^{\mathrm{T}} x_i + W_y^{\mathrm{T}} y_i, \\ &\text{FFS-II}: \begin{bmatrix} W_x^{\mathrm{T}} x_i \\ W_y^{\mathrm{T}} y_i \end{bmatrix}. \end{aligned} \tag{12}$$

The algorithm of RCA is summarized in Algorithm 1. The whole algorithm can be divided into three phases. In the first phase, a n by n matrix R_{corr} is initialized to zero. In the second phase, the correlation term set is constructed and is represented by the matrix R_{corr}. In the third phase, we obtain the two canonical projection matrices W_x and W_y with d columns by solving the generalized eigenvalue decomposition

Algorithm 1 RCA

Inputs: Training data $\mathcal{X} = \bigcup_{k=1}^{c} \mathcal{X}_k$, $\mathcal{Y} = \bigcup_{k=1}^{c} \mathcal{Y}_k$,
The size of correlation term set t,
The dimension of canonical subspace d,
Outputs: Two canonical projection matrices,
$W_x = [w_{x_1} w_{x_2} \cdots w_{x_d}]$ and $W_y = [w_{y_1} w_{y_2} \cdots w_{y_d}]$.
Initialize: $R_w = (0)_{n \times n}$;
1. **For** $l = 1$ **To** t **Do**
2. Let $\tilde{\mathcal{X}}^{(l)} = \emptyset$, $\tilde{\mathcal{Y}}^{(l)} = \emptyset$;
3. **for** k = 1 **To** c **Do**
4. Construct bootstrap samples $\tilde{\mathcal{X}}_k^{(l)}$, $\tilde{\mathcal{Y}}_k^{(l)}$ from $\mathcal{X}_k$ and $\mathcal{Y}_k$ through the two approaches, RCA-I or RCA-II;
5. Set $\tilde{\mathcal{X}}^{(l)} = \tilde{\mathcal{X}}^{(l)} \bigcup \tilde{\mathcal{X}}_k^{(l)}$, $\tilde{\mathcal{Y}}^{(l)} = \tilde{\mathcal{Y}}^{(l)} \bigcup \tilde{\mathcal{Y}}_k^{(l)}$;
6. **Loop**
7. Fill R_{corr} according to $\tilde{\mathcal{X}}^{(l)}$ and $\tilde{\mathcal{Y}}^{(l)}$;
8. **Loop**
9. Set $R = R_w + R_w^{\mathrm{T}}$;
10. Obtain d pairs of canonical variable by solving Eq. (11);

problem in Eq. (11). The process that constructs the special bootstrap sample will be repeated t times during the second phase. Bootstrap samples were formed over every class every time, which is different from standard bootstrap sample.

4 Kernelization of RCA

RCA is a linear learning model and it can not cope with complex non-linear problems. When dealing with those non-linear problems, a common technique is using kernel functions. In this section, we introduce the kernel gerneralization of RCA, KRCA. Kernel methods firstly map data into high dimensional feature spaces, then linear models are adapted. In the past years, various kernel based methods have been developed, such as KCCA [31], KPCA [32], KICA [33].

Suppose samples $[x_1, x_2, \ldots, x_n]$, $[y_1, y_2, \ldots, y_n]$ are mapped into feature spaces by two non-linear mapping functions ϕ_x, ϕ_y:

$$\begin{aligned}\tilde{X} &= [\phi_x(x_1), \phi_x(x_2), \ldots, \phi_x(x_n)]\\ \tilde{Y} &= [\phi_y(y_1), \phi_y(y_2), \ldots, \phi_y(y_n)].\end{aligned} \tag{13}$$

The canonical basis of RCA in the feature spaces can be expressed as $\tilde{w}_x = \tilde{X}\alpha$, $\tilde{w}_y = \tilde{Y}\beta$. RCA in the feature spaces can be rewritten as

$$\begin{aligned}&arg\max_{\alpha,\beta} \alpha^{\mathrm{T}}\tilde{X}^{\mathrm{T}}\tilde{X}R\tilde{Y}^{\mathrm{T}}\tilde{Y}\beta\\ &s.t.\begin{cases}\alpha^{\mathrm{T}}\tilde{X}^{\mathrm{T}}\tilde{X}\tilde{X}^{\mathrm{T}}\tilde{X}\alpha = 1,\\ \beta^{\mathrm{T}}\tilde{Y}^{\mathrm{T}}\tilde{Y}\tilde{Y}^{\mathrm{T}}\tilde{Y}\beta = 1.\end{cases}\end{aligned} \tag{14}$$

We have

$$\begin{cases}\tilde{X}^{\mathrm{T}}\tilde{X} = [\phi_x(x_i)^{\mathrm{T}}\phi_x(x_j)]_{i,j} \in R^{n\times n},\\ \tilde{Y}^{\mathrm{T}}\tilde{Y} = [\phi_y(y_i)^{\mathrm{T}}\phi_y(y_j)]_{i,j} \in R^{n\times n}.\end{cases} \tag{15}$$

Defining kernel function in the feature spaces: $k_x(x_i, x_j), k_y(y_i, y_j)$, and replacing the inner products in the above equation with kernel functions, we get

$$\begin{aligned}&arg\max_{\alpha,\beta} \alpha^{\mathrm{T}}K_xRK_y\beta\\ &s.t.\begin{cases}\alpha^{\mathrm{T}}K_x^2\alpha = 1,\\ \beta^{\mathrm{T}}K_y^2\beta = 1,\end{cases}\end{aligned} \tag{16}$$

where $K_x, K_y \in R^{n\times n}$ are kernel matrices and $K_{x_{ij}} = k_x(x_i, x_j)$, $K_{y_{ij}} = k_y(y_i, y_j)$.

The solution of Eq. (14) is similar to those of RCA. We omit it here due to space limit. After obtaining linear combination coefficients $(\alpha_i, \beta_i) \in R^n \times R^n$, we can solve canonical basis as follows:

$$\begin{aligned}W_\phi &= [\omega_{\phi_1}\omega_{\phi_2}\ldots\omega_{\phi_d}] = \tilde{X}[\alpha_1\alpha_2\ldots\alpha_d],\\ W_\varphi &= [\omega_{\varphi_1}\omega_{\varphi_2}\ldots\omega_{\varphi_d}] = \tilde{Y}[\beta_1\beta_2\ldots\beta_d],\end{aligned} \tag{17}$$

where d is the number of canonical basis.

5 Experiments

In this section, we evaluate the classification performances of the methods RCA-I and RCA-II on Multiple Features data set and Internet Advisements data set picked out from UCI repository and three faces databases. The first data set used is Multiple Features data set, which consists of 2 000 examples of ten handwritten digits ('0-9'), 200 examples for each digit. Each example is represented by six features sets. The six sets of features and number of attributes are:

1) profile correlations (216 attributes and called **Fac** for short);
2) Fourier coefficients of the character shapes (76, **Fou**);
3) Karhunen-Love coefficients (64, **Kar**);
4) morphological features (6, **Mor**);
5) pixel averages in 2 by 3 windows (240, **Pix**);
6) Zernike moments (47, **Zer**).

Any two of them can be used as two data views, so there will be 15 possible combinations in total.

The second data set used is the Internet Advisements data set ,which includes 3 279 web images (459 Ads and 2 820 Non-ads) with 1 558 attributes. Except four attributes with missing value, the remaining 1 554 attributes can be divided into five groups, which are used for describing image urls and text descriptions. These attributes are as follows:

1) 472 attributes from ancurl terms, abbreviated as **Anc**;
2) 111 attributes from alt terms, abbreviated as **Alt**;
3) 19 attributes from caption terms, abbreviated as **Cap**;
4) 495 attributes from origurl terms, abbreviated as **Org**;
5) 457 attributes from url terms, abbreviated as **Url**.

We also use three face databases to perform face recognition experiments: ORL, YALE, and CMU PIE database. ORL, also called AT&T database, consists of 400 images of 40 subjects, 10 images for each subject. These images are photographed in different times, with changing lightning, facial expressions. The size of each original image is 92 × 112 pixels, with 256 gray levels per pixel. YALE database

contains 165 gray scale images of 15 individuals. There are 11 images for each subject with different illumination conditions and expressions: center/left/right-light, wearing glasses or not, happy, normal, right-light, sad, sleepy, surprised, and wink. The CMU PIE database contains a huge collection of face images, under varying poses, illuminations and expressions. There are 68 subjects in PIE database, each with 13 different poses, 43 different illumination conditions, and four different expressions. A subset with frontal pose (C27) was used in experiments.

5.1 Experiment settings

We compare our methods with several related algorithms, e.g., classical CCA, partial least square (PLS) [27], locality preserving CCA (LPCCA) [16], discriminant CCA (DCCA). In face recognition experiments, two well-known techniques Eigenfaces [34] and Fisherfaces [35] are also taken for comparing. Both FFS-I and FFS-II are adapted to fuse lower dimensional related features.

In subspace learning, one important thing is to determine the dimension of subspace. Generally, performances might vary with different dimensions. In the experiments on Internet Advisements data set, the subspace dimensionality is set automatically by sorting the canonical variables coefficients in descending order, so that 95 percent correlation information is preserved after dimensionality reduction. In other experiments, we always set the subspace dimension to 30. For the views with feature numbers less than 30, such as **Mor** with six features in Multiple Features Data Set, we set the subspace dimension to the feature number of the view. Besides, we also study the differences and connections between RCA-I and RCA-II, and the effects of the size of correlation term set in RCA-I and RCA-II on them.

For kernel-based methods, the Gaussian is chosen as kernel function, i.e., $K(x_1, x_2) = e^{-(x_1-x_2)^2/2\sigma^2}$. The kernel widths σ^2 is determined through cross validation. The parameter space is defined as $\{2^{-4}, 2^{-3}, \ldots, 2^0, \ldots, 2^3, 2^4\} \times \sigma_0^2$, where σ_0^2 denotes the mean square distances of sample,

$$\sigma_0^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \| x_i - x_j \|^2 . \tag{18}$$

The parameter t is closely related with the performance of RCA. The parameter was chosen empirically according to Table 1.

Table 1 The settings of parameter t

Data set	RCA-I	RCA-II
Multiple features	100	20
Internet Ads	100	40
YALE	20	10
ORL	30	15
PIE(10)	20	10
PIE(15)	30	15
PIE(20)	40	20

Note: the numbers in the parentheses are the sizes of the training sets

5.2 Multiple features data set

In the experiment, half of the data set will be selected randomly for learning canonical subspaces and the rest for testing. Thus there are 1 000 training examples and 1 000 testing examples in total, 100 for each class. The task is set to predict the classes of the testing examples. Average recognition rates over ten independent trials were recorded, shown in Tables 2 and 3. The two tables correspond to FFS-I and FFS-II respectively. In the tables, the first column represents 15 possible

Table 2 Recognition rates on multiple features data set (FFS-I)

Data	CCA	PLS	LPCCA	DCCA	RCA-I	RCA-II	KCCA	KRCA-I	KRCA-II
Fac and **Fou**	0.872 0	0.873 3	0.904 7	**0.937 3**	**0.937 3**	0.927 3	0.913 1	0.761 4	0.765 1
Fac and **Kar**	0.962 0	0.945 3	0.958 3	0.966 8	0.964 7	**0.967 0**	0.958 8	0.939 2	0.941 5
Fac and **Mor**	0.766 7	0.860 0	0.788 0	0.884 8	0.877 7	0.870 7	0.876 3	**0.923 0**	0.920 0
Fac and **Pix**	0.940 7	0.949 3	0.949 3	0.961 6	0.961 3	**0.963 0**	0.953 3	0.934 7	0.937 3
Fac and **Zer**	0.850 7	0.910 0	0.878 0	0.948 4	**0.948 7**	0.942 3	0.945 3	0.943 7	0.941 7
Fou and **Kar**	0.895 3	**0.960 0**	0.911 7	0.929 6	0.921 3	0.914 0	0.928 0	0.862 0	0.867 0
Fou and **Mor**	0.755 3	0.742 0	0.731 7	0.810 7	**0.815 0**	0.812 0	0.797 0	0.768 3	0.764 0
Fou and **Pix**	0.821 0	**0.967 3**	0.801 0	0.905 9	0.888 0	0.890 0	0.927 3	0.871 7	0.857 7
Fou and **Zer**	0.823 0	0.791 3	0.824 7	**0.829 6**	0.818 0	0.825 7	0.824 0	0.726 0	0.729 0
Kar and **Mor**	0.778 3	0.869 0	0.815 7	0.873 2	0.866 0	0.856 3	0.910 0	**0.951 3**	0.951 3
Kar and **Pix**	0.963 3	0.966 0	**0.970 0**	0.926	0.921 7	0.926 7	0.917 0	0.936 0	0.943 3
Kar and **Zer**	0.882 7	0.858 0	0.923 3	0.934 4	0.933 3	0.931 3	0.912 7	**0.937 7**	0.936 7
Mor and **Pix**	0.723 3	0.829 3	0.738 3	0.856	0.849 7	0.833 3	0.904 7	**0.946 0**	0.942 0
Mor and **Zer**	0.722 3	0.740 7	0.707 7	0.788 3	0.788 3	0.784 3	0.726 7	0.793 7	**0.797 7**
Pix and **Zer**	0.818 3	0.888 0	0.873 3	0.911	0.906 0	0.899 3	0.914 3	0.937 3	**0.934 0**

Table 3 Recognition rates on multiple features data set (FFS-II)

Data	CCA	PLS	LPCCA	DCCA	RCA-I	RCA-II	KCCA	KRCA-I	KRCA-II
Fac and **Fou**	0.897 8	0.921 8	0.932 2	0.969 3	**0.969 4**	0.968 2	0.925 6	0.844 6	0.856 8
Fac and **Kar**	0.962 8	0.970 6	0.960 8	0.972 6	0.971 8	**0.974 4**	0.954 2	0.948 2	0.948 2
Fac and **Mor**	0.773 0	0.885 2	0.825 6	**0.935 6**	0.928 0	0.910 4	0.899 2	0.930 0	0.930 6
Fac and **Pix**	0.945 4	0.967 8	0.949 2	0.962	0.967 2	**0.971 2**	0.964 6	0.946 2	0.945 2
Fac and **Zer**	0.868 6	0.966 0	0.925 4	0.970 0	**0.971 2**	0.969 6	0.959 8	0.938 2	0.938 6
Fou and **Kar**	0.930 0	0.958 8	0.960 2	**0.961 0**	0.958 4	0.958 8	0.932 0	0.868 2	0.861 2
Fou and **Mor**	0.773 0	0.637 2	0.777 2	0.822 0	0.822 2	**0.822 8**	0.816 2	0.785 0	0.783 2
Fou and **Pix**	0.846 6	**0.966 4**	0.850 8	0.946 0	0.936 0	0.934 4	0.930 6	0.799 6	0.805 2
Fou and **Zer**	**0.846 4**	0.810 0	0.845 6	0.840 9	0.842 4	0.846 2	0.844 2	0.784 2	0.779 0
Kar and **Mor**	0.815 8	0.919 6	0.858 8	0.936	0.939 4	0.914 8	0.938 2	**0.956 8**	0.956 6
Kar and **Pix**	0.965 0	0.966 2	**0.969 4**	0.937 7	0.943 2	0.938 2	0.968 6	0.944 0	0.943 8
Kar and **Zer**	0.919 8	0.918 2	**0.962 6**	0.958 8	0.957 4	0.955 2	0.917 2	0.940 8	0.944 2
Mor and **Pix**	0.755 4	0.911 0	0.754 4	0.920 5	0.907 4	0.880 0	0.938 4	**0.948 2**	0.946 0
Mor and **Zer**	0.748 6	0.790 0	0.735 8	0.819 9	**0.820 6**	0.819 4	0.757 6	0.810 8	0.809 2
Pix and **Zer**	0.845 6	0.920 6	0.919 8	0.936 6	0.937 8	0.931 4	0.919 2	**0.941 0**	0.936 0

view combinations in order, i.e., one for view **Fac** and view **Fou**, fifteen for view **Pix** and view **Zer** and so on.

The recognition rates of CCA in Tables 2 and 3 can be seen as base lines. Our methods RCA and their kernelizations have better recognition performances in most cases, which implies that better class separation can be achieved by including within-class cross correlation terms into CCA process. Although the results of DCCA are similar to RCA, fewer correlation terms are needed in RCA. Besides, KCCA shows better performances than CCA, while it is time-consuming for kernelized methods to choose appropriate parameters. Meanwhile, different couples of views infect the classification performance, for example, most methods perform poor on the couples with **Mor**.

Figure 4 shows the influences of different sizes of correlation term sets on the methods RCA-I and RCA-II on multiple features data set. The top (Figs. 4(a) and 4(b)) and bottom (Figs. 4(c) and 4(d)) rows correspond to the two feature fusion strategies respectively. The horizontal axis represents the size of correlation term sets, denoted by t (Eqs. (8) and (9)), and the vertical axis represents classification accuracy. The range of t is 1 to 120. From Fig. 4, we can see that with the

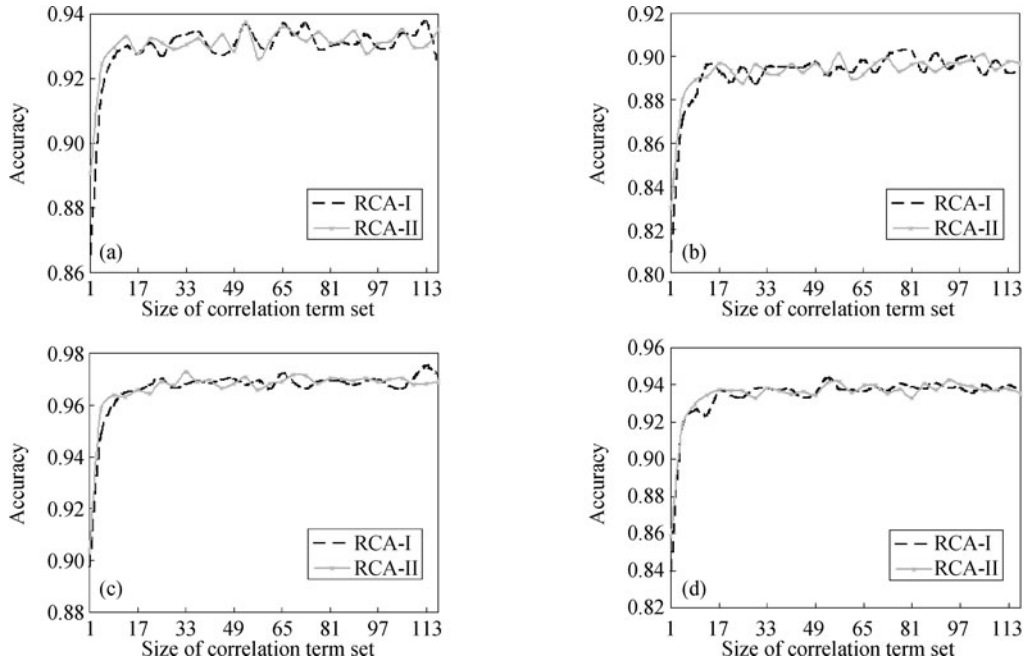

Fig. 4 Influences of correlation term set size on classification performances of the two methods RCA-I and RCA-II. (a) **Fac** and **Fou** (FFS-I); (b) **Fou** and **Pix** (FFS-I); (c) **Fac** and **Fou** (FFS-II); (d) **Fou** and **Pix** (FFS-II)

increasing size of the correlation term sets, RCA (both I and II) accomplishs better recognition performances. It is particular significant when t is smaller (1–10). It implies that more discriminant information can be retained when using more correlation terms. However, when t continues to increase, RCA tends to be more smooth. It is also clear that RCA-II achieves better recognition results than RCA-I when t is smaller. The reasons can be that RCA-II makes better use of prior information about the data sets than RCA-I by imposing extra constraints on the correlation term sets.

5.3 Internet Advisements data set

In this experiment, there are five experimental settings on Internet Advisements. In each experimental setting, one of the five attribute groups was picked out in turn as the first view, and the remaining attribute groups as the second view. Meanwhile, 230 positive samples and 230 negative samples were randomly chosen as training set, and the remaining 2 189 samples in the data set as testing set.

Due to the imbalance of the two classes and the huge number of attributes, we preprocessed the data using Principle Component Analysis to preserve 95 percent variance information in order to avoid small sample problem. From Tables 4 and 5, we can see that kernelized methods performed better than non-kernelized methods in Internet Advisements Data Set. Meanwhile, Kernelized RCA methods do better than KCCA. Except Kernelized methods, we can see that PLS performs a little better than others, and RCA methods perform a little better than DCCA. In summary, better performance can be achieved by including within-class cross correlation terms into CCA process.

5.4 Face recognition experiments

Some preprocessing steps had been done for images in these databases [36, 37]. Face area in each image was cropped and the final size was set to 64×64 pixels, which was considered as the first view. Images with different resolutions can provide information at different levels and can be regarded as another view. We resize each image to 32×32 pixels to form the second view. Then double Daubechies wavelet transformation is performed on original images, and the low-frequency images are used as the third view. We obtain the fourth view of local binary pattern (LBP) histograms of each image, which has been shown to be efficient patterns to represent face images [38]. In LBP setting, each image is divided into 4×4 local regions firstly, 16×16 pixels for each region. Then LBP histograms were calculated over all sixteen regions for every image [39].

Differing from previous experimental setting, we do not exhaust all of six possible view combinations and only three combinations of six are chosen. The original images (64×64) are always taken as the first view and the other image views will be taken in turn as the second view, because the original images are easier to obtain than others and the other views can be calculated from the original images.

For the YALE and ORL databases, the data sets were partitioned into equal size training sets and testing sets randomly. So there are seven training images for the YALE data set and five images for the ORL data set. For the larger CMU PIE database, three preconcerted partitions are provided. Ten, fifteen and twenty images, respectively, are picked out randomly to form training sets, and the remaining images are for testing. Thus there are nine settings in total for the PIE database. Experiments on the three database are repeated ten times independently, and average recognition rates are for comparing. The parameter t is set according to Table 1.

Two well-known face recognition techniques, Eigenfaces [34] and Fisherfaces [35] were also adopted for comparing.

Table 4 Recognition rates on Internet Advisements data set (FFS-I)

Data	CCA	PLS	LPCCA	DCCA	RCA-I	RCA-II	KCCA	KRCA-I	KRCA-II
Anc	0.733 2	0.757 1	0.731 9	0.735 9	0.734 5	0.735 3	0.865 8	0.893 3	**0.895 4**
Alt	0.763 0	0.765 3	0.764 0	0.752 9	0.761 2	0.761 7	0.900 1	0.907 4	**0.911 9**
Cap	0.728 4	0.768 7	0.734 3	0.744 4	0.768 3	0.762 4	0.857 9	**0.904 6**	0.904 2
Org	0.750 3	0.766 8	0.750 4	0.738 7	0.751 9	0.748 6	0.887 8	**0.909 8**	0.905 6
Url	0.759 6	0.773 9	0.757 4	0.763 7	0.768 1	0.766 9	0.904 8	0.913 4	**0.917 1**

Table 5 Recognition rates on Internet Advisements data set (FFS-II)

Data	CCA	PLS	LPCCA	DCCA	RCA-I	RCA-II	KCCA	KRCA-I	KRCA-II
Anc	0.742 7	0.742 2	0.737 3	0.749 8	0.744 4	0.745 3	0.878 5	0.902 8	**0.903 4**
Alt	0.754 9	0.755 7	0.751 0	0.766 9	0.754 3	0.751 0	0.895 5	**0.911 2**	0.910 1
Cap	0.729 5	0.769 8	0.740 9	0.746 8	0.770 2	0.769 7	0.879 7	**0.908 8**	0.904 4
Org	0.754 2	0.756 4	0.753 5	0.747 8	0.749 5	0.746 3	0.893 9	0.906 1	**0.906 3**
Url	0.757 1	0.763 5	0.752 8	0.763 3	0.756 7	0.757 8	0.899 2	0.910 5	**0.910 6**

For every view combination, we firstly extract 150 principal components using PCA from each view, then two low dimensional views are fused through FFS-I or FFS-II for face recognition with Eigenfaces and Fisherfaces.

Tables 6 and 7 are the recognition rates of different methods on the three face databases. The first column indicates the database. The sequential numbers after the names of databases represent different view combination settings, i.e., one for original images (64×64) and scaled images (32×32), two for original images and wavelet transformations of images, three for original images and LBP histograms of images. For PIE data set, the numbers in parenthesis denote the sizes of training sets. PCA and LDA represent Eigenfaces and Fisherfaces, respectively.

As shown in Tables 6 and 7, RCA achieves the best recognition performance on all three databases under two feature fusion strategies. The two implementations of RCA cannot outperform each other on all databases, e.g., on YALE, RCA-I gets better accuracies than RCA-II when adapting FFS-I and RCA-II get better accuracies than RCA-I when adapting FFS-II. DCCA also performs well on the three face databases, but more correlation terms are needed in the experiments. Kernelized methods still show unstable results due to difficulty of choosing kernel parameters appropriately. Compared with PCA and LDA, RCA can make full use of useful information behind multiple views.

Also, Fig. 5 shows the effect of the size of the correlation terms set on the two methods RCA-I and RCA-II on the three face databases. The third view combination setting of the three face databases is chosen in Fig. 5, i.e., original images and LBP histograms. From Fig. 5, RCA-I and RCA-II get better classification results with increasing size of

Table 6 Recognition rates on YALE, ORL and PIE data sets (FFS-I)

Data	CCA	PLS	LPCCA	DCCA	RCA-I	RCA-II	KCCA	KRCA-I	KRCA-II	PCA	LDA
YALE1	0.423 3	0.448 9	0.177 8	0.872 2	**0.880 0**	0.847 8	0.263 3	0.746 7	0.738 9	0.407 8	0.440 0
YALE2	0.537 8	0.531 1	0.342 2	0.802 2	**0.852 2**	0.818 9	0.344 4	0.748 9	0.732 2	0.636 7	0.714 4
YALE3	0.588 9	0.585 6	0.331 1	**0.892 2**	0.877 8	0.882 2	0.373 3	0.808 9	0.767 8	0.518 9	0.668 9
ORL1	0.875 0	0.875 0	0.860 5	0.926	0.926 5	**0.930 5**	0.619 0	0.836 0	0.825 0	0.844 0	0.908 0
ORL2	0.887 0	0.887 0	0.813 5	0.93	0.927 0	**0.932 5**	0.610 5	0.831 0	0.842 0	0.842 5	0.917 0
ORL3	0.892 0	0.892 0	0.878 0	0.962	**0.965 5**	0.965 0	0.847 5	0.901 0	0.906 5	0.774 0	0.938 0
PIE1(10)	0.850 7	0.850 7	0.856 5	0.922 4	0.924 4	**0.925 6**	0.813 4	0.855 8	0.855 0	0.846 1	0.914 3
PIE1(15)	0.911 4	0.911 4	0.912 6	0.951 4	**0.953 7**	0.953 4	0.882 3	0.904 1	0.903 7	0.900 4	0.943 4
PIE1(20)	0.939 1	0.939 1	0.943 3	0.964 6	0.964 7	**0.965 0**	0.919 8	0.933 4	0.930 3	0.933 0	0.957 4
PIE2(10)	0.849 8	0.849 8	0.872 7	0.923 9	**0.928 2**	0.928 2	0.808 8	0.867 6	0.867 3	0.845 7	0.923 6
PIE2(15)	0.907 7	0.907 7	0.923 4	0.949 9	**0.954 3**	0.954 0	0.886 6	0.913 1	0.913 0	0.905 9	0.950 9
PIE2(20)	0.936 7	0.936 7	0.946 7	0.961 1	0.965 0	**0.965 5**	0.911 7	0.936 7	0.936 5	0.934 7	0.962 4
PIE3(10)	0.871 7	0.871 7	0.869 3	0.954 3	**0.959 1**	0.956 8	0.883 4	0.861 8	0.860 1	0.851 7	0.927 6
PIE3(15)	0.923 9	0.923 9	0.922 4	0.973 8	**0.976 4**	0.975 9	0.937 0	0.907 8	0.911 0	0.907 6	0.955 6
PIE3(20)	0.952 7	0.952 7	0.948 7	0.984 1	**0.985 0**	0.985 0	0.960 6	0.935 0	0.936 0	0.934 7	0.970 8

Table 7 Recognition rates on YALE, ORL and PIE data sets (FFS-II)

Data	CCA	PLS	LPCCA	DCCA	RCA-I	RCA-II	KCCA	KRCA-I	KRCA-II	PCA	LDA
YALE1	0.418 9	0.447 8	0.173 3	0.868 8	0.886 7	**0.887 8**	0.263 3	0.831 1	0.820 0	0.411 1	0.584 4
YALE2	0.541 1	0.530 0	0.340 0	0.795 5	0.841 1	**0.845 6**	0.345 6	0.772 2	0.773 3	0.525 6	0.712 2
YALE3	0.588 9	0.583 3	0.338 9	0.887 7	0.882 2	**0.897 8**	0.376 7	0.843 3	0.847 8	0.573 3	0.701 1
ORL1	0.875 5	0.875 5	0.860 5	0.925	**0.931 0**	0.928 0	0.619 0	0.821 5	0.818 5	0.875 0	0.340 0
ORL2	0.887 0	0.887 0	0.814 0	0.929	0.925 5	**0.931 0**	0.608 0	0.841 0	0.840 0	0.887 0	0.148 0
ORL3	0.893 5	0.893 5	0.877 5	0.966	0.965 5	**0.969 0**	0.855 0	0.920 0	0.935 0	0.892 0	0.858 5
PIE1(10)	0.850 7	0.850 7	0.856 5	0.922 4	0.924 0	**0.924 8**	0.815 4	0.860 0	0.859 8	0.850 7	0.910 3
PIE1(15)	0.911 4	0.911 4	0.912 7	0.951 7	**0.953 4**	0.953 1	0.884 0	0.906 8	0.907 4	0.911 4	0.943 3
PIE1(20)	0.939 2	0.939 2	0.943 1	0.964 1	0.964 5	**0.965 3**	0.920 5	0.934 2	0.934 8	0.939 1	0.958 9
PIE2(10)	0.849 8	0.849 8	0.872 8	0.924 0	0.927 9	**0.928 0**	0.810 8	0.866 0	0.866 2	0.849 8	0.915 7
PIE2(15)	0.907 8	0.907 8	0.923 5	0.949 4	**0.954 5**	0.953 7	0.887 7	0.916 4	0.915 6	0.907 7	0.947 9
PIE2(20)	0.936 7	0.936 7	0.946 7	0.961 5	0.965 3	**0.965 7**	0.912 5	0.939 4	0.939 9	0.936 7	0.961 3
PIE3(10)	0.878 5	0.878 5	0.873 0	0.952 2	**0.958 4**	0.957 7	0.893 8	0.894 5	0.895 4	0.871 7	0.957 9
PIE3(15)	0.929 5	0.929 5	0.926 0	0.973 1	**0.978 6**	0.978 1	0.943 7	0.939 5	0.936 4	0.923 9	0.976 8
PIE3(20)	0.958 3	0.958 3	0.953 4	0.985 6	**0.986 9**	0.986 8	0.965 2	0.958 6	0.957 1	0.952 7	0.984 1

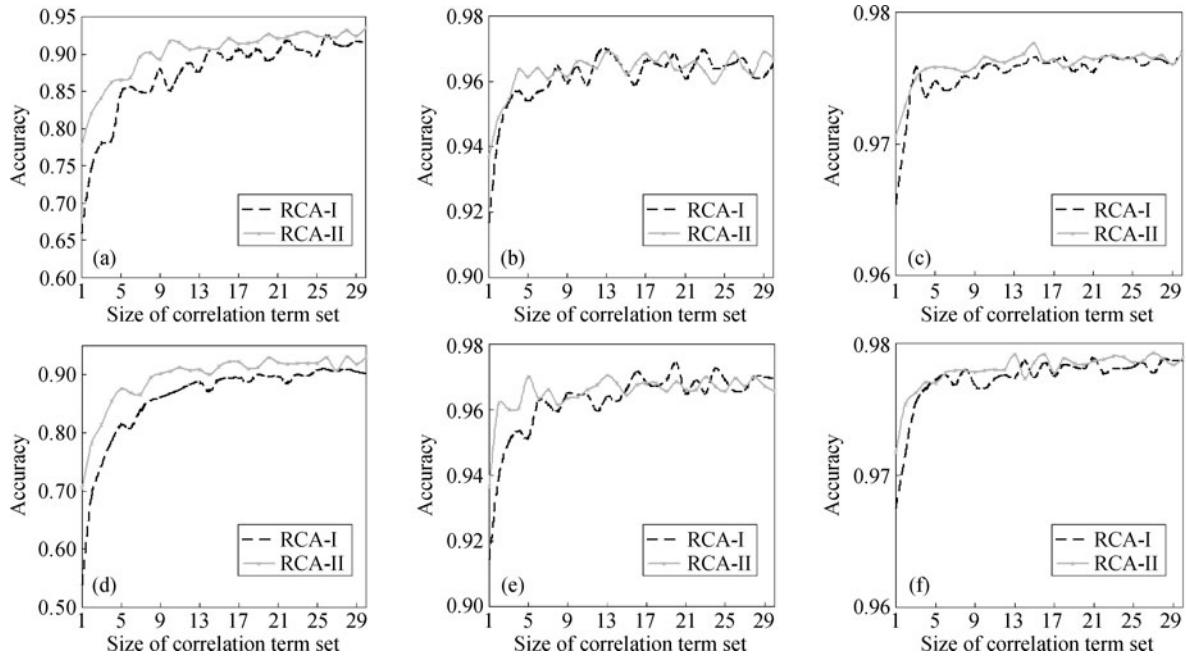

Fig. 5 Effect of correlation term set size on classification performances of the two methods RCA-I and RCA-II in face recognition experiments. (a) YALE3 (FFS-I); (b) ORL3 (FFS-I); (c) PIE3 (15) (FFS-I); (d) YALE3 (FFS-II); (e) ORL3 (FFS-II); (f) PIE3 (15) (FFS-II)

correlation term set and tend to be smooth when t is large enough. It is worth noting that when t is smaller, RCA-II could achieve better accuracies than RCA-I with the same correlation term set size.

6 Conclusion

In this paper, we proposed a multi-view dimensionality reduction method called canonical random correlation analysis (RCA), where not only the correlation between different views of a sample but also the cross-view correlations between within-class samples are considered. Moreover, we extended RCA to kernel RCA (KRCA) to extract the nonlinear correlations between different views. Experimental results on several multi-view data sets have validated the efficacy of our proposed methods.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61422204, 61473149) and the Jiangsu Natural Science Foundation for Distinguished Young Scholar (BK20130034).

Appendix

- $S = \{(x_i, y_i)\}_{i=1}^{n}$ observations with two views;
- $\mathcal{X}, \mathcal{Y}$ the respective sample for two views;
- w_x, w_y a pair of direction for two views;
- $E[\cdot]$ empirical expectation;
- C_{xy} between-sets covariance matrix;
- C_{xx}, C_{yy} within-sets covariance matrices;
- $X = [x_1 x_2 \ldots x_n]$ data matrix for view $\mathcal{X}$;
- $x_i y_j^{\mathrm{T}} (i = j)$ within-class correlation item;
- $\mathcal{X}_k, \mathcal{Y}_k$ the jth subset of respective view of S;
- $\tilde{\mathcal{X}}_k, \tilde{\mathcal{Y}}_k$ the jth subset in final correlation term set;
- $\tilde{\mathcal{X}}_k^{(l)}, \tilde{\mathcal{Y}}_k^{(l)}, l = 1, 2, \ldots, t$ t sets of bootstrap samples;
- $R_w \in R^{n \times n}$ weight matrix;
- $K_x, K_y \in R^{n \times n}$ kernel matrices.

References

1. Duda R O, Hart P E, Stork D G, Pattern Classification. 2nd ed. New York: Wiley-Interscience, 2000.
2. Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting on Association for Computational Lingustics. 1995, 189–196
3. Xia T, Tao D, Mei T, Zhang Y. Multiview spectral embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2010, 40(6): 1438–1446
4. Zheng H, Wang M, Li Z. Audio-visual speaker identification with multi-view distance metric learning. In: Proceedings of the 17th IEEE International Conference on Image Processing. 2010, 4561–4564
5. Wang M, Li H, Tao D, Lu K, Wu X. Multimodal graph-based reranking for Web image search. IEEE Transactions on Image Processing, 2012, 21(11): 4649–4661
6. Yu J, Wang M, Tao D. Semisupervised multiview distance metric learning for cartoon synthesis. IEEE Transactions on Image Processing,

2012, 21(11): 4636–4648

7. Long B, Philip SY, Zhang Z. A general model for multiple view unsupervised learning. In: Proceedings of the SIAM International Conference on Data Mining. 2008, 822–833
8. Han Y, Wu F, Tao D, Zhuang Y, Jiang J. Sparse unsupervised dimensionality reduction for multiple view data. IEEE Transactions on Circuits and Systems for Video Technology. 2012, 22(10): 1485–1496
9. Xie B, Mu Y, Tao D, Huang K. m-SNE: multiview stochastic neighbor embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2011, 41(4): 1088–1096
10. Hotelling H. Relation between two sets of variates. Biometrica, 1936, 28: 321–377
11. Diethe T, Hardoon D R, Shawe-Taylor J. Multiview fisher discriminant analysis. In: Proceedings of NIPS Workshop on Learning from Multiple Sources. 2008
12. Akaho S. A kernel method for canonical correlation analysis. In: Proceedings of the International Meeting of the Psychometric Society. 2001
13. Vía J, Santamaría I, Pérez J. A learning algorithm for adaptive canonical correlation analysis of several data sets. Neural Networks. 2007, 20(1): 139–152
14. Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. Neural Computation, 2004, 16(12): 2639–2664
15. Yang C, Wang L, Feng J. On feature extraction via kernels. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2008, 38(2): 553–557
16. Sun T, Chen S. Locality preserving CCA with applications to data visualization and pose estimation. Image and Vision Computing, 2007, 25(5): 531–543
17. Blaschko M B, Jacquelyn J A, Bartels A, Lampert C H, Gretton A. Semi-supervised kernel canonical correlation analysis with application to human fMRI. Pattern Recognition Letters, 2011, 32(11): 1572–1583
18. Blaschko M B, Lampert C H, Gretton A. Semi-supervised laplacian regularization of kernel canonical correlation analysis. Lecture Notes in Computer Science, 2008, 5211: 133–145
19. Golugula A, Lee G, Master S R, Feldman M D, Tomaszewski J E, Speicher D W, Madabhushi A. Supervised regularized canonical correlation analysis: integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery. BMC Bioinformatics, 2011, 12(1): 483
20. Thum A, Mönchgesang S, Westphal L, Lübken T, Rosahl S, Neumann S, Posch S. Supervised Penalized Canonical Correlation Analysis. 2014, arXiv preprint arXiv:1405.1534
21. Jing X Y, Hu R M, Zhu Y P, Wu S S, Liang C, Yang J Y. Intra-view and inter-view supervised correlation analysis for multi-view feature learning. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence. 2014
22. Jing X, Sun J, Yao Y, Sui Z. Supervised and unsupervised face recognition method base on 3CCA. In: Proceedings of International Conference on Automatic Control and Artificial Intelligence. 2012, 2009–2012
23. Guo S, Ruan Q, Wang Z, Liu S. Facial expression recognition using spectral supervised canonical correlation analysis. Journal of Information Science and Engineering, 2013, 29(5): 907–924
24. Shelton J A. Semi-supervised subspace learning and application to human functional magnetic brain resonance imaging data. Dissertation for the Doctoral Degree. Oxford: University of Oxford, 2010
25. Sun T, Chen S, Yang J, Shi P. A novel method of combined feature extraction for recognition. In: Proceedings of the 8th IEEE International Conference on Data Mining. 2008, 1043–1048
26. Majumdar A, Ward R. Robust classifiers for data reduced via random projections. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2010, 40(5): 1359–1371
27. Wegelin J A. A survey of partial least squares (PLS) methods, with emphasis on the two-block case. Department of Statistics, University of Washington, Technical Report. 2000, 371
28. Bourke P. Cross correlation. Auto Correlation–2D Pattern Identification, 1996
29. Theodoridis S, Koutroumbas K. Pattern Recognition. 3rd ed. New York: Academic Press, 2006
30. Sun Q, Zeng S, Liu Y, Heng P, Xia D. A new method of feature fusion and its application in image recognition. Journal of Pattern Recognition, 2005, 38(12): 2437–2448
31. Melzer T, Reiter M, Bischof H. Appearance models based on kernel canonical correlation analysis. Journal of Pattern Recognition, 2003, 36(9): 1961–1971
32. Shawe-Taylor J, Williams C K I, Cristianini N, Kandola J S. On the eigenspectrum of the gram matrix and the generalization error of kernel-PCA. IEEE Transactions on Information Theory, 2005, 51(7): 2510–2522
33. Bach F R, Jordan M I. Kernel independent component analysis. Journal of Machine Learning Research, 2002, 3: 1–48
34. Turk M, Pentland A. Eigenfaces for recognition. Journal of Cognitive Neuro Science, 1991, 3(1): 71–86
35. Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: recognition using class-specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711–720
36. He X, Cai D, Niyogi P. Lplacian score for feature selection. Advances in Neural Information Processing Systems. 2005, 18: 507–514
37. Cai D, He X, Hu Y, Han J, Huang T. Learning a spatially smooth subspace for face recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2007, 1–7
38. Ahonen T, Hadid A, Pietikainen M. Face recognition with local binary patterns. In: Proceedings of the 8th European Conference on Computer Vision. 2004, 469–481
39. Zhang J, Zhang D. A novel ensemble construction method for multiview data using random cross-view correlation between within-class examples. Pattern Recognition, 2011, 44(6): 1162–1171

Yanyan Zhang received the MS degree in computer science and application from Nanjing University of Aeronautics and Astronautics, China in 2010. Now she is a teaching assistant in PLA University of Science and Technology, China. Her current research interests include face recognition and sparse learning.

Jianchun Zhang received the MS degree in computer science and application from Nanjing University of Aeronautics and Astronautics, China in 2010. His research interests include pattern recognition and image processing.

Zhisong Pan received the BS degree in computer science and MS degree in computer science and application from PLA Information Engineering University, China, in 1991 and 1994 respectively, and the PhD degree in Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, China in 2003. From July 2006 to the present, he has led several key projects of intelligent data processing for the network management. His current research interests mainly include pattern recognition, machine learning and neural networks.

Daoqiang Zhang received the BS and PhD degrees in computer science from Nanjing University of Aeronautics and Astronautics (NUAA), China in 1999 and 2004, respectively. He is currently a professor in the Department of Computer Science and Engineering of NUAA. His research interests include machine learning, pattern recognition, data mining, and image processing. In these areas, he has published over 40 technical papers in refereed international journals or conference proceedings. He was nominated for the National Excellent Doctoral Dissertation Award of China in 2006, and won the best paper award at the 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06). He has served as a program committee member for several international and native conferences. He is also a member of Chinese Association of Artificial Intelligence (CAAI) Machine Learning Society.