PERFORMANCE EVALUATION METHODOLOGY FOR FACE RECOGNITION ALGORITHMS

by

Hyeonjoon Moon

A dissertation submitted to
the Faculty of the Graduate School of
the State University of New York at Buffalo
in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

June 1999

To my parents, Hyo-Shik Moon and Hyo-Sim Lee,
who gave me endless love,
to my wife Jyhyun, and my daughters Elaine and Rachel,
who inspired me to pursue my research.

Acknowledgments

I would like to express my gratitude to my advisor Dr. Peter D. Scott for his insightful guidance and suggestions on my dissertation. I would like to thank Dr. Nasser M. Nasrabadi for his constant suggestions, guidance and support throughout the duration of my research. Also, I would like to thank Dr. Mehrdad Soumekh, and Dr. Venugopal Govindaraju for serving my dissertation committee and their insightful comments on my research. I am very grateful to have Dr. P. Jonathon Phillips to be my outside reader, who have suggested a lot of valuable advice for the improvements of this dissertation.

There is no way I can begin to acknowledge the encouragement and love of my parents. I could not have finished the long journey of my Ph.D. work without their love and prayer. In recent years, my parents-in-law have also provided me invaluable supports. My lovely daughters, Elaine and Rachel, have enforced the necessary diversion for me to enjoy the non-academic world. I am especially blessed to have a lovely wife, Jyhyun Shin, to maintain a sweet home while I am tangled by the research works. They are certainly deserve my heartiest gratitude for their encouragement, patience and unfailing love.

Besides, I would like to thank a number of friends and family in Korea and United States, who have supported me by their love and prayer. Finally, I want to thank God who always provided and protected me. He is my best friend regardless of the mountain or valley of my life journey. Hence, this dissertation is dedicated to Him, the origin of life and wisdom.

Abstract

We present two fundamental performance evaluation methodologies for face recognition algorithms. Our experiments include (1) the development of an evaluation methodology based on an identification and verification model and (2) the investigation of design decisions for a principal component analysis (PCA) based face recognition system. Throughout the series of experiments, we present a robust and comprehensive evaluation methodology for face recognition algorithms that allows researchers to identify the relative strengths and weaknesses of their algorithms and that points out the directions for future research.

Two critical performance characteristics of face recognition algorithms are the identification and verification performance. We report performance results based on the identification and verification model for various face recognition algorithms. We identify the state of the art by direct quantitative assessment of different approaches. The results that we report are for images taken (1) on the same day, (2) on different days, (3) at least one year apart, and (4) under different lighting conditions.

PCA-based algorithms form the basis of numerous algorithms in the face recognition literature. PCA is a statistical technique and its incorporation into a face recognition system requires numerous design decisions. We explicitly state the design decisions by implementation of a generic modular PCA-based face recognition system. We make a comprehensive analysis of the different implementations for each module, as these affect the variations in performance.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Overview

Over the last decade, recognition of the human face from still and video images has been emerging as an active area of research [2, 45, 118]. For most humans, face recognition is not a difficult problem. However, the performance of computerized face recognition systems is another story, due to the large number of variations in facial appearance [9, 13, 42, 65] and the similarities of different faces. Therefore, face recognition presents a significant challenge and is one of the fundamental problems in computer vision and pattern recognition.

In addition to its importance to fundamental research, face recognition has numerous applications for surveillance, security, telecommunications, digital libraries, and human-computer interactions (for a survey of face recognition research, see Chellappa et al [22] and Samal et al [100]). A number of face recognition applications are being implemented in such areas as the control of access to restricted facilities or equipment, the credentialing of individuals for background and security checks, the monitoring of airports or border crossings, and the finding and logging of multiple appearances of individuals in surveillance videos. Other possible applications are for verifying identity at automatic teller

machines (ATMs) and matching photo identification records for fraud detection, including credit cards, passports, and driver's licenses.

Computer algorithms can serve as models for the human face recognition function [4, 6, 31, 58, 78, 110]. By directly comparing these models (algorithms) with human performance, one can assess which models are biologically plausible [51, 83, 114]. The closer the concordance between human and model performance, the greater the plausibility. However, the models need not be comprehensive; i.e., account for all aspects of face recognition. Rather, one can ascertain which properties of the human face processing system can be correctly modeled. So far, many face recognition systems have been implemented based on either the principal component analysis (PCA) approach (also known as the eigenfaces) or the dynamic link architecture (DLA) [61]. PCA encodes second order statistics of the face and can be enhanced using spatiotemporal constraints encoded as manifold trajectories that correspond to the views obtained as the face rotates in three-dimensional space. In DLA, elastic graph matching is attempted between locally derived forensic landmark grids, possibly encoded using Gabor wavelets [32, 123].

Solutions for the face recognition problem involve the segmentation of faces from cluttered scenes [68, 99, 107, 125] and the extraction of the features from the face [70, 73, 79, 82, 127] and their classification [10, 26, 60, 108]. The variability of applications poses a wide range of technical challenges at each stage of the recognition process. The accuracy of a face recognition algorithm is strongly affected by the limitations placed on the problem by image quality, cluttered backgrounds, lighting conditions, head rotations and scalings, facial expressions, and variations in appearance and partial occlusions. Therefore, the solutions for the face recognition problem have been synergetic efforts from fields such as signal processing [39, 91], pattern recognition [20, 36], machine learning [69, 74, 94, 101], neural networks [48, 53, 101], evolutionary computation [112, 117], neurosciences [58], and psychophysics studies of human

perception [12, 15, 16, 18, 55, 105, 126].

A new topic in face recognition is the interaction between predictive learning and performance evaluation. One has to develop both to assess performance on given data sets and to make predictions about future performance on unseen data sets. Statistical learning theory provides the means to estimate the guaranteed risk for testing on future facial imagery. This risk is formulated in terms of the empirical risk calculated during training and the complexity of the classification model underlying the face recognition system. Thus, performance depends on both the complexity of the classifier and the relative size and quality of the training versus test data sets. Also, one clearly has to develop standard databases and an evaluation methodology to assess and compare competing face recognition systems. Decision theory and receiver operating characteristic (ROC) provide the tools needed to quantify the level of performance displayed by specific face recognition systems.

1.2 Motivations for Research

There are two main categories of research to advance the state of the art in face recognition. The first category is the development of algorithms that can provide reliable solutions to face recognition problems. In algorithm development, a number of techniques have been proposed for preprocessing and enhancement [55, 56], detection of face and facial components in a scene [28, 29, 30, 43, 98, 107, 124], feature extraction algorithms [3, 26, 126], and classification techniques [67, 70, 92]. Not one of these procedures should be neglected, since each component is critical and performs as a part of the face recognition system.

The second category is the development of an evaluation methodology based on different scenarios and categories of images, or computational psychophysics studies using human performance. Recently, empirical evaluation techniques

have emerged as a serious research field in pattern recognition and computer vision. An empirical evaluation is defined as the development of a methodology for measuring the ability of algorithms to meet requirements for system level implementation. (See Table 1.1 for a list of papers that use empirical evaluation techniques for computer vision and pattern recognition.)

There are three fundamental approaches or categories in evaluation work. As with any classification, there is a risk that the categories will not necessarily be clean divisions. Evaluation work could belong to more than one category or not neatly fit into any category. However, the categorization provides insights that are useful for the development of an empirical evaluation of computer vision algorithms.

The first category is the problem of obtaining ground truth data where none are evident. Thus, a major component of the evaluation process is to develop a method of obtaining the ground truth data. The classic example of this problem is the development of evaluation methods for edge detectors. The question of what should be marked as an edge in a real image is often problematic. In this case, the human perception of an edge quality may be used, as it uses properties of an edge that are different from those needed for machine vision tasks.

The second category is the evaluation of a set of classifications by one group. The group wanting to do the evaluation will often not be able to get access to original implementations of all the algorithms of interest. Therefore, they have to implement some of the algorithms based on information in the literature. This introduces the possibility that the version of the algorithm evaluated will not be identical to the original developer's algorithm. However, implementation and evaluation of a set of algorithms by one group can at least establish performance for a baseline algorithm. Comparing an algorithm against such a baseline allows for an initial assessment.

The third category is the independent administration of evaluations. In an

Table 1.1: Empirical evaluation works for computer vision and pattern recognition research.

Authors [reference]	Applications	Description
Blue et al [11]	Fingerprint matching OCR applications	Pattern classifiers
Cho et al [23]	Edge detection	Bootstrap, perturbation strategy
Demigny and Kamle [33]	Edge detection, edge operators	Canny criteria, localization criterion
Heath et al [49]	Edge detector, variance analysis	Low-level processing, human rating
Hong and Jain [52]	Fingerprint matching, face recognition	Decision fusion, eigenface, minutiae
Jain and Zongker [54]	Feature selection, SAR image classification	Genetic algorithm, node pruning, texture models
Lindenbaum [62]	Object recognition, pose estimation	Localization, noise models, similarity measures
Lopez et al [64]	Ridge detection, valley detection	Comparative analysis, drainage patterns
Moon et al [71, 72]	Face recognition	Principal component analysis, nearest neighbor classifier
Randen and Husoy [92]	Texture classification	Filtering approaches
Shufelt [102]	Monocular building extraction	Comparative analysis, feature delineation
Shufelt [103]	Object recognition, building detection	Vanishing points, photogrammetry
Zhao et al [129]	Linear discriminant analysis	Classifiers

independent evaluation, one group collects a set of images, designs an evaluation protocol, provides images to the testees, and evaluates the results. This method is most desirable, since all algorithms are tested on the same assumptions, images, and scoring method. Independent evaluation by a noncompetitor gives a greater sense of impartiality to the results. In this method, the key point of the success is that the evaluation mechanism needs to be evolved and refined over time. Our research is based on this category, since this method provides a high degree of standardization and allows direct comparison between competing approaches. Additionally, independent evaluation helps to assess the state of the art and point out directions for future research.

In this dissertation, we present two major performance evaluation methodology for face recognition algorithms. Our research is focused on implementation and investigation of algorithm development as well as development of an evaluation methodology for face recognition. We did a number of experiments that included (1) the development of an evaluation protocol based on an identification and verification model and (2) the implementation of a principal component analysis (PCA) based face recognition system and the investigation of design decisions. The primary objectives of our research included the establishment of a standardized evaluation methodology for face recognition, the assessment of the state of the art in face recognition, and the presentation of a design methodology to identify future areas of research.

Two critical performance characteristics of face recognition algorithms are the identification and verification performance. In most face recognition literature, a large number of algorithms have reported outstanding recognition results (usually better than 95% correct identification) on relatively small databases (usually fewer than 50 individuals). To date, direct comparison was impossible between competing algorithms since the results were reported using different assumptions, databases, and evaluation methods. We reported performance results for various face recognition algorithms by developing a new evaluation

Table 1.2: PCA works for face recognition research.

Authors [reference]	Technique	Database
Abdi, Valentine, Edelman, and O'Toole [1, 113, 114]	Linear neural networks radial basis function	320 images of Japanese, Caucasian
Belhumeur, Hespanha, and Kriegman [8]	Fisher's linear discriminant, illumination invariance	Harvard Robotics Lab. 330 images of 5 people
Craw, Costen, and Kato [27]	Shape-free PCA, caricaturing	387 images of 27 people 14 images per person
Etemad and Chellappa [35, 128]	Discriminant eigenfeatures, evidential reasoning	FERET database (DB) 2000 images
Georghiades, Kriegman, and Belhumeur [38]	Illumination cone, generalized bas-relief transform	Harvard Robotics Lab. 660 images of 10 people
Hancock, Burton, and Bruce [46, 47]	Shape-free PCA, graph matching system	Aberdeen frame face DB 186 images of 50 people
Kirby and Sirovich [59, 106]	Karhunen-Loeve expansion, symmetric eigenfunctions	100 images (in the ensemble) 200 (extended ensemble)
Liu and Wechsler [63]	Probabilistic reasoning models, Bayes classifier, MAP classifier	FERET database 1107 images of 369 people
Moghaddam and Pentland [69]	Density estimation, maximum likelyhood	MIT Media Lab. 7562 mug shots of 3000 people
Penev and Atick [79]	Local feature analysis	FERET database
Pentland, Moghaddam, and Starner [81]	View-based method, parametric method	MIT Media Lab. 7562 mug shots of 3000 people
Swets and Weng [109]	Discriminant analysis, feature selection	Weizmann Institute Face DB 1614 images of 802 classes
Turk and Pentland [111]	PCA (eigenface)	MIT Media Lab. 2500 images of 16 people
Wilder, Phillips, Jiang, and Wiener [121]	Gray-scale projection, matching pursuit filter	1212 Infrared images of 101 people (without glasses)

methodology based on the identification and verification model. We identified the state of the art by direct quantitative assessment of different approaches. The results were reported for images taken on the same day, on different days, at least one year apart, and under different lighting conditions.

PCA-based algorithms form the basis for numerous algorithms and studies in the face recognition research. (See Table 1.2 for a list of papers that have used PCA-based techniques for face recognition.) PCA is a statistical technique and its incorporation into a face recognition system requires numerous design decisions. We explicitly state the design decisions by introducing a generic modular PCA-based face recognition system.

We performed two main experiments and report the results using identification and verification performance based on the standard set of facial images. In the first experiment, we presented comprehensive analysis of different implementations for each module that affect variations in performance. We explored the variations of the algorithm performance by changing the illumination normalization procedure, studying the effects of image compression using Joint Photographic Experts Group (JPEG) [80] and wavelet [116] techniques, varying the number of eigenvectors in the representation, and changing the distance measure in the classification process. In the second experiment, we examined variations in algorithm performance by computing algorithm performance on 100 randomly generated image sets of the same size.

Throughout the series of experiments, we present a robust and comprehensive evaluation methodology for face recognition algorithms. Our evaluation methodology allows researchers to identify the algorithms' relative strengths and weaknesses by direct quantitative assessment of different approaches and our evaluation points out the directions for future research.

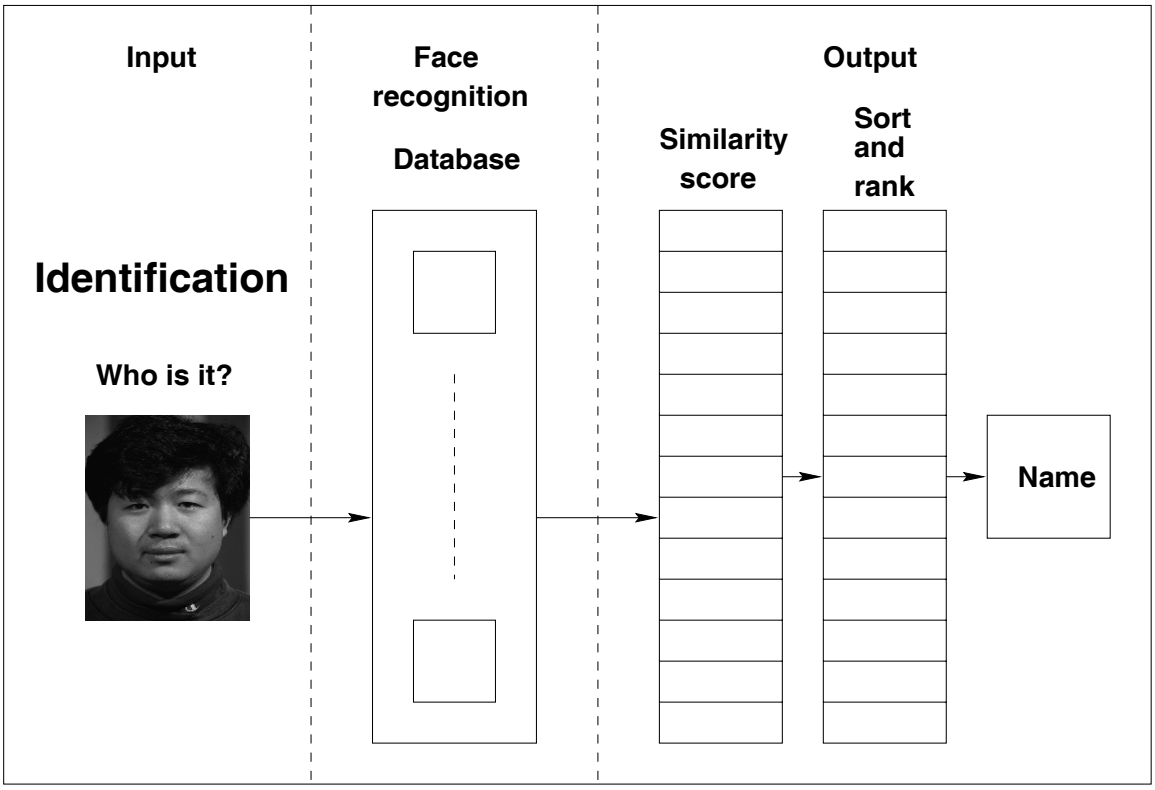


Figure 1.1: Identification scenario for face recognition.

1.3 Face Recognition Scenarios

A general face recognition task can be defined as identification or verification of one or more persons from still or video images. In most face recognition literature, the results have been reported to a single identification performance measure for a database of images; i.e., on database X, algorithm A correctly identifies faces n percent of the time (or more generally, a single probability of identification curve for database X). This implies that the identification performance on a single database is predictive of verification performance. In an identification application, an algorithm is presented with a face that it must identify. Meanwhile, in a verification application, an algorithm is presented with a face and a claimed identity, and the algorithm must accept or reject the claim. Since identification and verification are different problems, they need separate evaluation techniques.

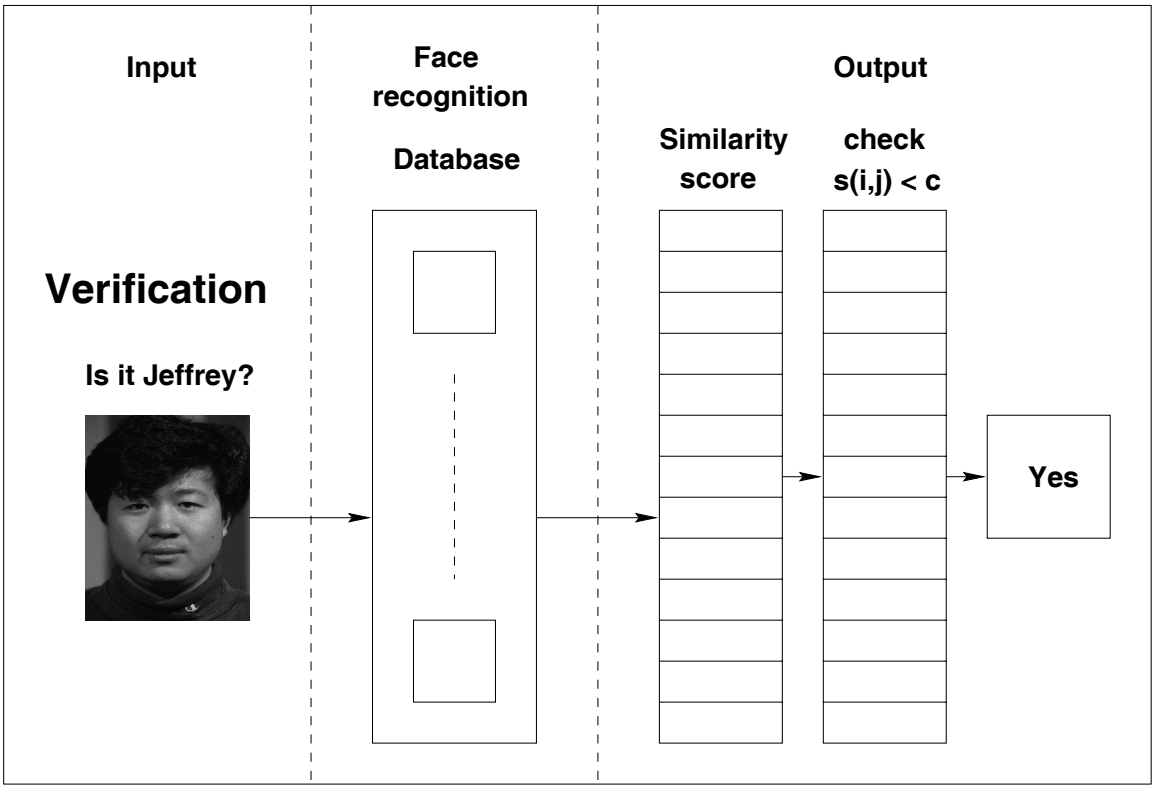


Figure 1.2: Verification scenario for face recognition.

In an identification scenario, the input is an image of an unknown individual (a probe) that is presented to a face recognition algorithm. The algorithm reports the closest matches from a collection of images of known individuals (a gallery) (see Figure 1.1). The performance of the algorithm is measured by its ability to correctly identify the person in the probe image. For example, an unknown facial image from a surveillance video would be a probe, and the face recognition system would display the images of the 100 people from the gallery in the database that most resembled the unknown individual. A possible application for this scenario would be to search electronic mug shots for the identity of a suspect.

In a verification scenario, the input is a face with a claimed identity. The face recognition algorithm either accepts or rejects the claimed identity based on the similarity score $s(i, j)$ and threshold value c (see Figure 1.2). In this case, the im-

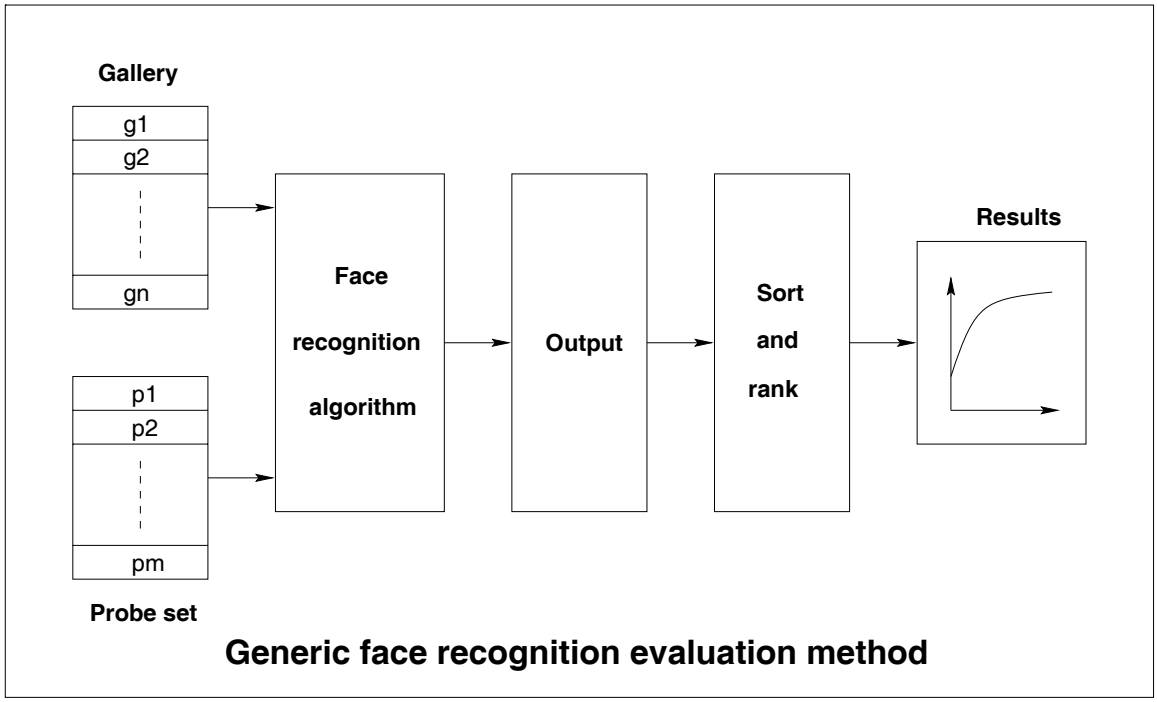


Figure 1.3: Schematic diagram for face recognition evaluation process.

portant system performance measures are the probabilities of false alarms and missed recognitions. A false alarm occurs when the algorithm reports that the person in a probe image is in the gallery when that person is not in the gallery. A missed recognition is when the algorithm reports that the person in the probe is not in the gallery when that person is in the gallery, or the algorithm identifies that person as the wrong person. The applications for this scenario include use with an automatic teller machine (ATM), the verification of identities for a passport or driver's license, or the control of access to buildings and computers. In the access control applications [60], when an individual walks up to a doorway, his or her image is captured, analyzed, and compared with the gallery of individuals approved for access. Alternatively, the system could monitor points of entry into a building or an airport, and search for terrorists or other criminals attempting to enter surreptitiously.

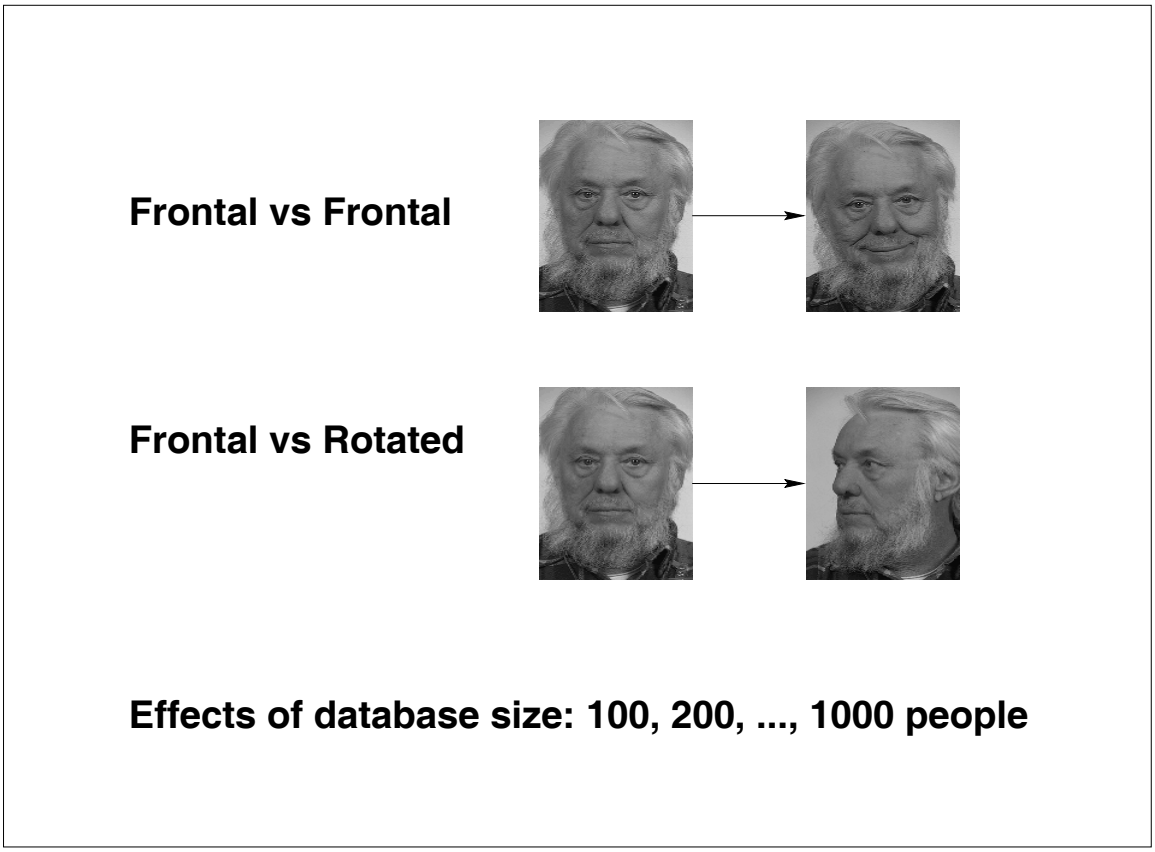


Figure 1.4: Examples of evaluation issues for face recognition.

1.4 Evaluation Issues for Face Recognition

We presented a schematic diagram of a general evaluation procedure for face recognition algorithms (see Figure 1.3). For each probe, the algorithm calculates the similarity score $s(i, j)$ for the gallery presented to the face recognition system. The similarity scores are sorted and ranked in proper order to generate performance results.

For the successful implementation of face recognition algorithms, it is crucial that the algorithm performance for intended applications can be estimated or predicted from known performance results. These predictions are used in making decisions in designing a demonstration system. To be able to make predictions, the performance evaluations for identification and verification sce-

narios and the characterization of performance in terms of the variability of the images in the database and the database size are necessary.

One of the key issues in face recognition evaluation is how does the size of the database affect performance (see Figure 1.4 for examples of evaluation issues for face recognition). The effect of database size is important because algorithms are usually developed and initially evaluated on databases that are smaller than the databases encountered in applications. To be able to make an intelligent choice of which algorithms are appropriate for an application, one would like to be able to predict performance on larger databases. We investigate the effects of database size and composition on identification and verification performance. Further questions can be raised about the age, gender [1, 14, 17, 21, 76, 77, 122], and race [75] distribution of the database.

1.5 Outline of the Dissertation

This dissertation is organized as follows. Chapter 2, The Face Recognition Technology (FERET) Program, is an overview of the FERET program. We describe the FERET program and give details of the FERET testing history, database, and the evaluation procedure.

Chapter 3, Performance Evaluation Methodology for Face Recognition Algorithms, presents research goals and the originality and contributions of the research, followed by a discussion of decision theory and design principles. We give a more detailed description of the major differences between the efforts for the old FERET test (performed in August 1994 and March 1995) and the new FERET test (performed in September 1996 and March 1997) based on our new evaluation methodology.

Chapter 4, An Identification Model for Face Recognition Algorithms, focuses on the identification model and reports the identification results for the performance of partially and fully automatic algorithms. A more detailed analysis and

discussion are given based on the variation in identification performance.

Chapter 5, A Verification Model for Face Recognition Algorithms, focuses on the verification model and reports the verification results on the performance of partially and fully automatic algorithms. A more detailed analysis and discussion are given based on the variation in verification performance.

Chapter 6, Analysis of PCA-Based Face Recognition Algorithms, describes a design methodology based on the key principles identified in chapters 4 and 5. The system was implemented with a modular design concept that uses normalization of images, feature selection by PCA representations of images, and a recognition process based on nearest-neighbor classifiers [39]. We explored a comprehensive analysis of the system components that includes the normalization, feature extraction, and classification. Fundamental problems that are specific to the recognition of human faces are addressed, and some solutions are proposed.

Chapter 7, Conclusions, summarizes the performance evaluation methodology, addresses the originality and contributions of research, and identifies areas for further face recognition research.

In the Appendix, we define terms used in these chapters and present mathematical representations of the fundamental techniques for our PCA-based face recognition system including histogram equalization, generation of the eigenface, and nearest-neighbor classifier.

Chapter 2

The Face Recognition Technology (FERET) Program

2.1 Overview of the FERET Program

Two critical requirements are necessary to support a reliable face recognition system. These include the collection of a large database of facial images and a standardized testing procedure. The FERET[1] program was designed to address both requirements through the collection of a large database of facial images and an independently administered testing procedure [85, 86, 87, 88, 93]. The FERET program has focused on three major tasks, which include (1) the collection of a large database of facial images and testing, (2) the development of a performance evaluation methodology for face recognition algorithms, and (3) the investigation of the technology base for a face recognition system.

In this dissertation, the main focus of our research includes the development of a new evaluation methodology for face recognition algorithms and the assess-

[1]The FERET program is sponsored by the Department of Defense Counterdrug Technology Development Program through the Defense Advanced Research Projects Agency (DARPA), with the U.S. Army Research Laboratory (ARL) serving as technical agent.

ment of the technology by implementation and investigation of a PCA-based face recognition system.

2.2 The FERET Testing History

To date, two major FERET tests have been conducted, which include (1) a test of the initial development phase of face recognition algorithms in August 1994 and March 1995 (which we will collectively call the *old test*), and (2) a test based on a new evaluation methodology for the face recognition algorithms evaluated in September 1996 and March 1997 (which we will collectively call the *new test*). The goals of these efforts were to measure overall progress in face recognition to assess the state of the art in face recognition, determine the maturity of face recognition algorithms from leading researchers, and develop an independently administered standardized evaluation method for face recognition algorithms.

The August 1994 test established an initial performance baseline for face recognition algorithms that could automatically locate, normalize, and identify faces from a database. The participants for the August 1994 test were the Massachusetts Institute of Technology (MIT), Rutgers University, the University of Illinois, the University of Southern California (USC), and The Analytic Science Company (TASC).

The March 1995 FERET test measured progress since August 1994 and evaluated the performance of algorithms on larger galleries. One emphasis of the March 1995 test was on probe sets that contained duplicate images. A set of images is referred to as a *duplicate* set if the person in the set is in a previously collected set (see Figure 2.1). The March 1995 test was significantly more difficult, since the number of duplicates increased from 60 to 463. For the March 1995 test, the participants were MIT, USC, and the Laboratory of Computational Neuroscience at Rockefeller University. When the performances

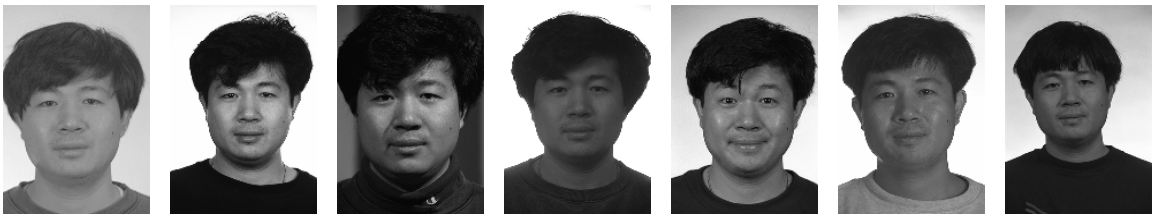


Figure 2.1: Examples of variations between collections (duplicate images).

of the August 1994 and March 1995 FERET tests were directly compared, absolute performance did not drop despite the increased difficulty of the March 1995 test. These results showed that the March 1995 test performance was an improvement over the August 1994 test (for details of these tests results and the FERET database, see Phillips et al [86, 88] and Rauss et al [93]).

To be able to characterize the performance of the two portions of the old test and open up the testing procedure to more algorithms, we devised two new test versions; the September 1996 and March 1997 tests (see Table 2.1 for details of the new test). We report the results for 12 algorithms for the new test, including 10 partially automatic algorithms and 2 fully automatic algorithms. Two of these algorithms were developed at the MIT Media Laboratory. The first was the same algorithm that was tested in March 1995. This algorithm was retested so that improvements made to it since March 1995 could be measured. The second algorithm was based on more recent work [66, 69]. Algorithms were also tested from Excalibur Corp. (Carlsbad, CA), Michigan State University (MSU) [109, 130], Rutgers University [119], the University of Southern California (USC) [123], and the University of Maryland (UMD) [35, 128, 130]. The algorithm from UMD was first tested in September 1996 and a modified version was tested in March 1997. For the fully automatic version of the test, algorithms from MIT and USC were evaluated.

The final two algorithms were our implementation of a normalized correlation and a principal components analysis (PCA) based algorithm [71, 111]. These

Table 2.1: List of algorithms that took the September 1996 test, broken out by versions taken and dates administered. (Note: MIT tested two algorithms.)

		Test Date		
Version of test	Algorithms	September 1996	March 1997	Baseline
Partially automatic (Eye coordinates given)	Baseline PCA [71, 111]			•
	Baseline Correlation			•
	Excalibur Corp.	•		
	MIT Media Lab	•(2)		
	MSU [109, 130]	•		
	Rutgers Univ. [119]	•		
	UMD [35, 128, 130]	•	•	
	USC		•	
Fully automatic	MIT Media Lab [66, 69]	•		
	USC [123]		•	

algorithms provide a performance baseline. In our implementation of the PCA-based algorithm, all images were (1) translated, rotated, and scaled so that the center of the eyes were placed on specific pixels; (2) faces were masked to remove background and hair; and (3) the nonmasked facial pixels were processed by a histogram equalization algorithm. The training set contained 500 faces from the development set of the FERET database. Faces were represented by their projection onto the first 200 eigenvectors and were identified by a nearest-neighbor classifier using the L_1 metric. For normalized correlation, the images were (1) translated, rotated, and scaled so that the center of the eyes were placed on specific pixels and (2) faces were masked to remove background and hair.

2.3 The FERET Database

The FERET database was established on the assumption that the evaluation of face recognition algorithms requires a common database of images for both development and testing. For the evaluation procedure to produce meaningful results, the images in the developmental database must resemble those on which algorithms are to be tested. The development and testing data sets must be similar in both quality and quantity. For example, if the test will consist of images of 1000 individuals, it is not appropriate for the development database to consist of 50 individuals. The algorithms tested will be only as good as the database from which they are developed. The FERET database has fulfilled the data requirements for both development and testing and has become the de facto standard for face recognition from still images [84, 88].

Before the existence of the FERET database, most research efforts addressed the results that came from the use of small databases that were developed under highly controlled conditions. Since the FERET database was developed to address a real-world problem, it was created to be more realistic, although still providing a semicontrolled environment over the type and nature of the images collected. The FERET database consists of two parts. The *development* portion is given to researchers while the *sequestered* portion is not released but is used to test the researchers' face recognition algorithms. The images in the development set are representative of the sequestered images. The sequestered images allow face recognition algorithms to be evaluated on images that they have not been exposed to before.

For the FERET database, the facial images were produced with a 35-mm camera with Kodak's color ultra film. These images were processed onto a CD-ROM by Kodak's multiresolution digital image technique. The facial images were created by retrieving the color images from the CD-ROM and converting them into 8-bit gray scale and tagged-image file format (TIFF) images. The dimensions

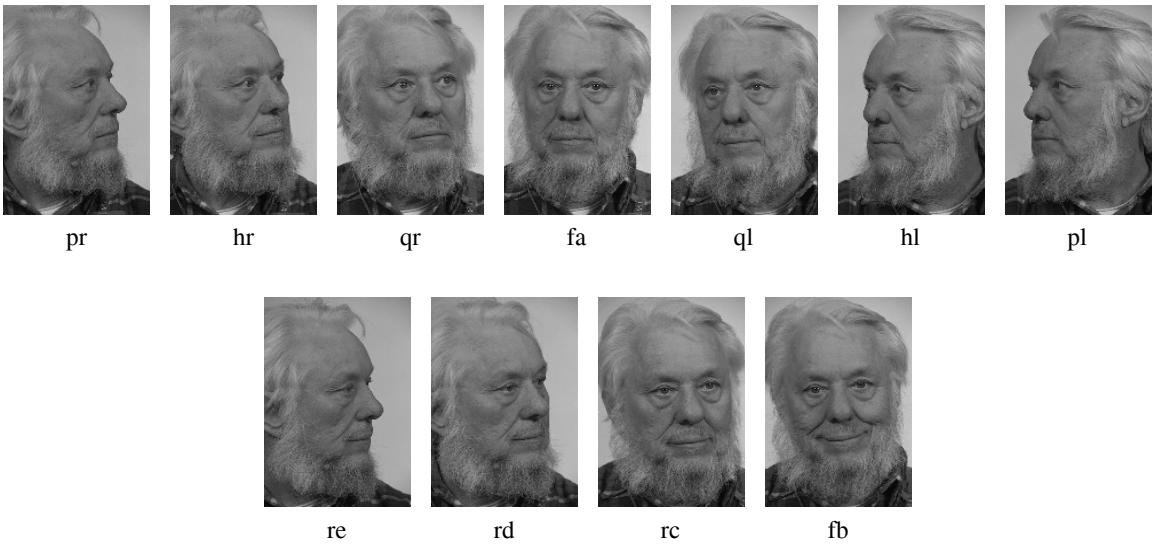


Figure 2.2: Typical set of images of one individual collected in one sitting.

of these images are 256 pixels wide and 384 pixels high. Each image was given a unique file name that encodes the image ground truth data. The fields include (1) the subject's identity, affixed to a five digit number; (2) the pose of the image, identified by two characters; (3) the date that the image was taken, given in a six-digit format; and (4) special variations flags. A fixed identity number was given for any one person so that any future images of that person would have the same identity number.

The facial images were collected in 15 different sessions between August 1993 and July 1996. To maintain a degree of consistency throughout the database, the same physical setup and location were used in each photography session. However, there were variations from session to session since the equipment had to be reassembled for each session. Images of an individual were acquired with the person in different poses and placed in sets of 5 to 11 images under relatively unconstrained conditions (see Figure 2.2). Some images of people in the database span nearly a year between the first sitting and the most recent one (see Figure 2.3).

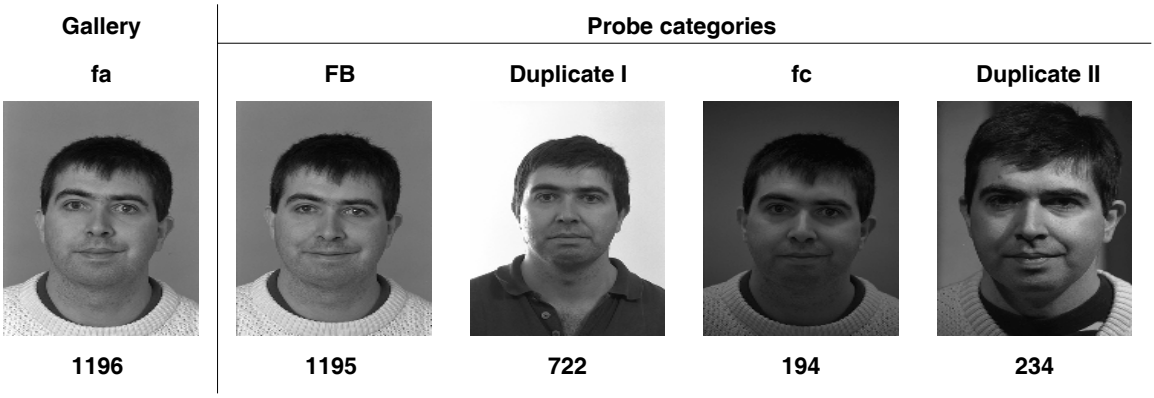


Figure 2.3: Examples of different categories of probes with number of images used. The duplicate I image was taken within one year of the **fa** image and the duplicate II and **fa** images were taken at least one year apart.

Images were taken at two frontal views (fa and fb), with a different facial expression requested for the second frontal view. For 200 sets of images, a third frontal view was used with a different camera and different lighting (this is referred to as the fc image). The remaining images were collected at six different aspects: right and left profile (pr, pl), right and left quarter profile (qr, ql), and right and left half profile (hr, hl). Additionally, images at five irregularly spaced extra locations (ra, rb, rc, rd, and re) were collected for some individuals. Some individuals were asked to put on their glasses or add some simple but significant variation to the images.

For the August 1994 test, the gallery consisted of 317 individuals among 673 sets of images with 5,000 total images from the FERET database. This required numerous collection activities and a large-scale effort to catalogue the images into a database. This database has been released to at least 50 different research groups for the development and performance evaluation of their algorithms. For the March 1995 test, the gallery consisted of 831 individuals among 1,109 sets of images with 8,525 total images.

By July 1996, 1,564 sets of images were in the database, for 14,126 total images. To support the September 1996 and March 1997 tests, an additional 456

sets of images were collected for the FERET database. Currently, the database contains files of 1,199 individuals and 365 duplicate sets of images. For some people, over two years elapsed between their first and most recent sittings, with some subjects being photographed multiple times. There were 91 duplicate sets where the time between the first and last sittings was at least 18 months. The development portion of the database, which consists of 503 images, was released to researchers. The remaining images were sequestered for testing by the federal government.

2.4 The FERET Testing Procedure

The FERET testing procedure was designed to establish a standardized evaluation methodology for face recognition algorithms. The FERET test measured performance of the face recognition algorithms, but was not concerned with the speed of the implementation, real-time implementation issues, and any speed and accuracy trade-offs. These issues need to be addressed in a fielded system and they were beyond the scope of the FERET test.

In figure 2.4, we present a schematic diagram of the FERET test procedure. The FERET test was administered at each research group's site under the supervision of a government representative. The processing time or number of workstations used are not taken into account because execution times can vary according to the machines used, network configuration, and the amount of time that the developers spent optimizing their code. These factors should be considered for the development of face recognition algorithms that could be incorporated into fieldable systems.

The images in the gallery and probe sets were selected from both the developmental and sequestered portions of the FERET database. Only images from the FERET database were included in the test. However, algorithm developers were not prohibited from using images outside the FERET database to develop or

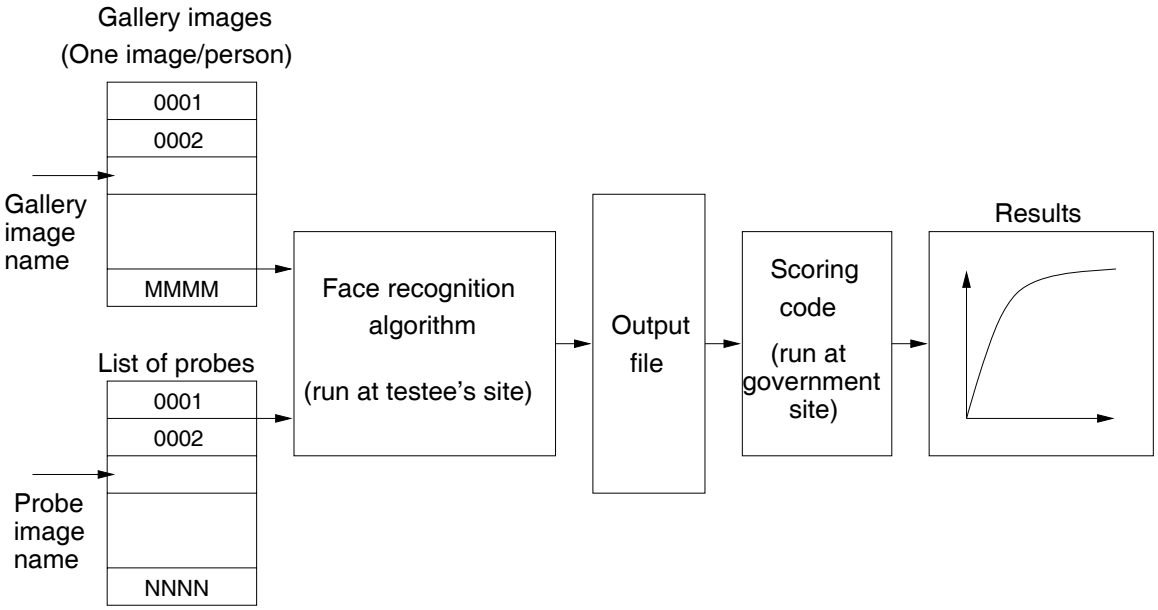


Figure 2.4: Schematic diagram of the FERET testing procedure.

tune parameters in their algorithms. To ensure that the matching was not done by file name with ground truth information, we generated random file names for the images given to the researchers for testing. The nominal pose of each face was provided to the researchers. The results of the test were recorded on 8-mm computer data tape. Finally, each test was processed by scoring code in government facilities and the results were presented based on each test category based on different scoring code.

For the old FERET test, three major tests were conducted based on three different gallery and probe sets. The first test is the large gallery test, which measured the ability of the algorithms to handle large databases. The differences between a probe image and a gallery image of a person include changes in time, scale, illumination, and pose. The first test examined the ability of algorithms to recognize faces from a gallery of 316 individuals. The second was the false-alarm test, which measured how well an algorithm rejects faces not in the gallery. The goal of the false-alarm tests is to see if an algorithm can successfully differentiate between probes that are in the gallery and those not in the gallery. The third test

developed a baseline for effects of pose changes on the algorithm's performance. A small set of rotation tests have been investigated to provide a baseline for pose variation. For each test, different gallery and probe sets have been created.

As part of the FERET program, our major focus was on the development of a new evaluation methodology that could overcome the limitations of the old test procedure. The primary objectives of our research based on a new evaluation methodology include (1) the development of a standardized evaluation methodology for face recognition, (2) the measurement of the progress in face recognition since the September 1994 and March 1995 tests, (3) the identification of state of the art by direct assessment of the competing face recognition algorithms, and (4) the identification of directions for future face recognition research.

Chapter 3

Performance Evaluation Methodology for Face Recognition Algorithms

3.1 Research Goals

The main goals of our research are (1) the development of a flexible and robust evaluation methodology for face recognition algorithms and (2) the assessment of the technology base by implementation and investigation of a PCA-based face recognition system. The availability of the FERET program database and testing procedure has made a significant difference and has served as a basis for our research and development of a face recognition system.

Our evaluation methodology provides a comprehensive picture of the state of the art in face recognition technology. This picture was created by the flexibility of our new evaluation methodology for different scenarios, categories of images, and versions of algorithms. The performance of face recognition algorithms could be further investigated without having a new set of tests for both

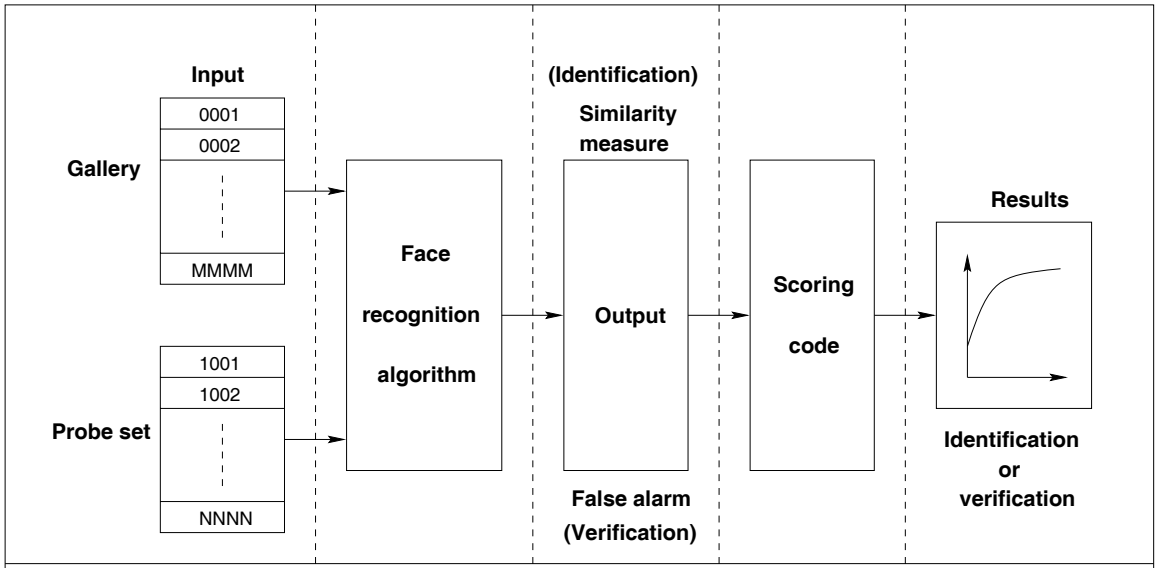
identification and verification scenarios. Also, we investigated a number of design decisions for a PCA-based face recognition system. These design decisions point out critical requirements for a fieldable face recognition system.

3.2 Originality and Contributions

The originality and contributions of our research were addressed throughout the series of experiments and in our comprehensive evaluation methodology. Before the FERET test, there was no method to evaluate or compare the competing face recognition algorithms. Various researchers collected their own databases images (often of fewer than 50 individuals) under conditions relevant to the aspects of the problems that they were examining. The independently administered FERET testing is based on a standard database and evaluation method. Thus, researchers could investigate the strengths and weaknesses of their algorithm by direct assessment with other competing algorithms.

Since our evaluation methodology is developed based on the same assumptions, database, and evaluation method, it is possible for researchers to report results by direct comparison among competing algorithms. More importantly, we clarify the state of the art in face recognition and identify general directions for future research. Our evaluation methodology allows the face recognition community to assess overall strengths and weaknesses in the field. The evaluation methodology allows comparison not only on the basis of the performance of an individual algorithm, but also on the aggregate performance of all algorithms tested. Through this type of assessment, the research community learns in an unbiased and open manner of the important technical problems to be addressed and the progress that is being made toward solving them.

We have presented a new evaluation methodology based on an identification and verification model (see Figure 3.1 for the comparisons of the features between the old and new FERET tests). Our evaluation methodology was de-



For: n test scenarios	Test categories	Flexiblility	Further investigation	New scenarios
Old: Need n different gallery and probe sets	Frontal, rotated false alarm	No, need each for ide. and ver.	Impossible	Re-test
New: Only one sets of gallery and probe set	fa, fc, dupl, dupII ROC, EER, etc.	Yes, one output for ide. and ver.	Possible	Can be examined without re-test

Figure 3.1: Comparisons of the features between old FERET test and new test based on our new evaluation procedure.

signed with flexibility to measure algorithm performance for both identification and verification tasks. The FERET database and testing procedure make it possible to independently evaluate face recognition algorithms. Because our new evaluation protocol has the ability to test an algorithm's performance on different tasks for multiple galleries and probe sets, it became the de facto standard for measuring the performance of face recognition algorithms. The results and analysis of the identification and verification performance scores are presented in Chapters 4 and 5. Also, we have implemented a PCA-based face recognition system and investigated a number of design decisions that show variations of different approaches. The comprehensive analysis of the results for the design decisions are presented in Chapter 6. These results address the critical factors that affect the performance of face recognition systems.

3.3 Decision Theory

There are many situations in which we have to make decisions based on observations or data that are random variables. As a random phenomenon, the combinations of signals and noise must be described statistically and analyzed in the framework of the theory of probability. The theory behind the solutions for these situations is known as *decision theory* or *hypothesis testing*. In communication or radar technology, decision theory or hypothesis testing is known as (signal) detection theory [50]. We present fundamental concepts of the conditional probability, binary decision theory, and various decision tests.

3.3.1 Conditional Probability

A. Definition

The *conditional probability* of an event A given event B, denoted by $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) > 0 \tag{3.1}$$

where $P(A \cap B)$ is the joint probability of A and B. Similarly,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad P(A) > 0 \tag{3.2}$$

is the conditional probability of an event B given event A. From Equations 3.1 and 3.2, we have

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{3.3}$$

Equation 3.3 is often quite useful in computing the joint probability of events.

B. Bayes' Rule

From Equation 3.3, we can obtain the following *Bayes' rule*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.4}$$

3.3.2 Hypothesis Testing

A. Definitions

A *statistical hypothesis* is an assumption about the probability law of random variables. Suppose we observe a random sample $(X_1, ..., X_n)$ of a random variable X whose probability density function (PDF) $f(\mathbf{x};\theta) = f(x_1, ..., x_n;\theta)$ depends on a parameter θ. We wish to test the assumption $\theta = \theta_0$ against the assumption $\theta = \theta_1$. The assumption $\theta = \theta_0$ is denoted by H_0 and is called the *null hypothesis*. The assumption $\theta = \theta_1$ is denoted by H_1 and is called the *alternative hypothesis*.

$$H_0 : \theta = \theta_0 \text{ : (null hypothesis)},$$

$$H_1 : \theta = \theta_1 \text{ : (alternative hypothesis)}.$$

A hypothesis is called *simple* if all parameters are exactly specified. Otherwise, it is called *composite*. Thus, suppose $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$; then H_0 is simple and H_1 is composite.

B. Hypothesis Testing and Types of Errors

Hypothesis testing is a decision process that establishes the validity of a hypothesis. We can think of the decision process as dividing the observation space R^n (Euclidean n-space) into two regions, R_0 and R_1. Let $\mathbf{x} = (x_1, ..., x_n)$ be the observed vector. Then if $\mathbf{x} \in R_0$, we will decide on H_0; if $\mathbf{x} \in R_1$, we decide

on H_1. The region R_0 is known as the *acceptance region*, R_1 is known as the *rejection* (or *critical*) *region*, since the null hypothesis is rejected. Thus, with the observation vector (or data), one of the following four actions can happen:

1. H_0 true; accept H_0.
2. H_0 true; reject H_0 (or accept H_1).
3. H_1 true; accept H_1.
4. H_1 true; reject H_1 (or accept H_0).

The first and third actions correspond to correct decisions, and the second and fourth actions correspond to errors. The errors are classified as

1. Type I error: Reject H_0 (or accept H_1) when H_0 is true.
2. Type II error: Reject H_1 (or accept H_0) when H_1 is true.

Let P_I and P_{II} denote, respectively, the probabilities of Type I and Type II errors:

$$P_I = P(D_1|H_0) = P(\mathbf{x} \in R_1; H_0) \tag{3.5}$$

$$P_{II} = P(D_0|H_1) = P(\mathbf{x} \in R_0; H_1) \tag{3.6}$$

where $D_i (i = 0, 1)$ denotes the event that the decision is made to accept H_i. P_I is often denoted by α and is known as the *level of significance*, and P_{II} is denoted by β, and $(1 - \beta)$ is known as the *power of the test*. Note that since α and β represent probabilities of events from the same decision problem, they are not independent of each other or of the sample size n. It would be desirable to have a decision process such that both α and β will be small. However, in general, a decrease in one type of error leads to an increase in the other type for a fixed sample size. The only way to simultaneously reduce both types of errors is to

increase the sample size. One might also attach some relative importance (or cost) to the four possible courses of action and minimize the total cost of the decision. The probabilities of correct decisions may be expressed as

$$P(D_0|H_0) = P(\mathbf{x} \in R_0; H_0) \tag{3.7}$$

$$P(D_1|H_1) = P(\mathbf{x} \in R_1; H_1) \tag{3.8}$$

As an example, the two hypothesis for radar signal detection can be defined as

H_0: no target exists,

H_1: target is present.

In this case, the probability of a Type I error $P_I = P(D_1|H_0)$ is often referred to as the *false alarm* probability (denoted by P_F), the probability of a Type II error $P_{II} = P(D_0|H_1)$ as the *miss* probability (denoted by P_M), and $P(D_1|H_1)$ as the *detection* probability (denoted by P_D). The cost of failing to detect a target cannot be easily determined. In general, we set a value of P_F that is acceptable and seek a decision test that constraint P_F to this value while minimizing P_D (or equivalently minimizing P_M). This is known as the *Neyman-Pearson* test [34].

3.3.3 Decision Tests

A. Maximum-Likelihood Test

Let $\mathbf{x}$ be the observation vector and $P(\mathbf{x}|H_i), i = 0, 1$, denote the probability of observing $\mathbf{x}$ given that H_i was true. In the *maximum-likelihood* test, the decision regions R_0 and R_1 are selected as

$$R_0 = \{\mathbf{x} : P(\mathbf{x}|H_0) > P(\mathbf{x}|H_1)\} \tag{3.9}$$

$$R_1 = \{\mathbf{x} : P(\mathbf{x}|H_0) < P(\mathbf{x}|H_1)\} \tag{3.10}$$

Thus, the maximum-likelihood test can be expressed as

$$d(\mathbf{x}) = \begin{cases} H_0, & \text{if } P(\mathbf{x}|H_0) > P(\mathbf{x}|H_1) \\ H_1, & \text{if } P(\mathbf{x}|H_0) < P(\mathbf{x}|H_1) \end{cases} \tag{3.11}$$

The above decision test can be rewritten as

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{3.12}$$

If we define the likelihood ratio $\Lambda(\mathbf{x})$ as

$$\Lambda(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \tag{3.13}$$

the maximum-likelihood test can be expressed as

$$\Lambda(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{3.14}$$

which is called the *likelihood ratio test*, and 1 is called the *threshold value* of the test. Note that the likelihood ratio $\Lambda(\mathbf{x})$ is also often expressed as

$$\Lambda(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} \tag{3.15}$$

B. MAP Test

Let $P(H_i|\mathbf{x}), i = 0, 1$, denote the probability that H_i was true given a particular value of $\mathbf{x}$. The conditional probability $P(H_i|\mathbf{x})$ is called an *a posteriori* (or posterior) probability; that is, a probability that is computed after an observation has been made. The probability $P(H_i), i = 0, 1$, is called an *a priori* (or prior) probability. In the *maximum a posteriori* (MAP) test, the decision regions R_0 and R_1 are selected as

$$R_0 = \{\mathbf{x} : P(H_0|\mathbf{x}) > P(H_1|\mathbf{x})\} \tag{3.16}$$

$$R_1 = \{\mathbf{x} : P(H_0|\mathbf{x}) < P(H_1|\mathbf{x})\} \tag{3.17}$$

Thus, the MAP test is given by

$$d(\mathbf{x}) = \begin{cases} H_0, & \text{if } P(H_0|\mathbf{x}) > P(H_1|\mathbf{x}) \\ H_1, & \text{if } P(H_0|\mathbf{x}) < P(H_1|\mathbf{x}) \end{cases} \tag{3.18}$$

which can be rewritten as

$$\frac{P(H_1|\mathbf{x})}{P(H_0|\mathbf{x})} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{3.19}$$

Using Bayes' rule (Equation 3.4), Equation 3.19 reduces to

$$\frac{P(\mathbf{x}|H_1)P(H_1)}{P(\mathbf{x}|H_0)P(H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{3.20}$$

Using the likelihood ratio $\lambda(\mathbf{x})$ defined in Equation 3.13, the MAP test can be expressed in the following likelihood ratio test as

$$\Lambda(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta = \frac{P(H_0)}{P(H_1)} \tag{3.21}$$

where $\eta = P(H_0)/P(H_1)$ is the threshold value for the MAP test. Note that when $P(H_0) = P(H_1)$, the maximum-likelihood test is also the MAP test.

C. Neyman-Pearson Test

As we mentioned before, it is not possible to simultaneously minimize both $\alpha(= P_I)$ and $\beta(= P_{II})$. The Neyman-Pearson test provides a workable solution to this problem in that the test minimizes β for a given level of α. Hence, the Neyman-Pearson test maximizes the power of the test $1 - \beta$ for a given level of significance α. In the Neyman-Pearson test, the critical (or rejection) region R_1 is selected such that $1 - \beta = 1 - P(D_0|H_1) = P(D_1|H_1)$ is maximum, subject to the constraint $\alpha = P(D_1|H_0) = \alpha_0$. This is a classical problem in optimization: maximizing a function subject to a constraint, which can be solved by the use of the Lagrange multiplier method. Thus, we construct the objective function

$$J = (1 - \beta) - \lambda(\alpha - \alpha_0) \tag{3.22}$$

where $\lambda \geq 0$ is a Lagrange multiplier. Then the critical region R_1 is chosen to maximize J. It can be shown that the Neyman-Pearson test can be expressed in terms of the likelihood ratio test as

$$\Lambda(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta = \lambda \tag{3.23}$$

where the threshold value η of the test is equal to the Lagrange multiplier λ, which is chosen to satisfy the constraint $\alpha = \alpha_0$.

D. Bayes' Test

Let C_{ij} be the cost associated with (D_i, H_j), which denotes the event that we accept H_i when H_j is true. Then, the average cost, which is known as *Bayes' risk*, can be written as

$$\bar{C} = C_{00}P(D_0, H_0) + C_{10}P(D_1, H_0) + C_{01}P(D_0, H_1) + C_{11}P(D_1, H_1) \tag{3.24}$$

where $P(D_i, H_j)$ denotes the probability that we accept H_i when H_j is true. By Bayes' rule (Equation 3.4), we have

$$\begin{aligned} \bar{C} &= C_{00}P(D_0|H_0)P(H_0) + C_{10}P(D_1|H_0)P(H_0) \\ &\quad + C_{01}P(D_0|H_1)P(H_1) + C_{11}P(D_1|H_1)P(H_1) \end{aligned} \tag{3.25}$$

In general, we assume that $C_{10} > C_{00}$ and $C_{01} > C_{11}$, since it is reasonable to assume that the cost of making an incorrect decision is greater than the cost of making a correct decision. The test that minimizes the average cost $\bar{C}$ is called the Bayes' test, and it can be expressed in terms of the likelihood ratio test as

$$\Lambda(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta = \frac{(C_{10} - C_{00})P(H_0)}{(C_{01} - C_{11})P(H_1)} \tag{3.26}$$

Note that when $C_{10} - C_{00} = C_{01} - C_{11}$, the Bayes' test (Equation 3.26) and the MAP test (Equation 3.21) are identical.

E. Minimum Probability of Error Test

If we set $C_{00} = C_{11} = 0$ and $C_{01} = C_{10} = 1$ in Equation 3.24, we have

$$\bar{C} = P(D_1, H_0) + P(D_0, H_1) = P_e \tag{3.27}$$

which is just the probability of making an incorrect decision. Thus, in this case, the Bayes' test yields the minimum probability of error, and Equation 3.26 becomes

$$\Lambda(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta = \frac{P(H_0)}{P(H_1)} \tag{3.28}$$

We see that the minimum probability of the error test is the same as the MAP test.

F. Minimax Test

We have seen that the Bayes' test requires the a priori probabilities $P(H_0)$ and $P(H_1)$. Frequently, these probabilities are not known. In such circumstance, the Bayes' test cannot be applied, and the following minimax (min-max) test may be used. In the minimax test, we use the Bayes' test that corresponds to the least favorable $P(H_0)$. In the minimax test, the critical region R_1^* is defined by

$$\max_{P(H_0)} \bar{C}[P(H_0), R_1^*] = \min_{R_1} \max_{P(H_0)} \bar{C}[P(H_0), R_1] < \max_{P(H_0)} \bar{C}[P(H_0), R_1] \tag{3.29}$$

for all $R_1 \neq R_1^*$. In other words, R_1^* is the critical region that yields the minimum Bayes' risk for the least favorable $P(H_0)$. Assuming that the minimization and maximization operations are interchangeable, we have

$$\min_{R_1} \max_{P(H_0)} \bar{C}[P(H_0), R_1] = \max_{P(H_0)} \min_{R_1} \bar{C}[P(H_0), R_1] \tag{3.30}$$

The minimization of $\bar{C}[P(H_0), R_1]$ with respect to R_1 is simply the Bayes' test, so that

$$\min_{R_1} \bar{C}[P(H_0), R_1] = \bar{C}^*[P(H_0)] \tag{3.31}$$

where $C^*[P(H_0)]$ is the minimum Bayes' risk associated with the a priori probability $P(H_0)$. Thus, Equation 3.30 states that we may find the minimax test by finding the Bayes' test for the least favorable $P(H_0)$; that is, the $P(H_0)$ that maximizes $\bar{C}[P(H_0)]$.

3.4 Design Principles

The traditional method of testing a face recognition algorithm is to provide the algorithm with two sets of images – the gallery and the probe set – that do not intersect [87]. Unfortunately, this method severely limits one's ability to analyze the data. To overcome this deficiency, we modified the evaluation protocol to allow for a more detailed analysis of face recognition algorithm performance. We designed the evaluation protocol so that algorithm performance could be computed for a variety of different galleries and probe sets.

In the new protocol, an algorithm is given two sets of images: the *target set* and the *query set*. We introduce this terminology to distinguish these sets from the gallery and probe sets that are used in computing performance statistics. The target set is given to the algorithm as the set of known facial images. The images in the query set are the unknown facial images to be identified. The FERET test had two fundamental design rules. The first was that there is only one image per person in the gallery. The second was that the representation used to encode the faces is learned from a subset of images in the gallery. These design rules test recognition methods that do not explicitly use class information.

3.4.1 Test Sets, Galleries, and Probe Sets

For each image q_i in the query set $\mathcal{Q}$, an algorithm reports the similarity $s_i(k)$ between q_i and each image t_k in the target set $\mathcal{T}$. The evaluation protocol is designed so that each algorithm can use a different similarity measure. We do not compare similarity measures from different algorithms. The key property of the new protocol, which allows for greater flexibility in scoring, is that for any two images s_i and t_k, we know $s_i(k)$. (In fact, designation of which set is the target and which is the query is arbitrary. By reformatting the output, we can change the roles of the target and query sets.)

This flexibility allows the evaluation methodology to be robust and comprehensive and is achieved by computing scores for virtual galleries and probe sets. A gallery $\mathcal{G}$ is a virtual gallery if $\mathcal{G}$ is a proper subset of the target set; i.e., $\mathcal{G} \subset \mathcal{T}$. Similarly, $\mathcal{P}$ is a virtual probe set if $\mathcal{P} \subset \mathcal{Q}$. For a given gallery $\mathcal{G}$ and probe set $\mathcal{P}$, the performance scores are computed by examination of the similarity measures $s_i(k)$, such that $q_i \in \mathcal{P}$ and $t_k \in \mathcal{G}$.

The virtual gallery and probe set technique allows us to characterize algorithm performance by different categories of images. The different categories include (1) rotated images, (2) duplicates taken within a week of the gallery image, (3) duplicates where the time between the images is at least one year, (4) galleries containing one image per person, and (5) galleries containing more than one image per person. We can create a gallery of 100 people and estimate the algorithm's performance at recognizing people in this gallery. Using this as a starting point, we can then create virtual galleries of 100, 200, ..., 1000 people and determine how performance changes as the size of the gallery increases. Another avenue of investigation is to create n different galleries of size 100, and calculate the variation in algorithm performance with the different galleries of this size.

To take full advantage of virtual galleries and probe sets, we placed multiple images of the same person in the target and query sets. If such images were marked as the same person, then the algorithms being tested could use the information in the evaluation process. To keep this from happening, we required that each image in the target set be treated as a unique face. In practice, this condition is enforced by giving every image in the target and query set a unique identification.

For each query image q_i, an algorithm output the similarity measure $s_i(k)$ for all images t_k in the target set. The output for each query image q_i was sorted by the similarity scores $s_i(\cdot)$. Since the target set is a subset of the query set, the test output contained the similarity score between all images in the target set.

Table 3.1: Size of galleries and probe sets for different probe categories.

Probe category	Duplicate I	Duplicate II	**FB**	**fc**
Gallery size	1196	864	1196	1196
Probe set size	722	234	1195	194

Except for the rotated and digitally modified images, the target and query sets were the same. Thus, the test output contained every target image matched with itself. To obtain a robust comparison of algorithms, it was necessary to calculate performance on a large number of galleries and probe sets. This allowed a detailed analysis of performance on multiple galleries and probe sets. (We did not present the results for the rotated or digitally modified images.)

To allow for a robust and detailed analysis, we report identification and verification scores for four categories of probes. The first probe category was the **FB** probes. For each set of images, there were two frontal images. One of the images was randomly placed in the gallery, and the other image was placed in the **FB** probe set. (This category is denoted by **FB** to differentiate it from the **fb** images in the FERET database.) The second probe category contained all duplicate frontal images in the FERET database for the gallery images. We refer to this category as the duplicate I probes. The third category was the **fc** (images taken the same day, but with a different camera and lighting). The fourth consisted of duplicates where there was at least one year between the acquisition of the probe image and corresponding gallery image. We refer to this category as the duplicate II probes. For this category, the gallery images were acquired before January 1995 and the probe images were acquired after January 1996. The size of the galleries and probe sets for the four probe categories are presented in Table 3.1. The **FB**, **fc**, and duplicate I galleries are the same. The duplicate II gallery is a subset of the other galleries. None of the individuals photographed for the gallery images wore glasses.

3.4.2 Performance Evaluation

Generally, face recognition algorithms consist of two main parts: (1) face detection [68, 69, 99, 107, 125] and normalization and (2) face identification and verification. Algorithms that consist of both parts are referred to as *fully automatic algorithms*, and those that consist of only the second part are *partially automatic algorithms*. The first version of the test dealt with partially automatic algorithms, and the test algorithms were given a list of images in the target and query sets, and the coordinates of the center of the eyes for images in the target and query sets. In the second version of the test, the coordinates of the eyes were not provided. (For details of the test, refer to Table 2.1.)

By comparing the performance between the two versions, one can estimate the performance of the face-locating and identifying portions of an algorithm. Both tasks are evaluated on the same sets of images. The target and query sets were the same for each version. The target set contained 3,323 images and the query set 3,816 images. All the images in the target set were frontal images. The query set consisted of all the images in the target set plus rotated images and digitally modified images. We designed the digitally modified images to test the effects of illumination and scale [87].

The basic models for evaluating the performance of an algorithm are the closed and open universes. In the closed universe, every probe is in the gallery, and in an open universe, some probes are not in the gallery. Both models reflect different and important aspects of face recognition algorithms and report different performance statistics. We report identification results using a closed universe model.

In an identification problem, the input to an algorithm is an unknown face, and the algorithm reports back the estimated identity of an unknown face from a database of known individuals. In the closed universe, every probe is in the gallery. The complement to the closed universe is the open universe where some

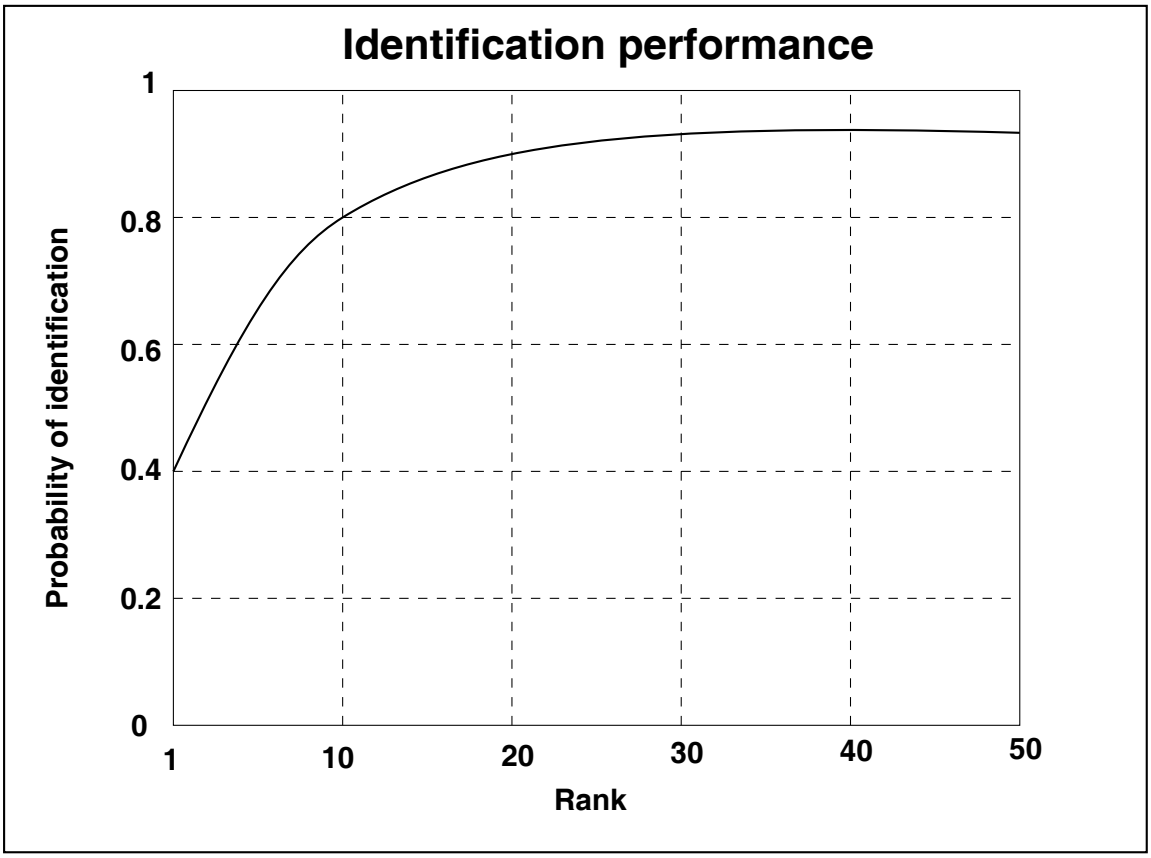


Figure 3.2: Example of identification performance.

probes are not in the gallery. The open universe model is used in a verification scenario where the input is a face with a claimed identity. In a verification problem, the algorithm either accepts or rejects the claimed identity [95, 96].

The closed universe model allows one to ask how good is an algorithm at identifying a probe image. The question is not always, "Is the top match correct?", but "Is the correct answer in the top n matches?" This lets one know how many images have to be examined to get the desired level of performance. For identification, the performance statistics are reported as cumulative match scores (see Figure 3.2). The rank is plotted along the horizontal axis, and the vertical axis is the probability of identification. The probability of identification can be calculated for any subset of the probe set. We calculated this score to

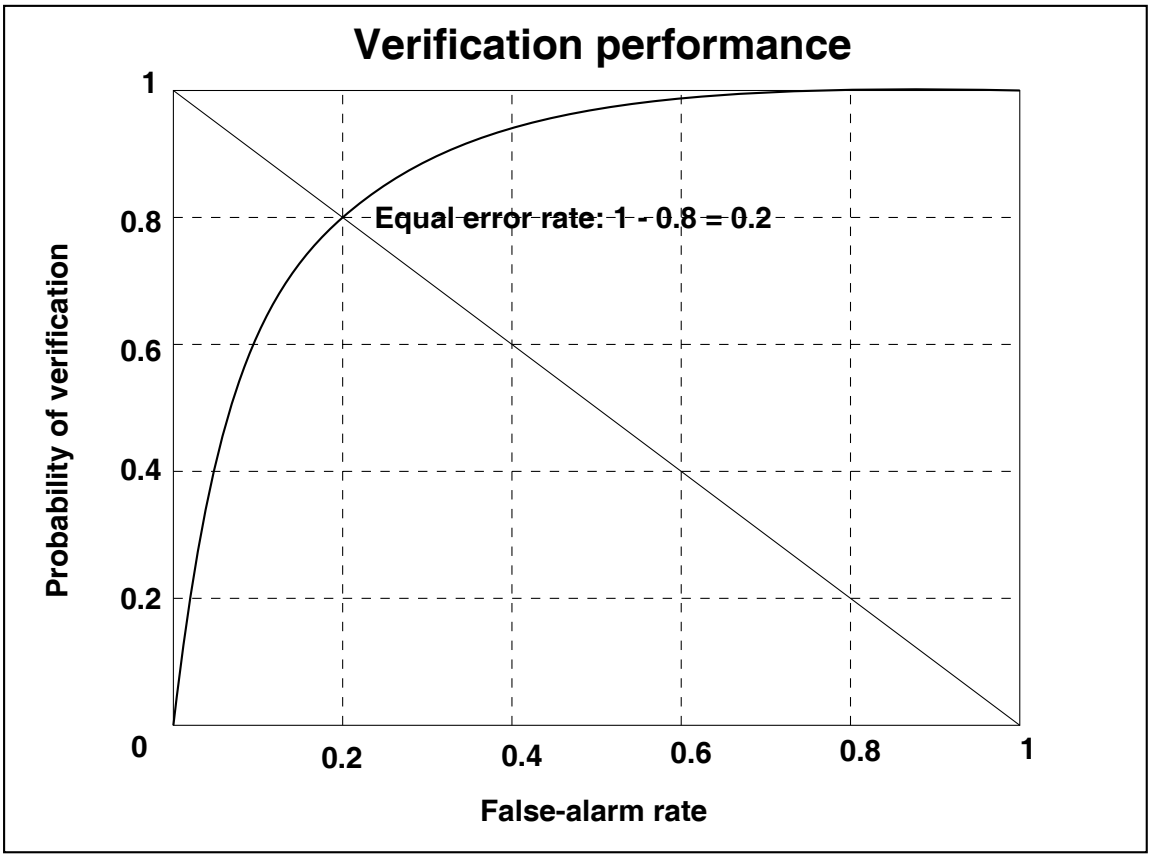


Figure 3.3: Example of verification performance.

evaluate an algorithm's performance on different categories of probes; i.e., rotated or scaled probes.

In an open universe test, the results are reported on a *receiver operating characteristic* (ROC) curve. ROC curves are being used to judge the discrimination ability of various statistical methods that combine various clues and test results for predictive purposes. In our experimental results, ROC presents the trade-off between the probability of false alarm and the probability of correct identification (see Figure 3.3). There are two classes of probes in an open universe. The first is made up of probes that are not in the gallery, which could generate false alarms. A false alarm occurs when an algorithm reports that one of these probes is in the gallery. The false-alarm rate is $P_F = \widehat{F}/F^*$, where F^* is

the number of probes not in the gallery and $\widehat{F}$ is the number of probes reported as false alarms. The second class is made up of probes that are in the gallery. Performance on these probes is characterized by P_I, the probability that a probe is correctly identified; thus, $P_I = \widehat{I}/I^*$, where I^* is the size of a set of probes and $\widehat{I}$ is the number of these probes that are correctly identified.

For a given algorithm, the choice of a suitable hit and false-alarm rate pair depends on a particular application. However, for performance evaluation and comparison among algorithms, the *equal error rate* (EER) is often quoted. The EER occurs at the threshold c, where the incorrect rejection and false-alarm rates are equal (incorrect rejection rate = $1 -$ verification rate).

Chapter 4

An Identification Model for Face Recognition Algorithms

4.1 Introduction

Over the last decade, face recognition has become an active area of research in computer vision, neuroscience, and psychology. Progress has advanced to the point that face recognition systems are being demonstrated in real-world settings [86]. The rapid development of face recognition is due to a combination of factors: active development of algorithms, the availability of a large database of facial images, and a method for evaluating the performance of face recognition algorithms. The FERET database and evaluation methodology address the latter two points and are de facto standards. There have been three FERET evaluations with the most recent being the September 1996 and March 1997 FERET test.

We report identification results using a closed universe model. The closed universe model allows one to ask how good an algorithm is at identifying a probe image. This lets one know how many images have to be examined to get the desired level of performance. In an identification problem, the input to an algorithm is an unknown face, and the algorithm reports back the estimated identity

of an unknown face from a database of known individuals (e.g., searching an electronic file of mug shots for the identity of a suspect).

4.2 Identification Model

Our new evaluation protocol was designed to assess and advance the state of the art and point out directions for future research. To succeed at this, the test design must solve the *three bears problem*; the test cannot be too hard or too easy, but has to be just right. If the test is too easy, the testing process becomes an exercise in *tuning* existing algorithms. If the test is too hard, the test is beyond the ability of existing algorithmic techniques. If the results from the test are poor, they do not allow for an accurate assessment of algorithmic capabilities.

The solution to the three bears problem is through the selection of images in the test set and the evaluation protocol. Tests are administered using an evaluation protocol that states the mechanics of the tests and the manner in which the tests will be scored. In face recognition, for example, the protocol states the number of images of each person in the test, how the output from the algorithm is recorded, and how the performance results are reported.

The characteristics and quality of the images are major factors in determining the difficulty of the problem being evaluated. For example, if the faces are in a predetermined position in the images, the problem is different from that for images in which the faces can be located anywhere in the image. In the FERET database, variability was introduced by the inclusion of images taken at different dates and locations. This resulted in changes in lighting, scale, and background.

The evaluation protocol is based on a set of design principles. Stating the design principle allows one to assess how appropriate the FERET test is for a particular face recognition algorithm. Also, the design principles help in de-

termining if an evaluation methodology for testing algorithm(s) for a particular application is appropriate. Before discussing the design principles, we state the evaluation protocol.

The second design principle is that training is completed before the start of the test. This forces each algorithm to have a general representation for faces, not a representation tuned to a specific gallery. Without this condition, virtual galleries would not be possible.

For algorithms to have a general representation for faces, they must be gallery (class) insensitive. Examples are algorithms based on normalized correlation or PCA. An algorithm is class sensitive if the representation is tuned to a specific gallery. Examples are straightforward implementation of Fisher discriminant analysis [35, 109]. The Fisher discriminant analysis technique was adapted to class insensitive testing methodologies by Zhao et al [128, 130], with performance results of these extensions being reported in this chapter.

The third design rule is that all algorithms tested compute a similarity measure between two facial images; this similarity measure was computed for all pairs of images in the test set. Knowing the similarity score between all pairs of images from the target and query sets allows for the construction of virtual galleries and probe sets.

The computation of an identification score is quite simple. Let $\mathcal{P}$ be a probe set and $|\mathcal{P}|$ the size of $\mathcal{P}$. We score probe set $\mathcal{P}$ against gallery $\mathcal{G}$, where $\mathcal{G}$ = $\{g_1, ..., g_M\}$ and $\mathcal{P}$= $\{p_1, ..., p_N\}$ by comparing the similarity scores $s_i(\cdot)$ such that $p_i \in \mathcal{P}$ and $g_k \in \mathcal{G}$. For each probe image $p_i \in \mathcal{P}$, we sort $s_i(\cdot)$ for all gallery images $g_k \in \mathcal{G}$. We assume that a smaller similarity score implies a closer match. If g_k and p_i are the same image, then $s_i(k) = 0$. The function $id(i)$ gives the index of the gallery image of the person in probe p_i; i.e., p_i is an image of the person in $g_{id(i)}$. A probe p_i is correctly identified if $s_i(id(i))$ is the smallest score for $g_k \in \mathcal{G}$. A probe p_i is in the top k if $s_i(id(i))$ is one of the k-th smallest scores $s_i(\cdot)$ for

gallery $\mathcal{G}$. Let R_k denote the number of probes in the top k. We reported $R_k/|\mathcal{P}|$, the fraction of probes in the top k. As an example, let $k = 5$, $R_5 = 80$ and $|\mathcal{P}|$ = 100. Based on the formula, the performance score for R_5 is 80/100 = 0.8.

In reporting identification performance results, we state the size of the gallery and the number of probes scored. The size of the gallery is the number of different faces (people) in the images that are in the gallery. For all results, there is one image per person in the gallery. Thus, the size of the gallery is also the number of images in the gallery. The number of probes scored (also, size of the probe set) is $|\mathcal{P}|$. The probe set may contain more than one image of a person and the probe set may not contain an image of everyone in the gallery. Every image in a probe is an image of a person in the corresponding gallery.

4.3 Identification Results

4.3.1 Partially Automatic Algorithm Performance

We report identification scores for four categories of probes (see section 3.4.1). The results for identification are reported as cumulative match scores. Table 4.1 shows the categories corresponding to the figures that present the results, type of results, and size of the gallery and probe sets. Figures 4.1 to 4.4 report the identification performance of four categories of probes; **FB**, duplicate I, **fc**, and duplicate II.

In Figures 4.5 and 4.6, we compare the difficulty of different probe sets. Whereas, Figures 4.1 to 4.4 report the identification performance for each algorithm, Figure 4.5 shows a single curve that is an average of the identification performance of all algorithms for each probe category. For example, the first rank score for duplicate I probe sets is computed from an average of the first rank score for all algorithms in Figure 4.2. In Figure 4.6, we present the current upper bound identification performance on partially automatic algorithms for

Table 4.1: Figures reporting identification results for partially automatic algorithms. Performance is broken out by probe category.

Figure No.	Probe category	Gallery size	Probe set size
4.1	**FB**	1,196	1,195
4.2	Duplicate I	1,196	722
4.3	**fc**	1,196	194
4.4	Duplicate II	864	234

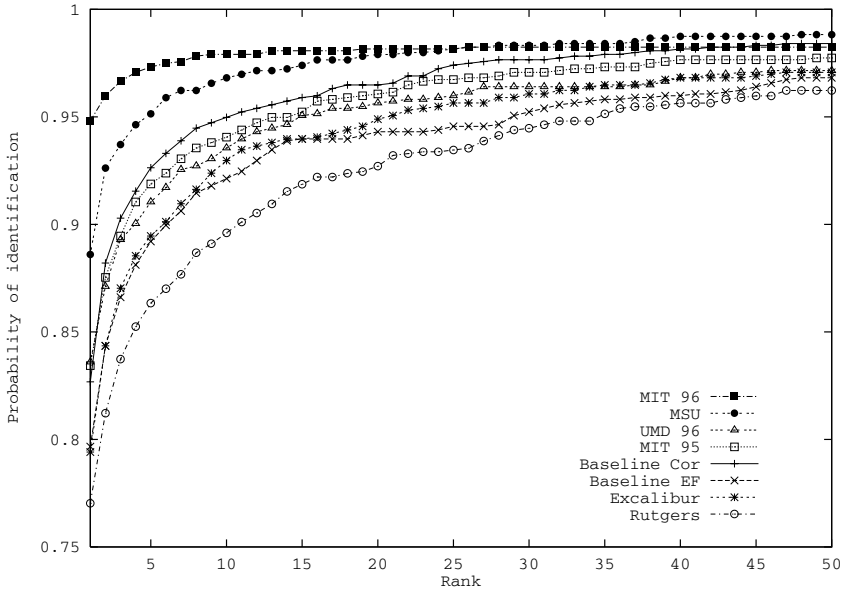
each probe category. For each category of probe, Figure 4.6 plots the algorithm with the highest top rank score (R_1).

4.3.2 Fully Automatic Algorithm Performance

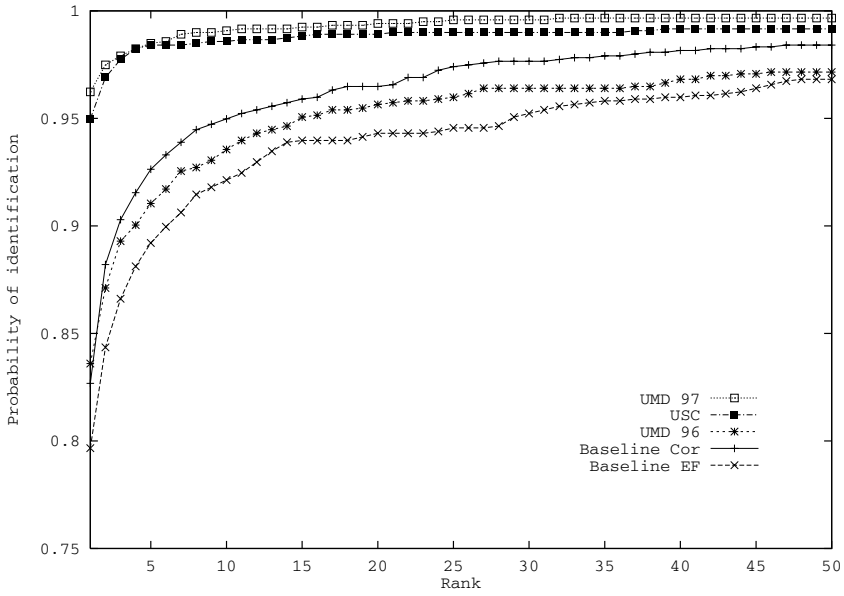
We report identification performance for the fully automatic algorithms of the MIT Media Lab and USC. To allow for a comparison between the partially and fully automatic algorithms, we plot the results for the partially and fully automatic algorithms in one graph. Figure 4.7 shows identification performance for **FB** probes and Figure 4.8 shows identification performance for duplicate I probes. Additionally, Figure 4.9 shows identification performance for **fc** probes and Figure 4.10 shows identification performance for duplicate II probes. (The gallery and probe sets are the same as in subsection 4.3.1.)

4.3.3 Variation in Identification Performance

From a statistical point of view, a face recognition algorithm estimates the identity of a face. Consistent with this view, we can ask about the variance in performance of an algorithm: For a given category of images, how does performance change if the algorithm is given a different gallery and probe set? In Tables 4.2 and 4.3, we show how algorithm performance varies if the people in

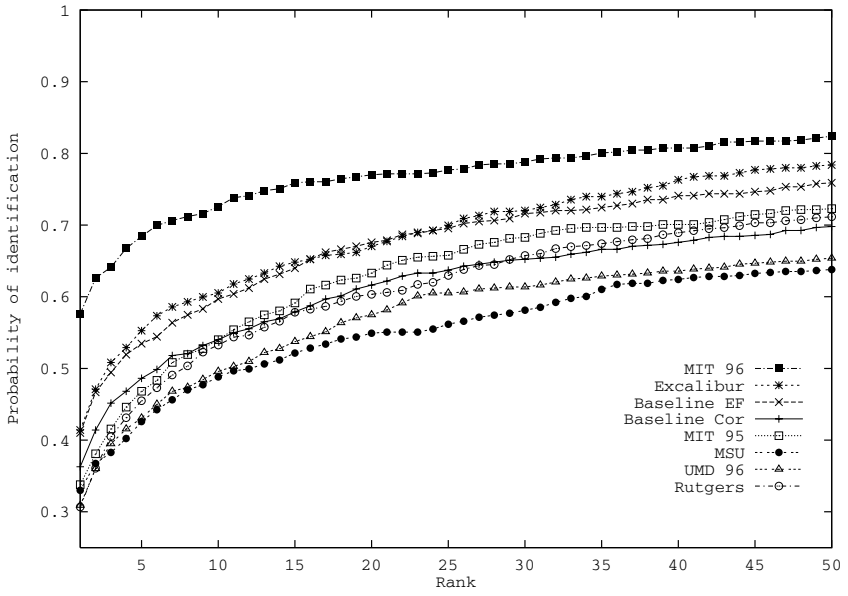


(a)

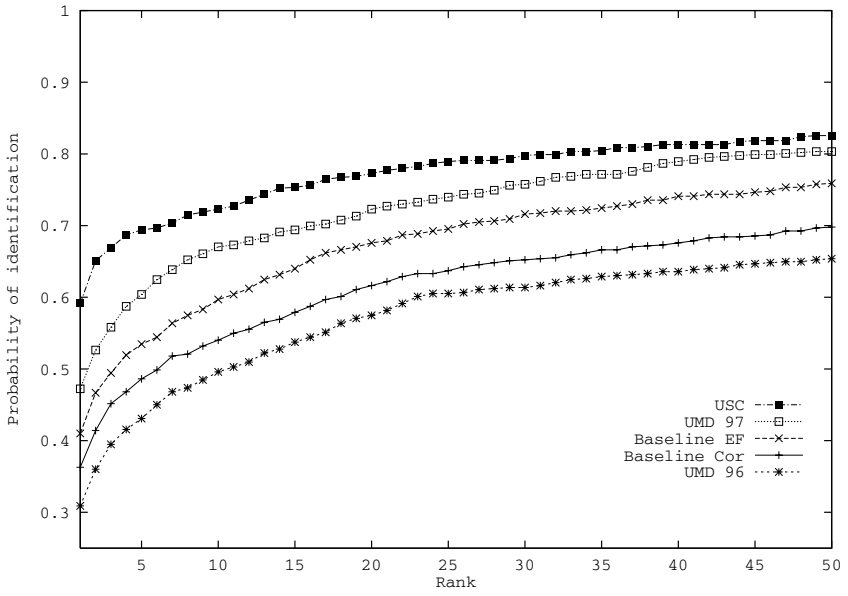


(b)

Figure 4.1: Identification performance for **FB** probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.

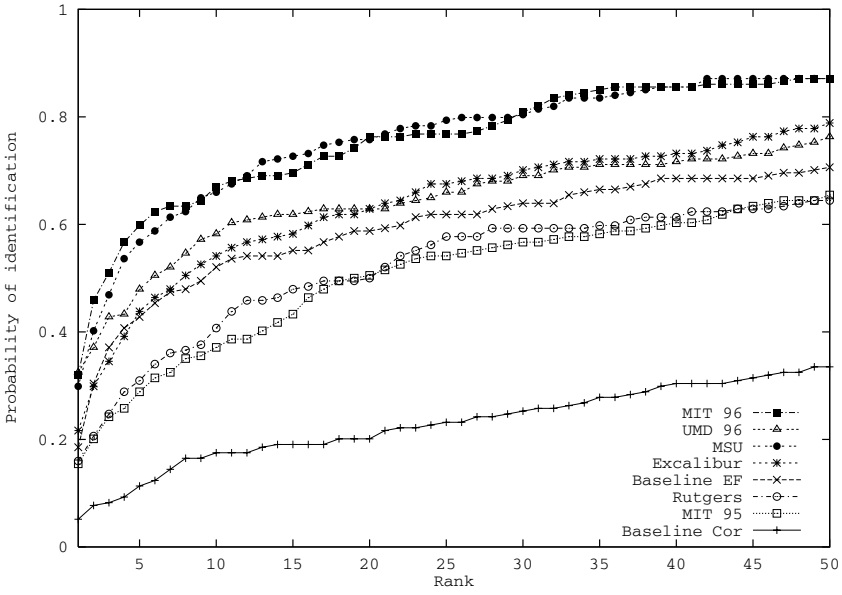


(a)

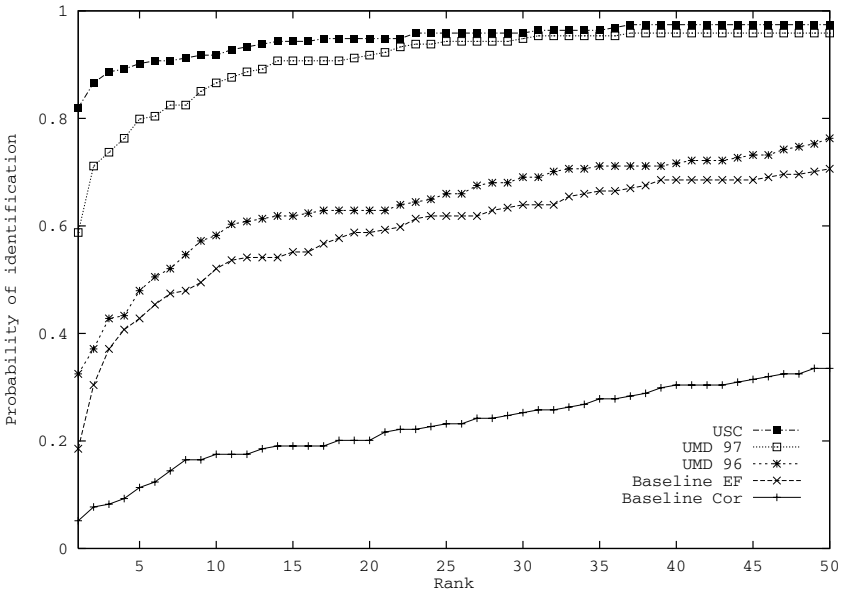


(b)

Figure 4.2: Identification performance for duplicate I probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.

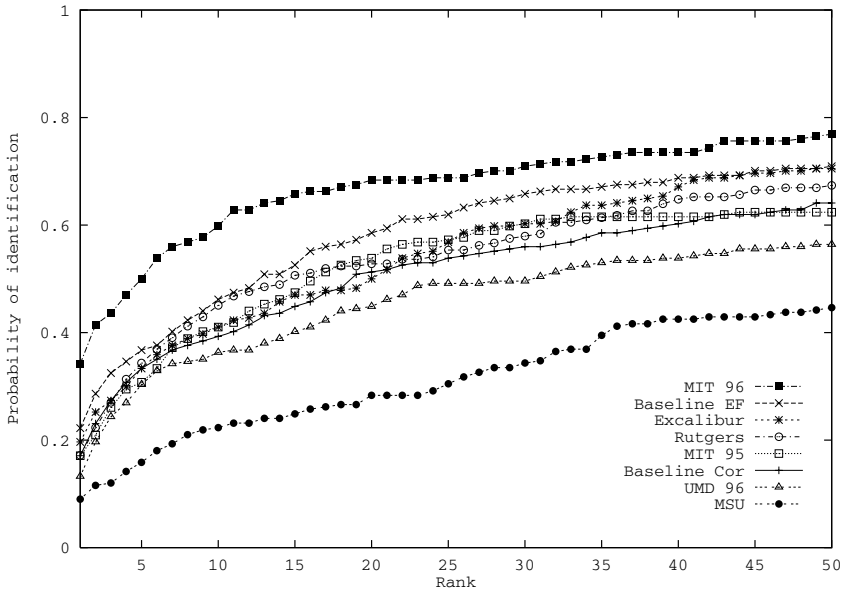


(a)

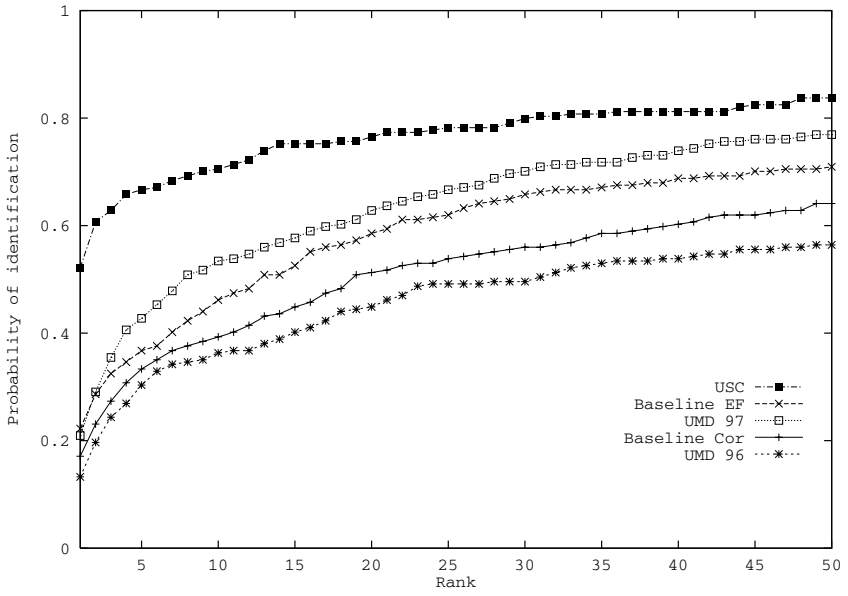


(b)

Figure 4.3: Identification performance for **fc** probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.



(a)



(b)

Figure 4.4: Identification performance for duplicate II probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.

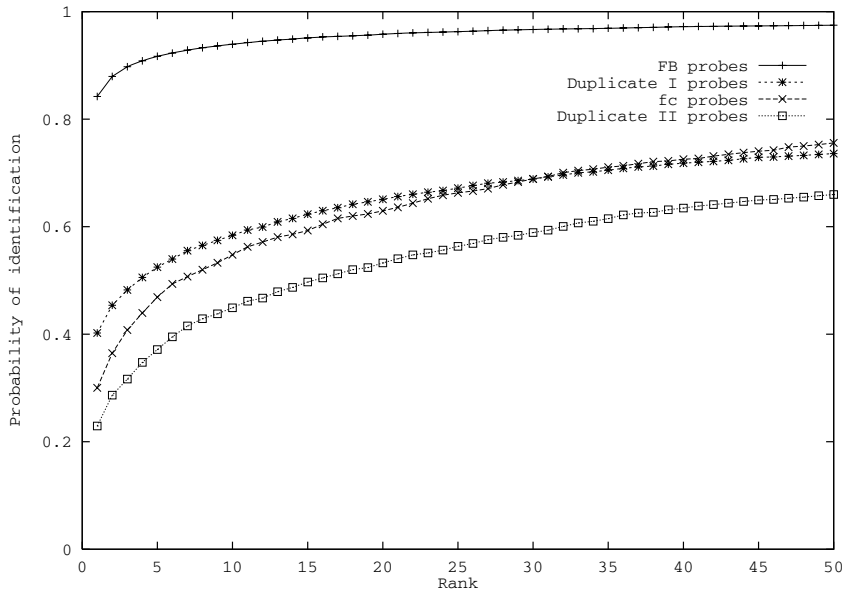


Figure 4.5: Average identification performance of partially automatic algorithms for each probe category.

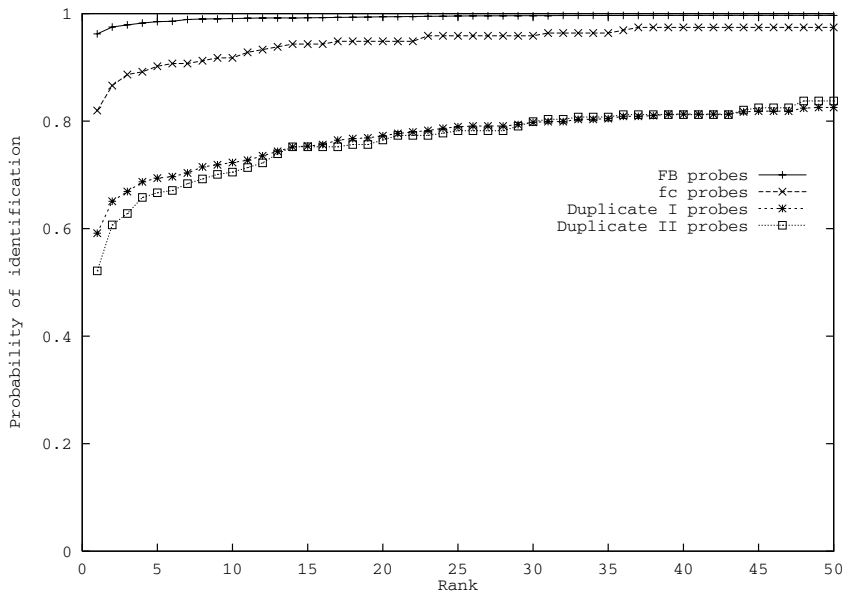


Figure 4.6: Current upper bound on identification performance of partially automatic algorithm for each probe category.

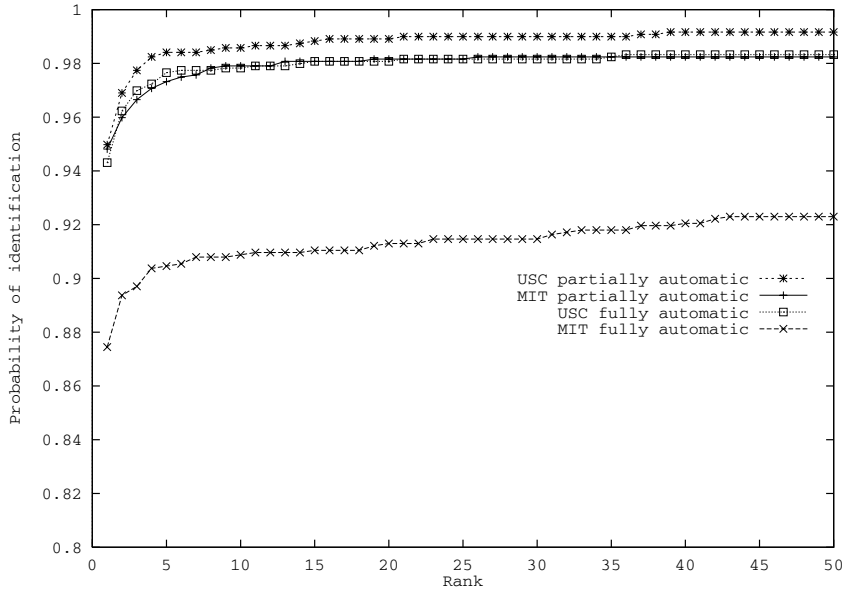


Figure 4.7: Identification performance of fully automatic algorithms against partially automatic algorithms for **FB** probes.

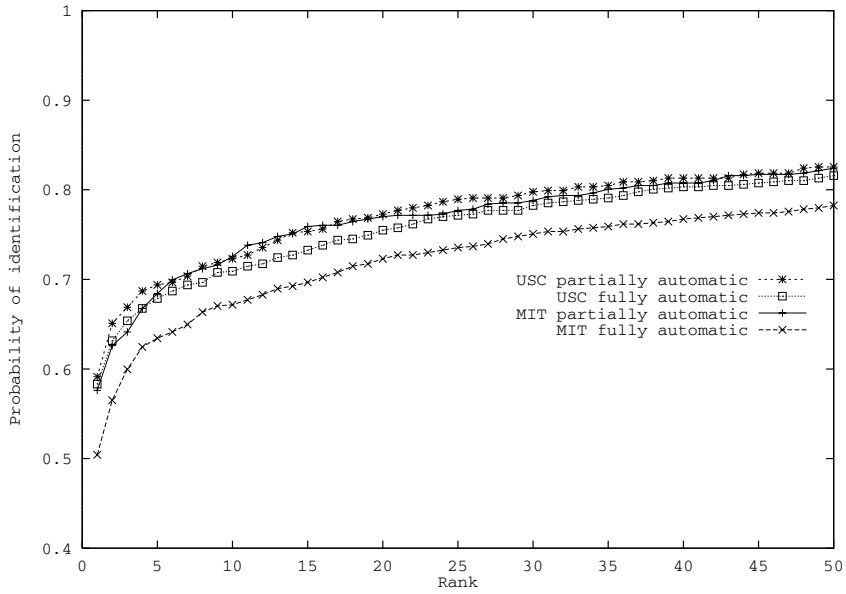


Figure 4.8: Identification performance of fully automatic algorithms against partially automatic algorithms for duplicate I probes.

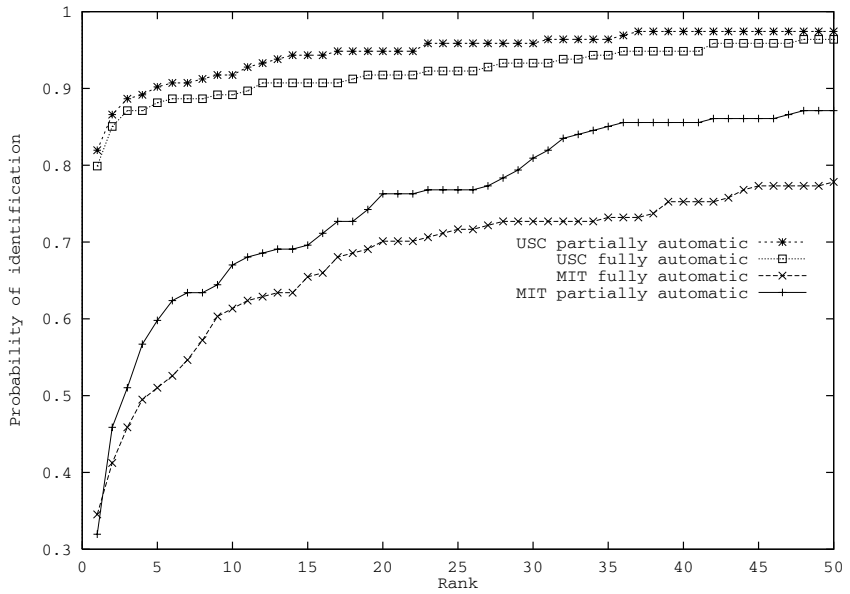


Figure 4.9: Identification performance of fully automatic algorithms against partially automatic algorithms for **fc** probes.

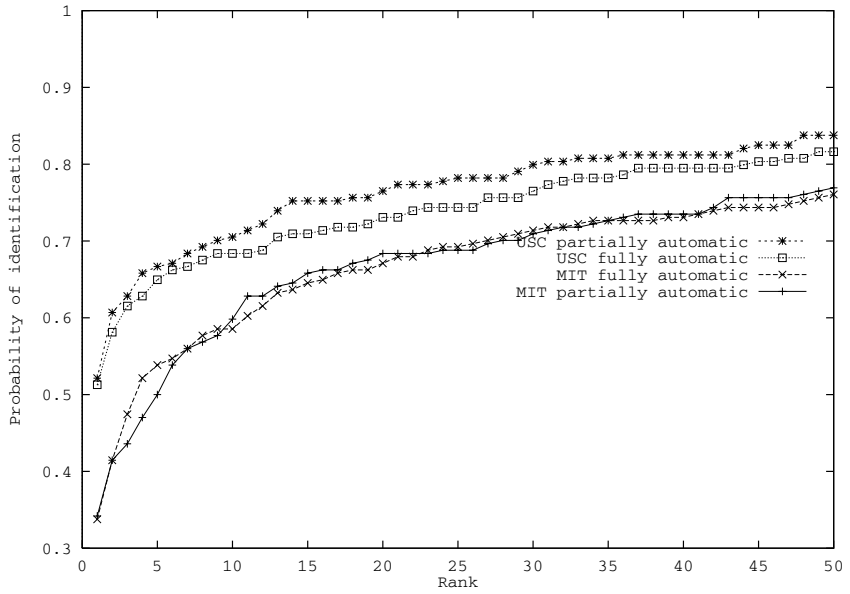


Figure 4.10: Identification performance of fully automatic algorithms against partially automatic algorithms for duplicate II probes.

the galleries change. For this experiment, we constructed six galleries of approximately 200 individuals, in which an individual was in only one gallery (the number of people in each gallery versus the number of probes scored is given in Tables 4.2 and 4.3). Results are for the partially automatic algorithms.

We have ordered algorithms by their top rank performance on each gallery; for example, in Table 4.2, the UMD March 1997 algorithm scored highest on gallery 1 and the baseline PCA and correlation tied for ninth place. Also included in the table is the average identification performance for all algorithms. Table 4.2 reports results for the **FB** probes. Table 4.3 is organized in the same manner as Table 4.2, except that duplicate I probes are scored. Tables 4.2 and 4.3 report results for the same gallery. The galleries were constructed by placing images in the galleries in the chronological order that the images were collected (the first gallery contains the first images collected and the sixth gallery contains the most recent images that were collected). In Table 4.3, mean age refers to the average time between the collection of images in the gallery and the corresponding duplicate probes. No scores are reported in Table 4.3 for gallery 6 because there are no duplicates for this gallery.

4.4 Discussions and Conclusions

Our new evaluation protocol makes it possible to independently evaluate algorithms. The protocol was designed to evaluate algorithms on different galleries and probe sets for different scenarios. In this chapter, we discuss how we computed the performance on identification tasks using this protocol. Our results show that factors that affect performance include scenario, date tested, and probe category.

The September 1996 and March 1997 test was the latest FERET test (the others were the August 1994 and March 1995 test [88]). One of the main goals of the FERET tests has been to improve the performance of face recognition

Table 4.2: Variations in identification performance on six different galleries on **FB** probes. Images in each gallery do not overlap. Ranks range from 1 to 10.

	Rank by top match					
	Gallery size / scored probes					
	200/200	200/200	200/200	200/200	200/199	196/196
Algorithms	Gallery 1	Gallery 2	Gallery 3	Gallery 4	Gallery 5	Gallery 6
Baseline PCA	9	10	8	8	10	8
Baseline Cor.	9	9	9	6	9	10
Excalibur Corp.	6	7	7	5	7	6
MIT Sep96	4	2	1	1	3	3
MIT Mar95	7	5	4	4	5	7
MSU	3	4	5	8	4	4
Rutgers	7	8	9	6	7	9
UMD Sep96	4	6	6	10	5	5
UMD Mar97	1	1	3	2	2	1
USC	2	3	2	2	1	1
Average score	0.935	0.857	0.904	0.918	0.843	0.804

algorithms, and this can be seen in the new FERET test. The first improvement in performance was with the MIT Media Lab September 1996 algorithm over the March 1995 algorithm; the second is the improvement that came with the UMD algorithm between September 1996 and March 1997.

By looking at the improvements in the algorithms over the series of FERET tests, one sees that substantial progress has been made in face recognition. The most direct method is to compare the performance of fully automatic algorithms on **fb** probes (the two earlier FERET tests only evaluated fully automatic algo-

Table 4.3: Variations in identification performance on five different galleries on duplicate probes. Images in each gallery do not overlap. Ranks range from 1 to 10.

	Rank by top match				
	Gallery size / scored probes				
	200/143	200/64	200/194	200/277	200/44
Mean age of probes (months)	9.87	3.56	5.40	10.70	3.45
Algorithms	Gallery 1	Gallery 2	Gallery 3	Gallery 4	Gallery 5
Baseline PCA	6	10	5	5	9
Baseline Cor.	10	7	6	6	8
Excalibur Corp.	3	5	4	4	3
MIT Sep96	2	1	2	2	3
MIT Mar95	7	4	7	8	10
MSU	9	6	8	10	6
Rutgers	5	7	10	7	6
UMD Sep96	7	9	9	9	3
UMD Mar97	4	2	3	3	1
USC	1	3	1	1	1
Average score	0.238	0.620	0.645	0.523	0.687

rithms. The best top rank score for **fb** probes on the August 1994 test was 78% on a gallery of 317 individuals, and for March 1995, the top score was 93% on a gallery of 831 individuals [88]. This compares to 87% in September 1996 and 95% in March 1997 (on a gallery of 1,196 individuals).

On duplicate I probes, MIT Media Lab improved from 39% (March 1995) to

51% (September 1996); USC's performance remained approximately the same at about 58% between March 1995 and March 1997. This improvement in performance was achieved while the gallery size increased and the number of duplicate I probes increased from 463 to 722. While increasing the number of probes does not necessarily increase the difficulty of identification tasks, we argue that the September 1996 duplicate I probe set was more difficult to process then the March 1995 set. The September 1996 duplicate I probe set contained the duplicate II probes, and the March 1995 duplicate I probe set did not contain a similar class of probes. Overall, the duplicate II probe set was the most difficult probe set.

Another goal of the FERET tests is to identify areas of strength and weakness in the field of face recognition. We addressed this issue by computing algorithm performance for multiple galleries and probe sets. From this evaluation, we concluded that algorithm performance is dependent on the gallery and probe sets. We observed variation in performance due to changing the gallery and probe set within a probe category, and by changing probe categories. The effect of changing the gallery while keeping the probe category constant is shown in Tables 4.2 and 4.3. For **fb** probes, the performance range is 80% to 94%; for duplicate I probes, the range is 24% to 69%. Equally important, Tables 4.2 and 4.3 show the variability in relative performance levels. For example, in Table 4.3, the UMD September 1996 duplicate performance varies between number three and nine. Similar results were found by Moon and Phillips [71] in their study of PCA-based face recognition algorithms. This shows that an area of future research could be to measure the effect of changing the gallery and probe sets and to statistically measure the characteristics of these variations.

Figures 4.5 and 4.6 show probe categories characterized by difficulty. These figures show that **fb** probes are the easiest to identify and duplicate II probes are the most difficult to identify. On average, duplicate I probes are easier to identify than **fc** probes. However, the best performance on **fc** probes is sig-

nificantly better than the best performance on duplicate I and II probes. This comparative analysis shows that future areas of research could address the processing of duplicate II probes and develop methods to compensate for changes in illumination.

The scenario being tested also contributes to the algorithm performance. For identification, the MIT Media Lab algorithm was clearly the best algorithm tested in September 1996. However, for verification, no algorithm was a top performer for all probe categories (see Rizvi et al [97]). Also, for the algorithms tested in March 1997, the USC algorithm performed better overall than the UMD algorithm for identification; however, for verification, UMD performed better overall [97]. This shows that performance on one task is not predictive of performance one another task.

The new FERET test shows that definite progress is being made in face recognition and that the upper bound in performance has not been reached. The improvement in performance documented in this chapter directly shows that the FERET series of tests have made a significant contribution to face recognition. This conclusion is indirectly supported by (1) the improvement in performance between the algorithms tested in September 1996 and March 1997, (2) the number of papers that use FERET images and report experimental results using FERET images, and (3) the number of groups that participated in the new FERET test.

Chapter 5

A Verification Model for Face Recognition Algorithms

5.1 Introduction

The verification of a person's identity is a potential area for applications of face recognition systems. In verification applications, a system confirms the claimed identity of a face presented to it. Proposed applications for verification systems include controlling access to buildings and computer terminals, confirming identities at automatic teller machines (ATMs), and verifying passport identities at immigration ports of entry. These applications have the potential to influence and impact our daily life.

For systems to be successfully fielded, it is critical that their performance is identified. To date, the performance of most algorithms has only been reported on identification tasks, which implies that characterization on identification tasks holds for verification. For face recognition systems to successfully meet the demands of verification applications, it is necessary to develop testing and scoring procedures that specifically address these applications.

A scoring procedure is one of two parts of an evaluation protocol. In the first part, an algorithm is executed on a test set of images and the output from executing the algorithm is written to a file(s). This produces the raw results. In the second part, a scoring procedure processes raw results and produces performance statistics. If the evaluation protocol and its associated scoring procedure are properly designed, the performance statistics can be computed for both identification and verification scenarios.

Our new performance evaluation methodology is designed for face recognition algorithms [84, 85]; it used images from the FERET database of facial images [88]. The new FERET test is the latest in a series of FERET tests to measure the progress, assess the state of the art, identify strengths and weaknesses of individual algorithms, and point out future directions of research in face recognition. Prior analysis of the FERET results has concentrated on identification scenarios. In this chapter, we present (1) a verification model for the new FERET test, and (2) results for verification performance scores.

5.2 Verification Model

In our verification model, a person in image p claims to be the person in image g. The system either accepts or rejects the claim. (If p and g are images of the same person, then we write $p \sim g$, otherwise, $p \not\sim g$.) Performance of the system is characterized by two performance statistics [104]. The first is the probability of accepting a correct identity; formally, the probability of the algorithm reporting $p \sim g$ when $p \sim g$ is correct. This is referred to as the verification probability, denoted by P_V (also referred to as the hit rate in the signal detection literature). The second is the probability of incorrectly verifying a claim; formally, the probability of the algorithm reporting $p \sim g$ when $p \not\sim g$. This is called the false-alarm rate and is denoted by P_F.

Verifying the identity of a single person is equivalent to a detection problem

where the gallery $G = \{g\}$. The detection problem consists of finding the probes in $p \in P$ such that $p \sim g$.

For a given gallery image g_i and probe p_k, the decision of whether an identity was confirmed or denied was generated from $s_i(k)$. The decisions were made by a *Neyman-Pearson* observer. A Neyman-Pearson observer confirms a claim if $s_i(k) \leq c$ and rejects it if $s_i(k) > c$. By the Neyman-Pearson theorem [44], this decision rule maximized the verification rate for a given false-alarm rate α. Changing c generated a new P_V and P_F. By varying c from its minimum to maximum value, we obtained all combinations of P_V and P_F. A plot of all combinations of P_V and P_F is an ROC (also known as the relative operating characteristic) [34, 44]. The input to the scoring algorithm was $s_i(k)$; thresholding similarity scores and computing P_V, P_F, and the ROCs were performed by the scoring algorithm.

The above method computed an ROC for an individual. However, we need performance over a population of people. To calculate an ROC over a population, we performed a round robin evaluation procedure for a gallery G. The gallery contained one image per person.

The first step generated a set of partitions of the probe set. For a given $g_i \in G$, the probe set P is divided into two disjoint sets D_i and F_i. The set D_i consisted of all probes p such that $p \sim g_i$, and F_i consisted of all probes such that $p \not\sim g_i$.

The second step computed the verification and false-alarm rates for each gallery image g_i for a given cut-off value c, denoted by $P_V^{c,i}$ and $P_F^{c,i}$, respectively. The verification rate was computed by

$$P_V^{c,i} = \begin{cases} 0, & \text{if } |D_i| = 0 \\ \frac{|s_i(k) \leq c \text{ given } p_k \in D_i|}{|D_i|} & \text{otherwise,} \end{cases}$$

where $|s_i(k) \leq c$, given $p \in D_i|$ was the number of probes in D_i such that $s_i(k) \leq c$.

The false-alarm rate is computed by

$$P_F^{c,i} = \begin{cases} 0, & \text{if } |F_i| = 0 \\ \frac{|s_i(k) \leq c, \text{ given } p_k \in F_i|}{|F_i|} & \text{otherwise.} \end{cases}$$

The third step computed the overall verification and false-alarm rates. This was a weighted average of $P_V^{c,i}$ and $P_F^{c,i}$. The overall verification and false-alarm rates are denoted by P_V^c and P_F^c, and were computed by

$$P_V^c = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{|D_i|}{\frac{1}{|G|}\sum_i |D_i|} P_V^{c,i} = \frac{1}{\sum_i |D_i|} \sum_{i=1}^{|G|} |s_i(k) \leq c, \text{ given } p_k \in D_i| \cdot P_V^{c,i}$$

and

$$P_F^c = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{|F_i|}{\frac{1}{|G|}\sum_i |F_i|} P_F^{c,i} = \frac{1}{\sum_i |F_i|} \sum_{i=1}^{|G|} |s_i(k) \leq c, \text{ given } p_k \in F_i| \cdot P_F^{c,i}.$$

The verification ROC was computed by varying c from $-\infty$ to $+\infty$.

The equal error rate (EER) occurs at the threshold c where the incorrect rejection and false-alarm rates are equal; that is, $1 - P_V^c = P_F^c$. In the verification scenario, the lower EER value means better performance results.

In reporting verification scores, we state the size of the gallery G, which was the number of images in the gallery set G and the number of images in the probe set P. All galleries contained one image per person, and probe sets could contain more than one image per person. Probe sets did not necessarily contain an image of everyone in the associated gallery. For each probe p, there existed a gallery image g such that $p \sim g$.

5.3 Verification Results

5.3.1 Partially Automatic Algorithm Performance

We report verification scores for four categories of probes (see section 3.4.1). The verification results are reported on ROCs. Table 5.1 shows the categories

Table 5.1: Figures reporting verification results for partially automatic algorithms. Performance is broken out by probe category.

Figure No.	Probe category	Gallery size	Probe set size
5.1	**FB**	1196	1195
5.2	Duplicate I	1196	722
5.3	**fc**	1196	194
5.4	Duplicate II	864	234

corresponding to the figures presenting the results, type of results, and size of the gallery and probe sets. Figures 5.1 to 5.4 report the verification performance of four categories of probes: **FB**, duplicate I, **fc**, and duplicate II.

For each probe category, there are two ROCs. The first ROC reports results for the two baseline algorithms and the algorithms tested in September 1996. The second ROC reports results for the two baseline algorithms, the algorithms tested in March 1997, and the UMD algorithm tested in September 1996. For the upper bounds, we reported the algorithm with a minimum EER in Table 5.2. We also report the average and best EER for each probe category in Figures 5.5 and 5.6.

The verification performance of algorithms from a particular group will improve and the performance levels of face recognition algorithms in general will improve over time. Thus, one should not compare test results from different test dates. This is illustrated by the improvement in the performance of the UMD algorithm between September 1996 and March 1997. In consideration of this fact, we present results for September 1996 and March 1997 on different ROCs.

In Figure 5.5, we compare the difficulty of different probe sets. Whereas, Figures 5.1 to 5.4 report the verification performance for each algorithm, Figure 5.5 shows a single curve that is an average of the verification performance

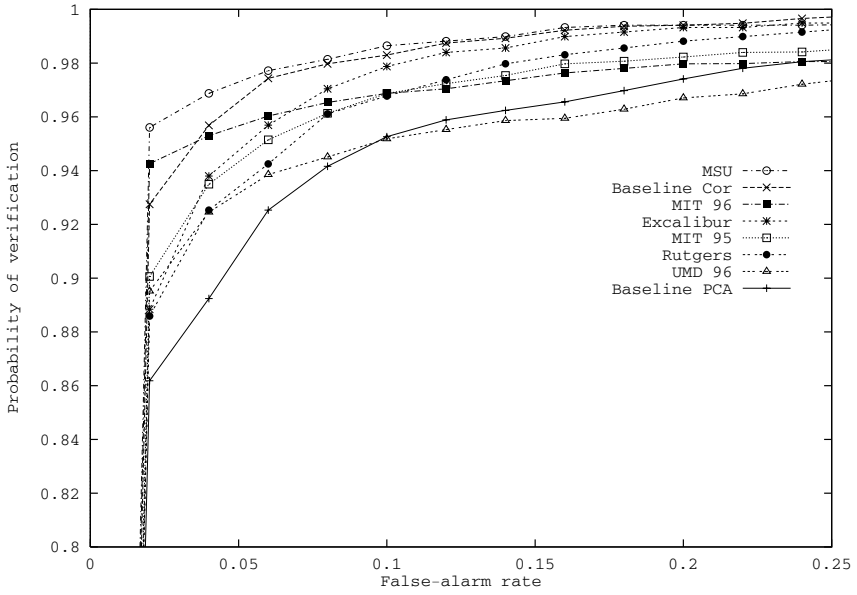
Table 5.2: Equal error rates (EER) by probe category.

	EER by probe category (%)			
Algorithms	**FB**	duplicate I	**fc**	duplicate II
Baseline PCA	7	19	15	22
Baseline correlation	4	21	23	27
Excalibur	5	16	14	24
MIT Mar95	5	20	25	26
MIT Sep96	4	20	26	26
MSU	3	23	11	31
Rutgers	6	18	17	21
UMD Sep96	7	22	16	23
UMD Mar97	1	12	8	14
USC	2	14	6	17
Average	4	19	16	23
Minimum	1	12	6	14

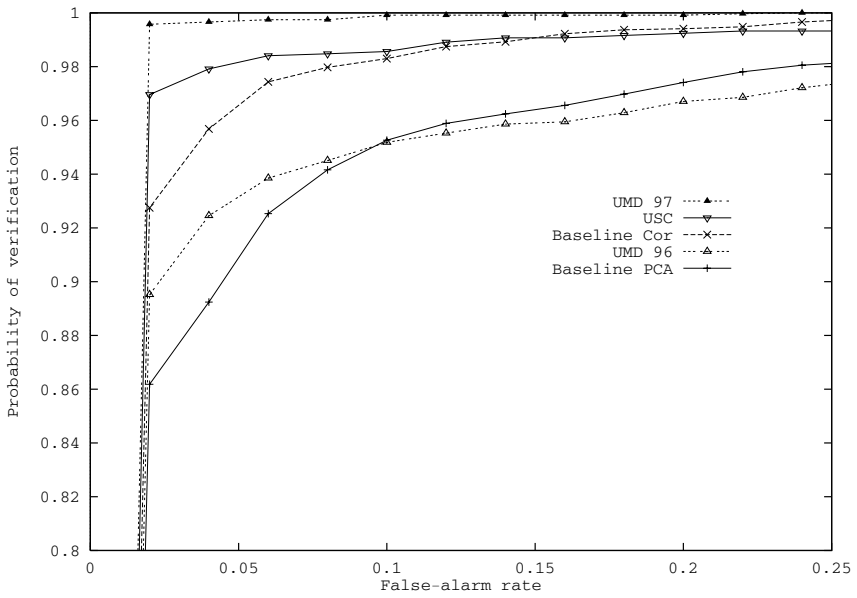
of all the algorithms. The average ROC is computed by averaging the P_V values for each P_F. The average performance provides an overall measure of the state of the art. For applications, one is interested in the currently achievable upper bound performance. In Figure 5.6, we present the current upper bound performance for each probe category in Figure 5.5.

5.3.2 Fully Automatic Algorithm Performance

In this subsection, we report on the verification performance for the fully automatic algorithms of the MIT Media Lab and USC. To allow for a comparison between the partially and fully automatic algorithms, we plot the results for the partially and fully automatic algorithms. Figure 5.7 shows the verification

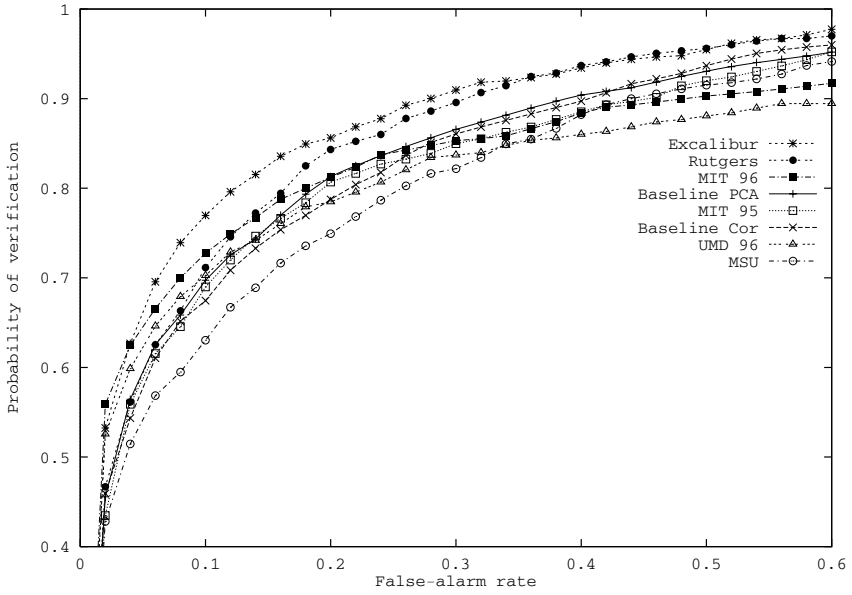


(a)

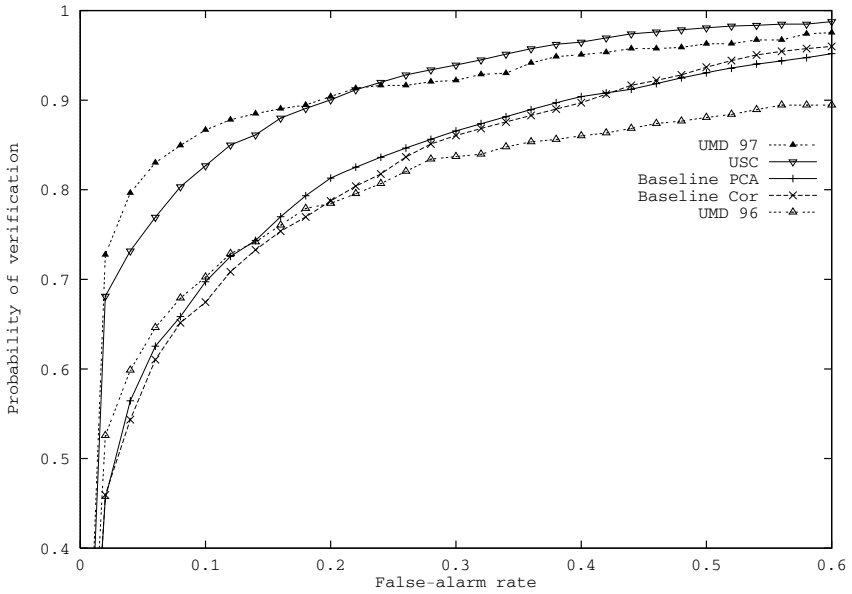


(b)

Figure 5.1: Verification performance for **FB** probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.

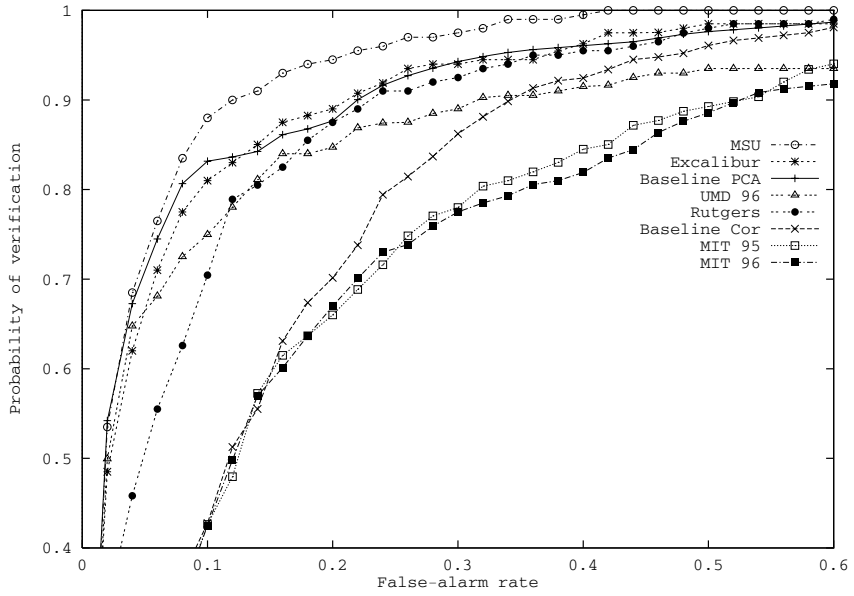


(a)

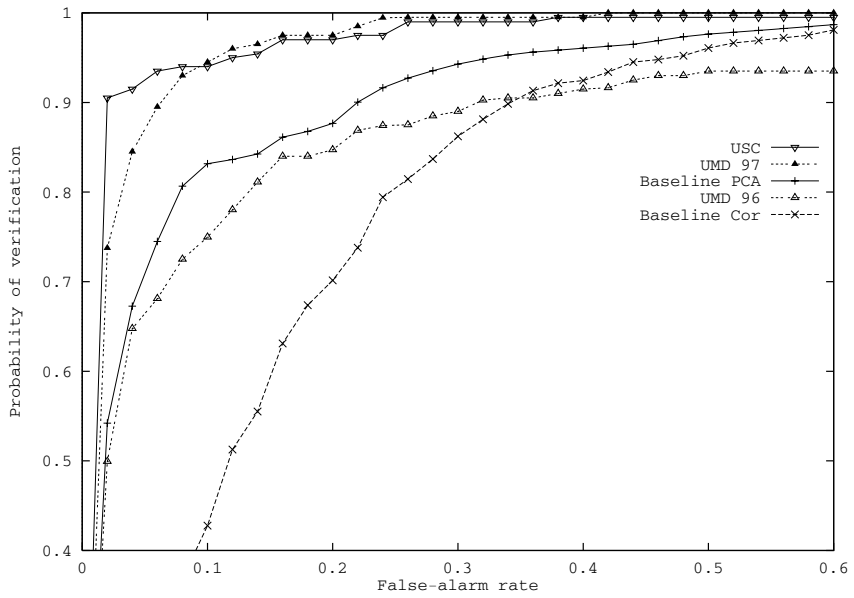


(b)

Figure 5.2: Verification performance for duplicate I probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.

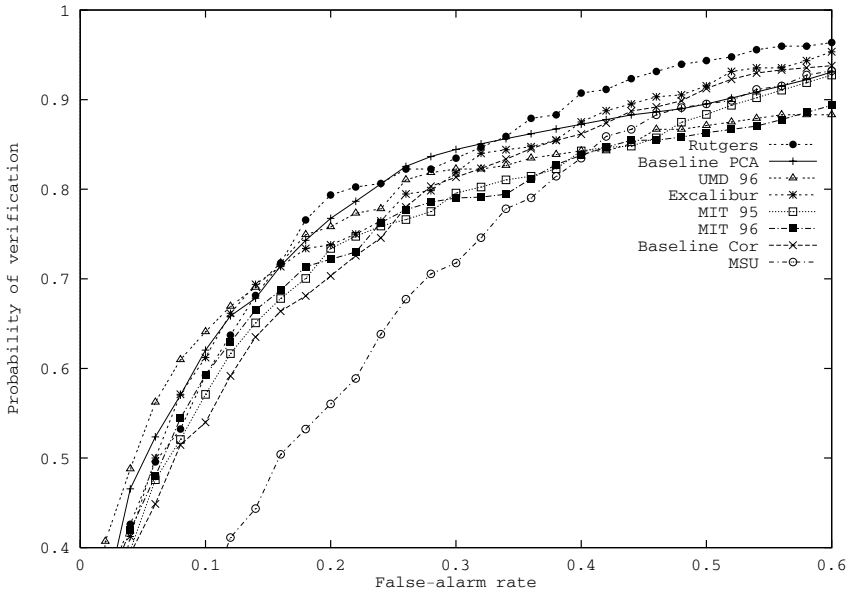


(a)

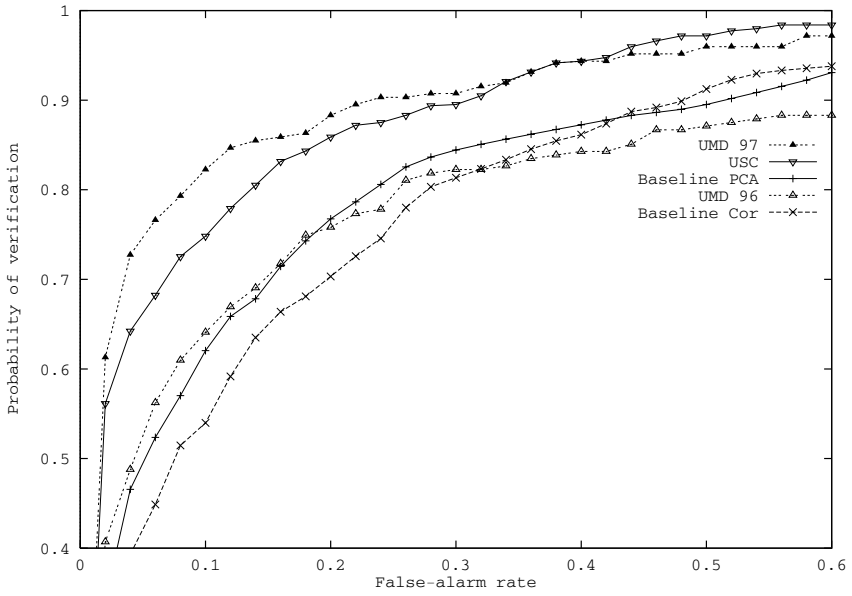


(b)

Figure 5.3: Verification performance for **fc** probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.



(a)



(b)

Figure 5.4: Verification performance for duplicate II probes. Partially automatic algorithms tested in (a) September 1996 and (b) March 1997.

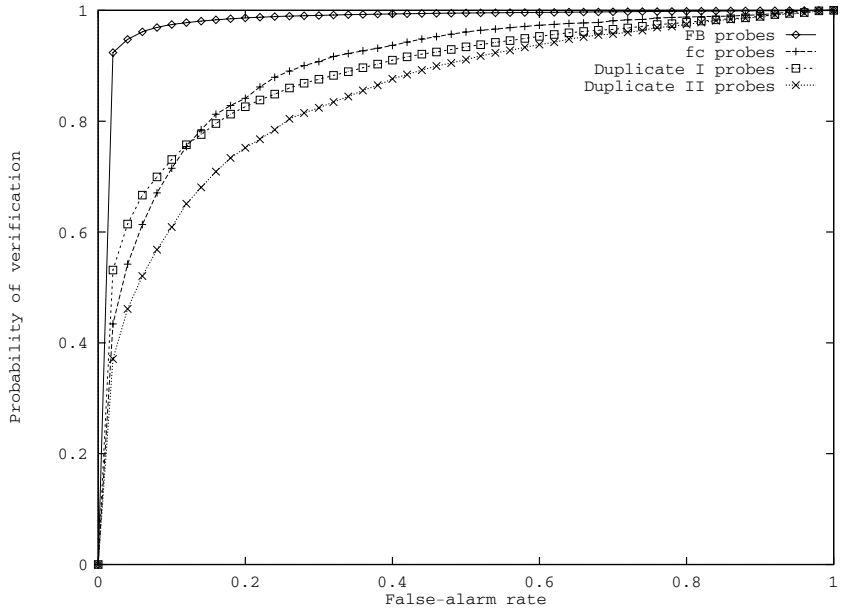


Figure 5.5: Average verification performance of partially automatic algorithms for each probe category.

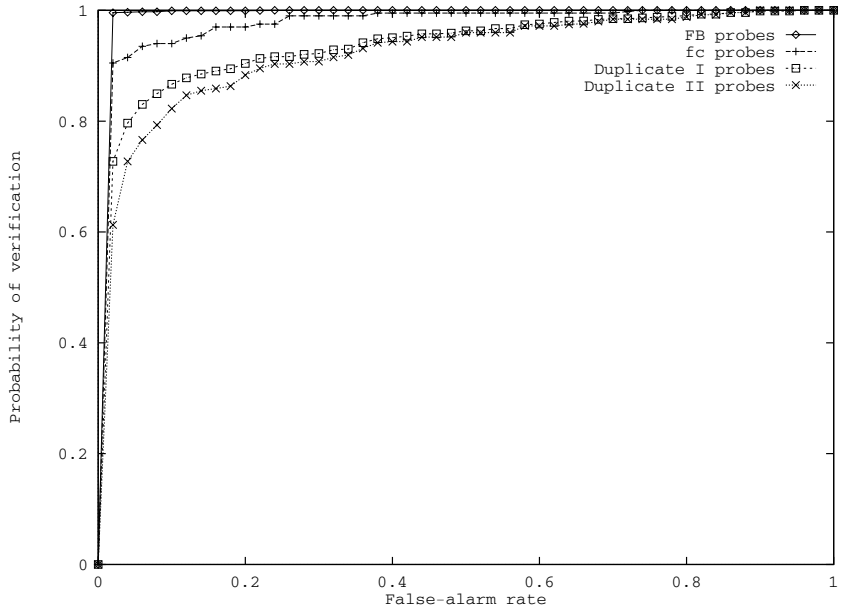


Figure 5.6: Current upper bound on verification performance of partially automatic algorithms for each probe category.

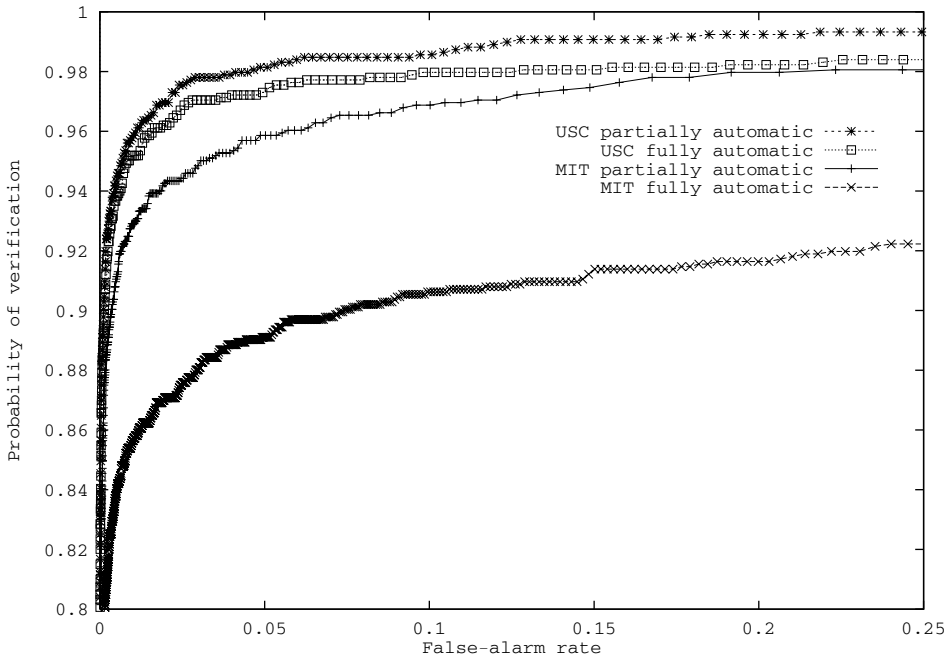


Figure 5.7: Verification performance of fully automatic algorithms against partially automatic algorithms for **FB** probes.

performance for **FB** probes and Figure 5.8 shows the verification performance for duplicate I probes. Additionally, Figure 5.9 shows the verification performance for **fc** probes and Figure 5.10 shows the verification performance for duplicate II probes.

5.4 Discussions and Conclusions

We have devised a verification scoring procedure for the new FERET test and reported results for this procedure. This allows for an independent assessment of face recognition algorithms in a key potential application.

This FERET test shows improvement in performance for both face recognition as a field and for individual algorithms. The improvement in the field is exhibited by the overall increase in performance of the algorithms tested between September 1996 and March 1997. An individual increase in performance

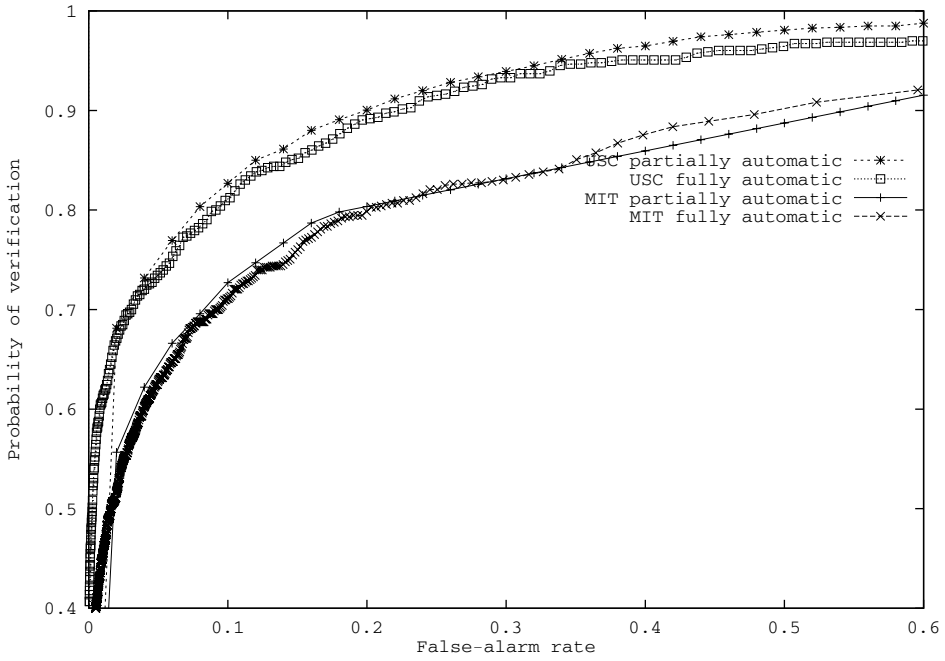


Figure 5.8: Verification performance of fully automatic algorithms against partially automatic algorithms for duplicate I probes.

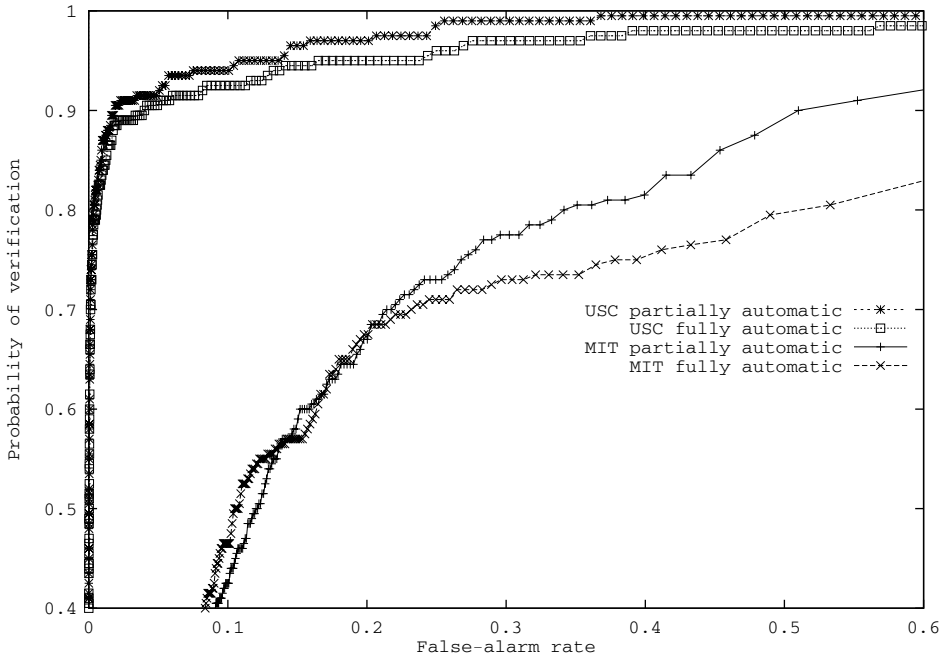


Figure 5.9: Verification performance of fully automatic algorithms against partially automatic algorithms for **fc** probes.

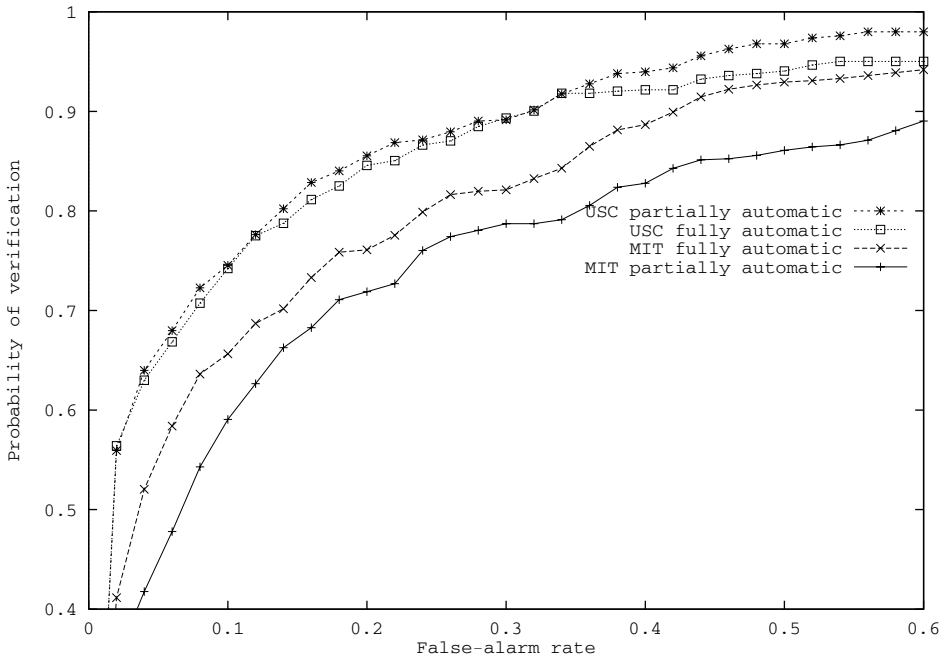


Figure 5.10: Verification performance of fully automatic algorithms against partially automatic algorithms for duplicate II probes.

is demonstrated by the improvement of the UMD algorithm. This increase shows that algorithms' performance should only be directly compared if they are tested at the same time. The September 1996 MIT algorithm was the top performer for the algorithms tested in September 1996. Among the algorithms tested in September 1996, no algorithm was among the top performers for all probe categories. This shows that relative performance on one task may not be predictive of relative performance on another task.

We broke out the performance for four categories of probes. Each category represents a different degree of difficulty. To estimate the degree of difficulty for each category, we compared the average and current upper bounds of performance for each category. For average performance, our results rank **FB** probes as easiest, duplicate II probes as most difficult, and **fc** and duplicate I probes as tied in the middle. For current upper bounds, duplicate I probes are more difficult than **fc** probes. Our results also show that we can expect that the best

performance will be significantly better than the average performance. Upper bound performance for all probe categories is superior to all average performance categories, except for **FB** probes.

The results in this chapter show that algorithm development is a dynamic process. Our evaluation methodology make an important contribution to face recognition and computer vision. This evaluation methodology will let researchers know the strengths and weaknesses of their algorithms. Thus, researchers will know where to concentrate their efforts to improve performance.

Chapter 6

Analysis of PCA-Based Face Recognition Algorithms

6.1 Introduction

Over the last several years, numerous face recognition algorithms have been developed based on principal component analysis (PCA). The main idea of PCA is to reduce the dimensionality of a data set while retaining most of the variation present in the data set [57]. PCA-based algorithms are the de facto benchmark for face recognition algorithms. Their popularity is due to their ease of implementation and their achievement of reasonable performance levels [84, 88, 96]. PCA serves as the basis for new face recognition algorithms [5, 35, 59, 67, 79, 111], a benchmark for comparison with new algorithms [7, 109, 121], and a computational model in psychophysics [47, 112, 115]. PCA-based algorithms have been applied in a broad spectrum of studies, including face detection [68, 107], face recognition [19, 24, 47, 111], and gender classification [25].

PCA is a statistical method for reducing the dimensionality of high dimensional data, where the data are represented as a vector. There is an accepted basic design for an algorithm built on PCA. However, the details of the basic

algorithm require a number of design decisions. These design decisions include detection of faces from a scene, preprocessing of facial images, feature extraction to represent a face, and the similarity measure for comparing faces. Each of these design decisions has an impact on the overall performance of the algorithm.

Some of these design decisions have been explicitly stated in the literature; e.g., the similarity measure for comparing two faces. However, a large number of decisions are not mentioned and are passed from researcher to researcher by word of mouth. For example, the illumination normalization and number of eigenfeatures that one chooses to include in a representation. Because the design details are not explicitly stated, a reader cannot assess the merits of a particular implementation and the associated claims. This can unnecessarily cast a shadow on the performance claims of a new algorithm when a PCA-based algorithm is used as a benchmark. Knowledge of the basic strengths and weaknesses of different implementations can provide insight and guidance in developing algorithms that build on PCA.

In this chapter, we present a generic modular PCA-based face recognition system. Our PCA-based face recognition system consists of normalization, PCA projection, and recognition modules. Each module consists of a series of basic steps, where the purpose of each step is fixed. However, we systematically vary the algorithm in each step. For example, the classifier step will always recognize a face, but we experiment with different classifiers. The selection of which algorithm is in each step is a design decision.

Based on the generic model for PCA-based algorithms, we evaluate different implementations. Because we use a generic model, we can change the implementation in an orderly manner and assess the impact on performance of each modification. We report identification and verification performance scores for each category of probes. We report performance results using top rank score for identification and equal error rate (EER) for verification. The algorithms are

evaluated with the September 1996 FERET testing procedure [84].

In experiment I, we performed a detailed evaluation of variations in the implementation. By testing on standard galleries and probe sets, the reader can compare the performance of our PCA implementations with the algorithms tested under the FERET program. In this experiment, we vary the illumination normalization procedure, the number of eigenvectors in the representation, and the similarity measure and we study the effects of compressing facial images on algorithm performance. The effects of image compression on recognition is of interest in applications where image storage space or image transmission time are critical parameters.

In algorithm evaluation, two critical questions are often ignored. First, how does performance vary with different galleries and probe sets. Second, when is a difference in performance between two algorithms statistically significant. In experiment two, we look at this question by randomly generating 100 galleries of the same size. We then calculate the performance on each of the galleries against **fb** and duplicate probes. Because we have 100 scores for each probe category, we can examine the range of scores and the overlap in scores among different implementations of the PCA algorithm.

6.2 PCA-Based Face Recognition System

6.2.1 Principal Component Analysis

Principal component analysis (PCA), which is also referred to as the Hotelling transform or the discrete Karhunen-Loève transform, is based on statistical properties of vector representations. It has several useful properties, such as decorrelation of data and optimization of compression error [41]. Kirby and Sirovich [59, 106] applied PCA to representing faces and Turk and Pentland [111] extended PCA to recognizing faces. We provide a brief summary of the funda-

mental theory of PCA below. (For further details on PCA, see Fukunaga [37] or Jolliffe [57]).

Assume a population of random vectors of the form

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} . \tag{6.1}$$

The *mean vector* and the *covariance matrix* of the vector population $\mathbf{x}$ are defined as

$$\mathbf{m_x} = E\{\mathbf{x}\} , \tag{6.2}$$

$$\mathbf{C_x} = E\{(\mathbf{x} - \mathbf{m_x})(\mathbf{x} - \mathbf{m_x})^T\} , \tag{6.3}$$

where $E\{\arg\}$ is the expected value of the argument, and T indicates vector transposition. Because $\mathbf{x}$ is n-dimensional, $\mathbf{C_x}$ is a matrix of order $n \times n$. Element c_{ii} of $\mathbf{C_x}$ is the variance of x_i, the ith component of the $\mathbf{x}$ vectors in the population, and element c_{ij} of $\mathbf{C_x}$ is the covariance between elements x_i and x_j of these vectors. The matrix $\mathbf{C_x}$ is real and symmetric. If elements x_i and x_j are uncorrelated, their covariance is zero, and therefore $c_{ij} = c_{ji} = 0$. For N vector samples from a random population, the mean vector and covariance matrix can be approximated from the samples by

$$\mathbf{m_x} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k \tag{6.4}$$

$$\mathbf{C_x} = \frac{1}{N} \sum_{k=1}^{N} (\mathbf{x}_k \mathbf{x}_k^T - \mathbf{m_x} \mathbf{m_x}^T) . \tag{6.5}$$

Because $\mathbf{C_x}$ is real and symmetric, we can always find a set of n orthonormal eigenvectors for this covariance matrix. A simple but foolproof algorithm to find these orthonormal eigenvectors for all real symmetric matrices is the Jacobi method [90]. The Jacobi algorithm consists of a sequence of orthogonal similarity transformations. Each transformation is just a plane rotation designed to annihilate one of the off-diagonal matrix elements. Successive transformations undo previously set zeros, but the off-diagonal elements get smaller and smaller, until the matrix is effectively diagonal (to the precision of the computer). We obtain the eigenvectors by accumulating the product of transformations during the process, while the main diagonal elements of the final diagonal matrix are the eigenvalues.

Let $\mathbf{e}_i$ and $\lambda_i, i = 1, 2, \ldots, n$, be the eigenvectors and the corresponding eigenvalues of $\mathbf{C_x}$, sorted in a descending order so that $\lambda_j \geq \lambda_{j+1}$ for $j = 1, 2, \ldots, n-1$. Let $\mathbf{A}$ be a matrix whose rows are formed from the eigenvectors of $\mathbf{C_x}$, such that

$$\mathbf{A} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_n \end{bmatrix} . \tag{6.6}$$

This $\mathbf{A}$ matrix can be used as a transformation matrix that maps the $\mathbf{x}$'s into vectors denoted by $\mathbf{y}$'s, as shown below:

$$\mathbf{y} = \mathbf{A}(\mathbf{x} - \mathbf{m_x}) . \tag{6.7}$$

Equation 6.7 is called the *Hotelling transform*. The $\mathbf{y}$ vectors resulting from this transformation have a zero mean vector; that is, $\mathbf{m_y} = \mathbf{0}$. The covariance matrix of the $\mathbf{y}$'s can be computed from $\mathbf{A}$ and $\mathbf{C_x}$ by

$$\mathbf{C_y} = \mathbf{A}\mathbf{C_x}\mathbf{A}^T \; . \tag{6.8}$$

Furthermore, $\mathbf{C_y}$ is a diagonal matrix whose elements along the main diagonal are the eigenvalues of $\mathbf{C_x}$; that is,

$$\mathbf{C_x} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & . & \\ & & & . \\ 0 & & & \lambda_n \end{bmatrix} . \tag{6.9}$$

Because the off-diagonal elements of $\mathbf{C_y}$ are zero, the elements of the $\mathbf{y}$ vectors are uncorrelated. Since the elements along the main diagonal of a diagonal matrix are its eigenvalues, $\mathbf{C_x}$ and $\mathbf{C_y}$ have the same eigenvalues and eigenvectors. In fact, the transformation of the $\mathbf{C_x}$ into $\mathbf{C_y}$ is the essence of the Jacobi algorithm described above.

Through the Hotelling transform, a new coordinate system is established. The origin of this new coordinate system is at the centroid of the population, $\mathbf{m_x}$, with new axes in the direction specified by the eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. The eigenvalue λ_i becomes the variance of component y_i along eigenvector $\mathbf{e}_i$. With its capability to realign unknown data into a new coordinate system based on the principal axes of the data, PCA is often used to achieve rotational invariance in image processing tasks.

On the other hand, we may want to reconstruct vector $\mathbf{x}$ from vector $\mathbf{y}$. Because the rows of $\mathbf{A}$ are orthonormal vectors, $\mathbf{A}^{-1} = \mathbf{A}^T$. Therefore, any vector $\mathbf{x}$ can be reconstructed from its corresponding $\mathbf{y}$ by the relation

$$\mathbf{x} = \mathbf{A}^T\mathbf{y} + \mathbf{m}_\mathbf{x} \, . \tag{6.10}$$

Instead of using all the eigenvectors of $\mathbf{C}_\mathbf{x}$, we may pick only K eigenvectors corresponding to the K largest eigenvalues and form a new transformation matrix, $\mathbf{A}_K$, of order $K \times n$. In this case, the resulting $\mathbf{y}$ vectors would be K-dimensional, and the reconstruction given in Equation 6.10 would no longer be exact. The reconstructed vector using $\mathbf{A}_K$ is

$$\hat{\mathbf{x}} = \mathbf{A}_K^T\mathbf{y} + \mathbf{m}_\mathbf{x} \, . \tag{6.11}$$

The mean square error between $\mathbf{x}$ and $\hat{\mathbf{x}}$ can be computed by the expression

$$\epsilon = \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{K} \lambda_j = \sum_{j=K+1}^{n} \lambda_j \, . \tag{6.12}$$

Because the λ_j's decrease monotonically, Equation 6.12 shows that we can minimize the error by selecting the K eigenvectors associated with the K largest eigenvalues. Thus the Hotelling transform is optimal in the sense that it minimizes the MSE between the vectors $\mathbf{x}$ and their approximations $\hat{\mathbf{x}}$.

In a PCA-based face recognition algorithm, the input is a training set $\mathbf{t}_1, ..., \mathbf{t}_W$ of N images such that the ensemble mean is zero ($\sum_i \mathbf{t}_i = 0$). Each image is interpreted as a point in $\Re^{n \times m}$, where the image is n by m pixels. PCA finds a representation in a $(W-1)$ dimensional space that preserves variance. PCA generates a set of $N-1$ eigenvectors $(e_1, ..., e_{N-1})$ and eigenvalues $(\lambda_1, ..., \lambda_{N-1})$. (In the face recognition literature, the eigenvectors can be referred to as *eigenfaces*.) We normalize the eigenvectors so that they are orthonormal. The eigenvectors are ordered so that $\lambda_i > \lambda_{i+1}$. The λ_i's are equal to the variance of the projection of the training set onto the ith eigenvector. The low order eigenvectors

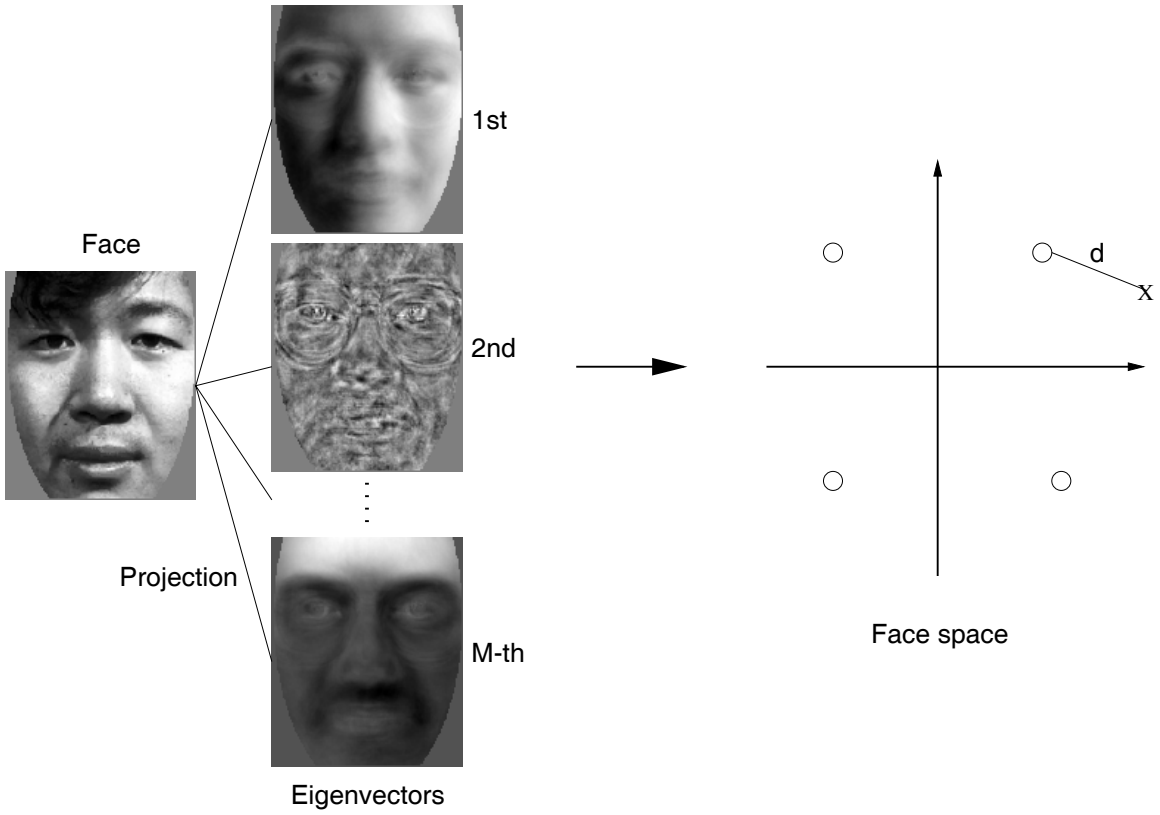


Figure 6.1: Representation of face as a point in face space. A face is represented by its projection onto a subset of M eigenvectors (the PCA generates a set of $N-1$ eigenvectors from N training images).

encode the larger variations in the training set (low order refers to the index of the eigenvectors and eigenvalues). The face is represented by its projection onto a subset of M eigenvectors, which we will call *face space* (see Figure 6.1). Thus the normalized face is represented as a point in an M dimensional face space. The dimensionality reduction is achieved when the face has been projected into the eigenvectors.

6.2.2 System Modules

Our face recognition system consists of three modules and each module is composed of a sequence of steps (see Figure 6.2). The first module normalizes

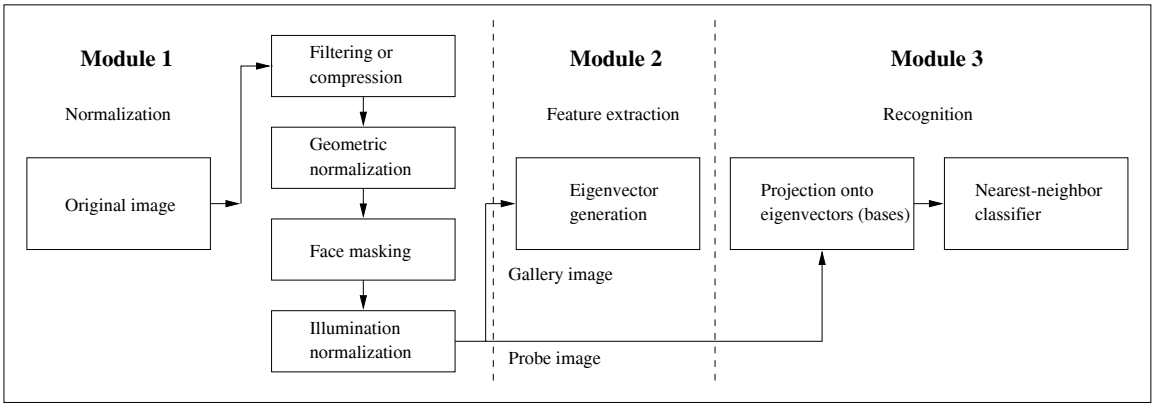


Figure 6.2: Block diagram of PCA-based face recognition system.

the input image. The goal of normalization is to transform facial images into a standard format that removes variations that can affect recognition performance. This module consists of four steps. Figure 6.3 shows the input and output of some of the steps in the normalization module.

The first step filters or compresses the original image. The image is filtered to remove high frequency noise in the image. An image is compressed to save storage space and reduce transmission time. The second step places the face in a standard geometric position by rotating, scaling, and translating the center of the eyes to a standard location. Even for the cooperating subjects, it is difficult to sustain the same position of the head and the same distance from the acquisition camera time after time. Therefore, integration of invariance to such changes is a compulsory part of any face recognition system. However, the extent of rotation and scaling depends on a particular application. The goal of this step is to remove variations in size, orientation, and location of the face. The third step masks out background pixels, hair, and clothes to remove unnecessary variations that can interfere with the identification process. The fourth module removes some of the variations in illumination between images. Changes in illumination are critical factors in algorithm performance. Variations in lighting conditions generally affect the image structure, making it more difficult to trace the original features. Integration of the invariance to changes in illumination

Figure 6.3: Input and output images of the normalization module.

improves the recognition performance.

The second module performs the PCA decomposition on the training set. This produces the eigenvectors (eigenfaces) and eigenvalues. We did not vary this module because we use the training set that was used for the FERET program for the generation of eigenvectors [84]. (The mathematical representation of the PCA algorithm for eigenvector generation is described in the appendix, section A.3.)

The third module identifies the face from a normalized image, and consists of two steps. The first step projects the image onto the eigen representation. The critical parameter in this step is the subset of eigenvectors that represent the face. The second step recognizes faces using a nearest-neighbor classifier. The critical design decision in this step is the similarity measure in the classifier. We presented performance results using $L1$ distance, $L2$ distance, angle between feature vectors, and Mahalanobis distance. Additionally, Mahalanobis distance was combined with $L1$, $L2$, and angle between feature vectors mentioned above.

6.3 Experiment I

The purpose of experiment I is to examine the effects of changing the steps in our generic PCA-based face recognition system. We do this by establishing a baseline algorithm and then varying the implementation of selected steps one

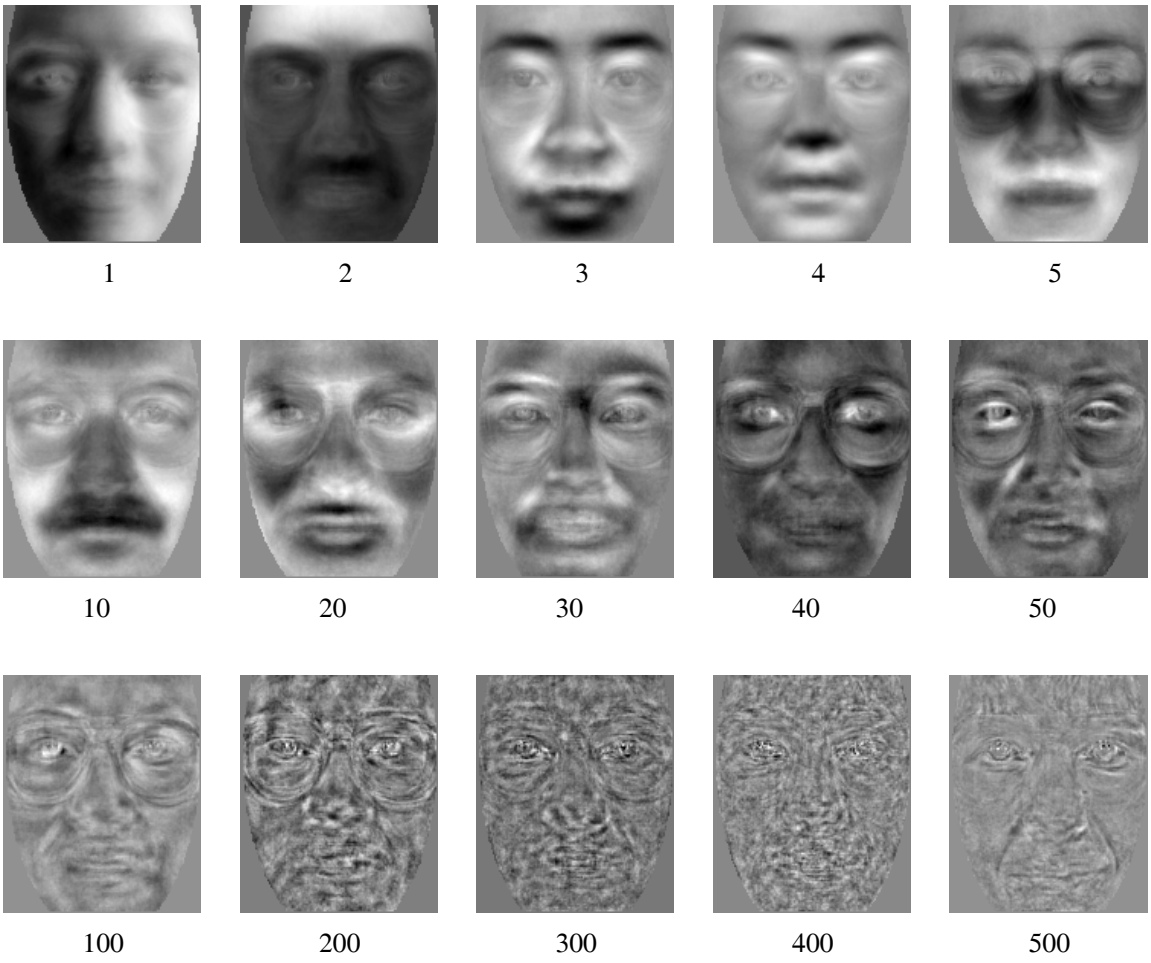


Figure 6.4: Examples of eigenvectors (eigenfaces) from PCA. The number below each image represents the order of eigenvalues.

at a time. Ideally, we would test all possible combinations of the variations. However, because of the number of combinations, this is not practical and we vary the steps individually.

The baseline algorithm has the following configuration: The images are not filtered or compressed. Geometric normalization consists of rotating, translating, and scaling the images so that the center of the eyes are on standard pixels. This is followed by masking the hair and background from the images. In the illumination normalization step, the nonmasked facial pixels were normalized

Table 6.1: Identification performance results for illumination normalization methods. Performance scores are the top rank match.

	Probe category			
Illumination normalization	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline	0.35	0.13	0.77	0.26
Original image	0.32	0.11	0.75	0.21
Histogram eq. only	0.34	0.12	0.77	0.24
$\mu = 0.0$, $\sigma = 1.0$ only	0.33	0.14	0.76	0.25

Table 6.2: Verification performance results for illumination normalization methods. Performance scores are equal error rate (EER).

	Probe category			
Illumination normalization	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline	0.24	0.30	0.07	0.13
Original image	0.25	0.31	0.07	0.14
Histogram eq. only	0.25	0.30	0.07	0.13
$\mu = 0.0$, $\sigma = 1.0$ only	0.25	0.29	0.07	0.14

by a histogram equalization algorithm. Then, the nonmasked facial pixels were transformed so that the mean is equal to 0.0 and the standard deviation is equal to 1.0. The geometric normalization and masking steps are not varied in the experiments.

The training set for the PCA consists of 501 images (one image per person), which produces 500 eigenvectors. The training set is not varied in this experiments. In the recognition module, faces are represented by their projection onto the first 200 eigenvectors and the classifier uses the L_1 norm. In Figure 6.4, we

present some examples of the eigenvectors (eigenfaces).

6.3.1 Variations in the Normalization Module

A. Illumination Normalization

We experimented with three variations to the illumination normalization step. For the baseline algorithm, the nonmasked facial pixels were transformed so that the mean was equal to 0.0 and the standard deviation was equal to 1.0 followed by a histogram equalization algorithm (for details, see the appendix, section A.2). In the first variation, the nonmasked pixels were not normalized (original image). For the second variation, the nonmasked facial pixels were normalized with a histogram equalization algorithm [40]. For the third variation, the nonmasked facial pixels were transformed so that the mean was equal to 0.0 and variance equal to 1.0. The identification and verification performance results from the illumination normalization methods are presented in Table 6.1 and 6.2.

B. Compressing and Filtering the Images

We examined the effects of JPEG and wavelet compression and low pass filtering (LPF) on recognition. For this experiment, the original images were compressed and then uncompressed before being fed into the geometric normalization step of the normalization module. For both compression methods,

0.1	0.1	0.1
0.1	0.2	0.1
0.1	0.1	0.1

Figure 6.5: A 3 x 3 mask showing actual coefficients for low pass filtering.

Table 6.3: Identification performance scores for low pass filter and JPEG and wavelet compressed images (0.5 bits/pixel compression). Performance scores are the top rank match.

	Probe category			
Normalization	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline	0.35	0.13	0.77	0.26
JPEG	0.35	0.13	0.78	0.25
Wavelet	0.36	0.15	0.79	0.25
LPF	0.36	0.15	0.79	0.24

Table 6.4: Verification performance score for low pass filter and JPEG and wavelet compressed images (0.5 bits/pixel compression). Performance scores are equal error rate (EER).

	Probe category			
Normalization	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline	0.24	0.30	0.07	0.13
JPEG	0.24	0.29	0.06	0.13
Wavelet	0.23	0.29	0.07	0.13
LPF	0.23	0.28	0.07	0.13

the images were compressed approximately 16:1 (0.5 bits per pixel). We experimented with other compression ratios and found that performance was comparable. The results are for eigenvectors generated from noncompressed images. We found that performance in this case was slightly better than on eigenvectors trained from compressed images. Because compression algorithms usually low pass filter the images, we decided to examine the effects on performance of low pass filtering the original image. The filter was a 3 x 3 spatial filter with a center

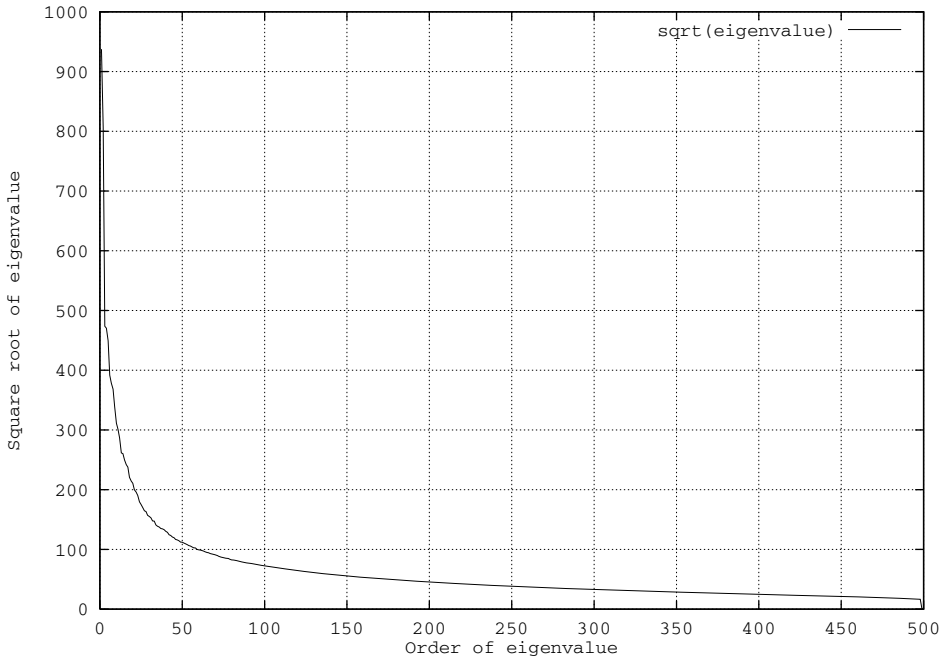


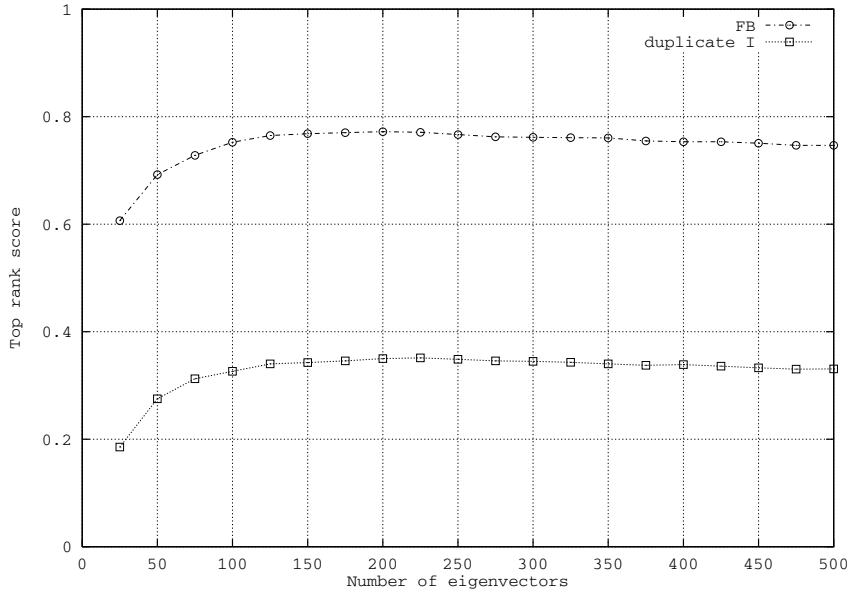
Figure 6.6: Distribution of eigenvalues based on their order.

value of 0.2 and the remaining values equal to 0.1 (see Figure 6.5). Table 6.3 and 6.4 report identification and verification performances for the baseline algorithm, JPEG and wavelet compression, and low pass filtering.

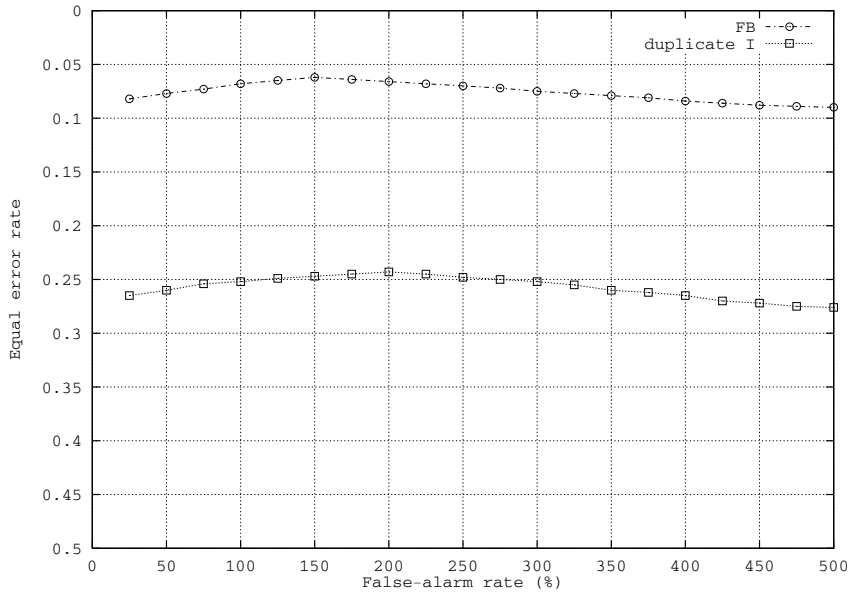
6.3.2 Variations in the Recognition Module

A. Number of Low Order Eigenvectors

The higher order eigenvectors that are associated with smaller eigenvalues encode small variations and noise among the images in the training set. One would expect from the exponentially decreasing eigenvalues that the higher order eigenvectors would not contribute to recognition (see Figure 6.6). We examined this hypothesis by computing performance as a function of the number of low order eigenvectors in the representation. Figure 6.7 shows (a) the top rank score and (b) the equal error rate for **FB** and duplicate I probes as the function of the number of low order eigenvectors included in the representation in face



(a)



(b)

Figure 6.7: Performance on **FB** and duplicate I probes based on number of low order eigenvectors used: (a) Identification and (b) verification performance score. (For verification, the y-axis (EER) is reversed so that the top equals 0.)

Table 6.5: Identification performance score with low order eigenvectors removed. Performance scores are the top rank match.

Number of low order eigenvectors removed	Probe category			
	Duplicate I	Duplicate II	**FB** probe	**fc** probe
0 (Baseline)	0.35	0.13	0.77	0.26
1	0.35	0.15	0.75	0.38
2	0.34	0.14	0.74	0.36
3	0.31	0.14	0.72	0.37

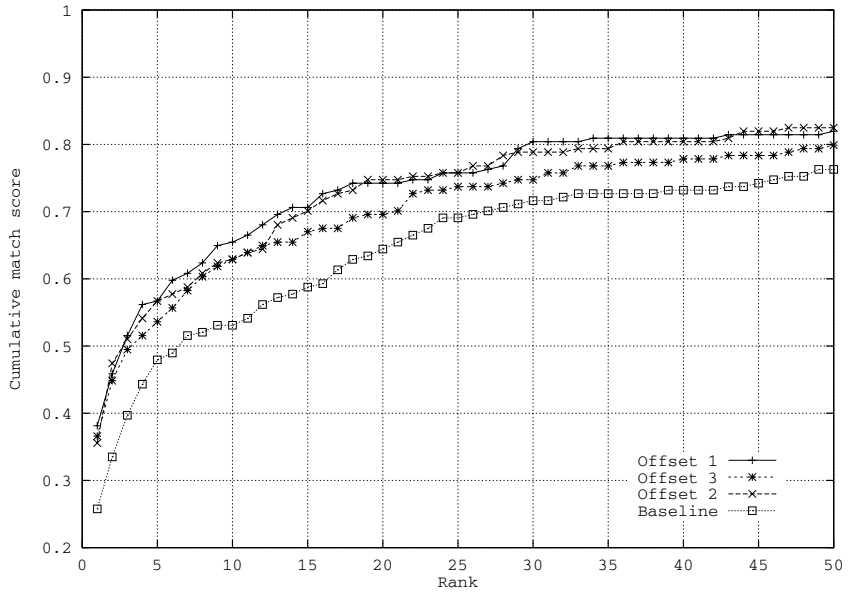
Table 6.6: Verification performance scores with low order eigenvectors removed. Performance scores are equal error rate (EER).

Number of low order eigenvectors removed	Probe category			
	Duplicate I	Duplicate II	**FB** probe	**fc** probe
0 (Baseline)	0.24	0.30	0.07	0.13
1	0.21	0.23	0.08	0.15
2	0.23	0.25	0.10	0.14
3	0.22	0.23	0.11	0.13

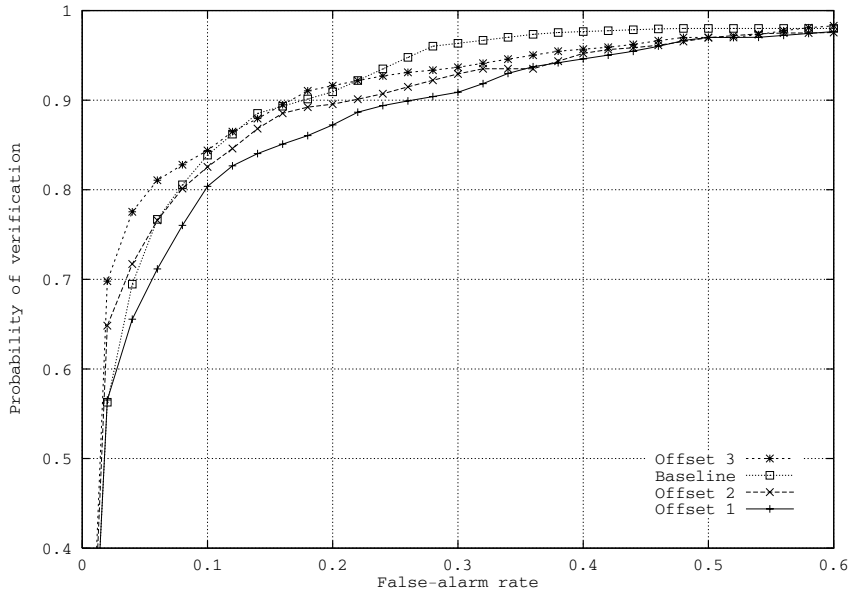
space. The representation consisted of $e_1, \ldots, e_n,\ n = 50, 100, \ldots, 500$, where e_is are the eigenvectors generated by the PCA decomposition.

B. Removing Low Order Eigenvectors

The low order eigenvectors encode gross differences within the training set. If the low order eigenvectors encode variations such as lighting changes, then performance may improve by removing the low order eigenvectors from the representation. We looked at this hypothesis by removing the first, second and



(a)



(b)

Figure 6.8: Performance on **fc** probes with first one, two, and three low order eigenvectors removed: (a) Identification and (b) verification performance score.

third eigenvector from the representation; i.e., the representation consisted of $e_i, \ldots, e_{200},\ i = 1, 2, 3, 4$. The identification and verification performance results from these variations are given in Tables 6.5 and 6.6. Table 6.6 shows that removing the first three eigenvectors resulted in an overall increase in verification performance of duplicate I and duplicate II probes. This increase is further highlighted in Figure 6.8 (b). Because there was a noticeable variation in performance for the **fc** probes among the different categories of probes, we report the cumulative match score and ROC for **fc** probes (see Figure 6.8).

C. Nearest-Neighbor Classifier

We experimented with seven similarity measures for the classifier. Their identification and verification performance results are listed in Tables 6.7 and 6.8. Details of the similarity measures are given in the appendix, section A.4. The performance scores for the **fc** probes show the most variation among the different categories of probes. In Figure 6.9, we report detailed identification and verification performance results for **fc** probes.

6.3.3 Discussions

In experiment I, we conducted a series of experiments that systematically varied the steps in each module based on our PCA-based face recognition system. The goal was to help to understand the effects of these variations on performance scores.

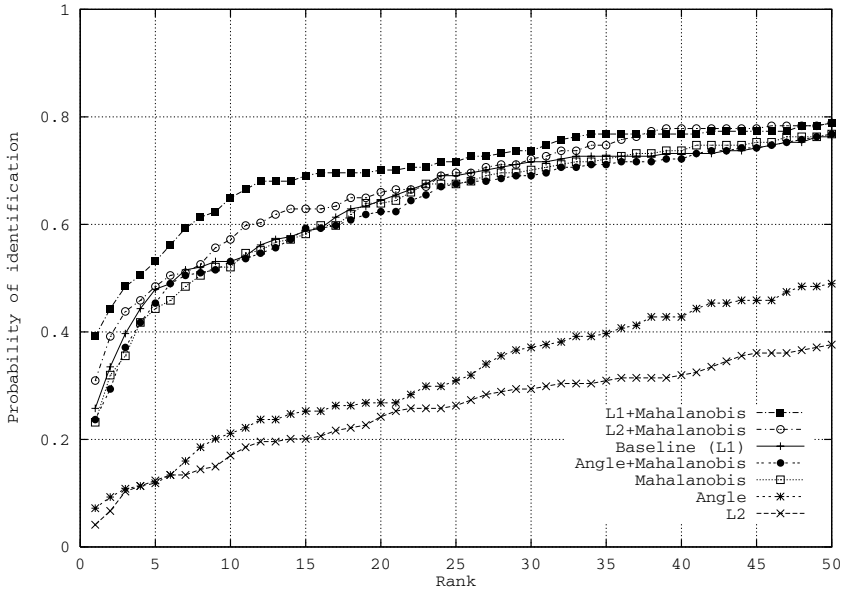
In the normalization module, we varied the illumination normalization and compression steps. The results show that performing an illumination normalization step improves identification performance (see Table 6.1), but which implementation is selected is not critical (see Table 6.2). The results also show that compressing or filtering the images does not significantly affect performance (see Tables 6.3 and 6.4).

Table 6.7: Identification performance scores based on different nearest-neighbor classifier. Performance scores are the top rank match.

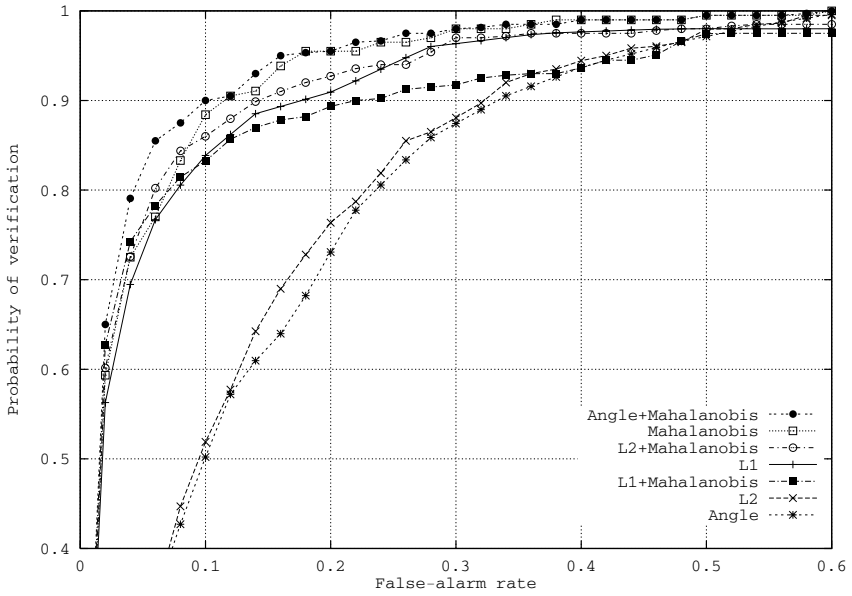
	Probe category			
Nearest-neighbor classifier	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline (L_1)	0.35	0.13	0.77	0.26
Euclidean (L_2)	0.33	0.14	0.72	0.04
Angle	0.34	0.12	0.70	0.07
Mahalanobis	0.42	0.17	0.74	0.23
L_1 + Mahalanobis	0.31	0.13	0.73	0.39
L_2 + Mahalanobis	0.35	0.13	0.77	0.31
Angle + Mahalanobis	0.45	0.21	0.77	0.24

Table 6.8: Verification performance scores based on different nearest-neighbor classifier. Performance scores are equal error rate (EER).

	Probe category			
Nearest-neighbor classifier	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline (L_1)	0.24	0.30	0.07	0.13
Euclidean (L_2)	0.21	0.26	0.05	0.22
Angle	0.19	0.22	0.05	0.22
Mahalanobis	0.11	0.12	0.04	0.11
L_1 + Mahalanobis	0.34	0.39	0.12	0.13
L_2 + Mahalanobis	0.25	0.30	0.07	0.12
Angle + Mahalanobis	0.11	0.12	0.03	0.10



(a)



(b)

Figure 6.9: Effects of nearest-neighbor classifier on performances for **fc** probes: (a) Identification and (b) verification performance score.

In the recognition module, we experimented with three classes of variations. First, we varied the number of low order eigenvectors in the representation from 50 to 500 by steps of 50. In Figure 6.6, the eigenvalues decrease exponentially. Figure 6.7 shows that performance increases until approximately 150 to 200 eigenvectors are in the representation, and then performance decreases slightly. Representing faces by the first 30 to 40% of the eigenvectors is consistent with results on other facial image sets that the authors have seen.

Second, low order eigenvectors were removed. Table 6.5 shows that removing the first eigenvector resulted in an overall increase in identification performance. For the identification performance, the largest increase was observed with the **fc** probes. This increase is further highlighted in Figure 6.8 (a). The low order eigenvectors encode the greatest variations among the training set. The most significant difference between the **fc** probes and the gallery images was a change in lighting. If the low order eigenvectors encode lighting differences, then this would explain the substantial increase in performance by removing the first eigenvector.

Third, the similarity measure in the nearest-neighbor classifier was changed. This variation showed the largest range of identification and verification performance. In Table 6.7, the identification performance of duplicate I probes performance ranged from 0.31 to 0.45, and for **fc** probes ranged from 0.07 to 0.39. In Table 6.8, the verification performance of duplicate I probes ranged from 0.11 to 0.34, and for **fc** probes ranged from 0.10 to 0.22. For duplicate I, duplicate II, and **FB** probes, the angle+Mahalanobis distance performed the best. For the **fc** probes, the L_1+Mahalanobis distance performed the best for identification and the angle+Mahalanobis distance performed the best for verification (see Figure 6.9). Because of the range of performance, it is clear that selecting the similarity measure for the classifier is the critical decision in designing a PCA-based face recognition system. However, the design decision is dependent on the type of images in the galleries and the probe sets that the system will process.

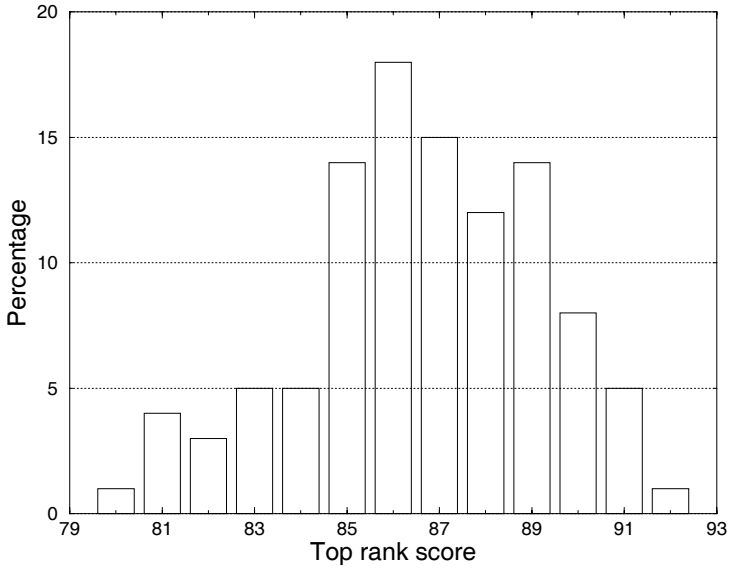
6.4 Experiment II

6.4.1 Variations in Galleries and Probe Set

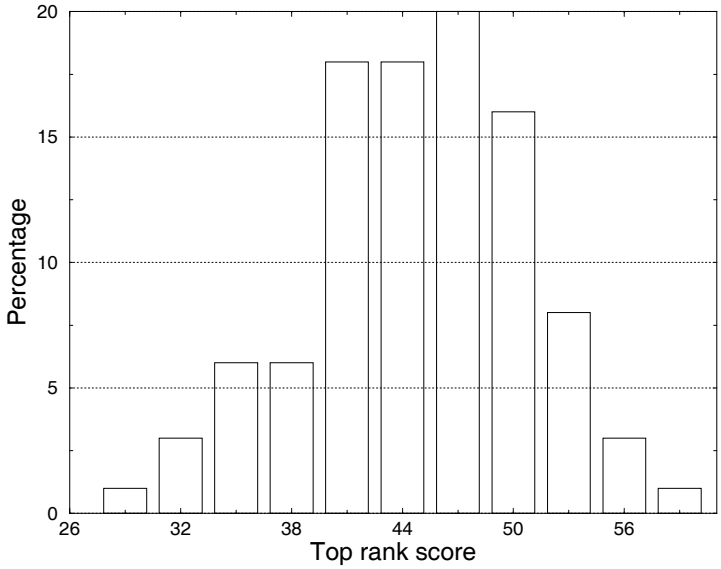
The comparison among algorithms in experiment I is based on the algorithm performance on four probe sets. The performance among the different probe sets cannot be directly compared since the number of probes in each category is different. The natural question is, "When is the difference in performance between two classifiers significant?"

To address this question, we randomly generated 100 galleries of 200 individuals, with one frontal view image per person. The galleries were generated without replacement from the **FB** gallery of 1,196 individuals in experiment I. Then we scored each of the galleries against the **FB** and duplicate I probes for each of the seven classifiers in experiment I. (There were not enough **fc** and duplicate II probes to compute performances for these categories.) For each randomly generated gallery, the corresponding **FB** probe set consisted of the second frontal view image for all images in that gallery; the duplicate I probe set consisted of all duplicate images in the database for each image in the gallery. We measured performance by the top rank score (the fraction of probes that were correctly identified).

For an initial look at the range in performance, we examine the baseline algorithm (L_1 similarity measure). There are similar variations for the six remaining distances. For each classifier and probe category, we had 100 different scores. In Figures 6.10 and 6.11, we present the histogram of top rank scores and equal error rates (%) for the baseline algorithm for both the **FB** and the duplicate I probe sets. For the top rank score, performance ranges from 0.80 to 0.92 for the **FB** probe and from 0.29 to 0.59 for the duplicate I probe. For equal error rate, performance ranges from 4.6 to 8.2 for the **FB** probe and from 18.8 to 33.2 for the duplicate I probe. This clearly shows a large range in performance of the 100 galleries.

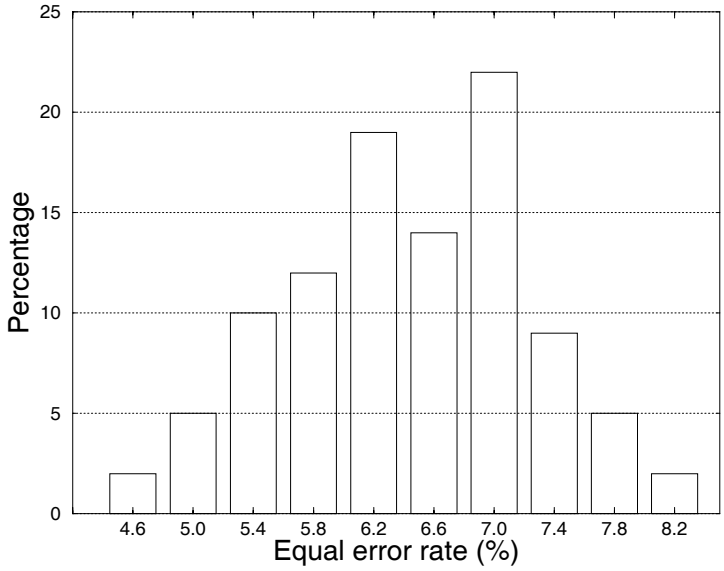


(a)

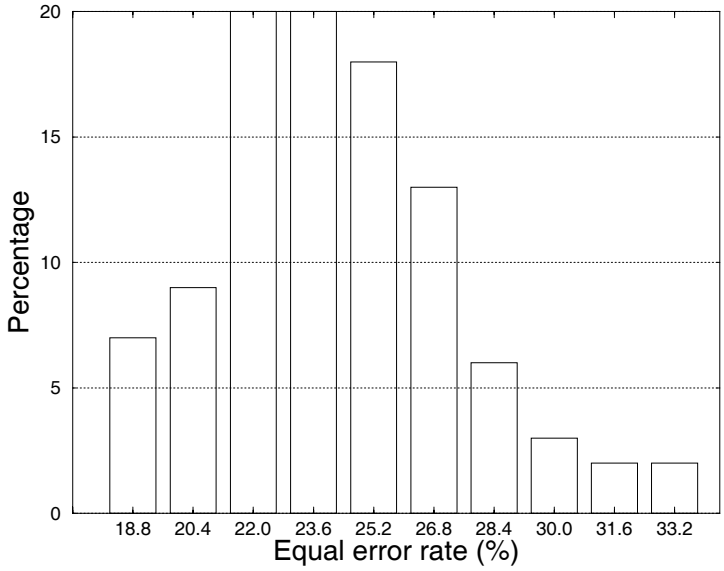


(b)

Figure 6.10: Histogram of top rank scores of the baseline algorithm ($L1$ similarity measure) (a) **FB** and (b) duplicate I probes.



(a)



(b)

Figure 6.11: Histogram of equal error rates (%) of the baseline algorithm ($L1$ similarity measure) (a) **FB** and (b) duplicate I probes.

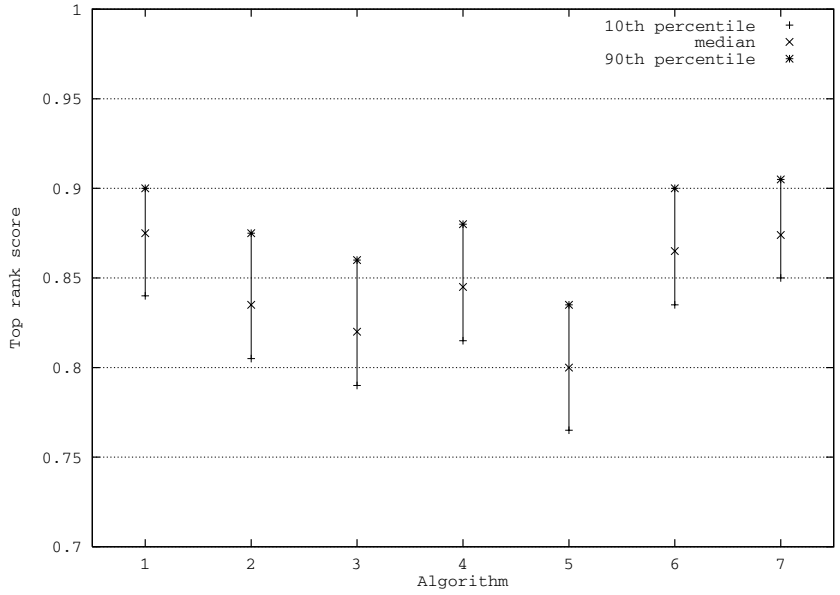
In Figure 6.12 and 6.13, we reported a truncated range of top rank scores and equal error rates (%) for the seven different nearest-neighbor classifiers for both the **FB** and duplicate I probe sets. For each classifier, the score is marked with the median by $\times$, the 10th percentile by $+$, and 90th percentile by $*$. We plotted these values because they are robust statistics. We selected the 10th and 90th percentile because they mark a robust range of scores and outliers are ignored. From these results, we get a robust estimate of the overall performance of each classifier.

6.4.2 Discussions

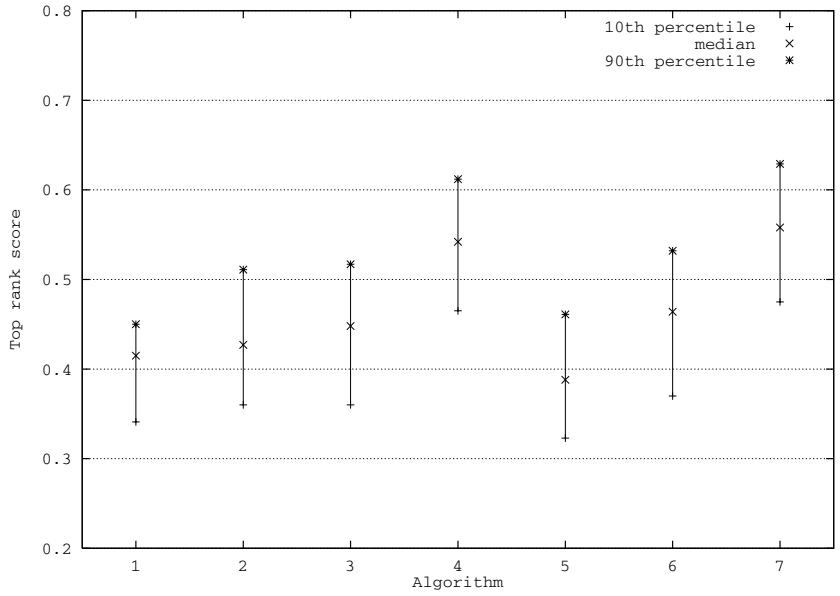
In experiment II, the main goal was to get a rough estimate of when the difference in performance is significant. From Figures 6.12 and 6.13, the range in identification and verification scores is approximately ± 0.06 about the median. This suggests that a reasonable threshold for measuring a significant difference in performance for the classifiers is ~ 0.12.

The results for duplicate I probes in experiment II are consistent with the results in experiment I. In Tables 6.7 and 6.8, the top classifiers were the Mahalanobis and angle+Mahalanobis. These two classifiers produce better performance than the other methods as shown in Figures 6.12 and 6.13. In both experiments, the L_1+Mahalanobis received the lowest identification and verification performance scores. This suggest that for duplicate I scores, the angle+Mahalanobis or Mahalanobis distance should be used. Based on the results of this experiment, the performance of smaller galleries can predict relative performance on larger galleries.

For the **FB** probes, there is not as sharp a division among classifiers. One possible explanation is that in experiment I, the top match scores for the **FB** probes did not vary as much as the duplicate I scores. There is consistency among the best scores (L_1, L_2+Mahalanobis, and angle+Mahalanobis). The re-

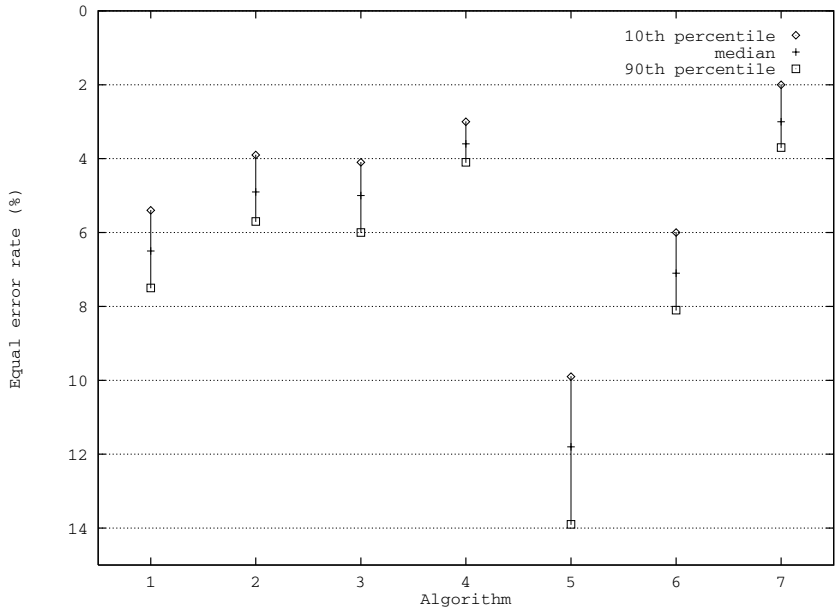


(a)

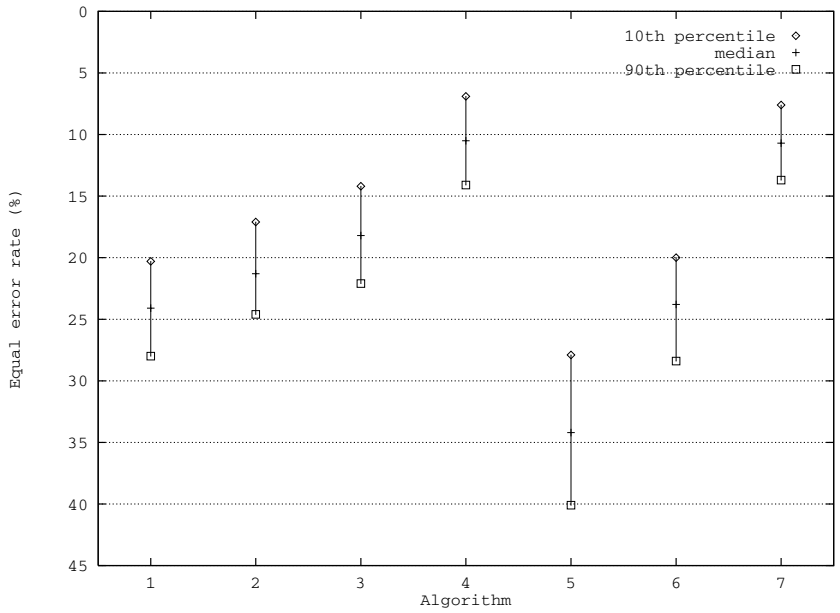


(b)

Figure 6.12: The range of top rank scores using seven different nearest-neighbor classifiers. The nearest-neighbor classifiers presented are (1) L_1, (2) L_2, (3) Angle, (4) Mahalanobis, (5) L_1+Mahalanobis, (6) L_2+Mahalanobis, and (7) Angle+Mahalanobis. (a) **FB** and (b) duplicate I probes.



(a)



(b)

Figure 6.13: The range of equal error rates (%) using seven different nearest-neighbor classifiers. The nearest-neighbor classifiers presented are (1) L_1, (2) L_2, (3) Angle, (4) Mahalanobis, (5) L_1+Mahalanobis, (6) L_2+Mahalanobis, and (7) Angle+Mahalanobis. (a) **FB** and (b) duplicate I probes.

maining classifiers' performances can be grouped together. The performance scores of these classifiers are within each other's error margins. We defined *error margins* as a robust range of performance scores. This suggests that either the L_1, L_2+Mahalanobis, or angle+Mahalanobis distance should be used.

6.5 Conclusions

The main goal of our experiment was to point out the critical design decisions for PCA-based face recognition system. We introduced a generic modular PCA-based face recognition systems and systematically varied the components to measure the impact of these variations. From the results throughout the series of experiments, we present two models for a PCA-based face recognition system. In the proposed models, our design decision includes processing steps with better performance in each module.

The choice of steps used in the proposed I system include (1) illumination normalization ($\mu = 0.0$ and $\sigma = 1.0$), (2) low-pass filtering (LPF), (3) removal of the first low order eigenvector, and (4) using the angle+Mahalanobis distance. The choice of steps used in the proposed II system includes (1) illumination normalization ($\mu = 0.0$ and $\sigma = 1.0$), (2) wavelet compression [0.5 bpp], (3) removal of the first low order eigenvector, and (4) using the L_1+Mahalanobis distance. The proposed I system addresses the effects of LPF with angle+Mahalanobis distance, while the proposed II system represents wavelet compression with L_1+Mahalanobis distance.

In Table 6.9, the identification performance score for the duplicate I probe is increased from 0.35 to 0.49 for proposed I method, and the duplicate II probe from 0.13 to 0.26 for both proposed I and II method (top rank score). The identification performance score for **FB** probe is slightly increased from 0.77 to 0.78 for both proposed I and II methods, and **fc** probe from 0.26 to 0.33 for proposed II method (top rank score). In Table 6.10, the verification performance for the

Table 6.9: Comparison of identification performance scores for baseline, proposed I ($\mu = 0.0$ and $\sigma = 1.0$, LPF, first low order eigenvector removed, angle+Mahalanobis distance), and proposed II ($\mu = 0.0$ and $\sigma = 1.0$, wavelet compression [0.5bpp], first low order eigenvector removed, $L1$+Mahalanobis distance) algorithm. Performance scores are the top rank match.

	Probe category			
Algorithm	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline	0.35	0.13	0.77	0.26
Proposed I	0.49	0.26	0.78	0.26
Proposed II	0.40	0.26	0.78	0.33

Table 6.10: Comparison of verification performance scores for baseline, proposed I ($\mu = 0.0$ and $\sigma = 1.0$, LPF, first low order eigenvector removed, angle+Mahalanobis distance), and proposed II ($\mu = 0.0$ and $\sigma = 1.0$, wavelet compression [0.5bpp], first low order eigenvector removed, $L1$+Mahalanobis distance) algorithm. Performance scores are equal error rate (EER).

	Probe category			
Algorithm	Duplicate I	Duplicate II	**FB** probe	**fc** probe
Baseline	0.24	0.30	0.07	0.13
Proposed I	0.11	0.21	0.07	0.15
Proposed II	0.20	0.22	0.07	0.10

duplicate I probe is improved from 0.24 to 0.11 for proposed I method, and the duplicate II probe improved from 0.30 to 0.21 for proposed I method (equal error rate). The verification performance score for the **FB** probe shows the same results for all three methods, and the **fc** probe improved from 0.13 to 0.10 for proposed II method (equal error rate).

Based on these results, the proposed algorithms show reasonably better performance for the duplicate I, duplicate II (for proposed I method) and **fc** probes (for proposed II method) than the baseline algorithm in both identification and verification scenarios. For the **FB** probes, both identification and verification results show almost identical performance scores for each method used. The results of identification are reported for four different categories of probes in Figures 6.14 and 6.15. Also, the verification performances are reported for four different categories of probes in Figures 6.16 and 6.17.

In our evaluation process, we introduced a modular design for PCA-based face recognition systems. This allowed us to systematically vary the steps and measure the impact of these variations on performance. Like PCA, the majority of the face recognition algorithms in the literature are view-based [67, 81, 120] and have the same basic architecture as our PCA-based system. By following the evaluation procedure presented in this chapter, algorithm designers can determine the optimal configuration of their face recognition system. We have come to four major conclusions from the series of experiments with PCA-based face recognition system.

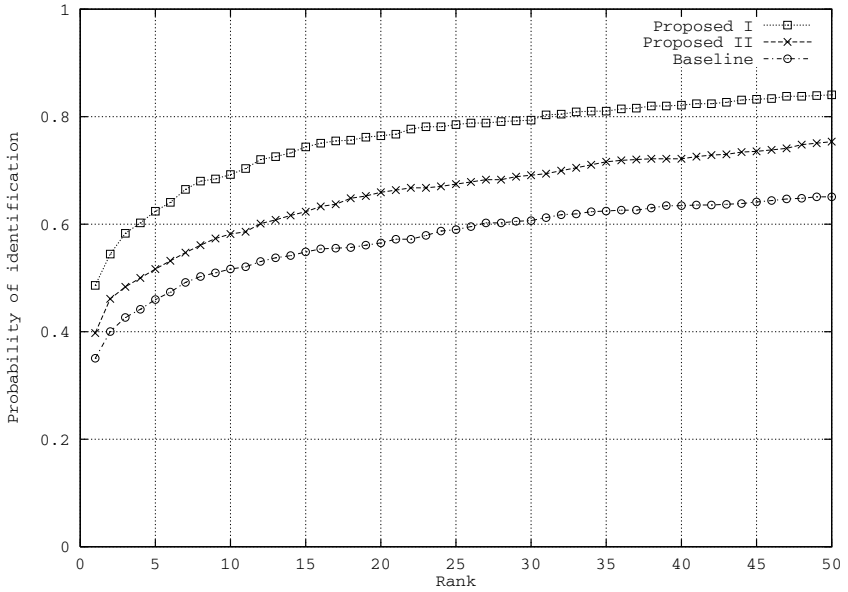
First, JPEG and wavelet compression algorithms do not degrade performance. This is important because it indicates that compressing images to save transmission time and storage costs will not reduce algorithm performance.

Second, selection of the nearest-neighbor classifier is the critical design decision in designing a PCA-based algorithm. The proper selection of a nearest-neighbor classifier is essential to improve performance scores. Furthermore, our experiments shows that similarity measures that achieve the best performance are not generally considered in the literature.

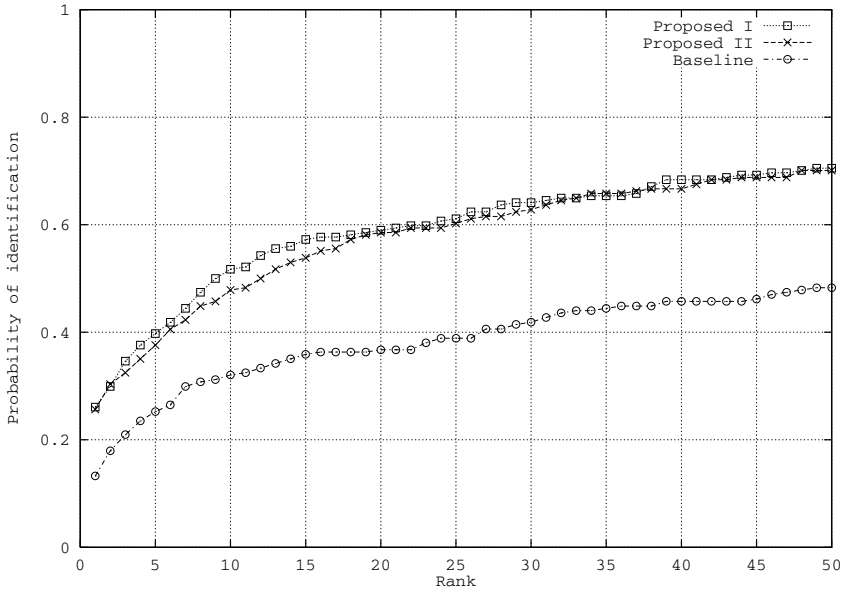
Third, the performance scores vary among the probe categories. This shows when one designs an algorithm, one needs to consider the type of images that the algorithm will process. We found that the **FB** and duplicate I probes are

least sensitive to system design decisions, while **fc** and duplicate II probes are the most sensitive.

Fourth, the performance within a category of probes can vary greatly. This leads to the recommendation that when comparing algorithms, the performance scores from a set of galleries and probe sets need to be examined. We generated 100 galleries and calculated performance against **fb** and duplicate probes. Then, we examined the range of scores and the overlap in scores among different implementations.

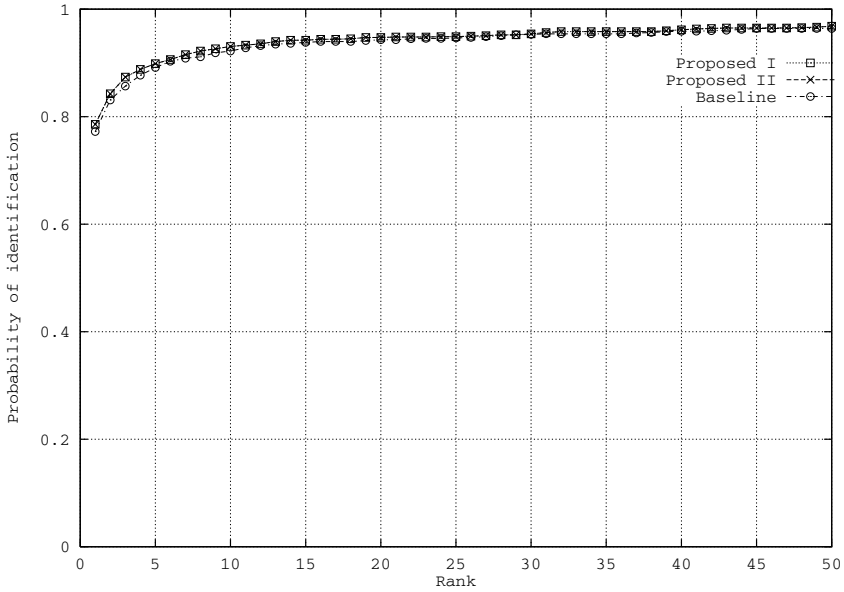


(a)

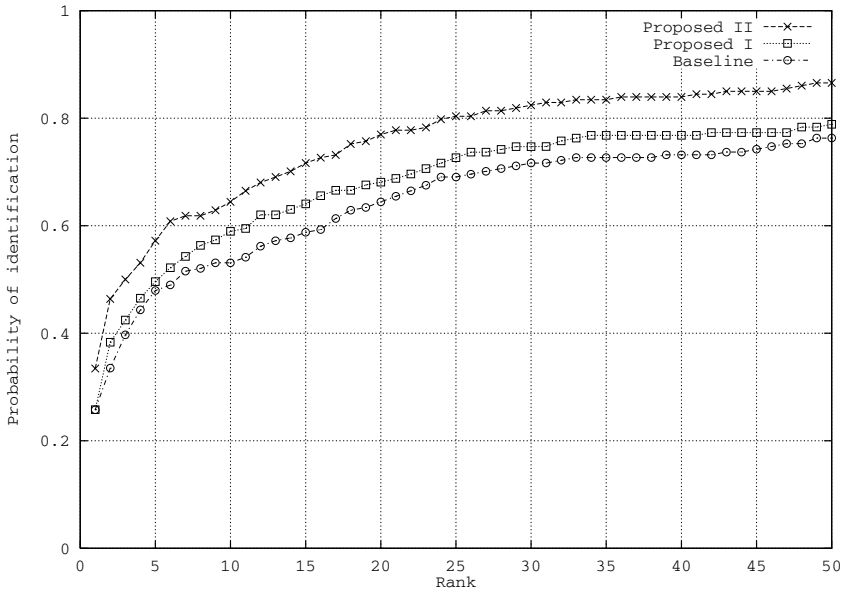


(b)

Figure 6.14: Identification performance comparison of baseline and proposed I, and proposed II algorithms. (a) duplicate I and (b) duplicate II probes.

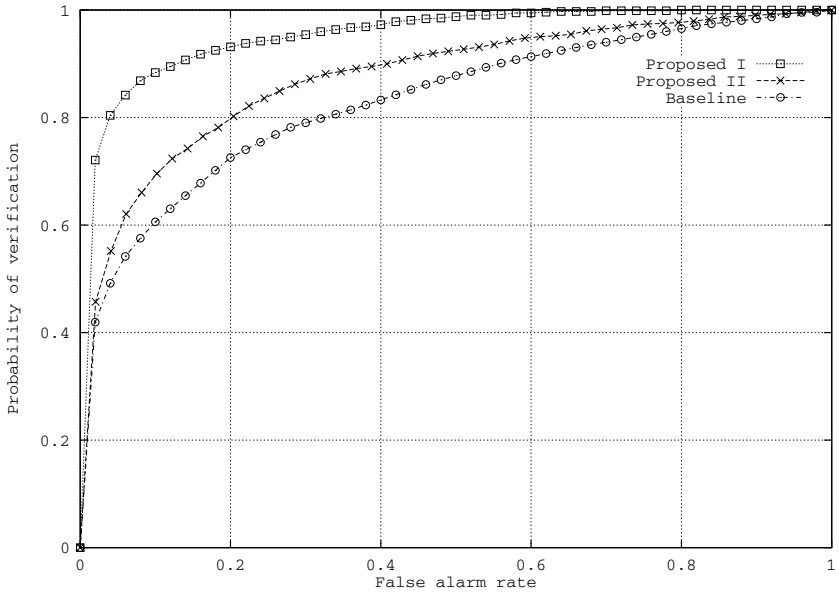


(a)

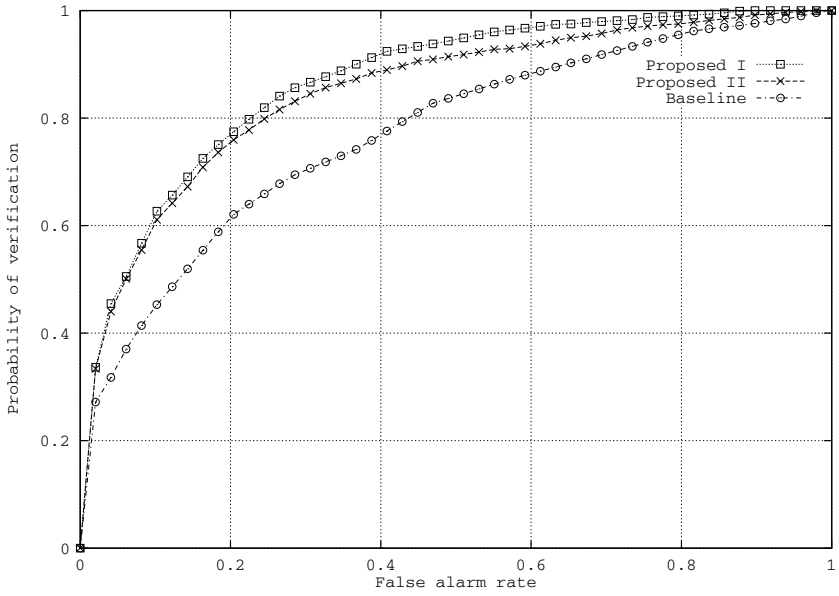


(b)

Figure 6.15: Identification performance comparison of baseline and proposed I, and proposed II algorithms. (a) **FB** and (b) **fc** probes.

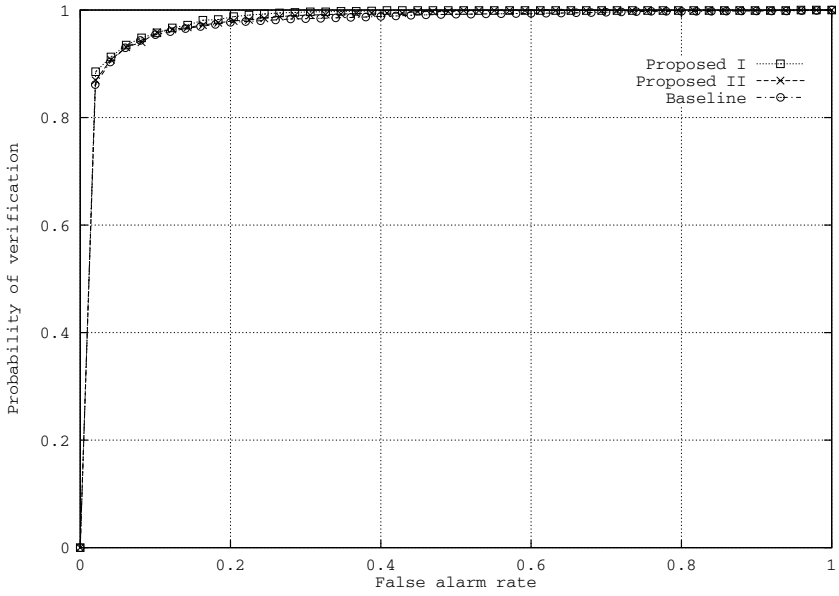


(a)

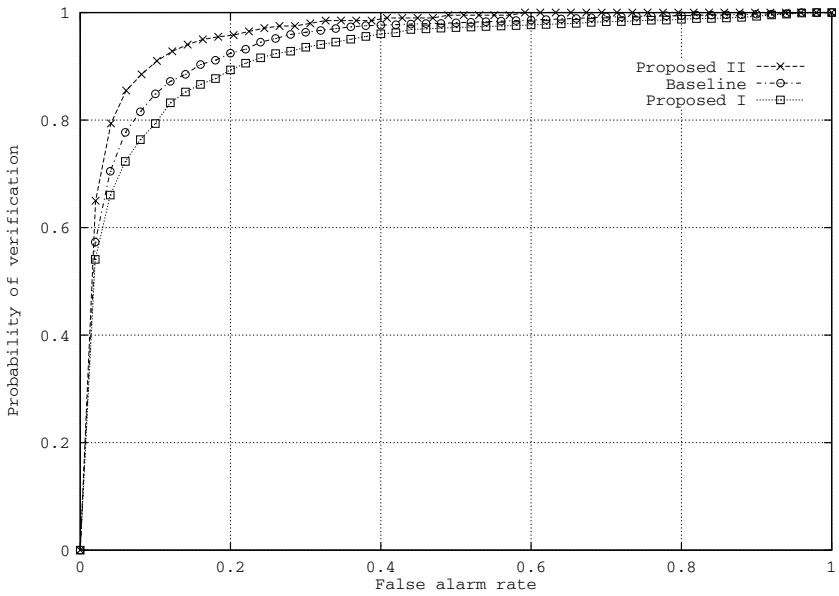


(b)

Figure 6.16: Verification performance comparison of baseline and proposed I, and proposed II algorithms. (a) duplicate I and (b) duplicate II probes.



(a)



(b)

Figure 6.17: Verification performance comparison of baseline and proposed I, and proposed II algorithms. (a) **FB** and (b) **fc** probes.

Chapter 7

Conclusions

In this dissertation, we have presented a performance evaluation methodology that is particularly suitable for the recognition of facial images. The automated recognition of human faces presents a significant challenge to the research community. Typically, most faces have a similar structure, since they only differ in minor details, and the appearance of a face can vary depending on the circumstances. The problem is further complicated by the often uncontrolled environment in which the facial image is acquired. In addition, each particular application that uses face recognition has limitations such as recognition scenarios, storage requirements, and transmission time. These limitations make face recognition one of the most difficult problems in computer vision and pattern recognition.

We developed a performance evaluation for face recognition algorithms based on the identification and verification model and presented a design methodology based on a generic modular PCA-based face recognition system. The main contributions of our research to the face recognition community are (1) the establishment of a standard methodology for evaluating face recognition algorithms, (2) an assessment of the state of the art in face recognition, and (3) the presentation of a design methodology for individual face recognition algorithms. The

design of our performance evaluation methodology using the identification and verification model and our generic modular PCA-based face recognition system have achieved the research objectives as described below.

First, the improvement in identification and verification performance shows directly that our performance evaluation methodology has made a significant contribution to face recognition technology. The results show that factors that affect the identification and verification performance include test scenario, date tested, and probe category. These evaluation efforts let researchers know the strengths and weaknesses of their algorithms and where improvements could be made. This is directly supported by the new FERET test, which presents improvements in identification and verification performance between the algorithms tested, the number of papers that use the experimental results based on our performance evaluation method, and the number of groups that participated.

Second, we have assessed the state of the art in face recognition by presenting identification and verification performance based on our new evaluation protocol. The new FERET test shows that definite progress is being made in face recognition and that the upper bound in performance has not been reached. We addressed various evaluation issues by computing algorithm performance for different probe categories and multiple galleries and probe sets. We observed variation in performance due to changing the gallery and probe set within a probe category and by changing probe categories. To estimate the degree of difficulty for each category, we compared the average and current upper bounds of performance for each category. For average performance, our results rank **FB** probes as easiest, duplicate II probes as most difficult, and **fc** and duplicate I probes as tied in the middle. For current upper bounds, duplicate I probes are more difficult than **fc** probes. Our results show that we can expect that the best performance will be significantly better than the average performance. Upper bound performance for all probe categories is superior to all average per-

formance categories except for **FB** probes.

Third, we have presented a design methodology of a generic modular PCA-based face recognition system and identified directions for future face recognition research. We have systematically varied the steps in our system and measured the impact of these variations on performance. By following our evaluation procedure, algorithm designers can determine the optimal configuration of their face recognition system. We have established a baseline algorithm and presented performance results by varying the implementation among the selected steps one at a time. Also, we have examined the effects of changing galleries to point out the variations in performance results.

Future research on our performance evaluation methodology includes (1) the establishment of baselines for human computer interaction; (2) the testing and evaluation of face recognition problems for aging, gender, and race; (3) the development of evaluation methods for real-time multimedia applications; and (4) the creation of generalized evaluation methods applicable for common computer vision and pattern recognition problems. Future research will require test designs and experiments that are more robust in design and content and that have databases with more variations. Examples of these variations are (1) images of individuals taken over an extended period of time; (2) images with a variety of features (e.g., glasses, facial hair, occlusions, rotations, and changes in illumination, etc.); (3) the effects of changing galleries and probe sets; and (4) multimedia databases, including audiovisual information within real-world settings.

Appendix A

Appendix

A.1 Definition of Terms

- **Duplicate**: an image of a person whose corresponding gallery image was taken on a different date.
- **Duplicate I**: all duplicate images for the gallery images.
- **Duplicate II**: duplicate images where there was at least one year between the acquisition of the probe image and corresponding gallery image.
- **FERET**: The Face Recognition Technology Program.
- **Gallery**: the collection of images of known individuals.
- **Probe**: an image of an unknown individual.
- **Probe set**: the collection of probes.
- **Query set**: the unknown facial images to be identified by the algorithm.
- **Target set**: the set of known facial images given to the algorithm.

A.2 Histogram Equalization

In histogram equalization, the goal is to obtain a uniform histogram for the output image [41, 89]. If the original image has pixel values I_{min} and I_{max}, while the available range is between 0 and $2^N - 1$, we transform $I(x, y)$ by the function

$$I(x, y) = (2^N - 1)(I(x, y) - I_{min})/(I_{max} - I_{min}),$$

where N is the number of bits with which the image has been digitized and $I(x, y)$ is the digitized gray level of the pixel with coordinates (x, y).

A.3 Generation of Eigenface

Let us consider a set of facial images for training $\{X_1, X_2, ..., X_M\}$. The mean face is defined by $\Psi = \frac{1}{M}\sum_{n=1}^{M} X_n$. Each face differs from the mean face by $\Phi_i = X_i - \Psi$. This set of vectors is used for the calculation of eigenfaces that finds a set of M orthonormal vectors, $\mathbf{u_n}$, that best describes the distribution of the data. The kth vector, $\mathbf{u_k}$, is chosen such that

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(\mathbf{u}_k^T \Phi_n)^2$$

is maximum, subject to

$$\mathbf{u}_l^T \mathbf{u}_k = \delta_{lk} = \begin{cases} 1, & \text{if } l = k \\ 0, & \text{otherwise} \end{cases}$$

The vectors $\mathbf{u}_k$ and scalars λ_k are the eigenvectors and eigenvalues, respectively, of the covariance matrix

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_n\Phi_n^T = AA^T,$$

where the matrix $A = [\Phi_1, \Phi_2, ..., \Phi_M]$.

Let us consider a set of N facial images $\{x_1, x_2, ..., x_N\}$ taking values in an n-dimensional image space, and assume that each image belongs to one of c classes. Let us also consider a linear transformation mapping the original n-dimensional image space into an m-dimensional feature space, where $m < n$. The new feature vectors $y_k \in \mathcal{R}^m$ are defined by the following linear transformation:

$$y_k = W^t x_k, k = 1, 2, ..., N,$$

where $W \in \mathcal{R}^{n\times m}$ is a matrix with orthonormal columns. If the total scatter matrix S_T is defined as

$$S_T = \sum_{k=1}^{N}(x_k - \mu)(x_k - \mu)^T,$$

where n is the number of sample images, and $\mu \in \mathcal{R}^n$ is the mean image of all samples, then after applying the linear transformation W^T, the scatter of the transformed feature vectors $\{y_1, y_2, ..., y_N\}$ is $W^T S_T W$. In PCA, the projection W_{opt} is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e.,

$$W_{opt} = \arg\max_W |W^T S_T W| = [w_1 w_2 ... w_m],$$

where $\{w_i | i = 1, 2, ..., m\}$ is the set of n-dimensional eigenvectors of S_T corresponding to the m largest eigenvalues. Since these eigenvectors have the same

dimension as the original images, they are referred to as eigenfaces [57]. If classification is performed using a nearest-neighbor classifier in the reduced feature space and m is chosen to be the number of images N in the training set, then the eigenface method is equivalent to the correlation method.

A.4 Nearest-Neighbor Classifier

We mathematically describe the similarity measure used in the nearest-neighbor classifiers. The variables $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ are k-dimensional vectors and $\mathbf{x}_i, \mathbf{y}_i$, and $\mathbf{z}_i$ are the ith components of the vectors.

A.4.1 L_1 Distance.

$$d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sum_{i=1}^{k} |x_i - y_i|$$

A.4.2 L_2 Distance.

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2 = \sum_{i=1}^{k} (x_i - y_i)^2$$

A.4.3 Angle Between Feature Vectors.

$$d(\mathbf{x}, \mathbf{y}) = -\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = -\frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_{i=1}^{k} (x_i)^2 \sum_{i=1}^{k} (y_i)^2}}$$

A.4.4 Mahalanobis Distance.

$$d(\mathbf{x}, \mathbf{y}, \mathbf{z}) = -\sum_{i=1}^{k} x_i y_i z_i,$$

and

$$_i = \sqrt{\frac{\lambda_i}{\lambda_i + \alpha^2}} \simeq \frac{1}{\sqrt{\lambda_i}}, \alpha = 0.25,$$

where λ_i = the eigenvalue of the ith eigenvector.

A.4.5 L_1 + Mahalanobis Distance.

$$d(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{k} |x_i - y_i| z_i$$

A.4.6 L_2 + Mahalanobis Distance.

$$d(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{k} (x_i - y_i)^2 z_i$$

A.4.7 Angle + Mahalanobis Distance.

$$d(\mathbf{x}, \mathbf{y}, \mathbf{z}) = -\frac{\sum_{i=1}^{k} x_i y_i z_i}{\sqrt{\sum_{i=1}^{k} (x_i)^2 \sum_{i=1}^{k} (y_i)^2}}$$

Bibliography

[1] H. Abdi, D. Valentin, B. Edelman, and A. J. O'Toole. More about the difference between men and women. *Perception*, 24:539–562, 1995.

[2] J. Atick, P. Griffin, and A. N. Norman. Face recognition from live video for real-world applications—now. *Advanced Imaging*, 10(5):58–62, May 1995.

[3] J. Atick, P. Griffin, and A. N. Redlich. Statistical approach to shape from shading: Reconstruction of three-dimensional surfaces from single two-dimensional images. *Neural Computation*, 8:1321–1340, 1996.

[4] J. Atick and A. N. Redlich. Convergent algorithm for sensory receptive field development. *Neural Computation*, 5:45–60, 1993.

[5] M. S. Barlett, H. M. Lades, and T. J. Sejnowski. Independent component representations for face recognition. In *Proceedings of the SPIE Symposium on Electronic Imaging; Science and Technology; Conference on Human Vision and Electronic Imaging III*, in press 1998.

[6] H. Barlow and W. Levick. Changes in the maintained discharge with adaptation level in the cat retina. *Journal of Physiology (London)*, 202:699–718, 1969.

[7] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs fisherfaces: Recognition using class specific linear projection. In *Proceedings of the 4th European conference on computer vision*, pages 45–58, 1996.

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

[9] D. J. Beymer. Face recognition under varying pose. Technical Report A.I. Memo No. 1461, MIT, December 1993.

[10] M. Bichsel. *Strategies of robust object recognition for the automatic identification of human faces*. PhD thesis, Univ. of Zurich, 1991.

[11] J. L. Blue, G. T. Candela, P. J. Grother, R. Chellappa, and C. L. Wilson. Evaluation of pattern classifiers for fingerprint and OCR applications. *Pattern Recognition*, 27(4):485–501, 1994.

[12] V. Bruce. Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, 73:105–117, 1982.

[13] V. Bruce. Stability from variation: The case of face recognition. The M. D. Vernon memorial lecture. *Quart. Journal of Exper. Psychology*, 47A(1):5–28, 1994.

[14] V. Bruce, A. M. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney. Sex discrimination: How do we tell the difference between male and female faces? *Perception*, 22:131–152, 1993.

[15] V. Bruce, A. Coombes, and R. Richards. Describing the shapes of faces using surface primitives. *Image and Vision Computing*, 11:353–363, 1993.

[16] V. Bruce, E. Hanna, N. Dench, P. Healey, and A. M. Burton. The importance of 'mass' in line drawings of faces. *Applied Cognitive Psychology*, 6:619–628, 1992.

[17] V. Bruce and S. Langton. The use of pigmentation and shading information in recognising the sex and identies of faces. *Perception*, 23, 1994.

[18] V. Bruce, T. Valentine, and A. D. Badderly. The basis of the 3/4 view effect in face recognition. *Applied Cognitive Psychology*, 1:109–120, 1987.

[19] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Trans. PAMI*, 15(10), 1993.

[20] C.J.C. Burges. A tuturial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, (submitted), 1998.

[21] A. M. Burton, V. Bruce, and N. Dench. What's the difference between men and women? Evidence from facial measurement. *Perception*, 22:153–176, 1993.

[22] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of face: A survey. *Proceedings of the IEEE*, 83:704–740, 1995.

[23] K. Cho, P. Meer, and J. Cabrera. Performance assessment through bootstrap. *IEEE Trans. PAMI*, 19(11):1185–1198, 1997.

[24] G. W. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. In *Proc. Int. Conf. Neural Network*, pages 322–325. Kluwer, Dordrecht, 1990.

[25] G. W. Cottrell and J. Metcalfe. Empath: Face, emotion, and gender recognition using holons. In D. Touretzky, editor, *Advances in Neural Information Processing*, volume 3, pages 564–571, San Mateo, CA, 1991. Morgan Kaufman.

[26] I. Cox, J. Ghosen, and P. Yianilos. Feature-based face recognition using mixture-distance. In *Proceedings Computer Vision and Pattern Recognition 1996*, pages 209–216, 1996.

[27] I. Craw, N. Costen, and T. Kato. How should we represent faces for automatic recognition. *IEEE Trans. PAMI*, (in press) 1998.

[28] Y. Dai and Y. Nakano. Extraction of facial images from a complex background using color information and SGLD matrices. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 238–242, 1995.

[29] Y. Dai, Y. Nakano, and H. Miyao. Extraction of facial images from a complex background using SGLD matrices. In *12th International Conference on Pattern Recognition*, pages 137–141, volume I, october 1994.

[30] T. Darrell, G. Gordon, J. Woodfill, and M. Harville. A virtual mirror using real-time robust face tracking. In *3rd International Conference on Automatic Face and Gesture Recognition*, pages 616–621, 1998.

[31] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, July 1985.

[32] J. G. Daugman. Complete discrete 2-D Gabor transform by neural networks for image analysis and compression. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, July 1988.

[33] D. Demigny and T. Kamle. A discrete expression of canny's criteria for step edge detector performance evaluation. *IEEE Trans. PAMI*, 19(11):1199–1211, 1997.

[34] J. P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

[35] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *J. Opt. Soc. Am. A*, 14:1724–1733, August 1997.

[36] T. E. Flick, L. K. Jones, R. G. Priest, and C. Herman. Pattern recognition using projection pursuit. *Pattern Recognition*, 23(12):1367–1376, 1990.

[37] K. Fukunaga. *Introduction to statistical pattern recongition*. Academic Press, Inc., San Diego, CA, 1990.

[38] A. Georghiades, D. Kriegman, and P. Belhumeur. Illumination cones for recognition under variable lighting: Faces. In *Computer Vision and Pattern Recognition 98*, 1998.

[39] A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.

[40] R. Gonzalez and P. Wintz. *Digital Image Processing*. Addison–Wesley Publishing Co., 87.

[41] R. Gonzalez and R. Woods. *Digital Image Processing*. Addison–Wesley Publishing Co., 1993.

[42] G. G. Gordon. Face recognition from frontal and profile views. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 47–52, 1995.

[43] V. Govindaraju, D. Sher, R. Srihari, and S. Srihari. Locating human faces in newspaper photographs. In *Proceedings of Computer Vision and Pattern Recognition*, pages 549–554, 1989.

[44] D. Green and J. Swets. *Signal Detection Theory and Psychophysics*. John Wiley & Sons Ltd., 1966.

[45] S. Gutta, J. Huang, D. Singh, I. Shah, B. Takacs, and H. Wechsler. Benchmark studies on face recognition. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 227–231, 1995.

[46] P.J.B. Hancock, V. Bruce, and A. M. Burton. A comparison of two computer-based face identification system with human perceptions of faces. *Vision Research*, 1997.

[47] P.J.B. Hancock, A. M. Burton, and V. Bruce. Face processing: Human perception and principal component analysis. *Memory & Cognition*, 24(1):26–40, 1996.

[48] S. Haykin. *Neural Networks: A Comprehensive Introduction*. Macmillan College Publishing Co., 1994.

[49] M. D. Heath, S. Sarkar, T. Sanocki, and K. W. Bowyer. A robust visual method for assessing the relative performance of edge-detection algorithms. *IEEE Trans. PAMI*, 19(12):1338–1359, 1997.

[50] C. W. Helstrom. *Elements of Signal Detection & Estimation*. PTR Prentice Hall, 1995.

[51] H. Hill, P. G. Schyns, and S. Akamatsu. Information and viewpoint dependence in face recognition. *Cognition*, 63(2):201–222, February 1997.

[52] L. Hong and A. Jain. Integrating faces and fingerprints for personal identification. *IEEE Trans. PAMI*, 20(12):1295–1307, 1998.

[53] N. Intrator. Combining exploratory projection pursuit and projection pursuit regression with application to neural networks. *Neural Computation*, 5:443–455, 1993.

[54] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. PAMI*, 19(2):153–158, 1997.

[55] A. Johnston, H. Hill, and N. Carmen. Recognising faces: Effects of lighting direction, inversion and brightness. *Perception*, 21:365–375, 1992.

[56] A. Johnston and P. J. Passmore. Shape from shading 1: Surface curvature and orientation. *Perception*, 23:169–189, 1994.

[57] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[58] J. Jones and L. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Cognitive Neurophys.*, 58:1233–1258, 1987.

[59] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. PAMI*, 12(1), 1990.

[60] W. Konen and E. Schulze-Kruger. ZN-Face: A system for access control using automated face recognition. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 18–23, 1995.

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

[62] M. Lindenbaum. An integrated model for evaluating the amount of data required for reliable recognition. *IEEE Trans. PAMI*, 19(11):1251–1264, 1997.

[63] C. Liu and H. Wechsler. Probabilistic reasoning models for face recognition. In *Computer Vision and Pattern Recognition 98*, 1998.

[64] A. M. Lopez, F. Lumbreras, J. Serrat, and J. J. Villanueva. Evaluation of methods for ridge and valley detection. *IEEE Trans. PAMI*, 21(4):336–347, 1999.

[65] T. Maurer and C. von der Malsburg. Single-view based recognition of faces rotated in depth. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 248–253, 1995.

[66] B. Moghaddam, C. Nastar, and A. Pentland. Bayesian face recognition using deformable intensity surfaces. In *Proceedings Computer Vision and Pattern Recognition 96*, pages 638–645, 1996.

[67] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *Proc. SPIE Conference on Automatic Systems for the Identification and Inspection of Humans*, volume 2277 (SPIE), pages 12–21, 1994.

[68] B. Moghaddam and A. Pentland. Maximum likelihood detection of faces and hands. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 122–128, 1995.

[69] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. PAMI*, 19(7):696–710, 1997.

[70] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: Probablistic matching for face recognition. In *3rd International Conference on Automatic Face and Gesture Recognition*, pages 30–35, 1998.

[71] H. Moon and P. J. Phillips. Analysis of PCA-based face recognition algorithms. In K. W. Bowyer and P. J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision*. IEEE Computer Society Press, Los Alamitos, CA, 1998.

[72] H. Moon and P. J. Phillips. Analysis of PCA-based face recognition algorithms. In *2nd Inter. Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 205–210, 1999.

[73] H. Murase and S. K. Nayar. Image spotting of 3D objects using the parametric eigenspace representation. In *Proceedings of the 9th Scandinavian conference on image analysis*, pages 325–332, 1995.

[74] H. Murase and S. K. Nayar. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[75] A. J. O'Toole, K. A. Deffenbacher, D. Valentin, and H. Abdi. Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, 22(2):208–224, 1994.

[76] A. J. O'Toole, K. A. Deffenbacher, D. Valentin, K. McKee, D. Huff, and H. Abdi. The perception of face gender: the role of stimulus structure in recognition and classification. *Memory & Cognition*, in press, 1996.

[77] A. J. O'Toole, T. Vetter, N. Troje, and H. Bultoff. Sex classification is better with three-dimensional head structure than with intensity information. *Perception*, 26:75–84, 1997.

[78] A. J. O'Toole, T. Vetter, H. Volz, and E. M. Salter. Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age. *Perception*, 26(6):719–732, 1997.

[79] P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[80] W. B. Pennebaker and J. L. Mitchell. *JPEG still image compression standard*. Van Nostrand Reinhold, 1993.

[81] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eignspaces for face recognition. In *Proceedings Computer Vision and Pattern Recognition 94*, pages 84–91, 1994.

[82] P. J. Phillips. Matching pursuit filters applied to face identification. *IEEE Trans. on Image Processing*, 7(8):1150–1164, 1998.

[83] P. J. Phillips, R. M. McCabe, and R. Chellappa. Biometric image processing and recognition. In *Proceedings of the IX European Signal Processing Conference (EUSIPCO-98)*, 1998.

[84] P. J. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings Computer Vision and Pattern Recognition 97*, pages 137–143, 1997.

[85] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation. In P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.

[86] P. J. Phillips and P. Rauss. The face recognition technology (FERET) program. In *Proceedings of Office of National Drug Control Policy, CTAC International Technology Symposium*, pages 8–11 — 8–20, August 1997.

[87] P. J. Phillips, P. Rauss, and S. Der. FERET (face recognition technology) recognition algorithm development and test report. Technical Report ARL-TR-995, U.S. Army Research Laboratory, 1996.

[88] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, 16(5):295–306, 1998.

[89] W. K. Pratt. *Digital Image Processing*. John Wiley & Sons, 1978.

[90] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Receipes in C, The Art of Scientific Computing*. Cambridge University Press, 1992.

[91] L. R. Rabiner and B. Gold. *Theory and application of digital signal processing*. Prentice-Hall, 1975.

[92] T. Randen and J.H. Husoy. Filtering for texture classification: A comparative study. *IEEE Trans. PAMI*, 21(4):291–310, 1999.

[93] P. Rauss, P. J. Phillips, A. T. DePersia, and M. Hamilton. The FERET (Face Recognition Technology) program. In *Surveillance and Assessment Technology for Law Enforcement*, volume 2935 (SPIE), pages 2–11, 1996.

[94] A. N. Redlich. Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, 5:289–304, 1993.

[95] S. Rizvi, P. J. Phillips, and H. Moon. The FERET verification testing protocol for face recognition algorithms. In *Automatic Face and Gesture Recognition*, pages 48–53, 1998.

[96] S. Rizvi, P. J. Phillips, and H. Moon. A verification protocol and statistical performance analyis for face recognition algorithms. In *Computer Vision and Pattern Recognition 98*, 1998.

[97] S. Rizvi, P. J. Phillips, and H. Moon. The feret verification testing protocol for face recognition algorithms. *Image and Vision Computing Journal*, (to appear) 1999.

[98] H. A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical Report CMU-CS-95-158, School of Computer Science, Carnegie Mellon University, July 1995.

[99] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. PAMI*, 20(11):23–28, 1998.

[100] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25(1):65–77, 1992.

[101] T. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 5:459–473, 1989.

[102] J.A. Shufelt. Performance evaluation and analysis of monocular building extraction from aerial imagery. *IEEE Trans. PAMI*, 21(4):311–326, 1999.

[103] J.A. Shufelt. Performance evaluation and analysis of vanishing point detection techniques. *IEEE Trans. PAMI*, 21(3):282–288, 1999.

[104] B. W. Silverman. *Density Estimation for Statistics and data analysis*. Chapman and Hall, 1986.

[105] B. W. Silverman and A. Young. *In the Eyes of the Beholder*. Oxford University Press, 1998.

[106] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A*, 4:519–524, 1987.

[107] K-K Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. PAMI*, 20(1):39–51, 1998.

[108] D. Swets and J. Weng. discriminant analysis and eigenspace partition tree for face and object recognition from views. In *2nd International Conference on Automatic Face and Gesture Recognition*, pages 192–197, 1996.

[109] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. PAMI*, 18(8):831–836, 1996.

[110] M. Tiech and P. Diament. Relative refractoriness in visual information processing. *Biol. Cybern.*, 38:187–191, 1980.

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

[112] D. Valentin, H. Abdi, and B. Edelman. What represents a face: A computational approach for the integration of physiological and psychological data. *Perception*, 26:1271–1288, 1997.

[113] D. Valentin, H. Abdi, B. Edelman, and A. J. O'Toole. Principal component and neural network analyses of face images: What can be generalized in gender classification? *Journal of Mathematical Psychology*, 42:175–195, 1997.

[114] D. Valentin, H. Abdi, and A. J. O'Toole. Categorization and identification of human face images by neural networks. *Journal of Biological Systems*, 2:413–430, 1994.

[115] T. Valentine, editor. *Cognitive and Computational Aspects of Face Recognition*. Routledge, 1995.

[116] M. Vetterli and J. Kovacevic. *Wavelets and subband coding*. Prentice Hall, 1995.

[117] H. Wechsler. *Computational Vision*. Academic Press, 1990.

[118] H. Wechsler, P. J. Phillips, V. Bruce, S. F. Soulie, and T. Huang. *Face Recognition: From Thoery to Application*. Springer-Verlag, 1998.

[119] J. Wilder. Face recognition using transform coding of gray scale projection projections and the neural tree network. In R. J. Mammone, editor, *Artifical Neural Networks with Applications in Speech and Vision*, pages 520–536. Chapman Hall, 1994.

[120] J. Wilder, P. J. Phillips, C. Jiang, and S. Wiener. Comparison of view-based face recognition algorithms on visible and infrared imagery. In *Surveillance and Assessment Technologies for Law Enforcement*, volume 2935 (SPIE), pages 36–44, 1996.

[121] J. Wilder, P. J. Phillips, C. Jiang, and S. Wiener. Comparison of visible and infrared imagery for face recognition. In *2nd International Conference on Automatic Face and Gesture Recognition*, pages 182–187, 1996.

[122] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition and gender determination. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 92–97, 1995.

[123] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. PAMI*, 17(7):775–779, 1997.

[124] C. R. Wren, A. Azarbayejani, T. Darrel, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. PAMI*, 19:780–785, 1997.

[125] G. Yang and T. S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[126] A. Young, D. Hay, K. McWeeny, B. Flude, and A. Ellis. Matching familiar and unfamiliar faces on internal and external features. *Perception*, 14:737–746, 1985.

[127] A. L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.

[128] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *3rd International Conference on Automatic Face and Gesture Recognition*, pages 336–341, 1998.

[129] W. Zhao, R. Chellappa, and N. Nandhakumar. Empirical performance analysis of linear discriminant classifiers. In *Computer Vision and Pattern Recognition 98*, 1998.

[130] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.