Maximum Margin Discriminant Analysis based Face Recognition

Kornél Kovács[1], *András Kocsor*[1] and *Csaba Szepesvári* [2]
[1]Research Group on Artificial Intelligence of the Hungarian Academy of Sciences,
University of Szeged, Aradi vértanúk tere 1., 6720 Szeged, Hungary
[2]Computer and Automation Research Institute of the Hungarian Academy of Sciences,
Kende u. 13-17, 1111 Budapest, Hungary

Abstract:
Face recognition is a highly non-trivial classification problem since the input is high-dimensional and there are many classes with just a few examples per class. In this paper we propose using a recent algorithm – Maximum Margin Discriminant Analysis (MMDA) – to solve face recognition problems. MMDA is a feature extraction method that is derived from a set of sound principles: (i) each feature should maximize information transmission about the classification labels, (ii) only the decision boundary should determine the features and (iii) features should reveal independent information about the class labels. Previously, MMDA was shown to yield good performance scores on a number of standard benchmark problems. Here we show that MMDA is capable of finding good features in face recognition and performs very well provided it is preceded by an appropriate preprocessing phase.

1 Introduction

Human face recognition is a special classification problem where the number of classes is high, there are only a few samples per class and the input space is high-dimensional. These properties make face recognition an especially challenging classification problem. Successful approaches to face recognition must exploit the inherent regularity of face images: a good classifier has to suppress within-class (intra-personal) differences while enhancing between-class (or extra-personal) differences. This is the basic idea underlying some subspace methods, the best known examples of which include the Fisherface method and Moghaddam and Pentland's Bayesian Face Recognition (BFR) methods [2, 6]. These methods work by projecting the space of images into a lower-dimensional space where classification is typically done by resorting to 1-nearest neighbour classification with an appropriately defined distance function. The difficulty is that since there are only a very few examples per class, the information of what needs to be suppressed/enhanced cannot be class specific still, the features should maximize the amount of information kept about the class labels.

Here we propose using a recent feature extraction method called Maximum Margin Discriminant Analysis (MMDA) [4] to find an appropriate feature space. MMDA projects input patterns onto the subspace spanned by the normals of a set of pairwise orthogonal margin maximizing hyperplanes. The method can be regarded as a non-parametric extension of Lin-

ear Discriminant Analysis (LDA) which makes no normality assumptions on the data but, instead, uses the principle that the separating hyperplane employed should only depend on the decision boundary. This principle is complemented by a deflation technique which extracts a sequence of orthogonal projection directions. The kernel mapping idea can also be used to derive a corresponding non-linear feature extraction method. Earlier it was shown that MMDA can produce high quality features [[4]]: the performance of classifiers built on the top of these features often exceeds the performance of other state-of-the-art methods. Hence it seems worthwhile to apply MMDA to the problem of extracting features in face recognition.

Unfortunately the direct application of MMDA alone to face recognition can not be expected to give good results. This is because MMDA was originally proposed for binary (two-class) classification problems. In [4] several extensions were proposed for multi-class problems, but all these extensions assume that a sufficiently large number of samples is available for each class and that the number of classes is small. Thus applying MMDA to face recognition is a non-trivial problem.

In this paper we introduce an appropriate preprocessing step and show that with this step MMDA can indeed be applied to extract useful features in face recognition. The main idea is to create appropriate binary classification subproblems which can then be used to derive the set of features. Experiments using the CSU Toolkit [3] and the FERET database show that the proposed method achieves the best performance amongst several competing linear subspace methods, with its gain increasing on harder tasks.

The paper is organized is as follows. In the next section MMDA is introduced. This is followed by a description of the proposed algorithm. In Section 4 the results of the empirical evaluation of the algorithm are presented. It is followed by conclusions and a summary of future work in Section 5.

2 MMDA

MMDA makes use of the principal idea underlying LDA, that of projecting the input data onto the normal of a given hyperplane which best separates the two classes and provides all the information a decision maker needs to classify the input patterns. However, at this point LDA places normality assumptions on the data, whereas in MMDA ones makes no such assumptions but uses margin maximizing hyperplanes (MMHs) instead. The rationale behind this choice is explained by the following points: (i) without additional information MMHs are likely to provide good generalization on future data, (ii) these hyperplanes are insensitive to small perturbations of correctly classified patterns lying further away from the ideal separating hyperplane, and (iii) they are insensitive to small changes in their parameters. In addition to these properties, MMHs are insensitive to the actual *probability distribution* of patterns lying further away from the decision boundary. Hence when a large mass of the data lies far away from the ideal decision boundary we can expect the new method to win against those methods

that minimize some form of average loss/cost since they necessarily take into account the full distribution of the input patterns.

The MMDA algorithm supplements the idea of projecting onto the space spanned by the normal of a margin maximizing hyperplane with a deflation technique which guarantees that all subsequent hyperplanes (and all subsequent normals) are orthogonal to each other. As a consequence, each successive feature extraction step extracts "new" information unrelated to information extracted in the previous steps, in a way analogous to what happens in Principal Component Analysis (PCA).

Deflation can be incorporated as a step to transform the data covariance matrix. However, as shown in [4] one can also incorporate a suitable orthogonality criterion in the equations defining the margin maximizing hyperplane. These two approaches have been shown to be equivalent. For the sake of simplicity, here we shall present here only the approach that uses deflation.

Let X, y be the training data, where $X = (x_1, \ldots, x_n)$ are the input patterns ($x_j \in \mathbb{R}^d$) and $y \in \{-1, +1\}^n$ are the corresponding target labels (i.e. the task is a binary classification task).

When the data is not separable one seeks to maximize the margin and minimize the misclassification error simultaneously [7]. This results in a quadratic programming problem. In order to introduce the corresponding equations formally, let us choose a positive real number C that we will use to weight the misclassification cost. Then the maximum margin separation (MMS) problem is defined as follows: Given (X, y, C) find a $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $\xi = (\xi_1, \ldots, \xi_n)^T \in \mathbb{R}^n$ such that[1)]

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \to \min \text{ s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \quad i = 1, \ldots, n. \tag{1}$$

MMDA now proceeds as follows: Given (X, y, C), find the solution of the MMS problem (X, y, C). Let this solution be (w_1, b_1). The first extracted feature component is $f_1(x) = w_1^T x$. Now transform the data by projecting it onto a space orthogonal to w_1. For simplicity we will assume that w_1 is normalized so $\|w_1\| = 1$. Then the projected data is given by $x_i' = x_i - (w_1^T x_i)w_1$. Let X' denote the matrix $(x_1', \ldots, x_n')$ and let (w_2, b_2) be the solution of the MMS problem (X', y, C). Then the second extracted feature component is $f_2(x) = w_2^T x'$, where $x' = x - (w_1^T x)w_1$. This procedure can be repeated as many times as desired. It then follows that all the extracted weights are orthogonal to each other [4], and hence $f_2(x) = w_2^T(x - (w_1^T x)w_1) = w_2^T x$. Similarly, if $w_3, \ldots, w_r$ ($r \leq d$) are the normals extracted up to step r then the ith feature value $f_i(x)$ can be computed from $f_i(x) = w_i^T x$.

In practice we could use existing support vector machine (SVM) code to find the margin

[1)] Here $\|w\|$ denotes the ℓ^2 norm of w.

maximizing hyperplanes. Typically these solve (1) via its Langrangian dual:

$$-\frac{1}{2}\alpha^T R\alpha + \alpha^T 1 \to \max, \quad \text{such that} \quad y^T\alpha = 0, \quad 0 \le \alpha \le C1, \tag{2}$$

where $C1 = (C, \ldots, C)^\top \in \mathbb{R}^d$, $\alpha \in \mathbb{R}^n$, and the comparison of vectors is made one component at a time. Further, in the above equation R is the R-matrix corresponding to (X, Y): $R = YX^TXY$, with $Y = \mathrm{diag}(y_1, \ldots, y_n)$ [7]. Given α, the solution of (2), the solution of the MMS problem (X, y, C) is recovered from $w = X\alpha$ and $b = 1^T\alpha$. We shall call (2) the dual MMS problem parameterized by (R, y, C).

Let X' be defined as before. Notice that (2) depends on the data vector X only through the matrix R. Thus the Langrangian dual defined for the transformed data X' has the same form as in (2), but R needs to be recalculated. It was shown in [4] that if X' is the data X projected onto a space orthogonal to the orthonormal system $W = (w_1, \ldots, w_r)$, where $W = XA$ for some matrix A, then the R-matrix corresponding to (X', Y) can be calculated from the relation $R' = Y\left(K - (KA)(KA)^T\right)Y$, where $K = X^TX$.

Hence it is possible to use existing SVM code to extract a sequence of orthogonal margin maximizing hyperplanes just by transforming the matrix R.

3 Using MMDA for Feature Extraction in Face Recognition

Since MMDA is defined for binary classification problems, with multi-class problems one needs to group certain classes together. In [4] it was suggested that MMDA should be used following a "one-vs.-all" approach: basically when the number of classes is m, MMDA is run m times with one class against all the others. This is a simple approach and in [4] it was suggested that although it is likely to be suboptimal, it can yield a sufficiently good performance even when compared with the more involved approach based on output-coding.

It should be mentioned that the one-vs.-all approach cannot produce useful features in face recognition tasks as here all the m subproblems are seriously skewed with one class having only a few elements and the other has lots. Such skewed distributions will yield highly correlated features for the independent subproblems since the subproblems are "well aligned" (it is easy to see that forcing independent features to be decorrelated does not help either due to the large overlap of the problems). The same conclusion holds for the features obtained using the output-coding approach since there classes are grouped together without taking into account their relations in the input space. In a typical subtask persons with very different faces could be grouped together while persons with similar faces might be assigned to different classes.

This gives us the idea of using the available data to create the binary classification subproblems for MMDA. The particular approach suggested here is to use information in the images to create the subproblems. One approach for doing just this is the following. Some method (like LDA or PCA) is used to generate a number of unrelated features. For each feature, the

projections of training images on a selected feature are computed. This produces a number of points on the real-line. Next, the persons in the training set are grouped into two groups so as to minimize the total within-group distortion between the previously calculated points on the real-line (this is a special case of k-means clustering and can be implemented efficiently). MMDA is then run on this binary classification problem (on the original, untransformed images) and the corresponding features are then saved. The process is continued with the next feature. The union of features extracted this way defines the extracted feature space. The proposed algorithm is listed below.

Algorithm 1 Feature Extraction by MMDA for Face Recognition

input: $(m, (x_1, y_1) \ldots, (x_N, y_N))$ // no. of subjects, list of face-image, person id pairs
$F := ()$; $X^i := \{x_j | y_j = i\}$, $i = 1, \ldots, m$; // images of person i
$(w_1, \ldots, w_n) := \mathrm{FE}((x_1, y_1) \ldots, (x_N, y_N))$; // extract n features using method FE
for $i \in \{1, \ldots, n\}$ **do**
  $z_j := w_i^T x_j$, $j = 1, \ldots, N$; // project images
  $Z^i := \{z_j | y_j = i\}$, $i = 1, \ldots, m$; // collect projected images of person i
  Find $(v_1, \ldots, v_m) \in \{-1, 1\}^m$ such that

$$\sum_{\substack{v_i=-1, v_j=-1 \\ i \neq j}} \sum_{\substack{z \in Z^i \\ z' \in Z^j}} (z - z')^2 + \sum_{\substack{v_i=1, v_j=1 \\ i \neq j}} \sum_{\substack{z \in Z^i \\ z' \in Z^j}} (z - z')^2$$

  is minimized.
  $F_0 := MMDA(\cup_{v_i=-1} X^i, \cup_{v_j=+1} X^j)$; // extract features using MMDA
  append$(F, F0)$; // append features to the list of features extracted so far
end for
return F

4 Experimental Evaluation

We used the CSU Face Identification Evaluation System to evaluate the performance of our algorithm on the FERET database [3]. Since the idea can be applied to other supervised feature extraction algorithms, we also decided to test it using LDA in place of MMDA, so that we could see the effect of the data-grouping procedure and the effect of MMDA separately. In addition, both PCA and LDA were used as the underlying feature-extraction methods. This gave rise to the algorithms LDA(LDA), LDA(PCA), MMDA(LDA) and MMDA(PCA). The CSU Toolkit allows one to choose a number of distance functions. We tried out a large number of choices, but only the results for the best distance functions are shown here. For the algorithms mentioned so far this was the so-called covariance distance function. The two other algorithms tested were PCA and a combination of LDA and MMDA method. For PCA the best results were obtained using the distance called MahCosine in the CSU Toolkit

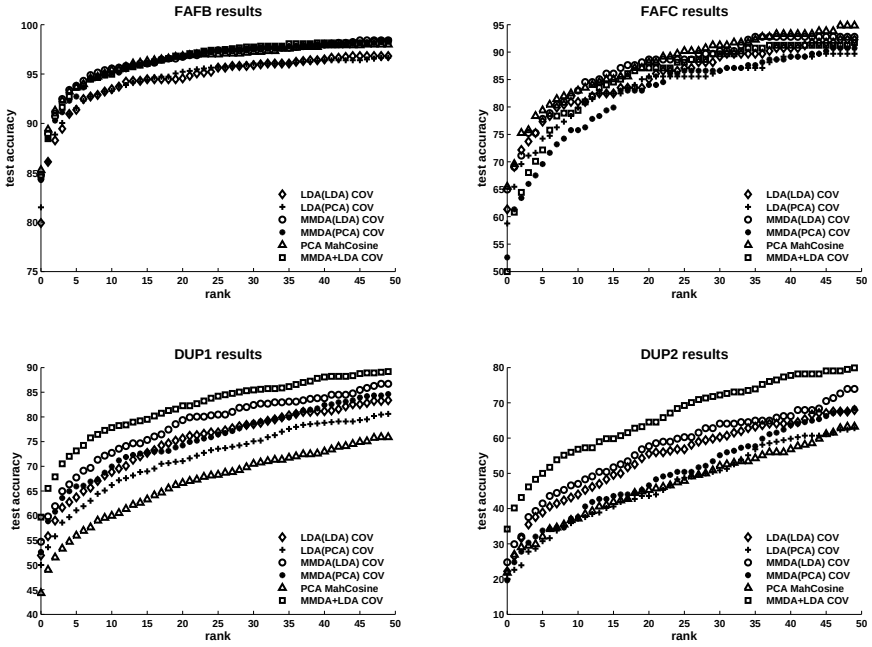

Figure 1: Results obtained for the standard probe sets of FERET.

(MahCosine is the cosine distance measured in the Mahalanobis space; for more details see [3]), while the combined approach, which contains the original LDA direction itself beside the derived MMDA one, employs also the above mentioned covariance distance.

Figure 4 shows the Cumulative Match Curve of recognition rate versus recognition rank that was obtained for the standard probe sets FAFB, FAFC, DUP1 and DUP2. For a description of these probe sets the reader should see Section 4.1 of [1]. In brief, FAFB contains images that were taken at the same time as the training (gallery) images, but the subjects were asked to assume a different facial expression than those in the gallery images. FAFC contains the images of subjects under significantly different lighting conditions (this is a smaller set than FAFB). DUP1 contains images taken between one minute and 1,031 days after the gallery image was taken, while DUP2 is a subset of DUP1 where the probe image was taken at least 18 months after the probe image.

According to the figure, MMDA based methods with the COV distance consistently achieved the best performance on all the tests. On FAFB all the algorithms achieved similar performances. The performance of MMDA(PCA) is significantly worse than that of MMDA(LDA) (especially for the harder probe sets) – in line with our expectation that LDA should yield better individual groupings than PCA. On FAFB and FAFC the best performance next to MMDA(LDA) was achieved by PCA MahCosine (i.e. a special Eigenface method), which is still hard to compete with on these test sets. Note, however, the serious degradation of

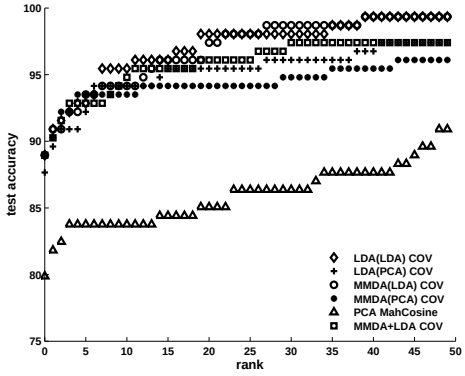

Figure 2: Results obtained for the one-image-per person test.

the performance of this method on the "harder" test sets DUP1 and DUP2. It appears that rejecting noise (the objective of PCA) is the most useful when images are taken close to each other in time (and hence possibly lie in a more concise subspace). Quite surprisingly the performance of LDA(LDA) comes closest to the performance of the MMDA algorithms. Still, MMDA-based algorithms have a considerable advantage over LDA(LDA) on all the datasets.

In face recognition one particularly critical issue is the number of images per person available in the training set. In a typical application only a few images might be available for each person. In addition we should not expect to rerun the feature extraction part when a new person is incorporated into the database. Hence we ran further tests in order to evaluate whether our algorithm can indeed suppress intra-personal changes while enhancing extra-personal differences under such stringent conditions. We took 200 persons and divided them into two disjoint sets: a set of training images (this set contained 70% of the images) and a set used for testing (this one had one image in the gallery per person, and as many images as the person had in the probe set). The results obtained are shown in Fig. 2 above. Here, MMDA(LDA) COV still came out the best, followed by LDA(LDA) COV.

5 Conclusions

In this paper we described the application of MMDA – a recent non-parametric feature extraction method – together with a clever preprocessing method to face recognition tasks. MMDA is a feature extraction method that is most suitable for binary classification problems where many samples are available for each of the classes. Since face recognition lies on the opposite end of the spectrum of classification problems (many classes, few samples per class) MMDA cannot be used directly with face recognition. The preprocessing method proposed is capable of creating a sufficiently large number of independent groupings of subjects such that subjects within the same group have similar images and the corresponding images can be fed into MMDA that returns some features.

In our experiments we found that the performance of MMDA rivals that of the best alternative methods on all the tests that we tried. Of particular interest is the result of a test where each person in the training set had a single image and persons in the test set had no images in the training set used to tune the features. It was found that MMDA performed the best on this test – the significance of this result is that situations like the tested one are likely to be found in practice.

We think that the results obtained so far are encouraging. However, there remain some important open questions. It would be important to examine the robustness of the results – this could be done using existing tools from the CSU toolkit. Moreover, the second experiment involved only 200 persons due to limited time and resources. It would be useful to know whether the advantages of using MMDA(LDA) are retained if the number of samples in the training set were to be increased. Also, there are a number of other ways to create (balanced) binary problems, e.g. by clustering the image space and then combining nearest neighbour clusters. It would be interesting to find out what the performance of MMDA/LDA is with such groupings (see [8] for such an approach applied to LDA). Then as argued in [5] a better input representation should have a significant impact on the performance of the algorithms as well. It would also be interesting to try the method in other domains like text classification that share some of the characteristics of face recognition problems.

References

[1] K. Baek, B.A.Draper, J.R. Beveridge, and K. She. PCA vs. ICA: A comparison on the FERET data set. In *Proc. of the Fourth International Conference on Computer Vision, Pattern Recognition and Image Processing*, pages 824–827, Durham, NC, USA, 2002.

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

[3] M. Teixeira D. Bolme, R. Beveridge and B. Draper. The csu face identification evaluation system: Its purpose, features and structure. In *International Conference on Vision Systems*, pages 304–311. Springer-Verlag, 2003.

[4] A. Kocsor, K. Kovács, and Cs. Szepesvári. Margin maximizing discriminant analysis. In *ECML/PKDD-2004*, pages 227–238, 2004.

[5] X. Liu, A. Srivastava, and D. Wang. Intrinsic generalization analysis of low dimensional representations. *Neural Networks*, 16(5-6):537–545, 2003.

[6] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000.

[7] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, NY, USA, 1995.

[8] M. Zhu and A.M. Martinez. Optimal subclass discovery for discriminant analysis. In *Proc. of IEEE Workshop on Learning in Computer Vision and Pattern Recognition (LCVPR)*, 2004.