LSDA Solution Schemes for Modelless 3D Head Pose Estimation

F. Dornaika[1,2], A. Bosaghzadeh[1]
[1] Dept. of Computer Science and Artificial Intelligence
Univ. of the Basque Country UPV/EHU, San Sebastian, Spain
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

B. Raducanu
Computer Vision Center
Bellaterra, Barcelona, Spain
bogdan@cvc.uab.es

Abstract

Locality Sensitive Discriminant Analysis (LSDA) is a recent linear manifold learning method used in pattern recognition and computer vision. Whenever LSDA is used for face image analysis, it suffers from a number of problems including the Small Sample Size (SSS) problem, and that its classification performance seems to be heavily influenced by its parameters. In this paper, we propose two novel solution schemes for LSDA. The first solution is a novel parameterless approach. The second solution is an exponential LSDA which can solve the SSS problem, in the sense that there is no need to perform the pre-stage of dimensionality reduction. The proposed solution schemes have been applied to the problem of modelless coarse 3D head pose estimation. They were tested on two databases FacePix and Pointing'04. They were conveniently compared with other state-of-the-art linear techniques. The experimental results confirm that our methods can give better results than the existing ones.

1. Introduction

In most computer vision and pattern recognition problems, the large number of sensory inputs, such as images and videos, are computationally challenging to analyze. In such cases it is desirable to reduce the dimensionality of the data while preserving the original information in its distribution, allowing for more efficient learning and inference. During the last few years, a large number of approaches have been proposed in order to compute the embedding of high dimensional spaces. We categorize these methods by their linearity. The linear methods, such as Principal Component Analysis (PCA) [14] and Multi-dimensional Scaling (MDS) [3], are evidently effective in observing the Euclidean structure. PCA projects the samples along the directions of maximal variances and aims to preserve the Euclidean distances between the samples. Unlike PCA which is unsupervised, Linear Discriminant Analysis (LDA) [7] is a supervised technique. One limitation of PCA and LDA is that they only see the linear global Euclidean structure. However, recent researches show that the samples may reside on a nonlinear submanifold, which makes PCA and LDA inefficient. The nonlinear methods such as Locally Linear Embedding (LLE) [11], Laplacian Eigenmaps [2], and Isomap [13], focus on preserving the local structures.

Linear Dimensionality Reduction (LDR) techniques have been increasingly important in pattern recognition [9, 16] since they permit a relatively simple mapping of data onto a lower-dimensional subspace, leading to simple and computationally efficient classification strategies. The goal of all dimensionality reduction techniques is to find a new set of features that represents the target concept in a more compact and robust way but also to provide more discriminative information. Many dimensionality reduction techniques can be derived from a graph whose nodes represent the data samples and whose edges quantify the similarity among pairs of samples [12]. Locality Preserving Projections (LPP) is a typical graph-based LDR method that has been successfully applied in many practical problems. LPP is essentially a linearized version of Laplacian Eigenmaps [2]. In [15], the authors proposed a linear discriminant method called Average Neighbors Margin Maximization (ANMM). It associates to every sample a margin that is set to the difference between the average distance to heterogeneous 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. A similar method based on similar and dissimilar samples was proposed in [1]. In many linear embedding techniques, the neighborhood relationship is measured by an artificial constructed adjacent graph. Usually, the most popular adjacent graph construction manner is based on the K nearest neighbor and ε-neighborhood criteria. Once an adjacent graph is constructed, the edge weights are assigned by various strategies such as 0-1 weights and heat kernel function. Unfortunately, such adjacent graph is artificially constructed in advance, and thus it does not necessarily uncover the intrinsic local geometric structure of the samples. The performance of the technique is seriously sensitive to the

neighborhood size K.

As can be seen, all graph-based linear techniques require several parameters that should set in advance or tuned empirically using tedious cross-validation processes. Many existing works consider the neighborhood size as a user-defined parameter. It is set in advance to the same value for all samples. Moreover, some discriminant linear techniques employ an additive objective function that include a balance parameter that should also be determined.

In [4], the authors proposed a method called Locality Sensitive Discriminant Analysis. It computes a linear mapping that simultaneously maximizes the local margin between heterogeneous samples and pushes the homogeneous samples closer to each other. In this paper, we propose two novel solution schemes for LSDA. The first solution is a novel parameterless approach that has two important characteristics: (i) while all spectral-graph based manifold learning techniques (supervised and unsupervised) are depending on several parameters that require manual tuning, ours is parameter-free, and (ii) it adaptively estimates the local neighborhood surrounding each sample based on the data similarity. The second solution is an Exponential LSDA (ELSDA). In addition to the above two characteristics, ELSDA solves the Small Sample Size problem in the sense that there is no need to perform the pre-stage of dimensionality reduction.

Besides, we apply the proposed method to the problem of coarse 3D head pose estimation. The remainder of the paper is organized as follows. Section 2 describes the proposed parameter-free Locality Sensitive Discriminant Analysis. Section 3 describes the proposed Exponential Locality Sensitive Discriminant Analysis (ELSDA). Section 4 presents the application which deals with coarse 3D head pose estimation from images. It provides some experimental results obtained with two databases: FacePix and Pointing'04. Throughout the paper, capital bold letters denote matrices and small bold letters denote vectors.

2. Proposed parameter-free LSDA

2.1. Two graphs and adaptive set of neighbors

We assume that we have a set of N labeled samples $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^D$. In order to discover both geometrical and discriminant structure of the data manifold, we build two graphs: the within-class graph G_w and between-class graph G_b. Let $l(\mathbf{x}_i)$ be the class label of $\mathbf{x}_i$. For each data point $\mathbf{x}_i$, we compute two subsets, $N_b(\mathbf{x}_i)$ and $N_w(\mathbf{x}_i)$. $N_w(\mathbf{x}_i)$ contains the neighbors sharing the same label with $\mathbf{x}_i$, while $N_b(\mathbf{x}_i)$ contains the neighbors having different labels. We stress the fact that unlike the classical methods for neighborhood graph reconstruction, our algorithm adapts the size of both sets according to the local sample point $\mathbf{x}_i$ and its similarities with the rest of samples. Instead of using a fixed size for the neighbors, each sample point $\mathbf{x}_i$ will have its own adaptive set of neighbors. The set is computed in two consecutive steps. First, the average similarity of the sample $\mathbf{x}_i$ is computed by the total of all similarities with the rest of the data set (Eq. (1)). Second, the sets $N_w(\mathbf{x}_i)$ and $N_b(\mathbf{x}_i)$ are computed using Eqs. (2) and (3), respectively.

$$AS(\mathbf{x}_i) = \frac{1}{N}\sum_{k=1}^{N} sim(\mathbf{x}_i, \mathbf{x}_k) \tag{1}$$

where $sim(\mathbf{x}_i, \mathbf{x}_k) \in [0, 1]$ is a real value that encodes the similarity between $\mathbf{x}_i$ and $\mathbf{x}_k$. Simple choices for this function are the Kernel heat and the cosine.

$$N_w(\mathbf{x}_i) = \{\mathbf{x}_j \mid l(\mathbf{x}_j) = l(\mathbf{x}_i), sim(\mathbf{x}_i, \mathbf{x}_j) > AS(\mathbf{x}_i)\} \tag{2}$$

$$N_b(\mathbf{x}_i) = \{\mathbf{x}_j \mid l(\mathbf{x}_j) \neq l(\mathbf{x}_i), sim(\mathbf{x}_i, \mathbf{x}_j) > AS(\mathbf{x}_i)\} \tag{3}$$

Equation (2) means that the set of within-class neighbors of the sample $\mathbf{x}_i$, $N_w(\mathbf{x}_i)$, is all data samples that have the same label of $\mathbf{x}_i$ and that have a similarity higher than the average similarity associated with $\mathbf{x}_i$. There is a similar interpretation for the set of between-class neighbors $N_b(\mathbf{x}_i)$. From Equations (2) and (3) it is clear that the neighborhood size is not the same for every data sample. This strategy adapts the set of neighbors according to the local density and similarity between data samples in the original space. It is worth mentioning that for real data sets the mean similarity is always a positive value. Since the concepts of similarity and closeness of samples are tightly related, one can conclude, at first glance, that our introduced strategy is equivalent to the use of an ε-ball neighborhood. It is worth noticing that there are two main differences: (i) the use of an ε-ball neighborhood requires a user-defined value for the ball radius ε, and (ii) the ball radius is constant for all data samples, whereas in our strategy the threshold (1) depends on the local sample.

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:

$$W_{w,ij} = \begin{cases} sim(\mathbf{x}_i, \mathbf{x}_j)\ if\ \mathbf{x}_j \in N_w(\mathbf{x}_i)\ or\ \mathbf{x}_i \in N_w(\mathbf{x}_j) \\ 0, otherwise \end{cases}$$

$$W_{b,ij} = \begin{cases} 1\ \ if\ \mathbf{x}_j \in N_b(\mathbf{x}_i)\ or\ \mathbf{x}_i \in N_b(\mathbf{x}_j) \\ 0, otherwise \end{cases}$$

2.2. Optimal mapping

A linear embedding technique is described by a matrix transform that maps the original samples $\mathbf{x}_i$ into low dimensional samples $\mathbf{A}^T\mathbf{x}_i$. The number of columns of $\mathbf{A}$ defines

the dimension of the new subspace. We aim to compute a linear transform, $\mathbf{A}$, that simultaneously maximizes the local margins between heterogenous samples and pushes the homogeneous samples closer to each other (after the transformation). Mathematically, this corresponds to:

$$\min_{\mathbf{A}} \frac{1}{2} \sum_{i,j} \|\mathbf{A}^T (\mathbf{x}_i - \mathbf{x}_j)\|^2 W_{w,ij} \tag{4}$$

$$\max_{\mathbf{A}} \frac{1}{2} \sum_{i,j} \|\mathbf{A}^T (\mathbf{x}_i - \mathbf{x}_j)\|^2 W_{b,ij} \tag{5}$$

Using simple matrix algebra, the above criteria become respectively:

$$\begin{aligned} J_{homo} &= \frac{1}{2} \sum_{i,j} \|\mathbf{A}^T (\mathbf{x}_i - \mathbf{x}_j)\|^2 W_{w,ij} && (6)\\ &= \mathbf{tr}\left\{\mathbf{A}^T \mathbf{X} (\mathbf{D}_w - \mathbf{W}_w) \mathbf{X}^T \mathbf{A}\right\} && (7)\\ &= \mathbf{tr}\left(\mathbf{A}^T \mathbf{X} \mathbf{L}_w \mathbf{X}^T \mathbf{A}\right) && (8)\\ J_{hete} &= \frac{1}{2} \sum_{i,j} \|\mathbf{A}^T (\mathbf{x}_i - \mathbf{x}_j)\|^2 W_{b,ij} && (9)\\ &= \mathbf{tr}\left\{\mathbf{A}^T \mathbf{X} (\mathbf{D}_b - \mathbf{W}_b) \mathbf{X}^T \mathbf{A}\right\} && (10)\\ &= \mathbf{tr}\left(\mathbf{A}^T \mathbf{X} \mathbf{L}_b \mathbf{X}^T \mathbf{A}\right) && (11) \end{aligned}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ is the data matrix, $\mathbf{D}_w$ denotes the diagonal weight matrix, whose entries are column (or row, since $\mathbf{W}_w$ is symmetric) sums of $\mathbf{W}_w$, and $\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w$ denotes the Laplacian matrix associated with the graph G_w.

Given two individual optimization objectives as in Eq. (4) and Eq. (5) , we may construct a difference criterion to maximize:

$$J = \alpha J_{hete} - (1 - \alpha) J_{homo} \tag{12}$$

where $0 < \alpha < 1$ is a tradeoff parameter. In practice, however, it is hard and tedious to choose an optimal value for this parameter. Therefore, instead of using the difference criterion (12), we formulate the objective as a quotient so that α is removed as follows:

$$J = \frac{J_{hete}}{J_{homo}} = \frac{\mathbf{tr}\left(\mathbf{A}^T \mathbf{X} \mathbf{L}_b \mathbf{X}^T \mathbf{A}\right)}{\mathbf{tr}\left(\mathbf{A}^T \mathbf{X} \mathbf{L}_w \mathbf{X}^T \mathbf{A}\right)} = \frac{\mathbf{tr}\left(\mathbf{A}^T \widetilde{\mathbf{S}}_b \mathbf{A}\right)}{\mathbf{tr}\left(\mathbf{A}^T \widetilde{\mathbf{S}}_w \mathbf{A}\right)} \tag{13}$$

where the symmetric matrix $\widetilde{\mathbf{S}}_b = \mathbf{X} \mathbf{L}_b \mathbf{X}^T$ denotes the locality preserving between class scatter matrix, and the symmetric matrix $\widetilde{\mathbf{S}}_w = \mathbf{X} \mathbf{L}_w \mathbf{X}^T$ denotes the locality preserving within class scatter matrix. We point out that the trace ratio optimization problem (13) can be replaced by the simpler yet inexact trace form:

$$\max_{\mathbf{A}} \mathbf{tr}\left\{\left(\mathbf{A}^T \widetilde{\mathbf{S}}_w \mathbf{A}\right)^{-1} \left(\mathbf{A}^T \widetilde{\mathbf{S}}_b \mathbf{A}\right)\right\} \tag{14}$$

The above optimization problem has a closed form solution. The columns of the sought matrix $\mathbf{A}$ are given by the generalized eigenvectors associated with the largest eigenvalues of the following equation:

$$\widetilde{\mathbf{S}}_b \mathbf{A} = \widetilde{\mathbf{S}}_w \mathbf{A} \Lambda$$

where Λ is the diagonal matrix of eigenvalues.

In many real world problems such as face recognition, both matrices $\mathbf{X}\mathbf{L}_b\mathbf{X}^T$ and $\mathbf{X}\mathbf{L}_w\mathbf{X}^T$ can be singular. This stems from the fact that sometimes the number of images in the training set, N, is much smaller than the number of pixels in each image, D. This is known as the Small Sample Size (SSS) problem. To overcome the complication of singular matrices, original data are first projected to a PCA subspace or a random orthogonal space so that the resulting matrices $\mathbf{X}\mathbf{L}_b\mathbf{X}^T$ and $\mathbf{X}\mathbf{L}_w\mathbf{X}^T$ are non-singular.

3. Exponential LSDA

As can be seen, solving the SSS problem relied in the past mainly on applying a PCA on the raw data. However, PCA eliminates the null space of the total covariance matrix of data. It is well know that this null space may contain some discriminant information. Therefore, by using PCA as an initial stage in LSDA some discriminant information will not be handed over to the framework of LSDA.

3.1. Matrix Exponential

The matrix exponential is widely used in applications such as control theory, and Markov chain analysis. In this section, the definition and properties of matrix exponential are introduced. Given an $n \times n$ square matrix $\mathbf{S}$, its exponential is defined as follows:

$$\exp(\mathbf{S}) = \mathbf{I} + \mathbf{S} + \frac{\mathbf{S}^2}{2!} + \ldots + \frac{\mathbf{S}^m}{m!} + \ldots$$

where $\mathbf{I}$ is the identity matrix with the size of n×n. The properties of matrix exponential are listed as follows:

- $\exp(\mathbf{S})$ is a finite matrix.
- $\exp(\mathbf{S})$ is a full rank matrix.
- If matrix $\mathbf{S}$ commutes with $\mathbf{T}$, i.e., $\mathbf{ST} = \mathbf{TS}$, then $\exp(\mathbf{S}+\mathbf{T}) = \exp(\mathbf{S})\exp(\mathbf{T})$.
- For an arbitrary square matrix $\mathbf{S}$, there exists the inverse of $\exp(\mathbf{S})$. This is given by:

$$(\exp(\mathbf{S}))^{-1} = \exp(-\mathbf{S})$$

- If $\mathbf{T}$ is a nonsingular matrix, then $\exp(\mathbf{T}^{-1}\mathbf{S}\mathbf{T}) = \mathbf{T}^{-1}\exp(\mathbf{S})\,\mathbf{T}$.
- If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are eigenvectors of $\mathbf{S}$ that correspond to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are also eigenvectors of $\exp(\mathbf{S})$ that correspond to the eigenvalues $e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n}$. It is also well known that the matrix is non-singular.

A wide variety of methods for computing $\exp(\mathbf{S})$ were analyzed in [10]. The scaling and squaring method is one of the best methods for computing the matrix exponential [8].

3.2. Exponential LSDA

The exponential version of LSDA is obtained by using the exponential of $\widetilde{\mathbf{S}}_b$ and $\widetilde{\mathbf{S}}_w$. Thus, the criterion to be maximized becomes:

$$\max_{\mathbf{A}} \mathbf{tr}\left\{\left(\mathbf{A}^T \exp(\widetilde{\mathbf{S}}_w)\,\mathbf{A}\right)^{-1}\left(\mathbf{A}^T \exp(\widetilde{\mathbf{S}}_b)\,\mathbf{A}\right)\right\} \quad (15)$$

The columns of the sought matrix $\mathbf{A}$ are given by the generalized eigenvectors associated with the largest eigenvalues of the following equation:

$$\exp(\widetilde{\mathbf{S}}_b)\,\mathbf{A} = \exp(\widetilde{\mathbf{S}}_w)\,\mathbf{A}\,\Lambda$$

where Λ is the diagonal matrix of eigenvalues. It should be noted that both matrices $\exp(\widetilde{\mathbf{S}}_b)$ and $\exp(\widetilde{\mathbf{S}}_w)$ are full rank matrices. This means that even in the case where the Small Sample Size problem occurs, the linear transform can be estimated without reducing the dimensionality of the data samples. We point out that we must normalize scatter matrices, $\widetilde{\mathbf{S}}_b$ and $\widetilde{\mathbf{S}}_w$, because $\exp(\widetilde{\mathbf{S}}_b)$ and $\exp(\widetilde{\mathbf{S}}_w)$ may involve large numbers. This normalization is carried out using the Frobenius norms of the matrices.

4. Coarse 3D head pose

4.1. Background

The majority of research work in 3D head pose estimation deals with tracking full rigid body motion (6 degrees of freedom) for a limited range of motion (typically +/-45 out-of-plane) and relatively high resolution images. Besides, such systems typically require a 3D model [5, 6] as well as its initialization. There is a tradeoff between the complexity of the initialization process, the speed of the algorithm and the robustness and accuracy of pose estimation. Although the model-based systems can run in real-time, they rely on frame-to-frame estimation and hence are sensitive to drift and require relatively slow and non-jerky motion. These systems require initialization and failure recovery. For situations in which the subject and camera are separated by more than a few feet, full rigid body motion tracking of fine head pose is no longer practical. In this case, model-less coarse pose estimation can be used. It can be performed on a single image at any time without any model given that some pose-classified ground truth data are learned a priori. Coarse 3D pose estimation can play an important role in many applications. For instance, it can be used in the domain of face recognition either by using hierarchical models or by generating a frontal face image.

Figure 1. Some samples in FacePix data set.

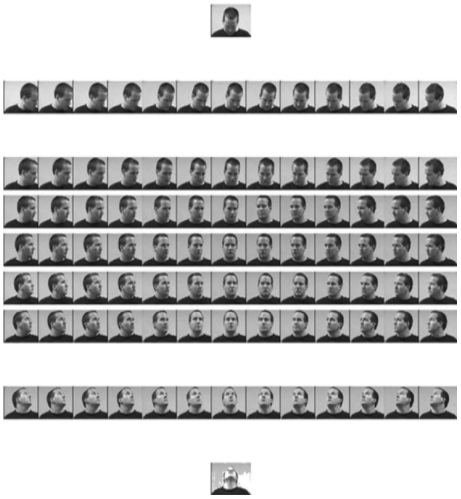


Figure 2. Some samples in Pointing'04 data set.

4.2. Databases

We evaluate the proposed methods with experiments on two public face data sets for face recognition and pose estimation.

The **FacePix** database includes a set of face images with pose angle variations. It is composed of 181 face images (representing yaw angles from $-90°$ to $+90°$ at 1 degree increments) of 30 different subjects, with a total of 5430 images. All the face images (samples) are 128 pixels wide and 128 pixels high. These images are normalized, such that the eyes are centered on the 57th row of pixels from the top, and the mouth is centered on the 87th row of pixels. Figure 1 provides examples extracted from the database, showing pose angles ranging from $-90°$ to $+90°$ in steps of $10°$. In our work, we downsample the set and only keep 10 poses in steps of $20°$. The images are resized to 32×32 pixels.

The **Pointing'04** Head-Pose Image Database consists of 15 sets of images for 15 subjects, wearing glasses or not and having various skin colors. Each set contains two series of

93 images of the same person at different poses (See Figure 2). In our work, we combine the two series into one single data set so that we can carry out tests on random splits. The pose or head orientation is determined by the pan and tilt angles, which vary from $-90°$ to $90°$ in steps of $15°$. Each pose has 2×15 images. In our work, we have extracted 65 poses (each pose has 2×15 images). The extracted 65 poses correspond to 13 yaw angles and 5 pitch angles. The images are resized to 32×32 pixels.

The ground truth data for this database are not as accurate as FacePix data set. Indeed, the method used for generating this data set belongs to directional suggestion category which assumes that each subject's head is in the exact same physical location in 3D space. Furthermore, it assumes that persons have the ability to accurately direct their head towards an object. The effect of this limitation will be obvious in the experimental results obtained with Pointing'04 data set.

4.3. Experimental results and method comparison

As mentioned earlier, the problem of coarse 3D head pose estimation can be cast into a classification problem. Estimating the pose class of a test face image is carried out in the new low dimensional space using the Nearest Neighbor classifier. For FacePix database, we have 10 different classes, each with 30 subjects. For each pose, l images are randomly selected for training and the rest are used for testing. For each given l, we average the results over 14 random splits. For every split, the pre-stage of dimensionality reduction (classical PCA) retained the top eigenvectors that correspond to 95% of the total variability. In general, the recognition rate varies with the dimension retained by the embedding method. In all our experiments, we recorded the best recognition rate for each algorithm.

Figure 3 depicts the recognition rate obtained with three different criteria: LSDA adopting the difference criterion (12), LSDA adopting the quotient criterion (14) and the ELSDA (15) when applied on FacePix database. These recognition rates depict the average recognition rates over 14 random splits of the data. The test sets were formed solely by unseen subjects. The number of training images l in Figures 3(a) and 3(b) was 5 and 20, respectively. As can be seen, ELSDA method has provided the best performance. Moreover, its recognition rate was high even when few dimensions were used. Besides, we can observe that the LSDA adopting the quotient criterion performed better than the LSDA adopting the difference criterion. The maximum dimension of LSDA is equal to the maximum dimension of the PCA pre-stage. Thus, for small training sets the maximum dimension has not exceeded 40. For the difference criterion, several trials have been performed in order to choose the optimal value for the parameter α. The results in Figure 3 correspond to those giving the best recognition rate in test sets. In the sequel, we report results obtained with the quotient criterion only.

Table 1 shows the recognition rates for different methods and for different numbers of training images, l. The methods used are: PCA, LPP, ANMM, the proposed parameterless LSDA with adaptive neighborhood size (fourth row), and the proposed ELSDA (fifth row). As can be seen, our proposed ELSDA achieved 90.1% recognition rate when 15 face images per pose/class (yaw angle) were used for training. We stress the fact that ANMM is one of the most powerful discriminant analysis methods recently developed. It should be noticed that the performance of the proposed method is close to that of ANMM. However, unlike ANMM which needs two user-defined parameters, our proposed methods do not need any parameter setting.

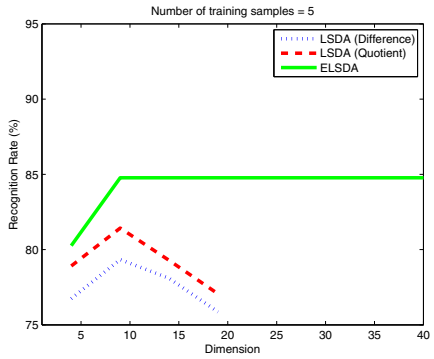
Table 2 shows the recognition rates for pitch and yaw angles obtained with PCA, LPP, ANMM, and the proposed methods when applied on Pointing'04 data set. For these methods, the linear mapping was learned using the 65 classes (poses) obtained by 13 yaw angles and 5 pitch angles. The recognition rates were computed separately for the pitch and yaw angles for all test images. As can be seen, our proposed method achieved the best performance. The recognition rates were relatively low since the ground-truth of Pointing'04 database was not accurate.

5. Conclusion

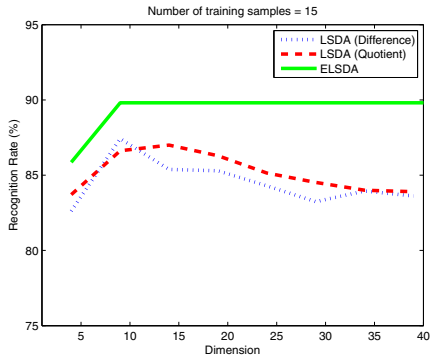
We developed two improved solution schemes for Locality Sensitive Discriminant Analysis. The first solution is a novel parameterless approach. The second solution is an exponential LSDA. The Exponential LSDA (ELSDA) is equivalent to transforming the original data into a new space by distance diffusion mapping, and then, the classical LSDA criterion is applied in such a new space. As a result of diffusion mapping, the margin between different classes is enlarged, which is helpful in improving classification accuracy. We used both proposed frameworks for coarse 3D head pose estimation. Experimental results demonstrate the advantage of the ELSDA over some state-of-art solutions. Future work will investigate the use of other image representations in order to estimate the coarse 3D head pose.

l	5	10	15
PCA	77.1% (20)	83.8% (35)	84.1% (25)
LPP	69.4% (20)	81.8% (20)	84.0% (30)
ANMM	82.7% (10)	84.6% (10)	88.9% (10)
LSDA	81.4% (10)	85.3% (10)	87.0% (15)
ELSDA	**85.4**% (10)	**88.0**% (10)	**90.1**% (10)

Table 1. Best average recognition accuracy (%) on FacePix set over 14 random splits. Each column corresponds to a fixed number of training images. The number appearing in parenthesis corresponds to the optimal dimensionality of the embedded subspace (at which the maximum average recognition rate has been reported).



(a)



(b)

Figure 3. Average recognition rate for the proposed LSDA schemes obtained with FacePix face data set using 5 training samples **(a)** and 15 training samples **(b)**.

Acknowledgment. This work was partially supported by the Spanish Government under the project TIN2010-18856.

References

[1] B. Alipanahi, M. Biggs, and A. Ghodsi. Distance metric learning vs. Fisher discriminant analysis. In *AAAI Conference on Artificial Intelligence*, 2008. 393

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. 393

[3] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag New York, 2005. 393

[4] D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. In *International Joint Conference on Artificial Intelligence*, 2007. 394

[5] F. Dornaika and J. Ahlberg. Face and facial feature tracking using deformable models. *International Journal of Image and Graphics*, 4(3):499–532, 2004. 396

[6] F. Dornaika and F. Davoine. On appearance based face and facial action tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(9):1107–1124, September 2006. 396

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1990. 393

	Pitch	Yaw
PCA	44.1% (70)	**46.7**% (70)
LPP	40.4% (25)	38.6% (25)
ANMM	44.7% (35)	43.9% (35)
LSDA	46.7% (30)	43.4% (30)
ELSDA	**47.7**% (30)	44.7% (30)

(a) training images/testing images 4/26.

	Pitch	Yaw
PCA	44.4% (85)	50.8% (90)
LPP	39.7% (30)	42.2% (25)
ANMM	44.1% (45)	48.9% (45)
LSDA	47.7% (20)	48.6% (20)
ELSDA	**48.4**% (25)	**50.1**% (25)

(b) training images/testing images 6/24.

	Pitch	Yaw
PCA	44.5% (90)	48.75% (65)
LPP	41.9% (45)	44.6% (35)
ANMM	46.6% (80)	50.5% (40)
LSDA	**50.2**% (20)	51.0 % (15)
ELSDA	49.5% (35)	**51.7**% (35)

(c) training images/testing images 10/20.

Table 2. Best average recognition accuracy (%) on Pointing'04 data set for pitch and yaw angles (over 10 random splits).

[8] N. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1196, 2005. 396

[9] A. M. Martinez and M. Zhu. Where are linear feature extraction methods applicable? *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(12):1934–1944, 2005. 393

[10] C. Moler and C. V. Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49, 2003. 396

[11] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 393

[12] B. Z. S. Yan, D. Xu and H.-J. Zhang. Graph embedding: A general framework for dimensionality reduction. In *Int. Conference on Computer Vision and Pattern Recognition*, 2005. 393

[13] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 393

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991. 393

[15] F. Wang, X. Wang, D. Zhang, C. Zhang, and T. Li. Marginface: A novel face recognition method by average neighborhood margin maximization. *Pattern Recognition*, 42:2863–2875, 2009. 393

[16] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extension: a general framework for dimensionality reduction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007. 393