IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No.2, pp. 228-233, 2001

PCA versus LDA

Aleix M. Martínez and Avinash C. Kak

Robot Vision Lab
School of Electrical and Computer Engineering
Purdue University, IN 47907-1285
{aleix, kak}@ecn.purdue.edu

Abstract

In the context of the appearance-based paradigm for object recognition, it is generally believed that algorithms based on LDA (Linear Discriminant Analysis) are superior to those based on PCA (Principal Components Analysis). In this communication we show that this is not always the case. We present our case first by using intuitively plausible arguments and then by showing actual results on a face database. Our overall conclusion is that when the training dataset is small, PCA can outperform LDA, and also that PCA is less sensitive to different training datasets.

Keywords: face recognition, pattern recognition, principal components analysis, linear discriminant analysis, learning from undersampled distributions, small training datasets.

1 Introduction

Many computer vision systems reported in the literature now employ the appearance-based paradigm for object recognition. One primary advantage of appearance-based methods is that it is not necessary to create representations or models for objects since, for a given object, its model is now implicitly defined by the selection of the sample images of the object.

When using appearance-based methods, we usually represent an image of size $n \times m$ pixels by a vector in an $n \cdot m$ dimensional space. In practice, however, these $(n \cdot m)$-dimensional spaces are too large to allow robust and fast object recognition. A common way to attempt to resolve this problem is to use dimensionality reduction techniques. Two of the most popular techniques for this purpose are: Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA, also known as Fisher Discriminant Analysis - FDA).

PCA has been used in face recognition [16, 5, 18, 9, 7], handprint recognition [11], human-made object recognition [12], industrial robotics [13], and mobile robotics [19]. LDA has been used in face recognition [17, 1] and mobile robotics [19]. LDA has also been proposed for generic object recognition [17], but results using a large database of objects have not been reported yet.

Of late, there has been a tendency to prefer LDA over PCA because, as intuition would suggest, the former deals directly with discrimination between classes, whereas the latter deals with the data in its entirety for the principal components analysis without paying any particular attention to the underlying class structure. *It is this tendency in the vision community that is subject to examination in this paper.*

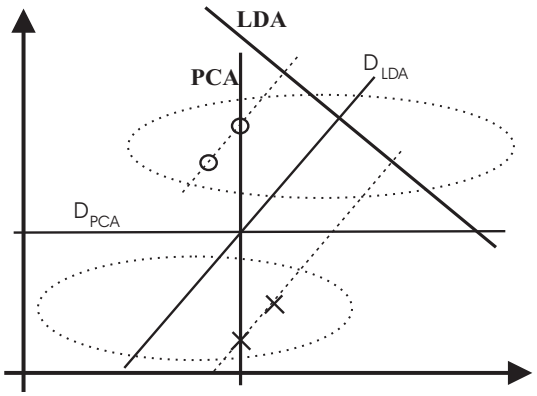

Figure 1: *There are two different classes embedded in two different "Gaussian-like" distributions. However, only two sample per class are supplied to the learning procedure (PCA or LDA). The classification result of the PCA procedure (using only the first eigenvector) is more desirable than the result of the LDA. D_{PCA} and D_{LDA} represent the decision thresholds obtained by using nearest-neighbor classification.*

In this paper we will show that the switch from PCA to LDA may not always be warranted and may sometimes lead to faulty system design, especially if the size of the learning database is small.[1] Our claim carries intuitive plausibility, as can be established with the help of Fig. 1. This figure shows two learning instances, marked by circles and crosses, for each class whose underlying (but unknown) distribution is shown by the dotted curve. Taking all of the data into account, PCA will compute a vector that has largest variance associated with it. This is shown by the vertical line labeled **PCA**. On the other hand, LDA will compute a vector which best discriminates between the two classes. This vector is shown by the diagonal line labeled **LDA**. The decision thresholds yielded by the nearest neighbor approach for the two cases are marked $\mathbf{D_{PCA}}$ and $\mathbf{D_{LDA}}$. As can be seen by the manner in which the decision thresholds intersect the ellipses corresponding to the class distributions, PCA will yield superior results.

Although examples such as the one depicted in Fig. 1 are quite convincing with regard to the claim that LDA is not always be superior to PCA, we still bear the burden of establishing our claim with the help of actual data. This we will do in the rest of this paper with the help of a face databases: the AR-face database (a publicly available data-set).

As additional evidence in support of our claim, we should also draw the attention of the reader to some of the results of the September 96 FERET competition [15]. In particular, we wish to point to the LDA results obtained by the University of Maryland [2] that compare unfavorably with respect to a standard PCA approach as described in [18]. A notable characteristic of the data used in such experiments was that only one or two learning samples per class were given to the system.

Of course, as one would expect, given large and representative learning datasets, LDA should outperform PCA. Simply to confirm this intuitive conclusion, we will also show results on the AR-database of faces. In this database, the sample size per class for learning is larger than was the case for the FERET competition.[2] For example, the database we will use to show LDA outperforming PCA has images of 13 different facial shots corresponding to different expressions or illumination conditions and/or occlusions for each subject.

[1] This is not to cast any aspersions on the system design employed by the previously cited contributions. Our claim has validity only when the size of the learning database is insufficiently large or non-uniformly distributed.

[2] Note that FERET deals with a very large number of classes, but the number of classes is not the issue in this paper. Our main concern here is with the problems caused by insufficient data per class available for learning.

(a) (b) (c)

Figure 2: *Localization and morphing: To morph an image, we must first localize the boundaries of the face, as shown by the overlaid horizontal and vertical lines in the left image in (a). We must also locate basic features such as the nose, as shown by the overlaid line running through the nose in the same image in (a), and the eyes, as shown by the white dots in the eyes in the right image in (a). Shown in (b) and (c) are two different examples of morphed faces.*

2 Localization and Morphing of Face Images

We will be comparing PCA and LDA with regard to only the identification of faces, independently of any localization and scale related issues. Therefore, we have manually carried out the localization step, followed by a morphing step so that each face occupies a fixed size array of pixels.

Formally, let us consider the set of N sample images $\mathbf{I}^i_{p\times q}$ (where p is the number of columns, q the number of rows and i the image number, i.e. $i = \{1, ..., N\}$).

We first manually localize the left, the right, the top and the bottom limits of the face as well as the left and the right eyes and the nose; as shown in Fig. 2(a). After localization, faces are morphed so as to fit a grid of size 85 by 60. Figs. 2(b) and (c) show the final results of morphing for two different subjects. As shown, after morphing the eye centers, the medial line of the nose, and the arc where the lips join, etc., are at the same pixel coordinates in all images. We will refer to these new images by $\hat{\mathbf{I}}^i_{n\times m}$, where n and m are the dimensions of the morphed image.

Each of these images can now be segmented by means of an oval-shaped mask centered at the middle of the morphed image rectangle. The pixels in the oval are vectorized into a t-dimensional vector $\mathbf{x}_i$ (where t corresponds to the number of pixels within the oval-shaped segment) by reading pixel values within the oval segment in a raster-scan manner. The vectors obtained in this manner from all N sample images will be denoted $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$.

3 The PCA Space

Given a t-dimensional vector representation of each face, the Principal Component Analysis (PCA) [4] can be used to find a subspace whose basis vectors correspond to the maximum-variance directions in the original space. Let $\mathbf{W}$ represent the linear transformation that maps the original t-dimensional space onto a f-dimensional feature subspace where normally $f \ll t$. The new feature vectors $\mathbf{y}_i \in \Re^f$ are defined by $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$, $i = 1, \ldots, N$. The columns of $\mathbf{W}$ are the eigenvalues $\mathbf{e}_i$ obtained by solving the eigenstructure decomposition $\lambda_i\mathbf{e_i} = \mathbf{Q}\mathbf{e_i}$, where $\mathbf{Q} = \mathbf{X}\mathbf{X}^T$ is the covariance matrix, and λ_i the eigenvalue associated with the eigenvector $\mathbf{e}_i$. Before obtaining the eigenvectors of $\mathbf{Q}$: (i) the vectors are normalized such that $\|\mathbf{x}_i\| = 1$ to make the system invariant to the intensity of the illumination source, and (ii) the average of all images is subtracted from all normalized vectors to ensure that the eigenvector with the highest eigenvalue represents the dimension in the eigenspace in which variance of vectors is maximum in a correlation sense. The covariance matrix $\mathbf{Q}$ is normally too large for an easy computation of the eigenvectors. Fortunately, several ways to get around this difficulty have been proposed [10, 16, 5, 12].

Pentland et al. [14] have empirically shown that superior face recognition results are achieved when the first three eigenvectors are not used (because the first three eigenvectors seem to represent changes in illumination). It has been recently shown that the elimination of more than 3 eigenvectors will, in general, worsen the results [8]. In this paper, we will also analyze how the elimination of these first three eigenvectors affects recognition performance.

4 The LDA Space

Linear Discriminant Analysis (LDA) [3, 4] searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). More formally, given a number of independent features relative to which the data is described, LDA creates a linear combination of these which yields the largest mean differences between the desired classes. Mathematically speaking, for all the samples of all classes we define two measures: (i) one called *within-class* scatter matrix, as given by

$$S_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\mathbf{x}_i^j - \mu_j)(\mathbf{x}_i^j - \mu_j)^T$$

where $\mathbf{x}_i^j$ is the ith sample of class j, μ_j is the mean of class j, c is the number of classes, and N_j the number of samples in class j; and (ii) the other is called *between-class* scatter matrix

$$S_b = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T$$

where μ represents the mean of all classes.

The goal is to maximize the between-class measure while minimizing the within-class measure. One way to do this is to maximize the ratio $\frac{det|S_b|}{det|S_w|}$. The advantage of using this ratio is that it has been proven [3] that if S_w is a non-singular matrix then this ratio is maximized when the column vectors of the projection matrix, $\mathbf{W}$, are the eigenvectors of $S_w^{-1}S_b$. It should be noted (and it is very easy to prove) that: (i) there are at most $c-1$ nonzero generalized eigenvectors, and so an upper bound on f is $c-1$, and (ii) we require at least $t+c$ samples to guarantee that S_w does not become singular (which is almost impossible in any realistic application). To solve this, [17] and [1] propose the use of an intermediate space. In both cases, this intermediate space is chosen to be the PCA space. Thus, the original t-dimensional space is projected onto an intermediate g-dimensional space using PCA and then onto a final f-dimensional space using LDA.

5 Experimental Results

The results to be presented in this section were obtained using the AR-face database [6].[3] This database consists of over 3200 color images of the frontal images of faces of 126 subjects. There are 26 different images for each subject. For each subject, these images were recorded in two different sessions separated by two weeks, each session consisting of 13 images. For illustration, these images for one subject are shown in Fig. 3. All images were taken by the same camera under tightly controlled conditions of illumination and viewpoint. Each image in the database consists of a 768×576 array of pixels, and each pixel is represented by 24 bits of RGB color values.

For the experiments reported in this section, 50 different individuals (25 males and 25 females) were randomly selected from this database. As stated earlier, images were morphed to the final 85×60 pixel arrays, segmented using an oval-shaped mask, and converted to gray-level images by adding all three color channels, i.e. $I = \frac{1}{3}(R+G+B)$.

[3] The AR database of face images is publicly available from `http://rvl1.ecn.purdue.edu/ARdatabase/ARdatabase.html`

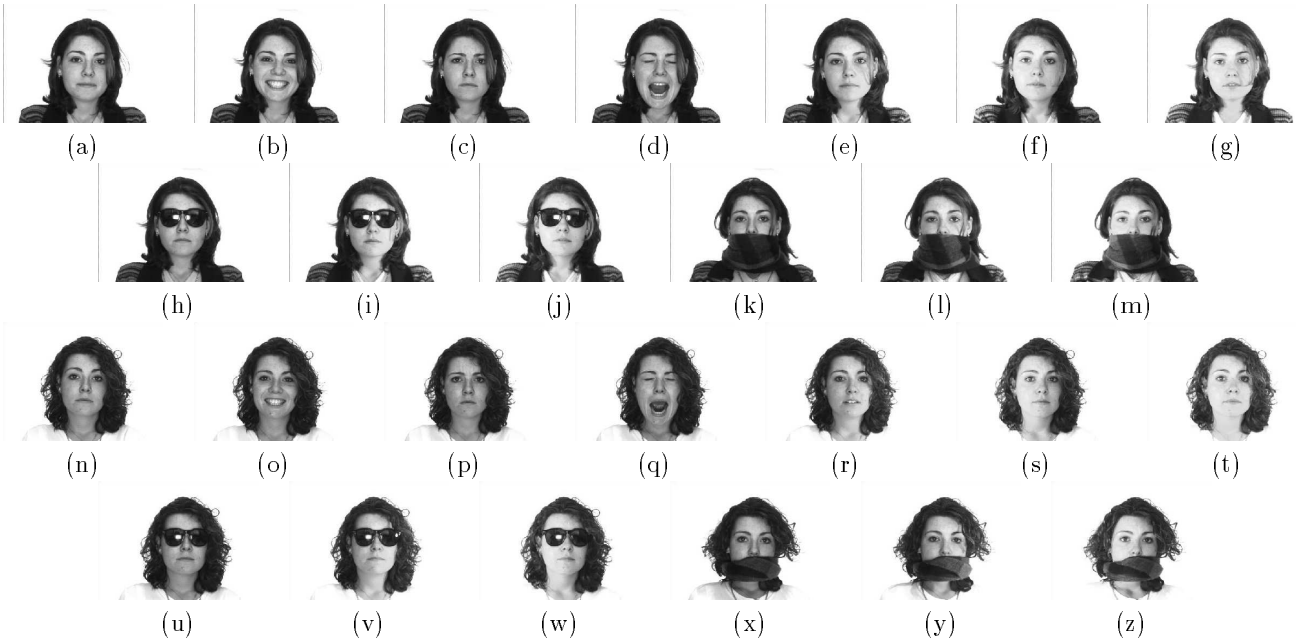

Figure 3: *Images of one subject in the AR face database. The images (a) through (m) were taken during one session and the images (n) though (z) at a different session.*

5.1 Small Training Data Sets

As discussed in the introduction, when a small (or non-representative) training data set is used, there is no guarantee that LDA will outperform PCA. In the introduction, this was justified on purely intuitive grounds with the help of Fig. 1. In this subsection, we study this effect on real data using the images of the AR face database.

To simulate the effects of a small training data set, our results here use two images per person for training and five for testing. In this subsection, only the non-occluded images recorded during the first of the two sessions are used. For example, for the subject shown in Fig. 3, only the images labeled (a) through (g) are used. Of the seven unoccluded images for each subject, there are obviously many different ways – a total of 21 – of selecting two for training and five for testing. We will use all these 21 different ways of separating the data into the training and the testing parts for the results reported here.

To each of the 21 different training and testing datasets created in the manner described above, we applied (i) PCA, (ii) PCA without the first three eigenvectors, and (iii) LDA. Testing was carried out by using the nearest-neighbor algorithm using the standard L_2-norm for the Euclidean distance. The datasets were indexed 1, 2, ..., 21, and the test results for the i^{th} dataset were represented by **Test#i**. In Fig. 4 we have shown the results for **Test#4**, **Test#6**, and **Test#9**. The horizontal coordinate in this figure represents the parameter f. Recall from Section 3 that f is the dimensionality of the final subspace in which face identification takes place. As was stated earlier, for LDA we also need to specify the value of the parameter g, which is the dimensionality of the intermediate space described in Section 3. Obviously, the value chosen for g would strongly affect the face recognition results. In order to make a fair comparison between PCA and LDA, for each value of f, we tried all possible values of g from a low of 15 to its maximum possible value of 50. The LDA results shown in Fig. 4 for each value of f are based on that value of g which yielded the best recognition rate.

We chose **Test#4**, **Test#6**, and **Test#9** for display in Figure 4 because each represents a different type

Method	$f = 1$	$f = 2$	$f = 3$	$f = 4$	$f = 5$	$f = 6$	$f = 7$	$f = 8$	$f = 9$	$f = 10$
PCA	6	9	13	9	9	9	7	4	4	3
PCA w/o 3	4	1	0	0	0	0	0	0	0	0
LDA	11	11	8	12	12	12	14	17	17	18

Table 1: *This table summarizes the results for all 21 ways of dividing the data into training and testing subsets. For the value of the dimensionality parameter f from 1 to 10. The top row shows the number of cases in which the basic PCA outperformed the other two algorithms, the middle row the number of cases in which the PCA without the first three eigenvectors did the best, and last row the number of cases for which LDA did the best.*

of comparative performance from the three algorithms tested. The performance curves for **Test#6** are typical of the datasets for which PCA outperformed LDA. The performance curves for **Test#4** are typical for those datasets for which PCA proved to be superior to LDA for some values of the dimensionality f and inferior for others. Finally, the performance curves in **Test#9** are typical for those datasets for which LDA outperformed PCA. That the same database should yield such different results is not surprising at all. Going back to Fig. 1, it is not difficult to visualize that if we altered the locations of the training samples shown there, we could get decision thresholds that would show either LDA outperforming PCA, or neither LDA or PCA yielding a clear separation between the underlying class distributions.

In Fig. 4, we have focussed on only low-dimensional spaces because we want to make a comparison of the *most* discriminant features for the LDA case with the *most* descriptive (in the sense of packing the most "energy") features for the PCA case. However, there is still the matter of whether or not the same conclusions hold in high-dimensional spaces that one would need to use for achieving sufficiently high recognition rates demanded by practical face recognition systems. Shown in Fig. 5 are results similar to those in Fig. 4 but when the number of dimensions is large. The three test cases shown in Fig. 5 were chosen so that one reflected PCA outperforming LDA (**Test#2**), the other LDA outperforming PCA (**Test#4**) , and third with neither conclusively outperforming the other (**Test#8**), as the number of dimensions was increased. In case the reader is wondering as to why the LDA curves do not go beyond $f = 40$, this is dictated by the following two considerations:

- The dimensionality of LDA is upper-bounded by $c-1$, where c is the number of classes, since that is the rank of the $S_w^{-1}S_b$ matrix. Since we used 50 classes, this gives us an upper bound of 49 for the dimensionality of the LDA space.
- The dimensionality of the underlying PCA space (from which the LDA space is carved out) cannot be allowed to exceed $N - c$ where N is the total number of samples available. This is to prevent S_w from becoming singular. Since we used 100 samples and since we have 50 classes, the dimensionality of the underlying PCA space cannot be allowed to exceed 50.

Since it makes no sense to extract a 49 dimensional LDA subspace out of a 50 dimensional PCA space, we arbitrarily hard-limited the dimensionality of the LDA space to 40.

Table 1 summarizes the results for all 21 cases of training and testing datasets for the case of low-dimensionality. And, Table 2 does the same for the case of high-dimensionality. For each value of the dimensionality parameter f, the top row shows the number of cases for which the basic PCA outperformed the other two algorithms, the middle row the number of cases for which PCA without the first three eigenvectors was the best, and the last row the number of cases for which LDA outperformed PCA.

It is interesting to note from Table 1 that if we limit the dimensionality f of the final subspace to between roughly 1 and 6, PCA (including PCA without the first three eigenvectors) can be expected to outperform LDA

Method	$f = 20$	$f = 30$	$f = 40$
PCA	3	2	2
PCA w/o 3	0	0	0
LDA	18	19	19

Table 2: *Same as in Table 1 but for high-dimensional spaces.*

almost just as frequently as the other way around. But, as established by the dataset for the **Test#6** (results shown in Fig. 4), we can also expect PCA to outperform LDA regardless of the value of the dimensionality parameter f. It would not be a stretch of the imagination to say that PCA outperforms LDA when it is more important to somehow learn the general appearance of a face from all the training samples supplied than how to best discriminate between faces of different subjects. For high-dimensional spaces, we can draw comparable conclusions, except that LDA has a greater chance of outperforming PCA for our data set. But note that this conclusion applied only to the specific data set used by us for the experiments reported here. One may end up with an entirely different conclusion for a different data set.

Another observation we wish to make is that while the suppression of the first three eigenvectors improves the performance of PCA in the presence of illumination variations, the LDA transform will usually do even better for small training datasets. This is the reason why the "PCA without the first three eigenvectors" wins so infrequently.

One last observation regarding small training datasets has to do with the relative behavior of PCA and LDA as the dimensionality parameter f becomes larger. The performance of both transforms gets better as the value of f increases. What's different between the two is that while the recognition rate with PCA saturates around 28% to 53% for $f = 10$ and around 44% to 75% for $f = 80$ for all different datasets, the performance of LDA can vary widely. For the experiments under discussion, the recognition rate obtained with LDA varied from a low of 31% to a high of 68% for $f = 10$ and 41 to 82 for $f = 40$.

5.2 Using Representative Samples Per Class

The previous subsection showed unequivocally that when the number of training samples per class is small, it is possible for PCA to outperform LDA. In this subsection, we reinforce what our intuition would naturally suggest: that when the number of learning samples is large and representative for each class, LDA will outperform PCA. The study in this subsection is carried out using all the 26 images for each subject in the AR database. That means that for a subject such as the one shown in Fig. 3, all the images labeled (a) through (z) are now used. As mentioned earlier, the first thirteen of these, (a) through (m), were taken in one session; these are used for training now. And the last thirteen, (n) through (z), were taken in a second session; these are used for testing.[4]

As for the experiments described in the preceding subsection, for the LDA algorithm we chose that value for the dimensionality parameter g which gave us the best final classification results for each value of f, the dimensionality of the final subspace in which classification is carried out. (Recall that g is the intermediate subspace needed in the implementation of LDA.) For each value of f, we tried all values of g from a low of 50 to the maximum allowed value of 600. The best final results were usually obtained for small values of g. This can be explained on the basis of the fact that the number of samples that a learning technique needs is proportional to the dimensionality of the dataset. This implies that in order to obtain good results with LDA, the total number

[4] In face recognition circles, this is referred to as the problem of recognizing "duplicates." A "duplicate" is an image of a face that is taken at a different time, weeks or even months later. Duplicate images for testing are taken under roughly the same illumination conditions and with similar occlusions as in the original set. Facial expressions should also be nearly the same.

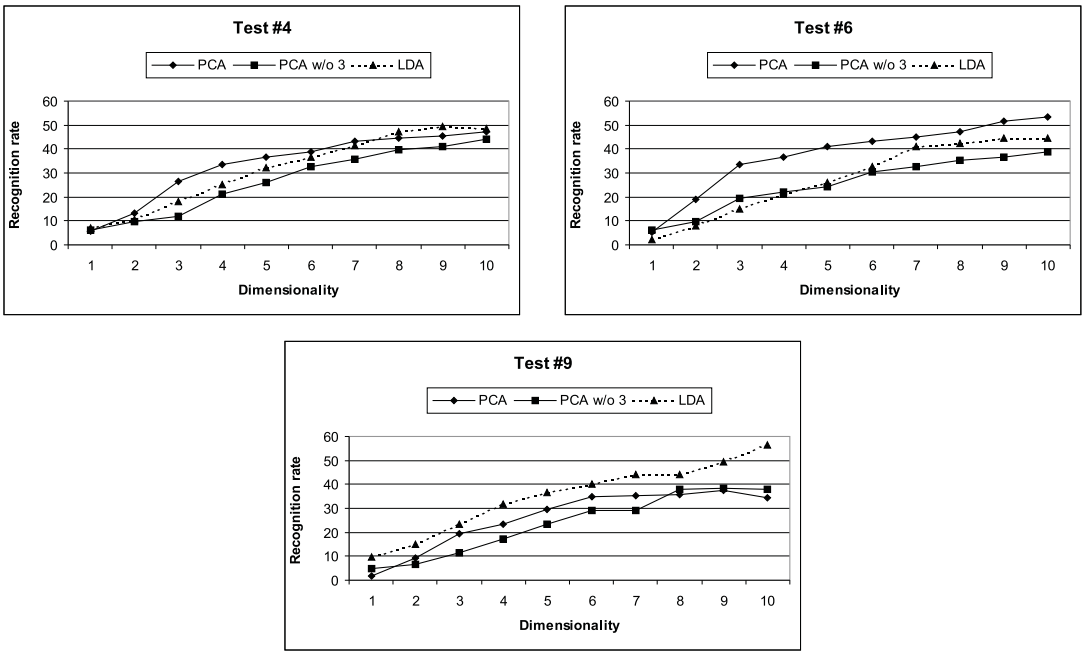

Figure 4: *Shown here are performance curves for three different ways of dividing the data into a training set and a testing set.*

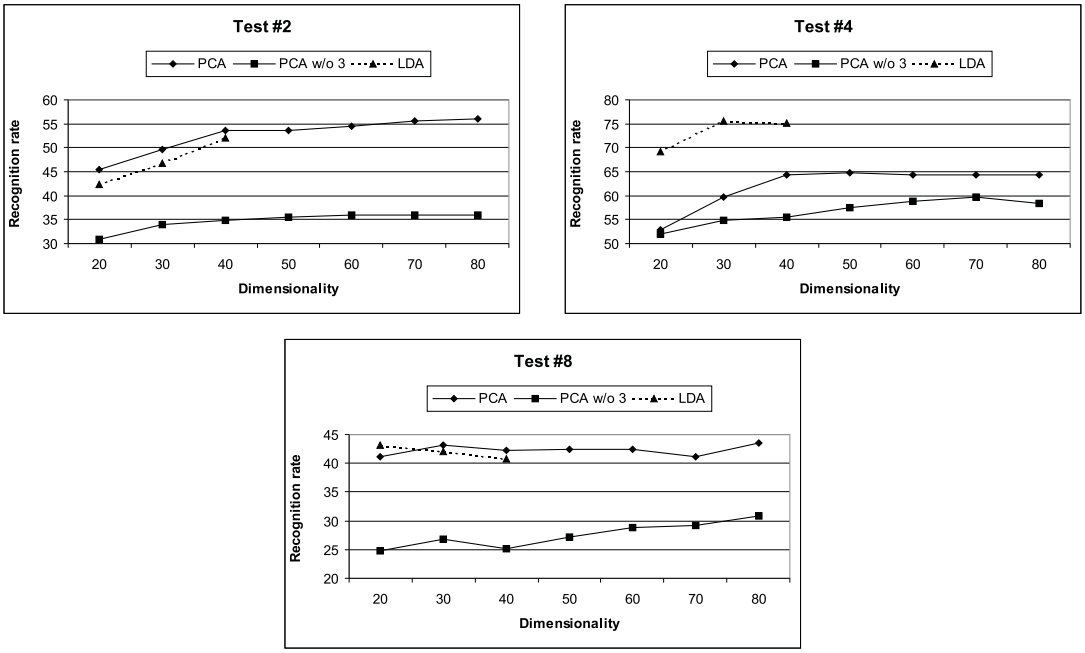

Figure 5: *Performance curves for the high-dimensional case.*

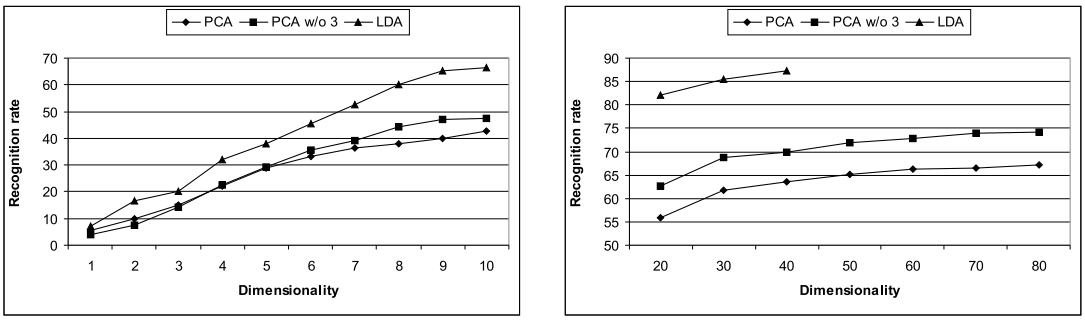

Figure 6: *Results obtained for each of the three algorithms while using 50 individuals (classes).*

of samples should be much larger than $t + c$, or, equivalently, much larger than the value of g. Fig. 6 shows the results. As was expected, LDA outperforms PCA when a large and representative training dataset is used.

6 Conclusions

Appearance-based methods are widely used in object recognition systems. Within this paradigm, PCA and LDA have been demonstrated to be useful for many applications such as face recognition. Although one might think that LDA should always outperform PCA (since it deals directly with class discrimination), empirical evidence suggests otherwise. This paper discusses the reasons for this seemingly anomalous behavior.

Our Fig. 1 illustrates how PCA might outperform LDA when the number of samples per class is small or when the training data non-uniformly sample the underlying distribution. In many practical domains, and especially in the domain of face recognition, one never knows in advance the underlying distributions for the different classes. So one could argue that in practice it would be difficult to ascertain whether or not the available training data is adequate for the job.

The experiments we report validate our claim. Several of our experiments show the superiority of PCA over LDA, while others show the superiority of LDA over PCA. When PCA outperforms LDA, the number of training samples per class is small, but not atypical of the data sizes used previously by some researchers.

References

[1] P.N. Belhumeour, J.P. Hespanha and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI**-19(7):711-720, 1997.

[2] K. Etemad and R. Chellapa, "Discriminant analysis for recognition of human face images," Journal of Optics of American A 14(8):1724-1733, 1997.

[3] R.A. Fisher, "The Statistical Utilization of Multiple Measurements," Annals of Eugenics, 8:376-386, 1938.

[4] K. Fukunaga, "Introduction to Statistical Pattern Recognition (second edition)," Academic Press, 1990.

[5] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI**-12(1):103-108, 1990.

[6] A.M. Martínez and R. Benavente, "The AR-face database," CVC Technical Report # 24, June 1998.

[7] A.M. Martínez, "Recognition of Partially Occluded and/or Imprecisely Localized Faces Using a Probabilistic Approach," Proc. of Computer Vision and Pattern Recognition, Vol. 1, pp. 712-717, Hilton Head, June 2000.

[8] H. Moon and P.J. Phillips, "Analysis of PCA-based Face Recognition Algorithms," Empirical Evaluation Techniques in Computer Vision (K.J. Bowyer and P.J. Phillips, Eds.), IEEE Computer Science 1998.

[9] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI**-19(7):696-710, 1997.

[10] H. Murakami and B.V.K. Vijaya Kumar, "Efficient Calculation of Primary Images from a Set of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI**-4(5):511-515, 1982.

[11] H. Murase, F. Kimura, M. Yoshimura and Y. Miyake, "An improvement of the auto-correlation matrix in pattern matching method and its application to handprinted 'HIRAGANA'," Transactions on IECE J64-D(3), 1981.

[12] H. Murase and S. Nayar, "Visual Learning and Recognition of 3-D Objects from Appearance," International Journal of Computer Vision, 14:5-24, 1995.

[13] S.K. Nayar, N.A. Nene and H. Murase, "Subspace Methods for Robot Vision," IEEE Transactions on Robotics and Automation **RA**-12(5):750-758, 1996.

[14] A. Pentland, T. Starner, N. Etcoff, N. Masoiu, O. Oliyide and M. Turk, "Experiments with eigenfaces," Looking at people, workshop of IJCAI 1993.

[15] P.J. Phillips, H. Moon, P Rauss and S.A. Rizvi, "The FERET evaluation methodology for face-recognition algorithms," Proceedings of the First International Conference on Audio and Video-based Biometric Person Authentification, Crans-Montana (Switzerland), 1997.

[16] L. Sirovich and M. Kirby, "A low-dimensional procedure for the characterization of human faces," J. Opt. Soc. Amer. A, 4(3):519-524, 1987.

[17] D.L. Swets and J.J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI**-18(8):831-836, 1996.

[18] M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal Cognitive Neuro-science, 3(1), 1991.

[19] J.J. Weng, "Crescepton and SHOSLIF: Towards Comprehensive Visual Learning," Early Visual Learning (S.K. Nayar and T. Poggio, Eds.), pp. 183-214, Oxford University Press, 1996.