A Discriminative Non-Linear Manifold Learning Technique for Face Recognition

Bogdan Raducanu[1] and Fadi Dornaika[2,3]

[1] Computer Vision Center, 08193 Bellaterra, Barcelona, Spain
bogdan@cvc.uab.es
[2] IKERBASQUE, Basque Foundation for Science
[3] University of the Basque Country, San Sebastian, Spain
fadi_dornaika@ehu.es

Abstract. In this paper we propose a novel non-linear discriminative analysis technique for manifold learning. The proposed approach is a discriminant version of Laplacian Eigenmaps which takes into account the class label information in order to guide the procedure of non-linear dimensionality reduction. By following the large margin concept, the graph Laplacian is split in two components: within-class graph and between-class graph to better characterize the discriminant property of the data. Our approach has been tested on several challenging face databases and it has been conveniently compared with other linear and non-linear techniques. The experimental results confirm that our method outperforms, in general, the existing ones. Although we have concentrated in this paper on the face recognition problem, the proposed approach could also be applied to other category of objects characterized by large variance in their appearance.

1 Introduction

In recent years, a new family of non-linear dimensionality reduction techniques for manifold learning has emerged. The most known ones are: Kernel Principal Component Analysis (KPCA) [1], Locally Linear Embedding (LLE) [2, 3], Isomap [4], Supervised Isomap [5], Laplacian Eigenmaps (LE)[6, 7]. This family of non-linear embedding techniques appeared as an alternative to their linear counterparts which suffer of severe limitation when dealing with real-world data: i) they assume the data lie in an Euclidean space, and ii) they may fail when the number of sample are too small. Opposite, the non-linear dimensionality techniques are able to discover the intrinsic data structure by exploiting the local topology, instead of general one. They attempt to optimally preserve the local geometry around each data sample while using the rest of the samples to preserve the global structure of the data.

The main contribution of our work is represented by a Discriminant LE (D-LE) algorithm, which exploits the class label information for mapping the original data in the embedded space. The use of labels allows us to split graph Laplacian associated with the data in two components: within-class graph and

between-class graph. Our proposed non-linear approach benefits from three important properties: (1) it is parameterless; (2) estimates adaptively the neighborhood around a sample, by exploiting the statistical significance of the data; and (3) it is discriminative - by using an objective function that simultaneously maximizes the local margin between heterogenous samples and pushes the homogeneous samples closer to each other.

These properties represent a significant advantage over other spectral-graph based manifold learning techniques because they require the setting of several parameters: (i) the width of the Gaussian Kernel, (ii) the size of neighborhood for non-full mesh graphs, and (iii) the blending parameter for combining two objective functions (e.g., the difference criterion used by a variant of Linear Discriminant Analysis). Therefore, all existing methods either fix these parameters in advance or perform tedious validation in order to select the best value for these parameters.

The combination between locality preserving property (inherited from the classical LE[1]) and the discriminative property (due to the large margin concept) represents a clear advantage for D-LE, compared with other non-linear embedding techniques, because it finds a mapping which maximizes the distance between data samples from different classes at each local area. In other words, it maps the points in an embedded space where data with similar labels fall close to each other and where the data from different classes fall far apart.

The adaptive selection of neighbors for the two graphs represents also an added value to our algorithm. It is well known that a sensitive matter affecting non-linear embedding techniques is represented by the adequate choice for neighborhood size. Setting a too high value for this parameter would result in a loss of local information, meanwhile a too low value could result in an overfragmentation of the manifold (problem known as 'short-circuiting). For this reason, setting an adequate value for this parameter is crucial in order to confer the approach topological stability.

The rest of the paper is organized as follows. Section 2 reviews some related work on manifold learning techniques. In section 3 we review, for the sake of completeness, the classical Laplacian Eigenmaps algorithm. Section 4 is devoted to the presentation of our new proposed algorithm. Section 5 presents some experimental results obtained on four face databases. Finally, section 6 contains our conclusions and guidelines for future work.

2 Related work

During the last few years, a large number of approaches have been proposed for constructing and computing an embedded subspace by finding an explicit or non-explicit mapping that projects the original data to a new space of lower dimensionality. These methods can be classified by their linearity. The non-linear methods such as Locally Linear Embedding (LLE), Laplacian Eigenmaps,

[1] By classical Laplacian Eigenmaps we refer to the algorithm introduced in [6].

Isomap, Hessian LLE (hLLE) [8] focus on preserving the geodesic distances which reflect the real geometry of the low-dimensional manifold. LLE formulates the manifold learning problem as a neighborhood-preserving embedding, which learns the global structure by exploiting the local symmetries of linear reconstructions. Isomap extends the classical Multidimensional Scaling (MDS) [9] by computing the pairwise distances in the geodesic space of the manifold. Essentially, Isomap attempts to preserve geodesic distances when data are embedded in the new low dimensional space. Based on the spectral decomposition of graph Laplacian, Laplacian Eigenmaps actually try to find Laplacian eigenfunction on the manifold.

The non-linear embedding methods have been successfully applied to some standard data sets and generated satisfying results in dimensionality reduction and manifold visualization. However, these approaches does not take into account the discriminant information that is usually available for many real world problems. Therefore, the application of these methods can be very satisfactory in terms of dimensionality reduction and visualization but can be fair for the tasks of classification. In [5], the authors propose a supervised version of Isomap. This version replaces pairwise Euclidean distances by a dissimilarity function that increases if the pair is heterogeneous and decreases otherwise.

The classical linear embedding methods (e.g., PCA, LDA, MDS, Maximum Margin Criterion (MMC)[10]) are demonstrated to be computationally efficient and suitable for practical applications, such as pattern classification and visual recognition. Recent proposed methods attempted to linearize some non-linear embedding techniques. This linearization is obtained by forcing the mapping to be explicit, i.e., performing the mapping by a projection matrix. For example, Locality Preserving Projection (LPP) [11–13] and Neighborhood Preserving Embedding (NPE) [14] can be seen as a linearized version of LE and LLE, respectively. The main advantage of the linearized embedding techniques is that the mapping is defined everywhere in the original space. However, since the embedding is approximated to a linear process, these methods ignore the geodesic structure of the true manifold. All these linear methods cannot reveal the perfect geometric structure of the non-linear manifold.

In [15], the authors exploit label information to improve Laplacian Eigenmaps. The proposed improvement affects the computation of the affinity matrix entries in the sense that an homogeneous pair of neighbors will have large values and heterogeneous pairs of neighbors will have a small value. Although, the authors show some performance improvement, the proposed method has two drawbacks. First, there is no guarantee that the heterogenous samples will be pushed away from each other. Second, the method has at least three parameters to be tuned. [16] proposed a linear discriminant method called Average Neighbors Margin Maximization (ANMM). This technique associates to every sample a margin that is set to the difference between the average distance to heterogenous neighbors and the average distance to the homogeneous neighbors. The linear transform is then derived by maximizing the sum of the margins in the embedded space.

3 Review of Laplacian Eigenmaps

Laplacian Eigenmaps is a recent non-linear dimensionality reduction techniques that aims to preserve the local structure of data [6]. Using the notion of the graph Laplacian, this non-supervised algorithm computes a low-dimensional representation of the data set by optimally preserving local neighborhood information in a certain sense. We assume that we have a set of N samples $\{\mathbf{y}_i\}_{i=1}^N \subset \mathbb{R}^D$. Define a neighborhood graph on these data, such as a K-nearest-neighbor or ϵ-ball graph, or a full mesh, and weigh each edge $\mathbf{y}_i \sim \mathbf{y}_j$ by a symmetric affinity function $W_{ij} = K(\mathbf{y}_i; \mathbf{y}_j)$, typically Gaussian:

$$W_{ij} = \exp(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{\beta}) \tag{1}$$

where β is usually set to the average of squared distances between all pairs.

We seek latent points $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^L$ that minimizes $\frac{1}{2}\sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|^2 W_{ij}$, which discourages placing far apart latent points that correspond to similar observed points. If $\mathbf{W} \equiv W_{ij}$ denotes the symmetric affinity matrix and $\mathbf{D}$ is the diagonal weight matrix, whose entries are column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$, then the Laplacian matrix is given $\mathbf{L} = \mathbf{D} - \mathbf{W}$. It can be shown that the objective function can also be written as (A similar derivation is given in section 4.2):

$$\frac{1}{2}\sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|^2 W_{ij} = tr(\mathbf{Z}^T \mathbf{L}\, \mathbf{Z}) \tag{2}$$

where the $N \times L$ matrix $\mathbf{Z}$ is given by $\mathbf{Z} = [\mathbf{x}_1^T; \ldots; \mathbf{x}_N^T]$. The i^{th} row of $\mathbf{Z}$ provides the vector $\mathbf{x}_i$—the embedding coordinates of the sample $\mathbf{y}_i$.

The matrix $\mathbf{Z}$ is the solution of the optimization problem:

$$\min_{\mathbf{Z}} tr(\mathbf{Z}^T \mathbf{L}\, \mathbf{Z}) \;\; s.t. \;\; \mathbf{Z}^T \mathbf{D}\, \mathbf{Z} = \mathbf{I}, \;\; \mathbf{Z}^T \mathbf{L}\, \mathbf{e} = \mathbf{0} \tag{3}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{e} = (1, \ldots, 1)^T$. The first constraint eliminates the trivial solution $\mathbf{Z} = \mathbf{0}$ (by setting an arbitrary scale) and the second constraint eliminates the trivial solution $\mathbf{e}$ (all samples are mapped to the same point). Standard methods show that the embedding matrix is provided by the matrix of eigenvectors corresponding to the smallest eigenvalues of the generalized eigenvector problem,

$$\mathbf{L}\, \mathbf{z} = \lambda\, \mathbf{D}\, \mathbf{z} \tag{4}$$

Let the column vectors $\mathbf{z}_0, \ldots, \mathbf{z}_{N-1}$ be the solutions of (4), ordered according to their eigenvalues, $\lambda_0 = 0, \ldots, \lambda_{N-1}$. The eigenvector corresponding to eigenvalue 0 is left out and only the next eigenvectors for embedding are used. The embedding of the original samples is given by the row vectors of the matrix $\mathbf{Z}$, that is,

$$\mathbf{y}_i \longrightarrow \mathbf{x}_i = (z_1(i), \ldots, z_L(i))^T \tag{5}$$

where $L < N$ is the dimension of the new space. From equation (4), we can observe that the dimensionality of the subspace obtained by Laplacian Eigenmaps is limited by the number of samples N.

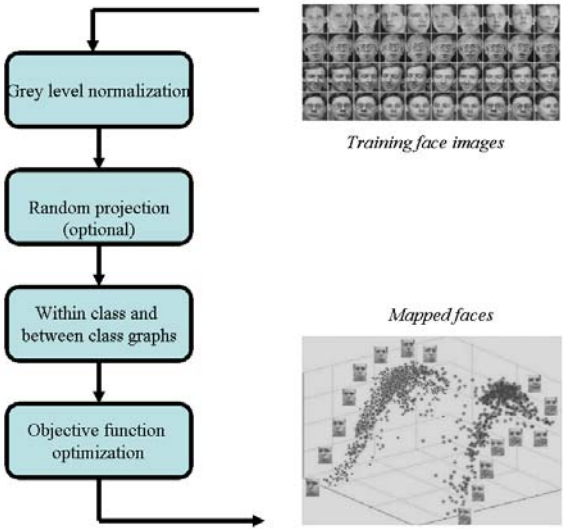


Fig. 1. Discriminant Laplacian Eigenmaps embedding for the face recognition problem.

4 Discriminant Laplacian Eigenmaps

While the LE may give good results for non-linear dimensionality reduction, it has not been widely used and assessed for the tasks of classification. Indeed, many experiments show that the recognition rate in the embedded space is highly depending on the choice of the neighborhood size in the reconstructed graph. Choosing the ideal size in advance can be a very difficult task. Moreover, the introduced mapping by LE does not exploit the discriminant information given by the labels of data. In this section, we present our Discriminant LE algorithm which has three important characteristics: (1) it is parameterless; (2) the neighborhood size is adaptive in the sense that this size is depending on the local density and similarity between data samples; and (3) the obtained embedding respects both discriminant and geometrical structure in data. In order to encode the similarity between two samples $\mathbf{y}_i$ and $\mathbf{y}_j$, we use Pearson's coefficient (normalized cross-correlation). Let p_{ij} denotes Pearson's coefficient associated with the pair $\mathbf{y}_i$ and $\mathbf{y}_j$. Furthermore, we map Pearson's coefficients to the interval $[0, 1]$ by using the following:

$$\bar{p}_{ij} = \frac{p_{ij} - min}{1 - min}$$

where min is the minimum of $p_{ij}s$ over the whole data set.

4.1 Two graphs and adaptive neighborhood size

In order to discover both geometrical and discriminant structure of the data manifold, we split the global graph in two components: the within-class graph

G_w and between-class graph G_b. Let $l(\mathbf{y}_i)$ be the class label of $\mathbf{y}_i$. For each data point $\mathbf{y}_i$, we compute two subsets, $N_b(\mathbf{y}_i)$ and $N_w(\mathbf{y}_i)$. $N_w(\mathbf{y}_i)$ contains the neighbors sharing the same label with $\mathbf{y}_i$, while $N_b(\mathbf{y}_i)$ contains the neighbors having different labels. We stress the fact that unlike the classical LE, our algorithm adapts the size of both sets according to the local sample point $\mathbf{y}_i$ and its similarities with the rest of samples. To this end, each set is defined for each sample point $\mathbf{y}_i$ and is computed in two consecutive steps. First, the average similarity of the sample $\mathbf{y}_i$ is computed by the total of all similarities with the rest of the data set (Eq. (6)). Second, the sets $N_w(\mathbf{y}_i)$ and $N_b(\mathbf{y}_i)$ are computed using Eqs. (7) and (8), respectively.

$$AS(\mathbf{y}_i) = \frac{1}{N}\sum_{k=1}^{N}\bar{p}_{ik} \tag{6}$$

$$N_w(\mathbf{y}_i) = \{\mathbf{y}_j \mid l(\mathbf{y}_j) = l(\mathbf{y}_i), \bar{p}_{ij} > AS(\mathbf{y}_i)\} \tag{7}$$

$$N_b(\mathbf{y}_i) = \{\mathbf{y}_j \mid l(\mathbf{y}_j) \neq l(\mathbf{y}_i), \bar{p}_{ij} > AS(\mathbf{y}_i)\} \tag{8}$$

Equation (7) means that the set of within-class neighbors of the sample $\mathbf{y}_i$, $N_w(\mathbf{y}_i)$, is all data samples that have the same label of $\mathbf{y}_i$ and that have a similarity higher then the average similarity associated with $\mathbf{y}_i$. There is a similar interpretation for the set of between-class neighbors $N_b(\mathbf{y}_i)$. From Equations (7) and (8) it is clear that the neighborhood size is not the same for every data sample. This mechanism adapts the set of neighbors according to the local density and similarity between data samples in the original space.

Each of the graphs mentioned before, G_w and G_b, is characterized by its corresponding affinity (weight) matrix $\mathbf{W}_w$ and $\mathbf{W}_b$, respectively. The matrices are defined by the following formulas:

$$W_{w,ij} = \begin{cases} \bar{p}_{ij}\ if\ \mathbf{y}_j \in N_w(\mathbf{y}_i)\ or\ \mathbf{y}_i \in N_w(\mathbf{y}_j) \\ 0, otherwise \end{cases}$$

$$W_{b,ij} = \begin{cases} \bar{p}_{ij}\ \mathrm{if}\ \mathbf{y}_j \in N_b(\mathbf{y}_i)\ or\ \mathbf{y}_i \in N_b(\mathbf{y}_j) \\ 0, otherwise \end{cases}$$

It is easy to show that the affinity matrix, $\mathbf{W}$, associated with the Laplacian Eigenmaps graph can be written as:

$$\mathbf{W} = \mathbf{W}_w + \mathbf{W}_b$$

4.2 Optimal mapping

One dimensional case Now consider the problem of mapping the within-class graph and between-class graph to a line so that connected points of G_w stay as

close together as possible while connected points of G_b stay as distant as possible. Let $\mathbf{z} = (x_1, x_2, \ldots, x_N)^T$ be such a map. Note that here every data sample is mapped to a real value. For the one dimension case, it is easy to see that the matrix of embedded data $\mathbf{Z}$ (introduced in Eq. (2)) reduces to the vector $\mathbf{z}$. A reasonable criterion for choosing a good map is to optimize the following two functions under some constraints:

$$\min \frac{1}{2} \sum_{i,j} (x_i - x_j)^2 \, W_{w,ij} \tag{9}$$

$$\max \frac{1}{2} \sum_{i,j} (x_i - x_j)^2 \, W_{b,ij} \tag{10}$$

Minimizing function (9) on within-class graph imposes a heavy penalty if neighboring samples $\mathbf{y}_i$ and $\mathbf{y}_j$ are mapped far apart while they are actually in the same class. Maximizing function (10) imposes a heavy penalty if neighboring samples $\mathbf{y}_i$ and $\mathbf{y}_j$ are mapped close together while they actually belong to different classes. The physical interpretation of optimizing (9) and (10) for face recognition is as follows. For each face image, it pulls the neighboring images of the same person towards it as near as possible, while simultaneously pushing the neighboring images of different people away from it as far as possible.

By simple algebra formulation, function (9) can be written as

$$\begin{aligned} & \frac{1}{2} \sum_{i,j} (x_i^2 \, W_{w,ij} + x_j^2 \, W_{w,ij} - 2 x_i \, W_{w,ij} \, x_j) \\ &= \sum_i x_i^2 \, D_{w,ii} - \sum_{i,j} x_i \, W_{w,ij} \, x_j \\ &= \mathbf{z}^T \, \mathbf{D}_w \, \mathbf{z} - \mathbf{z}^T \, \mathbf{W}_w \, \mathbf{z} \\ &= \mathbf{z}^T \, \mathbf{L}_w \, \mathbf{z} \end{aligned} \tag{11}$$

One can notice that the above function is similar to the function (2). However, this function only contains the Laplacian matrix, $\mathbf{L}_w$ associated with the within-class graph G_w.

Similarly, the function (10) can be reduced to:

$$\frac{1}{2} \sum_{i,j} (x_i - x_j)^2 \, W_{b,ij} = \mathbf{z}^T \, \mathbf{L}_b \, \mathbf{z}$$

We aim to find the optimal map $\mathbf{z}$ by simultaneously optimizing criteria (9) and (10).

These two objective functions can be combined into one single objective function:

$$\max \frac{\mathbf{z}^T \, \mathbf{L}_b \, \mathbf{z}}{\mathbf{z}^T \, \mathbf{L}_w \, \mathbf{z}} \tag{12}$$

The above problem has a closed form solution given by solving the following generalized eigenvalue problem:

$$\mathbf{L}_b \, \mathbf{z} = \lambda \, \mathbf{L}_w \, \mathbf{z} \tag{13}$$

Therefore, the optimal map $\mathbf{z}$ is simply the generalized eigenvector of (13) corresponding to the largest eigenvalue.

Multi dimensional case In this case, each data sample $\mathbf{y}_i$ is mapped into a vector $\mathbf{x}_i$. The aim is to compute the embedded coordinates $\mathbf{x}_i$ for each data sample. The derivation of the optimal mapping follows the same steps described above. The objective functions are:

$$\min \frac{1}{2} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|^2 W_{w,ij} \tag{14}$$

$$\max \frac{1}{2} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|^2 W_{b,ij} \tag{15}$$

Since $\frac{1}{2} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|^2 W_{w,ij} = tr(\mathbf{Z}^T \mathbf{L}_w \mathbf{Z})$, and $\frac{1}{2} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|^2 W_{b,ij} = tr(\mathbf{Z}^T \mathbf{L}_b \mathbf{Z})$, the objective function becomes:

$$\max \frac{tr(\mathbf{Z}^T \mathbf{L}_b \mathbf{Z})}{tr(\mathbf{Z}^T \mathbf{L}_w \mathbf{Z})} \tag{16}$$

where $\mathbf{Z}$ contains the unknown latent vectors $\mathbf{x}_i$ in its rows.

It is worthwhile to point out that the trace ratio optimization problem (16) can be replaced by the simpler yet inexact ratio trace form, i.e.:

$$\max tr[(\mathbf{Z}^T \mathbf{L}_w \mathbf{Z})^{-1} (\mathbf{Z}^T \mathbf{L}_b \mathbf{Z})] \tag{17}$$

which can be optimally solved by the generalized eigenvalue problem (13). Let the column vectors $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_L$ be the generalized eigenvectors according to their eigenvalue: $\lambda_1, \lambda_2, \cdots, \lambda_L$. Then, $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_L]$.

Fig. 2. Some samples in Extended Yale data set.

Fig. 3. Some samples in PF01 data set.

Fig. 4. Some samples in PIE data set.

5 Experimental results

In this section, we report the experimental results obtained from the application of our proposed algorithm to the problem of face recognition.

Face recognition is one of the most studied problems and a large literature has been devoted to this issue [17]. Face recognition represents an intuitive and non-intrusive method of recognizing people. Facial image data are often complex to understand and difficult to process due to their high variability in appearance. For this reason, it is mandatory to discover a meaningful low dimensional structure hidden in high dimensional observation data space [18]. Therefore, appearance-based face recognition was usually preceded by a given transform-based dimensionality reduction technique [19].

5.1 Face data sets

In this study, four face data sets are considered:

1. The UMIST face data set[2]. The UMIST data set contains 575 gray images of 20 different people. The images depict variations in head pose.
2. The Extended Yale Face Database B[3]. It contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. In our study, a subset of 1800 images has been used. Figure 2 shows some face samples in the Extended Yale Face Database B.
3. The PF01 face data set[4]. It contains the true-color face images of 103 people, 53 men and 50 women, representing 17 various images (1 normal face, 4 illumination variations, 8 pose variations, 4 expression variations) per person. All of the people in the database are Asians. There are three kinds of systematic variations, such as illumination, pose, and expression variations in the database. Some samples are shown in Figure 7.
4. The PIE face data set[5] contains 41,368 images of 68 people. Each person is imaged under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our study, we used a subset of the original dataset, considering 29 images per person. Some samples are shown in Figure 4.

[2] $http://www.shef.ac.uk/eee/research/vie/research/face.html$
[3] $http://vision.ucsd.edu/\sim leekc/ExtYaleDatabase/ExtYaleB.html$
[4] $http://nova.postech.ac.kr/special/imdb/imdb.html$
[5] $http://www.ri.cmu.edu/projects/project\_418.html$

5.2 Data preparation

Figure 1 illustrates the main steps of the application of D-LE to the problem of face recognition. The initial face data set is projected on the embedded face subspace using the D-LE algorithm, whose steps have been summarized by a 4 block-diagram (according to section 4). A face image is recognized using the nearest neighbor (NN) classifier applied in this low dimensional space.

To make the computation of the embedding more efficient, the dimensionality of the original data is reduced by applying random projections [20]. The main goal of random projections is to reduce the dimensionality of the original face data samples. It has a similar role to that of PCA yet with the obvious advantage that random projections do not need any training data.

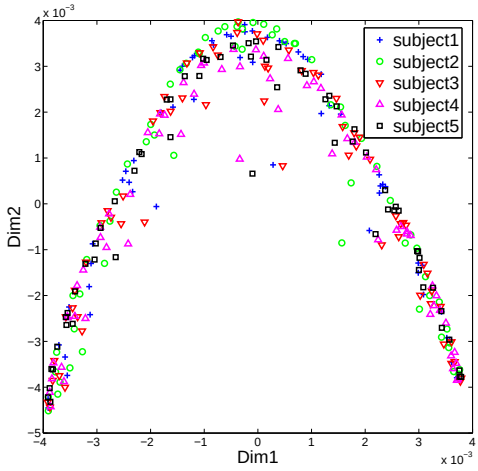
5.3 Visualization of the Embedding Process

Before presenting the quantitative evaluation of classification, it would be worthy to visualize the obtained embedded face data. To this end, we visualize some embedded samples using two methods: the classical LE and the proposed Discriminant LE. Figure 5.(a) visualizes the embedding of faces associated with five persons of the Extended Yale data set obtained with the classical LE. In this plot only the first two dimensions were used. Figure 5.(b) visualizes the embedding of the same five persons obtained with the proposed Discriminant LE. As can be seen, the intra and extra person variabilities are best presented in the embedded space obtained with the proposed D-LE.

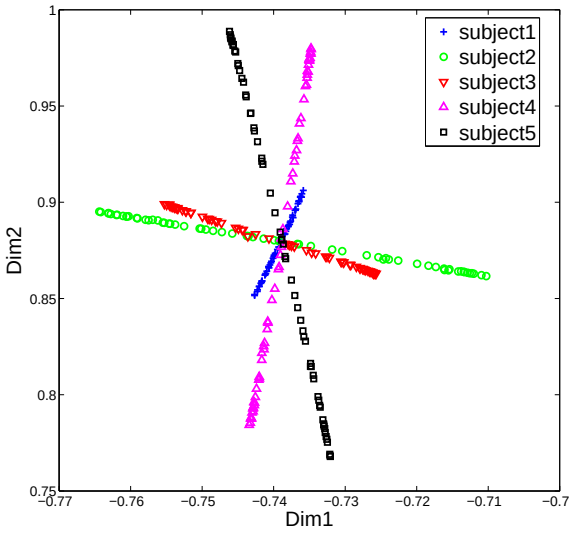
5.4 Evaluation methodology

We have compared our method with six different methods, namely PCA, LDA, ANMM, KPCA, Isomap and classical LE. For methods relying on neighborhood graphs (Isomap and LE), several trials have been performed in order to choose the optimal neighborhood size. The final values correspond to those giving the best recognition rate.

For each face data set and for every method, we conducted two groups of experiments for which the percentage of training samples was set to 30% and 50% of the whole data set, respectively. The remaining data was used for testing. The partition of the data set was done randomly. The best (average) performance obtained by these algorithms, based on a 10-fold cross-validation strategy, are shown in Table 1. The number appearing in parenthesis corresponds to the optimal dimensionality of the embedded subspace (at which the maximum recognition rate has been reported). We can observe that: i) the D-LE outperforms all other methods on three face data sets, ii) for UMIST face data set, the D-LE was outperformed by ANMM, KPCA, and PCA methods. This can be explained by the fact that the intra-class variability of UMIST set is due to face pose only. Therefore, the affinity matrix $\mathbf{W}_w$ in the denominator of quotient (16), $\mathbf{Z}^T(\mathbf{D}_w - \mathbf{W}_w)\,\mathbf{Z}$, was not stable enough so the resulting embedding obtained by maximizing this quotient was not as good as that obtained by KPCA



(a) Classical LE



(b) Proposed Discriminant LE

Fig. 5. Embedded faces of five persons of Extended Yale face data set.

and ANMM methods, iii) on the other hand, in the case of the PIE data set, the improvement brought by D-LE becomes very significant. This is due to the fact that the intra-class variation in the case of PIE data set (due to light variation and changes in facial expression) is high.

For a given embedding method, the recognition rate was computed for several dimensions belonging to $[1, L_{max}]$. For most of the tested methods L_{max} is equal to the number of samples used except for LDA and ANMM. For LDA, the maximum dimension is equal to the number of classes minus one. For ANMM the maximum dimension is variable since it is equal to the number of positive eigenvalues.[6]

Figures 6 and 7 illustrate the average recognition rate associated with Extended Yale, and PF01 data sets, respectively. The average recognition rate was computed (over ten folds) by PCA, KPCA, Isomap, ANMM, LE, and D-LE. The training/test percentage was set to 30%-70% for Extended Yale data set, and to 50%-50% for the PF01 data set. Since the maximum dimension for LDA is equal to the number of classes minus one, the corresponding curve was not plotted. Its rate was reported in Table 1. In [16], it is shown that the ANMM technique performs equally to or better than the following linear methods: Maximum Margin Criterion (MMC), Marginal Fisher Analysis (MFA), and Step non-parametric maximum margin criterion (SNMMC). Thus, the comparisons shown in Figures 6 and 7 contain implicitly those methods.

The maximum dimension depicted in the plots was set to a fraction of L_{max}, in order to guarantee meaningful results. Moreover, we can observe that after a given dimension the recognition rate associated with the three methods PCA, KPCA, and Isomap becomes stable. However, the recognition rate associated with LE and D-LE methods decreases if the number of used eigenvectors becomes large—a general trend associated with many non-linear methods. This means that the last eigenvectors do not have any discriminant information, lacking completely of statistical significance.

In conclusion, the advantage of classification based on non-linear dimensionality techniques is that only a relative small number of dimensions are required, compared with their linear counterparts (as it can be appreciated from Table 1). This is a very important result especially for the case when data lie in a very high dimensionality space (like hyperspectral images, for instance) because it allows a powerful compression of the data without any relevant loss of intrinsic information. Furthermore, they achieve very good results even with a small number of training samples.

6 Conclusions and Future Work

We proposed a parameterless non-linear dimensionality reduction technique, namely Discriminant Laplacian Eigenmaps (D-LE). Our algorithm benefits from the following important properties: i) it estimates adaptively the neighbors around a sample, by exploiting the statistical significance of the data; and ii) it is discriminative - by using an objective function that simultaneously maximizes the local margin between heterogenous samples and pushes the homogeneous samples closer to each other. For validation purposes, we applied our method to the

[6] This dimension is bounded by the the dimension of the input samples.

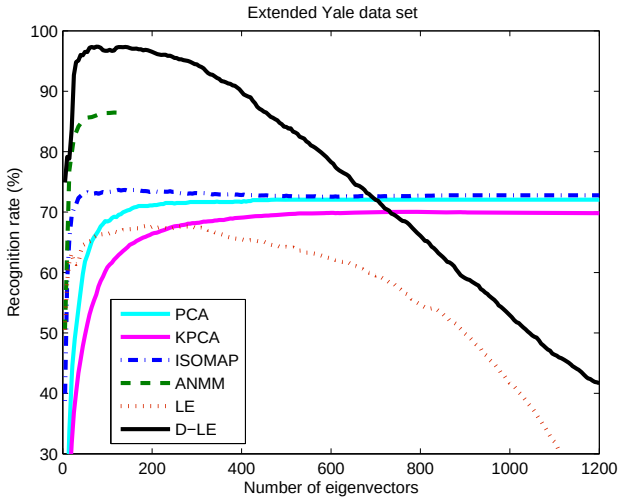


Fig. 6. Average recognition rate as function of the number of eigenvectors obtained with Extended YALE data set. The training/test percentage was set to 30%-70%

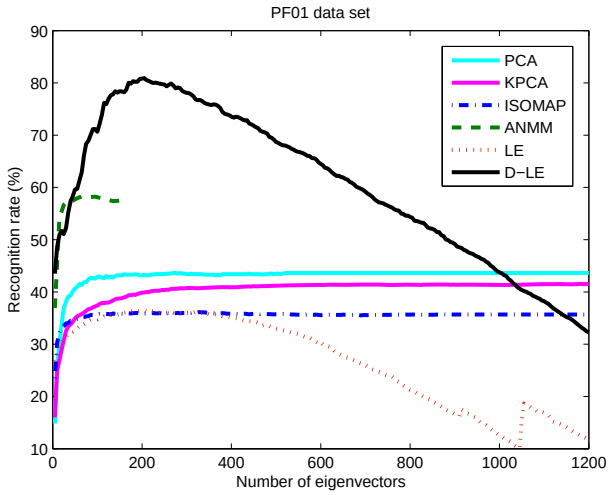


Fig. 7. Average recognition rate as function of the number of eigenvectors obtained with PF01 data set. The training/test percentage was set to 50%-50%.

30%-70%	UMIST	Extended Yale	PF01	PIE
PCA	88.08% (45)	72.06% (465)	33.00% (385)	30.76% (370)
LDA	85.35% (10)	88.00% (10)	62.16% (55)	63.38% (65)
ANMM	**92.10%** (88)	87.10% (139)	45.10% (169)	48.58% (162)
KPCA	89.82% (85)	70.04% (725)	34.91% (940)	39.25% (1030)
Isomap	84.11% (25)	73.69% (125)	31.39% (115)	36.47% (200)
LE	77.88% (40)	67.69% (185)	32.71% (170)	34.97% (330)
D-LE	89.15% (15)	**97.38%** (75)	**69.21%** (205)	**89.28%** (70)
50%-50%				
PCA	94.44% (65)	81.73% (395)	43.62% (270)	39.25% (330)
LDA	90.27% (15)	95.94% (25)	80.40% (30)	60.33% (65)
ANMM	98.2% (73)	94.20% (135)	55.00% (164)	66.00% (164)
KPCA	**95.79%** (85)	79.40% (820)	41.53% (1180)	50.34% (1190)
Isomap	91.63% (45)	79.23% (165)	36.13% (330)	45.02% (210)
LE	86.52% (40)	74.00% (445)	36.44% (200)	42.09% (385)
D-LE	93.54% (15)	**98.90%** (125)	**80.92%** (205)	**93.56%** (85)

Table 1. Best recognition accuracy obtained with four face data sets. The training/test percentage was set to 30%-70%, and 50%-50% for the top part and the bottom part, respectively.

face recognition problem. The experimental results obtained on four face data sets show that our approach outperforms many recent non-linear dimensionality reduction techniques. The proposed method is based on maximizing a certain local margin and is therefore intuitively related to the NN classifier that was used in the current study. Future work will be concentrated on two directions. First, we will investigate the generalization of the proposed method to other classifiers, such as Support Vector Machines (SVM) or Sparse Representation classifiers (SRC) [21]. Second, we will try to find for a given classification task the best set of obtained eigenvectors using the feature selection paradigm.

Acknowledgements

B. Raducanu is supported by the projects TIN2009-14404-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018), Ministerio de Ciencia e Innovacion, Spain.

References

1. B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998) 1299–1319.
2. S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
3. L. K. Saul, S. T. Roweis, Y. Singer, Think globally, fit locally: Unsupervised learning of low dimensional manifolds, Journal of Machine Learning Research 4 (2003) 119–155.

4. J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
5. X. Geng, D. Zhan, Z. Zhou, Supervised nonlinear dimensionality reduction for visualization and classification, IEEE Transactions on systems, man, and cybernetics-part B: cybernetics 35 (2005) 1098–1107.
6. M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.
7. P. Jia, J. Yin, X. Huang, D. Hu., Incremental Laplacian Eigenmaps by preserving adjacent information between data points, Pattern Recognition Letters 30 (16) (2009) 1457–1463.
8. D. Donoho, C. Grimes, Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, in: Proc. of the National Academy of Arts and Sciences, 2003.
9. I. Borg, P. Groenen, Modern Multidimensional Scaling: theory and applications, Springer-Verlag New York, 2005.
10. H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, in: Advances in Neural Information Processing Systems 16, 2003, pp. 157–165.
11. X. He, P. Niyogi, Locality preserving projections, in: Conference on Advances in Neural Information Processing Systems, 2003.
12. X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intelligence 27 (3) (2005) 328–340.
13. L. Zhang, L. Qiao, S. Chen, Graph-optimized locality preserving projections, Pattern Recognition 43 (2010) 1993–2002.
14. X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: IEEE International Conference on Computer Vision, 2005.
15. Q. Jiang, M. Jia, Supervised laplacian eigenmaps for machinery fault classification, in: World Congress on Computer Science and Information Engineering, 2009.
16. F. Wang, X. Wang, D. Zhang, C. Zhang, T. Li, Marginface: A novel face recognition method by average neighborhood margin maximization, Pattern Recognition 42 (2009) 2863–2875.
17. W. Zhao, R. Chellappa, A. Rosenfeld, P. Phillips, Face recognition: A literature survey, ACM Computing Surveys (2003) 399–458.
18. B. Draper, K. Baek, M. Bartlett, J. R. Beveridge, Recognizing faces with PCA and ICA, Computer Vision and Image Understanding 91 (2003) 115–137.
19. V. Dattatray, S. Raghunath, Advances in Face Image Analysis: Techniques and Technologies, IGI Golbal, 2011, Ch. Transform based feature extraction and dimensinality reduction techniques, pp. 120–136.
20. N. Goel, G. Bebis, A. Nefian, Face recognition experiments with random projections, in: SPIE Conference on Biometric Technology for Human Identification, 2005.
21. J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.