*Journal of Information Hiding and Multimedia Signal Processing* ©2014 ISSN 2073-4212
**Ubiquitous International** Volume **5**, Number **3**, July **2014**

# Shared Subspace Learning for Latent Representation of Multi-View Data

Zhenfeng Zhu, Linlin Du, Lei Zhang and Yao Zhao

Institute of Information Science
Beijing Jiaotong University,Beijing,100044,China
Beijing Key Laboratory of Advanced Information Science and Network Technology
Beijing, 100044, China

Received March, 2014; revised May, 2014

Abstract. *The pervasive existence of multi-view data has made conventional single view data analysis methods to confront with great challenge. To exploit new analysis technique for multi-view data has become one of active topics in the field of machine learning. From the point of shared subspace learning, this paper focuses on capturing the shared latent representation across multi-view by constructing the correlation in a shared subspace. Different from the classical canonical correlation analysis (CCA) , a more general model for learning a shared subspace was proposed, which not only provides an explicit latent representation but also can leverage favorably the contributions from different views for capturing the complementary information across multi-view. In order to preserve well the local structure of data in the both shared subspace and original multi-view feature spaces, a graph constraint is employed. Meanwhile, with the assist of prior class label, the boosted discriminative ability of the proposed multi-view analysis model can be achieved. The experimental results on the multi-view data retrieval and classification verify the effectiveness of the proposed model.*
**Keywords:** multi-view learning ,canonical correlation analysis (CCA), shared subspace learning , cross media

1. **Introduction.** The recent years have witnessed a growing emergence of multi-view data in the real world. Here, the multi-view is referred to as the data coming from diverse domains or sources but with an underlying consistent agreement among them. Meanwhile, each of views of a data can be characterized by a distinct attribute set or representation. Considering a situation of web image understanding, a web image can be described not only by the image content itself, but also at the same time by the surround document text. On the common assumption that each of views can complement each other, multi-view leaning is mainly concerned with exploiting the complementary information across the multiple views to improve the generalization ability of learning model, which is great different from those traditional machine learning methods on single view dada analysis. In most cases, multi-view learning has demonstrated its obvious advantage over the learning from single view [1, 4, 18].

From a technical point of view, multi-view learning can be categorized into Co-training [1, 2, 3, 4], multi-kernel learning [5, 6, 7]and shared subspace learning [8, 9, 10, 11, 23]. For a more complete survey of the literatures on multi-view learning, please refer to [12]. As one of semi-supervised learning methods, Co-training learning style was first proposed by Blum and Mitchell et al. [1] to alleviate the difficulty arising from the problem of

small size of labeled sample. Due to the encouraging success of Co-training in some applications, many substantial variants of it have devoted to derive effective and efficient learning performance [2, 3, 4]. Multi-kernel learning (*MLK*) aims to synthesize a unified kernel model by the linear or nonlinear combination of fixed kernels [5]. Considering the multi-view scenarios, one can construct each of base kernels on a corresponding single view of the input data. Consequently, the *MKL* approach can be naturally extended to handle multi-view data [6, 7].

Different from the former two multi-view learning mechanisms, the goal of shared subspace learning is to discover, for each multi-view data, a shared latent representation, such that the complementary information embedded in each of heterogeneous views can be well revealed [9]. One of the potential applications of shared subspace learning can be for cross-media retrieval and information fusion [21]. For instance, in the shared subspace, one can search a textual web page by taking an input image as query and vice versa. Classically, the most referred shared subspace learning method for multi-view data is canonical correlation analysis (*CCA*) [8, 13], which was first proposed by Hotelling [8].To satisfy the nonlinear condition, kernel canonical correlation analysis (*KCCA*) [14] has been proposed by the application of the kernel trick. By encoding the class label of data into a new view, *CCA* will be formulated as a least squares problem [15]. Some other variants of *CCA* can be found in [16, 17]. In [18], Sharma et al. developed a generalized multi-view analysis (*GMA*) method and showed that *CCA* is a special instance of *GMA*. By weighting joint matrix factorization, Yu et al. [19] proposed a shared subspace model (Multi-output Regularized Feature Projection, *MORP*) to build the correlation of multi-view data in the shared subspace. In the works of Salzmann et al. [20], they proposed to find a latent shared subspace in which the information is correctly factorized into shared and private parts across different views.

Compared with *CCA*, the superiorities of *MORP* are two-fold: 1) an explicit shared latent representation can be obtained; 2) the contributions from different views for capturing the complementary information across multi-view to form the latent representation can be well balanced. But the disadvantage of *MORP* is also obvious since it can't deal with the problem of out-of-sample efficiently. In this paper, we concentrated on the problem of learning a shared subspace for multi-view data. Specifically, a general shared subspace learning model was proposed to obtain a consistent representation of multi-view data, which fully considers both the merits of *CCA* and *MORP* models. In order to preserve the local geometrical structure of data in the both shared subspace and original multi-view feature spaces, a graph constraint is introduced. Meanwhile, with the assist of prior class label, the boosted generalization ability of the proposed multi-view analysis model has been achieved.

The rest of this paper is organized as follows. In the Section 2, we give a background review of related work. Section 3 presents the proposed general model for correlating multi-view data in a shared subspace. The experimental results and performance analysis are given in Section 4. Section 5 concludes the work of this paper.

2. **Problem Statement and Preliminaries.** Without loss of generality, we only consider the scenario of two- view in this paper. Given two sets of observations $X = [x_1, ..., x_n]^T \in \mathbb{R}^{n \times d_x}$ and $Y = [y_1, ..., y_n]^T \in \mathbb{R}^{n \times d_y}$, $\{x_i, y_i\}_{i=1,...,n}$ denotes a paired representations or views for the $i$-th sample. We assume that both $\{x_i\}_{i=1,...n}$ and $\{y_i\}_{i=1,...n}$ are centered, i.e. $\sum_{i=1}^{n} x_i = 0$ and $\sum_{i=1}^{n} y_i = 0$. We use $I_k$ for the $k \times k$ identity matrix, $|| C ||_F = \sqrt{\sum_{i,j} c_{i,j}^2}$ to represent the Fresenius norm of matrix $C$, and $Tr(\cdot)$ to denote the

trace of a symmetric matrix. In addition, for a matrix $A$, we use $a_{i.}$ to denote the $i$-th row and $a_{.j}$ the $j$-th column of $A$.

2.1. **Canonical Correlation Analysis (CCA).** Proposed by H. Hotelling in 1936 [8], Canonical Correlation Analysis ($CCA$) aims to seek pairs of basis vectors that maximize the correlation between the projections of the paired variables onto the corresponding basis vectors. The above correlation maximization problem can be formularized as follows:

$$\begin{aligned} &\arg max\ \ \rho = a^T \cdot C_{xy} \cdot b \\ &s.t.\ \ a^T \cdot C_{xx} \cdot a = 1,\ \ b^T \cdot C_{yy} \cdot b = 1 \end{aligned} \tag{1}$$

where $C_{xx} = X^T \cdot X$ and $C_{yy} = Y^T \cdot Y$ are the non-singular within-set covariance matrices and $C_{xy} = X^T \cdot Y$ is the between-sets covariance matrix, $a \in \mathbb{R}^{d_x \times 1}$ and $b \in \mathbb{R}^{d_y \times 1}$ are two corresponding projection vectors. Equivalently, this objective function (1) can be re-written in the form of matrix as:

$$\begin{aligned} &\arg min\ \ \Phi(A, B) = \|X \cdot A - Y \cdot B\|_F^2 \\ &s.t.\ \ A^T \cdot C_{xx} \cdot A = I_d,\ \ B^T \cdot C_{yy} \cdot B = I_d \end{aligned} \tag{2}$$

where $A = [a_{.1}, ...a_{.d}] \in \mathbb{R}^{d_x \times d}$, $B = [b_{.1}, ...b_{.d}] \in \mathbb{R}^{d_y \times d}$, $\{a_{.i}, b_{.i}\}_{i=1,...d}$ is a pair of projection basis vectors, $A^T \cdot C_{xx} \cdot A = I_d$, and $B^T \cdot C_{yy} \cdot B = I_d$ restrict the $d$ latent variables to be linearly independent. By some mathematical manipulation, the optimization problem given by Eq.(2) will lead to the following generalized eigenvalue problem:

$$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} = \lambda_i \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix} \cdot \begin{bmatrix} a_{.i} \\ b_{.i} \end{bmatrix} \tag{3}$$

It is worth to note that, for the sample $\{x_i, y_i\}$, $CCA$ does not provide explicitly a shared latent representation $u_{i.}$. In real application, the projections $x_i^T \cdot A$ and $y_i^T \cdot B$ have popularly been used to approximate $u_{i.}$, i.e. $u_{i.} = \frac{x_i^T \cdot A + y_i^T \cdot B}{2}$.

2.2. **Multi-Output Regularized Feature Projection.** Recently, multi-output regularized feature projection ($MORP$) has been proposed by Yu et al. [19] to build the correlation of multi-view data in the shared subspace by weighted joint matrix factorization. In their work, the encoded class label vector for each data is used to form its corresponding y view. However, it is straightforward to extend $MORP$ to real multi-view data. The objective function of $MORP$ model is given by:

$$\begin{aligned} &\arg min\ \ \Phi(P_x, P_y, U) = \beta\|X - U \cdot P_x\|_F^2 + (1-\beta)\|Y - U \cdot P_y\|_F^2 \\ &s.t.\ \ U^T \cdot U = I_d \end{aligned} \tag{4}$$

where $\beta$ is a balance weight to trade- off the contributions for reconstruction from both $X$ view and $Y$ view , respectively, $U = [u_{.1}, ..., u_{.d}] \in \mathbb{R}^{n \times d}$ and its $i$-th row $u_{i.}$ denotes the shared latent representation of the $i$-th sample in $d$-dimensional shared subspace, $P_x$ and $P_y$ are the corresponding loading matrixes. Here, the purpose of imposing the orthogonalization constraint on $U$ is to guarantee the independence among the variables in the shared subspace. As pointed out in [19], a generalized eigenvalue problem can be formulated to seek the optimal solution to the Eq.(4). Compared with the classical multi-view analysis method $CCA$, one can obtain explicitly a latent representation $u_{i.}$ for the sample $\{x_i, y_i\}$ . In addition, the balancing parameter $\beta$ can well leverage the contributions of each view for constructing the shared subspace. However, it is also obvious that the $MORP$ lacks the ability to deal with the problem of out-of-sample.

## 3. Correlating Multi-view Data in a Shared Subspace.

3.1. **A General Shared Subspace Model.** Motivated by *CCA* and *MORP*, we propose in this paper a more general model for correlating multi-view data in a shared subspace, which favorably takes both the advantages of *CCA* and *MORP*. Specifically,the proposed model is given as follows:

$$\begin{aligned} &\arg min\ \ \Phi(A,B,U) = \underbrace{(1-\beta)\|X\cdot A-U\|_F^2+\beta\|Y\cdot B-U\|_F^2}_{Reconstruction\ Term}+\mu \underbrace{\Psi(U)}_{Regularization\ Term} \\ &s.t.\ U^T\cdot U = I_d \end{aligned} \tag{5}$$

where $\beta$ and $\mu$ are two balancing parameters, the definition of $U$ is the same as in Eq.(4). Here, the reconstruction term based on joint matrix factorization term is used to correlating multi-view data in a shared subspace. Generally, as shown in the following subsections,the regularization term $\Psi(U)$ can possesses in most of cases an quadratic form, i.e. $\Psi(U) = Tr(U^T\cdot Q\cdot U)$ , where the symmetric matrix $Q$ holds some kind of special meaning according to the different definition of the regularization term. Thus, it can be found obviously in this case that the objective function $\Phi(A,B,U)$ in Eq.(5) is convex. The following *Lemma* shows a global optimum solution for the above convex optimization problem can be achieved.

**Lemma 3.1.** *Assume that $A^*$ and $B^*$ be the optimal solutions to Eq. (5) with $\Psi(U) = Tr(U^T\cdot Q\cdot U)$ , then we will have $A^* = (X^T\cdot X)^{-1}\cdot X^T\cdot U$ and $B^* = (Y^T\cdot Y)^{-1}\cdot Y^T\cdot U$ , with which Eq. (5) will be equivalent to the following maximization problem:*

$$\begin{aligned} &\arg max\ \ Tr[U^T\cdot(G-\mu Q)\cdot U] \\ &s.t.\ U^T\cdot U = I_d \end{aligned} \tag{6}$$

*where $G = (1-\beta)X\cdot(X^T\cdot X)^{-1}\cdot X^T+\beta Y\cdot(Y^T\cdot Y)^{-1}\cdot Y^T$ . Meanwhile, it also means that the optimal $U^*$ can be obtained by seeking $d$ eigenvectors of $G-\mu Q$, which correspond to the first $d$ biggest eigenvalues.*

*Proof.* For the constrained minimization problem given in Eq.(5), it can be transformed to a unconstrained form by introducing Lagrange multipliers:

$$\begin{aligned} \arg min\ \ L(A,B,U) &= (1-\beta)\|X\cdot A-U\|_F^2+\beta\|Y\cdot B-U\|_F^2+\mu Tr(U^T\cdot Q\cdot U) \\ &+Tr[\lambda\cdot(U^T\cdot U-I_d)] \end{aligned} \tag{7}$$

where $\lambda\in\mathbb{R}^{d\times d}$ is a symmetric matrix with $\lambda_{i,j}\geq 0$ being a Lagrange multiplier. Setting the derivative of $L$ w.r.t $A$ and $B$ to be zero, we have:

$$\begin{cases} \frac{\partial L}{\partial A} = 2(1-\beta)X^T\cdot X\cdot A-2(1-\beta)X^T\cdot U = 0 \\ \frac{\partial L}{\partial B} = 2\beta Y^T\cdot Y\cdot B-2\beta Y^T\cdot U = 0 \end{cases}$$

By some mathematical manipulation, it is not hard to get $A^* = (X^T\cdot X)^{-1}\cdot X^T\cdot U$ and $B^* = (Y^T\cdot Y)^{-1}\cdot Y^T\cdot U$ . Then, plugging these two optimal $A^*$ and $B^*$ back into Eq.(5), the optimization problem will become a maximization problem as shown in Eq.(6). This completes the proof. □

When we set to $\mu = 0$, the formulation (5) will come down to:

$$\begin{aligned} &\arg min\ \ \Phi(A,B,U) = \underbrace{(1-\beta)\|X\cdot A-U\|_F^2+\beta\|Y\cdot B-U\|_F^2}_{Reconstruction\ term} \\ &s.t.\ U^T\cdot U = I_d \end{aligned} \tag{8}$$

For the formulation (8), we can find that it holds close connection with *CCA*, especially when the optimal solutions of $A^*$,$B^*$ and $U^*$ satisfy $U^* = X\cdot A^*$ and $U^* = Y\cdot B^*$, they

will achieve agreement with each other. In addition, a latent shared representation can be provided explicitly, which shows an extremely significant difference from *CCA*. Hereafter, we name this formulation by explicit *CCA* or *eCCA* for short.

3.2. **Graph Regularized Shared Subspace Model.** In some recent studies for subspace learning, the preservation of the local structure of the data in modeling the subspace has received considerable attention and demonstrated good performance [21, 22] since it reflects the intrinsic structure property very well. Particularly, the popularly adopted method for preserving the local structure is graph constraint. In the following, we will introduce based on Eq.(5) a graph regularized shared subspace model, namely *GRSS*. Thus, in the shared subspace, the over-fitting problem of multi-view data can be avoided to some extent. Specifically, the proposed *GRSS* is modeled as:

$$\begin{aligned} &\arg min\ \ \Phi(A,B,U) = \underbrace{(1-\beta)\|X\cdot A-U\|_F^2+\beta\|Y\cdot B-U\|_F^2}_{Reconstruction\ term}+ \\ &\mu\ \underbrace{\sum_i\sum_j w_{i,j}\|\frac{u_{i.}}{\sqrt{d_i}}-\frac{u_{j.}}{\sqrt{d_j}}\|_F^2}_{Graph\ constrained\ regularization\ term} \\ &s.t.\ U^T\cdot U=I_d \end{aligned} \tag{9}$$

where $w_{i,j}$ is the edge weight to reflect the magnitude of strength of linkage between the $i$-th and the $j$-th samples and $d_i=\sum_j w_{i,j}$ . In our work, we denote the edge weight $w_{i,j}$ by $w_{i,j}=\frac{w^x_{i,j}+w^y_{i,j}}{2}$ with $w^x_{i,j}$ to be defined as:

$$w^x_{i,j}=\begin{cases} exp\left(-\frac{d^2(x_i,x_j)}{2\sigma_i\sigma_j}\right) & x_i\in\mathcal{N}_k(x_j)\ \text{or}\ \ x_j\in\mathcal{N}_k(x_i) \\ 0 & else \end{cases} \tag{10}$$

where $d(x_i,x_j)$ is a distance function and the Euclidean metric is used in this paper unless special specification, $\sigma_i=median[d(x_i,x_j)]_{x_j\in N_k(x_i)}$ , and $N_k(x_i)$ denotes the set of $k$ nearest neighbors of $x_i$ . The similar definition is for $w^y_{i,j}$ . Note that the graph constrained regularization term $\Psi(U)=\sum_i\sum_j w_{i,j}\|\frac{u_{i.}}{\sqrt{d_i}}-\frac{u_{j.}}{\sqrt{d_j}}\|_F^2$ is applied to unfold the underlying local geometrical structure of data.

By simplifying Eq.(9), we have:

$$\begin{aligned} &\arg min\Phi(A,B,U) = \underbrace{(1-\beta)\|X\cdot A-U\|_F^2+\beta\|Y\cdot B-U\|_F^2}_{Reconstruction\ term}+\mu\underbrace{Tr(U^T\cdot L_u\cdot U)}_{Regularization\ term} \\ &s.t.\ U^T\cdot U=I_d \end{aligned} \tag{11}$$

where $L_u=I-D^{-1/2}\cdot W\cdot D^{-1/2}$ is the normalized combinatorial Laplacian operator, $W=[w_{i,j}]_{i,j=1,...n}$ , and $D=diag[d_1,...d_n]$is a diagonal matrix. Since the regularization term $\Psi(U)=Tr(U^T\cdot L_u\cdot U)$ in Eq.(11) holds a quadratic form, following the Lemma 3.1, we can easily get the corresponding optimal solution to Eq.(11) with $Q=L_u$.

3.3. **Discriminant Shared Subspace Model (DSS).** Now, let's consider the classification task for multi-view data. In order to make the shared latent representation of multi-view data in the shared subspace to be more discriminative, the class label information is employed to form a discriminant regularization term. We call this shared subspace model with discriminative ability by discriminant shared subspace model, or *DSS* for short. Following the general shared subspace model as given in Eq.(5), we denote the

*DSS* model by :

$$\arg min\ \Phi(A,B,U) = \underbrace{(1-\beta)\|X\cdot A-U\|_F^2+\beta\|Y\cdot B-U\|_F^2}_{Reconstruction\ term}+ \tag{12}$$

$$\underbrace{\mu\,(\sum_{k=1}^{c}\sum_{i=1}^{n_k}\|u_{i.}-\hat{u}_k\|_F^2-\rho(\sum_{p=1}^{c}\sum_{q=1}^{c}\|\hat{u}_p-\hat{u}_q\|_F^2)}_{Discrininative\ regularization\ term}$$

$$s.t.\ U^T\cdot U=I_d$$

where $c$ is the number of class. For $k=1,...c$, $n_k$ is the number of samples of the $k$-th class and $\hat{u}_k$ denotes the sample mean from the $k$-th class.

For making the following illustrations more clearer, we further introduce some symbols by $S_w=(E_1-E_2)^T(E_1-E_2)$,$S_b=(E_2-E_3)^T(E_2-E_3)$,where $E_1=\begin{bmatrix} E^{[n_1]} & 0 & 0 & 0\\ 0 & \ddots & 0 & 0\\ 0 & 0 & E^{[n_c]} & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\in \mathbb{R}^{n\times n}$ with $E^{[n_k]}=[1,1,...1]^T\cdot[1,1,...1]\in\mathbb{R}^{n_k\times n_k}$ , $E_2=\begin{bmatrix} D_1 & 0 & 0 & 0\\ 0 & \ddots & 0 & 0\\ 0 & 0 & D_c & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\in\mathbb{R}^{n\times n}$ with $D_k=\frac{1}{n_k}E^{[n_k]}$, and $E_3=\frac{1}{n}E_1$.Using these symbols and by some mathematical manipulation on Eq.(12), we have:

$$\arg min\Phi(A,B,U) = \underbrace{(1-\beta)\|X\cdot A-U\|_F^2+\beta\|Y\cdot B-U\|_F^2}_{Reconstruction\ term}+ \tag{13}$$

$$\mu\underbrace{Tr(U^T\cdot S_w\cdot U-\rho U^T\cdot S_b\cdot U)}_{Discriminative\ regularization\ term}$$

$$s.t.\ U^T\cdot U=I_d$$

where $Tr(U^T\cdot S_w\cdot U)$ and $Tr(U^T\cdot S_b\cdot U)$ denote the total within- class scatter and between- class scatter in the shared subspace $U$ , respectively. According to Lemma 3.1, the optimal solution to Eq.(13) can be sought by similarly solving a generalized eigenvalue problem with $Q=S_w-\rho S_b$ .

## 4. **Experimental Results and Analysis.**

### 4.1. **Datasets and Experiment Setup.**

In our experiment, the evaluations on the performance of the proposed model are carried out on two datasets : the UCI handwritten digit dataset and the commercial video shot dataset . The details for these two datasets are listed in TABLE 1 and TABLE 2 , respectively.

The UCI handwritten digit dataset contains ten handwritten digits 0-9 with 200 samples for each of digits. For each sample, six sets of features are flourier coefficient, contour correlation characteristics, Karhunen-Love expansion coefficient, pixel average, Zernike moment , and morphological characteristics. The commercial video shot dataset was originally built for automatic commercial detection [24]. We randomly select a subset of it for performance evaluation, which consists of 900 commercial shots and 2100 non-commercial shots. For each shot, both audio and visual features are extracted to form the multi-view representations.

TABLE 1. **Details of the UCI handwritten digit dataset**

| | UCI handwritten digit dataset | | | | | |
|---|---|---|---|---|---|---|
| Features | *mfeat- fou* | *mfeat- fac* | *mfeat- kar* | *mfeat- pix* | *mfeat- zer* | *mfeat- mor* |
| Dimension | 76 | 216 | 64 | 240 | 47 | 6 |

TABLE 2. **Details of the commercial video shot dataset**

| | Commercial video shot dataset | |
|---|---|---|
| Features | *mfeat-aud* | *mfeat-vis* |
| Dimension | 473 | 473 |

4.2. **Performance Evaluation on GRSS Model.** We first evaluate the performance of GRSS model on the retrieval task of multi-view UCI handwritten digit dataset. The compared shared subspace algorithms include *CCA*, *MORP*, *eCCA* and *GRSS*. In this experiment, 90% samples are randomly selected to construct the retrieval dataset and the remaining samples are used to serve as query instances. The retrieval performances on group (fac, fou) and group (fou, kar) are shown in Fig. 1 and 2 .Compared with *CCA*, *MORP*, and *eCCA*, *GRSS* achieves the best performance. Such results further validates the effectiveness of the local structure preservation involved in *GRSS*.

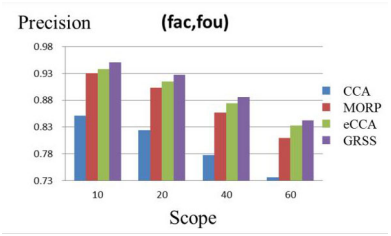


FIGURE 1. Retrieval performance comparisons on group (*mfeat-fac*, *mfeat-fou*)

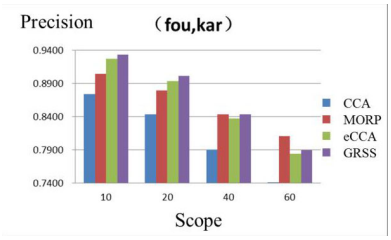


FIGURE 2. Retrieval performance comparisons on group (*mfeat-fou*, *mfeat-kar*)

4.3. **Performance Evaluation on DSS Model.** In essence, the proposed *DSS* model is a supervised shared subspace learning method. Consequently, in this experiment , we carry out performance evaluation on *DSS* model from the point of classification. For both the UCI handwritten digit dataset and the commercial video shot dataset, 80% samples are randomly selected to train the *DSS* model and the other 20% ones are used to construct the test dataset. Such process is repeated 10 times and the average classification accuracies are reported. For the UCI dataset, six groups of features are selected to form the multi-view data.

TABLE 3. **Classification performance comparisons on the UCI handwritten digit dataset**

| | X View | Y View | ***CCA*** | ***MORP*** | ***PLS*** | ***eCCA*** | ***DSS*** |
|---|---|---|---|---|---|---|---|
| Group 1 | mfeat-fac | mfeat-fou | 0.8970 | 0.9720 | **0.9805** | 0.9653 | 0.9760 |
| Group 2 | mfeat-fac | mfeat-kar | 0.9710 | 0.9663 | 0.9768 | 0.9703 | **0.9930** |
| Group 3 | mfeat-fou | mfeat-mor | 0.7870 | 0.7823 | 0.8020 | 0.7943 | **0.9850** |
| Group 4 | mfeat-fou | mfeat-pix | 0.8698 | 0.9728 | 0.9808 | 0.9713 | **0.9943** |
| Group 5 | mfeat-kar | mfeat-mor | 0.9110 | 0.9228 | 0.9203 | 0.9193 | **0.9928** |
| Group 6 | mfeat-kar | mfeat-zer | 0.8020 | 0.9595 | 0.9593 | 0.9560 | **0.9758** |

TABLE 4. **Classification performance comparisons on the UCI handwritten digit dataset**

| X View | Y View | ***CCA*** | ***MORP*** | ***PLS*** | ***eCCA*** | ***DSS*** |
|---|---|---|---|---|---|---|
| mfeat-aud | mfeat-vis | 0.9409 | 0.9457 | 0.9760 | 0.9810 | **0.9937** |

The classification results are illustrated in Table 3 and Table 4, respectively. As we can see, the proposed *DDS* model shows more powerful discriminative ability than other shared subspace methods on both two datasets. It is worth to note that *eCCA* is indeed a simplified version of *DSS* without employing discriminative regularization term (see Eq.(11) and Eq.(13)). The obvious advantage of *DSS* over *eCCA* shows the employment of discriminative information is much helpful for performance improvement.

4.4. **Impact of Parameter $\mu$.** As mentioned in Section 3.3, a parameter $\mu$ in both *GRSS* and *DSS* models is used to trade-off the reconstruction term and regularization term. Fig.3 shows the impact on retrieval performance on multi-view data group (fac, fou) by varying the parameter $\mu$ of *GRSS* . In this case, it can be found that $\mu = 1000$ gives rise to the best performance. As for the effect of parameter $\mu$ of *DSS* model, it is obvious from Fig.4 that $\mu = 0.5$ is the optimal.

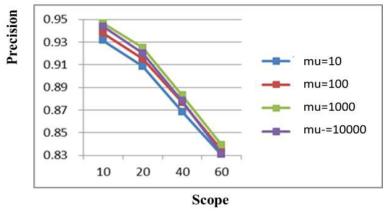


FIGURE 3. Impact of parameter $\mu$ of *GRSS* on retrieval performance

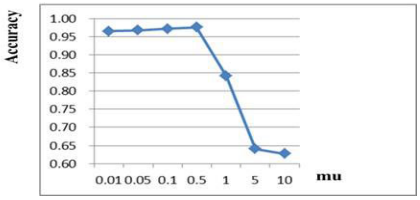


FIGURE 4. Impact of parameter $\mu$ of *DSS* on classification performance

5. **Conclusions.** The latent representation across multi-view has many potential applications. In this paper, we first propose a more general shared subspace learning model to capture the latent shared representation, which inherits nicely the merits of both the classical *CCA* and factorization based *MORP*. We also showed that the proposed shared subspace model is convex and to seek the global optimal solution comes down to solving a generalized eigenvalue problem. In order to well preserve the local structure of data in the both shared subspace and original multi-view feature space, a graph constraint is employed. Meanwhile, with the assist of prior class label, we extended the proposed general shared subspace model to a supervised one . Thus, the learned latent representation across multi-view can take much powerful discriminative ability. We carried out the experiments on the multi-view data retrieval and classification tasks to verify the effectiveness of the proposed model and showed promising results.

**Acknowledgment.** This work was supported in part by 973 Program (No.2012CB316400), National Natural Science Foundation of China (No.61025013, No.61172129), PCSIRT (No.IRT201206), Program for New Century Excellent Talents in University (No. 13-0661), and Fundamental Research Funds for the Central Universities (No.2012JBZ012). The authors would also like to thank anonymous reviewers for their constructive and valuable suggestions.

## REFERENCES


[1] A. Blum, T. M. Mitchell, Combining labeled, and unlabeled data with co-training, *Proc. of The 11th Annual Conference on Computational learning theory*, pp. 92-100, 1998.

[2] K. Nigam, R. Ghani, Analyzing the effectiveness, and applicability of co-training, *Proc. of the 9th International Conference on Information and Knowledge Management*, pp. 86-93, 2007.

[3] I. Muslea, S. Minton, and C. A. Knoblock, Active + semi-supervised learning = robust multi-view learning, *Proc. of the 19th International Conference on Machine Learning*, pp. 435-442, 2002.

[4] A. Kumar, H. Daumé III, A co-training approach for multi-view spectral clustering, *Proc. of International Conference on Machine Learning*, pp. 393-400, 2011.

[5] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, Learning the kernel matrix with semidefinite programming, *The Journal of Machine Learning Research*, vol. 5, pp. 27-74, 2004.

[6] F. R. Bach, G. R. G. Lanckriet and M. I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, *Proc. of the 21st international conference on Machine learning*, 2004.

[7] B. McFee, G. R. G. Lanckriet, Learning multi-modal similarity, *The Journal of Machine Learning Research*, vol. 12, pp. 491-523, 2011.

[8] H. Hotelling, Relation between two sets of variables, *Biometrika*, vol. 28, pp. 312-377, 1936.

[9] M. White, Y. Yu, X. Zhang, and D. Schuurmans, Convex multi-view subspace learning, *International Conference on Neural Information Processing Systems*, pp. 1682-1690, 2012.

[10] S. K. Gupta, D. Q. Phung, B. Adams, T. Tran, and S. Venkatesh, Nonnegative shared subspace learning and its application to social media retrieval, *Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1169-1178, 2010.

[11] S. Ji, L. Tang, S. Yu, and J. Ye, A shared-subspace learning framework for multi-label classification, *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 2, 2010.

[12] C. Xu, D. Tao, and C. Xu, A Survey on multi-view learning, *arXiv:1304.5634*, available at http://arxiv.org/abs/1304.5634., 2013.

[13] D. R. Hardoon, S. Szedmak, and J. S. Taylor, Canonical correlation analysis: An overview with application to learning methods, *Neural Computation*, vol. 16, no. 12, pp. 2639-2664, 2004.

[14] D. R. Hardoon, and J. S. Taylor, Convergence analysis of kernel canonical correlation analysis: theory and practice, *Machine Learning*, vol. 74, no. 1, pp. 23-38, 2009.

[15] L. Sun, S. Ji, and J. Ye, Canonical correlation analysis for multilabel classification: a least-squares formulation, extensions, and analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194-200, 2011.

[16] D. R. Hardoon, and J. Shawe-Taylor, Sparse canonical correlation analysis, *Machine Learning*, vol. 83, no. 3, pp. 331-353,2011.

[17] S. H. Lee, and S. Choi, Two-dimensional canonical correlation analysis, *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 735-738, 2007.

[18] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, Generalized multi-view analysis: a discriminative latent space, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2160-2167, 2012.

[19] S. Yu, K. Yu, V. Tresp, and H. P. Kriegel, Multi-output regularized feature projection, *IEEE Trans. Knowledge and Data Engineering*, vol. 18, no. 12, pp. 1600-1613, 2006.

[20] M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell, Factorized orthogonal latent spaces, *Proc. of International Conference on Artificial Intelligence and Statistics*, pp. 701-708, 2010.

[21] Z. Wang, C. Liu, T. Shi, and Q. Ding, Face-palm identication system on feature level fusion based on CCA, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 4, no. 4, pp. 272-279, 2013.

[22] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, Learning with local and global consistency, *Advances in Neural Information Processing Systems*, vol. 16, pp. 321-328, 2003.

[23] J. Chen, L. Tang, J. Liu, and J. Ye, A convex formulation for learning a shared predictive structure from multiple tasks, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1025-1038, 2013.

[24] N. Liu, Y. Zhao, and Z. Zhu, Exploiting visual-audio-textual characteristics for automatic TV commercial block detection and segmentation, *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 961-973, 2011.