Statistical Dimension Reduction Methods For Appearance-Based Face Recognition

by

Yangrong Ling

(Under the direction of Suchendra Bhandarkar)

Abstract

Two novel moment-based methods which are insensitive to large variation in lighting direction and facial expression are developed for appearance-based face recognition using dimension reduction methods in statistics. The two methods are based on Sliced Inverse Regression (SIR) (Li, 1991) and Sliced Average Variance Estimate (SAVE) (Cook and Weisberg, 1991) and termed as the Sirface method and the Saveface method, respectively. The Sirface method estimates the mean difference subspace while the Saveface method estimates the mean and covariance difference subspace. They produce well-separated classes in a low-dimensional subspace, even under severe variation in lighting and facial expression. In the subspace sense, the Sirface is equivalent to the Fisherface (Belhumeur *et al.*, 1997) and the Saveface is even more comprehensive. Since both methods produce the "optimal" (smallest) image subspaces, they can lower both the error rate and the computational expense.

Index words: Dimension-reduction, Subspaces, Sliced Inverse Regression (SIR), Sliced Average Variance Estimation (SAVE), Asymptotic Distribution, Permutation Tests, Appearance-based Vision, Face Recognition, Illumination Invariance, Fisher's Linear Discriminant

Statistical Dimension Reduction Methods For Appearance-Based Face Recognition

by

Yangrong Ling

B.S., Changchun University of Earth Science, China, 1994

M.S., Chinese Academy of Sciences, China, 1997

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2003

Statistical Dimension Reduction Methods For Appearance-Based Face Recognition

by

Yangrong Ling

Approved:

Major Professor: Suchendra Bhandarkar

Committee: Xiangrong Yin
Hamid Arabnia

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2003

DEDICATION

To my wife Qiqi and my son Evan

ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Suchendra Bhandarkar for his guidance in the research and preparation of this thesis. Special recognition is also given to Dr. Xiangrong Yin for his invaluable contributions to the research. Sincere thanks as well go to Hamid Arabnia for his help in the early stages of my work in Computer Science and serving on his thesis committee.

I sincerely appreciate my wife, Qiqi and my son Evan for their support and patience. For their love, encouragement, and endurance, I am profoundly grateful.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Background

During the past several years, numerous algorithms have been proposed for face recognition. While much progress has been made toward recognizing faces under small variations in lighting, facial expression and pose, reliable techniques for recognition under more extreme variations have proven elusive (Belhumeur *et al.*, 1997).

In this study, two face recognition algorithms which are insensitive to large variation in lighting direction and facial expression are developed using the dimension reduction methods in statistics. Note that lighting variability includes not only the light source intensity, but also direction and number of light sources.

Our approach to face recognition exploits two observations:

1) All of the images of a Lambertian surface, taken from a fixed viewpoint, but under varying illumination, lie in a 3D linear subspace of the high-dimensional image space (Belhumeur *et al.*, 1997).

2) Because of regions of shadowing, specularities, and facial expressions, the above observation does not exactly hold. In practice, certain regions of the face may exhibit deviation from the linear subspace, and, consequently, are less reliable for the purpose of recognition (Belhumeur *et al.*, 1997).

We make use of these observations by determining a linear projection of the input face images from the high-dimensional image space to a significantly lower dimensional

feature space which is insensitive both to variation in lighting direction and facial expression. Thus, optimal dimension reduction techniques are critical in such problems. We therefore import the dimension reduction concepts originally developed in statistics to the problem of face recognition. The two methods that we have developed are based on sliced inverse regression (SIR) (Li, 1991) and sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) and termed as the "Sirface" method and the "Saveface" method, respectively. Both methods produce well separated classes in a low-dimensional subspace, even under severe variation in lighting and facial expression. The subspace computed by the Sirface method is equivalent to the one obtained from the "Fisherface" method (Belhumeur *et al.*, 1997); and the subspace computed by the "Savefaces" method is equivalent to the one obtained by Quadratic Discriminant Analysis (QDA) which is a classical technique in the statistical area of classification and discriminant analysis. Both the Sirface method and the Saveface methods produce the "optimal"' (smallest dimension) feature subspace and result in a lower error rate and also reduced computational expense.

Organization of the thesis

The next chapter consists of a literature review of feature extraction and discriminant analysis techniques in computer vision. The two traditional methods for feature extraction and discriminant analysis; i.e., principal component analysis (PCA) and linear discriminant analysis (LDA) will be discussed in detail. Chapters 3 and 4 are manuscripts describing with the Sirface method and the Saveface method respectively. The Fisherface method and the Sirface method are compared in Chapter 3 and the Sirface

method and the Saveface method are compared in Chapter 4. Chapter 5 summarizes the work and concludes the thesis.

References

P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class-specific linear projection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 711-720.

R.D. Cook, and S. Weisberg, Discussion of Li (1991). *Journal of the American Statistical Association*, Vol. 86, pp. 328-332, 1991.

K.C. Li, Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association*, Vol. 86, pp. 316-342, 1991.

CHAPTER 2

RELATED WORK

Feature extraction and discriminant analysis in computer vision

Object recognition has long been a primary goal of computer vision, but it has turned out to be a very difficult endeavor (Turk, 2001). The novel techniques for data dimension reduction and discriminant analysis developed in our study have received some attention from statisticians but their application to problems in computer vision and pattern recognition have not been thoroughly investigated. Two practical problems in computer vision that could potentially benefit from these techniques include face recognition and hand gesture recognition. These problems are not only of academic interest within the research community in computer vision and pattern recognition but have significant practical applications as well. For example, the recognition of the human face under varying illumination conditions and camera viewpoints is not only a complex computer vision problem of considerable research interest but one with significant bearing on several application domains such as biometrics, multimedia, forensics, law enforcement, visual surveillance and content-based retrieval (Hallinan, *et al*., 1999). Similarly, vision-based hand gesture recognition has been used extensively in computer game navigation (Freeman, 1998), TV remote control (Freeman and Weissman, 1995), American Sign Language recognition (Starner *et al*., 1998), virtual navigation (Kadobayashi *et al*., 1998) and human-computer interaction (Pavlovic *et al*., 1997). Numerous algorithms have been reported in the recent research literature for face

recognition and gesture recognition; see surveys (Chellappa *et al*., 1995, Samil and Iyengar, 1992, Pavlovic *et al*., 1997).

There are two broad paradigms to face recognition and hand gesture recognition: *model-based* and *appearance-based*. Model-based approaches approximate the space of all possible face and hand gesture instances by means of a parametric model such as deformable templates (Brunelli and Poggio, 1993, Hallinan, 1995, Yuille and Hallinan, 1992) or by a collection of geometric features based on distance and angular measurements (Goldstein *et al*., 1971, Harmon *et al*., 1981, Harmon *et al*., 1978, Kaufman and Breeding, 1976, Wiscott *et al*., 1997). However, the biophysical complexity of the human anatomy, makes the formulation of an explicit and comprehensive parametric model a very challenging task. Consequently, most model-based approaches make certain simplifying assumptions about illumination, occlusion, surface rigidity, surface reflectance properties, etc. in the interest of keeping the model analytically tractable by limiting the number of parameters in the model. This limits the application of these models in many real-world situations where the underlying assumptions are often violated. Despite their economy of representation and invariance to illumination and viewpoint, model-based techniques are typically sensitive to errors in the feature extraction and measurement process (Cox *et al*., 1996).

There has been a growing interest in using appearance-based methods for human face and hand gesture recognition. This class of methods treats the input images as patterns or vectors in high-dimensional image space. The dimensionality of the image space is determined by the size of the input images i.e., if the images are of size $N \times N$ then the dimensionality of the image space is N^2. A key observation in appearance-based

methods is that the images of feasible faces or hand gestures lie in a complex low-dimensional subspace or eigenspace of the corresponding image space (Yang *et al*., 2000). Thus, if the input images were projected onto this low-dimensional subspace, then the features important for recognition could be captured while avoiding the "curse of dimensionality". A primary advantage of appearance-based methods is that it is not necessary to create explicit representations or models since the model is implicitly defined by the selection of sample images of the object.

The *Eigenface* approach developed by Turk and Pentland (Turk and Pentland, 1991a, 1991b) is based on the linear projection of the image space onto a low-dimensional eigenspace using Principal Component Analysis (PCA) for dimensionality reduction. The underlying assumption is that the set of images can be modeled by a multi-variate Gaussian distribution in the high-dimensional image space which can then be approximated by a linear combination of a smaller set of eigenvectors (termed as eigenfaces) which are derived by diagonalizing the covariance matrix of the multi-variate Gaussian distribution (Murase and Nayar, 1995). The Eigenface approach shows that a large variation in facial appearance can be modeled by a low-dimensional linear approximation (Kirby and Sirovich, 1990). However, since PCA maximizes the total scatter across all image classes (i.e., all images of all faces) it retains unwanted variations due to lighting and facial expression. Thus, while PCA projections are optimal for reconstruction from a low-dimensional basis, they are not optimal from a class discriminatory standpoint. The PCA projections are also sensitive to the training images used to generate the eigenfaces. The features derived from the PCA are also referred to as

the most expressive features (MEFs) in the context of content-based image retrieval (Hwang and Weng, 2000, Swets and Weng, 1996, Swets and Weng, 1999).

Belhumuer *et al.* (1997) have shown that the Fisher's Linear Discriminant Analysis (LDA) (Fisher, 1936) can produce well separated classes in a low-dimensional linear subspace even with significant variation in lighting and facial expression. Since LDA attempts to maximize the ratio of between-class scatter to within-class scatter, it was shown to outperform the PCA in terms of class discriminatory ability (Belhumeur *et al*., 1997). Martinez and Kak (2001), however, have cautioned that LDA outperforms PCA only when the training data set is large and representative enough. For small training data sets PCA was shown to outperform LDA and exhibit lower sensitivity to the composition of the training data set compared to the latter (Martinez and Kak, 2001).

Swets and Weng (1996, 1999) have combined optimal linear projection techniques based on PCA and LDA with a tree structure based on quasi-Voronoi space tessellation to achieve logarithmic retrieval complexity for content-based access to image databases. The most expressive features (MEFs) and the most discriminative features (MDFs) are computed at each level in the tree using PCA and LDA respectively. The resulting tree structure is shown to result in a hierarchical discriminant analysis (HDA) of the input high-dimensional space. In more recent work, Hwang and Weng (2000) generalize the idea of hierarchical discriminant analysis to that of hierarchical discriminant regression in order to unify classification and regression problems. The technique entails hierarchical clustering in both, input space and output space to generate a hierarchical discriminant regression (HDR) tree which can be used for coarse-to-fine classification. A sample-size-dependent negative log-likelihood (NLL) based distance

measure is introduced to deal with small-sample applications, large-sample applications and unbalanced sample applications. The concepts of HDA and HDR have also been applied to the problem of appearance-based hand gesture recognition from intensity image sequences where the gestures are limited to the American Sign Language (Cui and Weng, 1999, 2000). A related appearance-based approach to recognition of human motion uses the concept of motion history images (MHIs) and motion energy images (MEIs) to generate temporal view-based templates of the underlying motion (Bobick and Davis, 2001). A recognition scheme based on invariant moments is used to match the templates to stored instances of views of known actions. However, no attempt is made to reduce the dimensionality of the templates using projection methods.

In general, linear projection techniques such as LDA and PCA tend to perform poorly when cast shadows need to be taken into account or when the underlying surface deviates from the Lambertian assumption or has a non-constant albedo. Belhumeur and Kriegman (1998) have shown that under the assumption of a Lambertian surface with no shadowing, the set of face images illuminated by a single point source constitutes a 3-D linear subspace in the high-dimensional image space. But when the effect of cast shadows are taken into account, the set of face images under the Lambertian surface assumption and illuminated with multiple point light sources, constitute a convex polyhedral cone in high-dimensional image space termed as the illumination cone (Georghiades *et al*., 1998). If the shadowing effects are small, the illumination cone is flat enough to be approximated by a linear subspace (Georghiades *et al*., 2000). The illumination cone method has been shown to be more robust to pose variation and variation in facial

expression when compared to linear projection methods (Georghiades *et al*., 2000) since the former is inherently generative.

Neural networks have also been used successfully for face and gesture recognition within the appearance-based recognition paradigm (Li and Lu, 1999, Rowley *et al*., 1998a, 1998b). Neural networks have been successfully applied to the problem of frontal face detection (Rowley *et al*., 1998a) and also of face detection under unknown rotation (Rowley, 1998b). However, it is not clear whether neural networks can handle face or gesture recognition with a large number of degrees of freedom. Both, the training process and the recognition process become computationally far more complex because the size of the training set and the net's set of responses grow exponentially with the number of degrees of freedom.

Hidden Markov Models (HMMs) have also shown success, especially in temporal gesture recognition (Starner *et al*., 1998) and speech recognition (Rabiner, 1989). The HMM is a doubly stochastic process consisting of a probabilistic network with hidden and observable states. The hidden states drive the model dynamics and the probabilistic transitions between the hidden states are governed by a state transition matrix. An HMM is characterized by a state transition matrix, the probabilities of observed states and the initial state distribution. The training process entails the association of a distinct HMM with each discernable gesture and involves updating the parameters of the HMM so that chosen HMM best describes the spatio-temporal characteristics of the chosen gesture. The training is usually achieved by optimizing a maximum likelihood measure defined over a set of training examples for a specific gesture associated with the HMM. For first-order HMMs, efficient training algorithms based on dynamic programming

(termed as the Viterbi algorithm) can be designed (Rabiner, 1989). Complex gestures, however, cannot be adequately modeled by a first-order HMMs and entail higher-order HMMs which do not share the computational efficiency of the first-order HMM. Another potential drawback of the HMM is that the distributions of the observed states is typically modeled as a mixture of Gaussians (MoGs) in the interest of computational efficiency of the training procedure. Relaxation of the MoG assumption typically renders the training procedure computationally overwhelming. Also, in the original HMM, the probabilities of observed states are assumed to be stationary (i.e., time invariant), an assumption which may hold over a short time duration but not over the entire temporal interval characterizing the gesture. Nonstationary HMMs have been used for speech recognition but have found limited application in gesture recognition.

Support Vector Machines (SVMs) (Vapnik, 1998) whose foundations stem from statistical learning theory have been used with some success in pattern recognition in general and face recognition in particular. Intuitively, given a set of points (representing features) in high-dimensional space derived from two classes, a linear SVM determines a hyperplane leaving the largest possible fraction of points of a class on the same side of the hyperplane while maximizing the distance of either class from the hyperplane. This optimal separating hyperplane (OSH) is determined by a relatively small subset of points from the two classes termed as *support vectors*. The support vectors span a subspace of the original high-dimensional space and completely characterize the OSH. The support vectors, thus, condense all the information contained in the training set needed to classify the new data points. Computation of the OSH and the support vectors entails solving a constrained optimization problem using the method of Lagrangian multipliers.

Linear SVMs are best suited for cases where the underlying data set is linearly separable (Vapnik, 1998). Linear SVMs can be extended to non-linear SVMs where the notion of an OSH is generalized to that of an optimal separating hypersurface to handle cases that are not linearly separable (Vapnik, 1998). Non-linear SVMs however, come at a significant computational price since the training process needed to compute the optimal separating hypersurface is much more complex. SVMs in their canonical formulation are designed for binary classification. For more general *n*-ary classification (i.e., classification into $n > 2$ classes), an OSH needs to be computed for every pair of classes. Thus the space and time complexity of the SVM scales quadratically with *n*. An alternative would be to use a tournament scheme wherein the *n* OSHs separating each of the *n* classes from the remaining *n* -1 classes are computed. The test pattern is classified relative to the *n* OSHs and the final classification is done based on the outcome of the *n* classifications. Although linear in space and time complexity the tournament scheme has been reported to yield ambiguous classification (Cortes and Vapnik, 1995). SVMs have been applied to face recognition in conjunction with elastic graph matching (Tefas *et al*., 2001) and to 3D object recognition (Pontil and Verri, 1998).

Keren *et al*. (2001) describe a technique termed as *antifaces* for the detection of faces under a large class of linear transformations. The detection problem is solved by sequentially applying simple linear filters (detectors) which are designed to yield small results on the facial images and large results on random images. The detectors are designed such that their results are statistically uncorrelated resulting in a false alarm rate that diminishes exponentially with an increasing number of detectors. Although there is no formal learning procedure involved (as is common with other appearance-based

methods), computing the set of optimal detectors for large-size images is computationally complex. Moreover, the technique relies on a priori assumptions about the statistical distribution of the gray levels of random images which is key to circumventing the learning procedure and may not hold in certain cases. Moreover, the technique of antifaces is more suited for detection of faces rather than recognition of faces (Keren, 2001).

There is one important point worth noting about linear projection techniques such as PCA and LDA; there is no formal method for determining the minimum dimension of the projected subspace for optimal discriminant analysis or classification. In contrast, the techniques for feature extraction and discriminant analysis to be investigated are accompanied by a formal testing procedure to determine the minimum dimension of the projected subspace for optimal classification. Moreover, the proposed techniques could also be used as a preprocessor for some of the classification techniques described above such as SVMs, neural networks, HDR and HDA.

Two traditional methods for feature extraction and discriminant analysis: PCA and LDA

Let $\boldsymbol{X}$ be the $p \times 1$ input vector, typically referred to as the predictor vector in statistical regression/classification and $\boldsymbol{Y}$ be the output, typically indicative of class. $\boldsymbol{Y}$ is referred to as the class indicator or categorical variable in statistical regression/classification. Let $\boldsymbol{Y} = 1, \ldots, c$.

I. Principal component analysis (PCA)

Let $\mathbf{\Sigma}_X = Var(\boldsymbol{X})$ be the covariance matrix of the predictor vector $\boldsymbol{X}$. Let singular value decomposition be used to get the q eigenvectors corresponding to the q largest eigenvalues of $\mathbf{\Sigma}_X$ i.e. $\lambda_1, \ldots, \lambda_q$, where the eigenvalues are ordered in descending order of magnitude. These q eigenvectors form a $q < p$ dimensional subspace. Typically, the rest of the p-q eigenvalues are close to zero. Further analysis is carried out in the reduced dimensional subspace spanned by $\nu_1^T\boldsymbol{X}, \ldots, \nu_q^T\boldsymbol{X}$. PCA ignores the existence of $\boldsymbol{Y}$ (the categorical variable). PCA is also called the Karhunen-Loeve projection (Loeve, 1955). The following steps summarize the recognition process using PCA:

1. Create Eigenspace

1.a. Create the training data matrix ($\boldsymbol{X}$): Each of the training images is stored in a vector of size p

$$x^i = \left(x_1^i, \cdots, x_p^i\right)^T,\ 1 \le i \le n \ (n \text{ is the number of training images})$$

The training images are then combined into a data matrix of size $p \times n$.

$$X = \left(x^1, x^2, \cdots, x^n\right)$$

1.b. Compute the overall mean (μ): The overall mean image is a column vector such that each entry is the mean of all the corresponding pixels of the training images.

$$\mu = (\mu_1, \mu_2, \cdots, \mu_p)^T, \text{ where } \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^i,\ 1 \le j \le p$$

1.c. Create the centered data matrix ($\widetilde{X}$): Each of the training images must be centered. Subtracting the mean image from each of the training images centers the training images.

$$\widetilde{x}^i = x^i - \mu,\ 1 \le i \le n$$

Once the training images are centered, they are combined into a centered data matrix of size $p \times n$.

$$\widetilde{X} = \left(\widetilde{x}^1, \widetilde{x}^2, \cdots, \widetilde{x}^n\right)$$

1.d. Create the covariance matrix (Σ): The data matrix is multiplied by its transpose to create a covariance matrix.

$$\Sigma = \widetilde{X}\widetilde{X}^T$$

1.e. Compute the eigenvalues and eigenvectors of the covariance matrix: The eigenvalues $(\lambda_i, 1 \le i \le p)$ and corresponding eigenvectors $(v_i, 1 \le i \le p)$ are computed for the covariance matrix.

$$\Sigma v_i = \lambda_i v_i$$

Order the eigenvalues λi's from high to low. Keep only the eigenvectors associated with the non-zero eigenvalues.

2. Project the training images

Each of the centered training images $\widetilde{x}^i$ is projected onto the eigenspace spanned by the retained eigenvectors (V). To project an image onto the eigenspace, calculate the dot product of the image with each of the ordered eigenvectors.

$$\hat{x}^i = V^T \widetilde{x}^i$$

Therefore, the dot product of the image and the first eigenvector will be the first value in the new vector. The new vector of the projected image will contain as many values as the retained eigenvectors.

3. Identify test images

Each test image is first mean centered by subtracting the mean image, and is then projected into the same eigenspace**.**

$$\tilde{y}^i = y^i - \mu \text{, and}$$

$$\hat{y}^i = V^T \tilde{y}^i$$

The projected test image is compared to every projected training image and the training image that is found to be closest to the test image is used to identify or classify the training image.

II. Linear discriminant analysis (LDA)

Let μ_i and $\mathbf{\Sigma}_i$ be the mean and covariance for class *i*. Define:

$S_B = \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T$ and $S_w = \sum_{i=1}^{c} \sum_{X_k \in i} (X_k - \mu_i)(X_k - \mu_i)^T$ to be the between-class and within-class scatter matrices, respectively. We can assume that S_w is nonsingular, else use PCA to reduce its rank so that the reduced matrix is nonsingular. The goal of LDA is to find a matrix $\mathbf{\Gamma} = (\gamma_1, \ldots, \gamma_q)$ that maximizes the ratio $\frac{\gamma^T S_B \gamma}{\gamma^T S_W \gamma}$ under the constraint $\mathbf{\Gamma}^T S_w \mathbf{\Gamma} = \mathbf{I}$. When *c*=2, this technique is referred to as Fisher's linear discriminant analysis (LDA) (Fisher, 1936) whereas when *c*>2, it is called canonical covariate analysis (Mclachlan, 1992). Further analysis is carried out in the reduced subspace spanned by the columns of $\mathbf{\Gamma}^T \boldsymbol{X}$. Typically, $q < \min(c-1, p)$. The following steps summarizes the recognition process:

1. Compute the means: Compute the mean of the images in each class (μ_i) and the overall mean of all images (μ).

2. Center the images in each class: Subtract the mean of each class from the images in that class.

$$\forall x \in X_i, X_i \in X, \tilde{x} = x - \mu_i$$

3. Center the class means: Subtract the overall mean from the class means.

$$\widetilde{\mu}_i = \mu_i - \mu$$

4. Calculate the within class scatter matrix: The within class scatter matrix measures the amount of scatter between items within the same class. For the *ith* class a scatter matrix (S_i) is calculated as the sum of the covariance matrices of the centered images for that class.

$$S_i = \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T$$

where μ_i is the mean of the training images in class *i*, The within class scatter matrix (S_w) is the sum of all the scatter matrices.

$$S_w = \sum_{i=1}^{c} S_i$$

5. Calculate the between class scatter matrix: The between class scatter matrix (S_B) measures the amount of scatter between classes. It is calculated as the sum of the covariance matrices of the centered means of the classes, weighted by the number of images in each class.

$$S_B = \sum_{i=1}^{c} n_i (\mu_i - \mu)(\mu_i - \mu)^T$$

6. Solve the generalized eigenvalue problem: Solve for the generalized eigenvectors (V) and eigenvalues (Λ) of the within class and between class scatter matrices.

$$S_B V = \Lambda S_W V$$

7. Keep the first *C*-1 eigenvectors: Sort the eigenvectors by their associated eigenvalues from high to low and keep the first c-1 eigenvectors. These c-1 eigenvectors are the LDA basis vectors.

8. Project images onto eigenvectors: Project all the original (i.e. not centered) images onto the LDA basis vectors by calculating the dot product of the image with each of the LDA basis vectors.

LDA is a well-known technique in classification and discriminant analysis. The optimal situation is encountered when $\boldsymbol{X}$'s are normally distributed for each class i with equal covariance matrices i.e, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ for $i = 1, ..., c$. When it is not the case that all $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, information that is critical for classification could be lost. Hence the fact that $\boldsymbol{\Sigma}_i$ could be different for distinct classes needs to be considered.

References

P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class-specific linear projection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 711-720.

P.N. Belhumeur and D.J. Kriegman, What is the set of images of an object under all possible lighting conditions? *Intl. Jour. Computer Vision*, Vol. 28 No. 3, 1998, pp. 245-260.

A.F. Bobick and J.W. Davis, The recognition of human movement using temporal templates, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257-267, 2001.

R. Brunelli, and T. Poggio, Face recognition: features vs. templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 10, pp.1042-1053, 1993.

R. Chellappa, C. Wilson and S. Sirohey, Human and machine recognition of faces: a survey, *Proc. IEEE*, Vol. 83, No. 5, pp. 705-740, 1995.

C. Cortes and V. Vapnik, Vector support networks, *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995.

I. Cox, J. Ghosn and P. Yianilos, Feature-based face recognition using mixture distance, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 209-216, 1996.

Y. Cui and J. Weng, A learning-based prediction-and-verification segmentation scheme for hand sign image sequence, *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 21, No. 8, pp. 798-804, 1999.

Y. Cui and J. Weng, Appearance-based hand sign recognition from intensity image sequences, *Computer Vision and Image Understanding*, Vol. 78, pp. 157-176, 2000.

R.A. Fisher, The use of multiple measures in taxonomic problems, *Ann. Eugenics*, Vol. 7, 1936, pp. 179-188.

W.T. Freeman and C.D. Weissman, Television control by hand gestures, *Intl. Wkshp. Automatic Face and Gesture Recognition*, 1995.

W.T. Freeman, Computer vision for interactive computer graphics, *IEEE Computer Graphics and Appl*., vol. 18, No. 3, pp. 42-53, 1998.

A.S. Georghiades, D.J. Kriegman and P.N. Belhumeur, Illumination cones for recognition under variable lighting: faces, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998, pp. 52-58.

A.S. Georghiades, P.N. Belhumeur and D.J. Kriegman, From few to many: generative models for recognition under variable pose and illumination, *Proc. Intl. Conf. Automatic Face and Gesture Recognition*, 2000.

A. Goldstein, L. Harmon and A. Lesk, Identification of human faces, *Proc. IEEE*, Vol. 29, No. 5, pp. 748-760, 1971.

P. Hallinan, A deformable model for the recognition of human faces under arbitrary illumination, PhD thesis, Division of Applied Sciences, Harvard University, Cambridge, MA 1995.

P.L. Hallinan, G.G. Gordon, A.L Yuille, P. Giblin and D. Mumford, *Two- and Three-Dimensional Patterns of the Face*, A.K. Peters, Natick, MA, 1999.

L. Harmon, S. Kuo, P. Ramig and U. Raudkivi, Identification of human face profiles by computer, *Pattern Recognition*, Vol. 10, pp. 301-312, 1978.

L. Harmon, M. Kaun, R. Lasch, and P. Ramig, Machine identification of human faces, *Pattern Recognition*, Vol. 13, No. 2, pp. 97-110, 1981.

W.S. Hwang and J. Weng, Hierarchical discriminant regression, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, No. 1, pp. 1277-1293, 2000.

R. Kadobayashi, K. Nishimota and K. Mase, Design and evaluation of gesture Interface of an immersive walkthrough application for exploring cyberspace, *Intl. Conf. Automatic Face and Gesture Recognition*, 1998.

G. Kaufman and K. Breeding, The automatic recognition of human faces from profile silhouettes, *IEEE Trans. Systems, Man and Cybernetics*, Vol. 6, pp. 113-121, 1976.

D. Keren, M. Osadchy and C. Gotsman, Antifaces: A novel, fast method for image detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 23, No. 7, pp. 747-761, 2001.

M. Kirby and L. Sirovich, Application of the Karhunen-Loeve procedure for the chraracterization of human faces, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 12, No. 1, pp.103-108, 1990.

S. Li and J. Lu, Face recognition using the nearest feature line, *IEEE Trans. Neural Networks*, Vol. 10, No. 2, PP. 439-443, 1999.

M.M. Loeve, *Probability Theory*, NJ: Van Nostrand, 1955.

A.M. Martinez and A.C. Kak, PCA versus LDA, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 23, No. 2, pp. 228-233, 2001.

G.J. McLachlan, *Discriminant analysis and Statistical Pattern Recognition*, John Wiley, New York, 1992.

H. Murase and S.K. Nayar, Visual learning and recognition of 3D objects from appearance, *Intl. Jour. Computer Vision*, Vol. 14, No. 1, pp. 5-24, 1995.

V.I. Pavlovic, R Sharma and T.S. Huang, Visual interpretation of hand gestures for human- computer interaction; a review, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 677-695, 1997.

M. Pontil and A. Verri, Support vector machines for 3D object recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 20, No. 6, pp. 637-646, 1998.

L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, Vol. 77, pp.257-286, Feb. 1989.

H. Rowley, S. Baluja and T. Kanade, Neural network-based face detection, *IEEE Trans. Patt. Anal. Machine Intell*. Vol. 20, No. 1, pp. 23-38, 1998.

H. Rowley, S. Baluja and T. Kanade, Rotation invariant neural network-based face detection, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 38-44, 1998.

A. Samil and P. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, *Pattern Recognition*, Vol. 25, pp. 65-77, 1992.

T. Starner, J. Weaver and A. Pentland, Real-time American Sign Language Recognition using desk and wearable computer-based video, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 12, pp. 1371-1375, 1998.

D.L. Swets and J. Weng, Using discriminant eigenfeatures for image retrieval, *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 18, pp. 831-836, 1996.

D.L. Swets and J. Weng, Hierarchical discriminant analysis for image retrieval, *IEEE Pattern Anal. Mach. Intell*., Vol. 21, No. 5, pp. 386-401, 1999.

A. Tefas, C. Kotropoulos and I. Pitas, Using support vector machines to enhance the performance of elastic graph matching for frontal face authentification, *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 23, No. 7, pp. 735-746, 2001.

M. Turk, A random walk through eigenspace, *IEICE Trans. INF. & SYST.*, vol. E84-D, No. 12, 2001.

M. Turk and A. Pentland, Eigenfaces for recognition, *Jour. Cognitive Neuroscience*, vol. 3, No. 1, 1991.

M. Turk and A. Pentland, Face recognition using eigenfaces, *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 1991, pp. 586-591.

V.N. Vapnik, *Statistical Learning Theory*, John Wiley, New York, NY, 1998.

L. Wiscott, J. Fellous, N. Kruger and C. von der Malsburg, Face recognition by elastic bunch graph matching, *IEEE Trans.Pattern Anal. Mach. Intell*. Vol. 19, No. 7. pp. 775-779, 1997.

M.H. Yang, N. Ahuja and D. Kriegman, Face detection using a mixture of linear subspaces, *Proc. IEEE Intl. Conf. Face and Gesture Recognition*, 2000.

A.L. Yuille and P. Hallinan, Deformable templates, in *Active Vision* (A. Blake and A Yuille, eds), MIT Press, Cambridge, MA, 1992.

CHAPTER 3

SIRFACE vs. FISHERFACE: RECOGNITION USING CLASS SPECIFIC LINEAR PROJECTION AND FIRST ORDER STATISTICS[1]

[1] Ling, Y., Yin, X., and Bhandarkar, S., To be submitted to *Pattern Recognition*

Abstract

We develop a face recognition algorithm which is insensitive to large variation in lighting direction and facial expression using the dimension reduction methods in statistics. Taking a pattern classification approach, each pixel in an image is considered as a coordinate in a high-dimensional space. We linearly project the image into a subspace in a manner which discounts those regions of the face with large deviation, thus retains only those regions which are invariant to illumination and facial expression. Our projection method is based on Sliced Inverse Regression (SIR) (Li, 1991) and termed as the *Sirface* method. The Sirface method produces well separated classes in a low-dimensional subspace, even under severe variation in lighting and facial expression. In the subspace sense, Sirface is equivalent to the Fisherface method (Belhumeur *et al.*, 1997) but produces the optimal (i.e. with the fewest dimensions) subspace under the Fisherface projection and hence results in lower error rate and reduced computational expense.

Introduction

During the past several years, numerous algorithms have been proposed for face recognition. While much progress has been made toward recognizing faces under small variations in lighting, facial expression and pose, reliable techniques for recognition under more extreme variations have proven elusive (Belhumeur *et al.*, 1997).

In this paper, we outline a new approach for face recognition, one that is insensitive to large variations in lighting and facial expressions. Note that lighting

variability includes not only light source intensity, but also directions and number of light sources.

Our approach to face recognition exploits two observations:

1) All of the images of a Lambertian surface, taken from a fixed viewpoint, but under varying illumination, lie in a 3D linear subspace of the high-dimensional image space (Belhumeur *et al.*, 1997).

2) Because of regions of shadowing, specularities, and changes in facial expressions, the above observation does not exactly hold. In practice, certain regions of the face may exhibit deviation from the linear subspace, and, consequently, are less reliable for the purpose of recognition (Belhumeur *et al.*, 1997).

We make use of these observations by finding a linear projection of the faces from the high-dimensional image space to a significantly lower dimensional feature space which is insensitive to both variation in lighting direction and facial expression. Thus dimension reduction techniques are very useful in such problems. We then import the dimension reduction concepts originally developed in statistics to the problem of face recognition. Our method is based on sliced inverse regression (SIR Li, 1991) and for which we develop an algorithm called the *Sirface* method. The subspace computed by the Sirface method is equivalent to the one obtained from the Fisherface method (Belhumeur *et al*., 1997); hence it maximizes the ratio of between-class scatter to within-class scatter. But the Sirface method can further reduce the subspace dimension determined by the Fisherface method and result in a possible smaller dimensional subspace. This could lower both the error rate and computation expense.

While there are other methods such as correlation, the Eigenface method and linear subspace projection, the comparison among these methods along with the Fisherface method can be found in Belhumeur *et al*. (1997). The Fisherface method was shown to be superior to correlation, the Eigenface method and linear subspace projection. Thus the main point of this paper is to compare the Sirface method to the Fisherface method.

We should point out that Fisher's linear discriminant analysis (LDA) is a classical technique, especially in the areas of classification and discriminant analysis in statistics. Sliced inverse regression (SIR) (Li, 1991), on the other hand, is a more recent technique in statistical regression. The connection between regression and discriminant analysis was established recently (Kent, 1991, Cook and Yin, 2000). Details about the Fisherface method and the Sirface methods are provided in the following sections.

Methods

The face recognition problem can be simply stated: Given a set of face images labeled with the person's identity (*the learning set*) and an unlabeled set of face images from the same group of people (*the test set*), identify each person in the test images.

The standard procedure here is to use the learning set to establish some classification rules for the images and then apply these rules to classify the test set into the right image classes. Formally, let us consider a set of n sample images $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ taking values in a p-dimensional image space, and assume that each image belongs to one of c classes 1,..., c. Thus we need to establish the rules based on the p-dimensional image space and classify the $\boldsymbol{X}_i$'s to the right class. This is in fact a classification

problem. Since p is large, we would like to reduce the p dimensional image space to the smallest q dimensional image space, or more specifically find a $p \times q$ matrix $\boldsymbol{B}$ such that $\boldsymbol{B}^T\boldsymbol{X}$ is a new smaller q-dimensional subspace that retains all the classification information. This means that we will classify the images $\boldsymbol{X}_i$ into the same class regardless of whether or not we use the original p-dimensional image space or the reduced q-dimensional feature subspace. That is, the subspace spanned by the columns of $\boldsymbol{B}$ is a central discriminant subspace (Cook and Yin, 2001). In addition, using the reduced q-dimensional image space will have many advantages. For example, if $q \leq 3$, we can easily view the projected data. A reduced subspace will reduce the classification error rate and lower the computation expense. Although we have data in the original $\boldsymbol{X}$-scale, equivalently we can always transform them into an equivalent $\boldsymbol{Z}$-scale where

$$Z = \sum_X^{-\frac{1}{2}}(X - \mu_X)$$

and $\sum_X$ and μ_X are the covariance matrix and mean vector of $\boldsymbol{X}$. Here we assume that $\sum_X$ is nonsingular, otherwise we can first reduce the dimensionality of the original $\boldsymbol{X}$ using principal components analysis (PCA). The use of the $\mathbf{Z}$-scale allows easy comparison of various dimensionality reduction techniques.

In the next section, we examine the two pattern classification techniques for solving the face recognition problem. We approach this problem within the pattern classification paradigm, considering each of the pixel values in a sample image as a coordinate in a high-dimensional space (*the image space*).

The Fisherface Method

Let μ_i and $\sum_i$ be the mean and covariance for class *i*. Define

$$S_B = \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \sum_{i=1}^{c} n_i(\mu_i \mu_i^T)$$

and

$$S_w = \sum_{i=1}^{c} \sum_{Z_k \in i} (Z_k - \mu_i)(Z_k - \mu_i)^T$$

Note that in the ***Z***-scale, $\mu = 0$. We can always assume that $\boldsymbol{S}_W$ is nonsingular otherwise use PCA to reduce its rank so that the reduced dimensional subspace has nonsingular $\boldsymbol{S}_W$. Belhumeur *et al.* (1997) developed an algorithm called *Fisherface* to find a matrix $\Gamma = (\gamma_1,...,\gamma_q)$ that maximizes the following ratio:

$$\frac{\gamma^T S_B \gamma}{\gamma^T S_W \gamma}$$

under $\Gamma^T S_W \Gamma = I$. When $c = 2$, this technique is called Fisher's discriminant analysis (FDA). When $c > 2$, it is called canonical covariate analysis (Mclachlan, 1992).

The Fisherface method is a well-known technique in classification and discriminant analysis. The optimal situation is encountered when the ***X***'s are normally distributed for each class *i* with the same covariances for all classes, i.e. $\sum_i = \sum$ for $i = 1,...,c$. When not all $\sum_i = \sum$, information needed for classification could be lost, hence the difference among $\sum_i$ needs to be considered. Since the Fisherface method is based on the assumption of normal distribution, in cases where the assumption is not satisfied, it could lose important classification information.

The Sirface Method

Sliced inverse regression (SIR, Li 1991) was originally developed to reduce the data dimensionality in regression problems. Let ($\boldsymbol{Y}_i$, $\boldsymbol{X}_i$) i = 1, ..., n be a sample, where $\boldsymbol{Y}$ is a response variable and $\boldsymbol{X}$ is a predictor vector. Li (1991) considered the inverse mean of $\mathbf{E}(\boldsymbol{X}|\boldsymbol{Y})$ in the $\boldsymbol{Z}$-scale, by forming the matrix: Var($\mathbf{E}(\boldsymbol{Z}|\boldsymbol{Y})$). Singular value decomposition is used to find the minimum dimension of this matrix. We assuming that $\boldsymbol{S}_W$ is nonsingular, otherwise we can first reduce its rank to make it nonsingular using PCA as in Belhumeur *et al.* (1997). Kent (1991) mentioned that SIR is equivalent to LDA when $\boldsymbol{Y}$ is a categorical variable. Cook and Yin (2001) further developed this connection. For a slightly different matrix whose columns span the same subspace as SIR, Geisser (1977) proved a similar result.

In fact, for a categorical $\boldsymbol{Y}$, the SIR matrix is $M_{SIR} = \frac{1}{n} S_B$. For SIR, we only need to apply singular value decomposition to $\frac{1}{n} S_B$, that is, we need to find its d (the reduced dimension) non-zero eigenvalues and their corresponding eigenvectors. We call this method the *Sirface* method.

Comparison with The Fisherface Method

In the Sirface method, the subspace spanned by the d non-zero eigenvectors is the same as the subspace spanned by the c-1 eigenvectors in the Fisherface method. Hence using these d vectors does not result in loss of information. But the Fisherface method uses pre-specified m eigenvectors corresponding to the m largest eigenvalues. If $m<d$, then the Fisherface method may lose important classification information. If $m>d$, then the

Fisherface method uses redundant predictor vectors which may include noise. In the case of the Sirface method, there exists a formal method for determining the optimal dimensionality *d* of the reduced dimensional subspace based on input data. Thus the Sirfaces method is "optimal" in this sense.

The Sirface method can result in dimensionality reduction beyond that possible with the Fisherface method. In the absence of any further information, the Fisherface method is constrained to choose *c*-1 (where *c* is the number of classes) as the dimensionality of the reduced subspace. Any further reduction is done via exhaustive search. The Sirface method, on the other hand, yields a reduced dimensionality of *d* which is often less than *c*-1. While LDA, in general, is not robust to non-normal data (Krzanowski 1977), the Sirface method can help to find outliers and hence is much more robust (Cook and Yin, 2001).

Test for determining the optimal reduced dimension *d*

Let

$$M_{SIR} = \Gamma^T \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \Gamma$$

where $\mathbf{\Gamma}$ is a $p \times p$ orthogonal matrix whose columns are the eigenvectors $v_1, \cdots, v_p$ of $M_{SIR,}$ and $\mathbf{\Gamma}^{\mathrm{T}}=(\mathbf{\Gamma}_1, \mathbf{\Gamma}_0)$, $\mathbf{\Gamma}_1$ is a $p \times$ d matrix whose columns are the eigenvectors $v_1, \cdots, v_d$ of $M_{SIR,}$ $\mathbf{\Gamma}_0$ is a $p \times (p-d)$ matrix whose columns are the eigenvectors $v_{d+1}, \cdots, v_p$ of M_{SIR}. D is a $d \times d$ diagonal matrix whose elements $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of M_{SIR}. If *d* is known, we just use the estimate $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ of the sample matrix $\hat{M}_{SIR}$, and use their

corresponding eigenvectors $\hat{v}_1, \cdots, \hat{v}_d$. If *d* is unknown, an inference procedure about *d* is required. The relevant test statistic (Li, 1991) is

$$\hat{\Lambda}_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j$$

where $\hat{\lambda}_j$'s are the eigenvalues of the sample matrix $\hat{M}_{SIR}$.

Generally, the general asymptotic distribution for the SIR test statistic is a linear combination of independent chi-squared random variables each with one degree of freedom (Bura and Cook, 2001). For some special cases, such as normal predictors the test statistic has a central chi-square distribution (Li 1991, Cook 1998). The above result is a corollary of theorem 2 of Bura and Cook (2001).

Corollary 1: *If* $\boldsymbol{Z}|\boldsymbol{Y}$ *is normally distributed, then* $\hat{\Lambda}_d$ *has an asymptotic chi-squared distribution with (p-d)(c-d-1) degrees of freedom.*

Another choice is a permutation test which was first suggested by Cook and Weisberg (1991) and further developed by Cook and Yin (2001). We believe that this test is much more robust as demonstrated by Cook and Yin (2001), Yin (2000) and Yin and Cook (2002). When the predictors are normal, both the tests are in agreement; otherwise, the permutation test can recover the optimal reduced dimension but not the asymptotic test.

Let $\mathbf{U} = [\mathbf{u}_j]$ denote the $p \times p$ matrix of eigenvectors $\mathbf{u}_j$ of the kernel matrix **M**. Consider testing the hypothesis that $d \le m$ versus $d > m$. Partition $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ where $\mathbf{U}1$ is $p \times m$. The following proposition (Cook and Yin, 2001) provides a basis for constructing the permutation tests and for inference on *d*:

Proposition 1: *Let* $\boldsymbol{U}$ *be constructed as indicated previously.*

(i) If $(\boldsymbol{Y}, \boldsymbol{U}_1^T \boldsymbol{Z})$ $\perp\!\!\!\perp$ (independent of) $\boldsymbol{U}_2^T \boldsymbol{Z}$ *then* $m \geq d$.

(ii) Assume that $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}$. *Then* $m = d$ *if and only if* $(\boldsymbol{Y}, \boldsymbol{U}_1^T \boldsymbol{Z}) \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}$.

(iii) Assume that $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z} | \boldsymbol{Y}$. *If* $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$, *then* $m \geq d$.

Part (i) of this proposition may be the most important in practice because it requires no conditions. It says that if $(\boldsymbol{Y}, \boldsymbol{U}_1^T \boldsymbol{Z})$ is independent of $\boldsymbol{U}_2^T \boldsymbol{Z}$ (i.e., $f(Y, U_1^T Z, U_2^T Z) = f(Y, U_1^T Z) \cdot f(U_2^T Z)$) then we can discard the last p - m principal predictors $\boldsymbol{U}_2^T \boldsymbol{Z}$ without any loss of information on classification. We propose to test that possibility by comparing the observed test statistic $\hat{\Lambda}_m$ to its permutation distribution under the null hypothesis. This involves essentially re-computing $\hat{\Lambda}_m$ for each of a selected number of random permutations of the elements of the sample version of $\boldsymbol{U}_2^T \boldsymbol{Z}$, and then comparing the observed value to its permutation distribution to obtain the P-values v_m, $m = 0, \ldots, p\text{-}1$.

If v_m is large and non-significant then the subspace spanned by $\mathbf{U}_1$ (denoted as $S(\mathbf{U}_1)$) provides an upper bound on d. The smallest inferred upper bound is the one with the first non-significant P-value in the sequence $v_0, \ldots, v_{p-1}$.

Application of Proposition 1(i) to test the hypothesis that $d \leq m$ in practice involves the following general steps:

1. Compute the sample kernel matrix $\hat{M}$ for SIR and form the matrices of its eigenvectors $\hat{U}_1 = (\hat{u}_1, \ldots, \hat{u}_m)$ and $\hat{U}_2 = (\hat{u}_{m+1}, \ldots, \hat{u}_p)$.
2. Construct the vectors of sample principal predictors $\hat{V}_{1i} = \hat{U}_1^T \hat{z}_i$ and $\hat{V}_{2i} = \hat{U}_2^T \hat{z}_i$, $i=1, \ldots, n$.

3. Randomly permute the indices i of the $\hat{V}_{2i}$'s to obtain the permuted set $\hat{V}_{2i}^{*}$.

4. Construct the test statistic $\hat{\Lambda}_m^{*}$ based on the original data $\boldsymbol{Y}_i$, $\hat{V}_{1i}^{*}$ along with the permuted data $\hat{V}_{2i}^{*}$, i=1, … , n.

After repeating steps 3 and 4 a number of times, the P-value v_m is just the fraction of $\hat{\Lambda}_m^{*}$ that exceeds $\hat{\Lambda}_m$. Repeating steps 1–4 for m=0, … , p-1 gives the required series of P-values. We found this simple test to be quite useful in practice.

The theory behind the test computed under Proposition 1(i) guarantees only an upper bound on d so it is possible that we will end with more predictors than needed. Hence, additional assumptions are needed to eliminate that possibility. Proposition 1 (ii) requires that $\boldsymbol{U}_1^T \boldsymbol{Z}$ and $\boldsymbol{U}_2^T \boldsymbol{Z}$ be marginally independent, when $\boldsymbol{Z}$ is normally distributed. Using the test procedure sketched previously, the condition of Proposition 1(ii) allows us to infer directly about d rather than to infer about an upper bound.

A first approach to some discriminant analysis problems may involve the assumption that the conditional distribution of $\boldsymbol{Z}|\boldsymbol{Y}$ is normal. It may be reasonable in such cases to base a permutation test on the conditional independence statement in Proposition 1(iii) rather than on marginal independence $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}$ as in Proposition 1(ii). Assuming $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}|\boldsymbol{Y}$, $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$ implies $m \geq d$. Thus, failure to reject $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$ by using a Permutation test allows us to infer that the principal predictors $\boldsymbol{U}_2^T \boldsymbol{Z}$ can be discarded. The permutation algorithm sketched above can be adapted to test $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$.

The algorithm for the Sirface method

The following steps summarize the recognition process using the Sirface method:

1. Create the training data matrix (X): Each of the training images is stored in a vector

of size p

$$x^i = \left(x_1^i, \cdots, x_p^i\right)^T,\ 1 \le i \le n \ (n \text{ is the number of training images})$$

The training images are then combined into a data matrix of size $p \times n$.

$$X = \left(x^1, x^2, \cdots, x^n\right)$$

2. Compute the overall mean (μ): The overall mean image is a column vector such that each entry is the mean of all the corresponding pixels of the training images.

$$\mu = (\mu_1, \mu_2, \cdots, \mu_p)^T, \text{ where } \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^i,\ 1 \le j \le p$$

3. Create the centered data matrix ($\widetilde{X}$): Each of the training images must be centered. Subtracting the mean image from each of the training images centers the training images.

$$\widetilde{x}^i = x^i - \mu,\ 1 \le i \le n$$

Once the training images are centered, they are combined into a centered data matrix of size $p \times n$.

$$\widetilde{X} = \left(\widetilde{x}^1, \widetilde{x}^2, \cdots, \widetilde{x}^n\right)$$

4. Create the covariance matrix (Σ): The data matrix is multiplied by its transpose to create a covariance matrix.

$$\Sigma = \frac{1}{n}\widetilde{X}\widetilde{X}^T$$

5. Compute the eigenvalues and eigenvectors of the covariance matrix: The eigenvalues $(\lambda_i, 1 \le i \le p)$ (ordered from high to low) and corresponding eigenvectors $(v_i, 1 \le i \le p)$ are computed for the covariance matrix.

$$\Sigma v_i = \lambda_i v_i$$

6. Compute Z-scale images (Z): If Σ is nonsingular, then $Z = \Lambda^{-\frac{1}{2}}\Gamma^T \widetilde{X}$. Otherwise, find the number of positive eigenvalues (*k*) and

$$Z = \Lambda_k^{-\frac{1}{2}}\Gamma_k^T \widetilde{X}$$

$$\text{Where } \Lambda_k = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}, \ \Gamma_k = (v_1, \ldots v_k)$$

7. Compute the Sir matrix (M_{SIR}) in Z-scale

$$M_{SIR} = \frac{1}{n} S_B$$

Where S_B is the between scatter given by matrix $S_B = \sum_{i=1}^{c} n_i \mu_i \mu_i^T$, and μ_i is the mean for class *i* in Z-scale.

8. Test for the optimal reduced dimension of the subspace (*d*):

8.a Compute the eigenvalues $(\lambda_i, 1 \le i \le k)$ (ordered from high to low) and corresponding eigenvectors $(v_i, 1 \le i \le k)$ for the Sir matrix.

8.b Compute the test statistics

$$T = n \sum_{i=d+1}^{k} \lambda_i$$

T is asymptotic chi-squared distribution with (*k*-*d*)(*c*-*d*-1) degree of freedom.

8.c Compute p-value for *d*: If the computed p-value is greater than 5%, then we conclude that the reduced dimension is *d*.

9. Project the training images onto the reduced subspace: each of the training images in Z-scale is projected onto the reduced subspace.

10. Identify the test images: each test image is first transformed into Z-scale and then projected into the reduced subspace. The projected test image is compared to every projected training image and the training image that is found to be closest to the test image is used to identify the test image.

Experimental results

Because of the specific hypotheses that we want to test about the relative performance of the considered algorithms, many of the standard databases were inappropriate (Belhumeur *et al.*, 1997). In this study, we use a database from Yale University called The Yale Face Database B (Georghiades *et al.*, 2001). The database contains 5760 single light source images of 10 subjects, each seen under 576 viewing conditions (9 poses x 64 illumination conditions) (Figure 3.1). For every subject in a particular pose, an image with ambient (background) illumination was also captured. The images in the database were captured using a special-purpose illumination rig. This rig is fitted with 64 computer-controlled strobes. The 64 images of a subject in a particular pose were acquired at camera frame rate (30 frames/second) in about 2 seconds, so there is only small change in head pose and facial expression for those 64 (+1 ambient) images. The image with ambient illumination was captured without a strobe going off.

Five experiments are constructed using Yale Face Database B to test the Fisherface method and the Sirface method. For all experiments, classification was performed using a nearest neighbor classifier. The results tabulated in Table 1 have shown that, whereas in the case of the Fisherface method, a reduced dimensionality of c-1 (where c is the number of classes), is adequate, the Sirface method is capable of

Figure 3.1: Original (captured) images of a single individual from the Yale Face Database B, showing the variability due to illumination and pose. The images have been divided into four subsets (1 through 4 from top to bottom) according to the angle the light source direction makes with the camera axis. Every pair of columns shows the images of a particular pose (1 through 9 from left to right) (*from Georghiades et al., 2001*)

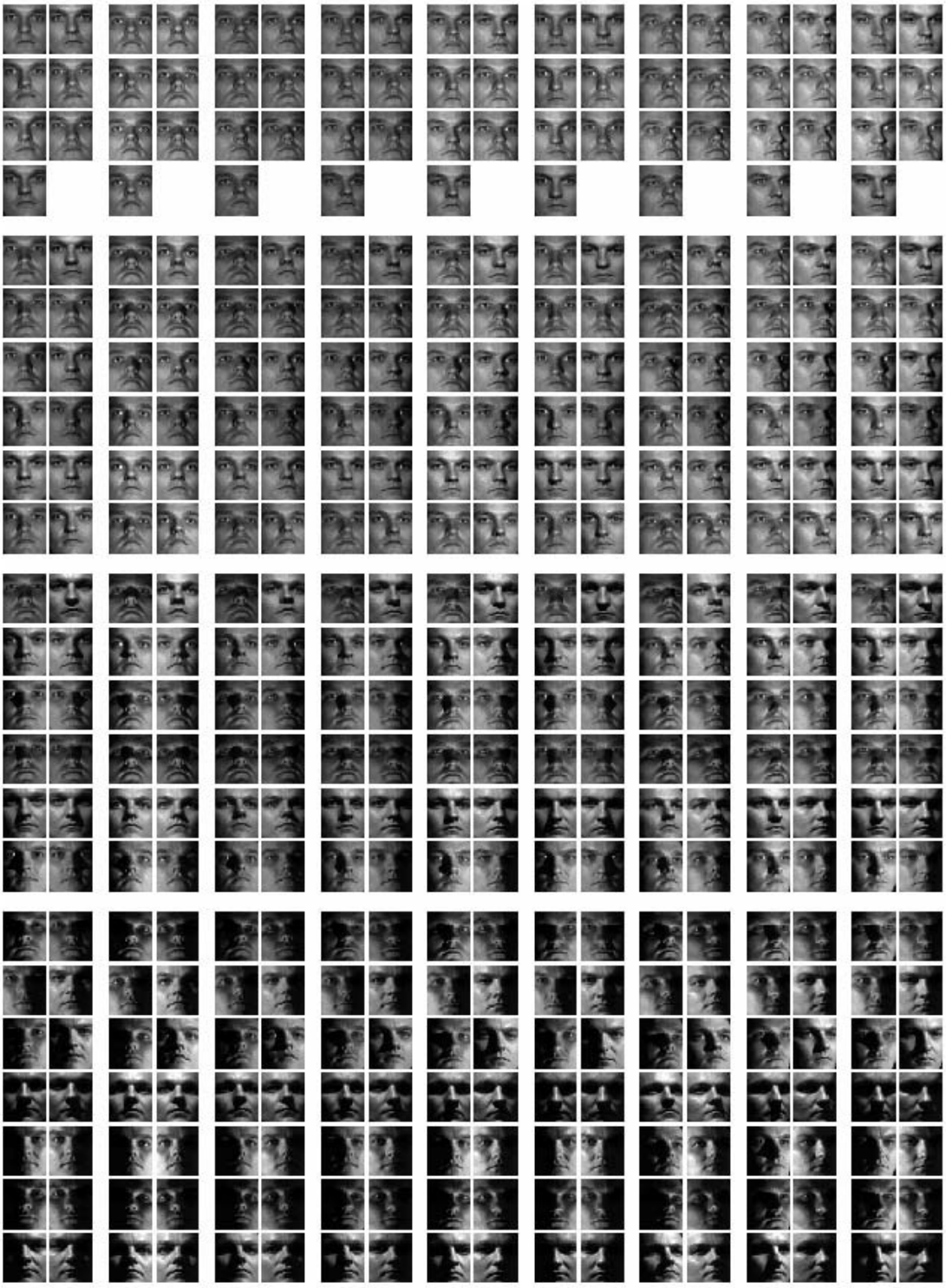


Table 3.1. Comparison of the Sirface and Fisherface methods on the Yale Face Database B

Dataset	N_{class}	$N_{image/class}$	N_{train}	N_{test}	d_{Fisher}	d_{Sir}	ε_{Fisher}	ε_{Sir}
1	20	5	100	140	19	9	5.0%	5.0%
2	20	5	100	120	19	9	7.5%	7.5%
3	20	3	60	140	19	6	20.0%	21.4%
4	10	5	50	70	9	5	4.3%	4.3%
5	10	7	70	70	9	7	1.5%	1.5%

N_{class}: number of classes, $N_{image/class}$: number of training images in each class, N_{train}: number of training images, N_{test}: number of test images, d_{Fisher}: reduced dimensionality resulting from the Fisherface method, d_{Sir}: reduced dimensionality resulting from the Sirface method, ε_{Fisher}: classification error of the Fisherface method, ε_{Sir}: classification error of the Sirface method

determining a reduced dimensionality much lower than c-1 without the need for exhaustive search and without compromising the classification accuracy. Also, the Sirface method is able to achieve over a 90% classification accuracy with as low as 5 training images per class (Table 1, Dataset 1, 2, 4, and 5). The classification accuracy of the Sirface method drops to 78.6% only when the number of training images per class is reduced to 3 (Table 1 Dataset 3). Generally, the classification accuracy decreases for both methods as the sample size of the training images decreases.

Conclusion

In this paper we proposed a novel technique for data dimensionality reduction in the context of appearance-based face recognition. This technique is based on Sliced inverse regression (SIR) and is termed as the "Sirface" method. Initial experiments on the Yale Face Database B show that the Sirface method can yield classification accuracy comparable to the well-known Fisherface method while resulting in dimensionality reduction beyond that possible with the Fisherface method. Whereas in the Fisherface method, the optimum reduced dimensionality can be determined only via exhaustive

search, the Sirface method has a formal technique for determining the optimum reduced dimensionality. Further testing of the Sirface method on a wider set of human faces is currently in progress.

References

P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class-specific linear projection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 711-720.

E. Bura and R.D. Cook, Estimating the structural dimension of regressions via parametric inverse regression, *Journal of Royal Statistical Society, B*., 63, No. 2, 20001, 393-410.

Cook, R. D., *Regression Graphics: Ideas for studying regressions through graphics,* New York: Wiley, 1998.

R.D. Cook, and X. Yin, Dimension reduction and visualization in discriminant analysis, invited paper with discussion in the *New Australia and New Zealand Journal of Statistics*, 2001, 43, No. 2, 147-199.

S. Geisser, Discrimination, allocatory and separatory, linear aspects, In: J. Van Ryzin, Ed., *Classification and Clustering*, Academic Press, New York, 301-330.

A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose, *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, No. 6, pp.643-660, 2001

R. Kadobayashi, K. Nishimota and K. Mase, Design and evaluation of gesture Interface of an immersive walkthrough application for exploring cyberspace, *Intl. Conf. Automatic Face and Gesture Recognition*, 1998.

W. J. Krzanowski, The performance of Fisher's Linear Discriminant function under non-optimal conditions, Technometrics, Vol. 19, No. 2, 1997.

J.T. Kent, Discussion of Li (1991), *Journal of the American Statistical Association*, vol.86, pp. 336-337, 1991.

K.C. Li, Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association*, Vol. 86, pp. 316-342, 1991.

G.J. McLachlan, *Discriminant analysis and Statistical Pattern Recognition*, John Wiley, New York, 1992.

X. Yin, and R.D. Cook, Dimension reduction for the conditional *k*-th moment in regression, *Journal of the Royal Statistical Society Ser B*, 64, 159-175, 2002.

CHAPTER 4

SAVEFACE: RECOGNITION USING CLASS SPECIFIC LINEAR PROJECTION AND SECOND ORDER STATISTICS[1]

[1] Ling, Y., Yin, X., and Bhandarkar, S., to be submitted to *Pattern Recognition*

Abstract

Using the dimension reduction methods in statistics, we develop a face recognition algorithm which is insensitive to large variation in lighting direction and facial expression. Taking a pattern classification approach, we consider each pixel in an image as a coordinate in a high-dimensional space and linearly project the image into a subspace in a manner which discounts those regions of the face with large deviation. Our projection method is based on Sliced Average Variance Estimation (SAVE) (Cook and Weisberg 1991) and termed as the *Saveface* method. It produces well-separated classes in a low-dimensional subspace, even under severe variation in lighting and facial expression. In the subspace sense, the Saveface method is more comprehensive than the Fisherface method (Belhumeur *et al.*, 1997) and the Sirface method. It produces the optimal (i.e., with the fewest dimension) subspace for the quadratic discriminant analysis (QDA) and hence results in a lower error rate and reduces computational expense.

Introduction

During the past several years, numerous algorithms have been proposed for face recognition. While much progress has been made toward recognizing faces under small variations in lighting, facial expression and pose, reliable techniques for recognition under more extreme variations have proven elusive (Belhumeur *et al.*, 1997).

In this paper, we outline a new approach for face recognition, one that is insensitive to large variations in lighting and facial expressions. Note that lighting variability includes not only light source intensity, but also direction and number of light sources.

Our approach to face recognition exploits two observations:

1) All of the images of a Lambertian surface, taken from a fixed viewpoint, but under varying illumination, lie in a 3D linear subspace of the high-dimensional input image space (Belhumeur *et al.*, 1997).

2) Because of regions of shadowing, specularities, and variation in facial expression, the above observation does not exactly hold. In practice, certain regions of the face may exhibit deviation from the linear subspace, and, consequently, are less reliable for the purpose of recognition (Belhumeur *et al.*, 1997).

We exploit these observations by finding a linear projection of the faces from the high-dimensional image space to a significantly lower dimensional feature space which in insensitive both to variation in lighting direction and facial expression. Thus, dimension reduction techniques are very useful in such problems. We then import the data dimension reduction concepts originally developed in statistics to the problem of face recognition. Our approach to face recognition is based on sliced average variance estimation (SAVE) (Cook and weisberg, 1991) for which we develop an algorithm called the *Saveface* method. The reduced-dimensional subspace computed by the Saveface method is equivalent to the subspace obtained via quadratic discriminant analysis (QDA) which is a classical technique, especially in the area of classification and discriminant analysis in statistics. Sliced average variance estimation, on the other hand, is a fairly new technique in statistical regression. The connection between SAVE and QDA has been established recently (Cook and Yin, 2001).

There are other methods for face recognition such as correlation, Eigenface and linear subspace projection. A comparison among these methods along with the Fisherface

method can be found in Belhumeur *et al*. (1997). The Fisherface method was shown to be superior to the correlation, Eigenface and linear subspace projection methods. An improved version of the Fisherface method is the Sirface method. Thus the main point of this paper is to compare the Saveface method to the Sirface method. Details about this method will be presented in the next several sections.

Method

The face recognition problem can be simply stated: Given a set of face images labeled with the person's identity (*the learning set*) and an unlabeled set of face images from the same group of people (*the test set*), identify each person in the test images.

The procedure here is to use the learning set to establish some classification rules for the training set images and then applies these rules to classify the test set images to the right images. Formally, let us consider a set of n sample images ($\boldsymbol{X_1}$, ..., $\boldsymbol{X_n}$) taking values in a p-dimensional image space, and assume that each image belongs to one of c classes (1, ..., c). Thus we need to establish the rules based on the p-dimensional image space and classify the $\boldsymbol{X}_i$'s in to the right class. This is in fact a classification problem. Since p is large, we'd like to reduce the p dimensional image space to smallest q-dimensional feature space. More specifically, we attempt to find a $p \times q$ matrix $\boldsymbol{B}$ so that $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}$ is a new smaller q-dimensional subspace that retains all the classification information. This means that we will classify the $\boldsymbol{X_i}$ into the same class regardless of whether or not we use the original p-dimensional image space or the reduced q-dimensional image subspace. In other words, the subspace spanned by the columns of $\boldsymbol{B}$ is a central discriminant subspace (Cook and Yin 2001). In addition, using the reduced q-

dimensional image space will have many advantages. For example, if $q \leq 3$, we can view the sampled data. A reduced dimensional subspace will reduce the error rate and reduce the computational expense. Although we have data in the original ***X***-scale space, equivalently we can always transform the data into a **Z**-scale space, where

$$Z = \sum_X^{-\frac{1}{2}} (X - \mu_x)$$

and $\mathbf{\Sigma}_x$ and μ_x are the covariance matrix and mean vector of ***X***. Here we assume $\mathbf{\Sigma}_x$ is nonsingular, otherwise we can reduce the original ***X*** using principal components analysis (PCA) first. The use **Z**-scale permits easy comparison of the various dimensionality reduction techniques.

In the next section, we examine the Saveface method for solving the face recognition problem. We approach this problem within the pattern classification paradigm, considering each of the pixel values in a sample image as a point in a high-dimensional space (i.e., *the image space*).

The Saveface method

Sliced Average Variance Estimation (SAVE, Cook and Weisberg, 1991) was originally developed for dimensionality reduction in statistical regression problems. Let ($\boldsymbol{Y}_i$, $\boldsymbol{X}_i$) i=1, ..., n be a sample, where ***Y*** is a response variable and ***X*** is a predictor vector. Cook and Weisberg (1991) considered the following matrix in the **Z**-scale: $E(I - \Sigma_{Z|Y})^2$. Then by using singular value decomposition one can find the minimum dimension of this matrix, by identifying the eigenvectors whose corresponding eigenvalues are nonzero. We term this method the *Saveface* method.

Let μ_i and $\mathbf{\Sigma}_i$ be the mean and covariance for class i where i=1, ..., c. If the response variable $\boldsymbol{Y}$=1, ..., c, then the SAVE matrix is given by

$$M_{SAVE} = \sum_{i=1}^{c} \frac{n_i}{n} (I - \Sigma_i)^2$$

Cook and Yin (2001) developed the connection between SAVE and the classical quadratic discriminant analysis (QDA) in the subspace sense. Note that M_{SAVE} itself does not need any normal distribution assumption on $\boldsymbol{Z}|\boldsymbol{Y}$. The subspace spanned by M_{SAVE} is $\boldsymbol{S}(\mathbf{I} - \mathbf{\Sigma}_i, \text{i} = 1, \ldots, c)$ (Cook and Critchley, 2000). Under normal distribution assumptions, Odell (1979), Decell, Odell and Coberly (1981), Tubbs, Coberly, and Young (1982), and Young and Odell (1984), Young, Marco, and Odell (1987) consider an equivalent subspace using different matrices. Based on Decell, Odell and Coberly (1981)'s results, Schott (1993) formulated a slightly different matrix under the normal distribution assumption, whose subspace is equivalent to the SAVE subspace.

Comparison with the Sirface method

Cook and Yin (2001) recently showed a direct link between SAVE and QDA which is a well-known technique in classification and discriminant analysis. The optimal situation for QDA is that the conditional variables $\boldsymbol{Z}|\boldsymbol{Y}$ = i are normally distributed for each class i with different covariances. When $\Sigma_i = \Sigma$ for i = 1, ..., c, the SAVE matrix reduces to the SIR matrix. When not all $\Sigma_i = \Sigma$, using the SIR matrix could result in loss of classification information, whereas the SAVE matrix captures all the classification information. On the other hand, SAVE is "optimal" for QDA, since it removes all the redundant directions. In

addition, in a manner similar to Sirface, Saveface only finds the smallest set of predictors while we still could choose different classifiers.

Generally speaking, SIR captures all information in the first (inverse) moment and SAVE captures all information in the first two (inverse) moments. In other words, the Sirface method estimates the mean difference subspace while the Saveface method estimates the mean and covariance difference subspace. Since $S(M_{SIR}) \subseteq S(M_{SAVE})$, Saveface is more comprehensive.

Test for determining the optimal reduced dimension *d*

Let

$$M_{SAVE} = \Gamma^T \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \Gamma$$

Where $\mathbf{\Gamma}$ is a $p \times p$ orthogonal matrix whose columns are the eigenvectors $v_1, \cdots, v_p$ of $M_{SAVE,}$ and $\mathbf{\Gamma}^T=(\mathbf{\Gamma}_1, \mathbf{\Gamma}_0)$, $\mathbf{\Gamma}_1$ is a $p \times$ d matrix whose columns are the eigenvectors $v_1, \cdots, v_d$ of $M_{SAVE,}$ $\mathbf{\Gamma}_0$ is a $p \times (p\text{-}d)$ matrix whose columns are the eigenvectors $v_{d+1}, \cdots, v_p$ of M_{SAVE}. D is a $d \times d$ diagonal matrix whose elements $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of M_{SAVE}. If *d* is known, we just use the estimate $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ of the sample matrix $\hat{M}_{SAVE}$, and use their corresponding eigenvectors $\hat{v} = \hat{v}_1, \cdots, \hat{v}_d$. If *d* is unknown, an inference procedure about *d* is required. The relevant test statistic (Li, 1991) is

$$\hat{\Lambda}_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j$$

Where $\hat{\lambda}_j$'s are the eigenvalues of the sample matrix $\hat{M}_{SAVE}$.

So far, there is no general asymptotic distribution for the SAVE test statistic. Cook and Lee (1999) developed a special asymptotic test for the binary classification problems where $\boldsymbol{Y} = 0,1$. Here we use a permutation test which was used in the Sirface method as first suggested by Cook and Weisberg (1991) and further developed by Cook and Yin (2001). We believe that this test is quite robust as demonstrated by Cook and Yin (2001), and Yin and Cook (2002). When the predictors are normal, both the asymptotic test and the permutation test are in agreement, otherwise, permutation test can recover the optimal reduced dimension but not the asymptotic test.

Let $\mathbf{U} = (\mathbf{u}_j)$ denote the $p \times p$ matrix of eigenvectors $\mathbf{u}_j$ of the population kernel matrix $\mathbf{M}$. Consider testing the hypothesis that $d \le m$ versus $d > m$. Partition $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ where $\mathbf{U}_1$ is $p \times m$. The following proposition (Cook and Yin 2001) provides a basis for constructing permutation tests and for inference on d:

Proposition 1: *Let* $\boldsymbol{U}$ *be constructed as indicated previously.*

(i) If $(\boldsymbol{Y}, \boldsymbol{U}_1^T \boldsymbol{Z}) \perp\!\!\!\perp$ (independent of) $\boldsymbol{U}_2^T \boldsymbol{Z}$ *then* $m \ge d$.

(ii) Assume that $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}$. *Then* $m = d$ *if and only if* $(\boldsymbol{Y}, \boldsymbol{U}_1^T \boldsymbol{Z}) \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}$.

(iii) Assume that $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z} | \boldsymbol{Y}$. *If* $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$, *then* $m \ge d$.

Part (i) of this proposition may be the most important in practice because it requires no conditions. It says that if $(\boldsymbol{Y}, \boldsymbol{U}_1^T \boldsymbol{Z})$ is independent of $\boldsymbol{U}_2^T \boldsymbol{Z}$ (i.e., $f(Y, U_1^T Z, U_2^T Z) = f(Y, U_1^T Z) \cdot f(U_2^T Z)$) then we can discard the last p - m principal predictors $\boldsymbol{U}_2^T \boldsymbol{Z}$ without any loss of information on classification. We propose to test that possibility by comparing the observed test statistic $\hat{\Lambda}_m$ to its permutation distribution under the null hypothesis. This involves essentially re-computing $\hat{\Lambda}_m$ for each of a selected number of random permutations of the elements of the sample version of $\boldsymbol{U}_2^T \boldsymbol{Z}$,

and then comparing the observed value to its permutation distribution to obtain the *P*-values v_m, $m = 0, \ldots, p\text{-}1$.

If v_m is large and non-significant then the subspace spanned by $\mathbf{U}_1$ (denoted as $S(\mathbf{U}_1)$) provides an upper bound on *d*. The smallest inferred upper bound is the one with the first non-significant *P*-value in the sequence $v_0, \ldots, v_{p-1}$.

Application of Proposition 1(i) to test the hypothesis that $d \leq m$ in practice involves the following general steps:

5. Compute the sample kernel matrix $\hat{M}$ for SIR and form the matrices of its eigenvectors $\hat{U}_1 = (\hat{u}_1, \ldots, \hat{u}_m)$ and $\hat{U}_2 = (\hat{u}_{m+1}, \ldots, \hat{u}_p)$.
6. Construct the vectors of sample principal predictors $\hat{V}_{1i} = \hat{U}_1^T \hat{z}_i$ and $\hat{V}_{2i} = \hat{U}_2^T \hat{z}_i$, $i=1, \ldots, n$.
7. Randomly permute the indices *i* of the $\hat{V}_{2i}$'s to obtain the permuted set $\hat{V}_{2i}^*$.
8. Construct the test statistic $\hat{\Lambda}_m^*$ based on the original data $\boldsymbol{Y}_i$, $\hat{V}_{1i}^*$ along with the permuted data $\hat{V}_{2i}^*$, $i=1, \ldots, n$.

After repeating steps 3 and 4 a number of times, the *P*-value v_m is just the fraction of $\hat{\Lambda}_m^*$ that exceeds $\hat{\Lambda}_m$. Repeating steps 1–4 for $m=0, \ldots, p\text{-}1$ gives the required series of *P*-values. We found this simple test to be quite useful in practice.

The theory behind the test computed under Proposition 1(i) guarantees only an upper bound on *d* so it is possible that we will end with more predictors than needed. Hence, additional assumptions are needed to eliminate that possibility. Proposition 1 (ii) requires that $\boldsymbol{U}_1^T \boldsymbol{Z}$ and $\boldsymbol{U}_2^T \boldsymbol{Z}$ be marginally independent, when $\boldsymbol{Z}$ is normally distributed.

Using the test procedure sketched previously, the condition of Proposition 1(ii) allows us to infer directly about d rather than to infer about an upper bound.

A first approach to some discriminant analysis problems may involve the assumption that the conditional distribution of $\boldsymbol{Z}|\boldsymbol{Y}$ is normal. It may be reasonable in such cases to base a permutation test on the conditional independence statement in Proposition 1(iii) rather than on marginal independence $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}$ as in Proposition 1(ii). Assuming $\boldsymbol{U}_1^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}_2^T \boldsymbol{Z}|\boldsymbol{Y}$, and $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$ implies $m \geq d$. Thus, failure to reject $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$ by using a permutation test allows us to infer that the principal predictors $\boldsymbol{U}_2^T \boldsymbol{Z}$ can be discarded. The permutation algorithm sketched above can be adapted to test $\boldsymbol{U}_2^T \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$.

The algorithm for the Saveface method

The following steps summarize the recognition process using the Sirface method:

1. Create the training data matrix (X): Each of the training images is stored in a vector of size p

$$x^i = \left(x_1^i, \cdots, x_p^i\right)^T,\ 1 \leq i \leq n \ (n \text{ is the number of training images})$$

The training images are then combined into a data matrix of size $p \times n$.

$$X = \left(x^1, x^2, \cdots, x^n\right)$$

2. Compute the overall mean (μ): The overall mean image is a column vector such that each entry is the mean of all the corresponding pixels of the training images.

$$\mu = (\mu_1, \mu_2, \cdots, \mu_p)^T, \text{ where } \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^i,\ 1 \leq j \leq p$$

3. Create the centered data matrix ($\widetilde{X}$): Each of the training images must be centered. Subtracting the mean image from each of the training images centers the training images.

$$\widetilde{x}^i = x^i - \mu,\ 1 \le i \le n$$

Once the training images are centered, they are combined into a centered data matrix of size $p \times n$.

$$\widetilde{X} = \left(\widetilde{x}^1, \widetilde{x}^2, \cdots, \widetilde{x}^n\right)$$

4. Create the covariance matrix (Σ): The data matrix is multiplied by its transpose to create a covariance matrix.

$$\Sigma = \frac{1}{n}\widetilde{X}\widetilde{X}^T$$

5. Compute the eigenvalues and eigenvectors of the covariance matrix: The eigenvalues $(\lambda_i, 1 \le i \le p)$ (ordered from high to low) and corresponding eigenvectors $(v_i, 1 \le i \le p)$ are computed for the covariance matrix.

$$\Sigma v_i = \lambda_i v_i$$

6. Compute Z-scale images (Z): If Σ is nonsingular, then $Z = \Lambda^{-\frac{1}{2}}\Gamma^T\widetilde{X}$. Otherwise, find the number of positive eigenvalues (*k*) and

$$Z = \Lambda_k^{-\frac{1}{2}}\Gamma_k^T\widetilde{X}$$

Where $\Lambda_k = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}$, $\Gamma_k = (v_1, \ldots v_k)$

7. Compute the Save matrix

$$M_{SAVE} = \sum_{i=1}^{c} \frac{n_i}{n}(I - \Sigma_i)^2$$

where I is identity matrix of size $k \times k$ and Σ_i is covariance matrix ($k \times k$) for class i in Z-scale.

8. Test for the optimal reduced dimension of the subspace (*d*):

8.a Compute the eigenvalues $(\lambda_i, 1 \le i \le k)$ (ordered from high to low) and corresponding eigenvectors $(v_i, 1 \le i \le k)$ for the Save matrix.

8.b Test the reduced dimension and define the reduced feature subspace: a permutation test based on Cook and Yin (2001)'s proposition is constructed to infer the reduced dimension.

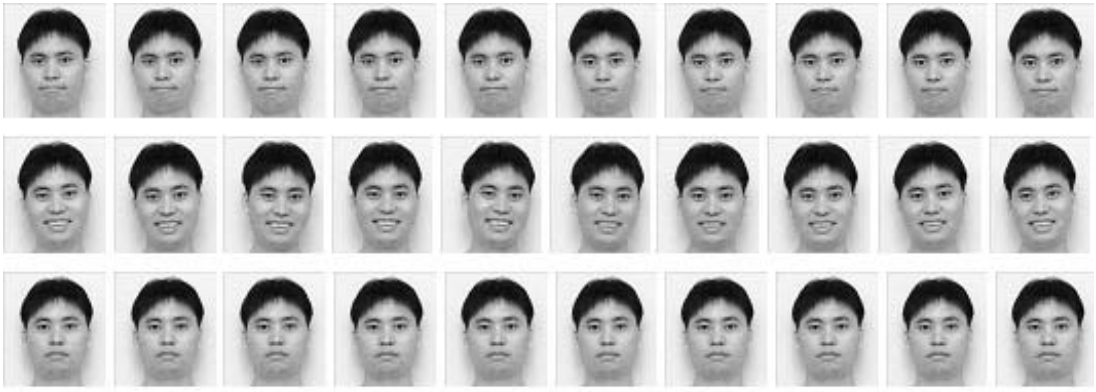
9. Project the training images onto reduced subspace: each of the training images in Z-scale is projected onto the reduced subspace.

10. Identify the test images: each test image is first transformed into the Z-scale and then projected onto the reduced subspace. The projected test image is compared to every projected training image and the training image that is found to be closest to the test image is used to identify the training image.

Experimental results

Since the Saveface method estimates both the mean and covariance difference subspace while the Sirface method only estimates the mean difference subspace, the assumption that the covariance matrices contain class-discriminantory information must be satisfied in the real face databases to show the improvement in the discriminatory capability of the Saveface method over the Sirface method. In other words, in such databases, the class-discriminatory information lies not only in the mean difference subspace but also in the covariance matrices. The Saveface method is able to recover the

Fig. 4.1. The face database for testing the Saveface and Sirface methods



class-discriminantory information. However, LDA and the Sirface method can only come up with a reduced dimension that is not capable of separating the classes. In this study, a special face database is constructed based on this assumption. The database contains 68 images of 3 facial expressions (angry, happy, and neutral) posed by a subject (Fig. 4.1). Three experiments are constructed using this face database to test the Saveface method and the Sirface method. For all experiments, classification was performed using a nearest neighbor classifier. The results tabulated in Table 4.1 and 4.2 have shown when the class-discriminatory information is contained not only in the mean difference subspace but also in the covariance difference subspace, by using the Saveface method, a very high classification accuracy can be achieved with a relatively higher reduced dimension than that obtained using the Sirface method. In addition, the classification accuracy could drop drastically if the reduced dimension is smaller than the optimal reduced dimension. This is because some important classification information lies in the discarded dimensions. In these experiments, the reason that the reduced dimensions for the Sirface method are much smaller than those of the Saveface method is because the number of expressions (classes) is very small.

Table 4.1. Test results for the Saveface method

Dataset	N_{class}	N_{train}	N_{test}	Positive eigenvalues	Reduced dimen.	Cumulative proportion of eigenvalues	Classification Accuracy
1	3	15	53	14	10	75.8%	64.2%
					11	83.2%	88.7%
					12	90.7%	100%
2	3	21	47	20	14	73.0%	64%
					15	78.1%	91.5%
					16	83.2%	91.5%
					17	88.4%	100%
3	3	30	38	29	25	88.5%	65.8%
					26	92.0%	97.4%
					27	95.5%	100%

Table 4.2. Test results for the Sirface method

Dataset	N_{class}	N_{train}	N_{test}	Reduced dimen.	Classification Accuracy
1	3	15	53	2	85.1%
2	3	21	47	2	83%
3	3	30	38	2	78.9%

Conclusion

In this paper we proposed a novel technique for data dimensionality reduction in the context of appearance-based face recognition. This technique is based on Sliced Average Variance Estimate (SAVE) regression and termed as the *Saveface* method. The experimental results show that the Saveface method appears to be more powerful than the Sirface method and LDA method when the class-discriminantory information is contained not only in the mean difference subspace but also in the covariance difference subspace of the real face database. Further experiments for extending these results to other large real face databases are necessary and will be investigated to test and validate the performance of the Saveface method.

References

P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class-specific linear projection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 711-720.

R.D. Cook, and S. Weisberg, Discussion of Li (1991). *Journal of the American Statistical Association*, Vol. 86, pp. 328-332, 1991.

R.D. Cook and H. Lee, Dimension reduction in regressions with a binary response, *Journal of the American Statistical Association*, Vol. 94, pp. 1187-1200, 1999.

R.D. Cook, and F. Critchley, Identifying outliers and regression mixtures graphically, *Journal of the American Statistical Association*, Vol. 95, pp. 735-749, 2000.

R.D. Cook, and X. Yin, Dimension reduction and visualization in discriminant analysis, invited paper with discussion in the *New Australia and New Zealand Journal of Statistics*, 2001, 43, No. 2, 147-199.

H.P. Decell, P.L. Odell, and W.A. Coberly, Linear dimension reduction and Bayes classification, *Pattern Recognition*, vol. 13, No. 3, pp. 241-243, 1981.

P.L. Odell, Amodel for dimension reduction in pattern recognition using continuos data, *Pattern Recognition*, vol. 11, pp. 51-54, 1979.

J.R. Schott, Determining the dimensionality in sliced inverse regression, *Journal of the American Statistical Association*, Vol. 89, pp. 141-148, 1994.

J.D. Tubbs, W.A. Coberly, and D.M. Young, Linear dimension reduction and Bayes classification with unknown population parameters, Pattern Recognition, vol. 15 No. 3, pp. 167-172, 1982.

X. Yin, *Dimension reduction using inverse third moments and central k-th moment subspaces*, PhD Thesis, Dept. of Statistics, University of Minnesota, 2000.

X. Yin, and R.D. Cook, Dimension reduction for the conditional *k*-th moment in regression, *Journal of the Royal Statistical Society Ser B*, 2002, 64 part 2, 159-175.

D.M. Young, V.R. Marco, and P.L. Odell, Quadratic discrimination: some results on optimal low-dimensional representation, *Journal of Statistical Planning and Inference*, 17, 1987.

CHAPTER 5

SUMMARY AND CONCLUSION

Two novel moment-based methods which are insensitive to large variation in lighting direction and facial expression are developed for appearance-based face recognition using dimension reduction methods originally developed in statistics. The two methods are based on Sliced Inverse Regression (SIR) and Sliced Average Variance Estimate (SAVE) and termed as the *Sirface* method and the *Saveface* method, respectively. The Sirface method estimates the mean difference subspace while the Saveface method estimates the mean and covariance difference subspace.

The Sirface method is compared to a traditional linear discriminant analysis (LDA) method termed as the *Fisherface* method in Chapter 3. Generally, LDA is optimal only under the assumption of normality with equal covariance matrices for each class. The dimension q of the reduced subspace under LDA satisfies the property $q \leq \min(c\text{-}1, p)$. However, the optimum value d of q can be arrived at only by an exhaustive search over a range of q. On the other hand, the Sirface method is optimal over a larger class of distributions (including the normal distribution) and there exists a statistical testing procedure for determining the optimum dimensionality d of the reduced subspace. Since an exhaustive search for determining the optimum value of q is averted, the Sirface method is much faster and more robust than LDA.

When the covariance matrices contain class-discriminatory information, LDA as well as the Sirface method will fail to capture all the necessary information even under the assumption of normality. One could resort to quadratic discriminant analysis (QDA)

which will capture the necessary second-order information. However, with many variables, QDA may not be a good choice since it is computationally very expensive. In contrast, the Saveface method discussed in Chapter 4 is a feature extraction technique that captures all the class-discriminatory information present in both the mean vectors and the covariance matrices regardless of the distribution of the underlying data. After the Saveface method computes the new predictors in reduced-dimensional space, it is not necessary to use a quadratic classifier and a simple classifier such as the nearest-neighbor classifier is sufficient. Thus the Saveface procedure captures more information than LDA and the Sirface method while avoiding the computational complexity of QDA. Besides, the Saveface method is optimal over a larger class of distributions including the normal distribution and also has an associated formal testing procedure which yields the optimum value d of the reduced subspace dimension q without having to resort to exhaustive search.

Besides face recognition, the proposed Sirface and Saveface methods could be used for other appearance-based computer vision problems such as hand gesture recognition, which has been used extensively in computer game navigation, TV remote control, American Sign Language recognition, virtual navigation, and human-computer interaction. Compared to algorithms based on traditional approaches to feature extraction and discriminant analysis, the proposed Sirface and Saveface methods are computationally more efficient, more robust, capable of capturing more complex discriminatory information and thus capable of producing better quality solutions under more general conditions. In this study, experiments for the Sirface method and the Saveface method are performed on a single face database. Further experiments for

extending the results to other large face databases are necessary and will be addressed in our future study. For complicated data where the discriminatory information exists in moments beyond the second order, none of the methods mentioned in this thesis i.e., PCA, LDA, the Sirface method, and the Saveface method, is able to capture the discriminatory information. In order to capture the discriminatory information present in moments beyond the second order, other high-moment based methods, such as sliced average third-moments, would be a good direction to investigate.