# Biometric Person Authentication for Access Control Scenario Based on Face Recognition

Hyeonjoon Moon* and Taekyoung Kwon

School of Computer Engineering
Sejong University, Seoul, 143-747, Korea
hmoon@sejong.ac.kr

**Abstract.** There are tremendous need increase for personal verification and identification in internet security, electronic commerce and access control in recent years. Also, as the demands for security in many applications such as data protection and financial transaction become an increasingly relevant issues, the importance of biometric technology is rapidly increasing. In this paper, we explored face recognition system for person authentication. We explicitly state the design decisions by introducing a generic modular PCA face recognition system. We designed implementations of each module, and evaluate the performance variations based on virtual galleries and probe sets. We perform experiments and report results using equal error rates (EER) based on verification scenario for access control applications.

## 1 Introduction

As the demands for security in many applications become an increasingly relevant issues, the importance of biometric authentication is rapidly increasing. In addition to research based on improving the performance of personal authentication and evaluation technologies, standardizations are emerging to provide a common interface and to permit an effective comparison and evaluation of different biometric technologies. We explored core technologies for uni-modal biometric based on face recognition system and provide performance evaluation technology for higher reliability of core biometric algorithms.

In this paper, we present a modular projection-based face recognition system [6] for access control applications. Our face recognition system consists of normalization, PCA [5] projection, and recognition modules [1,10]. Based on the modular design for projection-based algorithms, we evaluate different implementations [8,9]. Because we use a generic model, we can change the implementation in an orderly manner and assess the impact on performance of each modification [4]. We report verification performance score for each category of probes using equal error rate (EER) which is critical factor for access control applications.
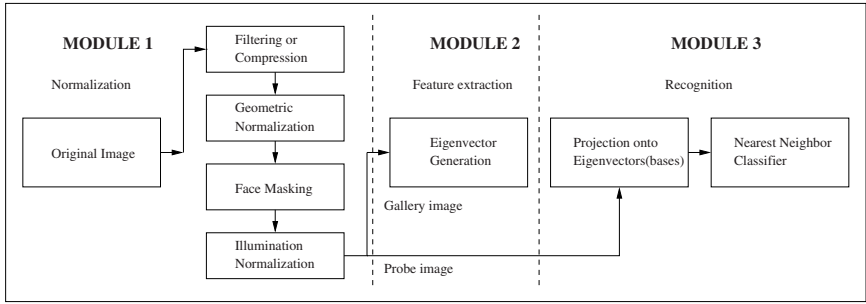
**Fig. 1.** Block Diagram of Projection-based Face Recognition System

In biometric person authentication, two critical questions are often ignored [7]. First, how does performance vary with different galleries and probe sets. Second, when is a difference in performance between two algorithms statistically significant. We systematically designed the modular based face recognition system and explored this question by analyzing randomly generated 100 galleries of the same size. We then calculate 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-based face recognition system.

## 1.1   System Modules

Our face recognition system consists of three modules and each module is composed of a sequence of steps (see Figure 1).

The first module normalizes 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. 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 eyes to standard locations. 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 which can interfere verification process. The fourth module removes some of the variations in illumination between images. Changes in illumination are critical factors in algorithm performance.

The second module performs the PCA decomposition on the training set. This produces the eigenvectors (eigenfaces) and eigenvalues. We use the training set which was used for FERET program for the generation of eigenvectors.

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

**Table 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 |

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, Mahalanobis distance. Additionally, Mahalanobis distance was incorporated with L1, L2, and angle between feature vectors mentioned above.

## 2   Test Design

### 2.1   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 of still images. There were variations in scale, pose, expression, and illumination of the face. The size of the galleries and probe sets for the four probe categories are presented in Table 1.

### 2.2   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, $p \nsim g$.) Performance of the system is characterized by two performance statistics. 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 $p \nsim g$. 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 [3], this decision rule maximized the verification rate for a given false alarm rate $\alpha$. Changing $c$ generated a new $P_V$ and $P_F$. By varying $c$ from it's 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 a receiver operating characteristic (ROC) (also known as the relative operating characteristic) [2,3]. The input to the scoring algorithm was $s_i(k)$; thresholding similarity scores, and computing $P_V$, $P_F$, and the ROCs was performed by the scoring algorithm.

The above method computed a ROC for an individual. However, we need performance over a population of people. To calculate a 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 $p \nsim g_i$.

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, which 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 was 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 $+\infty$.

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 $p \sim g$.

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

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

## 3   Experiment

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.

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.

First variation,the non-masked pixels were not normalized (original image). Second variation, the non-masked facial pixels were normalized with a histogram equalization algorithm. Third variation, the non-masked facial pixels were transformed so that the mean was equal to 0.0 and variance equal to 1.0. The original images were compressed and then uncompress prior to being feed into the geometric normalization step of the normalization module. For both compression methods, the images were compressed approximately 16:1 (0.5 bits per pixel).

In the normalization module, we varied the illumination normalization and compression steps. The results show that performing an illumination normalization step improves verification performance but which implementation that is selected is not critical. Based on our experiments, compression or filtering the images does not significantly effect performance.

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. Based on our experiments, the recognition performance increases until approximately 150–200 eigenvectors in the representation and then performance decreases slightly. Representing faces by the first 30–40% of the eigenvectors is consistent with results on other facial image sets that the authors have seen.

Second, the similarity measure in the nearest neighbor classifier was changed. This variation showed the largest range of verification performance. The range of performance variation shows that selecting the similarity measure for the classifier is the critical decision in designing a PCA-based face recognition system.

### 3.1   Variations in the Normalization Module

Our experiment includes three variations to 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. First variation,the non-masked pixels were

not normalized (original image). Second variation, the non-masked facial pixels were normalized with a histogram equalization algorithm. Third variation, the non-masked facial pixels were transformed so that the mean was equal to 0.0 and variance equal to 1.0.

## 3.2   Variations in the Recognition Module

**Nearest Neighbor Classifier.** Our experiment includes seven similarity measures for the classifier [6] and its verification performance results are listed in Tables 2. The performance score for **fc** probes shows most variation among different category of probes. In Figure 2, we reported detailed verification performance results for **fc** probes.

**Table 2.** Verification performance scores based on different nearest neighbor classifier. Performance scores are equal error rate (EER).

| Nearest neighbor classifier | Probe category | | | |
| --- | --- | --- | --- | --- |
| | 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 |

## 3.3   Variations in Galleries and Probe Set

The comparison among algorithms are based on 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 of probes are 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 image per person. The galleries were generated without replacement from the **FB** gallery of 1196 individuals. Then we scored each of the galleries against the **FB** and duplicate I probes for each of the seven classifiers. (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 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.

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.
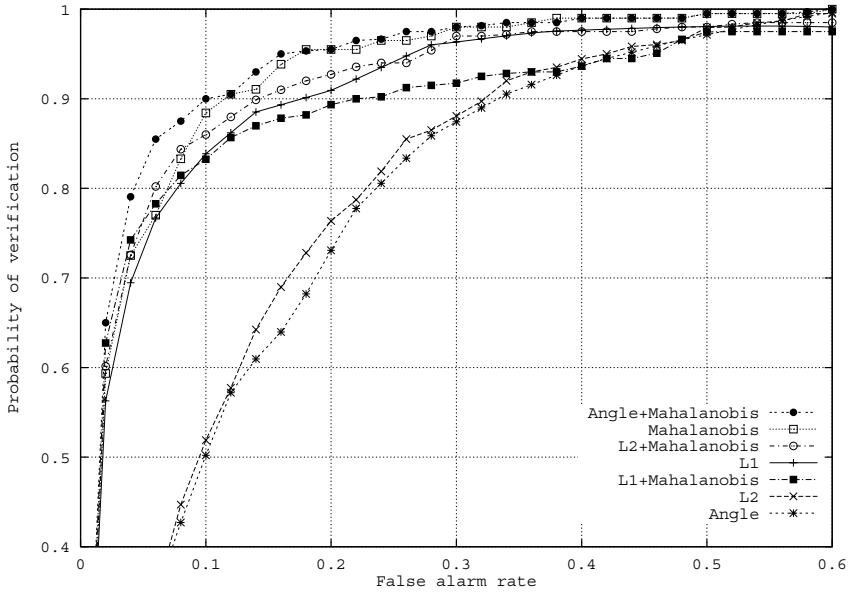
**Fig. 2.** Effects of nearest neighbor classifier on performances for **fc** probes

Performance ranges from 18.8 to 33.2 for equal error rate. This clearly shows a large range in performance of the 100 galleries.

In Figure 3, we reported a truncated range of equal error rates (%) for the seven different nearest neighbor classifiers for both **FB** and duplicate I probe sets. For each classifier, 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.

## 3.4   Discussion

The main goal of our experiment was to get a rough estimate of when the difference in performance is significant. From Figures 3, the range in verification score is approximately $\pm0.06$ about the median. This suggests a reasonable threshold for measuring significant difference in performance for the classifiers is $\sim 0.12$.

The top performing nearest neighbor classifiers were the Mahalanobis and angle+Mahalanobis. These two classifiers produces better performance than the other methods as shown in Figure 3. In this experiments, the $L_1$+Mahalanobis received the lowest verification performance scores. This suggest that for duplicate I scores that the angle+Mahalanobis or Mahalanobis distance should be used. Based on the results of this experiment, performance of smaller galleries can predict relative performance on larger galleries.
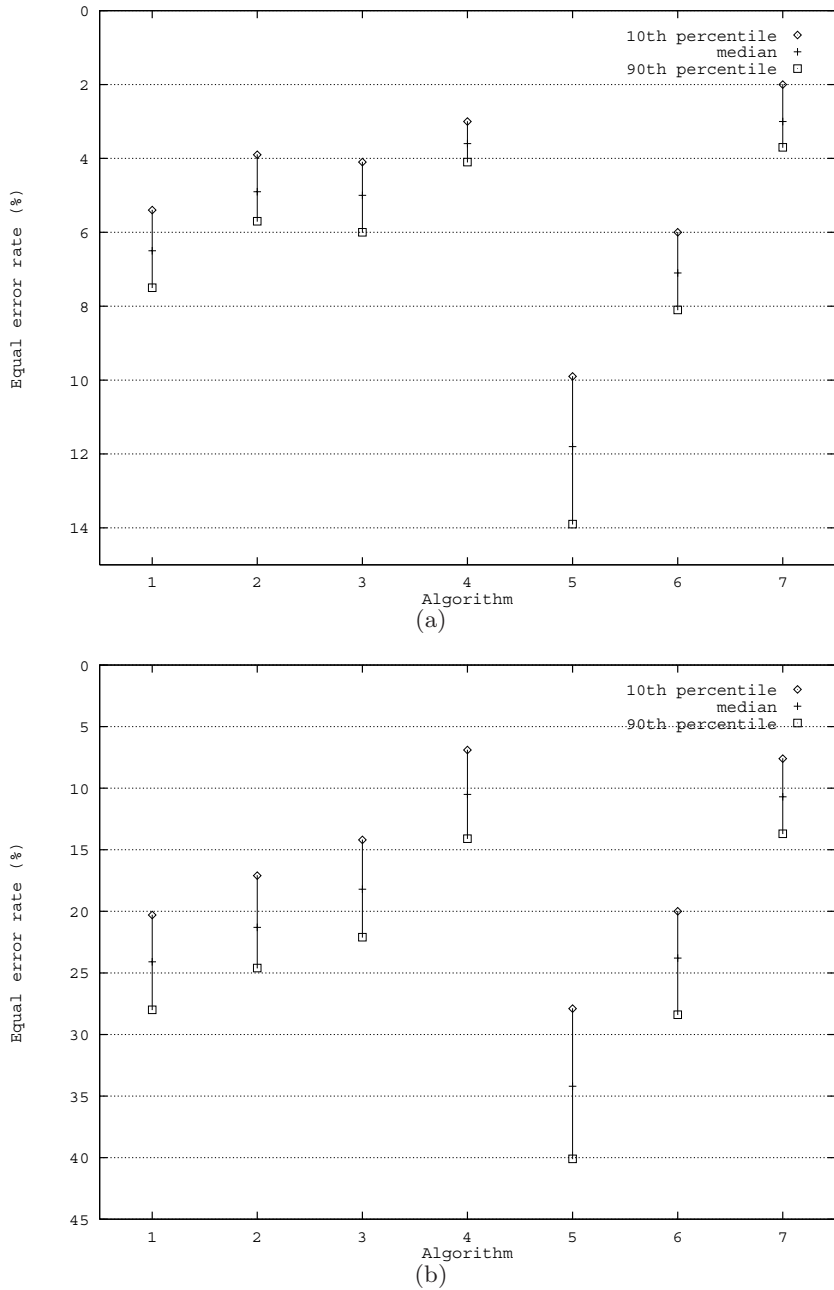
**Fig. 3.** The range of equal error rates (%) using seven different nearest neighbor classifiers. Note that the value of y-axis is in reverse order. The nearest neighbor classifiers presented are: (1) $L_1$, (2) $L_2$, (3) Angle, (4) Mahalanobis, (5) $L_1$+Mahalanobis, (6) $L_2$+Mahalanobis, and (7) Angle+Mahalanobis. (a) **FB** probes and (b) duplicate I probes.

**Table 3.** 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 [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 |

## 4   Conclusion

We proposed a biometrics person authentification system based on face recognition and evaluation procedure. In our experiment, we performed a detailed analysis of biometrics person authentification scenario based on face recognition system. The main goal of our experiment is to point out the directions for optimal configuration of PCA-based face recognition system. We introduced a modular design for PCA-based face recognition systems and systematically vary the core algorithms and measure the impact of these variations on performance. From the results throughout the series of experiments, we present two models for PCA-based face recognition system. In proposed models, our design decision includes processing steps with better performance in each module.

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

In Table 3, the verification performance for duplicate I probe is improved from 0.24 to 0.11 for Proposed I method, and duplicate II probe improved from 0.30 to 0.21 for Proposed I method (equal error rate). The verification performance score for **FB** probe shows same results for all three methods, and **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 duplicate I, duplicate II (for Proposed I method) and **fc** probes (for Proposed II method) than the baseline algorithm in verification scenario. For **FB** probes, verification results show almost identical performance scores for each method used.

From the series of experiments with optimal configuration of PCA-based face recognition system, we have come to three major conclusions.

First, image preprocessing and normalization (applying LPF, JPEG or wavelet compression) module 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 nearest neighbor classifier is essential to improve performance scores. Furthermore, our experiments shows similarity measures that achieve the best performance are not generally considered in the literature.

Third, the performance scores vary among the probe categories, and that the design of an algorithm need to consider 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.

# References

1. Brunelli, R., Poggio, T.: Face recognition: Features versus templates. IEEE Trans. PAMI, vol. 15(10) (1993)
2. Egan, J.P.: Signal Detection Theory and ROC Analysis. Academic Press, London (1975)
3. Green, D., Swets, J.: Signal Detection Theory and Psychophysics. John Wiley & Sons, New York (1966)
4. Jain, A., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. IEEE Trans. PAMI 19(2), 153–158 (1997)
5. Jolliffe, I.T.: Principal Component Analysis. Springer, Heidelberg (1986)
6. Moon, H.: Performance Evaluation Methodology for Face Recognition Algorithms. In: PhD thesis, Dept. of Computer Science and Engineering, SUNY Buffalo (1999)
7. Moon, H., Kim, J.: Biometrics identification and verification using projection-based face recognition system. In: Chae, K.-J., Yung, M. (eds.) Information Security Applications. LNCS, vol. 2908, Springer, Heidelberg (2004)
8. Moon, H., Phillips, P.J.: Computational and performance aspectsm of projection-based face recognition algorithms. Perception 30, 303–321 (2001)
9. Phillips, P.J., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. PAMI 22, 1090–1104 (2000)
10. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neuroscience 3(1), 71–86 (1991)