ON OPTIMAL SUBSPACES FOR APPEARANCE-BASED OBJECT RECOGNITION

Q. Wu‡, Z. Liu†, Z. Xiong‡, Y. Wang‡, T. Chen‡, and K. R. Castleman‡

‡Advanced Digital Imaging Research, LLC.
2450 South Shore Blvd., Suite 305, League City, TX 77573
Email: {qwu,wyp,chen,castleman}@adires.com
†Dept of Electrical Engineering, Texas A&M University,
College Station, TX 77843
Email: {liuzm,zx}@ee.tamu.edu

ABSTRACT

On the subject of optimal subspaces for appearance-based object recognition, it is generally believed that algorithms based on LDA (Linear Discriminant Analysis) are superior to those based on PCA (Principal Components Analysis), provided that relatively large training data sets are available [3, 5]. In this paper, We show that while this is generally true for classification with the nearest-neighbor classifier, it is not always the case with a maximum-likelihood classifier. We support our claim by presenting both intuitively plausible arguments and actual results on a large data set of human chromosomes. Our conjecture is that perhaps only when the underlying object classes are linearly separable would LDA be truly superior to other known subspaces of equal dimensionality.

1. INTRODUCTION

Appearance-based object recognition technology has become one of the relatively popular paradigms used in computer vision systems lately. Most of the research on automated human face recognition, for example, is based upon appearance-based object recognition and has reportedly achieved notable but limited success [5]. One obvious advantage of appearance-based methods is that the representations of objects are all embodied in the sample images, thus avoiding the need to define explicit features or models for the objects.

As image arrays, rather than feature vectors, are used in the appearance-based methods, projecting data from the original image space into a subspace for dimensionality reduction becomes necessary for practical reasons. Among various subspace approaches, PCA (Principal Components Analysis) [2] and LDA (Linear Discriminant Analysis) [3] have received considerable attention because of their well-defined optimal properties for representing the data and successful applications in face recognition [1-5], robotics [5], human-made object recognition [5], and image retrieval [6].

Since LDA is about a subspace optimized over discrimination between classes, while PCA is about a subspace optimized over representing the data in its entirety with least MSE (mean square error), it is generally believed that algorithms based on LDA are superior than those based on PCA [3, 4]. Only recently however, the issue of limitations of LDA has been addressed, and a study that indicates PCA can outperform LDA when the training data set is small has been reported [5].

In this paper, we point out an observation that there is yet another important limitation of LDA which we believe was previously unknown. We notice that while the superiority of LDA generally holds for classification with the nearest-neighbor classifier used in most previous LDA studies, it is not always the case with a maximum-likelihood classifier. Our claim can be accounted for by the following intuitively plausible arguments:

(1) The criterion function in LDA maximizes the ratio of the between-class scatter to within-class scatter. The dimensionality reduction transformation will result in a linear subspace with optimal linear discriminant functions, but not necessarily with optimal quadratic discriminant functions as used in the maximum likelihood classifier. This means that if the data are not linearly separable, LDA is not warranted to outperform PCA with the latter classifier.

(2) Fig.1 shows an example of a linearly nonseparable case with two classes of sample distributions. Clearly the classification decision will be different for the unknown object with the maximum-likelihood classifier versus the nearest-neighbor classifier.

In the following sections, we first review the representations of image data in PCA and LDA subspaces, along with their optimization criteria. We then review the principles

This work is supported by the NIH under the SBIR Grants 2R44 HD34329-02 and 2R44 CA76896-02. The authors would like to thank Dr. Janice L. Smith of DynaGene Laboratories for providing the chromosome images used in this study.

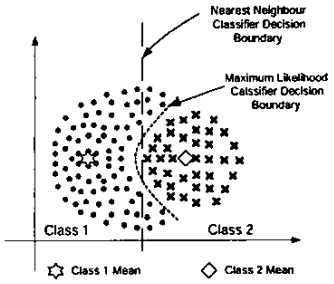

Fig. 1. Different decision boundaries of the nearest-neighbor and maximum-likelihood classifiers .

of the nearest neighbor and maximum likelihood classifiers. This is followed by our experimental results on a large data set of human chromosomes, to substantiate the claim in this paper.

2. PCA AND LDA SUBSPACES

2.1. The PCA Subspace

Suppose $\mathbf{x}$ is a zero-mean N-dimensional random vector representation of an input image, and $\mathbf{X}$ is a N-by-M data matrix whose columns are made of sample image vectors $\{\mathbf{x}_i\}, i = 1, ..., M$. PCA aims at projecting the data into a subspace whose basis vectors correspond to the maximum-variance directions in the original N-dimensional image space.

If we denote by $\mathbf{W}$ the linear transformation that projects the data from the original image space into a Q-dimensional subspace, the projected data vector in the subspace is $\mathbf{y} = \mathbf{W}\mathbf{x}$. The PCA transformation corresponds to the matrix $\mathbf{W}$ that is constructed so that its column vectors are the eigenvectors $\mathbf{w}_j$ of the covariance matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^\mathbf{T}$ arranged in the order of decreasing magnitude of the corresponding eigenvalues λ_i

$$\mathbf{C}\mathbf{w}_j = \lambda_i \mathbf{w}_j, \tag{1}$$

where λ_j is the eigenvalue associated with the eigenvector $\mathbf{w}_j, j = 1, ..., Q$, Q equals the smaller of N and M.

Because of the maximum-variance projection, PCA provides an optimal transformation for reconstructing the original image data from a lower-dimensional space with least MSE [10].

2.2. The LDA Subspace

Unlike PCA, LDA seeks a linear subspace that best discriminates among object classes rather than best resembles the original data. Specifically, LDA selects the transformation matrix $\mathbf{W}$ in such a way that the ratio of the between-class scatter and within-class scatter is maximized.

Suppose we define the between-class scatter matrix as

$$S_b = \sum_{i=1}^{c} n_i(\mu_i - \mu_x)(\mu_i - \mu_x)^T \tag{2}$$

and the within-class scatter matrix as

$$S_w = \sum_{i=1}^{c}\sum_{j=1}^{n_j} (X_j - \mu_i)(X_j - \mu_i)^T \tag{3}$$

where n_i is the number of samples in class i, c is the number of object classes, μ_i is the mean of class i, and μ_x is the mean of all classes. It has been proven in [7] that if S_w is nonsingular, the determinant ratio $\frac{|S_b|}{|S_w|}$ is maximized when the column vectors of the transformation matrix $\mathbf{W}$ are the generalized eigenvectors of $S_w^{-1}S_b$ corresponding to

$$S_b\mathbf{w}_i = \lambda_i S_w \mathbf{w}_i, i = 1, ..., m, \tag{4}$$

where $\lambda_i, i = 1, ..., m$ are the generalized eigenvalues, and m is the number of nonzero generalized eigenvectors, $m \leq c-1$. It has been suggested in [5] that at least $N+C$ samples are needed to ensure that S_w does not become singular.

3. NEAREST-NEIGHBOR AND MAXIMUM-LIKELIHOOD CLASSIFIERS

3.1. The Nearest-Neighbor Classifier

In pattern recognition, the nearest-neighbor and maximum-likelihood classifiers are among the most popular designs. The nearest-neighbor classifier is popular because it is a deterministic and non-parametric method that requires no knowledge of underlying probability density distributions of the object features, and it is based on a straightforward decision rule. For any test object feature vector $\mathbf{y}$, if we use ψ_i to represent the trained prototype of each class i, and $D(\mathbf{y}, \psi_i) = |\mathbf{y} - \psi_i|$ to denote the distance between the test object and each prototype, the nearest neighbor decision rule simply states that $\mathbf{y}$ is assigned to the class j if it is the nearest neighbor among all the prototypes, i.e.,

$$\mathbf{y} \rightarrow j, \quad if \quad D(\mathbf{y}, \psi_j) < D(\mathbf{y}, \psi_i), i = 1, ..., c, i \neq j. \tag{5}$$

The upper bound of the nearest-neighbor classifier error is known to be less than twice the Bayes error [10].

3.2. The Maximum-Likelihood Classifier

In contrast, the maximum-likelihood classifier is a probabilistic and parametric method. It is designed based on Bayes decision theory. This particular classifier commonly assumes that the object features have a multivariate normal distribution. During the classifier training phase, one characterizes the classes by first computing the Q-element mean vector μ_i for each class i, whose element is given by

$$\mu_{i,j} = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{i,j,k} \quad (6)$$

where Q is the number of features in use, n_i is the number of class i objects in the training set and $x_{i,j,k}$ is the measurement value for the jth feature of the kth object in class i. One then computes the Q-by-Q covariance matrix for each class. The covariance between the j_1th and j_2th feature is given, for class i, by

$$C_{i,j_1,j_2} = \frac{1}{n_i - 1}(\sum_{k=1}^{n_i} x_{i,j_1,k}x_{i,j_2,k} - n_i\mu_{i,j_1}\mu_{i,j_2}) \quad (7)$$

Since the object features are assumed to be multivariate normal distributed, the joint probability density function (*pdf*) for class i is given by

$$p_i(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^M|\mathbf{C}_i|}} exp[-\frac{1}{2}(\mathbf{x} - \mu_i)^T \mathbf{C}_i{}^{-1}(\mathbf{x} - \mu_i)] \quad (8)$$

If $\mathbf{x}$ is the feature vector of an unknown object, the likelihood of it belonging to class i is given by Bayes rule

$$L_i = P_i\frac{p_i(\mathbf{x})}{p(\mathbf{x})} \quad where \quad p(\mathbf{x}) = \sum_{i=1}^{c} p_i(\mathbf{x}) \quad (9)$$

where P_i is the a *priori* probability of occurrence of class i, and c is the number of classes. Under the maximum likelihood criteria, this unknown object is assigned into the class for which the likelihood is the highest.

4. EXPERIMENTAL RESULTS

In this section, we present our experimental results on a data set of human chromosomes. Automated chromosome classification has been an outstanding object recognition problem for decades [8]. The goal is to automate the laborious and expensive visual recognition of the chromosome images. Typically, the procedure starts from a metaphase cell image acquired through microscope imaging, to perform individual chromosome segmentation, rotation and straightening, feature measurement, and classification in sequence. The result is a so-called karyotype image in which all chromosomes in a cell are classified and copied onto the labelled slots corresponding to their 24 classes. Fig.4 shows images of a G-band metaphase cell and a karyotype of all the chromosomes in that cell.

The data set used in our experiments contains 15781 chromosomes from 342 normal cells. These images were collected from a data archive at the DynaGene Cytogenetics Laboratory in Houston, Texas, and are fairly representative of routine sample quality. All the chromosomes were segmented, rotated, and straightened using commercially available software [9]. They were then all resized to 10-by-100

Fig. 2. (a) A G-band metaphase cell image and (b) A karyotype of all the chromosomes in (a).

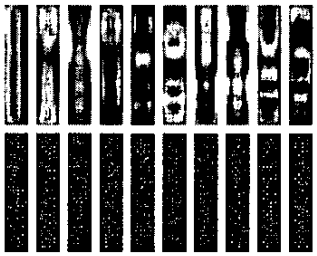

Fig. 3. The images of first 10 eigenvectors from the PCA (top) and LDA (bottom) subspaces.

image arrays by cubic spline interpolation, and normalized to have zero-mean and unit-variance, before being column-stacked to form 1-D data vectors as required by the PCA and LDA approaches. We used the common cross-validation method for classifier testing. Under this method the whole data set is split into two subsets, A and B, each containing approximately half of the data. The test results are averaged over using subset A for training and subset B for testing, and vice versa. During the training phase, the PCA and LDA subspaces described in section 2.1 and 2.2 are generated, and the projections of the training chromosome images onto the nonzero eigenvectors of each subspace are computed and gathered as the training feature vectors. For the testing phase, the same projections of the test chromosome images are carried out to obtain the test feature vectors. Fig.3 shows the images of first ten eigenvectors from the PCA (top) and LDA (bottom) subspaces.

Two series of experiments were conducted to evaluate the performances of both subspaces for chromosomes recognition with the two different classifiers. Fig.4 (a) and (b) plot the recognition curves of the PCA and LDA subspaces, with the NN (nearest-neighbor) and ML (maximum likelihood) classifiers respectively. The LDA dimensionality is limited by the number of chromosome classes, while the PCA dimensionality is limited by the lesser of image size

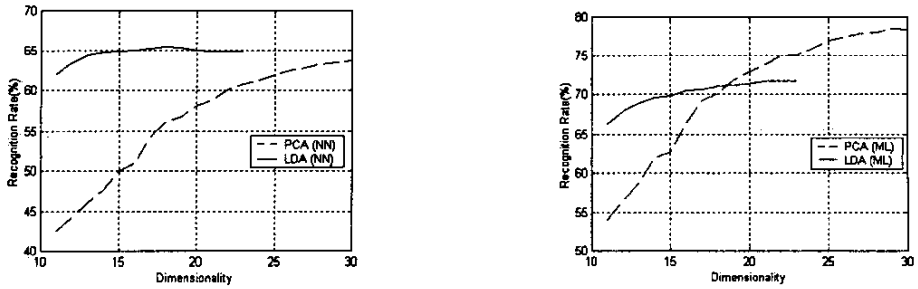

Fig. 4. (a) Performance of PCA and LDA subspaces with the nearest neighbor classifier. (b) Performance of PCA and LDA subspaces with the maximum likelihood classifier.

and number of training samples used. Without loss of generality, we only plot the curves within the indicated dimensionality interval (10~30) including the maximum LDA dimensionality (23), which suffices to establish our claim in this paper. These curves clearly show that with a relatively large data set and the same dimensionality, even though LDA performs consistently better than PCA with the NN classifier, it is not the case with the ML classifier, especially when the dimensionality goes higher. Table 1 shows the recognition results of both subspaces with the NN and ML classifiers, at the maximum LDA dimensionality. The confidence intervals are given to indicate the statistical significance of the results [10].

Table 1. Test results of recognition rate and confidence interval at the maximum LDA dimensionality.

Recognition Rate(%)	PCA	LDA
ML Classifier (%)	75.14	71.75
95% Conf. Int.(%)	-0.96 ~ +0.94	-1.00 ~ +0.98
NN Classifier (%)	61.23	64.91
95% Conf. Int. (%)	-1.08 ~ +1.07	1.06 ~ +1.04

5. CONCLUSIONS

PCA and LDA are widely used subspace methods for appearance-based object recognition. Since LDA optimizes over class separation, while PCA optimizes over representation with least MSE, the general perception is that LDA should always outperform PCA, and most previous studies support that with results obtained with the nearest neighbor classifier. In this paper, we show by both intuitively plausible arguments and experimental evidence that assuming the superiority of LDA is not warranted, especially with the maximum likelihood classifier, which outperforms the nearest neighbor classifier in both subspaces in the tests. Our conjecture is that perhaps only when the underlying object classes are linearly separable, would LDA become truly superior to other known subspaces of equal dimensionality, since under that scenario both classifiers will yield the same results.

6. REFERENCES

[1] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 12, no. 1, pp. 103–108, Jan 1990.

[2] Matthew Turk and Alex Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1994.

[3] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 19, no. 7, pp. 711–720, July 1997.

[4] Kamran Etemad and Rama Chellappa, "Discriminant analysis for recognition of human face images," *J. Optical Society of America*, vol. 14, pp. 1724–1733, August 1997.

[5] Aleix M.Martinez and Avinash C.Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 23, no. 2, pp. 228–233, Feburary 2001.

[6] Daniel L.Swets and John Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 18, no. 8, pp. 831–836, August 1996.

[7] R.A.Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.

[8] Jim Graham and Jim Piper, "Automatic Karyotype Analysis," *Methods in Molecular Biology*, 29:141-185, 1994.

[9] *MacKtype*: a software product of Applied Imaging Corporation, 2000.

[10] R.O. Duda, P.E. Hart, and D.G. Stork, "Pattern Classification," John Wiley & Sons, 2001.