Improving Face Recognition Performance Using a Hierarchical Bayesian Model

Ashwini Shikaripur Nadig

Submitted to the graduate degree program in
Electrical Engineering & Computer Science and the
Graduate Faculty of the University of Kansas
School of Engineering in partial fulfillment of the
requirements for the degree of Master of Science

Thesis Committee:

Dr. Brian Potetz: Chairperson

Dr. Prasad Kulkarni

Dr. Luke Huan

Date Defended

The Thesis Committee for Ashwini Shikaripur Nadig certifies
That this is the approved version of the following thesis:

Improving Face Recognition Performance Using a Hierarchical Bayesian Model

Committee:

Dr. Brian Potetz

Dr. Prasad Kulkarni

Dr. Luke Huan

Date Approved

Acknowledgements

I would like to take this opportunity to express my sincere thanks to my advisor, Dr. Brian Potetz for being an incredible mentor to me. But for his excellent support and guidance, I wouldnt have come so far. I also thank Dr. Luke Huan and Dr. Prasad Kulkarni for serving on my committee. I would like to thank my parents who always believed in me and supported me in all my endeavors. I also thank my sister, Pallavi, and brother-in-law, Karthik, for their support. Finally, I would like to thank all my friends who have always been there for me.

Abstract

Over the past two decades, face recognition research has shot to the forefront due to its increased demand in security and commercial applications. Many facial feature extraction techniques for the purpose of recognition have been developed, some of which have also been successfully installed and used. Principal Component Analysis (PCA), also popularly called as Eigenfaces has been used successfully and also is a de facto standard. Linear generative models such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) find a set of basis images and represent the faces as a linear combination of these basis functions. These models make certain assumptions about the data which limit the type of structure they can capture. This thesis is mainly based on the hierarchical Bayesian model developed by Yan Karklin of Carnegie Mellon University. His research was mainly focused on natural signals like natural images and speech signals in which he showed that for such signals, latent variables exhibit residual dependencies and non-stationary statistics. He built his model atop ICA and this hierarchical model could capture more abstract and invariant properties of the data. We apply the same hierarchical model on facial images to extract features which can result in an improved recognition performance over already existing baseline approaches. We use Kernelized Fisher Discriminant Analysis (KFLD) as our baseline as it is superior to PCA in a way that it produces well separated classes even under variations in facial expression and lighting. We conducted extensive experiments on the GreyFERET database and tested the performance on test sets with varying facial expressions. The results demonstrate the increase in performance that was expected.

Contents

List of Figures

Chapter 1

Introduction

Face recognition has been a hot research area over the past 30 years and has made significant advances thus far. This can be attributed to the need to secure and identify information and assets among millions. Many forms of identification technologies have been used over the years. The most common is the use of a Password/PIN (Personal Identification Number) for authentication. Many identification systems use attributed identifiers (name, SSN, bank acct no. etc), biographical identifiers (address, profession, education, etc) and biometric identifiers (photograph, fingerprints, etc). Since Passwords/PIN's, attributed identifiers and biographical identifiers carry a risk of forgery and theft, there has been an increased interest in the use of biometric identifiers as it is very difficult or impossible to tamper with an individual's biometric characteristics. The increased need for security and law enforcement has turned the spotlight towards biometric identification systems. The various biometrics that these systems can use are fingerprint, palm print, hand geometry, iris geometry, voice, gait and face. The question of which biometric to use is based on the specific application and also social and political factors. A single biometric cannot be ideal for all applications.

For instance fingerprint and iris recognition require the individual to pause and are subject to physical proximity and intrusiveness and needs the cooperation of the subjects. Such systems are appropriate for entry into high security areas and bank transactions. Face recognition systems on the other hand are unobtrusive and don't need the cooperation of the subjects which is very much needed in security and law enforcement applications (for example, to track a terrorist or a criminal in airports without the knowledge or cooperation of the subject). Apart from security and law enforcement applications, face recognition technology is being used in a variety of commercial applications such as video games, virtual reality, human-computer interaction and next generation smart environments.

1.1 Formal Definition

Given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces. Available collateral information such as race, age, gender, facial expression, or speech maybe used in enhancing recognition [37].

1.2 Steps in Face Recognition

There are four steps in the face recognition process: face segmentation or detection, normalization, feature extraction and recognition. All the four phases are important and interrelated.

- **Face Segmentation or Detection:** Given a probe image, the first task of a Face Recognition System (FRS) is to detect and separate the face region. Segmentation is easy when the background is clear but becomes a challenge when it is cluttered with other objects.

- **Normalization:** After detection, the face image needs to be normalized. This means that the face image should be standardized in terms of scale, rotation, illumination etc relative to the gallery images. For this, accurate locations of facial landmarks such as eyes, nose and mouth need to be provided to the FRS. The normalization step is very important for any FRS to perform well.

- **Feature extraction and Recognition:** During feature extraction, a face image is transformed into a simplified mathematical representation called the features and stored for recognition purposes. The feature extraction algorithm should be such that the extracted features should be able to represent the maximum distinctive information related to the face. A database of features is built in this way. When a probe image is presented to the FRS, it is translated to obtain its features and then compared with features in the database for recognition.

Many algorithms and techniques have been developed for feature extraction and recognition and this thesis mainly aims in developing a feature extraction technique which may be superior to the already existing techniques.

This thesis is organized as follows: in Chapter 2 we give the history and background of face recognition techniques. Chapter 3 gives a description of Karklin's hierarchical Bayesian model and the motivation behind this model. In Chapter 4, we give a brief description of the standard face recognition datasets and evaluation protocols including FERET, FRVT and FRGC. The experimental procedure and the results obtained are presented in Chapter 5 followed by the conclusions and possible future directions of work in the final chapter.

Chapter 2

Background

Over the past three decades, researchers from various disciplines like image processing, computer vision, pattern recognition, neural networks, computer graphics and psychology have been working extensively on the various aspects of face recognition. Earlier work in face recognition started in the 70's where key facial landmarks like the eyes, nose, mouth and the geometrical relationships between them were used for recognition purposes. But they performed poorly as such and deteriorated even more with variations in pose and illumination. During the 90's statistical approaches which treated face recognition as a pattern recognition problem were developed. In 1990, Kirby and Sirovich [16] showed that a face can be decomposed into representative eigenfaces (which are nothing but the principal components or eigenvectors) and the same face can be reconstructed using very few eigenfaces. In 1991 Turk and Pentland [32], used this approach and developed a recognition system based on PCA which showed good results for images with controlled lighting and orientation. Since then, significant amount of work has been carried out in this area and today there are commercial face recognition systems available. Feature extraction and classification techniques can be classified as:

- Holistic methods
- Feature-based (structural) matching methods
- Hybrid methods

2.1 Holistic methods:

These methods take the whole face image as the raw input. The major challenge faced by such methods is the high dimensionality of the data. For instance, for an image of size 600 X 500 the dimensionality is $3 * 10^5$ while we usually have only a few samples per subject which exacerbates the problem of dimensionality. The most popular method in this category used for dimensionality reduction is Principal Component Analysis (PCA) and there have been many extensions based on PCA.

Principal Component Analysis is a statistical method used for dimensionality reduction of data sets while retaining the majority of variations present in the data set. PCA was used for the first time for face recognition by Turk and Pentland [32] and they called it Eigenfaces. Each data vector (here a face) can be represented as a linear combination of orthogonal basis functions ϕ_i; $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \phi_i \approx \sum_{i=1}^{m} \alpha_i \phi_i$ (usually $m \ll n$). The basis functions can be obtained by solving the eigenvalue problem: $C\phi = \phi\Lambda$ where C is the covariance matrix of the data vectors. Each face can be represented as a vector of weights by projecting it onto a set of m eigenfaces called facespace. A face image is then represented as a point in an m-dimensional facespace. The eigenvalues are equal to the variance of the projection of the training set onto the eigenfaces. The eigenfaces are ordered with respect to the eigenvalues; the lower order eigenfaces capture the larger variations in the training set and the higher order eigenfaces capture the smaller variations mostly considered to be noise. A gallery of face images is represented as a vector of weights as explained above. When a probe image to be identified is given, it is first projected onto face space and then compared with the already existing database of gallery weights using a similarity measure (Euclidean, L1, etc). PCA performance is dependent on various design decisions like the number of eigenfaces retained, the similarity measure used and the image preprocessing techniques used and also the number of training samples used. The various design descisions and how they affect recognition performance is elucidated in [24].

The eigenfaces method was an important breakthrough for face recognition and since then it has been used as a de facto standard for benchmarking other algorithms. This method has shown significant performance for images under controlled conditions and is sensitive to changes in illumination, scale and pose. Turk and Pentland who were the first to use PCA for recognition used the Euclidean distance measure for classification. Since then many extensions and modifications to the basic eigenfaces approach have been proposed.

The standard eigenfaces was extended to a **Bayesian approach** [23] in which the Euclidean distance was replaced by a probabilistic measure of similarity. Two mutually exclusive classes: Ω_I representing intrapersonal variations between multiple images of the same individual and Ω_E representing extrapersonal variations in matching two different individuals are defined. Likelihood functions $\mathrm{P}(\Delta \mid \Omega_I)$ and $\mathrm{P}(\Delta \mid \Omega_E)$ are estimated for a given intensity difference $\Delta = I_1 - I_2$ and classification is done using the maximum a posteriori (MAP) rule i.e., two images are determined to belong to the same individual if $P(\Omega_I \mid \Delta) > P(\Omega_E \mid \Delta)$ and vice versa. A huge performance boost over the standard eigenfaces was reported and it was one of the top performers in the FERET 1996 evaluations.

View based eigenfaces [25] in which M separate eigenspaces are built each capturing the variations of N individuals in a common view (or orientation). When a probe image is given, the eigenspace to which it belongs is first determined. Then it is described using the eigenvectors of that eigenspace and used for recognition.

Linear Discriminant Analysis has been a very successful approach for face recognition [10]. The reason why Fisher Linear Discriminant Analysis (FLD) is superior to PCA is presented in [7] and they have showed that FLD outperforms PCA especially when challenging data sets having large illumination and expression variations are given. In PCA, the eigenvectors are selected so as to maximize the scatter of the projected samples. The drawback of this approach is that PCA maximizes not only the scatter due to between class variations but also the scatter due to within class variations which is mostly due to illumination. So different classes in the projected space will not be well clustered but smeared together. FLD finds the most discriminative projection in eigenspace by maximizing the ratio of the between-class scatter and the within-class scatter.

$$S_b = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{2.1}$$

where S_b is the between-class scatter matrix and;

$$S_w = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \tag{2.2}$$

where S_w is the within-class scatter matrix, μ_i is the mean image of class X_i, and

N_i is the number of samples in class X_i. The major drawback of this approach is the need to have multiple samples per class in the training set which is usually not available.

The basic LDA was extended to a **subspace LDA** [36] to improve the performance of basic LDA. PCA and LDA are combined to improve the generalization capability of LDA when only few samples per class are available. Face images are first projected to a lower dimensional facespace using PCA and then LDA is used as a linear classifier. The system was tested using FERET and it showed significant improvement over the basic LDA and also other competing FERET algorithms.

An evolution pursuit for face recognition using Genetic Algorithms (GA's) in determining the optimal basis for encoding human faces was presented in [20]. Other methods based on PCA are SVM based methods [11], 2D-PCA [34], ROCA [9].

Independent Component Analysis (ICA) is yet another linear feature extraction method like PCA. Face recognition by Independent Component Analysis which used the Infomax algorithm proposed by Bell and Sejnowski [8] is presented in [5]. Basis images found by PCA depend only on pair-wise relationships between pixels in the image database. Important information can be present in higher order relationships among pixels. Second-order statistics capture only the amplitude spectrum of images but not the phase spectrum. Higher-order statistics capture both. It is the phase information that has the structural information which can be a plus for recognition purposes. The differences between ICA and PCA and its advantages over PCA are listed below:

- ICA is a generalization of PCA which is sensitive to higher order statistics, not just the covariance matrix.

- PCA makes an assumption of gaussian sources that makes it inadequate when the sources are non-gaussian. It has been shown that many natural signals like speech and natural images are better described as linear combinations of ”high-kurtosis” or ”super-gaussian” sources. So PCA falls short when it comes to representing such signals.

- ICA basis vectors are statistically independent and not just linearly decorrelated as in PCA. Also they need not be orthogonal as in PCA.

A more detailed explanation of ICA is provided in the next chapter.

In [5], ICA was applied for face recognition and they showed that ICA outperforms PCA. On the contrary, [22] showed that ICA does not provide significant improvement over PCA. The experiments showed that the assumption of non-gaussian and independent components need not provide a good representation for face recognition. PCA was extended to Kernel PCA in [30]. Experiments were conducted on the USPS handwritten digit data set which showed that KPCA was able to extract non linear features which provided better recognition results. There are kernel extensions for FLD [6, 35] and ICA [4] but these methods are not within the scope of this thesis.

2.2 Feature-based (structural) matching methods:

In these methods, local features such as eyes, nose and mouth are extracted and their geometric relationships and statistics are used for classification purposes. The main challenges to face recognition performance are the distortions caused due to the variations in illumination, pose and expression. Most of the algorithms developed so far work very well for faces under controlled conditions but fail when subject to these variations. A person with the same expression may be unrecognizable under different illuminations. Gabor wavelets have been widely used for face recognition as they are robust to these distortions. There is also a biological relevance for the use of Gabor wavelets. The shapes of Gabor wavelets are shown to be similar to the receptive fields of simple cells in the primary visual cortex and they exhibit desirable properties of spatial locality and orientation selectivity. The survey paper on Gabor wavelets for face recognition [31] lists all the work that has been carried out in this direction and also the results of the various approaches. Gabor wavelets were first used in the Dynamic Link Architecture proposed by Lades et al. [17]. This was extended by Wiskott et al. [33] who proposed Elastic Bunch Graph Matching. Both DLA and EBGM can be quoted as examples for feature based approaches. Gabor features called ”jets” are extracted at various fiducial points in the face like the eyes, mouth and nose and elastic graph matching techniques are used for face representation and recognition.

There are holistic approaches using Gabor features too. [21] applied the Enhanced Fisher linear discriminant model (EFM) to the Gabor feature vector of face images and showed that it outperforms PCA and LDA. [19] applied Kernel PCA with a fractional power polynomial kernel to the Gabor feature vector. Interested readers can refer to [31] for a complete list of algorithms using Gabor features.

2.3 Hybrid Methods:

These methods are a mixture of both holistic and feature-based approaches. Another extension of the eigenfaces called Modular eigenfaces [25] computes eigenfeatures for facial features like eyes, nose and mouth separately. These eigenfaces are then augmented with these eigenfeatures and used for recognition.

Chapter 3

Improving face recognition using a hierarchical bayesian model

The hierarchical Bayesian model was developed by Yan Karklin for his Ph.D. dissertation [15].

3.1 Motivation for this model

Natural images comprise of complex, high dimensional data which are rich in statistical structure. Linear generative models such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been used to model such data (data refers to the images in this context) but they have certain shortcomings. To understand ICA better, a simple illustration of the cocktail-party problem is given. Imagine a room having two people who are speaking simultaneously. Two microphones are held at two separate locations. The microphones record two time signals $x_1(t)$ and $x_2(t)$ where t is the time index. Each of these recorded signals is a weighted combination of the original speech signals emitted by the two speakers in the room which is expressed as:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \tag{3.1}$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \tag{3.2}$$

where $s_1(t)$ and $s_2(t)$ are the speech signals emitted by the two speakers and a_{11}, a_{12}, a_{21}, a_{22} are some parameters that depend on the distance of the microphones from the speakers. Now the problem is to estimate the two speech signals $s_1(t)$ and $s_2(t)$ given only the recorded signals $x_1(t)$ and $x_2(t)$. This is related to the Blind Source Separation (BSS) method in which very little or nothing is known about the mixing matrix (i.e., the weights) or the source signals (hence the term 'Blind'). Certain general assumptions are made about both the weights and the source signals. For more details on ICA, please refer to [14]. Taking ICA as the base, Yan has built his hierarchical Bayesian model.

Starting with the basic ICA representation, the data (x) (in this case images) are generated as a linear combination of basis functions ($\mathbf{A}$) weighted by coefficients ($\mathbf{u}$),

$$x = \mathbf{A}\mathbf{u}. \tag{3.3}$$

The data likelihood for this model is

$$p(x) = p(\mathbf{u}) / \mid det(\mathbf{A}) \mid \tag{3.4}$$

The basis function $\mathbf{A}$ is adapted to maximize the data likelihood and the basis function coefficients ($\mathbf{u}$) which are unknown (latent) variables are assumed to be independent and identically distributed (i.i.d.),

$$p(\mathbf{u}) = \prod_i p(u_i). \tag{3.5}$$

The priors $p(u_i)$ are chosen to be fixed sparse distributions.

This model places a restriction on the type of structures that can be captured. But often, data are rich with statistical structure and latent variables of linear models adapted to these data exhibit residual dependencies. Another drawback of linear models is that they assume that the statistical regularities do not change; i.e. they assume stationary probability distributions. But actually, it has been shown that the statistics of the data do change as the physical properties of the environment or data acquisition conditions vary. While the stationary prior assumption gives a valid approximation of true density over a large enough corpus of training data, it does not reflect the variation across contexts that is observed in many signals.

To overcome the limitations of these linear models a hierarchical Bayesian model was proposed which can capture the nonlinear statistical regularities in non-stationary natural signals.

3.2 Karklin's Hierarchical Bayesian Model

The model is a generalization of the ICA; $x = \mathbf{Au}$ where the data x are generated by the linear combination of basis functions. The basis function coefficients $(\mathbf{u})$ are assumed to be sparsely distributed; a generalized gaussian distribution with zero mean is used:

$$p(u_i|\lambda_i, q_i) = \mathcal{N}(0, \lambda_i, q_i) = z_i exp(-\left|\frac{u_i}{\lambda_i}\right|^{q_i}) \tag{3.6}$$

where z_i is a normalizing constant. The parameter q_i determines the weight of the gaussian's tails. We use a value of 2 for q_i in this thesis. For ICA, λ_i is fixed to a constant as the basis functions in $\mathbf{A}$ themselves scale to fit the data.

Here, instead of assuming that the basis function coefficients ($\mathbf{u}$) are independent, the residual dependencies are modeled using the scale parameters of the prior (variance of the coefficients) which is nothing but λ, which is modeled as a nonlinear transformation of latent higher order variables.

$$log(\lambda/c) = \mathbf{Bv}. \tag{3.7}$$

where the logarithm of the scale parameter λ_i is described as a linear combination of a matrix of density components ($\mathbf{B}$) and density component coefficients ($\mathbf{v}$) where c is a defined constant.

The joint prior distribution of coefficients $\mathbf{u}$ is now expressed as

$$-\log p(\mathbf{u}|\mathbf{B}, \mathbf{v}) = \sum_i \left|\frac{u}{c\exp([\mathbf{Bv}]_\mathbf{i})}\right|^{q_i} + \log z \tag{3.8}$$

Basis function coefficients are assumed to be independent conditional on the higher-order variables, $p(\mathbf{u} \mid \mathbf{v}) = \prod p(u_i|\mathbf{v})$. This accounts for the dependence in the magnitudes of the basis function coefficients. Each density component represents a common deviation from the standard assumption of independence. A weighted combination of these density components describes a variety of possible joint probability distributions. Following is the bayes net representation of the hierarchical generative model.

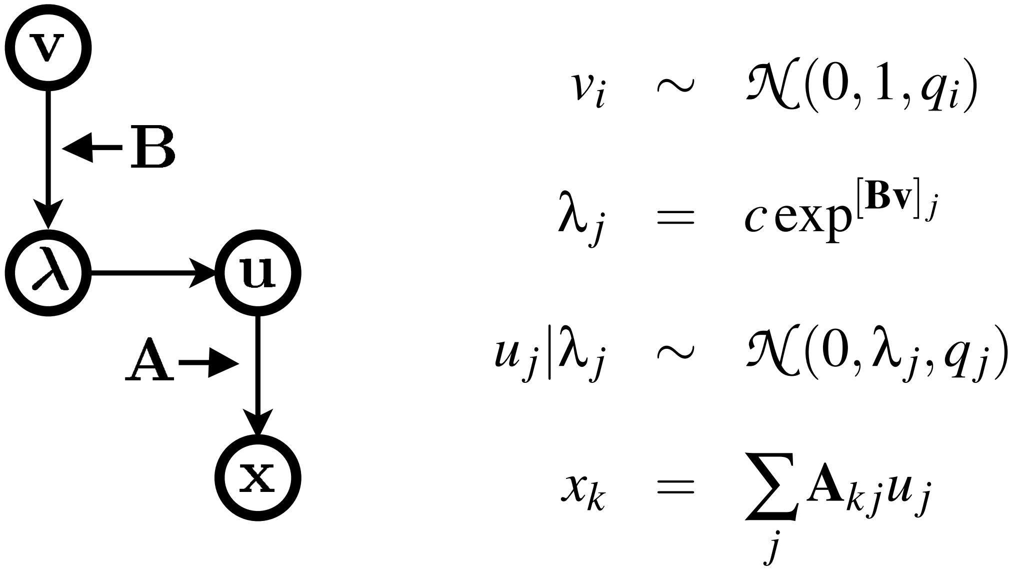
v
B
λ
u
A
x
$v_i \sim \mathcal{N}(0, 1, q_i)$
$\lambda_j = c \exp^{[\mathbf{Bv}]_j}$
$u_j|\lambda_j \sim \mathcal{N}(0, \lambda_j, q_j)$
$x_k = \sum_j \mathbf{A}_{kj} u_j$

Figure 3.1. Schematic of the hierarchical generative model (figure reproduced from [15])

Sparsely distributed random variables **v** specify (through a nonlinear transformation) the scale hyperparameters λ for the distribution of coefficients **u**. The data x are a linear combination of coefficients **u**. Matrices **A** and **B** are parameters that are adapted to the statistical distribution of the data.

Thus, the model forms a hierarchical representation in which the lower level codes data value precisely and the higher level represents more abstract and invariant properties of the signals. This model was basically used to capture the statistical structure of natural images. We are adapting the same model on face images instead of natural images to extract features which can be used to improve the already existing face recognition performance.

Chapter 4

Standard Datasets and Evaluation Protocols

In Chapter 2, we have seen the various techniques that have been developed for facial recognition. There was a need to fairly assess the performance of these algorithms for various application scenarios. To achieve this, a common database of images which was sufficiently large and an evaluation methodology were essential. In designing such evaluations, many factors had to be considered. One of the main factors was that the techniques developed were application dependent. A facial recognition technique can work well for a specific application but not so well for a different application scenario. So even the evaluations had to be designed for a specific application. Until 1993, face recognition research was in its infancy and there was no common database and evaluation protocol. Many of the researchers used to report their algorithm's performance as per their own assumptions and scoring methods on individually assembled datasets which usually used to be very small. Many papers also claimed high recognition performance on such small datasets which did not reflect the true capabilities of the algorithms.

Since then several large databases have been assembled and standard evaluations designed, the most important of them being the FERET (Face Recognition Technology) program [27,28], the FRVT vendor tests [2] and FRGC (Face Recognition Grand Challenge) [26]. A brief description of these are given below.

4.1 FERET

The Face Recognition Technology (FERET) program started in September 1993 and continued for a period of three years. It was funded by the Department of Defense (DoD) Counterdrug Technology Development Program. The main goal of the FERET program was to develop automatic face recognition capabilities that could be employed to assist security, intelligence and law enforcement personnel in the performance of their duties [28]. The program focused on three major tasks. The first task was to develop the face recognition algorithms. The second task was to collect a large database of facial images which was vital for the development and evaluation of the face recognition algorithms. The third task was to conduct government-monitored evaluations of the developed algorithms using standard testing protocols. The FERET database and evaluation protocol are de facto standards.

4.1.1 The FERET Database

Before FERET, there was no way of evaluating algorithms accurately or comparing the then existing algorithms in literature. Researchers used to assemble their own database of images according to their requirements and most of the databases used to be very small (< 50 individuals). Many papers also claimed to have very high recognition rates (> 95) on such small databases. One of the major elements of the FERET program was to develop a standard and sufficiently large database of facial images for both development and testing of algorithms.

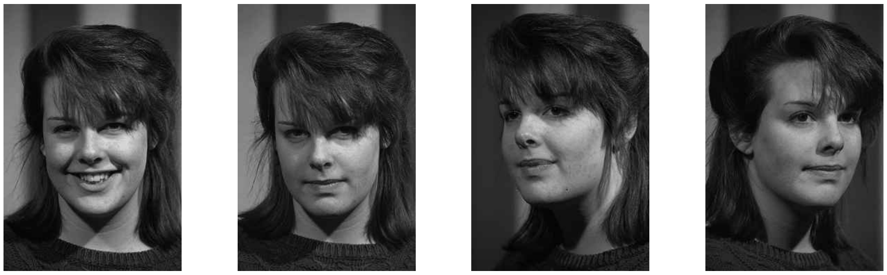

Figure 4.1. Example images of a single individual in the FERET database

The FERET database was collected in 15 sessions between August 1993 and July 1996. The images were collected in a semi-controlled environment. The same physical setup was used in each session to maintain consistency. The FERET database contains a total of 14,126 images of 1199 individuals and 365 duplicate sets of images. A duplicate image is the image of a person already in the database but acquired on a different day. This accounts for the effect of aging on recognition performance. Images of an individual were acquired in sets of 5 to 11 images. Two frontal views (fa and fb) each differing in facial expression were taken. For 200 sets of images, a third image (fc) was taken under different lighting. Duplicate images (Dup I and Dup II) of which Dup I was taken on a different day than the corresponding gallery image and Dup II taken over an year later. The remaining images were collected at various aspects between right and left profile. The database is divided into a development set provided to researchers and a sequestered set for testing.

4.1.2 The FERET Evaluation Protocol

Before FERET, there was no way of comparing face recognition algorithms as each researcher used his/her own assumptions, testing methods and images. The FERET database allowed algorithm development using a standard database. The FERET evaluations assessed the strengths and weaknesses of different approaches that could automatically locate, normalize and identify faces. A PCA based performance baseline was also established [24]. There were three evaluations conducted on August 1994, March 1995 and September 1996. Performance was computed for two tasks: identification and verification. In identification, an algorithm is given an unknown image, and it has to identify the corresponding gallery image. In verification, a probe image and the corresponding claimed identity in the gallery are given and the algorithm should verify whether the probe is the individual in the gallery. For verification results, please refer [29].

In the evaluation protocol, an algorithm is given two sets of images: the gallery set and the probe set. The gallery set is given as a set of known facial images. The probe set consists of unknown images. Let P be a probe set and G be a gallery set where P=$\{p_1, ...p_N\}$ and G=$\{g_1, ...g_M\}$. $\mid P \mid$ is the size of the probe set. For each image p_i in the probe set P, an algorithm reports a similarity $s_i(k)$ between p_i and each image g_k in the gallery set G. The probe set P is scored against gallery G, by comparing the similarity scores $s_i(.)$. It is assumed that a smaller similarity score implies a closer match. The function id(i) gives the index of the gallery image of the person in probe p_i. A probe p_i is correctly identified if $s_i(id(i))$ is the smallest score for $g_k \in G$. A probe p_i is in the top n if $s_i(id(i))$ is one of the nth smallest scores in $s_i(.)$ for gallery G. Let R_n denote the number of probes in the top n. Identification performance is reported as a cumulative match score - a graph in which the fraction of probes $R_n/|P|$ is plotted on the Y-axis and the rank on the X-axis. Here n is the rank.

Following are the top three performers in the September 96 evaluations:

- probabilistic eigenface from Massachusetts Institute of Technology [23]
- subspace LDA from University of Maryland [36] and
- Elastic Graph Matching from University of Southern California [33]

The FERET evaluations recognized three problem areas: recognizing duplicate images, recognizing under varying illumination and recognizing under pose variations.

4.2 Facial Recognition Vendor Test (FRVT)

Face Recognition Vendor Tests (FRVT) [2] provided independent government evaluations of commercially available and mature prototype face recognition systems. These evaluations were designed to provide U.S. government and law enforcement agencies with information to assist them in determining where and how facial recognition technology can best be deployed. There were three Facial Recognition Vendor Tests - FRVT 2000, 2002 and 2006 which were sponsored by Defense Advanced Research Projects Agency (DARPA), DoD Counterdrug Technology Development Program Office and National Institute of Justice (NIJ). With the end of FERET in 1997, facial recognition technology had emerged from its infancy to a prototype stage in universities and research labs. By the year 2000, there was a rapid development in not only the face recognition algorithms but also the supporting sytems and infrastructure necessary for commercial systems.

FRVT 2000 was especially designed to evaluate the capabilities of these commercial systems. It mainly focused on core technology evaluations and the product usablity. The FRVT 2000 test design was based on the September 96 FERET evaluation protocol. The algorithms were tested and compared using a standard database to allow for fairness. The product usability test examined system properties and usability. FRVT 2002 was designed to evaluate the technical progress since FRVT 2000 and also measure the performance on real-life large-scale databases. Similarly FRVT 2006 was designed to measure progress since FRVT 2002 and also determine if the goals of Face Recognition Grand Challenge (FRGC) were met.

4.3 Face Recognition Grand Challenge (FRGC)

Since FRVT 2002, many algorithms were developed that had the promise of improving recognition by an order of magnitude. These techniques include recognition from high resolution still images, 3D facial scans, Multi-sample images and pre-processing algorithms to correct for illumination and pose variations. The main goal of FRGC [3, 26] was to achieve this marked increase in recognition performance by developing algorithms for the above mentioned scenarios. The FRGC dataset is huge consisting of 50,000 images containing high resolution still images taken under controlled lighting and background, uncontrolled lighting and background and 3D scans. The FRGC evaluation protocol is based on the FERET and FRVT 2002 testing methodology.

The FRGC challenge problems consisted of six experiments. Experiment 1 measures recognition performance on frontal facial images taken under controlled lighting. In experiment 2, the effect of multiple still images on performance was evaluated. Experiments 3, 5 and 6 measure different implementations of 3D face recognition. In experiment 4, the target set consists of single controlled still images, and the query set consists of single uncontrolled still images.

Chapter 5

Experiments and Results

In this chapter, we give a detailed account of the experiments conducted and the results. The main goal of this thesis is to test the performance of our feature extraction method on frontal images with varying expression and size. We discuss the hierarchical feature extraction method using Karklin's model, the baseline face recognition approach and a description of the data set used. We start with a brief overview of the dataset used followed by a description of each step in the experimental process.

5.1 Dataset

The experimental data is divided into a training set used for training the algorithm and gallery and probe sets used for testing purposes. Having a sufficiently large number of images for training is very important. The faces images we used are a subset of the GreyFERET database. The training set consists of 1390 images of 695 subjects (2 samples per subject). The gallery and probe sets consists of 500 images each. Each image in the gallery set has a corresponding image in the probe set only differing in facial expression. The images we used for training comprise of two frontal images per subject - one with a neutral expression and the other with a variation in facial expression and some images have occlusions like spectacles. Example images from FERET are shown in the following figure. Since the main purpose of this thesis is to test the performance of our feature extraction technique with varying expressions, even the training and test sets are assembled to serve this purpose. A separate dataset of just the eyes (both left and right) were collected and stored for both the training and test sets.

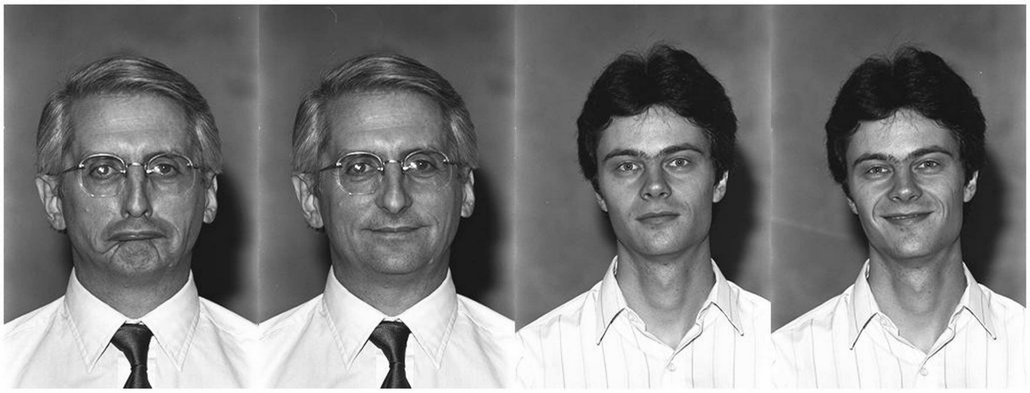

Figure 5.1. Images of two subjects from the training set with different facial expressions

5.2 Baseline Approach

The performance of any new algorithm can be gauged by comparing against already established baseline approaches. The FERET program has established PCA as the standard baseline. Fisher Linear Discriminant Analysis (FLD) and Subspace LDA were among the top performers in FERET. Based on this, we decided to use Kernelized Fisher Discriminant Analysis (KFLD) [30] as the baseline. A brief explanation of the KFLD method is given:

5.2.1 Kernelized Fisher Discriminant Analysis (KFLD)

Kernelized Fisher Discriminant Analysis is a non-linear generalization of Fisher Discriminant Analysis. In KFLD, data is first mapped non-linearly into some feature space $\Im$ and Fisher's discriminant is computed there, implicitly yielding a non-linear discriminant in input space [30]. Let ϕ be a non-linear mapping into some feature space $\Im$. To find the linear discriminant in $\Im$ we need to maximize

$$J(\omega) = \frac{\omega^T {S_b}^{\phi} \omega}{\omega^T {S_w}^{\phi} \omega} \tag{5.1}$$

where $\omega \in \Im$ and ${S_b}^{\phi}$ and ${S_w}^{\phi}$ are the between and within class scatter matrices just like in Fisher's discriminant but in $\Im$ space. If $\Im$ is very high or infinite dimensional, this will be an impossible task. To overcome this limitation kernels can be used which compute the dot products in feature space $(\phi(x) \cdot \phi(y))$ without mapping explicitly to $\Im$. Kernels which have been useful are Gaussian RBF, $k(x, y) = exp(|| \ x - y \ ||^2/c)$, or polynomial kernels $k(x, y) = (x \cdot y)^d$ where c and d are some positive constants. In kernel space we need to maximize

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha} \tag{5.2}$$

where $M = (M_1 - M_2)(M_1 - M_2)^T$ where $M_{ij} = \frac{1}{l_i} \sum_{k=1}^{l_i} k(x_j, {x_k}^i)$ and $N = \sum_{j=1,2} K_j (I - 1_{l_j}) {K_j}^T$, K_j is a $l \times l_j$ matrix with ${K_j}_{nm} = k(x_n, {x_m}^j)$, I is the identity matrix, 1_{l_j} the matrix with all entries $1/l_j$ and where α is related to w

by the equation:

$$w = \sum_{i=1}^{l} \alpha_i \phi(x_i) \tag{5.3}$$

5.3 Experimental Procedure

As already described in chapter 1, any face recognition system consists of the following steps:

- Face segmentation
- Normalization
- Feature extraction
- Recognition

Face segmentation is not within the scope of this thesis.

5.3.1 Normalization

Normalization is a very important step for any face recognition algorithm to perform well. Firstly, any color image is transformed into gray scale. We used the GreyFERET database where the images are already grayscale. Next, the luminance is normalized by linearly rescaling each image to the interval [0,1]. Coordinates for eye, nose and mouth locations are provided by the FERET database. All the images are standardized with respect to scale and rotation using the locations of eyes. Finally we crop the images to 151 X 119 pixels such that only the facial region is retained .

Each image is then vectorized and stored in a matrix such that the images are in rows and the pixels are in columns. Before usage, each image is subtracted

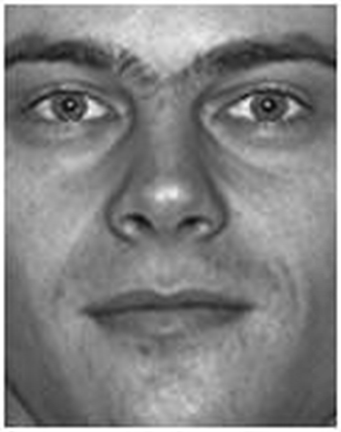

Figure 5.2. A normalized and cropped image

with the mean of all the images so as to maintain zero mean and unit variance across all the images.

5.3.2 Feature extraction

The feature extraction method we use can be categorized as a hybrid method. Hybrid methods are a mixture of both holistic and feature based methods. Our approach is to extract the hierarchical features from just the eye regions (both left and right) of the training images. The resulting hierarchical features are then concatenated to the raw pixel values of the cropped training images and KFLD is applied to obtain the final feature set. The goal of this step is to find the eigenvectors (say **VFisher**) of the scatter matrices in kernel space as described earlier. The columns of **VFisher** contains a set of basis images which can be used to represent the faces.

5.3.2.1 Karklin's hierarchical feature extraction method

A detailed explanation of Karklin's hierarchical Bayesian model is given in Chapter 3. Karklin's model is a generalization of Independent Component Analysis (ICA) and he develops his model atop ICA. We use a training method which is similar to Karklin's method. First the linear basis functions ($\mathbf{A}$) are adapted to the data using standard ICA using the FastICA package [1]. The FastICA package is a free MATLAB program that implements the fast fixed-point algorithm for independent component analysis and projection pursuit. The density components ($\mathbf{B}$) are learned on the coefficients of the fixed $\mathbf{A}$ by maximizing the posterior over the training data.

$$\log p(\mathbf{B}|\mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{A}) \propto \sum_n \log p(\mathbf{u}_n|\mathbf{B}, \hat{\mathbf{v}_n})p(\hat{\mathbf{v}_n})p(\mathbf{B})/|det\mathbf{A}| \tag{5.4}$$

This was obtained by performing gradient ascent. The learning method was unsupervised.

We fixed the number of independent components (IC's) to 25 since these IC's seemed to encode all the useful information. Moreover, the lesser the number of independent components, the greater the dimensionality reduction and faster the computation. We tried varying the number of density components used and observed that using more than 4 density components was redundant as they encoded more or less the same information. Hence the number of density components was fixed to 4.

5.3.3 Recognition

In this phase, the gallery and probe sets are used to evaluate the performance of our algorithm. We test only the recognition performance. The gallery and probe sets are so assembled such that each probe image has a corresponding gallery image. When a probe image to be recognized is given to the face recognition system, it should pull out the correct gallery image corresponding to the given probe image. There are three steps in this phase. Firstly, the higher order features ($\mathbf{v}$) of the eyes for both gallery and probe sets are computed using gradient ascent using the density components ($\mathbf{B}$) and linear basis functions ($\mathbf{A}$) computed in the training stage. These features are then appended to the normalized raw pixel values. The new gallery and probe sets are then projected onto kernel space. Each face can be represented as a vector of weights by projecting it onto a set of eigenvectors or fisherfaces (**VFisher**) derived in the training stage. We have a separate set of weights for both gallery and probe sets. Each probe image is then compared against the gallery images (using the weights) using a Euclidean (nearest neighbor) similarity measure. The gallery image with the least Euclidean distance is pulled out to be the match.

5.4 Results

In this section, we discuss the performance of our feature extraction method, i.e., KFLD applied to normalized data augmented with Yan's hierarchical features (let's call it KFLD+v). To test the robustness of our algorithm, we assembled 20 training and test sets which are different random permutations of our entire dataset. The training set consists of 1390 images and the test set consists of 500 images. We used the Gaussian kernel with $sigma = 25$. We fixed the value of sigma to 25 after testing the performance of KFLD at all other sigma values. The following plot shows the average performance taken over the 20 sets.

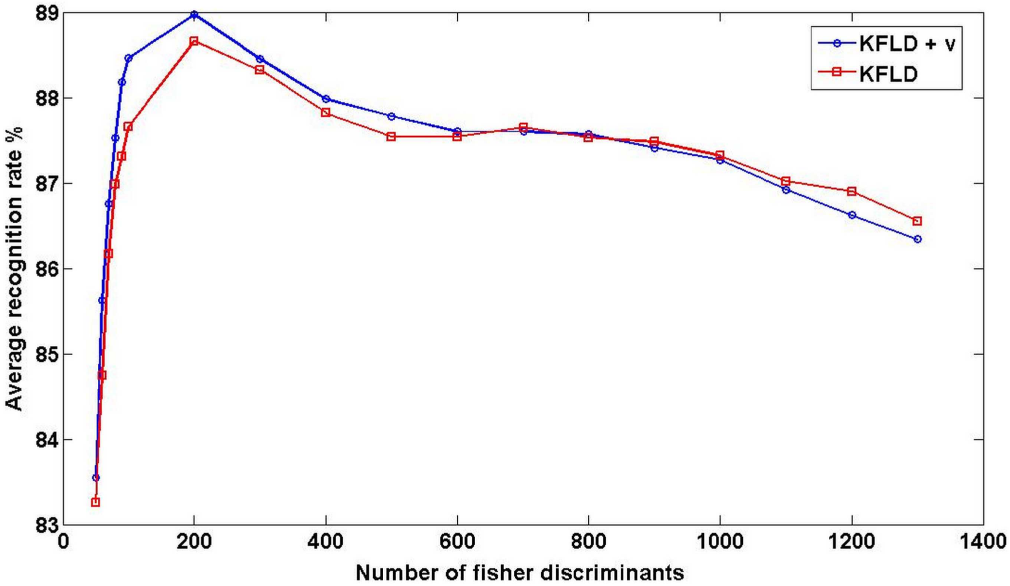

Figure 5.3. Plot showing the average recognition rate versus the number of principal components retained.

Clearly we see that our algorithm (KFLD+v) outperforms KFLD. KFLD+v shows an average maximum of 89% when compared to 88.6% of the basic KFLD while using just 200 principal components and just 8 additional features (4 for each eye).

The following plot shows the maximum performance improvement that was achieved by KFLD+v over basic KFLD.

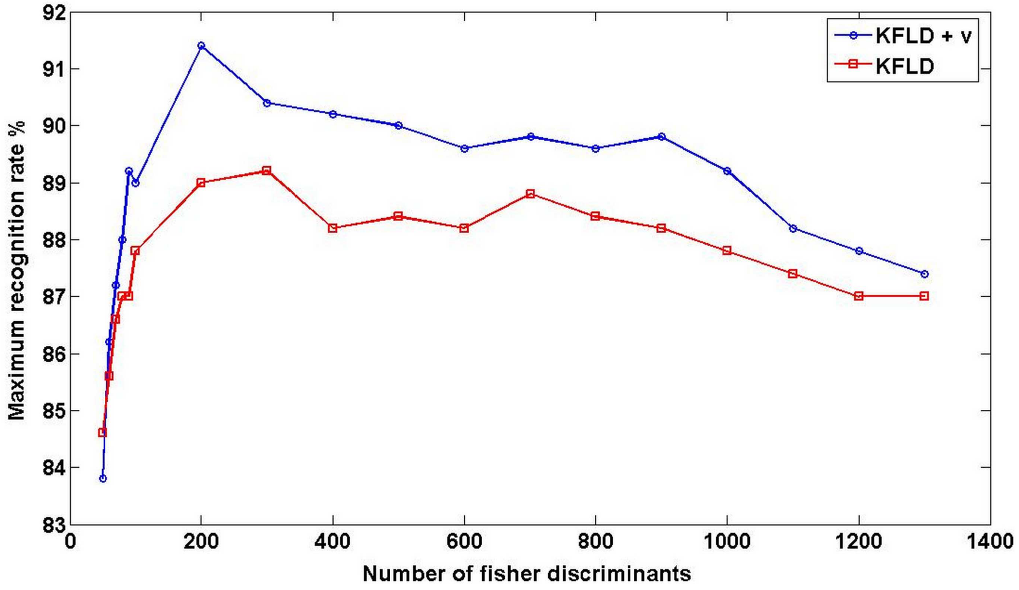

Figure 5.4. Plot showing the maximum recognition rate versus the number of principal components retained.

Adding the hierarchical features and then applying KFLD surely boosts the performance. We could see a maximum recognition rate of 91.4% as compared to a recognition rate of 89.2% for just KFLD. So a maximum improvement of over 2% was observed using KFLD + v.

The following plot shows the cumulative match scores. The vertical axis shows the average of the maximum recognition rate and the horizontal axis shows the rank.

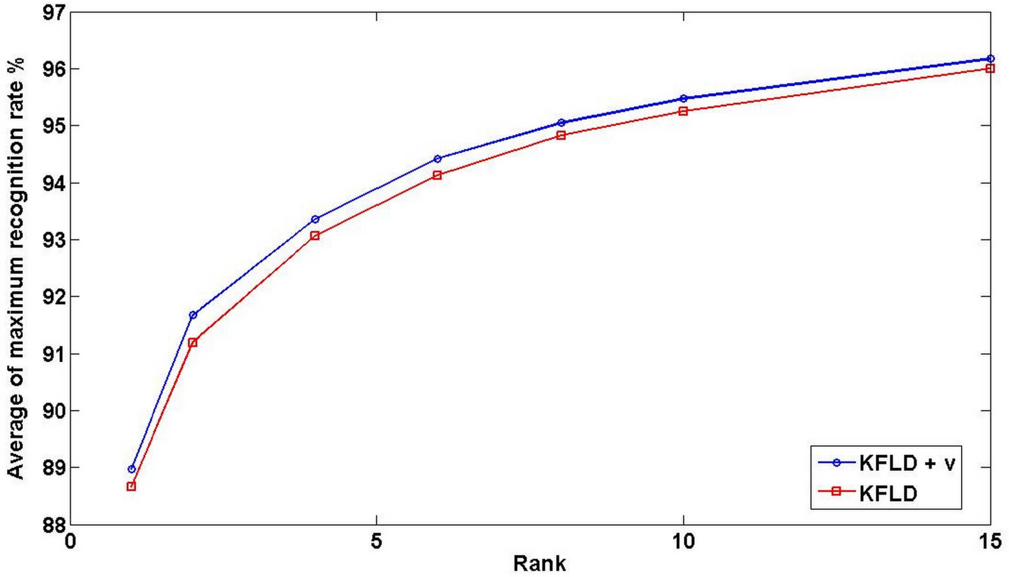

Figure 5.5. Plot showing the average of the maximum recognition rate versus the rank.

The plot shows the recognition rates for ranks 1 to 15. We can see that KFLD + v consistently outperforms KFLD.

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

Based on our extensive experiments and the results obtained, we have drawn the following conclusions:

- Substantial improvement in recognition rate could be achieved with the addition of just 8 new features (4 features for each eye). Hence we can say that, further boost in performance can be achieved if we include more features taken from other facial regions such as the mouth, nose etc.

- We could compute these hierarchical features with very little overhead in terms of computational time and memory.

- Humans are quite good at recognizing faces, but often have difficulty describing the features they use to discern them. Some features are simple and have straightforward physical interpretations (space between the eyes, etc), while others are more vague and sophisticated. These features are hard to describe, let alone simulate in a computer program. In order to be competitive with human capabilities, face recognition algorithms need to have a way of building up successively complex and sophisticated visual features, without relying on humans to communicate those features directly. Hierarchical graphical models are one promising way of doing that.

6.2 Future Directions

Following are the possible future avenues of research:

- Out of the 8 hierarchical features we compute, several of them may not be predictive and may be sensitive to unnecessary features due to variations in expression and illumination which may hinder recognition performance. Better recognition rates could be achieved using discriminative training approaches. Discriminative training is used to model the dependence of an unobserved variable (y) given an observed variable (x). This can be modelled using a conditional probability distribution $P(y|x)$. Generative models on the other hand model the joint probability distribution over all variables $P(x, y)$.

- Karklin's model has just one hierarchical layer atop ICA. We could add additional layers over this hierachical model and test whether we can obtain more useful features for recognition. Another technique which uses a similar layered approach is the Deep belief network. Deep Belief nets are probabilistic generative models that are composed of multiple layers of stochastic, latent variables. A deep belief is composed of simple learning modules each of which is a restricted Boltzmann machine that contains a layer of visible units that represent the data (images in our context) and a layer of hidden units that learn features that can capture higher-order structure in the data. Interested readers can refer to [12, 13, 18] for more details on Deep belief nets and Restricted Boltzmann Machines.

References

[1] http://www.cis.hut.fi/projects/ica/fastica/.

[2] http://www.frvt.org/.

[3] http://www.frvt.org/frgc/.

[4] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2003.

[5] M. Bartlett, J. Movellan, and T. Sejnowski. Face recognition by independent component analysis. 13(6):1450–1464, November 2002.

[6] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Comput.*, 12(10):2385–2404, 2000.

[7] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class-specific linear projection. 19(7):711–720, July 1997.

[8] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[9] F. De la Torre, R. Gross, S. Baker, and B. V. K. V. Kumar. Representational oriented component analysis (roca) for face recognition with one sample image per training class. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 266–273, Washington, DC, USA, 2005. IEEE Computer Society.

[10] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *Journal of Optical Society of America A*, 14:1724–1733, 1997.

[11] G. Guo, S. Z. Li, and K. Chan. Face recognition by support vector machines. In *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, page 196, Washington, DC, USA, 2000. IEEE Computer Society.

[12] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006.

[13] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006.

[14] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13(4-5):411–430, 2000.

[15] Y. Karklin and M. S. Lewicki. A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Comput.*, 17(2):397–423, 2005.

[16] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):103–108, 1990.

[17] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput.*, 42(3):300–311, 1993.

[18] H. Larochelle and Y. Bengio. Classification using discriminative restricted Boltzmann machines. pages 536–543, 2008.

[19] C. Liu. Gabor-based kernel pca with fractional power polynomial models for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(5):572–581, 2004.

[20] C. Liu and H. Wechsler. Evolutionary pursuit and its application to face recognition. *IEEE TRANS. PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 22(6):570–582, 2000.

[21] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Processing*, 11:467–476, 2002.

[22] B. Moghaddam. Principal manifolds and bayesian subspaces for visual recognition. *Computer Vision, IEEE International Conference on*, 2:1131, 1999.

[23] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, November 2000.

[24] H. Moon and P. Phillips. Computational and performance aspects of pca-based face recognition algorithms. *Perception*, 30(3):303–321, 2001.

[25] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. pages 84–91, 1994.

[26] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 947–954, Washington, DC, USA, 2005. IEEE Computer Society.

[27] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.

[28] P. J. Phillips, P. J. Phillips, S. Z. Der, P. J. Rauss, P. J. Rauss, O. Z. Der, and O. Z. Der. Feret (face recognition technology) recognition algorithm development and test results, 1996.

[29] S. A. Rizvi, P. J. Phillips, and H. Moon. The feret verification testing protocol for face recognition algorithms. In *FG '98: Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, page 48, Washington, DC, USA, 1998. IEEE Computer Society.

[30] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998.

[31] L. Shen and L. Bai. A review on gabor wavelets for face recognition. *Pattern Anal. Appl.*, 9(2):273–292, 2006.

[32] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[33] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775–779, 1997.

[34] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):131–137, 2004.

[35] M.-H. Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 215, Washington, DC, USA, 2002. IEEE Computer Society.

[36] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. pages 336–341, 1998.

[37] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.