Intra-View and Inter-View Supervised Correlation Analysis for Multi-View Feature Learning

Xiao-Yuan Jing[1,3], Rui-Min Hu[2,*], Yang-Ping Zhu[3], Shan-Shan Wu[3], Chao Liang[2], Jing-Yu Yang[4]

[1]State Key Laboratory of Software Engineering, School of Computer, Wuhan University, China
[2]National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, China
[3]College of Automation, Nanjing University of Posts and Telecommunications, China
[4]College of Computer Science, Nanjing University of Science and Technology, China
*Corresponding author: hrm1964@163.com

Abstract

Multi-view feature learning is an attractive research topic with great practical success. Canonical correlation analysis (CCA) has become an important technique in multi-view learning, since it can fully utilize the inter-view correlation. In this paper, we mainly study the CCA based multi-view supervised feature learning technique where the labels of training samples are known. Several supervised CCA based multi-view methods have been presented, which focus on investigating the supervised correlation across different views. However, they take no account of the intra-view correlation between samples. Researchers have also introduced the discriminant analysis technique into multi-view feature learning, such as multi-view discriminant analysis (MvDA). But they ignore the canonical correlation within each view and between all views. In this paper, we propose a novel multi-view feature learning approach based on intra-view and inter-view supervised correlation analysis (I^{2}SCA), which can explore the useful correlation information of samples within each view and between all views. The objective function of I^{2}SCA is designed to simultaneously extract the discriminatingly correlated features from both inter-view and intra-view. It can obtain an analytical solution without iterative calculation. And we provide a kernelized extension of I^{2}SCA to tackle the linearly inseparable problem in the original feature space. Four widely-used datasets are employed as test data. Experimental results demonstrate that our proposed approaches outperform several representative multi-view supervised feature learning methods.

Introduction

In real world applications, datasets are usually described with different views or representations. Multi-view feature learning refers to learning with multiple feature sets that reflect different characteristics or views of data, which is an vital research direction (Guo, 2013; Xu, Tao and Xu 2013; Wang, Nie and Huang 2013; Memisevic 2012; Wang, Li, and Ogihara 2012). Co-training and canonical correlation analysis are two representative and effective techniques in multi-view learning (Sun and Chao 2013). Co-training based methods (Kumar and Daume 2011a; Kumar and Daume 2011) are usually used for semi-supervised classification that combines both labeled and unlabeled data under multi-view setting. Canonical correlation analysis (CCA, Hardoon, Szedmak, and Shawe-Taylor 2004) has become an important technique in multi-view learning, since it can fully utilize the inter-view correlation. Multi-view CCA (MCCA, Rupnik and Shawe-Taylor 2010) is an unsupervised method. In this paper, we mainly study the CCA based multi-view supervised feature learning technique where the labels of training samples are known.

Recently, several supervised CCA based multi-view feature learning methods have been presented, such as multiple discriminant CCA (MDCCA, Gao et al. 2012), multiple principal angles (MPA, Su et al. 2012). These methods focus on investigating supervised correlation across different views. To deal with the linearly inseparable problem, researchers extend CCA to be kernel CCA (Sun et al. 2007; Leurgans, Moyeed, and Silverman 1993; Lai and Fyfe 2000; Bach and Jordan 2002). All methods mentioned above only reveal the linear or nonlinear correlation relationship between features of multiple views. However, they take no account of the intra-view correlation between samples, which is also an important part when exploiting supervised correlation among the samples. Therefore, in this paper, we need to simultaneously extract the discriminatingly correlated features from both inter-view and intra-view. The

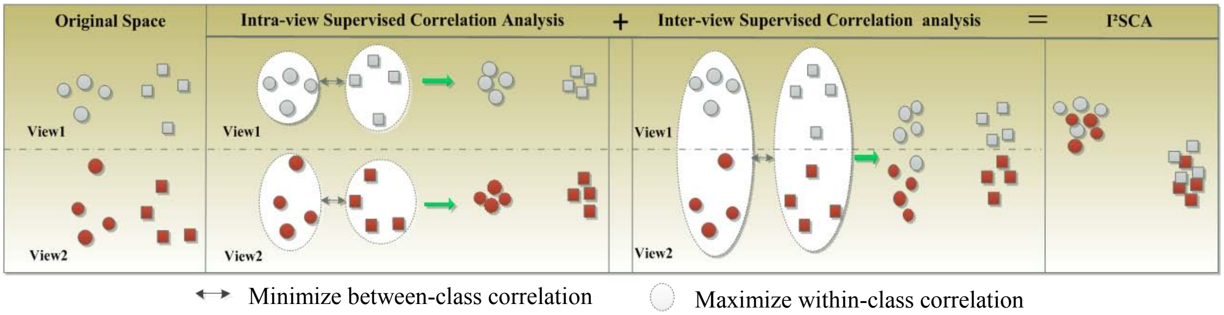

Original Space
Intra-view Supervised Correlation Analysis
+
Inter-view Supervised Correlation analysis
=
I²SCA
View1
View2
View1
View2
View1
View2
↔ Minimize between-class correlation
○ Maximize within-class correlation

Fig. 1. Conceptual illustration of the proposed I^{2}SCA, where different shapes (○ and □) stand for samples from different classes, and different colors (i.e. red and gray) stand for samples from different views.

effectiveness of the intra-view correlation samples can be further visualized in Fig. 1. As suggested in Figure 1, intra-view correlation "glues" intra-class samples better than inter-view correlation. Hence, their combination is able to further uncovered more discriminative information.

Researchers have also presented some discriminant analysis based multi-view feature learning methods, such as generalized multi-view analysis (GMA, Sharma et al. 2012), kernelized GMA (KGMA) and the multi-view fisher discriminant analysis (MFDA) (Diethe, Hardoon, and Shawe-Taylor 2008). MFDA also has a sparse version, that is, the SMFDA (Diethe, Hardoon, and Shawe-Taylor 2010). Moreover, Kan et al. (2012) presented a Multi-view discriminant analysis (MvDA) method, which can maximize the between-class variations and minimize the within-class variations of the learning common space from both intra-view and inter-view. Nevertheless, these methods ignore canonical correlation information from both intra-view and inter-view, which depicts the relationship among multiple views. Hence, we may improve the multi-view feature learning results if the discriminant correlation information among multiple views can be exploited.

For current multi-view supervised feature learning methods, there is still room for further improvements. We summarize the contributions of this paper as following points:

(1) We propose a novel multi-view feature learning approach based on intra-view and inter-view supervised correlation analysis (I^{2}SCA). It can explore the useful correlation information of samples within each view and between all views. The idea of the proposed approach is illustrated in Fig. 1.

(2) The objective function of I^{2}SCA is designed to make the within-class correlation from both inter-view and intra-view maximized, while make the between-class correlation from both inter-view and intra-view minimized. Moreover, our approach can obtain an analytical solution without iterative calculation.

(3) We provide a kernelized extension of I^{2}SCA, that is, KI^{2}SCA, to tackle the linearly inseparable problem in the original feature space.

The rest of this paper is organized as follows. In Section 2, we introduce the related works. In Section 3, we describe the proposed I^{2}SCA approach and its kernelized extension. Experimental results and analysis are provided in Section 4, and conclusion is drawn in Section 5.

Related work

CCA based Multi-view Feature Learning Methods

CCA finds linear combinations corresponding to two views, such that all transformed variables own maximum correlation. It cannot be directly applied to multi-view data, which leads to the formulation of multi-view CCA (MCCA, Rupnik and Shawe-Taylor 2010). It tends to obtain high correlations between all new variables simultaneously by optimizing the characteristics of the dispersion matrix of new variables. Multiset integrated CCA (MICCA, Yuan et al. 2010) describes an integrated correlation among multi-view variables based on generalized correlation coefficient. However, these methods are all unsupervised.

In order to utilize the supervised correlation across different views, supervised CCA methods were developed. Discriminant analysis of canonical correlations (DCC, Kim, Kittler, and Cipolla 2007) maximizes the within-class correlation and minimizes the between-class correlation for two sets of variables. Multiple discriminant CCA (MDCCA, Gao et al. 2012) was designed for multiple views, which demonstrates that CCA, MCCA and DCC are special cases of the DMCCA method. Multiple principal angles (MPA, Su et al. 2012) iteratively learns the view-specific projection transformations by following the traditional Fisher discriminant manner, and learns a global discriminative subspace on which the principal angles among multiple subspaces of same classes are minimized while those of different classes are maximized. The nonlinear extension of DCC, namely kernelized discriminative canonical correlation analysis (KDCCA, Sun et al. 2007), was presented to deal with linearly inseparable problem.

Discriminant Analysis based Multi-view Feature Learning Methods

Linear discriminant analysis (LDA, Belhumeur, Hespanha, and Kriegman 1997) technique is widely used in feature learning, which employs the famous Fisher criterion to minimize the within-class scatter while maximize the between-class scatter. To obtain multi-view counterparts of LDA and other feature learning methods, a generalized multi-view analysis (GMA, Sharma et al. 2012) framework was developed, and kernelized GMA (KGMA) was also given. Multi-view Fisher Discriminant Analysis (MFDA) (Diethe, Hardoon, and Shawe-Taylor 2008) learns the classifiers by making the predicted labels of these classifiers consistent with their real label. However, MFDA can only be applied for binary class problems. To address this problem, Chen and Sun (2009) extended MFDA to a multi-class form by using a hierarchical clustering technique. By incorporating multiple views of data in a perceptive transfer learning framework, Yang and Gao (2013) presented the multi-view discriminant transfer learning method. Moreover, Kan et al. (2012) presented a Multi-view discriminant analysis (MvDA) method, which can maximize the between-class variations and minimize the within-class variations of the learning common space from both intra-view and inter-view. MvDA is able to jointly acquire projection transforms by solving a generalized Rayleigh quotient.

Approach Description

Suppose that there are N views. Notation usage is provided in a summary as follows:

X_i : Sample set of the i^{th} view;

n : Number of samples from each view;

c : Number of classes in each view;

n_j : Number of samples from the j^{th} class in each view;

N : Number of views;

x^i_{jk} : The k^{th} sample from the j^{th} class in X_i ;

$\overline{x}^i$: Mean of all samples from X_i ;

$\hat{X}_i$: Mean-normalized X_i , where x^i_{jk} is normalized by using $\hat{x}^i_{jk} = x^i_{jk} - \overline{x}^i$;

S^i_b : Between-class scatter matrix of X_i ;

S^i_t : Total scatter matrix of X_i ;

C^{ij}_b : Between-class correlation between features of X_i and X_j ;

C^{ij}_w : Within-class correlation between features of X_i and X_j ;

C^{ij} : Supervised correlation between the samples of the i^{th} and j^{th} views;

w_i : Projective vector of X_i ;

W_i : Projective transformation of X_i , where W_i consists of d projective vectors;

Z_i : Projected features of samples set of the i^{th} view X_i .

Intra-view Supervised Correlation Analysis

Let $E_n = [1, \cdots, 1]^T$, and $A = diag\left(E_{n_1}, E_{n_2}, \cdots, E_{n_c}\right)$ denote a $n \times n$ symmetric, positive semi-definite, blocked diagonal matrix, where E_{n_k} is a $n_k \times n_k$ matrix with all its elements equal to 1. Assume that C^i_w and C^i_b denote the within-class correlation and between-class correlation of samples of the i^{th} view, respectively. Their definitions are given as follows:

$$\begin{aligned}
C^i_w &= \frac{\left(1 \Big/ \sum_{p=1}^{c} n_p^2\right) \sum_{p=1}^{c}\sum_{r=1}^{n_p}\sum_{t=1}^{n_p} w_i^T \left(x^i_{pr} - \overline{x}^i\right)\left(x^i_{pt} - \overline{x}^i\right)^T w_i}{\sqrt{\frac{1}{n}\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T \left(x^i_{pr} - \overline{x}^i\right)\left(x^i_{pr} - \overline{x}^i\right)^T w_i}\sqrt{\frac{1}{n}\sum_{p=1}^{c}\sum_{t=1}^{n_p} w_i^T \left(x^i_{pt} - \overline{x}^i\right)\left(x^i_{pt} - \overline{x}^i\right)^T w_i}} \\
&= \frac{n\sum_{p=1}^{c}\sum_{r=1}^{n_p}\sum_{t=1}^{n_p} w_i^T \hat{x}^i_{pr} x^{i\ T}_{pt} w_i}{\left(\sum_{p=1}^{c} n_p^2\right)\sqrt{\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T \hat{x}^i_{pr} x^{i\ T}_{pr} w_i}\sqrt{\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T \hat{x}^i_{pr} x^{i\ T}_{pr} w_i}} \\
&= \frac{n w_i^T \hat{X}_i A X_i^T w_i}{\left(\sum_{p=1}^{c} n_p^2\right) w_i^T \hat{X}_i X_i^T w_i}
\end{aligned} , \quad (1)$$

$$\begin{aligned}
C^i_b &= \frac{\left[1 \Big/ \left(n^2 - \sum_{p=1}^{c} n_p^2\right)\right] \sum_{p=1}^{c}\sum_{\substack{q=1 \\ q \neq p}}^{c}\sum_{r=1}^{n_p}\sum_{t=1}^{n_q} w_i^T \left(x^i_{pr} - \overline{x}^i\right)\left(x^i_{qt} - \overline{x}^i\right)^T w_i}{\sqrt{\frac{1}{n}\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T \left(x^i_{pr} - \overline{x}^i\right)\left(x^i_{pr} - \overline{x}^i\right)^T w_i}\sqrt{\frac{1}{n}\sum_{q=1}^{c}\sum_{t=1}^{n_q} w_i^T \left(x^i_{qt} - \overline{x}^i\right)\left(x^i_{qt} - \overline{x}^i\right)^T w_i}} \\
&= \frac{n\sum_{p=1}^{c}\sum_{\substack{q=1 \\ q \neq p}}^{c}\sum_{r=1}^{n_p}\sum_{t=1}^{n_q} w_i^T \hat{x}^i_{pr} x^{iT}_{qt} w_i}{\left(n^2 - \sum_{p=1}^{c} n_p^2\right)\sqrt{\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T \hat{x}^i_{pr} \hat{x}^{i\ T}_{pr} w_i}\sqrt{\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T x^i_{pr} x^{i\ T}_{pr} w_i}} \\
&= \frac{n\left(w_i^T \hat{X}_i \hat{E}_n E_n^{\ T} X_i^T w_i - w_i^T X_i A X_i^T w_i\right)}{\left(n^2 - \sum_{p=1}^{c} n_p^2\right) w_i^T \hat{X}_i X_i^T w_i} \\
&= -\frac{n w_i^T \hat{X}_i A X_i^T w_i}{\left(n^2 - \sum_{p=1}^{c} n_p^2\right) w_i^T \hat{X}_i X_i^T w_i}
\end{aligned} . \quad (2)$$

The supervised correlation between samples of the i^{th} view is thus defined as:

$$\begin{aligned}
C^i &= C^i_w - \alpha C^i_b \\
&= \left(\frac{n}{\sum_{p=1}^{c} n_p^2} + \frac{n\alpha}{n^2 - \sum_{p=1}^{c} n_p^2}\right)\frac{w_i^T \hat{X}_i A X_i^T w_i}{w_i^T \hat{X}_i X_i^T w_i}
\end{aligned} , \quad (3)$$

where $\alpha > 0$ is a tunable parameter that indicates the relative significance of the within-class correlation of intra-view C_w^i versus the between-class correlation of intra-view C_b^i. For intra-view supervised correlation analysis, we design the following objective function:

$$\max_{w_i,\, i=1,2,\cdots,N} \sum_{i=1}^{N} C^i . \tag{4}$$

The goal of Formula (4) is to preserve the useful within-class correlation from each view and eliminate the adverse between-class correlation from each view, which is favorable for classification.

Inter-view Supervised Correlation Analysis

Similarly, we define the within-class correlation C_w^{ij} and between-class correlation C_b^{ij} between the samples of the i^{th} and j^{th} views as follows:

$$\begin{aligned} C_w^{ij} &= \frac{\left(1\Big/\sum_{p=1}^{c} n_p^2\right)\sum_{p=1}^{c}\sum_{r=1}^{n_p}\sum_{t=1}^{n_p} w_i^T\left(x_{pr}^i-\bar{x}^i\right)\left(x_{pt}^j-\bar{x}^j\right)^T w_j}{\sqrt{\frac{1}{n}\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T\left(x_{pr}^i-\bar{x}^i\right)\left(x_{pr}^i-\bar{x}^i\right)^T w_i}\sqrt{\frac{1}{n}\sum_{p=1}^{c}\sum_{t=1}^{n_p} w_j^T\left(x_{pt}^j-\bar{x}^j\right)\left(x_{pt}^j-\bar{x}^j\right)^T w_j}} , \\ &= \frac{n w_i^T \hat{\hat{X}}_i A X_j^T w_j}{\left(\sum_{p=1}^{c} n_p^2\right)\sqrt{w_i^T \widehat{X_i X_i^T} w_i}\sqrt{w_j^T X_j X_j^T w_j}} \end{aligned} \tag{5}$$

$$\begin{aligned} C_b^{ij} &= \frac{\left[1\Big/\left(n^2-\sum_{p=1}^{c} n_p^2\right)\right]\sum_{p=1}^{c}\sum_{\substack{q=1\\q\neq p}}^{c}\sum_{r=1}^{n_p}\sum_{t=1}^{n_q} w_i^T\left(x_{pr}^i-\bar{x}^i\right)\left(x_{qt}^j-\bar{x}^j\right)^T w_j}{\sqrt{\frac{1}{n}\sum_{p=1}^{c}\sum_{r=1}^{n_p} w_i^T\left(x_{pr}^i-\bar{x}^i\right)\left(x_{pr}^i-\bar{x}^i\right)^T w_i}\sqrt{\frac{1}{n}\sum_{q=1}^{c}\sum_{t=1}^{n_q} w_j^T\left(x_{qt}^j-\bar{x}^j\right)\left(x_{qt}^j-\bar{x}^j\right)^T w_j}} . \\ &= -\frac{n w_i^T \hat{\hat{X}}_i A X_j^T w_j}{\left(n^2-\sum_{p=1}^{c} n_p^2\right)\sqrt{w_i^T \widehat{X_i X_i^T} w_i}\sqrt{w_j^T X_j X_j^T w_j}} \end{aligned} \tag{6}$$

The supervised correlation between the samples of the i^{th} and j^{th} views is defined as:

$$\begin{aligned} C^{ij} &= C_w^{ij} - \beta C_b^{ij} \\ &= \left(\frac{n}{\sum_{p=1}^{c} n_p^2} + \frac{n\beta}{n^2-\sum_{p=1}^{c} n_p^2}\right)\frac{w_i^T \hat{\hat{X}}_i A X_j^T w_j}{\sqrt{w_i^T \widehat{X_i X_i^T} w_i}\sqrt{w_j^T X_j X_j^T w_j}} \end{aligned} , \tag{7}$$

where $\beta > 0$ is a tunable parameter that indicates the relative significance of the within-class correlation of inter-view C_w^{ij} versus the between-class correlation of inter-view C_b^{ij}. Note that $C^{ij} = C^{ji}$.

For inter-view supervised correlation analysis, we design the following objective function:

$$\max_{w_i,\, i=1,2,\cdots,N} \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N} C^{ij} . \tag{8}$$

The goal of Formula (8) is to preserve the useful correlation of inter-view samples from the same class and eliminate the adverse correlation of inter-view samples from different classes.

I^{2}SCA Scheme

In order to effectively make full use of correlation information within each view and between different views, we combine intra-view and inter-view supervised correlation analysis, and design the following I^{2}SCA scheme:

$$\max_{w_p,\, p=1,2,\cdots,N} \sum_{i=1}^{N} C^i + \beta\sum_{p=1}^{N}\sum_{\substack{q=1\\q\neq p}}^{N} C^{pq} , \tag{9}$$

which is equivalent to

$$\begin{aligned} &\max_{W_p,\, p=1,2,\cdots,N} \xi_1\sum_{p=1}^{N}\sum_{\substack{q=1\\q\neq p}}^{N} w_p^T \widehat{X_p A} X_q^T w_q + \xi_2\sum_{p=1}^{N} w_p^T X_p A X_p^T w_p \\ &s.t. \quad w_p^T S_t^p w_p = 1,\ p = 1,2,\cdots,N \end{aligned} , \tag{10}$$

where $\xi_1 = \frac{1}{\sum_{p=1}^{c} n_p^2} + \frac{\alpha}{n^2-\sum_{p=1}^{c} n_p^2}, \xi_2 = \frac{1}{\sum_{p=1}^{c} n_p^2} + \frac{\beta}{n^2-\sum_{p=1}^{c} n_p^2}$. To get an analytical solution, we simplify Formula (10) by

$$\begin{aligned} &\max_{W_p,\, p=1,2,\cdots,N} \xi_1\sum_{p=1}^{N}\sum_{\substack{q=1\\q\neq p}}^{N} w_p^T \widehat{X_p A} X_q^T w_q + \xi_2\sum_{p=1}^{N} w_p^T X_p A X_p^T w_p \\ &s.t. \quad \sum_{p=1}^{N} w_p^T S_t^p w_p = N \end{aligned} , \tag{11}$$

which can be rewritten as

$$\max_{w} \ w^T G w \qquad s.t.\ w^T F w = N , \tag{12}$$

where

$$G = \begin{bmatrix} \xi_2 \widehat{X_1 A X}_1 & \xi_1 X_1 A X_2 & \cdots & \xi_1 X_1 A X_N \\ \xi_1 \widehat{X_2 A X}_1 & \xi_2 X_2 A X_2 & \cdots & \xi_1 X_2 A X_N \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1 \widehat{X_N A X}_1 & \xi_1 X_N A X_2 & \cdots & \xi_2 X_N A X_N \end{bmatrix},$$

$$F = \begin{bmatrix} S_t^1 & 0 & 0 & 0 \\ 0 & S_t^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & S_t^N \end{bmatrix} \text{ and } w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}.$$

Formula (12) can be formulated as a generalized eigenvalue problem:

$$Gw = \lambda F w . \tag{13}$$

Once the eigenvectors $w^k\ (k = 1,2,\cdots,d)$ associated with the first d largest eigenvalues of $F^{-1}G$ are obtained, we can get $w_1^k, w_2^k, \cdots, w_N^k$ from w^k. Let $W_j = \left[w_j^1, w_j^2, \cdots, w_j^d\right]$, where $j = 1,2,\cdots,N$. Then we can obtain the projected features Z_i by $Z_i = W_i^T X_i$, where $i = 1,2,\cdots,N$, and fuse these features as follows:

$$Z = \left[Z_1^T, Z_2^T, \cdots, Z_N^T\right]^T . \tag{14}$$

Finally, we use the nearest neighbor classifier with the cosine distance to classify Z.

Kernel I^{2}SCA

To tackle the linearly inseparable problem, I^{2}SCA can be extended by using kernel method to work in a nonlinear feature space instead of the original input space. Kernel trick has shown its effectiveness in many methods, including some kernel CCA methods (Sun et al. 2007; Leurgans, Moyeed, and Silverman 1993; Lai and Fyfe 2000; Bach and Jordan 2002). We first perform the kernel mapping for samples and then realize the I^{2}SCA in the mapped space to obtain projection transformation. Suppose there are N implicit mappings $\phi_i, i=1,2,...,N$, the samples mapped to the corresponding spaces are denoted by $\phi_i(X_i)=\left[\phi_i(x_1^i),\ldots,\phi_i(x_n^i)\right]$. Thus, the mapped multi-view data is $\{\phi_1(X_1),...,\phi_N(X_N)\}$. According to the kernelized method in (Sun et al. 2007), the basis vector pair can be represented as $w_i=\phi_i(X_i)\alpha$, where $\alpha,\beta\in R^n$ denote corresponding combination coefficient vectors. Note that the kernel trick $K_{x_i}=\phi_i(X_i)^T\phi_i(X_i)=k_x(x_p^i,x_q^i)$, $p=1,...,n,q=1,...,n$. Substituting $w_i=\phi_i(X_i)\alpha$ into Formula (9), we employ the kernel trick and kernel I^{2}SCA can be reformulated as

$$\max_{W_i,\iota=1,2,\cdots,N} \xi_1\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\alpha_i{}^T K_{x_i}AK_{x_j}\alpha_j+\xi_2\sum_{i=1}^{N}\alpha_i{}^T K_{x_i}AK_{x_j}\alpha_j \quad , \qquad (15)$$
$$s.t. \qquad \alpha_i{}^T K_{x_i}K_{x_i}\alpha_i=1,\ p=1,2,\cdots,N$$

where $\xi_1=\frac{1}{\sum_{p=1}^{c}n_p^2}+\frac{\alpha}{n^2-\sum_{p=1}^{c}n_p^2},\xi_2=\frac{1}{\sum_{p=1}^{c}n_p^2}+\frac{\beta}{n^2-\sum_{p=1}^{c}n_p^2}$, $K_{x_i}=\phi_i(X_i)^T\phi_i(X_i)=k_x(x_p^i,x_q^i)$ $,p=1,...,n,q=1,...,n$, $k_x(\cdot,\cdot)$ and $k_y(\cdot,\cdot)$ denote the corresponding Mercer kernels. $p=1,...,n,q=1,...,n$, $k_x(\cdot,\cdot)$ and $k_y(\cdot,\cdot)$ denote the corresponding Mercer kernels. We simplify Formula (15) by

$$\max_{W_i,\iota=1,2,\cdots,N} \xi_1\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\alpha_i{}^T K_{x_i}AK_{x_j}\alpha_j+\xi_2\sum_{i=1}^{N}\alpha_i{}^T K_{x_i}AK_{x_j}\alpha_j \quad . \qquad (16)$$
$$s.t. \qquad \sum_{i=1}^{N}\alpha_i{}^T K_{x_i}K_{x_i}\alpha_i=N,i=1,2,...,N$$

Similar to I^{2}SCA, Formula (16) can also be written as a generalized eigenvalue problem. Its solution can be solved by reference to Section 3.3.

It is noticed that in the experiments, the tunable parameters α and β are selected with 5-fold cross validation on the training set.

Experiments

In this section, we evaluate the proposed approaches with several related multi-view supervised feature learning methods including MDCCA (Gao et al. 2012), MPA (Su et al. 2012), MvDA (Kan et al. 2012) and KGMA (Sharma et al. 2012) on four public datasets. Note that KGMA refers to the kernel GMA, which outperforms GMA.

Recognition Performance Evaluation

Experiment on Multiple Feature Dataset

The multiple feature dataset (MFD) (Yuan et al. 2010) contains 10 classes of handwritten numerals, i.e. 10 numbers from 0 to 9. These digit characters are represented in terms of the next six feature sets, as shown in Table 1. There are 200 samples per class (for a total of 2000 samples) in each feature sets.

Table 1. Six sets of features of handwritten numerals in MFD

Pix: 240-dimension pixelaveragesfeaturein2_3 windows
Fac: 216-dimension profile correlations feature
Fou: 76-dimension Fourier coefficients of the character shapes feature
Kar: 64-dimension Karhunen–Loeve coefficients feature
Zer: 47-dimension Zernike moments feature
Mor: 6-dimension morphological feature

In this experiment, 100 samples per class (numeral) are randomly chosen as the training set, while the remaining samples are regarded as the testing set. Table 2 shows the average recognition rates of 20 random runs for all compared methods on MFD datasets.

Table 2. Average recognition rates (%) on MFD database

Method	MDCCA	MPA	MvDA	KGMA	I^{2}SCA	KI^{2}SCA
Result	95.52	94.87	95.53	95.64	**96.67**	**97.57**

From Table 2, we observe that I^{2}SCA improves the average recognition rate at least by 1.03% (=96.67-95.64), and KI^{2}SCA improves the average rate at least by 1.93% (=97.57-95.64) as compared with other methods. Here, to elaborate more about the improvements of our approach, we use the second digit precision. This made particular sense when data size grows big so that even a small portion of its improvement could gain significant economical concern. We also conduct the second digit precision in the following experiments.

Experiment on Coil-20 Dataset

The COIL-20 database (Murase and Nayar 1995) contains 7200 grayscale images of 100 objects (72 images per object) under various poses. Objects are rotated through 360 degrees and taken at the intervals of 5º. The size of each object image is 64×64 pixels.

In this experiment, 36 images per class are randomly chosen to form the training set, while the remaining images are regarded as the testing set. We extract Gabor transformation features (Grigorescu, Petkov and Kruizinga 2002), Karhunen-Loeve transformation features (Fukunaga and Koontz 1970) and Local Binary Patterns features (Ahonen, Hadid, and Pietikainen 2006) to form three sets of features. We perform the principal component analysis (PCA, Belhumeur, Hespanha, and Kriegman 1997) transformation to reduce their dimensions to 150, respectively.

Table 3. Average recognition rates (%) on Coil database

Method	MDCCA	MPA	MvDA	KGMA	I^2SCA	KI^2SCA
Result	95.53	95.05	96.07	96.71	**98.28**	**98.87**

Table 3 shows that I^2SCA improves the average recognition rate at least by 1.57% (=98.28-96.71), and KI^2SCA improves the average rate at least by 2.16% (=98.87-96.71) as compared with other methods.

Experiments on Multi-PIE Dataset

Multi-PIE dataset (Cai et al. 2006) contains more than 750,000 images of 337 people under various views, illumination and expressions (PIE). Here, a subset about 1632 samples from 68 classes in 5 poses (C05, C07, C09, C27, C29) is selected as test data. All images from Multi-PIE are 64×64 pixels. We perform PCA transformation to reduce their dimensions to 150, respectively.

Regular Face Recognition Experiment. We randomly choose 8 samples per class as the training samples, while the remaining samples are regarded as the testing set. Thus, the total number of training samples and testing samples is 544 and 1088, respectively. Table 4 shows the average recognition rates of all compared methods across of 20 random runs for all compared methods on Multi-PIE dataset. Table 4 shows that I^2SCA and KI^2SCA perform better in contrast with other related methods, which improve the average recognition rate at least by 1.09% (=97.54-96.45) and 1.78% (=98.03-96.45) , respectively.

Table 4. Average recognition rates (%) on Multi−PIE database

Method	MDCCA	MPA	MvDA	KGMA	I^2SCA	KI^2SCA
Result	95.21	94.77	95.84	96.45	**97.54**	**98.23**

Face Recognition across Pose. In this sub-section, we perform the classification experiment where the gallery and query set belongs to different views. We still randomly choose 8 samples per class as the training samples, while remaining ones from one view are used as gallery set and others from another view are probe set. Tables 5, 6 and 7 shows all the average recognition rates of 20 random runs for I^2SCA, KI^2SCA and KGMA.

Table 5. Average recognition rates (%) of KGMA on Multi−PIE

Gallery Probe	C05	C07	C09	C27	C29
C05	94.13	90.01	89.30	91.11	89.53
C07	87.53	93.97	84.33	88.63	88.16
C09	87.66	85.89	94.45	89.48	87.17
C27	88.97	88.90	88.66	93.01	88.61
C29	81.50	81.22	80.19	84.15	92.47

Table 6. Average recognition rates (%) of I^2SCA on Multi−PIE

Gallery Probe	C05	C07	C09	C27	C29
C05	94.66	90.28	89.28	91.11	90.13
C07	90.45	94.45	85.18	90.92	90.17
C09	85.73	88.18	94.76	90.46	88.68
C27	87.91	90.59	90.27	93.64	90.09
C29	86.05	86.22	85.71	87.90	93.13

Table 7. Average recognition rates (%) of KI^2SCA on Multi−PIE

Gallery Probe	C05	C07	C09	C27	C29
C05	95.82	91.93	91.17	91.53	92.22
C07	92.45	95.60	89.29	91.56	90.37
C09	88.73	90.34	95.59	89.27	88.61
C27	88.44	90.68	91.71	95.63	89.53
C29	87.63	88.06	87.10	88.32	94.16

Tables 5-7 show that most of recognition rates of I^2SCA and KI^2SCA are higher than those of KGMA.

Experiments on LFW Dataset

The LFW database (Huang et al. 2012) is a dataset for studying face recognition in unconstrained environments. This dataset contains a total of 13233 images and 5749 people. There are 1680 of the people pictured have two or more distinct photos. We crop each facial image to 60×60 pixels and all images will be transformed into grey-level images at first.

In this experiment, we choose the individuals who hold more than 14 photos in the dataset and 14 images per class are used to form the total sample set. Then we have 106 classes of samples. We randomly choose the 8 images per class to form the training set, while the remaining images are regarded as the testing set. We extract Gabor transformation features, Karhunen-Loeve transformation features and Local Binary Patterns features to form three sets of features. We perform the principal component analysis transformation to reduce their dimensions to 150, respectively.

Table 8 shows the average recognition rates of 20 random runs for all compared methods on LFW dataset.

Table 8. Average recognition rates (%) on LFW dataset

Method	MDCCA	MPA	MvDA	KGMA	I^2SCA	KI^2SCA
Result	85.57	84.97	87.74	88.39	**89.57**	**90.32**

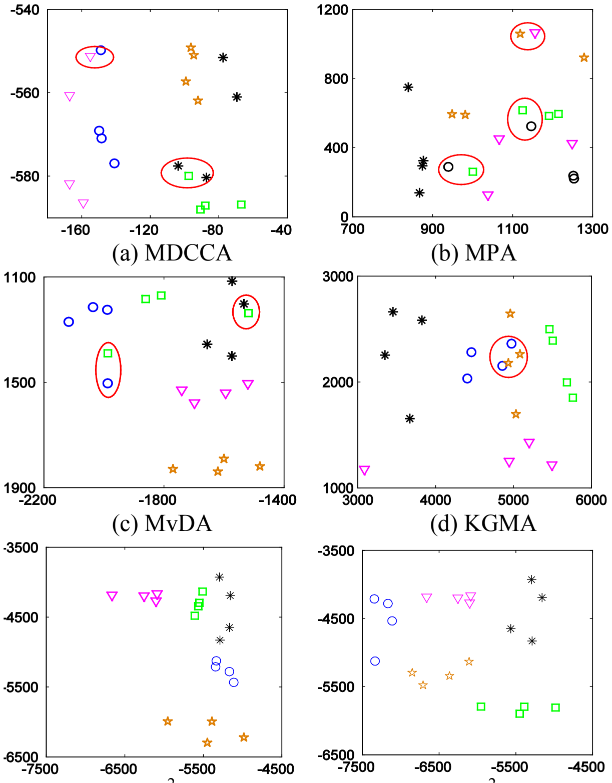
From table 8 we can find that both I^2SCA and KI^2SCA can perform better in the comparison with the other methods. I^2SCA improves the average recognition rate at least by 1.18%(=89.57-88.39) and KI^2SCA improves the average recognition rate at least by 1.93%(=90.32-88.39).

Further Experimental Analysis

Distribution of Samples

In order to analyze the separabilities of all methods, we provide the distribution of samples with two principal features extracted from 5 different views by using all related methods on Multi-PIE dataset. Here, we employ the PCA transform to obtain two principal features.

Fig. 2 shows the distribution of two principal features of 20 samples (from 5 different persons and 4 samples per person) extracted from the compared methods on the Multi−PIE dataset, where the markers (* , ☆, ○, □, and ▽) with different colors stand for 5 different persons. In Fig. 2, the red oval ring indicates the bad separability, which might lead to misclassification. Fig. 2 shows that the proposed approaches achieve preferable separabilities in comparison with other methods.



(a) MDCCA (b) MPA

(c) MvDA (d) KGMA

(e) I^{2}SCA (f) KI^{2}SCA

Fig. 2. Sample distributions of PIE dataset of 20 samples (5 different palms and 4 samples per palm) in the feature space. Here, the markers (* , ☆, ○, □, and ▽) with different colors stand for 5 different classes.

Statistical Significance

To statistically analyze the recognition results given in Table 2-4, we conduct a statistical test, i.e., Mcnemar's test (Draper, Yambor, and Beveridge 2006). This test can provide statistical significance between our proposed apporaches and other methods. Here, the Mcnemar's test uses a significance level of 0.05, that is, if the p-value is below 0.05, the performance difference between two compared methods is considered to be statistically significant. Table 9 shows the p-values between I^{2}SCA and other compared methods on Multiple Feature Dataset, Coil-20 Dataset and Multi-PIE dataset. And Table 10 shown the p-values between KI^{2}SCA and other compared methods on four datasets we utilized.

Table 9. P-values between I^{2}SCA and other compared methods on four datasets we utilized

Datasets	I^{2}SCA			
	MDCCA	MPA	MvDA	KGMA
MFD	1.32×10^{-8}	2.52×10^{-10}	3.29×10^{-10}	6.12×10^{-12}
Coil-20 Dataset	2.46×10^{-7}	2.47×10^{-6}	5.33×10^{-16}	7.58×10^{-24}
Multi-PIE Dataset	5.17×10^{-8}	1.59×10^{-8}	4.52×10^{-9}	2.42×10^{-10}
LFW Dataset	4.59×10^{-8}	1.38×10^{-8}	4.46×10^{-9}	2.91×10^{-10}

Table 10. P-values between KI^{2}SCA and other compared methods on four datasets we utilized

Datasets	KI^{2}SCA			
	MDCCA	MPA	MvDA	KGMA
MFD	3.41×10^{-9}	1.94×10^{-10}	5.47×10^{-10}	2.16×10^{-12}
Coil-20 Dataset	4.25×10^{-8}	8.47×10^{-6}	5.98×10^{-16}	5.78×10^{-24}
Multi-PIE Dataset	3.94×10^{-8}	2.95×10^{-8}	5.42×10^{-9}	4.84×10^{-10}
LFW Dataset	5.85×10^{-8}	8.32×10^{-8}	4.79×10^{-9}	9.03×10^{-10}

The two above tables show that the p-values of both I^{2}SCA and KI^{2}SCA are much less than 0.05, which demonstrated the statistical significance of our approaches.

Conclusion

In this paper, we propose a novel multi-view feature learning approach called intra-view and inter-view supervised correlation analysis (I^{2}SCA). It can fully explore the useful correlation information from both intra-view and inter-view. The objective function of I^{2}SCA can maximize the within-class correlation from both inter-view and intra-view, and simultaneously minimize the between-class correlation from both inter-view and intra-view. We provide a kernelized extension of I^{2}SCA, that is, kernel I^{2}SCA (KI^{2}SCA). The proposed approaches can obtain the analytical solutions without iterative calculation.

We employ the widely-used multiple feature dataset (MFD), COIL-20 dataset, Multi-PIE dataset and LFW dataset as the test data. As compared with several state-of-the-art multi-view supervised feature learning methods, our approaches achieve better recognition results. Besides, the demos of sample distributions in the feature space illustrate that the features learned by our approaches have better separability than that learned by other methods.

Acknowledgements

The work described in this paper was supported by the National Nature Science Foundation of China under Project Nos. 61231015, 61172173, 61272273 and 61233011, the Major Science and Technology Innovation Plan of Hubei Province (No. 2013AAA020), the Guangdong-Hongkong Key Domain Breakthrough Project of China (No. 2012A090200007), 333 Engineering of Jiangsu Province under Project No. BRA2011175.

References

Ahonen, T.; Hadid, A.; and Pietikainen, M. 2006. Face description with local binary patterns: application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12): 2037-2041.

Bach, F. R.; Jordan, M. I. 2002. Kernel independent component analysis. *Journal of Machine Learning Research,* 3: 1-48.

Belhumeur, P. N.; Hespanha, J. P.; and Kriegman, D. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7): 711-720.

Cai, D.; He. X.; Han, J.; and Zhang, H. J. 2006. Orthogonal laplacianfaces for face recognition. *IEEE Transactions on Image Processing*, 15(11): 3608-3614.

Chen, Q. N.; and Sun, S. L. 2009. Hierarchical multi-view fisher discriminant analysis. *Neural Information Processing*, 289-298.

Diethe, T; Hardoon, D. R.; and Shawe-Taylor, J. 2010. Constructing nonlinear discriminants from multiple data view. *Machine Learning and Knowledge Discovery in Databases*, 328-343.

Diethe, T.; Hardoon, D. R.; and Shawe-Taylor, J. 2008. Multiview fisher discriminant analysis. *NIPS workshop on learning from multiple sources*.

Draper, B. A.; Yambor, W. S.; and Beveridge, J. R. 2002. Analyzing PCA-based face recognition algorithms: eigenvector selection and distance measures. *Empirical Evaluation Methods in Computer Vision*, 1-15.

Fukunaga, K.; and Koontz, W. L. G. 1970. Application of the Karhunen-Loeve expansion to feature selection and ordering. *IEEE Transactions on Computers*, 19(4): 311-318.

Gao, L.; Qi, L.; Chen, E. Q.; and Guan, L. 2012. Discriminative multiple canonical correlation analysis for multi-feature information fusion. *IEEE International Symposium on Multimedia*, 36-43.

Grigorescu, S. E.; Petkov, N.; and Kruizinga, P. 2002. Comparison of texture features based on Gabor filters. *IEEE Transactions on Image Processing*, 11(10): 1160-1167.

Guo, Y. H. 2013. Convex subspace representation learning from multi-view data. *AAAI Conference on Artificial Intelligence*, 387-393.

Hardoon, D. R.; Szedmak. S.; and Shawe-Taylor, J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12): 2639-2664.

Huang, G. B.; Mattar, M. A.; Lee, H.; and Learned-Miller, E. 2012. Learning to align from scratch, *Neural Information Processing Systems,* 773-781.

Kan, M.; Shan, S. G.; Zhang, H. H.; Lao, S. H.; and Chen, X. L. 2012. Multi-view discriminant analysis. *European Conference on Computer Vision*: 808-821.

Kim, T. K.; Kittler, J.; and Cipolla, R. 2007. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6): 1005-1018.

Kumar, A.; and Daume III, H. 2011a. A co-training approach for multi-view spectral clustering. *International Conference on Machine Learning*, 393-400.

Kumar, A.; Rai, P.; and Daume III, H. 2011. Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing System*, 1413-1421.

Lai, P. L.; and Fyfe, C. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal on Neural System*, 10(5): 365-377.

Leurgans, S. E.; Moyeed, R. A.; and Silverman, B. W. 1993. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B (Methodological),* 55(3): 725-740.

Memisevic, R. 2012. On multi-view feature learning. *International Conference on Machine Learning,* 161-168.

Murase, H.; and Nayar, S. K. 1995. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision* 14(1): 5-24.

Nielsen, A. A. 2002. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Transactions on Image Processing*, 11(3): 293-305.

Rupnik, J.; and Shawe-Taylor, J. 2010. Multi-View Canonical correlation analysis. *International Conference on Data Mining and Data Warehouses*, 1-4.

Sharma, A.; Kumar, A.; Daume, H.; and Jacobs, D. W. 2012. Generalized multiview analysis: A discriminative latent space. *IEEE Conference on Computer Vision and Pattern Recognition*, 2160-2167.

Su, Y.; Fu, Y.; Gao, X.; and Tian Q. 2012. Discriminant learning through multiple principal angles for visual recognition. *IEEE Transactions on Image Processing*, 21(3): 1381-1390.

Sun, S. 2013. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8): 2031-2038.

Sun, S. L.; and Chao, G. Q. 2013. Multi-view maximum entropy discrimination. *International Joint Conference Artificial Intelligence*, 1706-1712.

Sun, T. K.; Chen, S. C.; Jin, Z.; and Yang J. Y. 2007. Kernelized discriminative canonical correlation analysis. *International Conference on Wavelet Analysis and Pattern Recognition*, 3: 1283-1287.

Wang, D.; Li, T.; and Ogihara, M. 2012. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. *AAAI Conference on Artificial Intelligence*, 683-689.

Wang, H.; Nie, F.; and Huang, H. 2013. Multi-view clustering and feature learning via structured sparsity. *International Conference on Machine Learning*, 28: 352-360.

Xu, C.; Tao, D. C.; and Xu, C. 2013. A survey on multi-view learning. CoRR abs/1304.5634.

Yuan, Y. H.; Sun, Q. S.; Zhou, Q.; and Xia. D. S. 2010. A novel multiset integrated canonical correlation analysis framework and its application in feature fusion. *Pattern Recognition*, 44(5): 1031-1040.

Yang, P.; and Gao, W. 2013. Multi-view discriminant transfer learning. *International Joint Conference Artificial Intelligence*, 1848-1854.