Learning Object Material Categories via Pairwise Discriminant Analysis

Zhouyu Fu and Antonio Robles-Kelly
NICTA*, RSISE Bldg. 115, Australian National University, Canberra, ACT 0200, Australia

Abstract

In this paper, we investigate linear discriminant analysis (LDA) methods for multiclass classification problems in hyperspectral imaging. We note that LDA does not consider pairwise relations between different classes, it rather assumes equal within and between-class scatter matrices. As a result, we present a pairwise discriminant analysis algorithm for learning class categories. Our pairwise linear discriminant analysis measures the separability of two classes making use of the class centroids and variances. Our approach is based upon a novel cost function with unitary constraints based on the aggregation of pairwise costs for binary classes. We view the minimisation of this cost function as an unconstrained optimisation problem over a Grassmann manifold and solve using a projected gradient method. Our approach does not require matrix inversion operations and, therefore, does not suffer of stability problems for small training sets. We demonstrate the utility of our algorithm for purposes of learning material catergories in hyperspectral images.

1. Introduction

The development of image sensor technology has made it possible to capture image data in hundreds of bands covering a broad spectrum of wavelength range. The rich information available in hyperspectral imagery has posed significant opportunities and challenges for feature extraction and classification. Many algorithms have been proposed for this purpose, such as Principal Component Analysis, Linear Discriminant Analysis, Decision Boundary , Projection Pursuit, and kernel methods[14]. All these algorithms treat the raw pixel spectra as input vectors in high dimensional spaces and look for linear or nonlinear mappings to the feature space, often with reduced dimensionality, by optimizing certain criterion, leading to statistically optimal solutions to classification.

Linear Discriminant Analysis (LDA) [9] is a classical method for linear dimensionality reduction and feature extraction. It can utilise label information for purposes of learning a lower dimensional space representation suitable for feature extraction, supervised learning and classification. Both LDA and the closely related Fisher's Linear Discriminant (FDA) [7] are concerned with learning the optimal projection direction for binary classes. The idea is to recover a linear feature transformation which maximises the variance between two classes while keeping the intraclass variances small.

These methods can be naturally generalised to handle multiclass classification tasks by introducing within and between-class scatter matrices to represent the average variance of each class and the distance between classes. As a result, the optimal transformation is obtained so as to maximise the between-class scatter while minimising the within-class dispersion. For the within-class scatter, LDA assumes the same conditional distribution for all the classes under study. This class conditional distribution is then modelled as a single Gaussian. This, in turn, implies that all the classes must have identical full rank covariances. Similarly, a single between-class scatter matrix is defined for all classes, which, hence, assumes each class is equally separable from the others.

Though effective, the assumptions above ignore the inherent inhomogeneities between different classes. This is particularly apparent in multiclass classification, where some classes are much harder to discriminate than others. As a result, well-separated classes can be over emphasised, whereas overlapping or neighbouring classes may not be well discriminated against one another. Therefore, the optimal feature transformation should aim at preserving the distances between well separated classes and, at the same time, place adequate emphasis on classes which are hard to discriminate upon.

This paper aims at casting the problem of multiclass classification in terms of a pairwise linear discriminant analysis based upon a measure of separability for every pair of classes. We call this new formulation of discriminant analysis *Pairwise Discriminant Analysis* (PDA). This approach gives rise to a new cost function for LDA based on the pairwise cost of separating any two classes in the feature space. Furthermore, the optimisation of the cost function can be approached by a gradient-based method on a Grassmann

*National ICT Australia is funded by the Australian Governments Backing Australia's Ability initiative, in part through the Australian Research Council.

1-4244-1180-7/07/$25.00 ©2007 IEEE

manifold akin to those in [16] and [18]. Moreover, our approach does not require the inversion of the class scatter matrix and, therefore, is not prone to instability for undersampled feature spaces.

The paper is organised as follows. Related work is presented in the following section. The cost function for our PDA method is developed in section 3. In section 4, we elaborate on the optimisation of the cost function. Experimental results are presented in section 5 and conclusions are given in section 6.

2. Motivation and Previous Work

To commence, let us define the generic problem of linear feature extraction for classification as treated in this paper. Given the sample set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in $\mathcal{R}^D$, the purpose is to find a transformation matrix $\mathbf{A} \in \mathcal{R}^{D\times d}$ that projects the input vector $\mathbf{x}_i$ onto the point $\mathbf{y}_i = \mathbf{A}^T\mathbf{x}_i$ in a lower dimensional space $\mathcal{R}^d (d \ll D)$ so as to maximise the separation between classes and the affinity within classes.

In computer vision and pattern recognition, this is often achieved by optimising a cost function over the linear combination of features which best separates two or more classes. One of the most popular methods for recovering these linear combinations for feature extraction in supervised learning is LDA. Here, a transformation matrix $\mathbf{A} \in \mathcal{R}^{D\times d}$ is determined so as to maximise the *Fisher criterion* given by

$$\begin{aligned} J_F(\mathbf{A}) &= tr((\mathbf{A}^T\mathbf{S}_w\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{S}_b\mathbf{A})) \\ \mathbf{S}_w &= \sum_{j=1}^{c} p_j\mathbf{S}_j = \sum_{j=1}^{c}\sum_{\mathbf{x}_i\in C_j} p_j(\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T \\ \mathbf{S}_b &= \sum_{j=1}^{c} p_j(\mathbf{m}_j - \overline{\mathbf{m}})(\mathbf{m}_j - \overline{\mathbf{m}})^T \\ &= \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} p_i p_j(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \end{aligned} \quad (1)$$

where c is the number of classes, $\mathbf{m}_j$ and $\overline{\mathbf{m}}$ are the class mean and sample mean vectors, respectively, and p_j is the prior probability of class j given by the contribution of the j^{th} class to the sample set $\mathcal{X}$. Also, in the equations above, $\mathbf{S}_w$ and $\mathbf{S}_b$ represent the within and between-class scatter matrices. Note that the matrix $\mathbf{S}_w$ can be regarded as the average class-specific covariance, whereas $\mathbf{S}_b$ can be viewed as the mean distance between all different classes. Thus, the purpose of Equation 1 is to maximise the between-class scatter while preserving within-class dispersion in the transformed feature space. This is effected by affine projecting the inverse intraclass covariance matrix and solving the generalised eigenvalue problem $\mathbf{S}_b\mathbf{A} = \lambda\mathbf{S}_w\mathbf{A}$. As the rank of S_b is $c-1$, the solution, for the c classes, is obtained by taking the eigenvectors corresponding to the largest $c-1$ eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$ [9].

Various extensions and improvements to LDA have been proposed in the literature. For instance, nonparametric discriminant analysis (NDA) [9] incorporates boundary information into between-class scatter. Mika *et al.* [19] and Boudat and Anour [2] have proposed kernel versions of LDA that can cope with severe non-linearity of the sample set $\mathcal{X}$. On the numerical stability and tractability of the LDA solution, there have also been a number of methods which aim at overcoming the singularity of the inverse intraclass covariance matrix inherent to undersampled feature spaces [8, 3, 4, 23, 22]. In a related development, Maximum Margin Criterion (MMC) [15] employs an optimisation procedure whose constraint is not dependent on the non-singularity of the within-class scatter matrix $\mathbf{S}_w$. Wang and Tang [21] have used dual subspaces to construct LDA classifiers which preserve most discriminative information.

Given the renewed interest and the vast work on LDA, it is somewhat surprising that there has not yet been paid much attention to the pairwise balance between different classes when seeking the optimal transformation in multiclass classification. This is ever more important since LDA methods maximise a cost function governed by the overall average class separability and, therefore, are likely to favor well-separated clusters.

To illustrate this phenomenon, a toy example is shown in Figure 1. Here, we represent three different classes in a two-dimensional space. We have denoted these classes C_1, C_2 and C_3. Note that C_1 and C_2 are very close to one another, and are both far away from C_3. Recall that the total between-class scatter matrix is the weighted average of the pairwise between-class scatters (see Equation 1). In this case, the distances between C_3 and the other two classes are larger than the distance between C_1 and C_2. As a result, the first principal direction, i.e. the direction corresponding to the eigenvector of the largest eigenvalue of $\mathbf{S}_w^{-1}\mathbf{S}_b$is dominated by the distances between C_3 and the other two classes, as indicated by dashed arrow. However, if we project the samples in this direction, there will be a noticeable overlap between the classes C_1 and C_2. It can be shown that the optimal projection direction for the three-class problem in the figure is given by the solid arrow. This is due to the fact that the projections of C_1, C_2 and C_3 onto the solid arrow direction are such that the interclass scatter is minimum. It is then clear that the pairwise relations between classes must be considered in order to achieve optimal class separation.

Moreover, LDA assumes the within-class scatters to be equal. Heteroscadastic Discriminant Analysis (HDA) [13] employs a maximum likelihood framework so as to account for dissimilar intra-class covariances. In a related development, Loog *et al.* [17] have proposed a weighted pairwise fisher scheme to balance the scatter between different

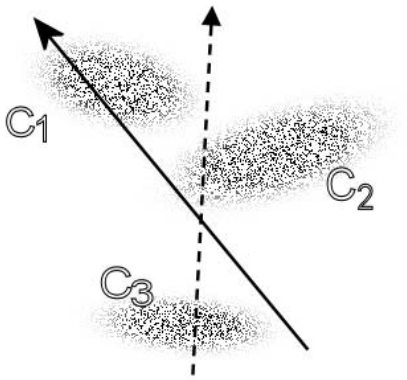


Fig. 1. Example of multiclass discriminant analysis. Three classes, the principal projection directions recovered by LDA (dotted arrow) and the desired optimal direction (solid arrow).

classes. A weight, which is derived based on the Mahalanobis distance between the class centroids, is applied to each pairwise between-class scatter matrix. This treatment has the effect of controlling the balance between pairs of classes. However, the within-class dispersion matrix S_w is considered to be equal for all the clusters.

3. Pairwise Discriminant Analysis

Hence, in order to gain over other LDA approaches for multiclass classification, we must learn a transformation to a subspace making use of a pairwise framework which accumulates the combination of costs for pairs of binary classes. Any multiclass classification problem can be converted to a number of binary ones by using a pairwise fusion framework [11]. This hinges in training a classifier for every two classes and making the final prediction based upon the combination of the decisions yield by the binary classifiers. Following this rationale, for multiclass discriminant analysis, we can also define costs for every pair of classes and combine them into a final target function in a cumulative fashion.

To do this, we require a criterion to measure the separability between two classes in the transformed feature space. The choice here is not unique. A straightforward way would be to use the same objective function as in Equation 1 with different between and within-class scatter matrices defined for every pair of classes. However, this involves a matrix inversion operation for each pairwise within-class scatter matrix. This is undesirable since it can result in numerical instability when small training datasets are available for any of the classes or categories under consideration. Another way would be to assume an underlying Gaussian distribution for each of the classes under study and employ information theoretic divergence measures, such as the Kullback-Leibler divergence or the Bhattacharyya distance with closed form solutions [5]. Unfortunately, this would still require matrix inversion operations and hence may be

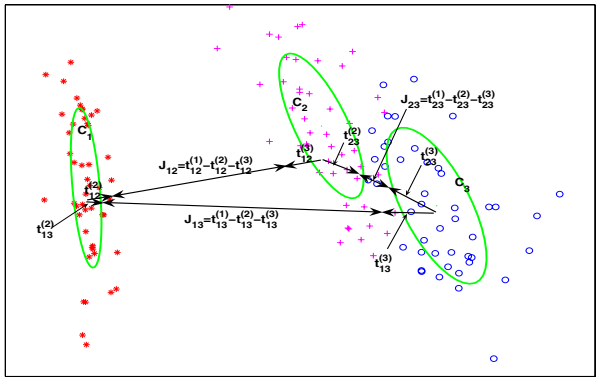


Fig. 2. Illustration of our proposed measure of class separation. "*"s, "+"s and "o"s denote features of classes 1, 2 and 3. $J_{i,j}$ denotes the separation measure for class i and class j.The larger $J_{i,j}$, the better separability for the two classes. In the sake of clarity, we have omitted the distances between class centroids.

unstable for applications with small training sets.

To overcome these problems, we propose a new measure of class distance which takes into account both, the centroid information and class variances in the transformed feature space. Our separation measure $J_{i,j}(\mathbf{A})$ is composed of three ingredients. The first of these is the term $t^{(1)}_{i,j}$, which denotes the L2 distance between the class centroids in the transformed subspace. The other two terms, $t^{(2)}_{i,j}$ and $t^{(3)}_{i,j}$, denote the variances of classes indexed i and j mapped onto the subspace projection matrix $\mathbf{A}$. That is, the projection of the scatters along the direction $\mathbf{A}^T\mathbf{m}_{i,j}$, which can be viewed as the line in the subspace connecting the centroids of classes i and j. Our class-distance measure $J_{i,j}(\mathbf{A})$ is then given by the distance between centroids minus the variances of the two classes along the line across them. To express $J_{i,j}(\mathbf{A})$ in matrix notation, let the mean vector and covariance matrix for class i be $\mathbf{m}_i$ and $\mathbf{S}_i$ respectively. We can then write

$$
\begin{aligned}
J_{i,j}(\mathbf{A}) &= t^{(1)}_{i,j} - t^{(2)}_{i,j} - t^{(3)}_{i,j} \qquad (2)\\
t^{(1)}_{i,j} &= ||\mathbf{A}^T\mathbf{m}_{i,j}|| = ||\mathbf{A}^T(\mathbf{m}_i - \mathbf{m}_j)||\\
t^{(2)}_{i,j} &= \frac{s^{(2)}_i}{t^{(1)}_{i,j}} = \frac{\sqrt{\mathbf{m}^T_{i,j}\mathbf{A}\mathbf{A}^T\mathbf{S}_i\mathbf{A}\mathbf{A}^T\mathbf{m}_{i,j}}}{||\mathbf{A}^T\mathbf{m}_{i,j}||}\\
t^{(3)}_{i,j} &= \frac{s^{(2)}_j}{t^{(1)}_{i,j}} = \frac{\sqrt{\mathbf{m}^T_{i,j}\mathbf{A}\mathbf{A}^T\mathbf{S}_j\mathbf{A}\mathbf{A}^T\mathbf{m}_{i,j}}}{||\mathbf{A}^T\mathbf{m}_{i,j}||}
\end{aligned}
$$

where have used the shorthand $\mathbf{m}_{i,j} = \mathbf{m}_i - \mathbf{m}_j$ and s_i denotes the projection of the i^{th} class scatter on the line connecting the class centroids.

In Figure 2 we illustrate the behaviour of our class separability measure making use of a 3-class dataset in a two-dimensional feature space. Here, we have denoted the samples for each class using a different marker. The corresponding terms, as given in Equation 2, are labelled for

all pairs of classes. From the figure, the physical meanings of the terms $t_{i,j}^{(1)}$, $t_{i,j}^{(2)}$ and $t_{i,j}^{(3)}$ become evident. Furthermore, the class separation $J_{i,j}(\mathbf{A})$ has a clear relation with respect to the margin between classes and, hence, has properties akin to those of margin-based classifiers, such as Support Vector Machines (SVMs) [20]. If the principal axes of the underlying distributions for both classes are aligned with the line through their centroids, our measure becomes equivalent to the class-margin. It is also worth noting that the proposed measure of separation does not involve any matrix inversion and is, thus, numerically more stable than other alternatives elsewhere in the literature. Note, however, that $J_{i,j}(\mathbf{A})$ is not a metric in the sense that it can be negative. Nonetheless, this is not a problem as we are not using it directly for optimisation purposes, rather it is treated as a variable in the objective function.

Having defined the separation measure for binary classes, we can now define the objective function for our multiclass PDA algorithm as the aggregation of the pairwise costs. As a result, our objective function becomes

$$\arg \min_{\mathbf{A}^T\mathbf{A}=\mathbf{I}} f(\mathbf{A}) = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} p_{i,j} g\left(J_{i,j}(\mathbf{A})\right) \quad (3)$$

$$g(x) = \frac{1}{1+\exp(\gamma(x-\mu))}$$

Here $J_{i,j}(A)$ is defined as in Equation 2, $g(x)$ is a nonlinear function which maps the class separability measures to pairwise costs and $p_{i,j} \propto n_i + n_j$ is the weight for the class pair i, j governed by the sizes n_i and n_j for the two classes.

Our choice of $g(x)$ has several desirable properties. First, it is monotonically decreasing so that larger separability values are always associated with lower costs. More importantly, $g(x)$ takes values in the range $(-\infty, \infty)$ to the bounded interval $[0, 1]$ in a nonlinear way such that the cost changes rapidly only for moderate values of $J_{i,j}(\mathbf{A})$. Note that, for extreme values of $J_{i,j}(\mathbf{A})$, the rate of change of $g(x)$ is small. Consider the case of well-separated classes or significantly overlapping ones which are *de facto* inseparable. For these classes, the function $g(x)$ will yield pairwise costs that approximate its asymptotic values. In the other hand, for neighbouring classes which are *separable*, the cost varies rapidly with respect to the measure $J_{i,j}(\mathbf{A})$. This, in turn, implies that the parameters μ and γ control the bandwidth of the function $g(x)$ and the position of its inflexion point. This is exemplified in Figure 3. Thus, the cost function can be fine-tuned to the application vehicle under consideration. In this paper, we choose these two parameters via cross validation through a grid-search over a discrete set of values.

For optimisation purposes, we have imposed the unitary constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ on the matrix $\mathbf{A}$. This is to ensure that the discriminant vectors are orthonormal. This has the effect of minimising the correlations between them while preserving the scale of the input space. Note that, although conventional LDA does not require $\mathbf{A}$ to be orthogonal, many alternatives in the literature [16, 22] assume orthogonality of the transformation matrix. From a more general perspective, both the unitary constraint and the nonlinear mapping $g(x)$ can be viewed as regularisers that prevent overfitting when optimising our objective function.

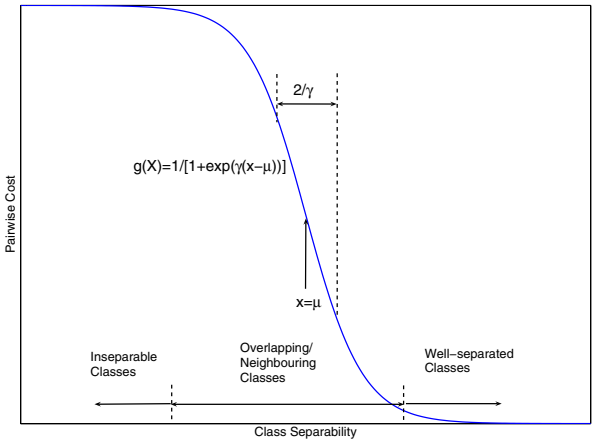


Fig. 3. Mapping from class separability to pairwise cost

4. Optimisation of the Cost Function

Equation 3 is a hard optimisation problem which is defined on a Grassmann manifold [10]. In addition to the unitary constraint, the objective function is invariant to any rotations in the transformed feature space, i.e. $f(\mathbf{A}) = f(\mathbf{AQ})$ for an arbitrary orthogonal matrix $\mathbf{Q} \in \mathcal{R}^{d\times d}$. Thus, we can build on recent developments in optimisation theory which have made possible the extension of unconstrained optimisation methods in the Euclidean space to Grassmann manifolds [6, 18]. In this paper, for optimisation on the Grassmann manifold, we use a projection-based, steepest descent method with backtracking line search based upon those proposed in [18, 16].

4.1. Projection-based Steepest Descent

Before introducing the optimisation algorithm, we require some formalism. In the manifold optimisation algorithm presented here, each point is only allowed to move along the descent direction within the tangent space. This is, the optimisation is governed by the tangent space defined at each point on a differentiable manifold embedded in the ambient Euclidean space. This tangent space is the natural generalisation of the tangent line in R^2 and the tangent plane in R^3 to the case of R^n. Moreover, following [6], for any matrix $\mathbf{X} \in R^{D\times d}$, its projection onto the tangent space is given by

$$P_T(\mathbf{X}) = P_\perp(\mathbf{X}) + P_\parallel(\mathbf{X}) \quad (4)$$

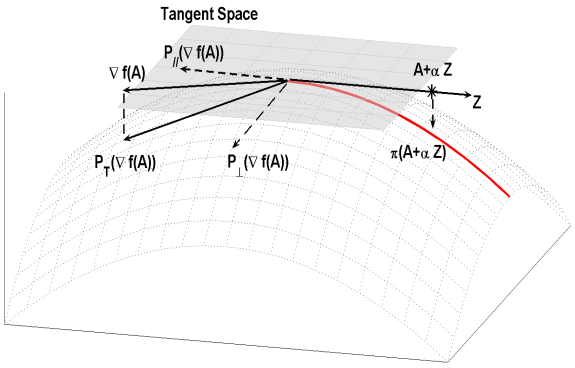


Fig. 4. Illustration of the steepest descent direction and various projections.

where $P_{\perp}(\mathbf{X}) = \frac{1}{2}\mathbf{A}(\mathbf{A}^T\mathbf{X} - \mathbf{X}^T\mathbf{A})$ and $P_{\parallel}(\mathbf{X}) = (\mathbf{I} - \mathbf{A}\mathbf{A}^T)\mathbf{X}$ are the the projections onto the so called *vertical* and *horizontal* spaces. In Figure 4, we show the geometry of these spaces on the manifold. Considering the rotation invariant condition, not all variations on the tangent space will contribute to the change of the objective function value. It can be shown that, on the Grassmann manifold, only the directions in the horizontal space will change the value of the objective function [6]. Hence, by substituting the gradient $\nabla f(\mathbf{A})$ of our objective function for $\mathbf{X}$ into Equation 4, the steepest descent direction on the Grassmann manifold becomes

$$\mathbf{Z} = -(\mathbf{I} - \mathbf{A}\mathbf{A}^T)\nabla f(\mathbf{A}) \tag{5}$$

With the steepest descent direction at hand, we can turn our attention to the line search method on the Grassmannn manifold. This is quite different from that in the Euclidean space, as line search must be made on the manifold itself. Hence, instead of searching along the line in the descent direction, we search along the geodesics on the manifold. This can be achieved by back-projecting the points on the line onto the geodesics. For the Grassmann manifold, let $\mathbf{X} \in \mathcal{R}^{D\times d}$ be a rank-d matrix as before. Specifically, if the QR decomposition of $\mathbf{X}$ is $\mathbf{X} = \mathbf{Q}\mathbf{R}$, then $\pi(\mathbf{X}) = \lfloor \mathbf{Q}\mathbf{I}_{D,d} \rfloor$. It can be shown that $\pi(\mathbf{X}) = \lfloor \mathbf{X} \rfloor = \lfloor \arg\min_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}} ||\mathbf{X}-\mathbf{Q}||^2 \rfloor$, where $\lfloor \mathbf{Q} \rfloor$ represents the subspace spanned by the columns of $\mathbf{Q}$ [18]. Therefore, the advantage of the treatment above is that we can use $\mathbf{Q}$ to represent $\pi(\mathbf{X})$ as an alternative to $\mathbf{X}$. This is due to the fact that the subspace spanned by the columns of $\mathbf{X}$ is the same subspace spanned by the first d columns of $\mathbf{Q}$ for the QR decomposition of $\mathbf{X}$.

4.2. Algorithm

With the theory above, we now proceed to introduce the optimisation algorithm used in this paper. The steps for the steepest descent method on the Grassmann manifold are described in Figure 5. Similarly to conventional gradient descent methods, the algorithm used here employs interleaved steps of gradient calculation and line search along the steepest descent direction until convergence.

---

1. t=0. Initialise $\mathbf{A}^{(0)} \in \mathcal{R}^{D\times d}$ with $\mathbf{A}^{(0)T}\mathbf{A}^{(0)} = \mathbf{I}$.
2. Compute the gradient $\nabla f(\mathbf{A}^{(t)})$ of the objective function $f(\cdot)$ at $A^{(t)}$ via Equation 6 and set the descent direction $\mathbf{Z}^{(t)} = -(\mathbf{I} - \mathbf{A}^{(t)T}\mathbf{A}^{(t)})\nabla f(\mathbf{A}^{(t)})$.
3. Evaluate $||\mathbf{Z}^{(t)}|| = tr(\mathbf{Z}^{(t)T}\mathbf{Z}^{(t)})$. Stop if $||\mathbf{Z}^{(t)}||$ is sufficiently small or the maximum number of iterations is reached.
4. Otherwise, perform line search along the direction of $\mathbf{Z}^{(t)}$ making use of the following rules
   - Repeat $\lambda = 2\lambda$ while $f(\mathbf{A}^{(t)}) - f(\pi(\mathbf{A}^{(t)} + 2\lambda\mathbf{Z}^{(t)})) \geq \lambda||\mathbf{Z}^{(t)}||$.
   - Repeat $\lambda = \frac{1}{2}\lambda$ while $f(\mathbf{A}^{(t)}) - f(\pi(\mathbf{A}^{(t)} + \lambda\mathbf{Z}^{(t)})) < \frac{1}{2}\lambda||\mathbf{Z}^{(t)}||$.
5. Do $\mathbf{A}^{(t+1)} = \pi(\mathbf{A}^{(t)} + \lambda\mathbf{Z}^{(t)})$
6. Stop if $\frac{||f(\mathbf{A}^{(t+1)} - f(\mathbf{A}^{(t)}||}{||f(\mathbf{A}^{(t)}||}$ is small enough. Otherwise do $t = t + 1$ and go to Step 2.

Fig. 5. Steepest Descent algorithm with Backtracking Line Search on the Grassmann Manifold

Several points deserve further elaboration here. First, the gradient $\nabla f(\mathbf{A})$ of the cost function $f(\mathbf{A})$ in Equation 3 is needed at each iteration, which can be expressedin closed form as follows

$$\begin{aligned}
\nabla f(\mathbf{A}) &= -\sum_i \sum_j \gamma p_{i,j}(J_{i,j}(\mathbf{A}) - \mu)g^2(J_{i,j}(\mathbf{A}))\nabla J_{i,j} \\
\nabla J_{i,j} &= \nabla t_{i,j}^{(1)} - \frac{(s_j + s_j)\nabla t_{i,j}^{(1)}}{t_{i,j}^{(1)^2}} - \frac{t_{i,j}^{(1)}(\nabla s_i + \nabla s_j)}{t_{i,j}^{(1)^2}} \\
\nabla t_{i,j}^{(1)} &= \frac{1}{t_{i,j}^{(1)}}\mathbf{m}_{i,j}\mathbf{m}_{i,j}^T\mathbf{A} \\
\nabla s_i &= \frac{2}{s_i}\,\mathrm{sym}(\mathbf{m}_{i,j}\mathbf{m}_{i,j}^T\mathbf{A}\mathbf{A}^T\mathbf{S}_i)\mathbf{A} \\
\nabla s_j &= \frac{2}{s_j}\,\mathrm{sym}(\mathbf{m}_{i,j}\mathbf{m}_{i,j}^T\mathbf{A}\mathbf{A}^T\mathbf{S}_j)\mathbf{A}
\end{aligned} \tag{6}$$

where $\mathrm{sym}(\mathbf{X})$ denotes the symmetry inducing operator $\frac{\mathbf{X} + \mathbf{X}^T}{2}$ for the matrix $\mathbf{X}$. The other terms are the same as they were first defined in Equation 2.

The line search in step 4 is adapted from the Armijo step size rule [1] for approximate search of the minimum

point along the geodesics on the manifold. Also, in our implementation we have employed conventional LDA [9], so as to initialise the algorithm. To this end, we compute the transformation matrix yield by LDA and impose unitary constraints. Recall that a unitary transformation matrix $\mathbf{A}$ is not a necessarily condition for LDA, thus, we use instead $\mathbf{A}^{(0)} = \pi(\mathbf{A})$, where $\pi(\mathbf{A})$ is the subspace presented in the previous section.

5. Experimental Results

In this section, we present results of our pairwise discriminant analysis algorithm for material identification and mapping on two hyperspectral images. The first of these is an image captured by the AVIRIS sensor system over an agricultural area in West Indiana, USA. Each pixel in the image is comprised of 220 bands in the range $375-2200$nm. The second data set was gathered over the Washington DC mall by a HYDICE system with the same number of spectral bands and similar spectral range. For both images, we have removed the water and atmospheric absorption bands in each pixel spectra and used the spectral values at the remaining 191 bands to form the feature vectors. We used a total of 9345 pixels from 9 classes in the first image data set and a total of 14311 pixels from 5 classes in the second image for our experiments. The pseudo-color images and the ground truth maps alongside with legends for both data sets are shown in Figure 6. Different colors indicate different terrain material types in the ground truth images.

Average Accuracy for Data Set 1

	5%	10%	20%
PCA	75.60 ± 18.33	79.51 ± 16.15	83.18 ± 13.06
LDA	76.32 ± 18.14	80.20 ± 15.94	83.94 ± 12.72
MMC	71.82 ± 21.94	74.23 ± 20.73	75.92 ± 19.77
PDA	$\mathbf{84.28 \pm 11.47}$	$\mathbf{87.83 \pm 9.02}$	$\mathbf{89.95 \pm 7.43}$

Average Accuracy for Data Set 2

	5%	10%	20%
PCA	78.25 ± 8.94	79.92 ± 6.94	82.06 ± 6.63
LDA	83.93 ± 7.64	86.73 ± 6.21	87.78 ± 5.93
MMC	81.75 ± 8.19	83.65 ± 6.78	84.53 ± 6.20
PDA	$\mathbf{87.03 \pm 6.27}$	$\mathbf{87.98 \pm 5.27}$	$\mathbf{89.28 \pm 4.36}$

Tab. 1. Average Within-class Accuracy for Both Data Sets

For both data sets, we have compared our results with those yield by Principal Component Analysis (PCA) [12], Linear Discriminant Analysis (LDA) [9] and the Maximum Margin Criterion (MMC) [15]. Throughout our experiments, we have used the same 1-nearest-neighbour classifier [5] and set the maximum number of iterations to 20. For our algorithm, as outlined in Figure 5, the error tolerance is 10^{-3}. For purposes of quantitative analysis, we are mainly interested in two criteria. The first of these is the overall testing error, defined by the total number of misclassified samples over the total number of testing samples. This measures the overall performance of the methods in the comparison group. Note that, however, in some cases the features tend to have good overall performance with very good results for most classes but poor results for the remaining few. Hence, the second criterion, the average within-class accuracy, is used here to measure the balance of classification results between different classes. The within-class accuracy is defined as the classification accuracy for each class over the number of classes. The larger its value, the lower the bias.

We have examined the cases where 5%, 10% and 20% of the samples are used for training and the remainder of the dataset is used for testing. For each case, we have repeated our experiment 10 times with different, randomly selected training pixels so as to recover the mean and standard deviation values for the error and the accuracy rates. In Figures 7(a) and 7(b), we show the overall testing errors of both data sets for the algorithms under study w.r.t. different training data sizes. The dark green, light green and yellow bars show the results for those cases in which 5%, 10% and 20% of the data has been used for training. The error bars indicate the standard deviation. The average within-class accuracies for both data sets are reported in Table 1.

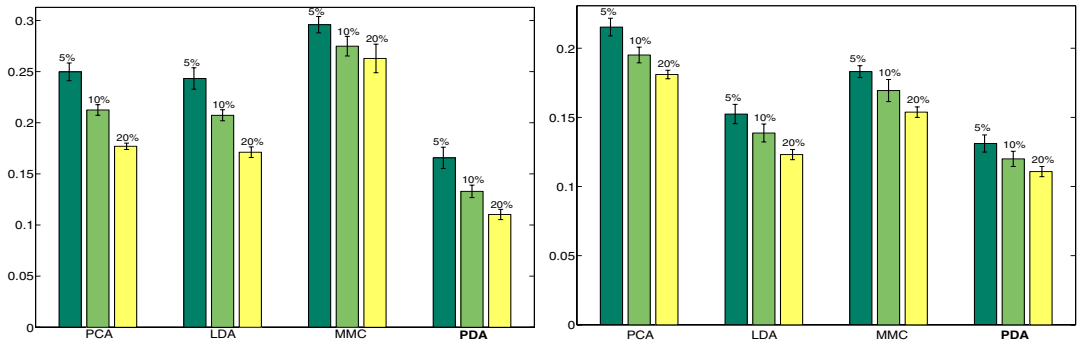
From the results, we can see that the proposed PDA algorithm achieves better results than the alternatives, both in terms of lower overall error rates and higher within-class accuracy. The improvement in performance is much more significant for smaller training sample sizes. The second-best result is achieved using traditional LDA, which gives a margin of improvement over PCA and MMC. This can be attributed to the fact that similar materials are hard to separate and, therefore, their corresponding classes show a high degree of overlap whereas different materials have very different spectra and are easy to classify. Our method balances differences between classes and favours discriminability between hard-to-classify materials. This is especially the case with the first data set, where classes formed by different vegetations have similar spectral profiles and are easily confused with one another with conventional approaches. Thus, despite the confusion caused by similar materials, our method achieves a significant improvement over the alternatives.

6. Conclusions

We have proposed a novel discriminant analysis algorithm for learning material class categories. The algorithm considers the pairwise relations between the classes and improves on the classification accuracy for different classes based on a nonlinear pairwise cost function. It does not require matrix inversion operations and is, hence, stable in dealing with small-sized training sets. The proposed algo-

(a) Image Data Set 1 (b) Image Data Set 2

Fig. 6. Example images for the datasets used in our experiments



(a) Error rates for image data set 1 (b) Error rates for image data set 2

Fig. 7. Testing errors for different algorithms and training data sizes.

rithm compares favourably with respect to traditional linear feature extraction methods in our experiments on a number of datasets.

References

[1] L. Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1), 1966. 5

[2] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000. 2

[3] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 2

[4] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new lda-based face recognition system which can solve the small sample size problem. 33:1713–1726. 2

[5] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973. 3, 6

[6] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1999. 4, 5

[7] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 1

[8] J. Friedman. Regularized discriminant analysis. *Journal American Statistical Assoc*, 94:165–175, 1989. 2

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990. 1, 2, 6

[10] J. Harris. *Algebraic Geometry, A First Course*. Springer, New York, 1992. 4

[11] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *Advances in Neural Information Processing Systems*, volume 10, 1998. 3

[12] I. Jolliffe. *Principal Component Analysis*. Springer, 1998. 6

[13] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26(4):283–297, 1998. 2

[14] D. Landgrebe. Hyperspectral image data analysis. *IEEE Signal Process. Mag.*, 19:17–28, 2002. 1

[15] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Neural Information Processing Systems*, volume 16, 2003. 2, 6

[16] D. Lin and X. Tang. Pursuing information projection on grassmann manifold. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 1727–1734, 2006. 2, 4

[17] M. Loog, R. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(7):762–766, 2001. 2

[18] J. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transaction on Signal Processing*, 50(3):635–650, 2002. 2, 4, 5

[19] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *In IEEE Neural Networks for Signal Processing Workshop*, pages 41–48, 1999. 2

[20] V. Vapnik. *The Statistical Learning Theory*. Springer, 1998. 4

[21] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proc. IEEE Computer Vision and Pattern Recognition*, 2004. 2

[22] J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005. 2, 4

[23] H. Yu and J. Yang. A direct lda algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001. 2