Multimodal Oriented Discriminant Analysis

Fernando De la Torre FTORRE@CS.CMU.EDU
Robotics Institute, Carnegie Mellon University. 5000 Forbes Av., Pittsburgh, PA 15217 USA

Takeo Kanade TK@CS.CMU.EDU
Robotics Institute, Carnegie Mellon University. 5000 Forbes Av., Pittsburgh, PA 15217 USA

Abstract

Linear discriminant analysis (LDA) has been an active topic of research during the last century. However, the existing algorithms have several limitations when applied to visual data. LDA is only optimal for Gaussian distributed classes with equal covariance matrices, and only classes-1 features can be extracted. On the other hand, LDA does not scale well to high dimensional data (over-fitting), and it cannot handle optimally multimodal distributions. In this paper, we introduce Multimodal Oriented Discriminant Analysis (MODA), a LDA extension which can overcome these drawbacks. A new formulation and several novelties are proposed:

- *An optimal dimensionality reduction for multimodal Gaussian classes with different covariances is derived. The new criteria allows for extracting more than classes-1 features.*
- *A covariance approximation is introduced to improve generalization and avoid over-fitting when dealing with high dimensional data.*
- *A linear time iterative majorization method is suggested in order to find a local optimum.*

Several synthetic and real experiments on face recognition show that MODA outperform existing linear techniques.

1. Introduction

Canonical Correlation Analysis (CCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), and Principal Component Analysis (PCA) are some examples of subspace methods (SM) useful for classification, dimensionality reduction and data modeling. These methods have been actively researched by the statistics, neural networks, machine learning and vision communities during the last century. The modeling power of SM can be especially useful when available data increases in features/samples, since there is a need for dimensionality reduction while preserving relevant attributes of the data. Another benefit of many subspace methods is that they can be computed as an eigenvalue or singular value type of problem, for which there are efficient numerical packages. An obvious drawback of SM is its linear assumptions; however, recently extensions based on kernel methods and latent variable models can overcome some of these limitations.

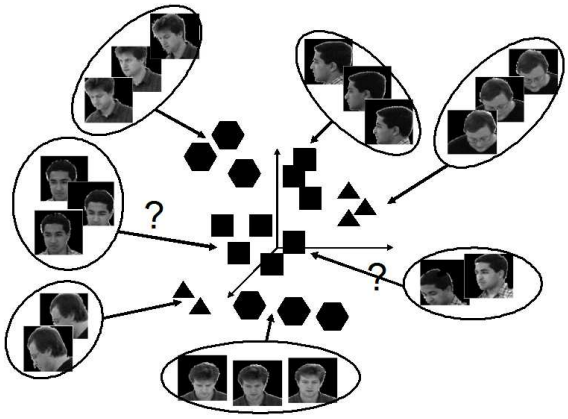

Figure 1. Projection onto a low dimensional space of face images for classification. Observe that the face distributions can be multimodal and with different covariances.

Among several classification methods (e.g. Support Vector Machines, decision trees), LDA remains a powerful preliminary tool for dimensionality reduction preserving discriminative features and avoiding the "curse of dimensionality". However, there exist several liminations of current LDA techniques (Fukunaga, 1990; Hastie et al., 2001). LDA is optimal only in the case

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

that all the classes are Gaussian distributed with equal covariances. Due to this assumption, the maximum number of features that can be extracted is the number of classes-1. Another common problem when dealing with high dimensional data is the small size problem (Yu & Yang, 2001), that is, the training set has more "dimensions" (pixels) than data samples [1]. In this situation LDA overfits and PCA techniques usually outperform LDA (Martinez & Kak, 2003). On the other hand, the computational/storage requirements of computing LDA directly from covariance matrices is impractical. In this paper we introduce Multimodal Oriented Discriminant Analysis (MODA), a new low dimensional discriminatory technique optimal for multimodal Gaussian classes with different covariances. MODA is able to efficiently deal with the small sample case and scales well to very high dimensional data avoiding overfitting effects. There is no closed form solution for the optimal values of MODA and an iterative majorization is proposed to seach for a local optimum. Finally, a new view and formulation of the LDA is introduced, which gives some new insights. Figure 1 illustrates the main purpose of this paper.

2. Linear Discriminant Analysis

The aim of LDA is to project the data into a lower dimensional space, so that the classes are as compact and as far as possible from each other. Many closed form solutions for LDA are based on the following covariance matrices:

$$\begin{aligned}\mathbf{S}_t &= \sum_{j=1}^{n}(\mathbf{d}_j-\mathbf{m})(\mathbf{d}_j-\mathbf{m})^T = \mathbf{D}\mathbf{P}_1\mathbf{D}^T \\ \mathbf{S}_w &= \sum_{i=1}^{c}\sum_{\mathbf{d}_j\in C_i}(\mathbf{d}_j-\mathbf{m}_i)(\mathbf{d}_j-\mathbf{m}_i)^T = \mathbf{D}\mathbf{P}_2\mathbf{D}^T \\ \mathbf{S}_b &= \sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T = \mathbf{D}\mathbf{P}_3\mathbf{D}^T\end{aligned}$$

$\mathbf{D}\in\Re^{d\times n}$ (see notation[2]) is the data matrix, where each column is a vectorized data sample. d denotes the number of features, n number of samples and c the number of classes. $\mathbf{m}=\frac{1}{n}\mathbf{D}\mathbf{1}_n$ is the mean vector for all the classes and $\mathbf{m}_i$ is the mean vector for the class i. n_i is the number of samples for class i and $\sum_{i=1}^{c} n_i = n$. $\mathbf{P}_i$ are projection matrices (i.e $\mathbf{P}_i^T=\mathbf{P}_i$ and $\mathbf{P}_i^2=\mathbf{P}_i$) with the following expressions:

$$\begin{aligned}\mathbf{P}_1=\mathbf{I}-\tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \quad \mathbf{P}_2=\mathbf{I}-\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T \\ \mathbf{P}_3=\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T-\tfrac{1}{n}\mathbf{1}_n\mathbf{1}_c\mathbf{G}^T \qquad (1)\end{aligned}$$

$\mathbf{G}\in\Re^{n\times c}$ is an dummy indicator matrix such that $\sum_j g_{ij}=1$, $g_{ij}\in\{0,1\}$ and g_{ij} is 1 if $\mathbf{d}_i$ belongs to class C_j.

$\mathbf{S}_b$ is the between-covariance matrix and represents the average of the distances between the mean of the classes. $\mathbf{S}_w$ represents the within-covariance matrix and it is a measure of the average compactness of each class. Finally $\mathbf{S}_t$ is the total covariance matrix. With the matrix expressions, it is straightforward to show that $\mathbf{S}_t=\mathbf{S}_w+\mathbf{S}_b$. The upper bounds on the ranks of the matrices are $c-1,\ n-c,\ n-1$ for $\mathbf{S}_b,\mathbf{S}_w,\mathbf{S}_t$ respectively.

Rayleigh like quotients are among the most popular LDA optimization criteria (Fukunaga, 1990). Some are: $J_1(\mathbf{B})=\frac{|\mathbf{B}^T\mathbf{S}_1\mathbf{B}|}{|\mathbf{B}^T\mathbf{S}_2\mathbf{B}|}$, $J_2(\mathbf{B})=tr((\mathbf{B}^T\mathbf{S}_1\mathbf{B})^{-1}\mathbf{B}^T\mathbf{S}_2\mathbf{B})$, $J_3(\mathbf{B})=\frac{tr(\mathbf{B}^T\mathbf{S}_1\mathbf{B})}{tr(\mathbf{B}^T\mathbf{S}_2\mathbf{B})}$, where $\mathbf{S}_1=\{\mathbf{S}_b,\mathbf{S}_b,\mathbf{S}_t\}$ and $\mathbf{S}_2=\{\mathbf{S}_w,\mathbf{S}_t,\mathbf{S}_w\}$. A closed form solution to previous minimization problems is given by a generalized eigenvalue problem $\mathbf{S}_1\mathbf{B}=\mathbf{S}_2\mathbf{B}\Lambda$. The generalized eigenvalue problem can be solved as a joint diagonalization, that is, finding a common basis $\mathbf{B}$, which diagonalizes simultaneously both matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ (i.e. $\mathbf{B}^T\mathbf{S}_2\mathbf{B}=\mathbf{I}$ and $\mathbf{B}^T\mathbf{S}_1\mathbf{B}=\Lambda$).

3. Oriented Discriminant Analysis

LDA is the optimal linear projection only in the case of having Gaussian classes with equal covariance matrix (Campbell, 1984; Duda et al., 2001; Hastie et al., 2001) (assuming enough training data). Fig. 2 shows a situation where two classes have almost orthogonal principal directions of the covariances and close means. In this pathological case, LDA chooses the worst possible discriminative direction where the classes are overlapped (it is also very numerically unstable), whereas ODA finds a better projection. In general, this situation becomes dangerous when the number of classes increases.

In order to solve this problem, several authors have proposed extensions and new views of LDA. Campbell (Campbell, 1984) derives a maximum likelihood approach to discriminant analysis. Assuming that all

[1] In this case the true dimensionality of the data is the number of samples, not the number of pixels.

[2] Bold capital letters denote a matrix $\mathbf{D}$, bold lower-case letters a column vector $\mathbf{d}$. $\mathbf{d}_j$ represents the j column of the matrix $\mathbf{D}$. All non-bold letters will represent variables of scalar nature. *diag* is an operator which transforms a vector to a diagonal matrix. $\mathbf{1}_k\in\Re^{k\times 1}$ is a vector of ones. $\mathbf{I}_k\in\Re^{k\times k}$ is the identity matrix. $tr(\mathbf{A})=\sum_i a_{ii}$ is the trace of the matrix $\mathbf{A}$ and $|\mathbf{A}|$ denotes the determinant. $||\mathbf{A}||_F=tr(\mathbf{A}^T\mathbf{A})=tr(\mathbf{A}\mathbf{A}^T)$ designates the Frobenious norm of a matrix. $N_d(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ indicates a d-dimensional Gaussian on the variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

the classes have equal covariance matrix, Campbell shows that LDA is equivalent to impose that the class means lie in a l-dimensional subspace. Following this approach, Kumar and Andreou (Kumar & Andreou, 1998) proposed heteroscedastic discriminant analysis, where they incorporate the estimation of the means and covariances in the low dimensional space. On the other hand, Saon et *al.* (Saon et al., 2000) define a new energy function to model the directionality of the data, $J(\mathbf{B}) = \prod_{i=1}^{c}(\frac{|\mathbf{B}^T\mathbf{S}_b\mathbf{B}|}{|\mathbf{B}^T\mathbf{\Sigma}_i\mathbf{B}|})^{n_i}$, where $\mathbf{\Sigma}_i$ is the class covariance matrix and $\mathbf{S}_b$ the between-class scatter covariance matrix.

Taking a different view point, Hastie et. al have proposed several LDA extensions by modifying the following regression problem (Hastie et al., 2001; Hastie et al., 1995):

$$E(\mathbf{B},\mathbf{V}) = ||\mathbf{V}\mathbf{G}^T - \mathbf{B}^T\mathbf{D}||_F \tag{2}$$

where $\mathbf{V} \in \Re^{k\times c}$ is a *scoring* matrix (Hastie et al., 1995). Similarly, Gallinari et. al (Gallinari et al., 1991) have also shown the connection between LDA and regression by minimizing $E_2(\mathbf{B},\mathbf{V}) = ||\mathbf{G}^T - \mathbf{V}^T\mathbf{B}^T\mathbf{D}||_F$. These approaches are appealing for several reasons. First, if the dummy matrix $\mathbf{G}$ contains 0 and 1's, the mapping gives a linear approximation of Bayes's posterior probability and if $g_{ij} = n_i/n$ then it returns classical LDA. On the other hand, Hastie et. al (Hastie et al., 2001; Hastie et al., 1995) have modified eq. 2 to take into account more than linear functions, for instance, they replace $\mathbf{B}^T\mathbf{D}$ by $\mathbf{B}\mathbf{f}(\mathbf{D})$, where $\mathbf{f}$ maps the original data (similar to kernel methods) introducing Flexible Discriminant Analysis (FDA) or Penalized Discriminant Analysis (PDA) by adding regularization terms to eq. 2 (e.g. $\mathbf{f}_2(\mathbf{B})$). Although similar in spirit, our work differs in several aspects; first we provide a new and probabilistic interpretation, we model directly the covariances in the original space rather than mapping the data to a higher dimensional space where usually the parameters and a functional form of the kernels need to be chosen, our method scales naturally with very high dimensional data and linear algorithms are developed to learn this model. We also show that ODA and MODA are consistent generalizations of LDA, whereas regression approaches have some limitations (Hastie et al., 2001; Gallinari et al., 1991) when the number of classes increases.

3.1. Maximizing Kullback-Leibler divergence.

In this section, we derive the optimal linear dimensionality reduction for Gaussian distributed classes with different covariances. A simple measure of distance between two Gaussian distributions $N(\mathbf{x};\boldsymbol{\mu}_i,\mathbf{\Sigma}_i)$ and

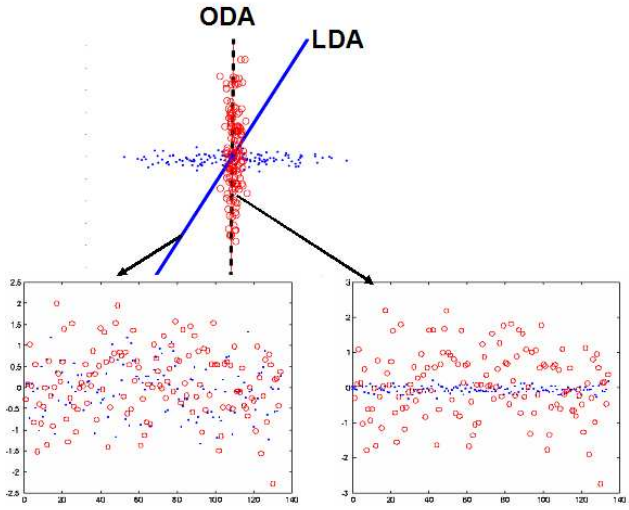

Figure 2. Projection onto LDA and ODA.

$N(\mathbf{x};\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)$ is given by the Kullback-Leibler (KL) divergence (Fukunaga, 1990):

$$\begin{aligned} KL_{ij} &= \int d\mathbf{x}\big(N(\mathbf{x};\boldsymbol{\mu}_i,\mathbf{\Sigma}_i) - N(\mathbf{x};\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)\big) \\ log\tfrac{N(\mathbf{x};\boldsymbol{\mu}_i,\mathbf{\Sigma}_i)}{N(\mathbf{x};\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)} &= tr(\mathbf{\Sigma}_i^{-1}\mathbf{\Sigma}_j + \mathbf{\Sigma}_j^{-1}\mathbf{\Sigma}_i - 2\mathbf{I}) \\ &+(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\mathbf{\Sigma}_j^{-1} + \mathbf{\Sigma}_i^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{aligned} \tag{3}$$

We assume that each class i is modeled as a gaussian $N(\boldsymbol{\mu}_i,\mathbf{\Sigma}_i)$ and the aim of ODA is to find a linear transformation $\mathbf{B} \in \Re^{d\times k}$, common to all the classes (i.e. $N(\mathbf{B}^T\boldsymbol{\mu}_i, \mathbf{B}\mathbf{\Sigma}_i\mathbf{B}^T)\ \forall i$) such that it maximizes the separability (Kullback-Leibler divergence) between the classes in the low dimensional space, that is:

$$\begin{aligned} E_3(\mathbf{B}) &= \textstyle\sum_{i=1}^{c}\sum_{j=1}^{c} KL_{ij} \propto \\ &\textstyle\sum_{i=1}^{c}\sum_{j=1}^{c} tr\big((\mathbf{B}^T\mathbf{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{\Sigma}_j\mathbf{B}) \\ &+(\mathbf{B}^T\mathbf{\Sigma}_j\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{\Sigma}_i\mathbf{B})\big) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \\ &\mathbf{B}\big((\mathbf{B}^T\mathbf{\Sigma}_j\mathbf{B})^{-1} + (\mathbf{B}^T\mathbf{\Sigma}_i\mathbf{B})^{-1}\big)\mathbf{B}^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{aligned} \tag{4}$$

After some algebraic arrangements (de la Torre & Kanade, 2005), the previous equation can be expressed in a more compact and enlightening manner:

$$\begin{aligned} G(\mathbf{B}) &= -\textstyle\sum_{i=1}^{c} tr\big((\mathbf{B}^T\mathbf{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})\big) \\ \mathbf{A}_i &= \textstyle\sum_{j\neq i}^{c}\big((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T + \mathbf{\Sigma}_j\big) \end{aligned} \tag{5}$$

Observe that a negative sign is introduced for convenience; rather than searching for a maximum, a minimum of $G(\mathbf{B})$ will be found.

Several interesting things are worth pointing out from eq. 5. If all covariances are the same (i.e. $\mathbf{\Sigma}_i = \mathbf{\Sigma}\ \forall i$), eq. 5 results in $tr\big((\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}(\mathbf{B}^T\sum_{i=1}^{c}\sum_{j\neq i}^{c}(\boldsymbol{\mu}_i -$

$\boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\mathbf{B})\big) + c(c-1)l$, which is exactly what LDA maximizes. ODA takes into account not just the distance between the means but also the orientation and magnitude of the covariance. In the LDA case, the number of extracted features cannot exceed the number of classes because the rank of $\mathbf{S}_b$ is $c-1$; however, ODA does not have this constraint and more features can be obtained. Unfortunately, due to different normalization factors $(\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}$, eq. 5 does not have a closed-form solution in terms of an eigenequation (not an eigenvalue problem).

4. Multimodal Oriented Discriminant Analsyis

In the previous section, it has been shown that ODA is the optimal linear transform for class separability in the case of Gaussian distributions with arbitrary covariances (full rank). However, in many situations the class distributions are not Gaussian. For instance, it is likely that the manifold of the facial appearance of a person under different illumination, expression, and poses is highly non-Gaussian. In this section, MODA, an extension of ODA that is able to model multimodal classes is described.

In order to model multimodal distributions, the training data for each class is first clustered using recent advances in multi-way normalized cuts (Yu & Shi, 2003). Once the input space has been clustered for each class, eq. 5 is modified to maximize the distances between the clusters of different classes, that is:

$$
\begin{aligned}
E_4(\mathbf{B}) &= -\sum_i \sum_{j\neq i} \sum_{r_1\in C_i} \sum_{r_2\in C_j} KL_{ij}^{r_1 r_2} = \\
&-\sum_i \sum_{j\neq i} \sum_{r_1\in C_i} \sum_{r_2\in C_j} tr\Big((\mathbf{B}^T\boldsymbol{\Sigma}_i^{r_1}\mathbf{B})^{-1} \\
&\mathbf{B}^T\big((\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})^T + \boldsymbol{\Sigma}_j^{r_2}\big)\mathbf{B}\Big) \qquad (6) \\
&= -\sum_i \sum_{r_1\in C_i} tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i^{r_1}\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})\big) \\
\mathbf{A}_i &= \sum_{j\neq i} \sum_{r_2\in C_j} (\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})^T + \boldsymbol{\Sigma}_j^{r_2}
\end{aligned}
$$

where $\boldsymbol{\mu}_i^{r_1}$ is the r_1 cluster of class i, and $r_1 \in C_i$ sums over all the clusters belonging to class i. $KL_{ij}^{r_1 r_2}$ denotes the Kullback-Leibler divergence between the r_1 cluster of class i and the r_2 cluster of class j. Observe that MODA looks for a projection matrix $\mathbf{B}$ which maximizes the KL divergence between clusters among all the classes, but it does not maximize the distance between the clusters of the same class.

As in the case of ODA, there is no closed expression for the maximum of eq. 6. However, if all the covariances are the same (i.e. $\boldsymbol{\Sigma}_i^{r_1} = \boldsymbol{\Sigma}\ \forall\ i, r_1$), there exists a closed form solution that can give a new insight into the method (see (de la Torre & Kanade, 2005) for more detail information).

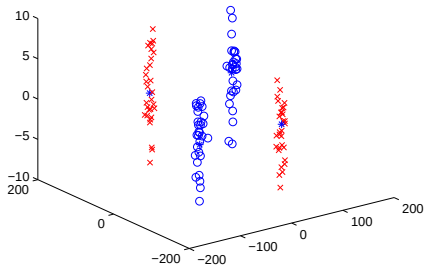

Figure 3. XOR problem

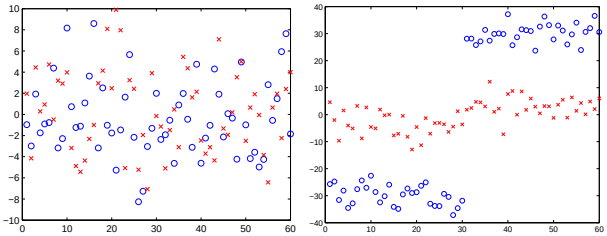

Figure 4. a) LDA b) MODA

Figure (3) shows four 3-dimensional Gaussians belonging to two classes (XOR problem). Each Gaussian has 30 samples generated with the same covariance. The means of the two classes is close to zero. Since the distribution for each class is multimodal and both classes have approximately the same mean, LDA cannot separate the classes well (fig. 4.a). Figure (4.b) shows how MODA is able to separate both classes. The figures show the projection into one dimension; the y-axis is the value of the projection and the x-axis is the sample number.

Hastie et. al (Hastie et al., 2001; Hastie et al., 1995) have introduced Mixture Discriminant Analysis (MDA) to overcome similar situations, however MODA differs in several aspects. First, it uses spectral graph methods to cluster the data because they accomodate better high dimensional data and are less prune to local minima in comparison with k-means type of algorithms. Secondly, it is not clear how MDA is able to model Gaussian distributions with different high dimensional covariances. Finally, MDA implicitely constraints the clusters of the same class to be far from each other, however MODA does not have this constraint (e.g 4.b).

5. Bound optimization

Eq. 6 is hard to optimize; second-order type of gradient methods (e.g. Newton or conjugate gradient) do not scale well with huge matrices (e.g. $\mathbf{B} \in \Re^{d\times l}$). Moreover, the second derivative of eq. 6 is quite complex. In this section, we use a bound optimization method called iterative majorization (Heiser, 1997; Leeuw, 1994; Kiers, 1995) able to monotonically reduce the value of the energy function. Although this type of optimization technique is not common in the vision/learning community, it is very similar to Expectation Maximization (EM) type of algorithms.

5.1. Iterative Majorization

Iterative majorization is a monotonically convergent method developed in the area of statistics (Heiser, 1997; Leeuw, 1994; Kiers, 1995), and it is able to solve relatively complicated problems in a straightforward manner. The main idea is to find a function, that makes it easier to minimize/maximize than the original (e.g. quadratic function) at each iteration. The first thing to do in order to minimize $G(\mathbf{B})$, eq. 6, is to find a function $L(\mathbf{B})$, which majorizes $G(\mathbf{B})$, that is, $L(\mathbf{B}) \geq G(\mathbf{B})$ and $L(\mathbf{B}_0) = G(\mathbf{B}_0)$, where $\mathbf{B}_0$ is the current estimate. The function $L(\mathbf{B})$ should be easier to minimize than $G(\mathbf{B})$. A minimum of $L(\mathbf{B})$, $\mathbf{B}_1$, is guaranteed to decrease the energy of $G(\mathbf{B})$. This is easy to show, since $L(\mathbf{B}_0) = G(\mathbf{B}_0) \geq L(\mathbf{B}_1) \geq G(\mathbf{B}_1)$. This is called the "sandwich" inequality by De Leeuw (Leeuw, 1994). Each update of the majorization will improve the value of the function, and if the function is bounded it will monotonically decrease the value of $L(\mathbf{B})$. Under these conditions it is always guaranteed to stop at a local optimum.

Iterative majorization is very similar to EM (Neal & Hinton, 1998) type of algorithms, which have been extensively used by the machine learning and computer vision communities. The EM algorithm is an iterative algorithm used to find a local maximum of the log likelihood, $log\ p(\mathbf{D}|\boldsymbol{\theta})$, where $\mathbf{D}$ is the data, $\boldsymbol{\theta}$ are the parameters. Rather than maximizing the log likelihood directly, EM uses Jensen's inequality to find a lower bound $log\ p(\mathbf{D}|\boldsymbol{\theta}) = log \int\ q(\mathbf{h}) \frac{p(\mathbf{D},\mathbf{h}|\boldsymbol{\theta})}{q(\mathbf{h})} d\mathbf{h} \geq \int\ q(\mathbf{h}) log \frac{p(\mathbf{D},\mathbf{h}|\boldsymbol{\theta})}{q(\mathbf{h})} d\mathbf{h}$, which holds for any distribution $q(\mathbf{h})$. The Expectation step, performs a functional approximation on this lower bound, that is, it finds the distribution $q(\mathbf{h})$, which maximizes the data and touches the log likelihood at the current parameter estimates $\boldsymbol{\theta}_n$. In fact, the optimal $q(\mathbf{h})$ is the posterior probability of the latent/hidden parameters given the data (i.e. $p(\mathbf{h}|\mathbf{D})$). The Maximization step maximizes the lower-bound w.r.t the parameters $\boldsymbol{\theta}$. The E-step in EM would be equivalent to the construction of the majorization function and the M-step just minimizes/maximizes this upper/lower bound.

5.2. Constructing a majorization function

In order to find a function which majorizes $G(\mathbf{B})$, the following inequality is used (Kiers, 1995), $||(\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-\frac{1}{2}}\mathbf{B}^T\mathbf{A}_i^{\frac{1}{2}} - (\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{\frac{1}{2}}(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)\mathbf{B}_n^T\mathbf{A}_i^{\frac{1}{2}}||_F \geq 0$, where we have assumed that the factorizations of $\mathbf{A}_i$ and $\mathbf{B}_i$ are possible, that is, $\mathbf{A}_i = \mathbf{A}_i^{\frac{1}{2}}\mathbf{A}_i^{\frac{1}{2}}$ and $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}$. Rearranging the previous equation derives in (apply $||\mathbf{A}||_F = tr(\mathbf{A}^T\mathbf{A})$):

$$\begin{aligned} tr((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})) \geq 2tr((\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1})(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B})) \\ -tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B}_n)(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}\big) \end{aligned} \quad (7)$$

By adding a sum to both sides of this inequality a function $L(\mathbf{B})$ which majorizes $G(\mathbf{B})$ is obtained:

$$\begin{aligned} G(\mathbf{B}) = -\textstyle\sum_i tr((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})) \leq L(\mathbf{B}) = \\ -\textstyle\sum_i 2tr((\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1})(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B})) + \\ tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B}_n)(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}\big) \end{aligned} \quad (8)$$

Effectively, it can easily shown that $L(\mathbf{B})$ majorizes $G(\mathbf{B})$ since $G(\mathbf{B}_n) = L(\mathbf{B}_n)$ and $L(\mathbf{B}) \geq G(\mathbf{B})$.

The function $L(\mathbf{B})$ is quadratic in $\mathbf{B}$ and hence easier to minimize. After rearranging terms a necessary condition for the minimum of $L(\mathbf{B})$ has to satisfy:

$$\begin{aligned} \tfrac{\partial L}{\partial \mathbf{B}} = \textstyle\sum_i -\mathbf{T}_i + \boldsymbol{\Sigma}_i\mathbf{B}\mathbf{F}_i = \mathbf{0} \\ \mathbf{F}_i = (\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B}_n)(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1} \\ \mathbf{T}_i = \mathbf{A}_i^T\mathbf{B}_n^T(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1} \end{aligned} \quad (9)$$

See (J. R. Magnus, 1999) for matrix derivatives. Finding the solution of eq. 9 involves solving the following system of linear equations $\sum_i \mathbf{T}_i = \sum_i \boldsymbol{\Sigma}_i\mathbf{B}\mathbf{F}_i$. A closed-form solution could be achieved by vectorizing eq. 9 with Kronecker products. However, the system would have dimensions of $(d\times l)\times(d\times l)$, which is not efficient in either space or time. Instead, a gradient descent algorithm which minimizes:

$$E_5(\mathbf{B}) = min_{\mathbf{B}}||\sum_i(\mathbf{T}_i - \boldsymbol{\Sigma}_i\mathbf{B}\mathbf{F}_i)||_F \quad (10)$$

is used. Due to the huge number of the equations to solve $(d \times l)$, an effective and linear time algorithm to solve for the optimum of eq. 10 is a normalized gradient descent:

$$\begin{aligned} \mathbf{B}^{n+1} = \mathbf{B}^n - \eta\tfrac{\partial E_5(\mathbf{B})}{\partial \mathbf{B}} \quad \mathbf{R}_k = \tfrac{\partial E_5(\mathbf{B})}{\partial \mathbf{B}} \\ \mathbf{R}_k = -\textstyle\sum_i \boldsymbol{\Sigma}_i\mathbf{B}\mathbf{F}_i^T + \sum_i\sum_j \boldsymbol{\Sigma}_i^T\boldsymbol{\Sigma}_j\mathbf{B}\mathbf{F}_i\mathbf{F}_j^T \end{aligned} \quad (11)$$

η is the step size needed to converge and it is estimated by minimizing $\eta = min_\eta || \sum_i \left(\mathbf{T}_i - \mathbf{\Sigma}_i(\mathbf{B} + \eta \mathbf{R}_k)\mathbf{F}_i \right) ||$. After some derivation, it can be shown that the optimal η is $\eta = \frac{\sum_i \sum_j tr(\mathbf{\Sigma}_i \mathbf{R}_k \mathbf{T}_i \mathbf{T}_j^T \mathbf{B}^T \mathbf{\Sigma}_j) - \sum_i tr\left(\mathbf{\Sigma}_i \mathbf{R}_k \mathbf{T}_i \mathbf{F}_i\right)}{\sum_i \sum_j tr(\mathbf{\Sigma}_i \mathbf{R}_i \mathbf{T}_i \mathbf{T}_j^T \mathbf{R}_j^T \mathbf{\Sigma}_j)}$ (de la Torre & Kanade, 2005).

6. Dealing with high dimensional data

When applying any classifier to visual data, a major problem is the high dimensionality of the images. Several strategies are necessary to get good generalization, such as feature selection or dimensionality reduction techniques (PCA, LDA, etc). In this context LDA or MODA can be a good initial step to extract discriminative features. However, as it is well known, dimensionality reduction techniques such as LDA, that preserve discriminative power cannot handle very well the case that $n << d$ (more pixels than training data), which is the typical. For instance, an image of 100×100 pixels will correspond to feature vectors of 10000 dimensions, which will induce covariance matrices of 10000×10000. To make the covariance full rank, at least 10000 independent samples would be necessary, and even that will be a poor estimate. In this scenario, working with huge covariance matrices presents two major problems: computational tractability (storage, efficiency and rank decificiency) and generalization.

To solve the computational aspect, one straightforward approach is to realize that if $d >> n$, the true dimensionality of $\mathbf{D} \in \Re^{d \times n}$ is n. Therefore, we can project into the first n principal components without losing any discriminative power. Besides the computational aspects, the second and more important problem is the lack of generalization when too few samples are available. As noticed by Hugues (Hughes, 1968), the fact of increasing the dimensionality would have to enhance performance for recognition (more information is added), but due to the lack of training data this will rarely occur. Fukunaga (Fukunaga & Hayes, 1989) studied the effects of finite data set in linear and quadratic classifiers, and concluded that the number of samples should be proportional to the dimension for linear classifiers and square for quadratic classifiers. In this case, LDA over-fits the data and does not generalize well to new samples. One way to understand over-fitting is to consider eq. 2. There are $O(k \times n)$ equations and $O(d \times k)$ unknowns ($\mathbf{B}$) [3]. Without enough training data, eq. 2 is an underdetermined system of equations with ∞ solutions. In other words, if there are more features than training samples, directly minimizing LDA will result in a dimensionality reduction that will act as a associative memory rather than learning anything (no regression is done), and no good generalization will be achieved.

In order to be able to generalize better than LDA and not suffer from storage/computational requirements, our solution approximates the covariance matrices as the sum of outer products plus a scaled identity matrix $\mathbf{\Sigma}_i \approx \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T + \sigma_i^2 \mathbf{I}_d$. $\mathbf{U}_i \in \Re^{d \times l}$, $\mathbf{\Lambda}_i \in \Re^{l \times l}$ is a diagonal matrix. The parameters σ_i^2, $\mathbf{U}_i$, $\mathbf{\Lambda}_i$ are estimated following a fitting approach which minimizes $E_c(\mathbf{U}_i, \mathbf{\Lambda}_i, \sigma_i^2) = ||\mathbf{\Sigma}_i - \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T - \sigma_i^2 \mathbf{I}_d||_F$. After opitmizing parameters, it can be shown (de la Torre & Kanade, 2005) that: $\sigma_i^2 = tr(\mathbf{\Sigma}_i - \mathbf{U}_i \hat{\mathbf{\Lambda}}_i \mathbf{U}_i^T)/d - l$, $\mathbf{\Lambda}_i = \hat{\mathbf{\Lambda}}_i - \sigma_i^2 \mathbf{I}_d$, where $\hat{\mathbf{\Lambda}}_i$ are the eigenvalues of the covariance matrix $\mathbf{\Sigma}_i$ and $\mathbf{U}_i$ the eigenvectors. The same expression could be derived using probabilistic PCA (Moghaddam & Pentland, 1997; Tipping & Bishop, 1999).

It is worthwhile to point out two important aspects of the previous factorizations. Factorizing the covariance as the sum of outer products and a diagonal matrix is an efficient (in space and time) manner to deal with the small sample case. Observe that to compute $\mathbf{\Sigma}_i \mathbf{B} = \mathbf{U}_i \mathbf{\Lambda}_i (\mathbf{U}_i^T \mathbf{B}) + \sigma_i^2 \mathbf{B}$ storing/computing the full $d \times d$ covariance is not required. On the other hand, the original covariance has $d(d+1)/2$ free parameters, and after the factorization the number of parameters is reduced to $l(2d - l + 1)/2$ (assuming orthogonality of $\mathbf{U}_i$), so that much less data is needed to estimate these parameters and hence it is not so prone to over-fitting. Also, the spectral properties of the matrix are not altered; the eigenvectors of $\mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T + \sigma_i^2 \mathbf{I}_d$ are the same as $\mathbf{\Sigma}_i$, and the set of eigenvalues will be $\zeta_1 = \sigma_i^2 + \lambda_1$, $\zeta_2 = \sigma_i^2 + \lambda_2$, $\zeta_{(l+1)} = \sigma_i^2, \cdots, \zeta_d = \sigma_i^2$, where λ_i are the eigenvalues of the sample covariance.

7. Experiments

7.1. Toy Problem

In order to verify that under ideal conditions ODA outperforms LDA, we tested ODA on a toy problem. 200 samples for five 20-dimensional (d=20) Gaussian classes were generated. Each sample for class c was generated as $\mathbf{y}_i = \mathbf{B}_c \mathbf{c} + \boldsymbol{\mu}_c + \mathbf{n}$, where $\mathbf{y}_i \in \Re^{20 \times 1}$, $\mathbf{B}_c \in \Re^{20 \times 7}$, $\mathbf{c} \sim N_7(\mathbf{0}, \mathbf{I})$ and $\mathbf{n} \sim N_{20}(\mathbf{0}, 2\mathbf{I})$. The means of each class are $\boldsymbol{\mu}_1 = 4\mathbf{1}_{20}$, $\boldsymbol{\mu}_2 = \mathbf{0}_{20}$, $\boldsymbol{\mu}_3 = -4[\mathbf{0}_{10}\ \mathbf{1}_{10}]^T$, $\boldsymbol{\mu}_4 = 4[\mathbf{1}_{10}\ \mathbf{0}_{10}]^T$ *and* $\boldsymbol{\mu}_5 = 4[\mathbf{1}_5\ \mathbf{0}_5\ \mathbf{1}_5\ \mathbf{0}_5]^T$. The basis $\mathbf{B}_c$ are random matrices, where each element has been generated from $N(0, 5)$. A weak orthogonality between the covariance matrices (i.e. $tr(\mathbf{B}_i^T \mathbf{B}_j) = 0\ \forall i \neq j$) is imposed with a Gram-Schmidt approach, i.e. $\mathbf{B}_j = \mathbf{B}_j -$

[3] Orthogonality of $\mathbf{B}$ is not assumed.

$\sum_{i=1}^{j-1} tr((\mathbf{B}_i\mathbf{B}_i)^{-1}\mathbf{B}_j^T\mathbf{B}_i)\mathbf{B}_i \ \forall j = 2\cdots 5$. The covariance matrices are approximated as $\mathbf{\Sigma}_i = \mathbf{U}_i\mathbf{\Lambda}_i\mathbf{U}_i^T + \sigma_i^2\mathbf{I}$, such that they preserve 90% of the energy.

In the test set, a linear classifier is used, that is, a new sample $\mathbf{d}_i$ is projected into the subspace by $\mathbf{x}_i = \mathbf{B}^T\mathbf{d}_i$ and it is assigned to the class that has smallest distance, $(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i)\hat{\mathbf{\Sigma}}_i^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i) + log|\hat{\mathbf{\Sigma}}_i|$, where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{\Sigma}}_i$ are the low-dimensional estimates of the mean and class covariance. Table 1 shows the average recognition rate of LDA and ODA over 50 trials. For each trial and each basis, the algorithm is run five times from different initial conditions (perturbing the LDA solution), and the best solution is chosen. As can observed from table 1, ODA always outperforms LDA and it is able to extract more features.

Basis	1	2	3	4	5	6
LDA	0.46	0.69	0.74	0.78	NA	NA
ODA	0.46	0.77	0.85	0.90	0.94	0.97

Table 1. Average over 50 trials

It is well known, that in the case of having a small number of samples, classical PCA can outperform LDA (Martinez & Kak, 2003). We run the same experiment as before but with a feature size of 152 (i.e. d=152) and just 40 samples per class. The results can be seen in table 2.

Basis	1	2	3	4	5	6
PCA	0.20	0.42	0.53	0.66	0.75	0.82
LDA	0.20	0.37	0.57	0.78	NA	NA
ODA	0.20	0.67	0.81	0.90	0.95	0.97
PCLDA	0.20	0.50	0.79	0.85	NA	NA
PCODA	0.20	0.70	0.84	0.91	0.95	0.97

Table 2. Average over 50 trials

PCLDA holds for PCA+LDA (preserving 95% of the energy) and PCMODA for PCA+ODA. Even, in the small sample case, ODA still outperforms all the other methods. Also, by projecting onto PCA, LDA avoids overfitting.

7.2. Face Recognition from Video

Face recognition is one of the classical pattern recognition problems that suffers from noise, limited number of training data and the face under pose/illumination changes describes non-linear manifolds. These facts make face recognition a good candidate for MODA.

A face database has been collected using our omnidirectional-meeting-capturing device (de la Torre et al., 2005). The database consist on 23 people recorded over two different days under different illumintation conditions. Figure 5 shows images of some people in the database, variations are due to facial expression, pose, scale and illumination conditions. The training set consists of the data gathered on the first day under three different illumination conditions (varying lights in the recording room), scale and expression changes. We have around 500 images per person in the training set and a similar number for the testing.The testing data consist of the recordings of the second day (a couple of weeks later) under similar conditions. Figure 6 illustrates the recognition performance using PCA, LDA and MODA, similarly table 3 provides some detailed numerical values for different number of basis.

Figure 5. Some training samples for 10 classes.

In this experiment, each class has been clustered into two clusters to estimate $\mathbf{B}$. Once $\mathbf{B}$ is calculated, the Euclidean distance for the nearest neighbourhood is used. Several metrics have been tested (e.g. Mahalanobis, Euclidean, Cosine, etc) and the Euclidean distance performed the best in our experiments. For the same number of bases, MODA outperforms PCA/LDA. Also, observe that LDA can extract only classes-1 features (22 features), whereas MODA can extract many more features. In this experiment, each sample is classified independently; however, using temporal information can greatly improve the recognition performance; Refer to (de la Torre et al., 2005) for more details.

Basis	2	5	10	20	30	50
PCA	0.12	0.26	0.43	0.55	0.58	0.59
LDA	0.21	0.36	0.48	0.56	NA	NA
MODA	0.23	0.38	0.50	0.59	0.61	0.63

Table 3. Recognition performance of PCA/LDA/MODA

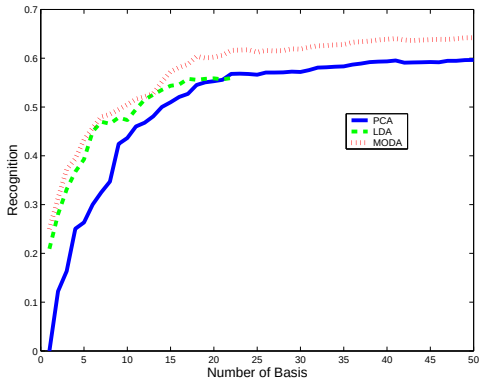

Figure 6. PCA/LDA/MODA

Acknowledgments

This work has been partially supported by National Business Center of the Department of the Interior under a subcontract from SRI International and U.S. Department of Defense contract N41756-03-C4024, NIMH Grant R01 MH51435.

References

Campbell, N. A. (1984). Canonical variate analysis - a general formulation. *Australian Journal of Statistics*, *26*, 86–96.

de la Torre, F., & Kanade, T. (2005). Multimodal oriented discriminant analysis. *tech. report CMU-RI-TR-05-03, Robotics Institute, Carnegie Mellon University, January 2005.*

de la Torre, F., Vallespi, C., Rybski, P. E., Veloso, M., & Kanade, T. (2005). Omnidirectional video capturing, multiple people tracking and recognition for meeting monitoring.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification.* John Wiley and Sons Inc.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition, second edition.* Academic Press.Boston, MA.

Fukunaga, K., & Hayes, R. (1989). Effects of sample size in classifier desing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*.

Gallinari, P., Thiria, S., Badran, F., & Fogelman-Soulie, F. (1991). On the relations between discriminant analysis and multilayer perceptrons. *Neural Networks*, *4*, 349–360.

Hastie, T., Tibshirani, R., & Buja, A. (1995). Flexible discriminant and mixture models. *Neural Networks and Statistics. J. Kay and D. Titterington, Eds.*

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction.* Springer.

Heiser, J. (1997). *Convergent computation by iterative majorization; theory and applications in multidimensional data analysis.* Krzanowski ed. Oxford University Press.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognition,. *IEEE Transactions on Information Theory*, *14*, 55–63.

J. R. Magnus, H. N. (1999). *Matrix differential calculus with applications in statistics and econometrics.* John Wiley.

Kiers, H. A. L. (1995). Maximization of sums of quotients of quadratic forms and some generalizations. *Psychometrika*, *60*, 221–245.

Kumar, N., & Andreou, A. (1998). Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, *26*, 283 – 297.

Leeuw, J. D. (1994). *Block relaxation algorihtms in statistics.* H.H. Bock, W. Lenski, M. Ritcher eds. Information Systems and Data Analysis. Springer-Verlag.

Martinez, A., & Kak, A. (2003). Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 228–233.

Moghaddam, B., & Pentland, A. (1997). Probabilistic visual learning for object representation. *Pattern Analysis and Machine Intelligence*, *19*, 137–143.

Neal, R., & Hinton, G. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models.* Kluwer.

Saon, G., Padmanabhan, M., Gopinath, R., & Chen, S. (2000). Maximum likelihood discriminant feature spaces. *ICASSP.*

Tipping, M., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, *61*, 611–622.

Yu, H., & Yang, J. (2001). A direct lda algorith for high-dimensional data– with applicatins to face recognition. *Pattern Recognition*, *34*, 2067–2070.

Yu, S., & Shi, J. (2003). Multiclass spectral clustering. *ICCV.*