Multi-view Uncorrelated Linear Discriminant Analysis with Applications to Handwritten Digit Recognition

Mo Yang and Shiliang Sun

Abstract— **Learning from multiple feature sets, which is also called multi-view learning, is more robust than single view learning in many real applications. Canonical correlation analysis (CCA) is a popular technique to utilize information from multiple views. However, as an unsupervised method, it does not exploit the label information. In this paper, we propose an algorithm which combines uncorrelated linear discriminant analysis (ULDA) with CCA, named multi-view uncorrelated linear discriminant analysis (MULDA). Due to the successful application of ULDA, which seeks optimal discriminant features with minimum redundancy in the single view situation, it could be expected that the recognition performance would be enhanced. Experiments on handwritten digit data verify this expectation with results outperform other related methods.**

I. Introduction

LEARNING from multiple feature sets, which is also called multi-view learning, is a rapid growing direction in machine learning with well theoretical basis and great practical success [1]. This learning mechanism emerged recently, largely motivated by a phenomenon of real data, that is, the same object can be observed at different viewpoints to generate multiple distinct samples. For instance, web pages can be described by urls and caption text. In content-based web-image retrieval, an object can be simultaneously described by the text surrounding the image and the visual features from the image. Moreover, even natural different 'views' do not exist, manufactured splits of features can still improve the performance in various applications.

A critical issue of multi-view learning is to effectively utilize the information stemming from different sources to improve its application performance. An effective method is information fusion, which can be realized through obtaining a common space for multiple views. Feature extraction is a common way to obtain this kind of subspace.

Canonical correlation analysis (CCA), first proposed by Hotelling [2], is the most popular technique to extract features in multi-view learning. It works on paired datasets to find two linear transformations each for one view such that the two transformed variables are most correlated. Kernel CCA (KCCA) [3][4] is a nonlinear extension of CCA by means of the kernel trick, which corresponds to performing CCA in a kernel-induced feature space. Locality preserving CCA (LPCCA) [5] is another nonlinear extension of CCA, which was introduced to discover the local manifold structure of the data by forcing nearby points in the original feature space to be close in the transformed subspace as well. Other typical approaches include bilinear model (BLM) [6] and partial least squares (PLS) [7]. However, all the aforementioned methods are unsupervised methods, i.e., without using label information, which may limit them from recognition performance.

To overcome this deficiency, many supervised methods for multi-view learning have been proposed in the past years. Linear discriminant analysis (LDA) [8] is an effective supervised method in single-view learning. It seeks an optimal linear transformation that maps the data into a subspace, in which the within-class distance is minimized and simultaneously the between-class distance is maximized, thus achieving maximum discrimination. Following the way LDA preserves class structure, discriminant CCA (DCCA) [9] was proposed to exploit discriminant structure by taking within-class correlation terms into account. It maximizes the difference of within-class and between-class correlations across two views. Similarly, inspired by the great performance of DCCA, random correlation ensemble (RCE) [10] was proposed to incorporate discriminant information into CCA by using random cross-view correlation between within-class examples and construct a lot of feature extractors to do multi-view ensemble learning. In [11][12], multiview Fisher discriminant analysis (MFDA) was proposed to learn classifiers in different views by maximizing the consistency between the predicted labels of these classifiers. However, it can only be applicable in binary classification. To deal with this problem, Chen and Sun [13] used a hierarchical clustering approach to extend MFDA to a multi-class scenario, namely hierarchical MFDA (HMFDA). In [14], common discriminant feature extraction (CDFE) was proposed to learn two transforms simultaneously by incorporating both empirical discriminative power and local consistency.

As mentioned above, to guarantee the recognition performance, preserving discriminant structure is a very important property in feature extraction. In other words, in the scenario of multi-view learning, both inter-view and intra-view discriminant information mean a lot to ensure the classification ability in the common space. DCCA and RCE take cross-view correlation between within-class examples into account, which means inter-view class structure was preserved, while intra-view data structure is ignored yet. The other methods mentioned above have similar deficiency as well. Multi-view discriminant analysis (MvDA) [15] is an effective method to cope with this problem. It maximizes the difference between

Mo Yang and Shiliang Sun are with the Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai 200241, P. R. China (email: momo.yang12@gmail.com, slsun@cs.ecnu.edu.cn).

This work is supported by the National Natural Science Foundation of China under Projects 61370175 and 61075005, and Shanghai Knowledge Service Platform Project (No. ZF1213).

the within-class variation and the between-class variation which are calculated from the samples from all views. It uses the same way to represent inter-view correlation and intra-view correlation, which can be cast as a natural extension of LDA with all the transformed feature sets (e.g. different views) regarded as a large data set.

In this paper, we propose a new approach called multi-view uncorrelated linear discriminant analysis (MULDA), which extracts mutually uncorrelated features in each view and computes transformations of each view to project them into a common subspace. Inspired by the effectiveness of CCA and LDA, we formulate our objective function with a simple and natural combination of these two methods. Additionally, because of the fact that the feature vectors extracted by the uncorrelated LDA (ULDA) [16][17][18][19] could contain minimum redundancy and the successful application of ULDA in various applications in the past years, we extend the LDA part to ULDA with a uncorrelated constraint added into our objective function. Similar to [16], it can be solved with a sequence of generalized eigenvalue problems.

The remainder of this paper is organized as follows. Section II gives a brief review of some related work. The formulations and solutions of the proposed MULDA are presented in Section III. Section IV performs handwritten digit recognition experiments on the multiple features data set. Finally, we conclude this paper and discuss some future works in Section V.

II. Background

In this section, we give a brief review of CCA, LDA and ULDA.

A. Canonical Correlation Analysis

Canonical correlation analysis was first proposed by Hotelling [2] to find a common space for two views such that the correlations between these transformed feature sets are maximized.

Given a data set with two views $\{(x_1, y_1), ..., (x_n, y_n)\}$, and $X = [x_1, ..., x_n], Y = [y_1, ..., y_n]$. CCA seeks to find two projection directions w_x and w_y, one for each view, to maximize the following linear correlation coefficient:

$$\frac{cov\left(w_x^T X, w_y^T Y\right)}{\sqrt{var\left(w_x^T X\right) var\left(w_y^T Y\right)}} = \frac{w_x^T C_{xy} w_y}{\sqrt{\left(w_x^T C_{xx} w_x\right)\left(w_y^T C_{yy} w_y\right)}}, \tag{1}$$

where covariance matrix C_{xy}, C_{xx} and C_{yy} are defined as

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x\right)\left(y_i - m_y\right)^T, \tag{2}$$

$$C_{xx} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x\right)\left(x_i - m_x\right)^T, \tag{3}$$

$$C_{yy} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - m_y\right)\left(y_i - m_y\right)^T, \tag{4}$$

with m_x and m_y being the means from the two views, respectively,

$$m_x = \frac{1}{n}\sum_{i=1}^{n} x_i,\ m_y = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{5}$$

Since w_x, w_y are scale-independent, (1) is equivalent to

$$\begin{array}{ll} \max_{w_x, w_y} & w_x^T C_{xy} w_y \\ \text{s.t.} & w_x^T C_{xx} w_x = 1,\ w_y^T C_{yy} w_y = 1. \end{array} \tag{6}$$

Applying Lagrangian multiplier method on (6), the optimization problem of CCA can be solved by a generalized eigenvalue problem as follows:

$$\begin{bmatrix} & C_{xy} \\ C_{yx} & \end{bmatrix}\begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}\begin{bmatrix} w_x \\ w_y \end{bmatrix}. \tag{7}$$

The generalized eigenvalue λ reflects the degree of correlation between projections. Suppose retaining d pairs of projection vectors (w_x, w_y) corresponding to the largest eigenvalues, the transformations of each view to a common space will be $W_x = [w_{x1}, ..., w_{xd}], W_y = [w_{y1}, ..., w_{yd}]$.

The classical CCA can only exploit the linear relationships between feature sets. When dealing with the nonlinear problem, KCCA [3] would be effective and LPCCA [5] is an alternative option.

B. Linear Discriminant Analysis

Linear discriminant analysis is a powerful technique for dimensional reduction, which was first proposed in [8]. It aims to find an optimal transformation that maps the data into a lower-dimensional space in which the within-class distance is minimized and the between-class distance is maximized simultaneously, thus achieving maximum discrimination.

Given a data matrix $X \in R^{m\times n}$ with each column corresponding to a data point. Assuming $X = [x_1, x_2, ..., x_n] = [X_1, X_2, ..., X_k]$, where $x_j \in R^m\ (1 \le j \le n)$ represents a data point, n is the sample size, k is the number of classes and $X_i \in R^{m\times n_i}$ denotes the subset of all the samples in class i with n_i being the number of data in this subset. So we have $\sum_{i=1}^{k} n_i = n$. Classical LDA computes a linear transformation $G \in R^{m\times l}$ that maps each column x_i of X in the m-dimensional space to a vector q_i in the l-dimensional space:

$$G : x_i \in R^m \rightarrow q_i = G^T x_i \in R^l\ (l \le m). \tag{8}$$

In LDA, three scatter matrices, called $with\text{-}class$, $between\text{-}class$ and $total$ scatter matrices are defined as follows:

$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in X_i}\left(x - m^{(i)}\right)\left(x - m^{(i)}\right)^T, \tag{9}$$

$$S_b = \frac{1}{n}\sum_{i=1}^{k} n_i\left(m^{(i)} - m\right)\left(m^{(i)} - m\right)^T, \tag{10}$$

$$S_t = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - m\right)\left(x_j - m\right)^T. \tag{11}$$

Based on these scatter matrices, the Fisher criterion function can be defined as:

$$F(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi}, \tag{12}$$

and an alternative criterion for classical LDA is:

$$F(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_t \varphi}. \tag{13}$$

Fisher's vector φ for (12) is the eigenvector corresponding to the maximum eigenvalue of $S_w^{-1} S_b$, and (13) can be solved analogously. Then the linear transformation G mentioned above is formulated by the first l eigenvectors corresponding to the largest eigenvalues.

C. Uncorrelated Linear Discriminant Analysis

Uncorrelated linear discriminant analysis was first proposed in [16] to find the optimal discriminant vectors that are S_t-orthogonal. To be specific, suppose $(r-1)$ vectors $\varphi_1, \varphi_2, ..., \varphi_{r-1}$ are obtained, then the rth vector φ_r is the one that maximizes the criterion (12), subject to the constraints: $\varphi_r^T S_t \varphi_i = 0$, $i = 1, ..., r-1$.

In [16], φ_i is found successively as follows: The j-th discriminant vector φ_j of ULDA is the eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem:

$$P_j S_b \varphi_j = \lambda_j S_w \varphi_j, \tag{14}$$

where

$$\begin{aligned} &P_1 = I_m, \\ &P_j = I_m - S_t D_j^T (D_j S_t S_w^{-1} S_t D_j^T)^{-1} D_j S_t S_w^{-1} (j > 1), \\ &D_j = [\varphi_1, \varphi_2, ..., \varphi_{j-1}]^T \ (j > 1), \\ &I_m = diag(1, 1, ..., 1) \in R^{m \times m}. \end{aligned} \tag{15}$$

III. Multi-view Uncorrelated Linear Discriminant Analysis

Correlated information between multiple views can provide useful information for building robust classifiers. Additionally, discriminant features of each view are important for recognition. So we propose a approach to incorporate CCA and LDA, which is called multi-view linear discriminant analysis (MLDA). It aims to achieve maximum correlation between different views and discrimination of each view simultaneously, so that the performance in this transformed common space would be enhanced. Moreover, motivated by the fact that uncorrelated features with minimum redundancy are highly desirable in many applications, we add a constraint into our objective function. Due to this constraint, the extracted feature vectors are mutually uncorrelated in each view. This procedure can be seen as an extension of the LDA ingredient in MLDA. The purpose of our method is to take advantage of both ULDA and CCA, so that useful features can be exploited for multi-view application.

In this section, we first introduce the formulation of multi-view linear discriminant analysis and its solution. Then we present our new approach multi-view uncorrelated linear discriminant analysis and describe the explicit derivation of the final solution.

A. Multi-view Linear Discriminant Analysis

Similar to the notations in the last section, assume we have a two-view data set $\{(x_1, y_1), ..., (x_m, y_m)\} \in R^p \times R^q$, and $X = [x_1, x_2, ..., x_n] = [X_1, X_2, ..., X_k]$, $Y = [y_1, y_2, ..., y_n] = [Y_1, Y_2, ..., Y_k]$, where p and q represent the dimension of X and Y. MLDA seeks to maximize the following objective function:

$$\frac{w_x^T S_{b_x} w_x}{w_x^T S_{t_x} w_x} + \frac{w_y^T S_{b_y} w_y}{w_y^T S_{t_y} w_y} + 2\gamma \frac{w_x^T C_{xy} w_y}{\sqrt{(w_x^T C_{xx} w_x)\left(w_y^T C_{yy} w_y\right)}}, \tag{16}$$

in which we use the common classical LDA criterion (13) to exploit discriminant vectors in each view. $\gamma > 0$ is a tunable parameter to balance the relative significance between the CCA part and the LDA part in (16)

By constraining the factors in the denominator to have value 1, (16) can be formulated as the following optimization problem:

$$\begin{aligned} &\max_{w_x, w_y} \quad w_x^T S_{b_x} w_x + w_y^T S_{b_y} w_y + 2\gamma w_x^T C_{xy} w_y \\ &\text{s.t.} \quad\quad w_x^T S_{t_x} w_x = 1, \ w_y^T S_{t_y} w_y = 1, \end{aligned} \tag{17}$$

where the matrices S_{b_x} and S_{b_y} are constructed according to (10), S_{t_x} and S_{t_y} are computed following (11), and C_{xy} is constructed according to (2). From (3), (4) and (11) we can find that they all represent the covariance matrix of a data set. Thus we replace C_{xx} and C_{yy} with S_{t_x} and S_{t_y} respectively for simplicity.

Through optimizing (17), the correlation between different views and the discrimination of each view can be maximized simultaneously. By using Lagrangian multiplier techniques, we can transform this constrained optimization problem (17) to a generalized multivariate eigenvalue problem of the following form:

$$\begin{bmatrix} S_{b_x} & \gamma C_{xy} \\ \gamma C_{yx} & S_{b_y} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \begin{bmatrix} S_{t_x} & \\ & S_{t_y} \end{bmatrix} \begin{bmatrix} \lambda_x w_x \\ \lambda_y w_y \end{bmatrix}, \tag{18}$$

which has appeared in the solution of [20] and can be solved by an alternation method [21].

In [22], a general multi-view feature extraction approach called generalized multiview analysis (GMA) was proposed for cross-view classification and retrieval. It has a similar formulation with (17), which is relaxed by coupling the constraints with a parameter to obtain a closed-form solution. Similarly, for the sake of convenience, we couple the constraints in (17) with $\sigma = \frac{tr(S_{t_x})}{tr(S_{t_y})}$, such that the constraints are transformed to a single constraint $w_x^T S_{t_x} w_x + \sigma w_y^T S_{t_y} w_y$. In the remainder, we will use this kind of relaxed version to derive our closed-form solution.

B. Multi-view Uncorrelated Linear Discriminant Analysis

It has been proved that uncorrelated features with minimum redundancy are desirable in many applications [16][17][18][19]. Inspired by the fact that ULDA can be successfully combined with other learning methods to obtain better performance [23], we add the uncorrelated constraint $w_r^T S_t w_j = 0$, $j = 1, 2, ..., r-1$ into MLDA, such that the

extracted feature vectors will be mutually uncorrelated in each view.

Let (w_{x1}, w_{y1}) be the vector pair solved by MLDA corresponding to the maximum eigenvalue. Suppose $r-1$ vector pairs (w_{xj}, w_{yj}), $j = 1, 2, ..., r-1$, of the two-view data set are obtained. MULDA seeks to find the rth feature vector pair (w_{xr}, w_{xr}) of data set X and Y which optimize the objective function (17) with the following conjugated orthogonality constraints:

$$w_{xr}^T S_{t_x} w_{xj} = w_{yr}^T S_{t_y} w_{yj} = 0 \quad (j = 1, 2, ..., r-1). \tag{19}$$

The optimization problem of MULDA can be formulated as:

$$\begin{array}{ll} \max_{w_{xr}, w_{yr}} & w_{xr}^T S_{b_x} w_{xr} + w_{yr}^T S_{b_y} w_{yr} + 2\gamma w_{xr}^T C_{xy} w_{yr} \\ \text{s.t.} & w_{xr}^T S_{t_x} w_{xr} + \sigma w_{yr}^T S_{t_y} w_{yr} = 1, \\ & w_{xr}^T S_{t_x} w_{xj} = w_{yr}^T S_{t_y} w_{yj} = 0 \\ & (j = 1, 2, ..., r-1), \end{array} \tag{20}$$

where w_{xr} and w_{xr} represent the rth discriminant vector of data set X and Y, respectively.

The corresponding Lagrangian function of (20) is

$$\begin{aligned} L(w_{xr}, w_{yr}) = {} & w_{xr}^T S_{b_x} w_{xr} + w_{yr}^T S_{b_y} w_{yr} + 2\gamma w_{xr}^T C_{xy} w_{yr} \\ & -\lambda \left(w_{xr}^T S_{t_x} w_{xr} + \sigma w_{yr}^T S_{t_y} w_{yr} - 1 \right) \\ & - \textstyle\sum_{j=1}^{r-1} 2\alpha_j w_{xr}^T S_{t_x} w_{xj} \\ & - \textstyle\sum_{j=1}^{r-1} 2\beta_j w_{yr}^T S_{t_y} w_{yj}. \end{aligned} \tag{21}$$

Taking its derivatives with respect to w_{xr} and w_{yr} to be zero, we have

$$S_{b_x} w_{xr} + \gamma C_{xy} w_{yr} - \lambda S_{t_x} w_{xr} - \sum_{j=1}^{r-1} \alpha_j S_{t_x} w_{xj} = 0, \tag{22}$$

$$S_{b_y} w_{yr} + \gamma C_{yx} w_{xr} - \lambda \sigma S_{t_y} w_{yr} - \sum_{j=1}^{r-1} \beta_j S_{t_y} w_{yj} = 0. \tag{23}$$

Multiplying the left-hand side of (22) and (23) by w_{xr}^T and w_{yr}^T, respectively, we obtain

$$2\lambda = w_{xr}^T S_{b_x} w_{xr} + w_{yr}^T S_{b_y} w_{yr} + 2\gamma w_{xr}^T C_{xy} w_{yr}, \tag{24}$$

which means 2λ represents the value of the objective function in (20).

Multiplying the left-hand side of (22), respectively, by w_{xi}^T, we obtain a set of $r-1$ expressions:

$$\begin{aligned} & w_{xi}^T S_{b_x} w_{xr} + \gamma w_{xi}^T C_{xy} w_{yr} - \textstyle\sum_{j=1}^{r-1} \alpha_j w_{xi}^T S_{t_x} w_{xj} = 0 \\ & (i = 1, 2, ..., r-1), \end{aligned} \tag{25}$$

which can be expressed in another form

$$\begin{aligned} & \begin{bmatrix} w_{x1}^T \\ w_{x2}^T \\ \vdots \\ w_{x(r-1)}^T \end{bmatrix} S_{b_x} w_{xr} + \gamma \begin{bmatrix} w_{x1}^T \\ w_{x2}^T \\ \vdots \\ w_{x(r-1)}^T \end{bmatrix} C_{xy} w_{yr} \\ & - \begin{bmatrix} w_{x1}^T \\ w_{x2}^T \\ \vdots \\ w_{x(r-1)}^T \end{bmatrix} S_{tx} \begin{bmatrix} w_{x1}^T \\ w_{x2}^T \\ \vdots \\ w_{x(r-1)}^T \end{bmatrix}^T \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{r-1} \end{bmatrix} = 0. \end{aligned} \tag{26}$$

Let

$$\begin{aligned} \vec{\alpha} &= \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{r-1} \end{bmatrix}^T, \\ D_x &= \begin{bmatrix} w_{x1} & w_{x2} & \cdots & w_{x(r-1)} \end{bmatrix}^T, \end{aligned} \tag{27}$$

so that (25) can be represented in a single matrix relation

$$D_x S_{b_x} w_{xr} + \gamma D_x C_{xy} w_{yr} = D_x S_{t_x} D_x^T \vec{\alpha}. \tag{28}$$

Thus, we obtain

$$\vec{\alpha} = \left(D_x S_{t_x} D_x^T \right)^{-1} \left(D_x S_{b_x} w_{xr} + \gamma D_x C_{xy} w_{yr} \right). \tag{29}$$

Symmetrically, let

$$\begin{aligned} \vec{\beta} &= \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_{r-1} \end{bmatrix}^T, \\ D_y &= \begin{bmatrix} w_{y1} & w_{y2} & \cdots & w_{y(r-1)} \end{bmatrix}^T, \end{aligned} \tag{30}$$

then we get

$$\vec{\beta} = \left(D_y S_{t_y} D_y^T \right)^{-1} \left(D_y S_{b_y} w_{yr} + \gamma D_y C_{yx} w_{xr} \right). \tag{31}$$

Using (27), (22) can be rewritten as

$$S_{b_x} w_{xr} + \gamma C_{xy} w_{yr} - \lambda S_{t_x} w_{xr} - S_{t_x} D_x^T \vec{\alpha} = 0. \tag{32}$$

Substituting (29) into (32), we have

$$\begin{aligned} & \left[I - S_{t_x} D_x^T \left(D_x S_{t_x} D_x^T \right)^{-1} D_x \right] \left(S_{b_x} w_{xr} + \gamma C_{xy} w_{yr} \right) \\ & = \lambda S_{t_x} w_{xr}. \end{aligned} \tag{33}$$

Analogously, from (23) and (31) we have

$$\begin{aligned} & \left[I - S_{t_y} D_y^T \left(D_y S_{t_y} D_y^T \right)^{-1} D_y \right] \left(S_{b_y} w_{yr} + \gamma C_{yx} w_{xr} \right) \\ & = \lambda \sigma S_{t_y} w_{yr}. \end{aligned} \tag{34}$$

Let

$$\begin{aligned} P_x &= I - S_{t_x} D_x^T \left(D_x S_{t_x} D_x^T \right)^{-1} D_x, \\ P_y &= I - S_{t_y} D_y^T \left(D_y S_{t_y} D_y^T \right)^{-1} D_y, \end{aligned} \tag{35}$$

then we derive a generalized eigenvalue problem:

$$\begin{bmatrix} P_x & \\ & P_y \end{bmatrix} \begin{bmatrix} S_{b_x} & \gamma C_{xy} \\ \gamma C_{yx} & S_{b_y} \end{bmatrix} \begin{bmatrix} w_{xr} \\ w_{yr} \end{bmatrix} = \lambda \begin{bmatrix} S_{t_x} & \\ & \sigma S_{t_y} \end{bmatrix} \begin{bmatrix} w_{xr} \\ w_{yr} \end{bmatrix}. \tag{36}$$

Equivalently,

$$\tilde{P} \tilde{S}_b \hat{w}_r = \lambda \tilde{S}_t \hat{w}_r, \tag{37}$$

from which we can obtain the rth uncorrelated feature vector $\hat{w}_r = \begin{bmatrix} w_{xr}^T & w_{yr}^T \end{bmatrix}$, which is the eigenvector corresponding to the maximum eigenvalue of (37). Matrices $\tilde{P}$, $\tilde{S}_b$ and $\tilde{S}_t$ are the corresponding matrices in (36).

With d obtained vector pairs (w_{xj}, w_{yj}), $j = 1, 2, ..., d$ after d iterations, let $W_x = [w_{x1}, w_{x2}, ..., w_{xd}]$, $W_y = [w_{y1}, w_{y2}, ..., w_{yd}]$. The combined feature extraction can be performed according to the following two strategies [24]:

$$I)\ Z = \begin{bmatrix} W_x & \\ & W_y \end{bmatrix}^T \begin{bmatrix} X \\ Y \end{bmatrix}, \tag{38}$$

$$II)\ Z = \begin{bmatrix} W_x \\ W_y \end{bmatrix}^T \begin{bmatrix} X \\ Y \end{bmatrix}, \tag{39}$$

and d satisfies the constraints $1 \leq d \leq min(p, q)$ and $1 \leq d \leq k$. Both of them are usable. In our experiments, we apply the first strategy to fuse extracted features. The main algorithm is given in Algorithm 1.

Algorithm 1 Multi-view uncorrelated linear discriminant analysis

Input:
Training data X, Y;
Reduced dimensions d;
Parameter λ;
Output:
Transformed data Z;

1: Construct matrices C_{xy}, S_{b_x}, S_{b_y}, S_{t_x}, S_{t_y} as in (2), (10), (11).
2: $\sigma \leftarrow \frac{tr(S_{t_x})}{tr(S_{t_y})}$.
3: Initialize $D_x = \emptyset$, $D_y = \emptyset$
4: **for** $r = 1$ **to** d **do**
5: Construct matrices P_x, P_y as in (35);
6: Obtain the rth vector pair (w_{xr}, w_{yr}) by solving (36);
7: Set $D_x = D_x \cup w_{xr}$, $D_y = D_y \cup w_{yr}$;
8: **end for**
9: $W_x \leftarrow D_x$, $W_y \leftarrow D_y$;
10: Extract features according to (38);
11: **return** Z.

As our projection vectors are solved by generalized eigenvalue decomposition, and in some cases $\tilde{S}_t$ could be singular, such that (37) can not be applied directly, we add a regularizer to $\tilde{S}_t$ [25] in our experiments.

IV. Experiments on Handwritten Digit Recognition

In this section, we evaluate the effectiveness of our method MULDA on handwritten digit recognition. Section IV.A describes our data set. Section IV.B examines the effect of the number of reduced dimensions on the recognition performance of MULDA. In Section IV.C, we compare MULDA with DCCA and *k*-nearest-neighbor (KNN), in terms of recognition accuracy. After feature extraction using MULDA and DCCA, the KNN classifier with $K = 3$ is employed.

A. Data Set

The multiple features data set, which is available from the UCI repository, consists of features of handwritten digits ('0'-'9') extracted from a collection of Dutch utility maps. 200 samples per class (for a total of 2,000 samples) have been digitized in binary images. Six sets of features, which respectively describe the digits from different views, are included. The six feature sets and number of features in each set are listed as follows: 1) Fourier coefficients of the character shapes (**FOU**,76); 2) Profile correlations (**FAC**,216); 3) Karhunen-Love coefficients (**KAR**,64); 4) Pixel averages in 2 x 3 windows (**PIX**,240); 5) Zernike moments (**ZER**,47); 6) Morphological features (**MOR**,6).

Any two of them are picked out to construct view X and view Y, so that there are total 15 pairs of different combinations and each combination forms a two-view data set. For each class, 100 pairs of feature vectors are randomly picked out for training, and the remaining are for test. We report averaged results after 20 random experiments. In the implementation of MULDA, we use 5-fold cross validation to select tuning parameter γ among $[0, 1000]$ for each two-view data set.

B. Effect of the Number of Reduced Dimensions on MULDA

In this experiment, the effect of the number of reduced dimensions on the recognition performance of MULDA is studied. We run MULDA by keeping the first $\tilde{d}$ dimensions only, where $1 \leq \tilde{d} \leq min(p, q)$ and $1 \leq \tilde{d} \leq k$. The recognition results on the combination of **PIX** and **ZER** are shown in Fig. 1, where the horizontal axis represents the reduced dimensions and the vertical axis represents the recognition accuracy. We can observe that the accuracy increases monotonically as the number of reduced dimensions increases, until $\tilde{d} = k - 1$ is reached. This observation is consistent with the theory in [8], that the optimal dimensionality of feature space is $k - 1$. Most observations of other two-view data sets are similar, so we do not present them here. For the two-view data sets in which feature set **MOR** is included, the reduced dimensions is set to be the dimension of **MOR**. Otherwise, in the following experiment, we set the reduced dimensions of MULDA to be $k - 1$.

C. Comparison of Recognition Performance

In this section, we present experimental results which compare MULDA with other two algorithms, DCCA and KNN. The reduced dimensions of DCCA is set to be the same with MULDA. KNN denotes the method that apply 3-NN classifier directly on two-view data sets. The results are summarized in Table I.

To make the comparison results more intuitive, we summarize the results in Fig. 2. Each number in horizonal axis corresponds to a two-view data set in Table I. We can observe from Fig. 2 that in most cases, the recognition performance of MULDA is better than DCCA and KNN. And in those cases that KNN is superior, MULDA is more competitive with KNN than DCCA. These comparative results confirm that in most instances, MULDA is able to extract a small number of features in each view and fuse them without loss of classification accuracy. Moreover, as MULDA removes the redundancy in the original features while achieving maximum correlation between different views and discrimination

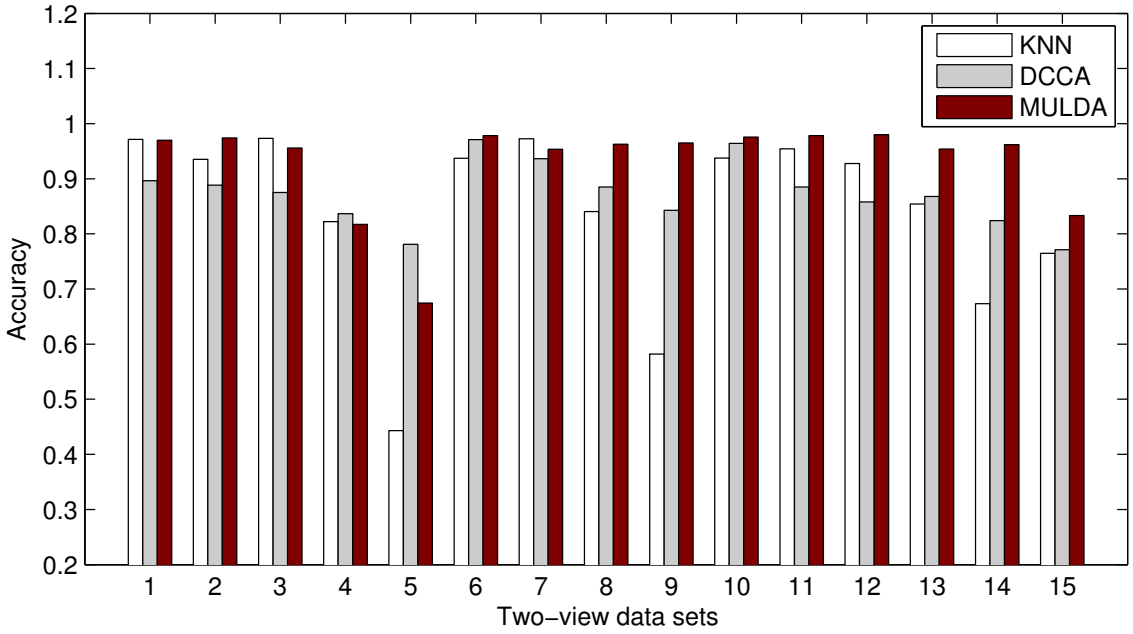

Fig. 2. Comparison of recognition performance among KNN, DCCA and MULDA

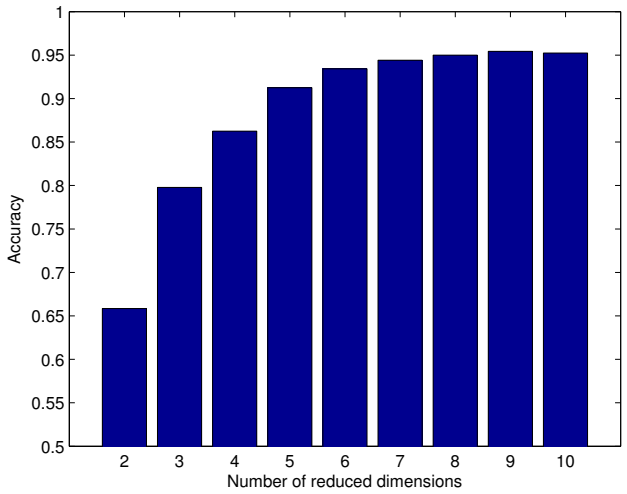

Fig. 1. Effect of the number of reduced dimensions on the recognition performance of MULDA for the combination of **PIX** and **ZER**

TABLE I

RECOGNITION ACCURACIES ON MULTIPLE FEATURE DATA SET.

X	Y	KNN	DCCA	MULDA
FOU	**KAR**	**0.9714**	0.8964	0.9699
FOU	**FAC**	0.9353	0.8885	**0.9740**
FOU	**PIX**	**0.9733**	0.8752	0.9558
FOU	**ZER**	0.8223	**0.8367**	0.8174
FOU	**MOR**	0.4432	**0.7812**	0.6745
KAR	**FAC**	0.9372	0.9710	**0.9781**
KAR	**PIX**	**0.9726**	0.9365	0.9534
KAR	**ZER**	0.8407	0.8851	**0.9626**
KAR	**MOR**	0.5820	0.8427	**0.9651**
FAC	**PIX**	0.9377	0.9643	**0.9757**
FAC	**ZER**	0.9543	0.8851	**0.9782**
FAC	**MOR**	0.9277	0.8581	**0.9796**
PIX	**ZER**	0.8542	0.8677	**0.9539**
PIX	**MOR**	0.6735	0.8242	**0.9618**
ZER	**MOR**	0.7649	0.7711	**0.8331**

in each view, the accuracy rate can be improved in the common space in many situation. DCCA utilizes label information by maximizing the difference between within-class and between-class correlations across two views. However, it misses the discriminative information within each view, which is very important for classification. For this reason, the recognition performance of DCCA is usually not as good as MULDA, though it can get the best performance on some rare occasions.

V. CONCLUSION AND FUTURE WORK

In this paper, we develop MULDA, an efficient algorithm that combines ULDA and CCA to simultaneously take advantage of these two algorithms. Different from previous work, both intra-view class structure and inter-view correlation are considered in our method. Additionally, the feature vectors extracted by our method are mutually uncorrelated in the common space, which means we can remove the redundancy in the original features while achieving maximum correlation between different views and discrimination in each view.

Comparative experiments on handwritten digit recognition verify the effectiveness of MULDA. The experimental results show that MULDA outperforms other related works in most cases.

In the implementation, we derive the closed-form solution of MULDA based on a relaxation of the constraints. Some deviations may be caused by this approximation. In our future work, we will study the effect of this approximation. Besides, for large and high-dimensional data sets, our algorithm may be computationally expensive, since each feature vector corresponds to a generalized eigenvalue decomposition problem, which encourages us to exploit more efficient closed-form solution, such as [17].

Additionally, inspired by [23], incorporating sparsity into our algorithm will be one future focus. And many studies imply that a non-linear extension of feature extraction methods can improve performance especially when the data has weak linear separability [26]. In the future, we also plan to extend the current work to deal with the nonlinearity.

REFERENCES

[1] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, pp. 2031-2038, 2013.

[2] H. Hotelling, "Relations between two sets of variates," *Biometrica*, vol. 28, pp. 321-377, 1936.

[3] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge university press, 2004.

[4] J. Shawe-Taylor, S. Sun, "Kernel methods and support vector machines," *Book chapter for E-Reference Signal Processing*, Elsevier, 2013.

[5] T. Sun, S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing*, vol. 25, pp. 531-542, 2007.

[6] J. B. Tenenbaum, W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, pp. 1247-1283, 2000.

[7] R. Rosipal, N. Kramer, "Overview and recent advances in partial least squares," *Subspace, Latent Structure and Feature Selection*, vol. 3940, pp. 34-51, 2006.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[9] T. Sun, S. Chen, J. Yang, P. Shi, "A novel method of combined feature extraction for recognition," *Proceedings of the International Conference on Data Mining*, pp. 1043-1048, 2008.

[10] J. Zhang, D. Zhang, "A novel ensemble construction method for multi-view data using random cross-view correlation between within-class examples," *Pattern Recognition*, vol. 44, pp. 1162-1171, 2011.

[11] T. Diethe, D. R. Hardoon, J. Shawe-Taylor, "Multiview fisher discriminant analysis," *NIPS Workshop on Learning from Multiple Sources*, 2008.

[12] T. Diethe, D. Hardoon, and J. Shawe-Taylor, "Constructing nonlinear discriminants from multiple data views," *Machine Learning and Knowledge Discovery in Databases*, vol. 6321, pp. 328-343, 2010.

[13] Q. Chen and S. Sun, "Hierarchical multi-view fisher discriminant analysis," *Lecture Notes in Computer Science*, vol. 5864, pp. 289-298, 2009.

[14] D. Lin, X. Tang, "Inter-modality Face Recognition," *Proceedings of the European Conference on Computer Vision*, vol. 3954, pp. 13-26, 2006.

[15] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, "Multi-view discriminant analysis," *Proceedings of the European Conference on Computer Vision*, vol. 7572, pp. 808-821, 2012.

[16] Z. Jin, J Y. Yang, Z S. Hu, Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognition*, vol. 34, pp. 1405-1416, 2001.

[17] J. Ye, T. Li, T. Xiong, R. Janardan, "Using uncorrelated discriminant analysis for tissue classication with gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, pp. 181-190, 2004.

[18] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research*, vol. 6, pp. 483-502, 2005.

[19] J. Ye, R. Janardan, Q. Li, H. Park, "Feature extraction via generalized uncorrelated linear discriminant analysis," *Proceedings of the International Conference on Machine Learning*, pp. 113, 2004.

[20] J. Rupnik, J. Shawe-Taylor, "Multi-View Canonical Correlation Analysis," *Proceedings of the Conference on Data Mining and Data Warehouses*, pp. 1-4, 2010.

[21] P. Horst, "Relations amongm sets of measures," *Psychometrika*, vol. 26, pp. 129-149, 1961.

[22] A. Sharma and A. Kumar, H. Daume, D. W. Jacobs, "Generalized Multiview Analysis: A discriminative latent space," *Proceedings of the Computer Vision and Pattern Recognition*, pp. 2160-2167, 2012.

[23] X. Zhang, D. Chu, "Sparse uncorrelated linear discriminant analysis," *Proceedings of the International Conference on Machine Learning*, pp. 45-52, 2013.

[24] Q. Sun, S. Zeng, Y. Liu, P. Heng, D. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, pp. 2437-2448, 2005.

[25] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, pp. 165-175, 1989.

[26] Z. Liang, P. Shi, "Uncorrelated discriminant vectors using a kernel method," *Pattern Recognition*, vol. 38, pp. 307-310, 2005.