A Supervised Combined Feature Extraction Method for Recognition

Tingkai Sun, Songcan Chen, Jingyu Yang and Peifei Shi

Abstract

Multimodal recognition is an emerging technique to overcome the non-robustness of the unimodal recognition in real applications. Canonical correlation analysis (CCA) has been employed as a powerful tool for feature fusion in the realization of such multimodal system. However, CCA is the unsupervised feature extraction and it does not utilize the class information of the samples, resulting in the constraint of the recognition performance. In this paper, the class information is incorporated into the framework of CCA for combined feature extraction, and a novel supervised method of combined feature extraction for multimodal recognition, called discriminative canonical correlation analysis (DCCA), is proposed. The experiments of text categorization, face recognition and handwritten digit recognition show that DCCA outperforms some related methods of both unimodal recognition and multimodal recognition.

1. Introduction

Most of the state-of-the-art pattern recognition methods are unimodal, e.g., audio-only speech recognition, and some commercially available unimodal recognition systems work well in reasonably good conditions. However, the performance of such systems may unpredictably deteriorate under some noisy conditions. When the non-robustness of unimodal recognition is noticed in real applications, as a result, multimodal recognition emerges and has been gained more and more attentions [1,2]. Here the term *modality* in the context of both unimodal recognition and multimodal recognition originally stands for the source of information or sensory channel [15].

For multimodal recognition, it is a critical issue to effectively utilize the information stemming from different sources to improve the recognition performance. An effective solution to this problem is information fusion, which is defined as the synergistic use of information from diverse sources to improve overall understanding of a phenomenon or the recognition of an object [3]. By the proper approach, information fusion can make use of the complementary information to emphasize the useful information for the problem at hand, meanwhile it also can reduce the uncertainties to some extent [3]. Pan et al [4] studied the multisensory information fusion in the Bayesian inference framework, that is, given *n* pairwise samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ coming from *c* classes $\{\omega_i\}_{i=1}^{c}$, a new pairwise sample ($\boldsymbol{x}$,$\boldsymbol{y}$) should be classified according to its *a posteriori* conditional probability, which is computed by the Bayes' rule

$$P(\omega_i \mid \boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_i) P(\omega_i)}{\sum_{j=1}^{c} p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_j) P(\omega_j)} \quad (1)$$

Since the denominator is common and the *priori* probability $P(\omega_i)$ is easily to be estimated, so the task turns to be how to effectively estimate the *priori* joint probability distribution function (pdf) $p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_i)$ for the recognition task. However, in the case of high dimensional and highly-coupled signals, the direct estimating pdf $p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_i)$ is a hard task. An alternative approach to this problem is 1) mapping the high dimensional signals to low-dimensional subspace by a linear mapping, 2) in the low-dimensional subspace, it easy to estimate the pdf, and 3) turning back to the high dimensional and obtaining the estimated pdf $p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_i)$, which is satisfied with the maximum entropy constraint. Pan et al [4] found that when the data distribution is Gaussian, the optimal linear mapping for the estimating of $p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_i)$ exactly corresponds to a CCA problem using the samples in class ω_i! So in this sense, the works in [4] laid the mathematical foundation of CCA for feature fusion. Unfortunately, for some applications, in which *c* is large, the separate CCAs based on the samples in each class are fussy computational tasks; what is worse, when the number of the samples $\boldsymbol{x}$($\boldsymbol{y}$) in ω_i is small, the estimated $p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_i)$ based on to too few samples in ω_i may be imprecise. Alternatively, Sun et

al. [5] employ CCA to extract features from *all the samples* $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, and directly fuse the extracted features for recognition. The advantage of doing so [5] is that it can obtain the global solution at once rather than separately estimating the class pdf $p(\boldsymbol{x}, \boldsymbol{y} \mid \omega_i)$, however, CCA is unsupervised feature extraction method, and in doing so the class label information is not exploited, resulting in the limitation of the recognition performance.

To remedy this shortcoming of CCA, in this paper, the class information is incorporated into the framework of CCA for combined feature extraction, and a novel method of combined feature extraction for multimodal recognition, called discriminative canonical correlation analysis (DCCA), is proposed. The experiments of text categorization, face recognition and handwritten digit recognition show that DCCA outperforms some related methods of both unimodal recognition and multimodal recognition.

2. Review of canonical correlation analysis

Given *n* pairs of mean-normalized *pairwise* samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n \in \Re^p \times \Re^q$ coming from *c* classes, CCA aims to find pairs of projection directions $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ that maximize the correlation between the random variable $x = \boldsymbol{w}_x^T \boldsymbol{x}_i$ and $y = \boldsymbol{w}_y^T \boldsymbol{y}_i$, i =1,…*n*. More formally, CCA can be described as the following problem:

$$
\begin{aligned}
(\boldsymbol{w}_x, \boldsymbol{w}_y) &= \arg\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{E[xy]}{\sqrt{\mathrm{var}[x]\,\mathrm{var}[y]}} \\
&= \arg\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\sum_{i=1}^n \boldsymbol{w}_x^T \boldsymbol{x}_i \boldsymbol{y}_i^T \boldsymbol{w}_y}{\sqrt{\sum_{i=1}^n \boldsymbol{w}_x^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{w}_x} \cdot \sqrt{\sum_{i=1}^n \boldsymbol{w}_y^T \boldsymbol{y}_i \boldsymbol{y}_i^T \boldsymbol{w}_y}} \\
&= \arg\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T \boldsymbol{X}\boldsymbol{Y}^T \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{w}_x} \cdot \sqrt{\boldsymbol{w}_y^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{w}_y}}
\end{aligned} \tag{2}
$$

Solving this optimization problem, it is easy to obtain the following equation:

$$
\begin{pmatrix} & \boldsymbol{X}\boldsymbol{Y}^T \\ \boldsymbol{Y}\boldsymbol{X}^T & \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{X}\boldsymbol{X}^T & \\ & \boldsymbol{Y}\boldsymbol{Y}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} \tag{3}
$$

where the generalized eigenvalue λ is exactly the correlation between the random variable *x* and *y*. Suppose there are at most *r* non-zero generalized eigenvalues λ corresponding to (3), once the vector pairs $(\boldsymbol{w}_{xi}, \boldsymbol{w}_{yi})$, *i*=1,…,*d*, corresponding to the first *d* largest generalized eigenvalues are obtained, let $\boldsymbol{W}_x$=[$\boldsymbol{w}_{x1}$,..., $\boldsymbol{w}_{xd}$], $\boldsymbol{W}_y$=[$\boldsymbol{w}_{y1}$,..., $\boldsymbol{w}_{yd}$], the combined feature extraction and the feature fusion can be performed in the following ways [5]:

I) $$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{W}_x & 0 \\ 0 & \boldsymbol{W}_y \end{pmatrix}^T \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \tag{4}$$

II) $$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{W}_x \\ \boldsymbol{W}_y \end{pmatrix}^T \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \tag{5}$$

which hereafter are called feature fusion strategy I and II (FFS-I and -II), respectively. Sun et al [5] studied FFS-I and –II using CCA and apply them to pattern recognition.

3. Discriminative canonical correlation analysis

Using CCA, the correlated information $\boldsymbol{w}_x^T \boldsymbol{x}_i$ and $\boldsymbol{w}_y^T \boldsymbol{y}_i$,*i*=1,…,*n*, are extracted and fused for recognition [5]. However, the class information of the samples is not exploited, resulting in the limitation of the recognition performance of CCA. In fact, CCA was originally proposed for modeling [6] rather than recognition, and correlation λ indicates the predictability between $\boldsymbol{w}_x^T \boldsymbol{x}_i$ and $\boldsymbol{w}_y^T \boldsymbol{y}_i$,*i*=1,…,*n*. In fact, CCA was more applied to modeling and prediction, such as image retrieval [7] and parameter estimation [8]. If the features are to be extracted for recognition, the class information of the samples should be exploited to extracted more discriminative features. To this end, we incorporate the class information in the framework of CCA for combined feature extraction, and propose a novel method of combined feature extraction for multimodal recognition, called discriminative canonical correlation analysis (DCCA), which are detailed as follows.

Given *n* pairs of mean-normalized pairwise samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n \in \Re^p \times \Re^q$ coming from *c* classes, DCCA can be formulated as the following optimization problem:

$$
\begin{aligned}
&\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \left(\boldsymbol{w}_x^T \boldsymbol{C}_w \boldsymbol{w}_y - \eta \cdot \boldsymbol{w}_x^T \boldsymbol{C}_b \boldsymbol{w}_y \right) \\
&\text{s.t. } \boldsymbol{w}_x^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{w}_x = 1,\ \boldsymbol{w}_y^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{w}_y = 1
\end{aligned} \tag{6}
$$

where the matrices $\boldsymbol{C}_w$ and $\boldsymbol{C}_b$ are constructed to measure the within-class similarity and the between-

class correlations similarity, respectively (detailed definition are given below), and η >0 a tunable parameter that indicates the relative significance of the within-class similarity $\boldsymbol{w}_x^T \boldsymbol{C}_w \boldsymbol{w}_y$ versus the between-class similarity $\boldsymbol{w}_x^T \boldsymbol{C}_b \boldsymbol{w}_y$. Let

$$\boldsymbol{X} = \left[\boldsymbol{x}_1^{(1)}, ..., \boldsymbol{x}_{n_1}^{(1)}, ..., \boldsymbol{x}_1^{(c)}, ..., \boldsymbol{x}_{n_c}^{(c)} \right] \quad (7)$$

$$\boldsymbol{Y} = \left[\boldsymbol{y}_1^{(1)}, ..., \boldsymbol{y}_{n_1}^{(1)}, ..., \boldsymbol{y}_1^{(c)}, ..., \boldsymbol{y}_{n_c}^{(c)} \right] \quad (8)$$

$$\boldsymbol{e}_{n_i} = [\underbrace{0,...0}_{\sum_{j=1}^{i-1} n_j}, \underbrace{1,...1}_{n_i}, \underbrace{0,...0}_{n-\sum_{j=1}^{i} n_j}]^T \in R^n \quad (9)$$

$$\boldsymbol{1}_n = [1,...1]^T \in R^n \quad (10)$$

where $\boldsymbol{x}_j^{(i)}$ denotes the *j*th sample in the *i*th class, so does $\boldsymbol{y}_j^{(i)}$, and n_i denotes the number of samples of $\boldsymbol{x}_j^{(i)}$ or $\boldsymbol{y}_j^{(i)}$ in the *i*th class. The matrix $\boldsymbol{C}_w$ is defined as

$$\begin{aligned} \boldsymbol{C}_w &= \sum_{i=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} \boldsymbol{x}_k^{(i)} \boldsymbol{y}_l^{(i)T} \\ &= \sum_{i=1}^{c} \left(\boldsymbol{X} \boldsymbol{e}_{n_i} \right) \left(\boldsymbol{Y} \boldsymbol{e}_{n_i} \right)^T \\ &= \boldsymbol{X} \boldsymbol{A} \boldsymbol{Y}^T \end{aligned} \quad (11)$$

Where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{1}_{n_1 \times n_1} & & & \\ & \ddots & & \\ & & \boldsymbol{1}_{n_i \times n_i} & \\ & & & \ddots \\ & & & & \boldsymbol{1}_{n_c \times n_c} \end{bmatrix} \in \Re^{n \times n} \quad (12)$$

is a symmetric, positive semidefinite, blocked diagonal matrix, and rank($\boldsymbol{A}$)=*c*.

On the other hand, the matrix $\boldsymbol{C}_b$ is defined as

$$\begin{aligned} \boldsymbol{C}_b &= \sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \boldsymbol{x}_k^{(i)} \boldsymbol{y}_l^{(j)T} \\ &= \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \boldsymbol{x}_k^{(i)} \boldsymbol{y}_l^{(j)T} - \sum_{i=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} \boldsymbol{x}_k^{(i)} \boldsymbol{y}_l^{(i)T} \\ &= \left(\boldsymbol{X} \boldsymbol{1}_n \right) \left(\boldsymbol{Y} \boldsymbol{1}_n \right)^T - \boldsymbol{X} \boldsymbol{A} \boldsymbol{Y}^T \\ &= -\boldsymbol{X} \boldsymbol{A} \boldsymbol{Y}^T \end{aligned} \quad (13)$$

The last "=" holds due to the fact that the samples have been mean-normalized so that both $\boldsymbol{X1}_n=\boldsymbol{0}$ and $\boldsymbol{Y1}_n=\boldsymbol{0}$ hold. Comparing (13) with (11), the difference between $\boldsymbol{C}_w$ and $\boldsymbol{C}_b$ is only one negative sign, so the objective of (6) turns to be $(1+\eta)\boldsymbol{w}_x^T \boldsymbol{C}_w \boldsymbol{w}_y$, and this optimization problem is independent of the parameter η, so η can be omitted. Thus DCCA can be formulated as:

$$\begin{aligned} &\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \boldsymbol{w}_x^T \boldsymbol{X} \boldsymbol{A} \boldsymbol{Y}^T \boldsymbol{w}_y \\ &\text{s.t.} \quad \boldsymbol{w}_x^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{w}_x = 1, \boldsymbol{w}_y^T \boldsymbol{Y} \boldsymbol{Y}^T \boldsymbol{w}_y = 1 \end{aligned} \quad (14)$$

Using the Lagrangian multiplier technique, it is easy to obtain the corresponding primary equation of DCCA as follows:

$$\begin{pmatrix} & \boldsymbol{X} \boldsymbol{A} \boldsymbol{Y}^T \\ \boldsymbol{Y} \boldsymbol{A} \boldsymbol{X}^T & \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{X} \boldsymbol{X}^T & \\ & \boldsymbol{Y} \boldsymbol{Y}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} \quad (15)$$

Once the vector pairs $(\boldsymbol{w}_{xi}, \boldsymbol{w}_{yi})$, *i*=1,…,*d*, corresponding to the first *d* largest generalized eigenvalues are obtained, let $\boldsymbol{W}_x$=[$\boldsymbol{w}_{x1}, ..., \boldsymbol{w}_{xd}$], $\boldsymbol{W}_y$=[$\boldsymbol{w}_{y1}, ..., \boldsymbol{w}_{yd}$], the combined feature extraction and the feature fusion can be performed according to FFS-I and –II, respectively, where *d* satisfies the constraints $d \leq \min(p,q)$ and $d \leq c$. Based on the extracted features in this way, any classifier, e.g., the nearest-neighbor classifier, can be used for recognition.

For the extracted features using DCCA, the following conclusion holds:

Theorem 1. Let $\boldsymbol{\xi}_i = \boldsymbol{w}_{xi}^T \boldsymbol{X}, \boldsymbol{\zeta}_i = \boldsymbol{w}_{yi}^T \boldsymbol{Y}$ denote the extracted features using DCCA, they satisfy that: $< \boldsymbol{\xi}_i, \boldsymbol{\xi}_j >= \delta_{ij}$, $< \boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j >= \delta_{ij}$, where δ_{ij} denotes the Kronecker symbol, i.e., $\delta_{ij} = 1$ if *i*=*j*, and 0 otherwise. Besides, $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are matrix $\boldsymbol{A}$-orthonormal, i.e., $\boldsymbol{\xi}_i \boldsymbol{A} \boldsymbol{\zeta}_j^T = \lambda_i \delta_{ij}$.

Proof: let $\boldsymbol{C}_x = \boldsymbol{X} \boldsymbol{X}^T$, $\boldsymbol{C}_y = \boldsymbol{Y} \boldsymbol{Y}^T$, and the main equation of DCCA (15) can be decoupled as:

$$\begin{cases} \boldsymbol{C}_w \boldsymbol{C}_y^{-1} \boldsymbol{C}_w^T \boldsymbol{w}_x = \lambda^2 \boldsymbol{C}_x \boldsymbol{w}_x \\ \boldsymbol{C}_w^T \boldsymbol{C}_x^{-1} \boldsymbol{C}_w \boldsymbol{w}_y = \lambda^2 \boldsymbol{C}_y \boldsymbol{w}_y \end{cases} \quad (16)$$

Let $\boldsymbol{H} = \boldsymbol{C}_x^{-\frac{1}{2}} \boldsymbol{C}_w \boldsymbol{C}_y^{-\frac{1}{2}} \in R^{p \times q}$, $\boldsymbol{u} = \boldsymbol{C}_x^{\frac{1}{2}} \boldsymbol{w}_x$, $\boldsymbol{v} = \boldsymbol{C}_y^{\frac{1}{2}} \boldsymbol{w}_y$, and rank($\boldsymbol{H}$)=*r*, Eqs. (16) turn to be

$$\begin{cases} \boldsymbol{H} \boldsymbol{H}^T \boldsymbol{u} = \lambda^2 \boldsymbol{u} \\ \boldsymbol{H}^T \boldsymbol{H} \boldsymbol{v} = \lambda^2 \boldsymbol{v} \end{cases} \quad (17)$$

which exactly corresponds the singular value decomposition (SVD) of matrix $\boldsymbol{H}$. Let the SVD of $\boldsymbol{H}$ be $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^T$, where $\boldsymbol{u}_i, \boldsymbol{v}_i$ are the i-th column vector of the orthonormal matrix $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Thus $\boldsymbol{w}_{xi} = \boldsymbol{C}_x^{-\frac{1}{2}} \boldsymbol{u}_i$ and $\boldsymbol{w}_{yi} = \boldsymbol{C}_y^{-\frac{1}{2}} \boldsymbol{v}_i$, i=1,…,d. So

$< \boldsymbol{\xi}_i, \boldsymbol{\xi}_j >= \boldsymbol{w}_{xi}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{w}_{xj} = \boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$,

$< \boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j >= \boldsymbol{w}_{yi}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{w}_{yj} = \boldsymbol{v}_i^T \boldsymbol{v}_j = \delta_{ij}$, and

$\boldsymbol{\xi}_i \boldsymbol{A} \boldsymbol{\zeta}_j^T = \boldsymbol{w}_{xi}^T \boldsymbol{X}\boldsymbol{A}\boldsymbol{Y}^T \boldsymbol{w}_{yj} = \boldsymbol{u}_i^T \boldsymbol{H} \boldsymbol{v}_j = \lambda_i \delta_{ij}$. □

From ***Theorem*** 1, we can know that for DCCA, the features extracted in the same modality (i.e., X or Y sample space) are statistically uncorrelated each other. That's to say, $< \boldsymbol{\xi}_i, \boldsymbol{\xi}_j >= 0$ and $< \boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j >= 0$ if $j \neq i$. So DCCA eliminates the redundant information in the same modality. According to the theory of the statistical pattern recognition, the features with less correlation, or without correlation, will benefit to the subsequent recognition.

4. Experiments and analysis

In this section, we will evaluate the ability of DCCA to combined feature extraction for recognition. To this end, firstly an artificial problem is studied to test the validity of DCCA, then the experiments of text categorization, face recognition and handwritten digit recognition are performed to evaluate the recognition performance of DCCA by comparison with some related recognition methods, i.e., CCA and partial least squares (PLS), which is also used for combined feature extraction and recognition [9].

4.1. Artificial problem

Consider a binary class problem. Let $\boldsymbol{X}$=[$\boldsymbol{X}_1$, $\boldsymbol{X}_2$], $\boldsymbol{Y}$=[$\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$], where $\boldsymbol{X}_i$, $\boldsymbol{Y}_i$, i=1,2, denote the samples of the i-th class coming from two data sets, respectively, among which the following relationship holds, i.e., $\boldsymbol{y}_i = \boldsymbol{W}^T \boldsymbol{x}_i + \boldsymbol{b} + \varepsilon_i$, where ε_i is the imposed Gaussian noise. Fig.1(a) shows the data distribution, and Fig.1(b), 1(c) and 1(d) in turn shows the extracted features ($\boldsymbol{w}_x^T \boldsymbol{x}_i$, $\boldsymbol{w}_y^T \boldsymbol{y}_i$) using CCA, PLS and DCCA.

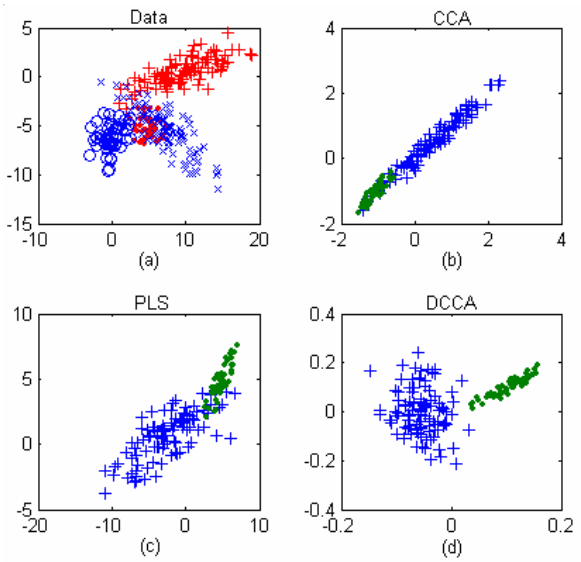


Fig.1. Artificial problem. (a) shows the sample distributions, where the symbols + ·× and ○ in turn denote the samples of class 1 and 2 in $\boldsymbol{X}$ set, of class 1 and 2 in $\boldsymbol{Y}$ set, respectively. (b)-(d) show the first pair of features extracted by CCA (b), by PLS (c) and by DCCA (d), respectively, where the horizontal coordinate denotes x component and vertical coordinate y component, and the symbols + and · denote class 1 and 2, respectively.

In Fig.1(b), CCA reveals the proximately linear correlation between the canonical components, yet the overlapping appears between classes, and this may results in misclassifications. In Fig.1(c), the overlapping between classes also appears to some degree for the features extracted by PLS. In contrast, in Fig.1(d), the samples belonging to the two classes are well separated from each other, and this indicates that: 1) both CCA and PLS are more suitable for modeling the linear model rather than recognition, 2) the features extracted by DCCA are more suitable for recognition.

4.2 Experiment of text categorization

The WebKB hypertext dataset (available at http://www.cs.cmu.edu/afs/cs/project/theo-11/www/-wwkb/) is employed in the experiment of text categorization. WebKB consists of 1051 web pages collected from web sites of computer science departments of four famous universities in U.S. The 1051 pages were manually classified into the categories of *course* (230 pages) and *non-course* (821 pages). Each page corresponds to two views, i.e., *fulltext* (the text on the web pages, referred to as sample in $\boldsymbol{X}$ set) and *inlinks* (the anchor text on the hyperlinks pointing to the page, referred to as sample in $\boldsymbol{Y}$ set). The original

hypertext documents are pre-processed by skipping html tokens, toss stop-words and stemming, resulting in 1854-dimensional vector for each fulltext document and a 106-dimensional vector for each inlinks document. The entries of these vectors denote the term-frequencies in the corresponding document. 120 and 400 pages in class *course* and class *non-course*, respectively, are randomly selected for training, and the remaining pages are used for test. Thus the total sizes of training set and test set are 520 and 531, respectively, and then the term frequency / inverse document frequency (TF-IDF) [10] vector corresponding to each document is computed. In this experiment, the proposed DCCA and other methods of combined feature extraction, such as CCA and partial least squares (PLS), are compared. Further, some frequently used text classifiers, such us Naïve Bayes[11], k-nearest neighbor [12] (k-NN), class mean vector [10] (CMV), are also exployed for comparision. The random experiment are repeated 100 times, and the average results are reported in Table 1 and 2, respectively.

Table 1. The recognition accuracies of some unimodal classifiers

Method	Recognition accuracy	
	fulltext	inlinks
Naïve Bayes	0.9083	0.8753
k-NN	**0.9448**	**0.9467**
CMV	0.9098	0.8881

Table 2. The recognition accuracies of some multimodal classifiers

Method	Recognition accuracy	
	Ratio-1	Ratio-2
DCCA	**0.9574**	**0.9522**
CCA	0.9213	0.9235
PLS	0.9203	0.9215

* Ratio-1 and -2 correspond to FFS-I and –II, respectively.

From Table 1, *k*-NN method outperforms Naïve Bayes and CMV, and From Table 2, we can see that DCCA outperforms not only CCA and PLS, but also all the related unimodal classifiers in Table 1.

4.3 Experiment of face recognition

The well-known ORL face dataset contains 400 human face images of 40 persons, each providing 10 different images, taken at different times and with varying facial expressions (smile/no smile, open/closed eyes), facial details (with or without glasses) and poses. The images are in upright, frontal position with tolerance for some tilting and rotation of up to 20 degree. All images are grayscale with 256 levels and normalized to 112×92 pixels. In each experiment, 5 images of each person are randomly selected for training, and the remaining 5 images for test. The random experiments are repeated 10 times. In this experiment, the famous Eigenface [13] and Fisherface [14] methods are selected as benchmark methods for comparison. In addition, for DCCA, CCA and PLS, the Daubechies wavelet transform is performed on images, and the resultant low-frequent images are specified as another set of data. Fig.2 shows 5 images of one person and the corresponding 5 low-frequent images.

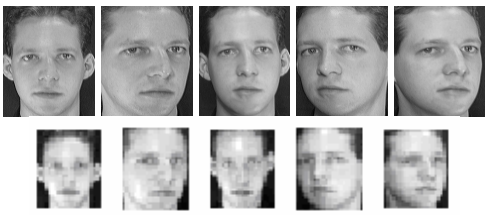


Fig.2 face images (upper row) and the low-frequent images (bottom row) of ORL face dataset

Table 3 tabulates the recognition results on ORL dataset. From Table 3 we can find that DCCA outperforms not only Eigenface, Fisherface, but also CCA and PLS methods.

Table 3. The recognition results on ORL

Method	Recognition accuracy
Eigenface	0.9355
Fisherface	0.9065
DCCA	**0.9495**[1] / **0.9485**[2]
CCA	0.9011[1] / 0.9088[2]
PLS	0.9395[1] / 0.9405[2]

*superscript 1, 2 correspond to FFS-I and –II, respectively.

4.4 Experiment of handwritten digit recognition

Multiple Features database (available at http://www.ics.uci.edu/~mlearn/MLSummary.html) consists of features of handwritten numerals ('0'-'9', total 10 classes) extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2000 patterns) have been digitized in binary images of size 30×48. Digits are represented in terms of Fourier coefficients (76 dimensions, referred to as FOU,76), profile correlations (FAC,216), Karhunen-Love coefficients (KAR,64), pixel averages (PIX,240), Zernike moments (ZER,47) and morphological features (MOR,6), respectively.

In experiments, any two datasets of Multiple Features database are picked out to construct the $\boldsymbol{X}$ and $\boldsymbol{Y}$ set for CCA, PLS and DCCA methods, thus there are total C_6^2 =15 pairs of different dataset combinations. For each combination, 100 pairs of feature vectors per class are randomly selected for training, the remaining 1000 pairs for test. The random experiment is repeated 10 times. Table 4 and 5 (separately corresponding to FFS-I and FFS-II) tabulate the recognition results using CCA, PLS and DCCA. We can find that in most cases, DCCA outperforms CCA and PLS in terms of the recognition accuracy.

Table 4. The recognition results (using FFS-I) on Multiple Features database

$\boldsymbol{X}$	$\boldsymbol{Y}$	Recognition accuracy		
		DCCA	CCA	PLS
FAC	FOU	**0.9813**	0.8785	0.9394
FAC	KAR	**0.9789**	0.9598	0.9397
FAC	MOR	**0.9302**	0.7656	0.8789
FAC	PIX	**0.9752**	0.9476	0.9396
FAC	ZER	**0.9772**	0.8623	0.9570
FOU	KAR	0.9687	0.9195	**0.9698**
FOU	MOR	**0.8278**	0.7633	0.4389
FOU	PIX	0.9662	0.8431	**0.9756**
FOU	ZER	**0.8543**	0.8351	0.8119
KAR	MOR	**0.9253**	0.8158	0.6234
KAR	PIX	0.9497	0.9641	**0.9753**
KAR	ZER	**0.9638**	0.9211	0.8289
MOR	PIX	**0.9100**	0.7602	0.7078
MOR	ZER	**0.8097**	0.7452	0.7154
PIX	ZER	**0.9544**	0.8398	0.8401

Table 5. The recognition results (using FFS-II) on Multiple Features database

$\boldsymbol{X}$	$\boldsymbol{Y}$	Recognition accuracy		
		DCCA	CCA	PLS
FAC	FOU	**0.9560**	0.8673	0.9394
FAC	KAR	**0.9752**	0.9603	0.9410
FAC	MOR	**0.9077**	0.7596	0.8716
FAC	PIX	**0.9718**	0.9472	0.9433
FAC	ZER	**0.9589**	0.8542	0.9521
FOU	KAR	0.9393	0.8969	**0.9714**
FOU	MOR	**0.8089**	0.7567	0.4398
FOU	PIX	0.9373	0.8270	**0.9761**
FOU	ZER	**0.8367**	0.8239	0.8110
KAR	MOR	**0.8928**	0.7857	0.6314
KAR	PIX	0.9493	0.9643	**0.9751**
KAR	ZER	**0.9383**	0.9081	0.8245
MOR	PIX	**0.8799**	0.7263	0.7071
MOR	ZER	**0.7943**	0.7258	0.6983
PIX	ZER	**0.9310**	0.8232	0.8375

The proposed DCCA stems from the framework of the combined feature extraction using CCA, and it outperforms the latter in terms of the recognition accuracy. Let us analyze their difference that could benefit to the recognition. For CCA, the features to be fused is *pairwise* $\boldsymbol{w}_{xi}^T\boldsymbol{x}$ and $\boldsymbol{w}_{yi}^T\boldsymbol{y}$, and the correlation between these two random variate can be written as $\mathrm{corr}(\boldsymbol{w}_{xi}^T\boldsymbol{x}, \boldsymbol{w}_{yi}^T\boldsymbol{y}) = \lambda_i$ [5], if the correlation between them is too high (even be perfect correlation, i.e., $\lambda = 1$, in the extreme case), it makes no sense to fuse them that contain too much redundant information. For DCCA, what about the correlation $\mathrm{corr}(\boldsymbol{w}_{xi}^T\boldsymbol{x}, \boldsymbol{w}_{yi}^T\boldsymbol{y})$ will be? In fact, $\mathrm{corr}(\boldsymbol{w}_{xi}^T\boldsymbol{x}, \boldsymbol{w}_{yi}^T\boldsymbol{y}) = <\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i>$, From ***Theorem*** 1, we can only know that $\boldsymbol{\xi}_i \boldsymbol{A} \boldsymbol{\zeta}_i^T = \lambda_i$. To compare the correlation $\mathrm{corr}(\boldsymbol{w}_{xi}^T\boldsymbol{x}, \boldsymbol{w}_{yi}^T\boldsymbol{y})$ of DCCA with that of CCA, we numerically compute them in this experiment. For instance, in the first combination, FAC and FOU, the correlation between the pairwise features are computed and illustrated in Fig. 3. Note that in this case, there are total 76 pairs of features for CCA, and only 9 pairs for DCCA, respectively. From Fig.3, we can see that the correlation between the *i*th, *i*=1,…,9, pair of features of DCCA is less than the correlation between the *i*th, *i*=1,…,9, pair of features of CCA. However, the recognition performance of DCCA is better than that of CCA. In other words, the features extracted by DCCA are more discriminative than those extracted by CCA. The further analysis implies that for DCCA, the recognition accuracy increases monotonously with the number, *d*, of the feature pairs (see Fig.4a); and for CCA, the recognition accuracy first increases and then decreases with the increase of the feature pairs (see Fig.4b). In other words, for CCA, some features are harmful to the recognition task. Moreover, we analyze the other 14 combinations and find things are very similar. In fact, in this paper, we find that for DCCA, the recognition performance always changes monotonously with the number, *d*, of the feature pairs, and the best recognition result is reached when *d* is set

to *c*-1. This characteristic makes DCCA easy to use in recognition tasks.

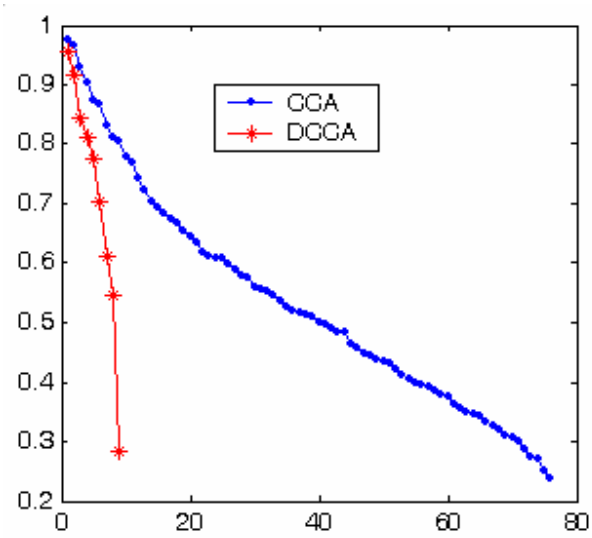
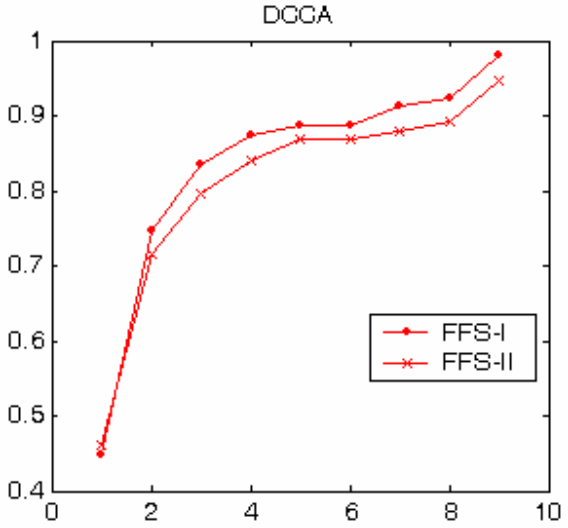
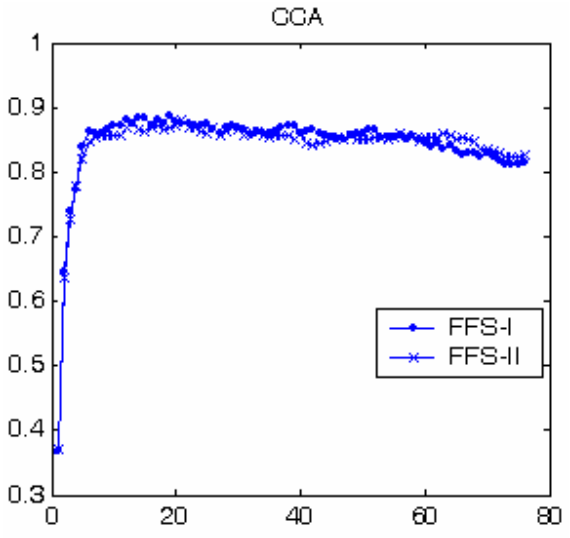


Fig.3 The correlation between the pairwise features in DCCA and CCA. The horizontal coordinate denotes the serial number of the pairwise features, and the vertical coordinate denotes the correlation.



(a)



(b)

Fig. 4 The change of the recognition accuracy (the vertical coordinate) w.r.t. the number of the feature pairs (the horizontal coordinate). (a) and (b) correspond to DCCA and CCA, respectively.

5. Conclusion and discussion

As an effective method of combined feature extraction, CCA can extract features between two sets of samples, and the features can be fused for the subsequent recognition. The related study verified the usefulness of CCA for recognition. However, the class information of the samples is not exploited by CCA, resulting in the limitation of the recognition performance. In this paper, we incorporate the class information into the framework of combined feature extraction and propose discriminative CCA (DCCA). The experimental results of the text categorization, face recognition and handwritten digit recognition show that DCCA outperform some related methods of both unimodal recognition and multimodal recognition. In addition, DCCA is a linear feature extraction method. Although the related work show that if it is kernelized using so-called *kernel trick*, better recognition performance can be achieved, yet the choice of the kernel and kernel parameter(s) are still troublesome, resulting in heavy computational tasks. In contrast, DCCA can be easily computed and applied to multimodal recognition problem. The next step of our aim is to generalize this method to the cases of more modalities.

Acknowledgement

References

[1] A. Ross, A. K. Jain. Multimodal biometrics: an overview. In: *Proc. of 12th European Signal Processing Conference* (EUSIPCO), Vienna, 2004, pp. 1221-1224.

[2] M.Sargin, E. Erzin, Y. Yemez, et al. Multimodal speaker identification using canonical correlation analysis. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, 1:I-613 - I-616.

[3] H. Pan. A Bayesian fusion approach and its application to integrating audio and visual signals in HCI. [ph.D. Dissertations], University of Illinois at Urbana-Champaign, 2001.

[4] Hao Pan, Z-P. Liang, Thomas S. Huang. Estimation of the joint probability of multisensory signals. *Pattern Recognition Letters*, 2001, 22(13):1431-1437.

[5] Q. Sun, S. Zeng, Y. Liu, P-A. Heng, D-S. Xia. A new method of feature fusion and its application in image recognition. *Pattern Recognition*, 2005, 38(12): 2437-2448.

[6] H. Hotelling, Relations between two sets of variates. *Biometrika*, 1936, 28:321–377.

[7] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation* 2004, 16: 2639-2664.

[8] T. Melzer, M. Reiter, H. Bischof, Appearance models based on kernel canonical correlation analysis, *Pattern Recognition*, 2003, 36(9):1961–1971.

[9] J. Wegelin. A survey of partial least squares (PLS) methods, with emphasis on the two-block case. *Technical Report* No.371, Department of Statistics, University of Washing, 2000.

[10] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 2002, 34(1):1-47.

[11] J. Rennie. Improving multi-class text classification with naive Bayes. [Master thesis], Massachusetts Institute of Technology, 2001.

[12] B. Dasarathy. Nearest neighbor (NN) norms: NN pattern classification techniques. Las Alamitos, California, IEEE Computer Society Press, 1990.

[13] M. Turk, A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 1991, 3(1): 71-86.

[14] P. N. Belhumeour, J. P. Hespanha, D. J. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, 19(7):711-720.

[15] C. Chibelushi, F. Deravi, J. Mason. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 2002, 4(1):23-37.