Multimodal Oriented Discriminant Analysis

Fernando De la Torre Takeo Kanade

CMU-RI-TR-05-03

January 2005

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

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), an 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 LDA techniques.

Contents

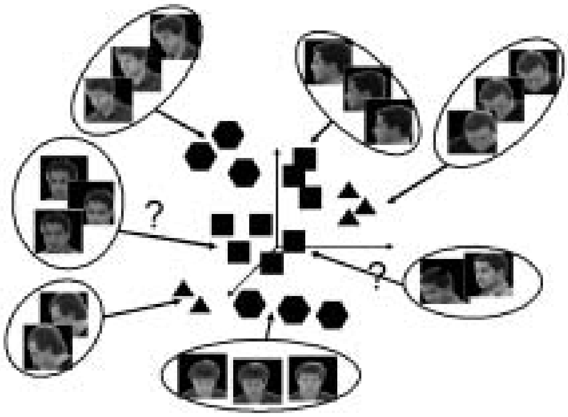

Figure 1: Classification of face images from video sequences by projecting onto a low dimensional space. Observe that the face distributions can be non-gaussians and with different covariances.

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. In particular, SM have been very successful in computer vision to solve problems such as structure from motion [29], detection/recognition [30] or face tracking [5, 23]. 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[1]. 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.

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”. In particular, LDA has been extensively used for classification problems such as speech/face recognition or multimedia information retrieval [4, 2, 9, 12, 31, 32, 34, 22]. However, there exist several liminations of current LDA techniques. LDA is optimal only in the case that all the classes are Gaussian distributed with equal covariances (multimodal distributions are not modeled). Due to this assumption, the maximum number of features that can be extracted is the number of classes-1. Another common problem in computer vision applications is the small size problem [32, 34], that is, the training set

[1] Also many times it is helpful to find a new coordinate system (e.g. Fourier transform).

has more ”dimensions” (pixels) than data samples [2]. In this situation LDA overfits and PCA techniques usually outperform LDA [22]. 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 not 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 most discriminant analysis methods is to project the data into a space of lower dimension, so that the classes are as compact and as far as possible from each other. Several optimization criteria are possible to compute LDA, and most of them are based on relations between the following covariance matrices, which can be expressed conveniently in matrix form as[3]:

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

where $\mathbf{D}\in\Re^{d\times n}$ is the data matrix, $\mathbf{m}=\frac{1}{n}\mathbf{D}\mathbf{1}_n$ is the mean vector for all the classes and $\mathbf{m}_i$ is the mean vector for the class i. $\mathbf{P}_i$ are projection matrices (i.e $\mathbf{P}_i^T=\mathbf{P}_i$ and $\mathbf{P}_i^2=\mathbf{P}_i$) with the following expressions:

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

$\mathbf{G}\in\Re^{n\times c}$ 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. c denotes the number of classes and n the number

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

[3] 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 $\mathbf{D}$. d_{ij} denotes the scalar in the row i and column j of the matrix $\mathbf{D}$ and the scalar i-th element of a column vector $\mathbf{d}_j$. d_{ji} is the i-th scalar element of the vector $\mathbf{d}^j$. 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. $\mathbf{e}_i$ is the i column of the identity matrix (i.e. $[0\,0\,0\cdots1\cdots0\,0]^T$), $N_d(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ indicates a d-dimensional Gaussian on the variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

of samples. $\mathbf{S}_b$ 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 [12]. 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 $\mathbf{S}_2 = \{\mathbf{S}_w, \mathbf{S}_t, \mathbf{S}_w\}$. Other constrained optimization formulations are possible [12]. 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 $\mathbf{B}^T\mathbf{S}_1\mathbf{B} = \Lambda$).

2.1 Another view onto LDA

Previous Rayleigh quotient optimization procedures are not easy to modify when incorporating new constraints (e.g temporal constraints or geometric invariance). Consider the following weighted between-class covariance matrix, $\hat{\mathbf{S}}_b = \mathbf{D}\mathbf{G}\mathbf{G}^T\mathbf{D}^T = \sum_{i=1}^{C}(\frac{n_i}{n})^2(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$, that favors the classes with more samples. Following previous work on neural networks [14, 21], it can be shown that maximizing $J_4(\mathbf{B}) = tr((\mathbf{B}^T\mathbf{S}_t\mathbf{B})^{-1}\mathbf{B}^T\hat{\mathbf{S}}_b\mathbf{B})$ is equivalent to minimize:

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

Optimizing over $\mathbf{V}$ results in $E(\mathbf{B}) = ||\mathbf{G}^T - \mathbf{G}^T\mathbf{D}^T\mathbf{B}(\mathbf{B}^T\mathbf{D}\mathbf{D}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{D}||$, and after some arrangements it can be shown [14, 21] that $E(\mathbf{B})$ is proportional to $-tr(((\mathbf{B}^T\mathbf{D}\mathbf{D}^T\mathbf{B})^{-1})\mathbf{B}^T\mathbf{D}\mathbf{G}\mathbf{G}^T\mathbf{D}^T\mathbf{B})$. In this case, it is assumed that $\mathbf{D}$ is zero mean and that $rank(\mathbf{D}) = d < n$. This approach is appealing for several reasons. 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. Also, Baldi and Hornik have shown that the surface has a unique local minimum [1], although several inflexion points. Observe, that the LDA problem is posed as one of hetero-associative memory, which could be solve efficiently in small data cases with the generalized SVD [8]. Finally gradient descent methods could be applied efficiently to optimize 2 where there is a great deal of data.

On the other hand, if $\mathbf{D}$ is zero mean, and all the classes are equally probable, LDA can be computed by maximizing:

$$E(\mathbf{B}) = max_{\mathbf{B}}\frac{tr(\mathbf{B}^T\mathbf{M}\mathbf{M}^T\mathbf{B})}{tr(\mathbf{B}^T\mathbf{D}\mathbf{D}^T\mathbf{B})} \tag{3}$$

where $\mathbf{M} \in \Re^{d\times c}$ is a matrix, such that each column, $\mathbf{m}_i$ contains the mean of the class i. In the previous expression it is possible to introduce two auxiliary variables $\mathbf{C}_1, \mathbf{C}_2$,

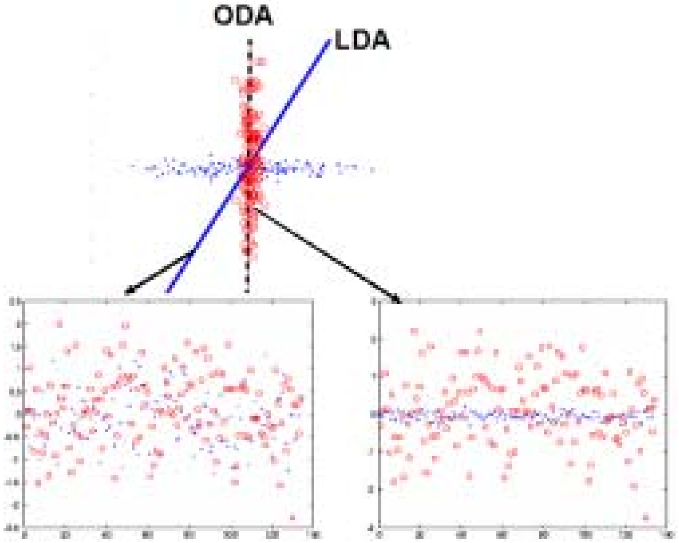

Figure 2: Projection onto LDA direction and ODA.

which would give us a new insight into the LDA problem, that is:

$$E(\mathbf{B}, \mathbf{C}_1, \mathbf{C}_2) = min_{\mathbf{B}, \mathbf{C}_1, \mathbf{C}_2} \frac{||\mathbf{M} - \mathbf{B}\mathbf{C}_1||_F}{||\mathbf{D} - \mathbf{B}\mathbf{C}_2||_F} \tag{4}$$

where $||.||_F$ denotes the Frobenious norm of a matrix (valid for any unitary invariant norm). Eq. 4 shows that LDA can be expressed as the ratio of two generative models; the denominator preserves the subspace of the distances between the centers, whereas the denominator is optimal for the null space of the data (which is not in the direction of the mean of the classes). Using the formulation of eq. 4 robustness to sample outliers could be introduced as in [6]. Alternatively Fidler and Leonardis [10] achieve robustness to intra-sample outliers using subsampling.

3 Oriented Discriminant Analysis

LDA is the optimal discriminative projection only in the case of having Gaussian classes with equal covariance matrix [3, 9] (assuming enough training data). LDA will not be optimal if the classes have different covariances. Fig. 2 shows one 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 [3] derives a maximum likelihood approach to discriminant analysis. Assuming that all 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 [19] 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.* [27] define a new energy function to model the directionality of the data, $J(\mathbf{B}) = \prod_{i=1}^{c}(\frac{|\mathbf{B}^T\mathbf{S}_b\mathbf{B}|}{|\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B}|})^{n_i}$, where $\boldsymbol{\Sigma}_i$ is the class covariance matrix and $\mathbf{S}_b$ the between-class scatter covariance matrix. In this paper, we extend previous approaches by deriving a probabilistic interpretation of the optimal discriminant analysis in the case of having classes with different covariances, and multimodal distributions. Also, our method scales well with high dimensional data and efficient algorithms are developed.

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 $N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ and $N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ is given by the Kullback-Leibler (KL) divergence [12]:

$$\begin{aligned} KL_{ij} &= \frac{1}{2}\int d\mathbf{x}\big(N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) - N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\big) log\frac{N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} \\ &= \frac{1}{2}tr(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i - 2\mathbf{I}) + \frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\Sigma}_j^{-1}+\boldsymbol{\Sigma}_i^{-1})(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \end{aligned} \tag{5}$$

The aim of ODA is to find a linear transformation $\mathbf{B}$, common to all the classes (i.e. $N(\mathbf{B}^T\boldsymbol{\mu}_i,\mathbf{B}\boldsymbol{\Sigma}_i\mathbf{B}^T)\ \forall i$), such that it maximizes the separability between the classes in the low dimensional space, that is :

$$\begin{aligned} E(\mathbf{B}) = \textstyle\sum_{i=1}^{c}\sum_{j=1}^{c} KL_{ij} \propto \sum_{i=1}^{c}\sum_{j=1}^{c} tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\boldsymbol{\Sigma}_j\mathbf{B}) \\ +(\mathbf{B}^T\boldsymbol{\Sigma}_j\mathbf{B})^{-1}(\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})\big) + (\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T \\ \mathbf{B}\big((\mathbf{B}^T\boldsymbol{\Sigma}_j\mathbf{B})^{-1} + (\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}\big)\mathbf{B}^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \end{aligned} \tag{6}$$

After some simple algebraic arrangements, the previous equation can be expressed in a more compact and enlightening manner:

$$\begin{aligned} G(\mathbf{B}) &= -\textstyle\sum_{i=1}^{c} tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})\big) \\ \mathbf{A}_i &= \textstyle\sum_{j\neq i}^{c}\big((\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T + \boldsymbol{\Sigma}_j\big) \end{aligned} \tag{7}$$

Observe that a negative sign is introduced for convenience; rather than searching for a maximum, a minimum of $G(\mathbf{B})$ will be found. $\mathbf{A}_i$ can be rewritten as: $\mathbf{A}_i = \mathbf{M}\mathbf{P}_i\mathbf{M}^T + \sum_{j\neq i}^{c}\boldsymbol{\Sigma}_j$, where $\mathbf{M}\in\mathbf{R}^{d\times c}$ is a matrix such that each column is the mean of each class, and $\mathbf{P}_i = \mathbf{I}_c + c\mathbf{e}_i\mathbf{e}_i^T - \mathbf{e}_i\mathbf{1}_c^T - \mathbf{1}_c\mathbf{e}_i^T \in \mathbf{R}^{c\times c}$. Several interesting things are worth pointing out from eq. 7. If all covariances are the same (i.e. $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}\ \forall i$), eq. 7 results in $tr\big((\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}(\mathbf{B}^T\sum_{i=1}^{c}\sum_{j\neq i}^{c}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\mathbf{B})\big) + c(c-1)l$, 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 $(\mathbf{B}^T\mathbf{\Sigma}_i\mathbf{B})^{-1}$, eq. 7 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 [33]. Fig. 3.a shows an example of clustering a set of faces from a video sequence, each row is a cluster which mostly corresponds to different poses. Once the input space has been clustered for each class, eq. 7 is modified to maximize the distances between the clusters of different classes, that is:

$$\begin{aligned} E(\mathbf{B}) &= \tfrac{-1}{2}\textstyle\sum_i \sum_{j\neq i} \sum_{r_1\in C_i} \sum_{r_2\in C_j} tr\Bigg((\mathbf{B}^T\mathbf{\Sigma}_i^{r_1}\mathbf{B})^{-1} \\ &\qquad \mathbf{B}^T\big((\boldsymbol{\mu}_i^{r_1}-\boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1}-\boldsymbol{\mu}_j^{r_2})^T+\mathbf{\Sigma}_j^{r_2}\big)\mathbf{B}\Bigg) \\ &= \tfrac{-1}{2}\textstyle\sum_i \sum_{r_1\in C_i} tr\big((\mathbf{B}^T\mathbf{\Sigma}_i^{r_1}\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})\big) \\ \mathbf{A}_i &= \textstyle\sum_{j\neq i}\sum_{r_2\in C_j}(\boldsymbol{\mu}_i^{r_1}-\boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1}-\boldsymbol{\mu}_j^{r_2})^T+\mathbf{\Sigma}_j^{r_2} \end{aligned} \tag{8}$$

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. Observe that MODA looks for a projection $\mathbf{B}$ 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. 8. However, if all the covariances are the same (i.e. $\mathbf{\Sigma}_i^{r_1} = \mathbf{\Sigma}\ \forall\ i, r_1$), there exists a closed form solution that can give a new insight into the method. In appendix A, it is shown that in this case, eq. 8 becomes $2Ktr\big((\mathbf{M}^T\mathbf{M}(\mathbf{P}_M - \sum_i k_i diag(\mathbf{g}_i) - \mathbf{G}\mathbf{G}^T)(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\big)$, which has a closed-form solution.

Figure (3.b) 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.

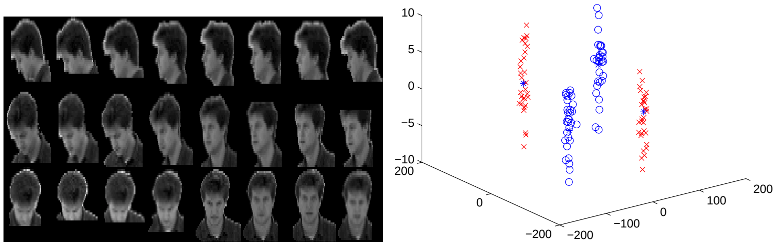

Figure 3: a) Cluster in different poses. b)XOR problem

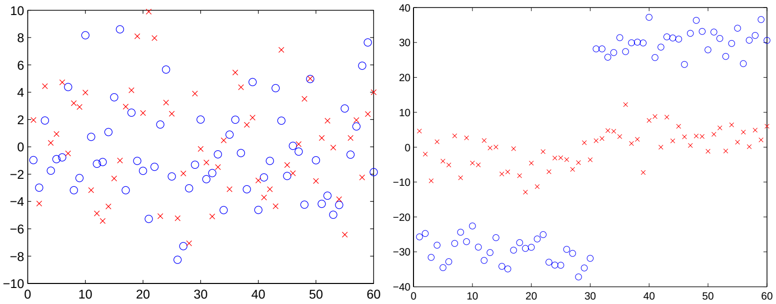

Figure 4: a) LDA b) MODA

5 Bound optimization

Eq. 8 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. 8 is quite complex. In this section, we use a bound optimization method called iterative majorization [15, 20, 18] 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 [15, 20, 18], 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. 8, 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 [20]. 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 [25] 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 [18], $||(\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-\frac{1}{2}}\mathbf{B}^T\mathbf{A}_i^{\frac{1}{2}} - (\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{\frac{1}{2}}(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)\mathbf{B}_n^T\mathbf{A}_i^{\frac{1}{2}}||_F \geq 0$, 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 previous equation derives in:

$$\begin{aligned} tr((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})) &\geq 2tr((\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1})(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B})) \\ &-tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B}_n)(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}\big) \end{aligned} \tag{9}$$

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

$$\begin{aligned} G(\mathbf{B}) = -\textstyle\sum_i tr((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}_i\mathbf{B})) \leq L(\mathbf{B}) = -\textstyle\sum_i 2tr((\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1})(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B})) + \\ tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B}_n)(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}\big) \end{aligned} \tag{10}$$

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:

$$\begin{aligned} \tfrac{\partial L}{\partial \mathbf{B}} &= \textstyle\sum_i -\mathbf{T}_i + \boldsymbol{\Sigma}_i\mathbf{B}\mathbf{F}_i = \mathbf{0} \\ \mathbf{F}_i &= (\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1}(\mathbf{B}_n^T\mathbf{A}_i\mathbf{B}_n)(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1} \\ \mathbf{T}_i &= \mathbf{A}_i^T\mathbf{B}_n^T(\mathbf{B}_n^T\boldsymbol{\Sigma}_i\mathbf{B}_n)^{-1} \end{aligned} \tag{11}$$

Finding the solution of eq. 11 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. 11 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:

$$E(\mathbf{B}) = min_{\mathbf{B}} || \sum_i (\mathbf{T}_i - \mathbf{\Sigma}_i \mathbf{B} \mathbf{F}_i) ||_F \tag{12}$$

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. 12 is a normalized gradient descent:

$$\begin{gathered} \mathbf{B}^{n+1} = \mathbf{B}^n - \eta \tfrac{\partial E(\mathbf{B})}{\partial \mathbf{B}} \;\; \mathbf{R}_k = \tfrac{\partial E(\mathbf{B})}{\partial \mathbf{B}} \\ \mathbf{R}_k = -\textstyle\sum_i \mathbf{\Sigma}_i \mathbf{B} \mathbf{F}_i^T + \sum_i \sum_k \mathbf{\Sigma}_i^T \mathbf{\Sigma}_k \mathbf{B} \mathbf{F}_i \mathbf{F}_k^T \end{gathered} \tag{13}$$

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

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. A more interesting approach, Direct LDA methods [32, 4], discard the null space of $\mathbf{S}_1$, which contains no discriminative information (i.e. $\mathbf{S}_1 \mathbf{B} = \mathbf{0}$), and then find the transformation that diagonalizes $\mathbf{S}_2$.

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 [17], 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 [13] 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. A similar conclusion has been obtained by Raudys and Jain [26], that the complexity of the classifier increases exponentially with the dimensionality of the data. 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 $O(c \times n)$ equations and $O(d \times k)$ unknowns ($\mathbf{B}$) [4]. 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.

Several regularized solutions have been proposed in order to alleviate the lack of training data [34, 16]. Hoffbeck and Landgrebe [16] have proposed a combination of class covariance and common covariance matrix, that is, the new covariance matrix $C(\mathbf{\Sigma}_i)$ will be, $C(\mathbf{\Sigma}_i) = \alpha_1 diag(\mathbf{\Sigma}_i) + \alpha_2 \mathbf{\Sigma}_i + \alpha_3 \mathbf{S} + \alpha_4 diag(\mathbf{S})$, where $\mathbf{S} = \frac{1}{L}\sum_i^L \mathbf{\Sigma}_i$ and $\sum_i \alpha_i = 1$. Zhao [34] suggested adding a regularization term $\Sigma_b + k\mathbf{I}_d$, where k is a small constant this will modify only the eigenvalues and will preserve the same directions (eigenvectors).

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 $\mathbf{\Sigma}_i \approx \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T + \sigma_i^2 \mathbf{I}_d$. $\mathbf{U}_i \in \Re^{d \times k}$, $\mathbf{\Lambda}_i \in \Re^{k \times k}$ is a diagonal matrix. In order to estimate the parameters σ_i^2, $\mathbf{U}_i$, $\mathbf{\Lambda}_i$, a fitting approach is followed by minimizing $E_c(\mathbf{U}_i, \mathbf{\Lambda}_i, \sigma_i^2) = ||\mathbf{\Sigma}_i - \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T - \sigma_i^2 \mathbf{I}_d||_F$. By making derivatives w.r.t $\mathbf{U}_i, \sigma_i^2$ and $\mathbf{\Lambda}_i$ and setting them to zero, it is easy to show that the parameters have to satisfy:

$$\mathbf{U}_i \mathbf{\Lambda}_i = (\mathbf{\Sigma}_i - \sigma_i^2 \mathbf{I}_d)\mathbf{U}_i \quad \sigma_i^2 = \frac{tr(\mathbf{\Sigma}_i - \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T)}{tr(\mathbf{I}_d)} \tag{14}$$

Taking into account that $\mathbf{\Sigma}_i$ and $(\mathbf{\Sigma}_i - \sigma_i^2 \mathbf{I}_d)$ have the same eigenvectors and the eigenvalues are related by σ_i^2, it is easy to show that: $\sigma_i^2 = tr(\mathbf{\Sigma}_i - \mathbf{U}_i \hat{\mathbf{\Lambda}}_i \mathbf{U}_i^T)/d - k$, $\mathbf{\Lambda}_i = \hat{\mathbf{\Lambda}}_i - \sigma_i^2 \mathbf{I}_d$, where $\hat{\mathbf{\Lambda}}_i$ are the eigenvalues of the covariance matrix $\mathbf{\Sigma}_i$. The same expression could be derived using probabilistic PCA [24, 28].

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 original sample covariance.

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

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. $\mathbf{B}_j = \mathbf{B}_j - \sum_{i=1}^{j-1} tr((\mathbf{B}_i\mathbf{B}_i)^{-1}\mathbf{B}_j^T\mathbf{B}_i)\mathbf{B}_i\ \forall j = 2\cdots 5$. The covariance matrices are approximated as $\boldsymbol{\Sigma}_i = \mathbf{U}_i\boldsymbol{\Lambda}_i\mathbf{U}_i^T + \sigma_i^2\mathbf{I}$, 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, $(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i)\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i) + log|\hat{\boldsymbol{\Sigma}}_i|$, where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ are the low-dimensional estimates of the mean and class covariance. Table 7.1 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 7.1, 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 [22]. 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 7.1.

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.

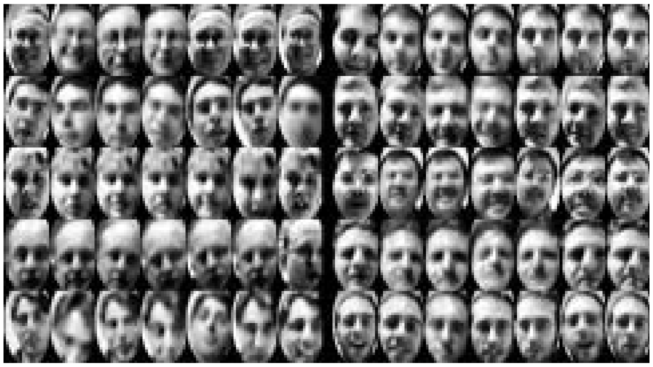

Figure 5: Training data.

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 database of 23 people has been collected using our omnidirectional-meeting-capturing device [7]. The database consist on 23 people recorded over two different days under different illumintation conditions. Figure 7.2 shows some images of 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. 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 7.2 provides some detailed numerical values for different number of basis.

Basis	2	5	10	20	25	30	50
PCA	0.12	0.26	0.43	0.55	0.56	0.58	0.59
LDA	0.21	0.36	0.48	0.56	NA	NA	NA
MODA	0.23	0.38	0.50	0.59	0.60	0.61	0.63

Table 3: Recognition performance of PCA/LDA/MODA

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

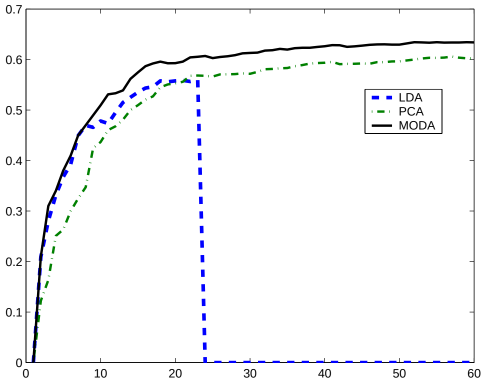

Figure 6: PCA/LDA/MODA

can greatly improve the recognition performance; Refer to [7] for more details.

8 Discussion and future work

In this paper we have introduced Multimodal Oriented Discriminant Analysis (MODA), a new discriminant analysis method that extends classical linear discriminant analysis by modeling class covariances and multimodal manifolds. Several synthetic and real experiments confirm that MODA always outperforms classical LDA. Even when there are few samples, MODA can perform better than PCA by factorizing the covariances.

However, several issues remain unsolved; there is need for faster optimization algorithms that can find global optimal solutions. It would be interesting to train MODA with Adaboost or boosting [11] techniques that use a greedy strategy to look for local optima. On the other hand, in the context of object recognition from video, one of the most important steps is registration and being able to deal with outliers. Further research needs to be done in order to address these problems.

A Appendix A

In general eq. 8 does not have a closed form solution; however, in the case that $\mathbf{\Sigma}_i^{r_1} = \mathbf{\Sigma}\ \forall i, r_1$ an eigensolution does exist. Let be k_i the number of clusters for class i and $K = \sum_{i=1}^{c} k_i$ the total number of clusters. $\mathbf{M} \in \Re^{d \times K}$ is a matrix such that each column contains the mean of each cluster and each class. $\mathbf{G} \in \Re^{K \times c}$ is a dummy indicator matrix, such that $\sum_j g_{ij} = 1$, $g_{ij} \in \{0, 1\}$ and g_{ij} is 1 if $\mathbf{m}_i$ belongs to class j. $\mathbf{P}_M = \mathbf{I}_d - \frac{1}{K}\mathbf{1}_d\mathbf{1}_d^T$ is a projection matrix. Using these definitions and taking into account that $\boldsymbol{\mu}_i^{r_1} = \mathbf{M}\mathbf{e}_{(\sum_{j=1}^{i-1} k_i)+r_1}$, it can be shown that:

$$\begin{aligned} E_1 = \textstyle\sum_{i=1}^{C}\sum_{j=1}^{C}\sum_{r_1 \in C_i}\sum_{r_2 \in C_j} \\ tr\big((\mathbf{B}^T(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})^T\mathbf{B})(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\big) \\ = 2K tr(\mathbf{B}^T\mathbf{M}^T\mathbf{M}\mathbf{P}_M\mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}) \end{aligned} \tag{15}$$

E_1 computes the sum of the distances of the means among all the clusters for all the classes. However, the distances between the clusters of the same class should be subtracted. The sum of distances between the clusters in a class, that is when $i = j$, is given by:

$$\begin{aligned} E_2 = \textstyle\sum_{i=1}^{C}\sum_{r_1 \in C_i}\sum_{r_2 \in C_i}^{C} \\ tr\big((\mathbf{B}^T(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})^T\mathbf{B})(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1} = \\ 2tr\Bigg((\mathbf{B}^T(\mathbf{M}^T\mathbf{M}(\textstyle\sum_{i=1}^{c} k_i diag(\mathbf{g}_i) - \mathbf{G}\mathbf{G}^T)\mathbf{B})(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\Bigg) \end{aligned} \tag{16}$$

Where recall that $\mathbf{g}_i$ is the i column of $\mathbf{G}$. Subtracting E_1 from E_2:

$$\begin{aligned} E_3 = E_1 - E_2 = \textstyle\sum_{i=1}^{C}\sum_{j \neq i}^{C}\sum_{r_1 \in C_i}\sum_{r_2 \in C_j} \\ tr\big((\mathbf{B}^T(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})^T\mathbf{B})(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\big) = 2K \\ tr\Bigg(\big(\mathbf{B}^T(\mathbf{M}^T\mathbf{M}(\mathbf{P}_M - \textstyle\sum_i k_i diag(\mathbf{g}_i) - \mathbf{G}\mathbf{G}^T))\mathbf{B}\big)(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\Bigg) \end{aligned} \tag{17}$$

results in eq. 8 if the covariances are the same. The solution of eq. 17 can be computed as a standard generalized eigenvalue problem.

Acknowledgments

This work has been partially supported by National Business Center of the Department of the Interior under a subcontract from SRI International, U.S. Department of Defense contract N41756-03-C4024, NIMH Grant R01 MH51435 and DARPA HumanID program under ONR contract N00014-00-1-0915.

References

[1] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2:53–58, 1989.

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

[3] N. A. Campbell. Canonical variate analysis - a general formulation. *Australian Journal of Statistics*, 26:86–96, 1984.

[4] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new ldabased face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conference Computer Vision*, pages 484–498, 1998.

[6] F. de la Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision.*, In press, 2002.

[7] F. de la Torre, C. Vallespi, P. E. Rybski, M. Veloso, and T. Kanade. Omnidirectional video capturing, multiple people tracking and recognition for meeting understanding. Technical report, Robotics Institute, Carnegie Mellon University, January 2005.

[8] K. I. Diamantaras. *Principal Component Neural Networks (Theory and Applications)*. John Wiley & Sons, 1996.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons Inc., 2001.

[10] S. Fidler and A. Leonardis. Robust lda classification by subsampling. In *Workshop on Statistical Analysis in Computer Vision*, 2003.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Dept. of Statistics, Stanford University Technical Report, 1998.

[12] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition*. Academic Press.Boston, MA, 1990.

[13] K. Fukunaga and R. Hayes. Effects of sample size in classifier desing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):873–885, 1989.

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

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

[16] J. Hoffbeck and D. Landgrebe. Covariance matrix estimation and classification with limited training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):763–767, 1996.

[17] G. Hughes. On the mean accuracy of statistical pattern recognition,. *IEEE Transactions on Information Theory*, 14:55–63, 1968.

[18] H. A. L. Kiers. Maximization of sums of quotients of quadratic forms and some generalizations. *Psychometrika*, 60(2):221–245, 1995.

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

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

[21] D. G. Lowe and A. Webb. Optimized feature extraction and the bayes decision in feed-forward classifier networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4):355–364, 1991.

[22] A. Martinez and A. Kak. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2003.

[23] I. Matthews and S. Active appearance models revisited. In *tech. report CMU-RI-TR-03-02, Robotics Institute, Carnegie Mellon University, April,*, 2003.

[24] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *Pattern Analysis and Machine Intelligence*, 19(7):137–143, July 1997.

[25] R. Neal and G. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer, 1998.

[26] S. Raudys and A. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3b):252–264, 1991.

[27] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *ICASSP*, 2000.

[28] M. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, 61:611–622, 1999.

[29] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. Jorunal of Computer Vision.*, 9(2):137–154, 1992.

[30] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal Cognitive Neuroscience*, 3(1):71–86, 1991.

[31] S. G. Y. Li and H. Liddell. Recognising trajectories of facial identities using kernel discriminant analysis. *Image and Vision Computing*, 21(13-14):1077–1086, 2003.

[32] H. Yu and J. Yang. A direct lda algorith for high-dimensional data– with applicatins to face recognition. *Pattern Recognition*, 34(10):2067–2070, 2001.

[33] S. Yu and J. Shi. Multiclass spectral clustering. In *ICCV*, 2003.

[34] W. Zhao. Discriminant component analysis for face recognition. In *ICPR*, pages 818–821, 2000.