Perception, 2001, volume 30, pages 303–321

DOI:10.1068/p2896

Computational and performance aspects of PCA-based face-recognition algorithms

Hyeonjoon Moon
Department of Electrical and Computer Engineering, State University of New York at Buffalo, Amherst, NY 14260, USA; e-mail: moon@acsu.buffalo.edu

P Jonathon Phillips
National Institute of Standards and Technology, Gaithersburg, MD 20899, USA; e-mail: jonathon@nist.gov
Received 16 March 2000

Abstract. Algorithms based on principal component analysis (PCA) form the basis of numerous studies in the psychological and algorithmic face-recognition literature. PCA is a statistical technique and its incorporation into a face-recognition algorithm requires numerous design decisions. We explicitly state the design decisions by introducing a generic modular PCA-algorithm. This allows us to investigate these decisions, including those not documented in the literature. We experimented with different implementations of each module, and evaluated the different implementations using the September 1996 FERET evaluation protocol (the de facto standard for evaluating face-recognition algorithms). We experimented with (i) changing the illumination normalization procedure; (ii) studying effects on algorithm performance of compressing images with JPEG and wavelet compression algorithms; (iii) varying the number of eigenvectors in the representation; and (iv) changing the similarity measure in the classification process. We performed two experiments. In the first experiment, we obtained performance results on the standard September 1996 FERET large-gallery image sets. In the second experiment, we examined the variability in algorithm performance on different sets of facial images. The study was performed on 100 randomly generated image sets (galleries) of the same size. Our two most significant results are (i) changing the similarity measure produced the greatest change in performance, and (ii) that difference in performance of $\pm 10\%$ is needed to distinguish between algorithms.

1 Introduction

Computer algorithms can serve as models for the human face-recognition function. Directly comparing these models (algorithms) with human performance allows the assessment of which models are biologically plausible. The closer the concordance between human and model performance, the greater the plausibility. The models need not be comprehensive, ie account for all aspects of face recognition. Rather, one can ascertain which properties of the human face-processing system are correctly modeled. For example, model A could correctly predict the effects of changes in illuminations, whereas model B could correctly account for changes in pose.

The starting point for numerous studies and lines of investigation were (and still are) algorithms based on principal component analysis (PCA) (also known in the literature as eigenfaces). PCA-based algorithms are popular because of the ease of implementing them and their reasonable performance levels (Phillips et al 1997, 2000; Rizvi et al 1998). Because of their popularity, PCA-based algorithms have become the de facto benchmark algorithm.

PCA-based algorithms have been the basis of numerous research projects in both psychophysics and computer vision. They have served as benchmarks for comparison with new algorithms (Belhumeur et al 1997; Phillips 1999b; Swets and Weng 1996; Wilder et al 1996; Zhao et al 1998), computational models in psychophysics (Hancock et al 1996; O'Toole et al 2000; Valentin et al 1997; Valentine 1995), and the basis for face-recognition algorithms (Barlett et al 1998; Etemad and Chellappa 1997; Kirby and Sirovich 1990; Liu and Wechsler 1999; Moghaddam and Pentland 1994, 1998; O'Toole et al 1993;

Penev and Atick 1996; Turk and Pentland 1991). PCA algorithms have been applied in a broad spectrum of studies including face detection (Moghaddam and Pentland 1995; Sung and Poggio 1998), face recognition (Brunelli and Poggio 1993; Fleming and Cottrell 1990; Hancock et al 1996; Turk and Pentland 1991), and sex classification (Abdi et al 1995; Cottrell and Metcalfe 1991; O'Toole et al 1997). Psychologists and neuroscientists had an active interest in PCA as a model for face processing prior to its adoption by the computer vision community for automatic face recognition (Cottrell and Metcalfe 1991; Fleming and Cottrell 1990; O'Toole et al 1988, 1991).

PCA is a statistical method for reducing the dimensionality of a data set while retaining the majority of the variation present in the data set (Jolliffe 1986). Because PCA is a statistical method for handling and analyzing data, a PCA-based face-recognition algorithm needs an algorithmic supporting structure. Constructing this supporting structure requires a number of critical design decisions. Each of these design decisions has the potential to affect the overall performance of the face-recognition algorithm. Some of these design decisions have been stated explicitly in the literature—for example, the similarity measure in the nearest-neighbor classifier. However, a large number of decisions are not mentioned and are passed from researcher to researcher by word of mouth. Two examples are the methods for normalizing illumination and the number of eigenvectors included in the representation. Because the design details are not stated explicitly, a reader cannot assess the merits of a particular implementation and the associated claims. This can unnecessarily cast a shadow on performance claims of studies where a PCA algorithm is used as a model or a benchmark. For example, does a PCA-based algorithm fail (or succeed) to explain observed data because of a faulty design decision? Or is the failure (or success) based on underlying properties of PCA? Knowledge of the basic strengths and weaknesses of different implementations can provide insight and guidance in designing studies or developing algorithms that build on PCA.

In this paper, 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.

Using the generic model for PCA-based algorithms, we evaluate different implementations. The generic model allows us to change the implementation in an orderly manner and to assess the impact on performance of each modification.

Algorithm evaluations are conducted by following an evaluation protocol. An evaluation protocol states how the test is conducted. This includes the quality of images and the number of images in the training and testing sets, how the algorithms are tested, how the results from algorithms are formatted, how the results are scored, and what scores are computed.

Some basic terms are introduced here to describe our evaluation protocol. The *gallery* is the set of known individuals. The images used to test the algorithms are called *probes*. The identity of the face in a probe is not known to the algorithm. A probe is either a new image of an individual in the gallery or an image of an individual not in the gallery. *Duplicates* are probes of individuals in the gallery that are taken on a different date, or under different conditions than the images stored in the gallery. To compute performance, one needs both a gallery and probe set. The probes are presented to an algorithm, and the algorithm returns the best match between each probe and the images in the gallery, or, more generally, ranks the gallery by similarity to each probe. Algorithm identification performance is reported on a cumulative match characteristic (CMC) (see section 3.2 for details). The estimated identity of a probe is the best match.

Computational algorithms must solve two problems that map easily onto the psychological tasks of recognition and identification. Recognition is the task of determining whether or not the face in a probe is of a person in the gallery. Identification is the task of determining which individual is the best match to the probe. Note that identification can be performed regardless of whether or not a face has been recognized. In the psychology literature, identification is referred to as a forced-choice experiment.

Finally, a very common task performed by computational algorithms, but less commonly performed by humans, is verification. Verification is a special case of recognition. In verification, an algorithm or person is presented with a probe and a claimed identity for the probe. The claim is either accepted or rejected, or, more generally, a confidence in the validity of the claim is generated. Verification results are reported on a receiver operator characteristic (ROC) (Macmillan and Creelman 1991).

The contents of the galleries and probe sets are described in the evaluation protocol. If the evaluation protocol is appropriately designed, performance scores can be calculated for multiple galleries and probe sets. We report results for the standard galleries and probe sets described in the September 1996 FERET evaluation protocol (Phillips et al 2000). The September 1996 FERET evaluation was the last of three FERET evaluations, which independently evaluated automatic face-recognition algorithms (Phillips et al 1997, 1998, 2000; Rizvi et al 1998). The FERET evaluation and its associated database have become the de facto standards in the automatic face-recognition community. 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. The FERET protocol allows one to measure identification and verification performance. In this paper we report identification results.

Computation of CMCs and ROCs requires a similarity measure between all probes and gallery images. From a complete set of the similarity measures, identification and verification performance can be computed. Knowing a rating of similarity between all probes and gallery images is the point that distinguishes most algorithm evaluations from psychological studies. Algorithms can easily compute a complete set of similarity ratings or measures, whereas most psychological studies do not explicitly measure these data. This results in psychological studies reporting performance for a single point on a ROC or the top-rank score for forced-choice experiments. The top-rank score is a single performance point on a CMC. It is not possible to draw a connection between a single point on a ROC and a single point on a CMC. Therefore, under these conditions, identification and verification are distinct problems.

For algorithms, identification and verification appear to be substantially different results. However, if one has a complete set of similarity measures, there is a direct connection between the two. Phillips (1999a) showed a duality between identification and verification, and, under the duality relationship, identification performance is an upper bound for verification performance. Or, more precisely, the cumulative match characteristic curve is an upper bound for the ROC. Thus, because we compute a complete set of similarity measures for each algorithm, we are generating information relevant to both forced-choice and verification style experiments.

By analyzing a complete set of similarity measures one can study the structure underlying facial processing. This was shown in O'Toole et al (2000), where algorithm and human performance were compared. The comparison was done at the level and similarity and typicality of individual faces, and showed a common bimodal structure for both humans and algorithms for the perception of faces. The study included performance data from seven of the algorithms reported in this paper (the classifier variations in section 4.3.3).

In this paper, we present the results of two experiments. In each experiment we investigated a different aspect of measuring algorithm performance. In experiment 1,

we systematically varied the components in our generic algorithm model. This allowed us to determine which design decisions have the greatest impact on performance. We varied the illumination normalization procedure, the number of eigenvectors in the representation, and the similarity measure; and we studied the effects of compressing facial images on algorithm performance. The effects of image compression on recognition are of interest in applications where image storage space or image transmission time are critical parameters.

One of the key parameters in algorithms is image quality. We characterize image match quality by the time between acquisition of the gallery and probe images of a person, and changes in illumination. These factors have a major impact on performance, and in numerous applications are major sources of variation among images.

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 2, we looked at this question by randomly generating 100 galleries of the same size. We then calculated performance on each of the galleries against two probe sets. The first set consisted of probes taken on the same day as the corresponding gallery image. This set represents algorithm performance under optimal conditions, and provides an upper bound for performance of the algorithms tested. The second set consisted of probes taken on different days than the corresponding gallery images. This examined performance under realistic conditions. 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.

2 PCA-based face-recognition system

In this section we discuss each of the components in our generic PCA-based algorithm.

2.1 *Principal component analysis (PCA)*

PCA is a statistical dimensionality-reduction method, which produces the optimal linear least-squares decomposition of a training set. Kirby and Sirovich (1990) applied PCA to representing faces and Turk and Pentland (1991) extended PCA to recognizing faces. [For further details on PCA, see Fukunaga (1972), Jolliffe (1986), or Valentin et al (1994).] In a PCA-based face-recognition algorithm, the input is a training set, $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N$ of N facial images such that the ensemble mean of the training set is zero $\left(\sum_i \boldsymbol{t}_i = 0\right)$. In computing the PCA representation, each image is interpreted as a point in $\mathbb{R}^{n \times m}$, where each image is n by m pixels. PCA finds the optimal linear least-squares representation in $(N-1)$-dimensional space, with the representation preserving variance.[1] The PCA representation is characterized by a set of $N-1$ eigenvectors $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{N-1})$ and eigenvalues $(\lambda_1, \ldots, \lambda_{N-1})$. 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_is are equal to the variance of the projection of the training set onto the ith eigenvector. Thus, the low-order eigenvectors encode the larger variations in the training set (low order refers to the index of the eigenvectors and eigenvalues). The higher-order eigenvectors encode smaller variations in the training set. Because these features encode smaller variations, it is commonly assumed that they represent noise in the training set. Because of this assumption and empirical results, higher-order eigenvectors are excluded from the representation. Faces are represented by their projection onto a subset of $M \leqslant N-1$ eigenvectors, which we will call *face space* (see figure 1). Thus, a facial image is represented as a point in an M-dimensional face space. The dimension M is a

[1] Representation is $(N-1)$-dimensional because the requirement that $\sum_i \boldsymbol{t}_i = 0$ removes one degree of freedom.

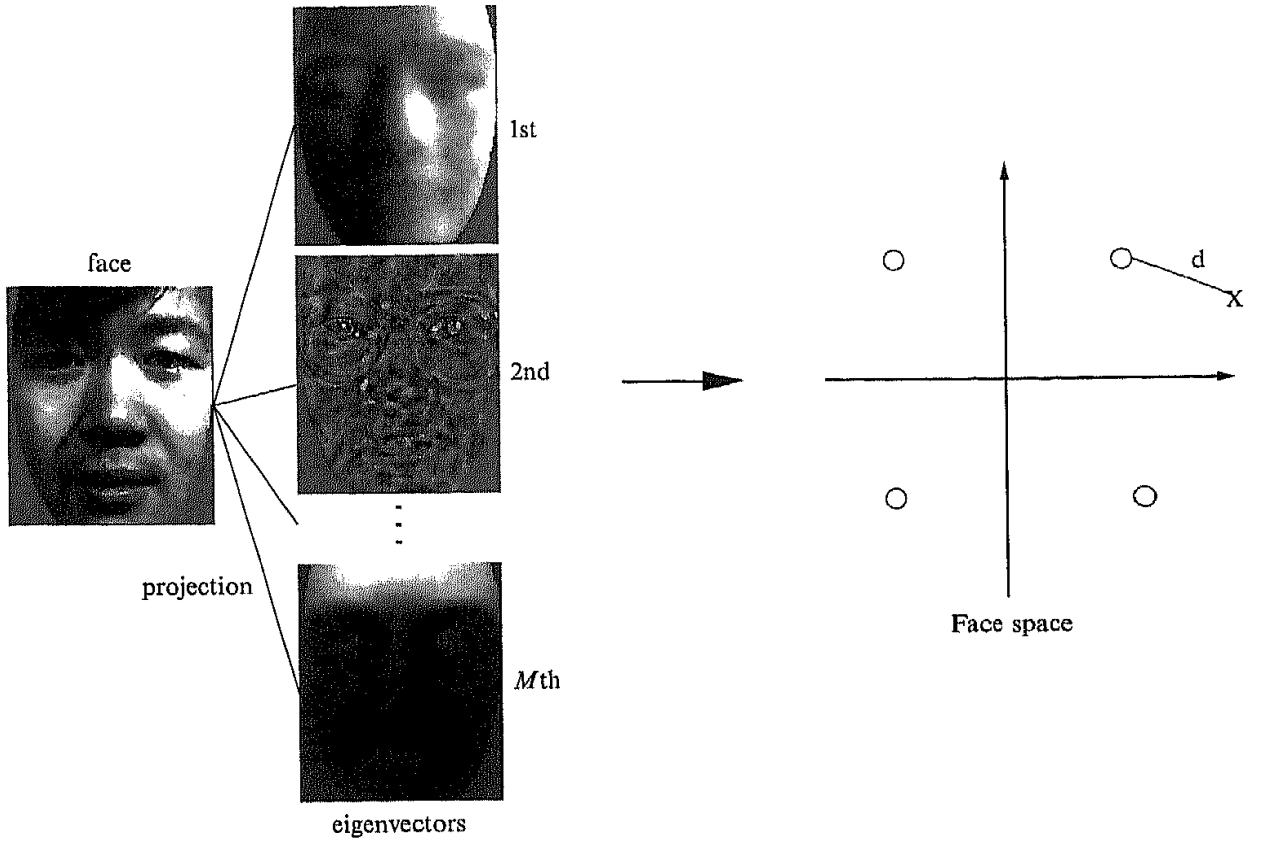

Figure 1. Representation of a face as a point in face space. A face is represented by its projection onto a subset of M eigenvectors. A set of facial images becomes a set of points 'O' in face space. The point marked by 'X' is a probe and is identified as the person in the gallery image nearest 'X'.

design decision that is discussed in the paper. A gallery of K facial images is represented as K points $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_K\}$ in face space.

A probe is identified by first projecting it into face space and then comparing the projection to all gallery images. We denote a probe by $\boldsymbol{p}_i$. A probe is compared to gallery images by a similarity measure. The similarity between probe $\boldsymbol{p}_i$ and gallery image $\boldsymbol{g}_k$ is denoted by $s_i(k)$. Two possible similarity measures are the Euclidean and L_1 distances between $\boldsymbol{p}_i$ and $\boldsymbol{g}_k$.

The identity of a probe is determined to be the gallery face, $\boldsymbol{g}_k$, that minimizes the similarity measure between $\boldsymbol{p}_i$ and the $\boldsymbol{g}_k$s. In this paper we assume that there is one image per person in the gallery, and $\boldsymbol{g}_{k^*}$ uniquely references the identity of the person. This recognition technique is called a nearest-neighbor classifier—a probe is identified as the person in the gallery image nearest the probe in face space.

2.2 *System modules*

Our face-recognition system consists of three modules and each module is composed of a sequence of steps (see figure 2). The first module normalizes the input image. The goal of the normalization module is to transform the facial image into a standard format that removes or attenuates variations that can affect recognition performance. This module consists of four steps; figure 3 shows the results of processing for some of the steps in the normalization module. The first step low-pass filters or compresses the original image. Images are filtered to remove high-frequency noise. 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 eyes to standard locations. The goal of this step is to remove variations in size, orientation, and location of the face in an image. The third step masks background pixels, hair, and clothes.

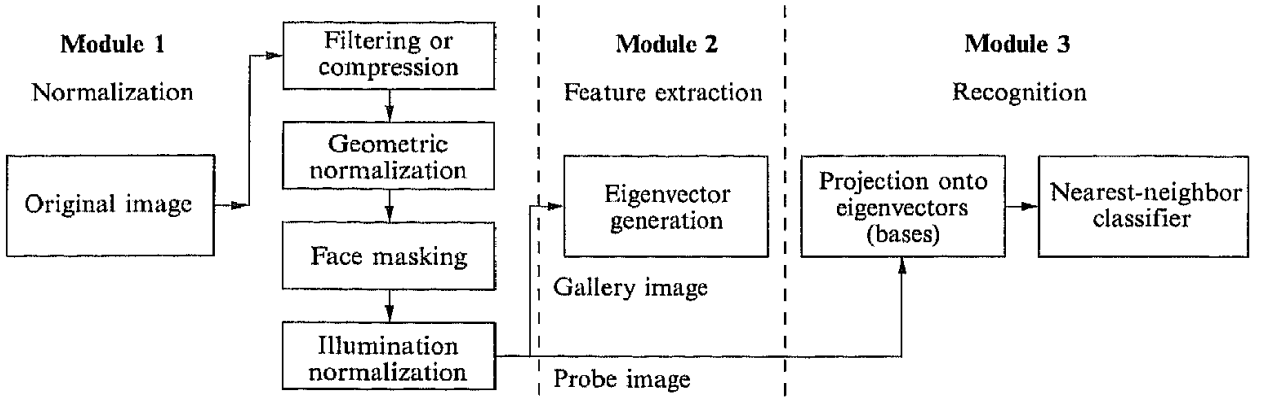

Figure 2. Block diagram of PCA-based face-recognition system.

Figure 3. An original image and the results of several steps in the normalization module.

This prevents image variations that are not directly related to the face from interfering with the identification process. The fourth step attenuates illumination variation among images, which is a critical factor in algorithm performance.

The second module performs the PCA decomposition on the training set, which produces the eigenvectors and eigenvalues. We did not vary this module. All experiments were conducted with the same training set, which was the one that was used for the PCA-baseline algorithm in the September 1996 FERET evaluation (Phillips et al 2000).

The third module identifies the face in a normalized image, and consists of two steps. The first step projects the image into face space. The critical parameter in this step is the subset of eigenvectors used to represent the face. The second step identifies faces with a nearest-neighbor classifier. Or, more precisely, the classifier ranks the gallery images by similarity to the probe. The critical design decision in this step is the similarity measure in the classifier. We presented performance results using L_1 distance, L_2 distance, angle between feature vectors, and Mahalanobis distance. Additionally, we created three new similarity measures by combining the Mahalanobis distance with the L_1, L_2, and angle similarity measures.

3 Test design

3.1 *FERET database*

The FERET database provides a common database of facial images for both development and testing of face-recognition algorithms and has become the de facto standard for face recognition from still images (Phillips et al 1998, 2000).

The images in the FERET database were initially acquired with a 35-mm camera. The film used was color Kodak Ultra. The film was processed by Kodak and placed onto a CD-ROM via Kodak's multiresolution technique for digitizing and storing digital imagery.

The colour images were retrieved from the CD-ROM and converted into 8-bit gray-scale images.[2]

The facial images were collected in 15 sessions between August 1993 and July 1996. Sessions lasted one or two days, and the location and setup did not change during a session. To maintain a degree of consistency throughout the database, the photographer used the same physical setup in each photography session. However, because the equipment had to be reassembled for each session, there was variation from session to session. This results in variations in scale, pose, expression, and illumination of the face (see figure 4). For details of the FERET database, refer to Phillips et al (1996, 1998).

Figure 4. Categories of images (example of variations).

In the FERET database, images of individuals were acquired in sets of 5 to 11 images. Each set includes two frontal views (**fa** and **fb**); a different facial expression was requested for the second frontal image. For 200 sets of images, a third frontal image was taken with a different camera and different lighting (**fc**).

One emphasis of the database collection was obtaining images of individuals on different days (duplicate sets). A *duplicate* is defined as an image of a person whose corresponding gallery image was taken on a different date or under different conditions, eg wearing glasses or with hair pulled back. The database contains 365 duplicate sets of images. For 91 duplicate sets, the time between the first and last sittings was at least 18 months.

3.2 *Design rule*

To obtain a robust comparison of algorithms, it is necessary to calculate performance on a large number of galleries and probe sets. To allow scoring on multiple galleries and probe sets, we designed a new evaluation protocol. In our protocol, during the evaluation 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 galleries and probe sets that are used in computing performance statistics.

An algorithm reports the similarity $s_i(k)$ between all query images $\boldsymbol{q}_i$ in the query set $\mathcal{Q}$ and all target images $\boldsymbol{u}_k$ in the target set $\mathcal{T}$. This property allows greater flexibility in scoring and producing a detailed analysis of performance on multiple galleries and probe sets. We can calculate performance for galleries that are subsets of the target set ($\mathcal{G} \subset \mathcal{T}$) and for probe sets that are subsets of the query set ($\mathcal{P} \subset \mathcal{Q}$). For a given gallery $\mathcal{G}$ and probe set $\mathcal{P}$, the performance scores are computed by examining similarity measures $s_i(k)$ for query images $\boldsymbol{q}_i$ that are in the probe set ($\boldsymbol{q}_i \in \mathcal{P} \subset \mathcal{Q}$) and for target image $\boldsymbol{u}_k$ that are in the gallery ($\boldsymbol{u}_k \in \mathcal{G} \subset \mathcal{T}$).

[2] Certain commercial equipment may be identified in order to adequately specify or describe the subject matter of this work. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment identified is necessarily the best available for the purpose.

In this paper we report identification results. In identification, one asks how good an algorithm is 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 a desired level of performance. The performance statistics are reported as cumulative match scores, which are plotted as a cumulative match characteristic (CMC).

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 $\mathcal{G} = \{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_K\}$ and $\mathcal{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{|\mathcal{P}|}\}$ by comparing the similarity scores $s_i(k)$ for $\boldsymbol{p}_i \in \mathcal{P}$ and $\boldsymbol{g}_k \in \mathcal{G}$. For each probe image $\boldsymbol{p}_i \in \mathcal{P}$, we sort $s_i(\cdot)$ for all gallery images $\boldsymbol{g}_k \in \mathcal{G}$. We assume that a smaller similarity score implies a closer match. The function $\mathrm{id}(i)$ gives the index of the gallery image of the person in probe $\boldsymbol{p}_i$, ie $\boldsymbol{p}_i$ is an image of the person in $\boldsymbol{g}_{\mathrm{id}(i)}$. A probe $\boldsymbol{p}_i$ is correctly identified if $s_i[\mathrm{id}(i)]$ is the smallest score for $\boldsymbol{g}_k \in \mathcal{G}$. A probe $\boldsymbol{p}_i$ is in the top n if $s_i[\mathrm{id}(i)]$ is one of the n smallest scores $s_i(\cdot)$ for gallery $\mathcal{G}$. Let R_n denote the number of probes in the top n. We report $R_n/|\mathcal{P}|$, the fraction of probes in the top n. The CMC is a plot with the rank n on the horizontal axis and the cumulative match score $R_n/|\mathcal{P}|$ on the vertical axis. The value $R_1/|\mathcal{P}|$ is the top rank score or the fraction of probes correctly identified.

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) contained in the gallery. For all results that we report, there is one image per person in the gallery; thus, the size of the gallery is the number of images in the gallery. The number of probes scored (also, size of the probe set) is $|\mathcal{P}|$. For all runs we computed a CMC. However, for most runs we only report the top rank score unless the top rank score is not representative of the CMC. The probe set may contain more than one image of a person and the probe set may contain an image of everyone in the gallery. Every image in the probe set has a corresponding image in the gallery. (Thus, there cannot be any false alarms.)

4 Experiment 1

The purpose of experiment 1 was to examine the effects of changing the steps in our generic PCA-based face-recognition system. We did this by establishing a baseline algorithm and then varying the implementation of selected steps one at a time. Ideally, we would test all possible combinations of variations. However, because of the number of combinations, this is not practical, and so we varied the steps individually.

The baseline algorithm has the following configuration. The images were not filtered or compressed. Geometric normalization consisted of rotating, translating, and scaling the images so the center of the eyes were on standard pixels. This was followed by masking the hair and background from the images. In the illumination normalization step, the non-masked facial pixels were normalized by a histogram-equalization algorithm (Pratt 1978). Then, the non-masked facial pixels were transformed so that the mean, μ, was equal to 0.0 and standard deviation, σ, was equal to 1.0. The geometric normalization and masking steps were not varied in the experiments in this paper.

The training set for the PCA consisted of 501 images (one image per person), which produced 500 eigenvectors. The training set was not varied in the experiments in the paper. In the recognition module, faces were represented by their projections onto the first 200 eigenvectors and the classifier used the L_1 norm.

4.1 *Test sets, galleries, and probe sets*

All images were from the FERET database, and the testing was done with the September 1996 FERET protocol. In this protocol, the target set contained 3323 images and the query set 3816 images. All the images in the target set were frontal images. The query set consisted of all the images in the target set.

To allow for a robust and detailed analysis, we report identification scores for four categories of probes (see figure 4 for examples of the four categories). The size of the galleries and probe sets for the four probe categories are presented in table 1. For three of the probe categories, performance was computed by using the same gallery. For the fourth category, a subset of the first gallery was used. The first gallery consisted of images of 1196 people with one image per person. For the 1196 people, the target and query sets contain **fa** and **fb** images from the same set. (The FERET images were collected in sets, and in each session there are two frontal images, **fa** and **fb**, see section 3.1.) One of these images was placed in the gallery and the other was placed in the **FB** probe set. The **FB** probes were the first probe category. (This category is denoted by **FB** to differentiate it from the **fb** images in the FERET database.) (Note: the query set contained all the images in the target set, so the probe set is a subset of the query set.) Thus, the **FB** probe set consisted of probe images taken on the same day and under the same illumination conditions as the corresponding gallery image.

Table 1. Size of galleries and probe sets for the four probe categories.

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

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** probes (images taken the same day as the corresponding gallery image, but with a different camera and lighting). The fourth category consisted of duplicates for which there was at least one year between the acquisition of the probe image and the corresponding gallery image; ie the gallery images were acquired before January 1995 and the probe images were acquired after January 1996. We refer to this category as the duplicate II probes. The gallery for the **FB**, duplicate I, and **fc** probes was the same. The gallery for duplicate II probes was a subset of 864 images from the gallery for the other categories. The smaller-sized gallery insured that there was at least one year between acquisition of gallery images and probes.

4.2 *Variations in the normalization module*

4.2.1 *Illumination normalization.* We experimented with three variations in the illumination normalization step. For the baseline algorithm, the non-masked facial pixels were transformed so that the mean was equal to 0.0 and standard deviation was equal to 1.0 followed by a histogram-equalization algorithm. For the first variation, the non-masked pixels were not normalized (original image). For the second variation, the non-masked facial pixels were normalized with a histogram-equalization algorithm. For the third variation, the non-masked facial pixels were transformed so that the mean was equal to 0.0 and standard deviation equal to 1.0. The eigenvectors were regenerated for each of the illumination normalization variations. The performance results from the illumination normalization methods are presented in table 2.

Table 2. Performance results for illumination normalization methods. Performance scores are the top rank matches. μ = mean; σ = standard deviation.

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

4.2.2 *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 prior to being processed by the geometric normalization step of the normalization module. For both compression methods, the images were compressed approximately 16 : 1 (0.5 bits/pixel). We experimented with other compression ratios and found that performance was comparable. The results are for eigenvectors generated from non-compressed or filtered images. We found that regenerating the eigenvectors reduced performance. 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 3×3 spatial filter with a center value of 0.2 and the remaining values equal to 0.1. Table 3 reports performance for the baseline algorithm, JPEG and wavelet compression, and low-pass filtering.

Table 3. Performance score for low-pass filter, JPEG, and wavelet compressed images (0.5 bits/pixel compression). Performance scores are the top rank matches.

Normalization	Probe category			
	duplicate I	duplicate II	**FB**	**fc**
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

4.3 *Variations in the recognition module*

4.3.1 *Number of low-order eigenvectors.* The higher-order eigenvectors, which are associated with smaller eigenvalues, encode small variations and noise among the images in the training set. One would expect that the higher-order eigenvectors would not contribute to recognition, and removing them from the representation would improve performance. We examined this hypothesis by computing performance as a function of the number of low-order eigenvectors in the representation. 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. Figure 5 shows the top rank score for **FB** and duplicate I probes as the function of the number of low-order eigenvectors included in the representation in face space. This shows that performance increases as the first 150 eigenvectors were added to the representation. For eigenvectors 150 to 225 there was very little change in performance, and after 225 eigenvectors were included, performance slowly decreased.

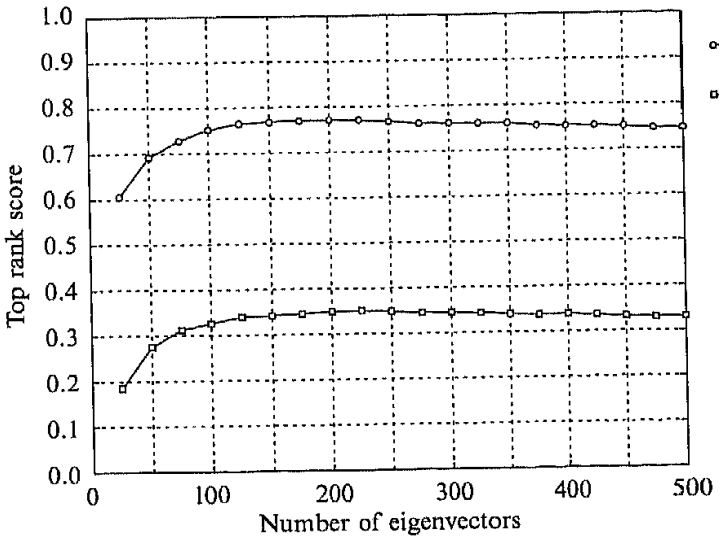

Figure 5. Performance on duplicate I and **FB** probes based on number of low-order eigenvectors in the representation. (Number of images in gallery = 1196; number of **FB** images in probe set = 1195; number of duplicate I in probe set = 722.)

4.3.2 *Removing low-order eigenvectors.* The low-order eigenvectors encode gross differences among the images in 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 1st, 2nd, 3rd, and 4th eigenvectors from the representation; ie the representation consisted of e_i, ..., e_{200}, $i = 1, 2, 3, 4, 5$. The performance results from these variations are given in table 4. Among the different category of probes, there is a noticeable variation in performance for **fc** probes as shown in figure 6.

Table 4. Performance of the baseline algorithm with low-order eigenvectors removed. Performance scores are the top rank matches.

Number of low-order eigenvectors removed	Probe category			
	duplicate I	duplicate II	**FB**	fc
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
4	0.20	0.09	0.50	0.22

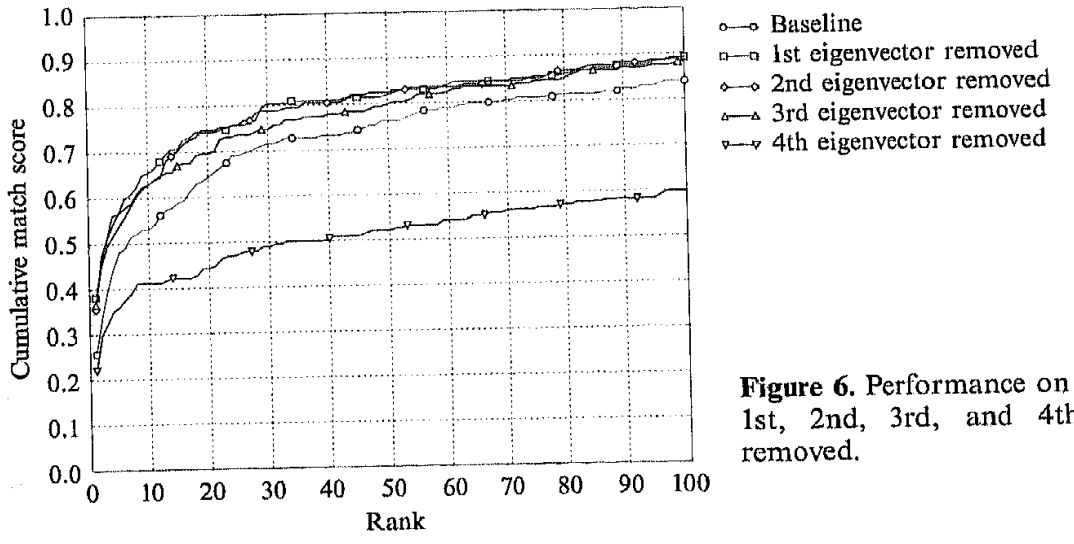

Figure 6. Performance on **fc** probes with 1st, 2nd, 3rd, and 4th eigenvectors removed.

4.3.3 *Nearest-neighbor classifier.* We experimented with seven similarity measures for the classifier, which are listed in table 5, along with their performance results. (Details of the similarity measures are given in appendix A.) Among the four categories of probes, the **fc** probes show the most variation in performance across the seven classifiers. Because of this variation, we present the cumulative match scores for the **fc** probes in figure 7.

Table 5. Performance scores based on different nearest-neighbor classifier. Performance scores are the top rank matches.

Nearest-neighbor classifier	Probe category			
	duplicate I	duplicate II	**FB**	**fc**
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

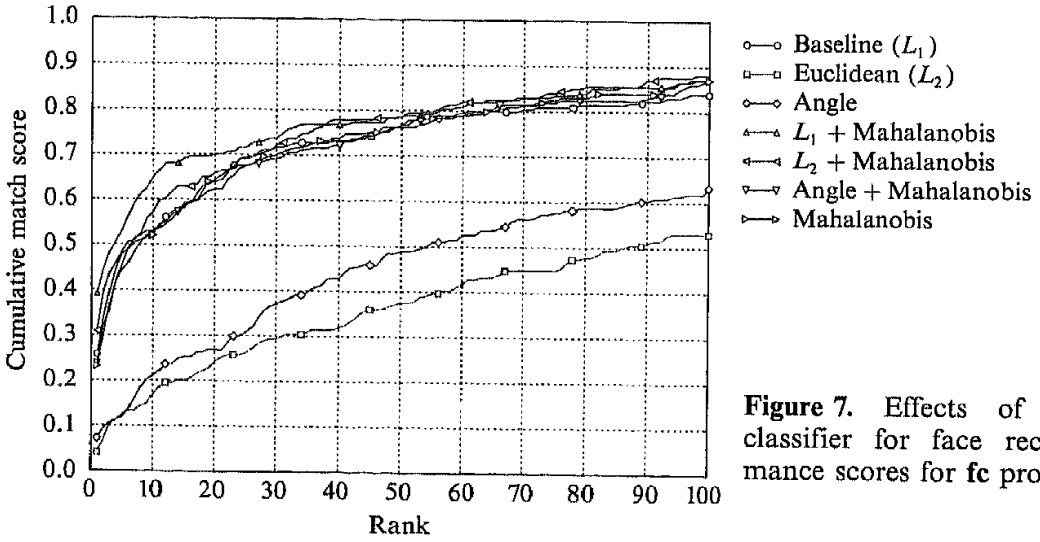

Figure 7. Effects of nearest-neighbor classifier for face recognition. Performance scores for **fc** probes.

4.4 *Discussion*

In experiment 1, we conducted experiments that systematically varied the steps in each module on the basis of our PCA-based face-recognition system. The goal of this experiment was to understand the effects of these variations on performance.

In the normalization module, we experimented with varying the illumination normalization and compression steps. The results show that performing an illumination normalization step improves performance, but the particular implementation is not critical. The results also show that compressing or filtering the images does not affect performance significantly.

In the recognition module, we experimented with the three classes of variations. First, we varied the number of low-order eigenvectors in the representation from 50 to 500 by steps of 50. Figure 5 shows that performance increases until approximately 200 eigenvectors are included in the representation and then performance decreases slightly. Representing faces by the first 40% of the eigenvectors is consistent with results on other facial image sets.

Second, removing the 1st eigenvector resulted in an overall increase in performance (table 4). The largest increase was observed with the **fc** probes. This increase is further highlighted in figure 6. 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 1st eigenvector.

Third, changing the similarity measure in the nearest-neighbor classifier produced the largest variation in performance. For duplicate I probes, performance ranged from 0.31 to 0.45, and for **fc** probes the range was from 0.07 to 0.39. 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. But, this distance was the worst for the duplicate I probe.

Both removing low-order eigenvectors from the representation and changing the similarity measure improved performance over the baseline algorithm. This naturally raises the following question: what is the effect of combining an alternative similarity measure and removing the low-order eigenvectors? Does the improvement in performance for both variations come from exploiting the same property in facial images, and combining them will not improve performance? Or, does each method exploit different properties, and will combining the two variations increase performance beyond that achieved for individual variations? To examine this question, we looked at removing the low-order eigenvectors for the angle + Mahalanobis and L_1 + Mahalanobis similarity measures. We selected the angle + Mahalanobis measure because it exhibited the best performance for the duplicate I and II, and **FB** probes, and the L_1 + Mahalanobis measure because it had the best performance for the **fc** probes. The results appear in table 6.

Table 6. Performance of the angle + Mahalanobis classifier and of the L_1 + Mahalanobis classifier with low-order eigenvectors removed. Performance scores are the top rank matches.

Number of low-order eigenvectors removed	Probe category			
	duplicate I	duplicate II	**FB**	**fc**
Angle + Mahalanobis classifier				
0	0.45	0.21	0.77	0.24
1	0.45	0.22	0.77	0.46
2	0.44	0.21	0.77	0.47
3	0.44	0.19	0.79	0.46
4	0.44	0.19	0.79	0.43
L_1 + Mahalanobis classifier				
0	0.31	0.13	0.73	0.39
1	0.30	0.13	0.73	0.39
2	0.30	0.13	0.72	0.41
3	0.30	0.12	0.72	0.40
4	0.29	0.12	0.72	0.40

For the L_1 + Mahalanobis similarity measure there was only a slight increase in performance for **fc** probes. However, for the angle + Mahalanobis similarity measure there was a substantial increase in performance for **fc** probes from 0.24 for no eigenvectors removed to 0.47 for two eigenvectors removed. This was an improvement over all L_1 + Mahalanobis similarity measure results. For both classifiers, there was a slight change in performance for the three remaining probe categories. This was consistent with the results in table 4 when low-order eigenvectors were removed from the baseline algorithm.

In combining removal of the low-order eigenvectors with changes in the similarity measure we found an overall increase in performance for **fc** probes with the angle + Mahalanobis similarity measure. Thus, for this similarity measure, combining the two variations increased performance over implementing just one of the variations.

Because of the variation in performance, it is clear that selecting the similarity measure for the classifier is the critical decision in designing a PCA-based face-recognition system. The second critical decision is deciding if removing the low-order eigenvectors is appropriate for the selected classifier. However, both these decisions are dependent on the type of images in the galleries and probe sets that the system will process.

5 Experiment 2

In experiment 1, for some variations in components, the range of performance was small, whereas, for others, the range was considerable, ie the nearest-neighbor classifier. The natural question is: When is the difference in performance between two variations significant? In experiment 2 we examine this question by quantifying the range of performance for each of the similarity measures in the previous experiment on 100 galleries. We selected the similarity measures because they had the greatest effect on performance of the variations studied in the previous experiment.

To address this question, we randomly generated 100 galleries of 200 individuals, with one frontal image per person. Each gallery was generated without replacement from the **FB** gallery of 1196 individuals in experiment 1. (Thus, there was overlap between galleries.) Then we scored each of the galleries against the **FB** and duplicate I probes for each of the seven classifiers in experiment 1. (There were not enough **fc** and duplicate II probes for all random galleries to compute performance statistics.) For each randomly generated gallery, the corresponding **FB** probe set consisted of the second frontal image for all images in that gallery; the duplicate I probe set consisted of all duplicate images for each image in the gallery. We measured performance by the top rank score using the fraction of probes that were correctly identified.

For an initial look at the range in performance, we examined the baseline algorithm (L_1 similarity measure). For each classifier and probe category, we had 100 different scores. Figure 8 presents the histogram of top rank scores for the baseline algorithm for both **FB** and duplicate I probe sets. This shows a range in performance ranges from 0.80 to 0.92 for **FB** probes, from 0.29 to 0.59 for duplicate I probe. There is clearly a large range in performance across the 100 galleries. There were similar distributions of scores for the six remaining similarity measures.

We summarize performance with a truncated range of top rank scores for the seven different nearest-neighbor classifiers in figure 9. Figure 9a shows the range for **FB** probes and figure 9b for duplicate I probes. For each classifier, we mark the median by $\times$, the 10th percentile by $+$, and 90th percentile by $*$. We plotted these statistics because they are robust and insensitive to outliers. From these studies, we get a robust estimate of the overall performance of each classifier.

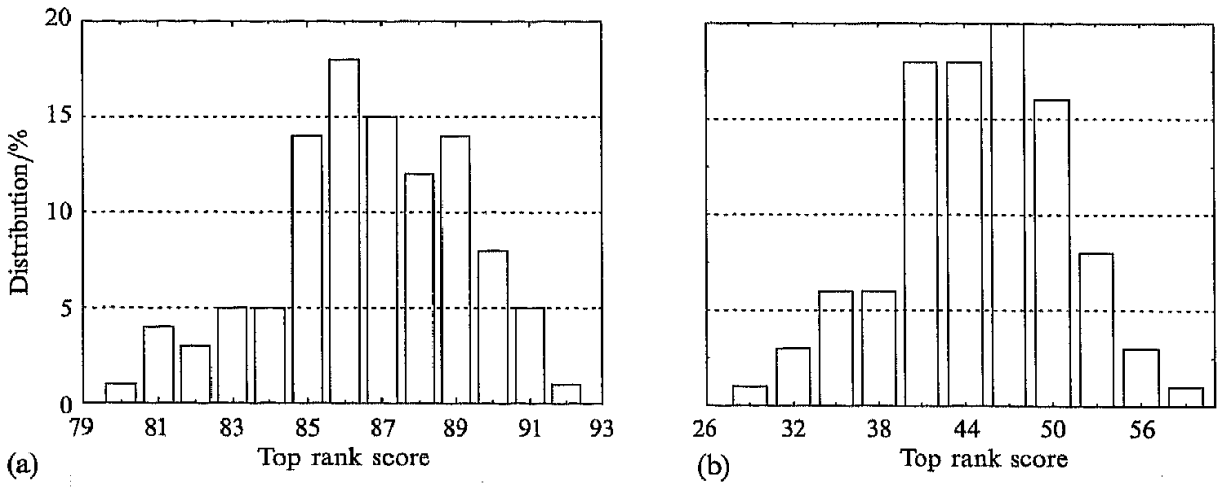

Figure 8. Histogram of top rank scores of the baseline algorithm (L_1 similarity measure) for (a) **FB** probes and (b) duplicate I probes.

5.1 *Discussion*

The main goal of experiment 2 was to estimate when the difference in performance was significant. From figure 9, the range in scores is approximately ± 0.05 about the median for all 14 runs. This suggests a reasonable threshold for measuring significant difference in performance for the classifiers is ~ 0.10.

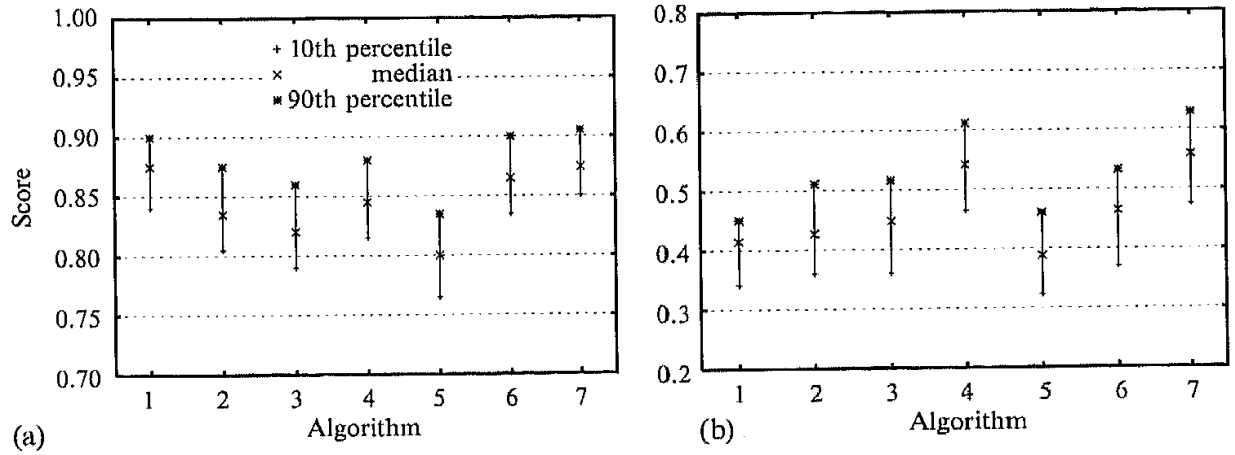

Figure 9. The range of top rank scores from seven different nearest-neighbor classifiers. The nearest neighbor-classifiers presented are: (1) L_1, (2) L_2, (3) angle, (4) Mahalanobis, (5) L_1 + Mahalanobis, (6) L_2 + Mahalanobis, and (7) angle + Mahalanobis. (a) **FB** probes and (b) duplicate I probes.

The results for duplicate I probes in experiment 2 are consistent with the results in experiment 1. In table 5, the top classifiers were the Mahalanobis and angle + Mahalanobis and these two classifiers produce better performance than the other methods as shown in table 7. In both experiments, the L_1 + Mahalanobis received the lowest scores. This suggests that for duplicate I scores the angle + Mahalanobis or Mahalanobis distance should be used. It follows from the results of this experiment that performance of smaller galleries can predict relative performance on larger galleries.

For **FB** probes, there was not as sharp a division among classifiers. One possible explanation is that in experiment 1 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 performance of the remaining classifiers can be grouped together. The performance scores of these classifiers are within each other's error margins. This suggests that either the L_1, L_2 + Mahalanobis, or angle + Mahalanobis distance should be used.

6 Conclusion

We have presented a design methodology of configuring PCA-based algorithms based on empirical performance results. The heart of the methodology is a generic modular design for PCA-based face-recognition systems. This allowed us to systematically vary the components and measure the impact of these variations on performance. Our experiments show that the quality and type of images to be processed are the driving factors in determining the performance of PCA-based systems.

On the basis of these experiments, we propose a new algorithm that is a combination of the variations studied. The components of the proposed algorithm are:

- perform illumination normalization ($\mu = 0.0$ and $\sigma = 1.0$),
- low-pass filter the images,
- remove the first low-order eigenvector, and
- employ the angle + Mahalanobis similarity measure.

Table 7 presents the identification scores for the baseline and proposed algorithms, and the combined variation of angle + Mahalanobis classifier and removing the first two eigenvectors. For **FB** probes, the scores for all three algorithms are not significantly different. The proposed algorithm has better performance scores for duplicate I and II probe sets. The algorithm with angle + Mahalanobis classifier and removing the first two eigenvectors has better performance scores for **fc** probes. This shows that a substantial increase in performance can be achieved over the baseline algorithm, and the design of the best algorithm is not necessarily one of the standard configurations in the literature.

Table 7. Comparison of baseline and proposed algorithms, and combination angle + Mahalanobis and removal of first two eigenvectors. Performance scores are the top rank matches.

Algorithm	Probe category			
	duplicate I	duplicate II	**FB**	**fc**
Baseline	0.35	0.13	0.77	0.26
Proposed	0.49	0.26	0.78	0.26
Angle + Mahalanobis and remove two eigenvectors	0.44	0.21	0.77	0.47

Another important observation from these results is that the effect on performance of combining variations is nonlinear. This is illustrated by two cases from our experiments. In the first case, combining the angle + Mahalanobis similarity measure with removal of the leading eigenvectors produced an increase in performance greater than the individual variations for **fc** probes. For **fc** probes, changing to the baseline algorithm (L_1) to angle + Mahalanobis similarity resulted in a decrease in performance from 0.26 to 0.24, and removing the leading eigenvector resulted in an increase in performance from 0.26 to 0.38. The combination of these two variations resulted in performance of 0.47, which is greater than the sum of the individual variations. In case two, which is the other end of the spectrum, we combined the L_1 + Malahanobis distance and removing the leading eigenvectors. Both variations individually increased performance for **fc** probes, but combined they did not produce a larger change.

From the series of experiments with PCA-based face-recognition systems, we have come to five major conclusions.

First, the selection of the nearest-neighbor classifier is the most critical design decision for PCA-based algorithms. Proper selection of the nearest-neighbor classifier is essential to achieve the best possible performance scores. Furthermore, we have looked at similarity measures that achieve better performance than those generally considered in the literature.

Second, for the performance difference between two algorithms to be significant, there needs to be at least a 0.10 difference in the cumulative match scores.

Third, performance scores vary among the probe categories. Thus, the design of an algorithm needs to take into account the type of images that the algorithm will process. 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 suggests that, when comparing algorithms, performance scores from multiple 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.

Fifth, 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.

For psychophysics studies, our conclusions have a number of implications. First, face-recognition studies should include a range of image qualities. For example, when measuring the concord between humans and algorithms, the results should be based on experiments for more than one type of facial image. Second, the details of an algorithm implementation can have significant impact on results and conclusion. By pointing out the most significant variations in an implementation, the accord between these variations and humans can be measured. An example of this is found in O'Toole et al (2000), which included the different classifier variations in a study. This study showed that the classifier makes a difference in how faces are perceived. More significantly, the classifiers fell into the same two classes as humans. Without studies like the one in this paper, one would not have been able to easily determine what variations of a PCA-based algorithm should be included in studies like O'Toole et al. This could result in researchers failing to observe key properties of how humans and algorithms perceive and process faces.

Acknowledgements. The authors thank Alice O'Toole for many helpful and insightful comments. The work reported here is part of the Face Recognition Technology (FERET) program, which was sponsored by the US Department of Defense Counterdrug Technology Development Program. Portions of this work were done while Jonathon Phillips was at the US Army Research Laboratory. Jonathon Phillips acknowledges the support of the National Institute of Justice and the Defense Advance Research Projects Agency.

References

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

Barlett M S, Lades H M, Sejnowski T J, 1998 "Independent component representations for face recognition" *Proceedings of the SPIE: Conference on Human Vision and Electronic Imaging III* **3299** 528 – 539

Belhumeur P, Hespanha J, Kriegman D, 1997 "Eigenfaces vs fisherfaces: Recognition using class specific linear projection" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** 711 – 720

Brunelli R, Poggio T, 1993 "Face recognition: Features versus templates" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15** 1042 – 1052

Cottrell G W, Metcalfe J, 1991 "Empath: Face, gender and emotion recognition using holons", in *Advances in Neural Information Processing Systems 3* Eds R P Lippman, J E Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann Publishers) pp 564 – 571

Etemad K, Chellappa R, 1997 "Discriminant analysis for recognition of human face images" *Journal of the Optical Society of America A* **14** 1724 – 1733

Fleming M, Cottrell G W, 1990 "Categorization of faces using unsupervised feature extraction", in *Proceedings of the International Joint Conference on Neural Networks* volume 2 (Ann Arbor, MI: IEEE Neural Networks Council) pp 65 – 70

Fukunaga K, 1972 *Introduction to Statistical Pattern Recognition* (Orlando, FL: Academic Press)

Hancock P J B, Burton A M, Bruce V, 1996 "Face processing: human perception and principal component analysis" *Memory & Cognition* **24** 26 – 40

Jolliffe I T, 1986 *Principal Component Analysis* (Berlin: Springer)

Kirby M, Sirovich L, 1990 "Application of the Karhunen – Loeve procedure for the characterization of human faces" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12** 103 – 108

Liu C, Wechsler H, 1999 "Comparative assessment of independent component analysis (ICA) for face recognition", in *2nd International Conference on Audio- and Video-based Biometric Person Authentication* (College Park, MD: Department of Computer Science, University of Maryland) pp 211 – 216

Macmillan N A, Creelman C D, 1991 *Detection Theory: A User's Guide* (Cambridge: Cambridge University Press)

Moghaddam B, Pentland A, 1994 "Face recognition using view-based and modular eigenspaces" *Proceedings of the SPIE: Conference on Automatic Systems for the Identification and Inspection of Humans* **2277** 12–21

Moghaddam B, Pentland A, 1995 "Maximum likelihood detection of faces and hands", in *International Workshop on Automatic Face and Gesture Recognition* Ed. M Bichsel (Zurich: Multimedia Laboratory, University of Zurich) pp 122–128

Moghaddam B, Pentland A, 1998 "Beyond linear eigenspaces: Bayesian matching for face recognition", in *Face Recognition: From Theory to Applications* Eds H Wechsler, P J Phillips, V Bruce, F Fogelman Soulie, T S Huang (Berling: Springer) pp 230–243

O'Toole A J, Abdi H, Deffenbacher K A, Bartlett J C, 1991 "Classifying faces by race and sex using an autoassociative memory trained for recognition", in *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* Eds K J Hammond, D Gentner (Hillsdale, NJ: Lawrence Erlbaum Associates) pp 847–851

O'Toole A J, Abdi H, Deffenbacher K A, Valentin D, 1993 "Low-dimensional representation of faces in higher dimensions of the face space" *Journal of the Optical Society of America A* **10** 405–411

O'Toole A J, Millward R B, Anderson J A, 1988 "A physical system approach to recognition memory for spatially transformed faces" *Neural Networks* **1** 179–199

O'Toole A J, Phillips P J, Cheng Y, Ross B, Wild H A, 2000 "Face recognition algorithms as models of human face processing", in *Proceedings of IEEE Fourth International Conference on Face and Gesture Recognition* (Los Alamitos, CA: IEEE Computer Society Press)

O'Toole A J, Vetter T, Troje N, Bülthoff H, 1997 "Sex classification is better with three-dimensional head structure than with intensity information" *Perception* **26** 75–84

Penev P, Atick J, 1996 "Local feature analysis: a general statistical theory for object representation" *Network: Computation in Neural Systems* **7** 477–500

Phillips P J, 1999a "On performance statistics for biometric systems", in *AutoID'99 Proceedings* pp 111–116

Phillips P J, 1999b "Support vector machines applied to face recognition", in *Advances in Neural Information Processing Systems 11* Eds M S Kearns, S A Solla, D A Cohn (Cambridge, MA: MIT Press) pp 803–809

Phillips P J, Moon H, Rauss P, Rizvi S, 1997 "The FERET evaluation methodology for face-recognition algorithms", in *Proceedings of Computer Vision and Pattern Recognition 97* (Los Alamitos, CA: IEEE Computer Society Press) pp 137–143

Phillips P J, Moon H, Rizvi S, Rauss P, 2000 "The FERET evaluation methodology for face-recognition algorithms" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** 1090–1104

Phillips P J, Rauss P, Der S, 1996 *FERET (face recognition technology) Recognition Algorithm Development and Test Report* Technical Report ARL-TR-995, U.S. Army Research Laboratory, Adelphi, MD

Phillips P J, Wechsler H, Huang J, Rauss P, 1998b "The FERET database and evaluation procedure for face-recognition algorithms" *Image and Vision Computing Journal* **16** 295–306

Pratt W K, 1978 *Digital Image Processing* (New York: John Wiley & Sons)

Rizvi S, Phillips P J, Moon H, 1998 "A verification protocol and statistical performance analysis for face recognition algorithms", in *Computer Vision and Pattern Recognition 98* (Los Alamitos, CA: IEEE Computer Society Press) pp 833–838

Sung K-K, Poggio T, 1998 "Example-based learning for view-based human face detection" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** 39–51

Swets D, Weng J, 1996 "Using discriminant eigenfeatures for image retrieval" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18** 831–836

Turk M, Pentland A, 1991 "Eigenfaces for recognition" *Journal of Cognitive Neuroscience* **3** 71–86

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

Valentin D, Abdi H, O'Toole A J, Cottrell G W, 1994 "Connectionist models of face processing: a survey" *Pattern Recognition* **27** 1209–1230

Valentine T (Ed.), 1995 *Cognitive and Computational Aspects of Face Recognition* (London: Routledge)

Wilder J, Phillips P J, Jiang C, Wiener S, 1996 "Comparison of visible and infrared imagery for face recognition", in *2nd International Conference on Automatic Face and Gesture Recognition* (Los Alamitos, CA: IEEE Computer Society Press) pp 182–187

Zhao W, Krishnaswamy A, Chellappa R, Swets D, Weng J, 1998 "Discriminant analysis of principal components for face recognition", in *Face Recognition: From Theory to Applications* Eds H Wechsler, P J Phillips, V Bruce, F Fogelman Soulie, T S Huang (Berlin: Springer) pp 73–85

Appendix

We mathematically describe the similarity measure used in the nearest-neighbor classifiers. The variables x, y, and z are k-dimensional vectors and x_i, y_i, and z_i are the ith components of the vectors.

A1 L_1 distance:

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

A2 L_2 distance:

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

A3 Angle between feature vectors:

$$d(x, y) = -\frac{x \cdot y}{\|x\| \|y\|} = -\frac{\sum_{i=1}^{k} x_i y_i}{\left[\sum_{i=1}^{k} (x_i)^2 \sum_{i=1}^{k} (y_i)^2\right]^{1/2}}$$

A4 Mahalanobis distance:

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

$$z_i = \frac{1}{\lambda_i^{1/2}},$$

where λ_i = eigenvalue of ith eigenvector. The values z_i are used in the following three distances.

A5 L_1 + Mahalanobis distance:

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

A6 L_2 + Mahalanobis distance:

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

A7 Angle + Mahalanobis distance:

$$d(x, y) = -\frac{\sum_{i=1}^{k} x_i y_i z_i}{\left[\sum_{i=1}^{k} (x_i)^2 \sum_{i=1}^{k} (y_i)^2\right]^{1/2}}.$$