Graph Regularized Multiview Marginal Discriminant Projection

Heng Pan[1,2], Jinrong He[1,2], Yu Ling[1,2], Lie Ju[1,2],Guoliang He[3]
[1] College of Informaiton Engineering, Northwest A&F University,712100 Yangling, China
[2] Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, 712100 Yangling, China
3 School of Computer, Wuhan University, 430070 Wuhan, China

Abstract. Multi-view data has become commonplace in today's computer vision applications, for the same object can be sampled through various viewpoints or by different instruments. The large discrepancy between distinct even heterogenous views bring the challenge of handling multi-view data. To obtain intrinsic common representation shared by all views, this paper proposes a novel multi-view algorithm called Multiview Marginal Discriminant Projection (MMDP), which is a supervised dimensionality reduction method for searching latent common subspace across multiple views. MMDP takes both inter-view and intra-view discriminant information into account and can preserve the global geometric structure and local discriminant structure of data manifold. Furthermore, the performance of MMDP is improved via imposing graph embedding as a regularization term to give a penalization of the local data geometric structure violation, which is called Graph regularized Multiview Marginal Discriminant Projection (GMMDP). The extensive experimental results on face recognition tasks demonstrate the effectiveness and robustness of MMDP and GMMDP. Finally, this paper excavates a new application scenario of multi-view learning and introduce it including the proposed GMMDP into solving hyperspectral image classification (HIC) problem, which leads to a satisfactory result.

Keywords: Marginal Discriminant Projection; Graph Laplacian; Manifold Regularization; Multiview Learning; Dimensionality Reduction; Hyperspectral Images Classification

1 Introduction

Multiview data has become commonplace in today's computer vision applications, for the same object can be sampled through various viewpoints or by different instruments. Compared to single-view data, the multi-view data contains more information of the observed object, but it not necessarily leads to a better result, since some information may be irrelevant and redundant for the classification tasks.

Therefore, it is essential to use effective multi-view learning methods to process multi-view data, which has become a hot issue for scholars to discuss and research in recent years. So far, various multi-view learning methods have been proposed and generally they can be divided into three categories [1]: 1) co-training [2], 2) multiple kernel learning [3], 3) common subspace learning [4]. Co-training is a semi-supervised learning method and Multiple kernel learning takes advantage of the kernel technique in dealing with non-linear pattern analysis problems. Common subspace learning aims to find a potential common subspace, then carry out the task of classification, cluster and so on after projecting each view into the low dimensional common subspace by adjusting the directions.

For a further step, common subspace learning can be divided into two classes: unsupervised and supervised. Canonical correlation analysis (CCA) [5, 6] is one of the most classical method in unsupervised learning whose goal is to maximize the correlation coefficient between two views' samples. However, CCA is limited by two views data. When it comes to multi-view problem, the Multiview Canonical Correlation Analysis (MCCA) [7,8] has been born, which aims to find v projection directions of v views, so that the sum of pair wise correlation coefficient will be largest after projection.

Moreover, substantial supervised multi-view learning methods have been proposed. A supervised multi-view learning framework Generalized Multiview Analysis (GMA) is proposed in [9], in which the discriminant information incorporated. Utilizing this framework, Linear Discriminant Analysis (LDA) [10] and Marginal Fisher Analysis (MFA) [11] can be extended to Generalized Multiview LDA (GMLDA) and Generalized Multiview MFA (GMMFA) respectively. Recently, Kan et al. proposed Multiview Discriminant Analysis (MvDA) [12], a multi-view learning method with consideration of both inter-view and intra-view. Then in [13], MvDA is improved to Mulitview Discriminant Analysis with View-Consistency (MvDA-VC). Based on MvDA, MvDA-VC focus on narrowing the distance between different views to improve the performance. In addition to the methods mentioned above, subspace learning still is the hotspot of multi-view learning, such as [14, 15] and so on.

To obtain intrinsic common representation shared by all views, this paper proposes a novel multi-view learning method based on Marginal Discriminant Projection (MDP) [16], a supervised linear dimensionality reduction algorithm that gives a new definition of data boundaries and obtains the low dimensional representation with the best discrimination performance. In [17], MDP is extended to Orthogonal Marginal Discriminant Projection (OMDP). Following the current multi-view strategy for cross-view recognition and MDP's work, this paper proposes the multi-view method MMDP, which can be concluded to a supervised method for searching latent common subspace across multiple views. MMDP takes both inter-view and intra-view discriminant information into account and can maintain the geometric construction of the data manifold.

Furthermore, recent studies have introduced spectral graph theory and manifold learning theory into dimensionality reduction field. In [18], LDA have be generalized to Graph Regularized LDA. In [19], Cai proposes Graph Regularized Nonnegative Matrix Factorization (GNMF), which explicitly considers the local invariance. Furthermore, hypergraph is used in [20]. Inspired by these works, we attempt to improve the performance of MMDP via imposing a graph as regularization term to give a penalization of the data geometric structure breaking. Thus, we generalize MMDP to GMMDP. The GMMDP's idea is shown in Figure 1. GMMDP calculates a Graph regularization term for each view and finds v directions to project all samples into a discriminant common subspace. After extracting the common information shared by all views, the subsequent work like classification and regression become easier to carry out.

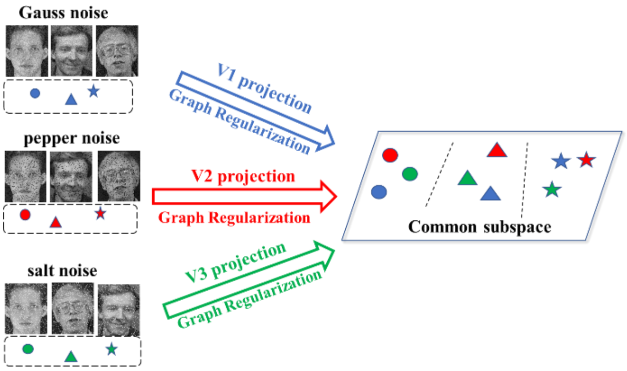


Figure 1. The basic ideal of GMMDP.
(Colors represent views and shapes represent classifications.)

Multi-view data do exist extensively in the real world, thus taking advantage of multi-view learning to handle these data is quite significant. Recently, many researches have come out to solve these. In [21, 22, 23], the idea of multi-view learning is applied to landmark search in the form of multi-modal. [24] learns from multiple social networks. [25, 26] bring it into the healthcare field, [14, 27] bring it into micro-videos and [15, 28] bring it into image memorability prediction.

In this paper, we excavate a new application scenario of multi-view learning and introduce it including our proposed GMMDP into solving hyperspectral image classification (HIC) problem. HIC is one of the key issues in the research of hyperspectral remote sensing technology that is widely used in agricultural planting, urban planning, mineral identification and other fields. However, the high dimensionality and redundant information lying in hyperspectral images tend to weaken the generalization ability of the classifier. Therefore, dimensionality reduction is often used to eliminate the inter-spectral correlation. Support vector machine (SVM) [29] is subsequently used to classify hyperspectral images after dimension reduction, which improves the accuracy significantly. However, the traditional spectral dimension reduction classification algorithm only considers the spectral features of pixels and ignores its spatial information. Recent years, HIC algorithms based on spatial spectral feature fusion have become a research hotspot. Neighborhood spatial feature, discrete Gabor transform [30], and discrete wavelet transform are very common in spatial feature extraction. There exist many kinds of spatial spectrum feature fusion strategies. For example, [31] proposed a hyperspectral decision fusion classification method based on PCA and windowed wavelet transform; [32] uses modified tensor locality preserving projection for hyperspectral images dimensionality reduction. With the research in deep learning field, [33, 34, 35] introduced convolutional neural networks into the HIC problem and achieved excellent realization.

We use the multi-view learning methods to study the hyperspectral image spectral-spatial feature problem by regarding the spectral features as one view, and the spatial features as another view. In the process of feature fusion, the redundant information of inter-view and intra-view are reduced simultaneously. We construct a framework to apply the multi-view learning to HIC as shown in Figure 2. Firstly, two views are constructed by extracting the spectral and spatial features from hyperspectral images, following normalization and PCA dimensionality reduction. Then multi-view learning method is applied to find a common subspace between views for classifying. Because of the same pixel in different views may be divided into different categories, we finally get the classification results after a decision fusion.

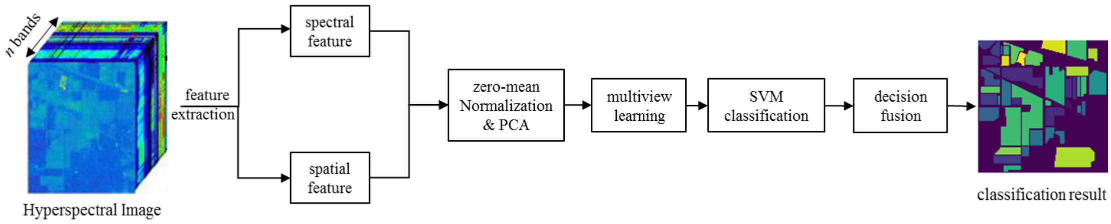


Figure 2 Overview of multi-view learning for hyperspectral classification

Totally speaking, our contribution in this paper can be summarized as follows:

(1) We propose a novel supervised multi-view learning algorithm named MMDP, which considers both inter-view and intra-view discriminant information and can maintain the geometric construction of the data manifold.

(2) We improve the performance of MMDP via imposing a graph as regularization term to give a penalization of the data geometric structure breaking and generalize MMDP to GMMDP.

(3) The extensive experiments on face recognition tasks have confirmed the effectiveness and robustness of the proposed MMDP and GMMDP methods.

(4) We excavate a new application scenario of multi-view learning and construct a framework to introduce multi-view learning into hyperspectral image classification problem, which leads to a satisfactory result.

The rest of this paper is organized as follows. Section 2 provides a review on related works. We elaborate on the proposed MMDP and GMMDP in Section 3. Then, performance evaluation and comparison on face recognition and hyperspectral images classification tasks is presented in Section 4. Finally, we conclude the paper in Section 5 and take an acknowledgement in Section 6.

2 Related Works

2.1 Multi-view Canonical Correlation Analysis

CCA [6] is a popular unsupervised dimensionality reduction technique which is similar to Principal Component Analysis (PCA), with an additional assumption that the data consists of feature vectors that arose from two sources (two views) that share some information. CCA aims to find two linear transforms w_1, w_2 and project the samples from two views into the common subspace for getting the maximal correlation between $w_1^T X_1$ and $w_2^T X_2$. Equation following as below:

$$\begin{gathered}\max_{w_1,w_2} w_1^T X_1 X_2^T w_2 \\ s.t. w_1^T X_1 X_1^T w_1 = 1, w_2^T X_2 X_2^T w_2 = 1\end{gathered} \tag{1}$$

where X_1 and X_2 are data matrix for each view. Eqn. 1 can be solved by resorting the eigenvalue decomposition with the Lagrange multiplier.

For multi-view cases, a pair-wise strategy which maximize the sum of correlations of each two views is feasible, thus CCA is extended to Multi-view CCA in [11] by finding v linear transforms $\{w_1, \ldots, w_v\}$ for v views, so the total correlation is maximal after projecting samples into a common subspace. The objective function of MCCA is formulated as follow:

$$\begin{gathered}\max_{w_1,\ldots,w_v} \sum_{i<j} w_i^T X_i X_j^T w_j \\ s.t. w_i^T X_i X_i^T w_i = 1, i = 1, 2, \ldots, v\end{gathered} \tag{2}$$

where $X_i \in R^{d_i \times n}$ is the data matrix of the i^{th} view with n samples of d_i dimension.

2.2 Generalized Multiview Analysis

GMA is a general multi-view feature extraction approach proposed in [12], who has some properties such as supervised, generalizable, multi-view, efficient, kernelizable and Domain-Independent. GMA exploits the fact that most popular supervised and unsupervised feature extraction techniques are the solution of a special form of a quadratic constrained quadratic program (QCQP), which can be solved efficiently as a generalized eigenvalue problem. GMA solves a joint, relaxed QCQP over different feature spaces to obtain a single (non)linear subspace. Intuitively, GMA is a supervised extension of Canonical Correlational Analysis (CCA), which is useful for cross-view classification and retrieval. Defining A_i as the between-class scatter matrix and B_i as the within-class scatter matrix, then introduce the balance parameters $\mu_i, \lambda_i, \gamma_i$, GMA's objective function is:

$$\begin{gathered}\max_{w_1,\ldots,w_v} \sum_{i=1}^{v} \mu_i w_i^T A_i w_i + \sum_{i<j} 2\lambda_{ij} w_i^T X_i X_j^T w_j \\ s.t. \sum_i \gamma_i w_i^T B_i w_i = 1\end{gathered} \tag{3}$$

where $X_i \in R^{d_i \times n}$ is the data matrix of the i^{th} view with n samples of d_i dimension. Eqn. 3 can be ultimately derived into this formula:

$$\tilde{A}\hat{w} = \lambda \tilde{B}\hat{w} \tag{4}$$

which is a standard generalized eigenvalue problem solved by any eigen-solver. GMA is a general framework for multi-view analysis which can be applied on many single-view methods such as LDA [13] and MFA [14].

2.3 Multiview Discriminant Analysis

In [16], Kan proposes the MvDA approach, which seeks for a discriminant common subspace for multiple views in a non-pairwise manner by jointly learning multiple view-specific linear transform. MvDA considers both inter-view and intra-view variations leading to a more discriminative common space where the between-class variations are maximum. Specifically, MvDA is formulated to jointly solve the multiple linear transforms by optimizing a generalized Rayleigh quotient. For v views, MvDA can be formulated as:

$$(w_1^*, w_2^*, ..., w_v^*) = \arg\max_{w_1,...,w_v} Tr(\frac{W^T DW}{W^T SW}) \tag{5}$$

that $W = [w_1^T, w_2^T, \dots, w_v^T]^T$ and $w_i \in R^{d_i \times r}$ where d_i is the dimension of i^{th} view and r is the subspace dimension. S is a block matrix and S_{jr} is defined as below with $\mu_{ij}^{(x)} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} x_{ijk}$:

$$S_{jr} = \begin{cases} \sum_{i=1}^{c} (\sum_{k=1}^{n_{ij}} x_{ijk} x_{ijk}^T - \frac{n_{ij} n_{ij}}{n_i} \mu_{ij}^{(x)} \mu_{ij}^{(x)T}), j == r \\ -\sum_{i=1}^{c} \frac{n_{ij} n_{ir}}{n_i} \mu_{ij}^{(x)} \mu_{ir}^{(x)T}, otherwise \end{cases} \tag{6}$$

D is a block matrix as well and D_{jr} is defined as below:

$$D_{jr} = (\sum_{i=1}^{c} \frac{n_{ij} n_{ir}}{n_i} \mu_{ij}^{(x)} \mu_{ir}^{(x)T}) - \frac{1}{n} (\sum_{i=1}^{c} n_{ij} \mu_{ij}^{(x)}) (n_{ir} \mu_{ir}^{(x)})^T \tag{7}$$

MvDA can be solved analytically through generalized eigenvalue decomposition.

3 Proposed Method

To facilitate the reading, we summarize the main symbols of this section with a notation table as follows. Other mathematical symbols are defined in concrete contexts. Usually lowercase letters represent variables or vectors, and uppercase letters represent matrices.

table 1 notation table of proposed method

symbol	description	symbol	description	symbol	description
x	feature vector of sample	y	label vector	n	number of samples
d	dimension of samples	v	number of views	c	number of classes
r	dimension after projection	X	data matrix	V	transform matrix
N	sum of samples from all views	I	identity matrix	W	weight matrix
D	degree matrix	L	Laplacian matrix	S	scatter matrix

3.1 Multi-view Marginal Discriminant Projection

Based on MDP's work, this paper proposes a multi-view learning approach MMDP which extend the original MDP to acquire a capability of dealing with multiple views problem. Before we present the detail formulation of MMDP, we introduce some basic definition given in [14]. For any of two samples x_i and x_j, their distance is $dist(x_i, x_j) = \|x_i - x_j\|_2$. For each class c_i, we define the within-class distance as $dist(c_i) = dist(x_a{}^i, x_b{}^i)$ that $x_a{}^i, x_b{}^i \in c_i$ and $dist(x_a{}^i, x_b{}^i) \geq dist(x_i, x_j), \forall x_i, x_j \in c_i$. And for each pair c_i and c_j, we define the between-class distance as $dist(c_i, c_j) = dist(x_j{}^i, x_i{}^j)$ that $x_j{}^i$ belong to class c_i, $x_i{}^j$ belong to class c_j and $dist(x_j{}^i, x_i{}^j) \geq dist(x_i, x_j), \forall x_i \in c_i, \forall x_j \in c_j$.

For single view, MMDP expects when data projected into the low-dimensional subspace, samples of same class will be closer, and samples of different classes will be farther. In another word, MMDP wants to minimize the sum of within-class distance and maximize the sum of between-class distance, and define it as the boundary *J* of the view which is kind of different with original MDP:

$$J = \frac{\sum_{i \neq j} dist(x_i{}^j, x_j{}^i)}{\sum_{i=1}^{c} dist(c_i)} \tag{8}$$

MMDP aims to make *J* of each view maximum after projection. For the sake of maintaining the global geometrical structure of data, we introduce an orthogonal constraint that $V^T V = I$, and the objective function can be described as follows:

$$V = \arg\max_{V^T V = 1} \frac{\sum_{i \neq j} \|V^T x_j^i - V^T x_i^j\|_2^2}{\sum_{i=1}^{c} \|V^T x_a^i - V^T x_b^i\|_2^2} \tag{9}$$

where *V* is the transform matrix MMDP seeks for. To get a simple expression, we boil it down to a graph embedding framework and introduce between-classes similarity weight matrix $W^{(b)}$ and within-class similarity weight matrix $W^{(w)}$:

$$\mathrm{W}_{ij}^{(b)} = \begin{cases} 1, & \text{samples are } x_i^j \text{and } x_j^i \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$\mathrm{W}_{ij}^{(w)} = \begin{cases} 1, & \text{samples are } x_a^i \text{and } x_b^i \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Therefore, Eqn. 9 can be reformulated as follows:

$$V = \arg\max_{V^T V = 1} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \|V^T x_i - V^T x_j\|^2 W_{ij}^{(b)}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \|V^T x_i - V^T x_j\|^2 W_{ij}^{(w)}} \tag{12}$$

that Eqn5 can be simplify as follows:

$$V = \arg\max_{V^T V = 1} tr\left(\frac{V^T X L^{(b)} X^T V}{V^T X L^{(w)} X^T V}\right) \tag{13}$$

where $L^{(b)} = D^{(b)} - W^{(b)}, L^{(w)} = D^{(w)} - W^{(w)}$ is Laplacian Matrix.

Then we talk about *v* views case. Firstly, we hope the summation of each view boundary J_i will be maximal and get the preliminary objective function:

$$[\hat{V}_1, \ldots, \hat{V}_v] = \arg\max \sum_{i=1}^{v} tr\left(\frac{V_i^T S_i^{(b)} V_i}{V_i^T S_i^{(w)} V_i}\right) \tag{14}$$

$\hat{V}_i$ is the optimal projection direction of i^{th} view, and $S_i^{(b)} = X_i L_i^{(b)} X_i^T, S_i^{(w)} = X_i L_i^{(w)} X_i^T$. Eqn. 14 can be cast as a form of quadratically constrained quadratic program (QCQP) and be reformulated as follows:

$$[\hat{V}_1,...,\hat{V}_v] = \arg\max \sum_{i=1}^{v} V_i^T S_i^{(b)} V_i, \quad s.t. V_i^T S_i^{(w)} V_i = 1, i = 1,2,...,v \tag{15}$$

Now we introduce a constraint for the relationship of the samples in different views. We hope to maximize covariance between the samples from different views. This leads to a closed form solution and better preserve the between class variation as argued in [9]:

$$[\hat{V}_1,...,\hat{V}_v] = \arg\max_{i \neq j} \sum_{i=1}^{v} \sum_{j=1}^{v} V_i^T X_i X_j^T V_j \tag{16}$$

here X_i is the data matrix of i^{th} view.

Combining with Eqn. 15 and Eqn. 16, we ultimately get the objective function of MMDP:

$$[\hat{V}_1,...,\hat{V}_v] = \arg\max \sum_{i=1}^{v} V_i^T S_i^{(b)} V_i + \sum_{i=1}^{v} \sum_{j=1}^{v} a_{ij} V_i^T X_i X_j^T V_j, \quad s.t. V_i^T S_i^{(w)} V_i = 1, i = 1,2,...,v \tag{17}$$

The matrix form of Eqn. 17 is:

$$\hat{V} = \arg\max \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix}^T \begin{bmatrix} S_1^{(b)} & \cdots & \alpha_{1v} X_1 X_v^T \\ \vdots & \ddots & \vdots \\ \alpha_{v1} X_v X_1^T & \cdots & S_v^{(b)} \end{bmatrix} \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix}, \quad s.t. \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix}^T \begin{bmatrix} S_1^{(w)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S_v^{(w)} \end{bmatrix} \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix} = 1 \tag{18}$$

Here, $\hat{V} = [\hat{V}_1, \cdots, \hat{V}_v]$. Eqn. 18 can be equivalently simplified as follows where $\tilde{A}$ and $\tilde{B}$ are the square symmetric matrix:

$$\hat{V} = \arg\max V^T \tilde{A} V, \quad s.t. V^T \tilde{B} V = 1 \tag{19}$$

This is a standard generalized eigenvalue problem that can be solved using any eigen-solver.

3.2 Graph Regularized Multiview Marginal Discriminant Projection

Recent studies in spectral graph theory [36] and manifold learning theory [37] have demonstrated that the local geometric structure can be effectively modeled through a nearest neighbor graph on a scatter of data points. Inspired by prior work in graph regularization, we generalize our MMDP to consider the geometric relationship via adding a regularization term to each view and name this model GMMDP.

We can construct an affinity graph per view where each vertex corresponds to a data point. Graph-building needs to consider two aspects: whether putting an edge between two points and what is the weight of the edge. For first problem, we only put edges between two homogenous data points, which considers the discriminant structure of data. For second problem, there are three ways for computing weight of edge W_{uv} [19]:

0-1 Weighting: if nodes *u* and *v* are connected, $W_{uv} = 1$.

Heat Kernel Weighting: if nodes *u* and *v* are connected, $W_{uv} = e^{-\frac{|x_u - x_v|^2}{\theta}}$. Heat Kernel has an intrinsic connection to the Laplace Beltrami operator on differentiable functions on a manifold.

Dot-product Weighting: if nodes *u* and *v* are connected, $W_{uv} = x_u^T x_v$.

In this paper, we choose the second add-edge scheme and choose heat kernel weighting as weighting scheme. There is a parameter θ in heat kernel weighting and we make it self-adapting by:

$$\theta = \frac{1}{n_{c_i}} \sum_{i \neq j} dist(x_i, x_j) \tag{20}$$

, where x_i, x_j belong to same class, n_{c_i} is the number of i^{th} class. Simply speaking, θ is equal to the mean value of all distance between each pair points belonging to same class. This weighting scheme has an intuitionistic geometric interpretation, that the distance between two points is closer, the weight of the edge is heavier and vice versa.

After finishing the nearest neighbor graph, we can use the following regularization term for each view to preserve the geometric information:

$$R_v = \frac{1}{2}\sum_{c \in C}\sum_{i,j \in c}(y_i - y_j)^2 W_{ij}^c = V_v^T X_v (D_v - W_v) X_v^T V_v = V_v^T X_v L_v X_v^T V_v = V_v^T G_v V_v \tag{21}$$

where X_v is the data matrix of v^{th} view and D_v is a diagonal matrix whose entries are column sum of W_v. Therefore, $L_v = D_v - W_v$ is a Laplacian matrix. By minimizing the regularization term, we can get the objective function of each view as following:

$$V_i = \arg\max tr\left(\frac{V_i^T S_i^{(b)} V_i}{V_i^T (S_i^{(w)} + \lambda G_i) V_i}\right) \tag{22}$$

where the $\lambda \geq 0$ is a parameter controls the smoothness of geometric information preserving.

Combining Eqn11 and Eqn. 22, it leads to the final objective function of GMMDP as following:

$$\hat{V} = \arg\max \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix}^T \begin{bmatrix} S_1^{(b)} & \cdots & \alpha_{1v} X_1 X_v^T \\ \vdots & \ddots & \vdots \\ \alpha_{v1} X_v X_1^T & \cdots & S_v^{(b)} \end{bmatrix} \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix}, \; s.t. \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix}^T \begin{bmatrix} S_1^{(w)} + \lambda G_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S_v^{(w)} + \lambda G_v \end{bmatrix} \begin{bmatrix} V_1 \\ \vdots \\ V_v \end{bmatrix} = 1 \tag{23}$$

which can be solved using any eigen-solver.

The algorithm of GMMDP can be summarized as:

Input: $[X_1, X_2, ..., X_v]$: v data matrices, each column per sample.

$[y_1, y_2, ..., y_v]$: v label vectors.

r: embedding dimensionality

Output: $[V_1, V_2, ..., V_v]$: v projection matrices.

Algorithm:

1. Calculate between-classes similarity weight matrix $S_i^{(b)}$ and within-class similarity weight matrix $S_i^{(w)}$ for each view.
2. Calculate Laplacian matrices of $S_i^{(b)}$ and $S_i^{(w)}$.
3. Build heat kernel weighting graph and calculate Laplacian matrix for each view.
4. Calculate Eqn. 23 and eigenvalue decomposition, select r eigenvectors according to r biggest eigenvalues.

Complexity analysis

In step 1, calculation of $W_i^{(b)}$ and $W_i^{(w)}$ involve calculating distance between each sample, whose complexity is $O(dn^2)$. Thus, the computation complexity of $S_i^{(b)}$ and $S_i^{(w)}$ is $O(dn^2)$.

In step 2, Laplacian matrices of $S_i^{(b)}$ and $S_i^{(w)}$ can be calculated by $O(n^2)$.

In step 3, the computation complexity of graph regularization term is $O(n^2)$ because the distances have been calculated in step 1.

In step 4, the computation complexity of eigenvalue decomposition is $O(N^3)$. The efficiency of GMMDP is mainly subject to this step.

4 Experiments

4.1 Parameters setting

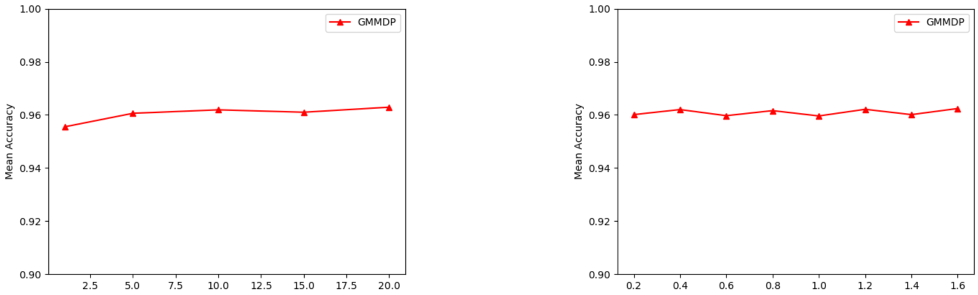
Generally, there are some parameters of our proposed methods need to be setup before further work, we have parameters α_{ij}, $1 \leq i, j \ \leq v$ in both MMDP and GMMDP and one more parameter λ in GMMDP. We determine these parameters experimentally with the Multi-PIE dataset described in Section 4.2.1. To simplify the adjusting work, each α_{ij} shares the same value and then we only have two parameters α and λ to test. Experiments are carried on with control variate method that one parameter fixed and the others varying. We achieve the results under 75% training set, 100 embedding dimensionality and k-NN classifier which are shown in Table 2, Table 3 and Figure 3.

Table 2 mean accuracy on Multi-PIE with alpha varying

alpha	***1***	***5***	***10***	***15***	***20***
accuracy	95.55%	96.06%	96.19%	96.10%	96.29%

Table 3 Mean accuracy on Multi-PIE with lambda varying

lambda	***0.2***	***0.4***	***0.6***	***0.8***	***1***	***1.2***	***1.4***	***1.6***
accuracy	96.01%	96.20%	95.97%	96.16%	95.96%	96.21%	96.01%	96.24%



a) mean accuracy with alpha varying **b) mean accuracy with lambda varying**

Figure 3 mean accuracy with parameters varying

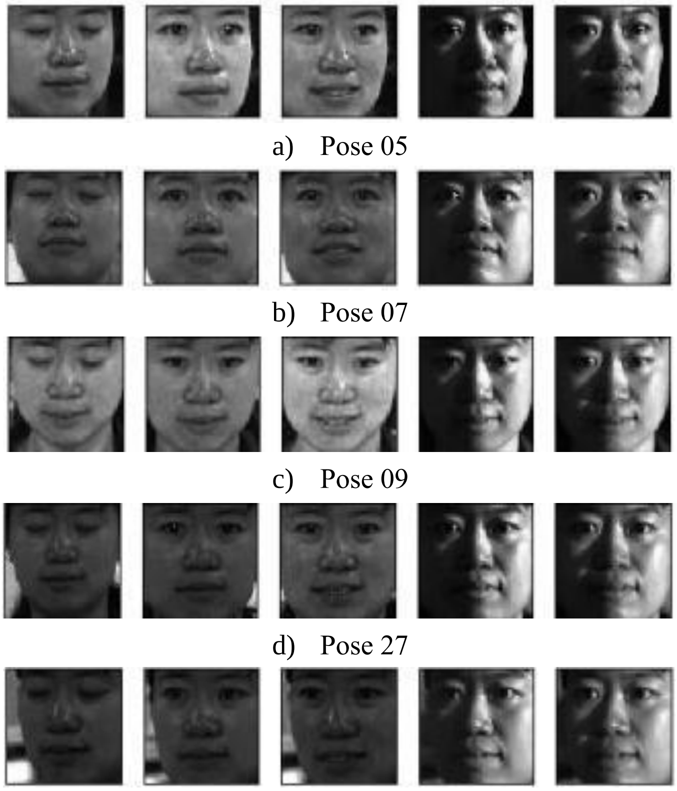
From Table 2, Table 3 and Figure 3 can we find that the change of parameters only causes subtle fluctuations to the classification results. For the sake of simplicity and generalization ability, we fix parameters $\alpha_{\mathrm{ij}} = 10$, $1 \leq i, j \ \leq v$ in MMDP and $\lambda = 1$ for GMMDP in both face recognition and hyperspectral images classification tasks.

4.2 Face recognition tasks

In this section, MMDP and GMMDP are evaluated on two face recognition tasks which are composed of multiple view data, i.e., face recognition across as multiple angles and multiple noise.

4.2.1 Datasets

The CMU Multi-PIE face database contains more than 750,000 images of 337 people recorded in up to four sessions over the span of five months. To test our GMMDP, we choose five views (pose05, pose07, pose09, pose 27, pose 29) from Multi-PIE, and for each pose we select 1632 subjects' face images of 64 people and 24 face images per people. Every image is 64×64 pixels in 256 levels and vary with facial expression and lighting difference. Part of the data set is shown in Figure 4.



a) Pose 05

b) Pose 07

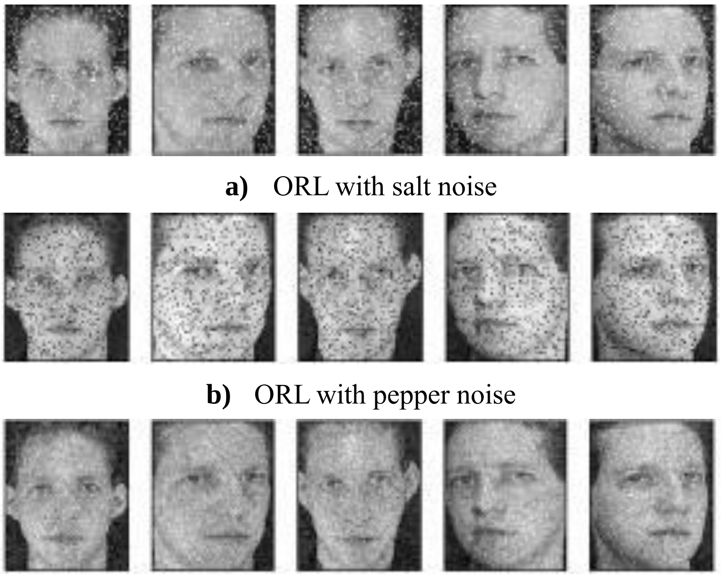
c) Pose 09

d) Pose 27

e) Pose 29

Figure 4 Portion data of Multi-PIE

The ORL Database of Faces contains 10 different images of each of 40 distinct subjects. The resolution of each image is 92×112 pixels, with 256 grey levels per pixel. In the real world, we may get the image data interfered by noise for kinds of reasons. To measure the robustness of GMMDP with dealing with noisy data, we construct the multi-view data through adding different type of noise into the ORL face dataset, thus get three views as shown in Figure 5, which mingle salt noise, pepper noise and Gauss noise.



a) ORL with salt noise

b) ORL with pepper noise

c) ORL with Gauss noise

Figure 5 Portion data of ORL

4.2.2 Experimental settings

In all face recognition experiments, we use principal component analysis (PCA) [38] to embed dimensionality for reducing the complexity and use Leave-One-Out method to split data set, which randomly select the same scale samples in every class and every view. We use the train set to find v projection directions by multi-view learning methods and then project original data. Finally, we use the train set in common subspace to train classifier and test set to predict. k-Nearest Neighbor (k-NN) [39] is chosen as the classifier. Besides, for GMMFA approach, we fix $\alpha = 10$, $\mu = 1, \gamma = 1, k_1 = 10, k_2 = 5$. Experiments are repeated 20 times and we calculate the average value as mean accuracy.

4.2.3 Face Recognition across as multiple angles

Face recognition across as multiple pose is evaluated on Multi-PIE dataset which take each pose as a view, thus totally five views are used. We get the recognition results shown in Table 4 and Figure 6a. Merge-view means directly merge multiple views into a single view to train the classifier, which is obviously not effective can we see in the diagram. There is a 26.67% discrepancy with the multi-view method at the maximum. MCCA behaves poorly by other multi-view methods for the reason of not making use of discriminated information. MvDA and GMMFA have equal shares of performance on this dataset. Our MMDP outperforms MvDA and GMMFA in most of objective dimensional subspace. GMMDP slightly improve the performance of MMDP and gains about 2% over other existing discriminated multi-view methods, for the sake of geometric information preserving.

4.2.4 Face Recognition across as multiple noise

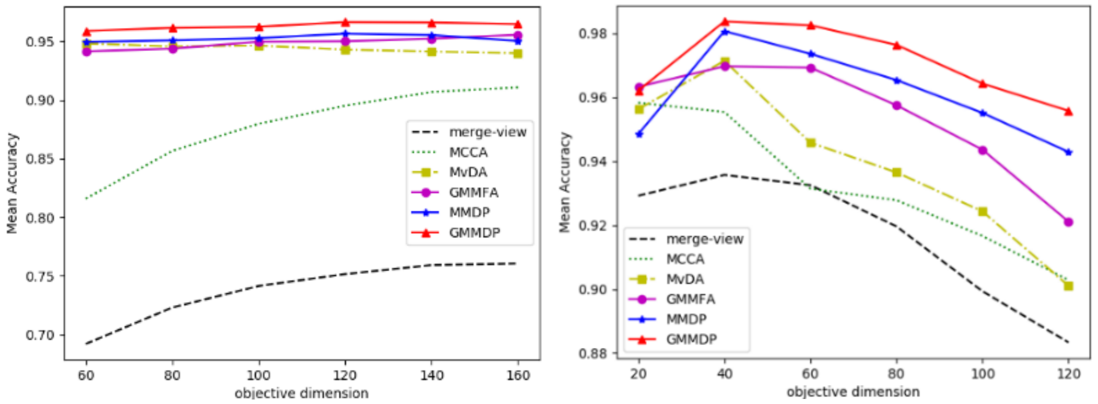
In this experiment, the results are shown in Table 5 and Figure 6b. The situation is similar to the previous experiment while the mean accuracy continues the downward trend with the embedding dimensions increasing. GMMDP shows a better performance among all multi-view algorithms. The experiment once again verifies the role that the graph regularization term played on enhancing performance. Through data is influenced by different kinds of noise, GMMDP still keeps an outstanding classification ability.

table 4 Mean Accuracy on Multi-PIE with different embedding dimensionalities

Method	***60***	***80***	***100***	***120***	***140***	***160***
Merge-view	69.20%	72.30%	74.14%	75.15%	75.92%	76.06%
MCCA	81.58%	85.65%	87.96%	89.51%	90.67%	91.07%
MvDA	94.81%	94.55%	94.63%	94.29%	94.12%	93.99%
GMMFA	94.13%	94.37%	94.96%	95.00%	95.22%	95.55%
MMDP	94.94%	95.08%	95.26%	95.65%	95.53%	95.03%
GMMDP	**95.87%**	**96.15%**	**96.24%**	**96.63%**	**96.60%**	**96.46%**

table 5 Mean Accuracy on ORL with different embedding dimensionalities

Method	***20***	***40***	***60***	***80***	***100***	***120***
Merge-view	92.92%	93.57%	93.25%	91.96%	89.93%	88.33%
MCCA	95.83%	95.54%	93.14%	93.60%	94.71%	93.61%
MvDA	95.63%	97.14%	94.58%	93.65%	92.43%	90.10%
GMMFA	**96.33%**	96.97%	96.93%	95.75%	94.36%	92.11%
MMDP	94.85%	98.07%	97.36%	96.54%	95.51%	94.29%
GMMDP	96.21%	**98.26%**	**98.22%**	**97.54%**	**96.43%**	**95.53%**



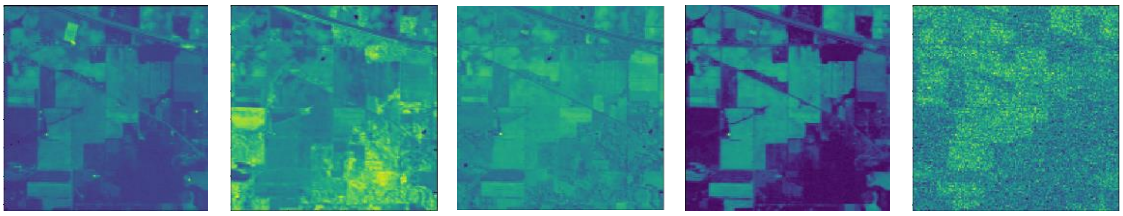
(a) Multi-PIE (b) ORL

Figure 6. Mean Accuracy with different embedding dimensionalities

4.3 Hyperspectral images classification tasks

4.3.1 Dataset

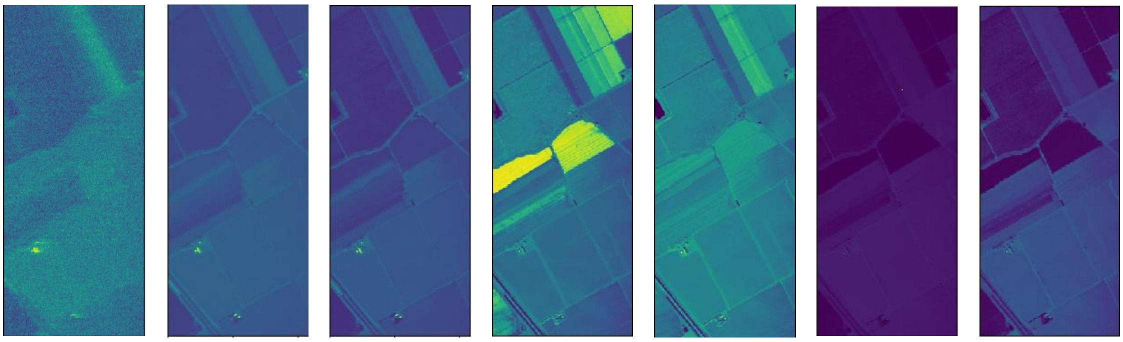
Indian Pines dataset comes from the hyperspectral data of 200 bands acquired by AVIRIS sensors in Indiana. The image size of each band is 145 × 145 pixels and the spatial resolution is 20 meters. The original Indian Pines dataset has 16 categories. The images of Indian Pines in different bands are shown as Figure 7. In this paper, 9 categories with more pixels are selected, and a total of 9234 samples are used for experiments. The experiment used the Leave-One-Out method to divide the data, with 15% of the data used for training and 85% of the data used for testing.



a) 10 b) 50 c) 100 d) 150 e) 200

Figure 7 The images of Indian Pines in different bands

Salinas Valley dataset is derived from the 224 bands of hyperspectral data acquired by the AVIRIS sensor in California, each with an image size of 512 x 217 pixels and a spatial resolution of 3.7 meters. The images of Salinas Valley in different bands are shown as Figure 8. The data has 16 categories, 54129 samples, and the experiment also uses the leave-out method to divide the data, 15% of which is used for training and 85% of the data is used for testing.



a) 1 b) 5 c) 10 d) 50 e) 100 f) 150 g) 200

Figure 8 The images of Salinas Valley in different bands

4.3.2 The multi-view features of hyperspectral images

Spectral View. A hyperspectral image consists of 3-dimensional data, where each pixel is a vector. We see $X = \{x_1, x_2, \dots, x_m\}$ as the spectral vector set, here $x_i \in R^{d\times 1}$ is a spectral feature vector of a sample pixel, m is the number of samples and d is the number of bands. In the hyperspectral image classification, PCA is usually performed to eliminate the inter-spectral correlation and reduce computational complexity. The set of all spectral vectors consists the spectral view.

Spatial View. With the deepening of the research on the real spatial distribution of ground objects, more and more spatial feature extraction methods have been proposed, including neighborhood spatial features, Gabor transform, wavelet transform, morphological profile [40] etc. This paper discusses two spatial feature extraction strategies to get the spatial view for subsequent multi-view learning.

1) *Neighborhood Spatial Features*. A naive spatial feature extraction method takes the target pixel as center and chooses $n \times n$ neighborhoods' spectral feature to form the spatial features:

$$k_i = [x_1^T, x_2^T, \dots, x_{n\times n-1}^T]^T \tag{24}$$

here $k_i \in R^{((n\times n-1)\times d)\times 1}$. However, this method greatly improves the dimensionality of the spatial features, which not only increases computational complexity, but also leads to over-fitting in the classification.

Referring to the idea of graph construction in manifold learning, we first calculate the spatial influence coefficient of each pixel in the $n \times n$ neighbor regrading to the target pixel, getting $w_j = e^{-\frac{|x_i - x_j|^2}{\theta}}$ with $\theta = \frac{1}{n\times n-1}\sum_{j=1}^{n\times n-1}|\boldsymbol{x}_i - \boldsymbol{x}_j|^2$. Then the spatial feature is defined as:

$$k_i = \frac{1}{\sum_{j=1}^{n\times n-1} w_j} \sum_{j=1}^{n\times n-1} w_j x_j \tag{25}$$

here $k_i \in R^{d\times 1}$, and obtain the spatial view $K = \{k_1, k_2, \dots, k_m\}$. The above spatial features have an intuitive geometric interpretation that the neighborhood pixels where the far contribute little and the near contribute much to the spatial feature.

2) *Discrete Wavelet Features*. Frequency domain transform is a commonly used method in image processing. Wavelet transform provides local analysis and refinement capabilities. The spatial feature can be obtained by performing local discrete wavelet transform on the images per hyperspectral bands. Centering on the target pixel, a $2n\times 2n$ window is selected, then a discrete wavelet transform is performed on the elements in the window to obtain wavelet coefficients LL_p、LH_p、HL_p、HH_p. Since the low-frequency components in the spectrum concentrate the energy of the image, which represents the overview of the image, LL_p can be selected as the spatial feature of the target pixel under the band p, and the final spatial feature of the target pixel is obtained thorough stacking the features under each band.

4.3.3 Settings

In the experiment, neighborhood spatial method and wavelet transform method are used to obtain spatial information as comparisons. After that, standard deviation normalization and PCA are performed on both spectral and spatial data. To reduce the parameters, experiments keep the subspace dimension consistent with the PCA dimension. Because the spectral and spatial features of pixels are treated as two data samples, there are cases where the same pixel is divided into distinct categories. We use spatial result as the final classification result, because in the image, spatially geometrically adjacent pixels are more likely to belong to the same class while spectral similar pixels are not necessarily the same substance due to the phenomenon of “different things with the same spectrum”. The parameters of GMMDP are same with the ones which fixed in face recognition tasks.

Experiments use MvDA and GMMFA as the comparison algorithm for multi-view learning and use traditional LDA-MLE and SVM-RBF as single view comparison algorithms. Gaussian kernel is uniformly selected for SVM as the classifier. In addition to the accuracy rate, we also use Kappa coefficients to measure the performance of the algorithm. Kappa is a discrete multivariate method that is widely used to evaluate the classification accuracy and error matrix of remote sensing images. Kappa is defined as follows.

$$Kappa = \frac{p_o - p_c}{1 - p_c} \tag{26}$$

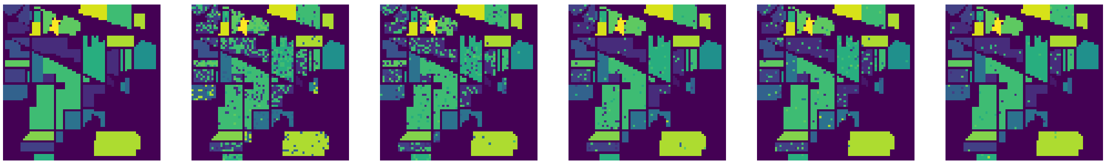
Here p_o is observed accuracy and p_c is chance agreement. Experiments are repeated 20 times and the mean value is taken.

4.3.4 Results and Discussion

The overall classification accuracy and Kappa coefficient on Indian Pines dataset are shown in Table 6 where traditional single-view methods and multi-view methods are compared. From Table 6 can we find that the multi-view learning significantly improves the accuracy of hyperspectral images classification, no matter which spatial feature extraction method is used. GMMDP using neighborhood spatial features outperforms traditional LDA-MLE with 35.22% accuracy higher and 42.17% Kappa higher as well as outperforms SVM-RBF with 18.79% accuracy higher and 22.49% Kappa higher. This illustrates the importance of spatial information on the one hand and proves that multi-view learning can improve the classification accuracy of hyperspectral images significantly on the other hand. Among supervised multi-view learning algorithms, although the gap is not large, GMMDP still achieves the best accuracy with 96.58% and 96.48%, and so does Kappa. In addition, the accuracy of multi-view learning will be influenced by spatial feature extraction methods. Neighborhood spatial extraction method proposed in our paper is better than the wavelet features no matter of accuracy or Kappa. Because it considers the different contribution of each surrounding pixel, but wavelet features not. Figure 9 visually shows the different performance during various classification algorithms.

Table 6 The classification result for Indian Pines dataset

	LDA-MLE	SVM-RBF	Spatial Feature	MvDA	GMMFA	GMMDP
Accuracy	61.36%	77.79%	neighborhood	96.33%	93.63%	**96.58%**
			wavelet	85.48%	83.11%	**96.48%**
Kappa	53.81%	73.49%	neighborhood	95.69%	92.50%	**95.98%**
			wavelet	82.68%	79.93%	**96.00%**



(a)ground truth (b)LDA-MLE (c)SVM-RBF (d)MvDA (e)GMMFA (f)GMMDP

Figure 9 Classification maps for Indian Pines dataset

The experimental results on Salinas Valley dataset are shown in Table 7. The performs of GMMDP using neighborhood spatial features is better than the traditional LDA-MLE method that accuracy is 8.8% higher and Kappa is 9.83% higher. Compare to SVM-RBF, GMMDP gets 5.93% accuracy higher and 6.63% Kappa higher. Although GMMDP underperforms slightly MvDA when use neighborhood spatial features, it still shows the best performance using wavelet features. Figure 10 visually shows the different performance during various classification algorithms.

Table 7 The classification result for Salinas Valley dataset

	LDA-MLE	SVM-RBF	Spatial Feature	MvDA	GMMFA	GMMDP
Accuracy	89.84%	92.71%	neighborhood	**98.95%**	98.47%	98.64%
			wavelet	95.94%	96.24%	**96.29%**
Kappa	88.66%	91.86%	neighborhood	**98.83%**	98.30%	98.49%
			wavelet	95.47%	95.80%	**95.85%**

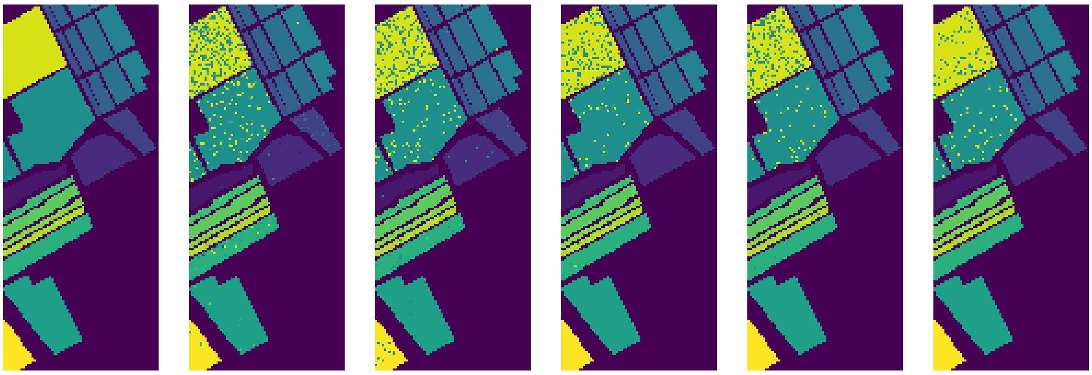

(a)origin (b)LDA-MLE (c)SVM-RBF (d)MvDA (e)GMMFA (f)GMMDP

Figure 10 Classification maps for Salinas Valley dataset

5 Conclusion

To address the multiple views data, this paper proposes a novel multi-view method MMDP, which takes both inter-view and intra-view discriminant information into account and can maintain the geometric construction of the data manifold. For a further work, we generalize MMDP to GMMDP via considering the geometric relationship. The following two experiments on face recognition across as multiple angles and multiple noise verified its effectiveness and robustness on finding latent common subspace. Besides, we excavate a new application scenario of multi-view learning and construct a framework to deal with HIC problem. Experiments on the open dataset shows that multi-view learning has a significant advantage in spectral-spatial classification of hyperspectral images. Among all multi-view dimensionality reduction algorithms, our GMMDP has the highest classification accuracy.

6 Acknowledgement

This work was partially supported by China Postdoctoral Science Foundation (No. 2018M633585), Natural Science Basic Research Plan in Shaanxi Province of China (No. 2018JQ6060), the Doctoral Starting up Foundation of Northwest A&F University (No.2452015302), and Students Innovation Training Project of China (No. 201710712064).

References

1. Zhao J, Xie X, Xu X, et al. Multi-view learning overview: Recent progress and new challenges[J]. Information Fusion, 2017, 38(11):43-54.
2. Blum A. Combining labeled and unlabeled data with co-training[C]// Eleventh Conference on Computational Learning Theory. ACM, 1998:92-100.
3. Wang H Q, Sun F C, Cai Y N, et al. On multiple kernel learning methods[J]. Acta Automatica Sinica, 2010, 36(36):1037-1050.
4. Yang P, Huang K, Liu C L. A multi-task framework for metric learning with common subspace[J]. Neural Computing & Applications, 2013, 22(7-8):1337-1347.
5. Hotelling H. Relations Between Two Sets of Variates[M]// Breakthroughs in Statistics. Springer New York, 1992:321-377.
6. Akaho S. A kernel method for canonical correlation analysis[J]. In Proceedings of the International Meeting of the Psychometric Society (IMPS2001, 2006, 40(2):263-269.
7. Nielsen A A. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data.[J]. IEEE Transactions on Image Processing, 2002, 11(3):293-305.
8. Rupnik J, Shawe-Taylor J. Multi-View Canonical Correlation Analysis[J]. Taylor.
9. Sharma A, Kumar A, Daume H, et al. Generalized Multiview Analysis: A discriminative latent space[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012:2160-2167.
10. Zhao W, Phillips P J. Subspace Linear Discriminant Analysis for Face Recognition[J]. 1999.
11. Mika S, Rätsch G, Weston J, et al. Fisher discriminant analysis with kernels[C]// Neural Networks for Signal Processing Ix, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.IEEE,2002:41-48.
12. Kan M,Shan S,Zhang H, et al. Multi-view Discriminant Analysis[C]// European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012:808-821.
13. Kan M,Shan S,Zhang H, et al.Multi-View Discriminant Analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(1):188.
14. Jing P, Su Y, Nie L, et al. Low-rank Multi-view Embedding Learning for Micro-video Popularity Prediction[J]. IEEE Transactions on Knowledge & Data Engineering, 2018, PP(99):1519-1532.
15. Jing P, Su Y, Nie L, et al. A Framework of Joint Low-rank and Sparse Regression for Image Memorability Prediction[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2018, PP(99):1-1.
16. He JR, Ding LX, Li ZK, Hu QH. Margin discriminant projection for dimensionality reduction. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):826−838 (in Chinese). http://www.jos.org.cn/1000-9825/4571.htm
17. Huang S, Yang D, Zhou J, et al. Graph regularized linear discriminant analysis and its generalization[J]. Pattern Analysis & Applications, 2015, 18(3):639-650.
18. He J, Wu D, Xiong N, et al. Orthogonal margin discriminant projection for dimensionality reducti on[J]. Journal of Supercomputing, 2016, 72(6):2095-2110.
19. Cai D, He X, Han J, et al. Graph Regularized Nonnegative Matrix Factorization for Data Representation[J]. IEEE Transactions on Pattern Analysis&Machine Intelligence,2011,33(8):1548-1560.
20. Lei Z, Shen J, Liang X, et al. Unsupervised Topic Hypergraph Hashing for Efficient Mobile

Image Retrieval[J]. IEEE Transactions on Cybernetics, 2016, PP(99):1-14.

21. Zhu L, Huang Z, Liu X, et al. Discrete Multi-modal Hashing with Canonical Views for Robust Mobile Landmark Search[J]. IEEE Transactions on Multimedia, 2017, PP(99):1-1.
22. Zhu L, Shen J, Jin H, et al. Landmark Classification With Hierarchical Multi-Modal Exemplar Feature[J]. IEEE Transactions on Multimedia, 2015, 17(7):981-993.
23. Xie L, Jin H, Shen J, et al. Content-based Landmark Search via Multi-modal Hypergraph Learning[J]. IEEE Transactions on Cybernetics, 2015, 45(12):2756.
24. Nie L, Song X, Chua T S. Learning from Multiple Social Networks[J]. Synthesis Lectures on Information Concepts Retrieval & Services, 2016, 8(2):118.
25. Nie L, Zhang L, Meng L, et al. Modeling Disease Progression via Multisource Multitask Learners: A Case Study With Alzheimer's Disease[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 28(7):1508-1519.
26. Nie L, Zhang L, Yang Y, et al. Beyond Doctors: Future Health Prediction from Multimedia and Multimodal Observations[C]// ACM International Conference on Multimedia. ACM, 2015:591-600.
27. Nie L, Wang X, Zhang J, et al. Enhancing Micro-video Understanding by Harnessing External Sounds[C]// ACM on Multimedia Conference. ACM, 2017:1192-1200.
28. Jing P, Su Y, Nie L, et al. Predicting Image Memorability Through Adaptive Transfer Learning from External Sources[J]. IEEE Transactions on Multimedia, 2017, PP(99):1-1.
29. Baesens B, Viaene S, Gestel T V, et al. Least squares support vector machine classifiers: an empirical evaluation[J]. Access & Download Statistics, 2000:1-16.
30. Ye Zhen, Bai Lin, Nian Yongjian. Hyperspectral Image Classification Algorithm Based on Gabor Feature and Locality-Preserving Dimensionality Reduction[J]. Acta Optica Sinica, 2016, 36(10):1028003.
31. Ye Z, He M. PCA and windowed wavelet transform for hyperspectral decision fusion classification[J]. Journal of Image & Graphics, 2015.
32. Deng Y J, Li H C, Pan L, et al. Modified Tensor Locality Preserving Projection for Dimensionality Reduction of Hyperspectral Images[J]. IEEE Geoscience & Remote Sensing Letters, 2018, PP(99):1-5.
33. Hu W, Huang Y, Wei L, et al. Deep Convolutional Neural Networks for, Hyperspectral Image Classification[J]. Journal of Sensors, 2015, 2015(2):1-12.
34. Lin L, Song X. Using CNN to Classify Hyperspectral Data Based on Spatial-spectral Information[M]// Advances in Intelligent Information Hiding and Multimedia Signal Processing. Springer International Publishing, 2017.
35. Mei S, Ji J, Hou J, et al. Learning Sensor-Specific Spatial-Spectral Features of Hyperspectral Images via Convolutional Neural Networks[J]. IEEE Transactions on Geoscience & Remote Sensing, 2017, 55(8):4520-4533.
36. Timothy P. Van Voorhis. The quadratically constrained quadratic program /[J]. Georgia Institute of Technology, 1997.
37. Fan R K C. Spectral graph theory[M]. Published for the Conference Board of the mathematical sciences by the American Mathematical Society, 1997.
38. Moore, B. Principal component analysis in linear systems: Controllability, observability, and model reduction[J]. Automatic Control, IEEE Transactions on, 1981, 26(1):17-32.
39. Keller J M, Gray M R, Givens J A. A fuzzy K-nearest neighbor algorithm[J]. IEEE Transactions

on Systems Man & Cybernetics, 201

40. R Bao, J Xia , Z Xue, P Du, M Che. Ensemble Classification for Hyperspectral Imagery based on Morphological Attribute Profiles[J]. Remote Sensing Technology and Application, 2016.