Modeling and Predicting Face Recognition System Performance Based on Analysis of Similarity Scores

Peng Wang, *Member*, *IEEE*,
Qiang Ji, *Sr. Member*, *IEEE*, and
James L. Wayman, *Sr. Member*, *IEEE*

Abstract—This paper presents methods of modeling and predicting face recognition (FR) system performance based on analysis of similarity scores. We define the performance of an FR system as its recognition accuracy, and consider the intrinsic and extrinsic factors affecting its performance. The intrinsic factors of an FR system include the gallery images, the FR algorithm, and the tuning parameters. The extrinsic factors include mainly query image conditions. For performance modeling, we propose the concept of "perfect recognition," based on which a performance metric is extracted from perfect recognition similarity scores (PRSS) to relate the performance of an FR system to its intrinsic factors. The PRSS performance metric allows tuning FR algorithm parameters offline for near optimal performance. In addition, the performance metric extracted from query images is used to adjust face alignment parameters online for improved performance. For online prediction of the performance of an FR system on query images, features are extracted from the actual recognition similarity scores and their corresponding PRSS. Using such features, we can predict online if an individual query image can be correctly matched by the FR system, based on which we can reduce the incorrect match rates. Experimental results demonstrate that the performance of an FR system can be significantly improved using the presented methods.

Index Terms—Face recognition, similarity scores, performance modeling, performance prediction, image quality.

1 Introduction

With current applications of computer vision systems to the automatic recognition of humans from various forms of imagery, algorithmic and system performance modeling and prediction is receiving increasing attention because of the impact such systems have on both collective security and personal privacy [14]. One form of automated human recognition currently receiving great research interest is face recognition (FR), which is the subject of ongoing US government evaluation campaigns [1]. This paper presents generic methods to model and predict the face recognition system performance based on an analysis of similarity scores. Here, the "performance" of a face recognition system is defined only as its accuracy in correctly matching face images. Other aspects of performance, such as speed, cost, availability, and maintainability, are not considered in this paper.

Many methods have been developed to improve a single recognition algorithm and to make the most use of each image [15]. However, all of the current methods have their own limitations in that they may show good accuracy under some environments, but may also show deteriorated performance under other environments due to many affecting factors. Some testing has shown that failures in recognizing persons are strongly associated with both the conditions of the collection environment and the properties of the subjects being recognized [7]. All the factors affecting the face recognition system performance can be categorized into two kinds: intrinsic and extrinsic factors. The intrinsic factors of an FR system include its gallery set, the algorithm, and its parameters.[1] The extrinsic factors are primarily image conditions of queries, including face rotation angles, expressions, and illumination.

In this paper, instead of trying to improve a specific algorithm, we present generic methods to model face recognition performance and to predict if query images can be correctly recognized by an FR system. By doing that, the performance of an FR system can be improved in the following aspects: 1) the FR system parameters can be automatically tuned only using gallery images for improved performance based on performance modeling, 2) given query images, their recognition results can be predicted to identify images that may be mismatched, so that further actions, such as human intervention, can be taken to improve recognition accuracy, and 3) the query image parameters (i.e., face alignment in this paper), can be optimized to achieve better recognition accuracy.

Specifically, our methods are based on an analysis of similarity scores output from face recognition systems. First, we present a performance metric by analyzing similarity scores calculated from intrinsic factors of an FR system. We show that the obtained performance metric can guide offline system parameter tuning without acquiring additional data. Second, we present a similarity score-based method to predict online the recognition results of query images into two cases: correct match and incorrect match. A recognition result of a query image is a correct match if the top match of the query image corresponds to a gallery image of the correct face. Usually, the recognition failure caused by an incorrect match has high cost in biometric systems, and our method, through performance prediction, intends to identify the query images whose recognitions are most likely failed. Humans can be alerted about these images for further actions, therefore improving the system performance. Third, we present a method, which is also based on analysis of similarity scores, to adjust alignment of query images (i.e., eye positions) online to achieve better recognition accuracy.

The rest of this paper is organized as follows: Related work is reviewed in Section 2. In Section 3, performance modeling for face recognition is introduced. The method of predicting face recognition results is presented in Section 4. Experimental results are discussed in Section 5. The paper concludes in Section 6 with a summary and discussion of future work.

2 Related Work

Much work has been done already on performance modeling and prediction of biometric systems such as fingerprint recognition [11], iris recognition [10], [9], and face recognition [8], [7], [3]. Specifically, in the work by Tabassi et al. [11], the quality of a fingerprint image is defined as the normalized distance between matching and nonmatching similarity scores. An 11-dimensional feature vector is extracted to identify the presence of feature points, e.g., minutia, and outliers. Then, a Neural Network classifier is trained using the extracted feature vector to predict query images into five levels of quality. Their method shows that images of higher quality usually have better accuracy in recognition, but their method cannot predict individual recognition results. Schmid et al. provide a probabilistic estimation of the lower performance bound of iris recognition algorithms based on an analysis of the Hamming distance between query and gallery iris images [9]. However, the lower bounds obtained from both the Chernoff

[1] The definitions of terms used for face recognition, such as gallery, query, and rank 1 recognition, comply with FERET and FRGC specifications [8], [7].

- *P. Wang is with the Department of Radiology, Section of Biomedical Image Analysis, University of Pennsylvania, 3600 Market Street, Suite 380, Philadelphia, PA 19104. E-mail: wpeng@ieee.org.*
- *Q. Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: qji@ecse.rpi.edu.*
- *J.L. Wayman is with the Office of Graduate Studies and Research, San Jose State University, San Jose, CA 95192. E-mail: JLWayman@aol.com.*

Manuscript received 31 Jan. 2006; revised 24 June 2006; accepted 22 Sept. 2006; published online 18 Jan. 2007.
Recommended for acceptance by S. Prabhakar, J. Kittler, D. Maltoni, L. O'Gorman, and T. Tan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org and reference IEEECS Log Number TPAMISI-0070-0106.
Digital Object Identifier no. 10.1109/TPAMI.2007.1015.

0162-8828/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

bound theory and the Large Deviation theory only provide approximate error estimates. They cannot be used to predict either the individual recognition results or the performance of systems that do not use likelihood ratio methods for recognition.

In face recognition, some empirical evaluation methods design specific training and testing sets, such as experimental settings in FERET and FRGC [8], [7], to study the system performance. Although such methods can directly assess the performance of an FR system under general circumstances, they cannot perform online performance prediction. To understand the factors affecting the FR algorithms, Givens et al. use a generalized linear model to analyze how some human characteristics, including age, race, gender, skin, glasses, and expression, affect the face recognition accuracy [3]. Their model needs to explicitly identify each affecting factor, which is an extremely difficult task in practical implementation. In addition, the identified factors can never be exhaustive.

Some other work predicts system performance using similarity scores. Li et al. propose to cluster the similarity scores into different sets, and then use the distance among the sets as features. AdaBoost is then used to select and combine these features to detect misclassification of a face recognition system [5]. AdaBoost usually needs a large number of features and many training samples. For example, more than 10,000 samples are used in [5]. However, such a large number of training samples are difficult to collect for practical systems. The similarity scores are also used to predict closed-set Cumulative Match Characteristic (CMC) curves with a small set of gallery data [4], [13]. In these methods, the rank k recognition results are modeled using parametric models whose parameters are estimated from a small gallery set. These methods work well only when gallery and query images are obtained under the same conditions, and they cannot predict individual recognition results.

3 Performance Modeling

3.1 Model of Face Recognition Systems

There is no shortage of algorithmic approaches to face recognition [15]. Typically, the function of an FR system is to map a query image to a label that represents its identity. In an FR system, the gallery set is denoted as $G = \{g_1, g_2, \ldots, g_n\}$, consisting of n gallery images whose identity is known to the algorithm. The query set is denoted as $Q = \{x_1, x_2, \ldots, x_m\}$, consisting of m query images whose identity is unknown to the algorithm. A face recognition algorithm measures the similarity between query images and each gallery image. For rank k recognition, which is generally used in closed-set searches, the system outputs labels of the gallery images corresponding to the k top matches [8]. In this paper, unless otherwise specified, we define recognition rate as the rate of correct matches at rank 1. To quantitatively evaluate the performance of a closed-set FR system, the Cumulative Matching Characteristic (CMC) curve is used [8].

For most FR systems, the similarity score plays an important role since it connects both intrinsic and extrinsic factors of an FR system. The similarity score is denoted as $S(x_i, g_j)$ or $S(i, j)$, for the comparison between the query x_i and the gallery g_j, and a larger similarity score means a potentially closer match. In our method of closed-set analysis, all the similarity scores of a query image x_i are sorted in a descending order, and are further normalized to the range [0, 1], thus the similarity scores of using different gallery images and different measurement methods can be compared. After sorting and normalization, the set of similarity scores for data x_i are represented as $\mathbf{S}_i = \{S(i, j_1) = 1, S(i, j_2), \ldots, S(i, j_n) = 0\}$, where j_k is the index of the gallery image corresponding to the $k\mathrm{th}$ sorted similarity score. Under the assumption that, for each query image, there is one and only one gallery image from the same person, we call the largest similarity score the "matching" score and the remaining similarity scores "nonmatching" scores.

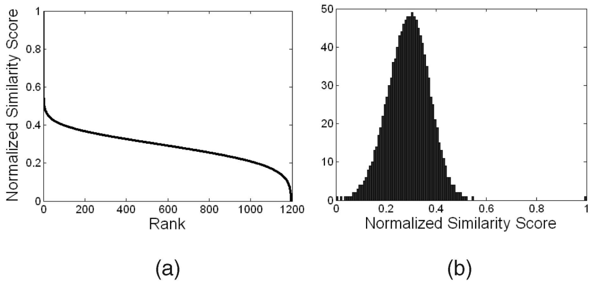


Fig. 1. Normalized similarity scores of a single image: (a) The normalized similarity scores sorted in a descending order. (b) Histogram of the normalized similarity scores.

3.2 Perfect Recognition

Given all the intrinsic factors of an FR system, including gallery images and algorithm parameters, its performance depends only upon the extrinsic factors, i.e., characteristics of query images. Empirical evaluation methods require a large set of query images with ground truth to evaluate the performance, and the resulting analysis cannot generalize to images taken under different environments. In this work, we utilize statistical analysis of similarity scores to characterize the relationship between the intrinsic factors of an FR system and its performance under different environments. To this end, the concept of "perfect recognition" is introduced.

In "perfect recognition," the gallery set G is duplicated to produce the query set Q, i.e., $Q = G = \{g_1, \ldots, g_n\}$. The "perfect recognition" uses the duplicated set for recognition, and obtains similarity scores of the each query image x_i: $\mathbf{S}_i = \{S(g_i, g_1), S(g_i, g_2), \ldots, S(g_i, g_n)\}$, $i = 1, \ldots, n$. Such similarity scores are called "Perfect Recognition Similarity Scores" (PRSS). For simplicity, the normalized and sorted PRSS are also denoted as $\{S(i, j_1), \ldots, S(i, j_n)\}$, $i = 1, \ldots, n$, where $S(i, j_k)$ corresponds to the $k\mathrm{th}$ largest perfect recognition similarity score in $\mathbf{S}_i$. Since PRSS encode information of all the intrinsic factors affecting the FR system performance, including gallery images and FR algorithms, they can be used to model the part of the FR's performance due to intrinsic factors.

3.3 Performance Metric from Similarity Scores

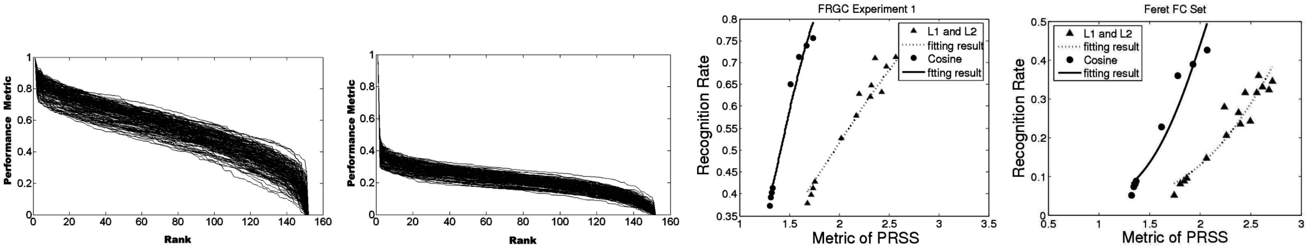
An example of PRSS is shown in Fig. 1, where the FR system uses FERET gallery data and a PCA-based recognition algorithm. It shows that the nonmatching scores (less than 1) are distinctively separated from the matching score (equal to 1). To quantitatively characterize the difference between matching and nonmatching scores for data x_i, a performance metric f_i is defined as

$$f_i = exp\left\{\frac{S(i, j_1) - \mu_i^{nm}}{\sigma_i^{nm}}\right\}, \tag{1}$$

where $S(i, j_1)$ is the matching score, and μ_i^{nm} and σ_i^{nm} are the mean and standard deviation of nonmatching scores $S(i, j_k)$, $k = 2, \ldots, n$. The mean of all f_is,

$$f = \frac{\sum_i f_i}{n},$$

is used to summarize the entire set of PRSS. f can be used to model the relationships between the performance of an FR system and its intrinsic factors. To quantitatively demonstrate this, we design the following experiments using PCA-based recognition systems. In a PCA-based FR system, commonly used similarity measurement methods include L1, L2, and Cosine. Also, each element of the



(a) (b) (c) (d)

Fig. 2. Relationship between f and recognition performance. (a) and (b) Plots of PRSS using different parameter sets. The horizontal axis is the rank of PRSS, and the vertical axis is the corresponding PRSS value: (a) $\mathrm{dim} = 40$, $\mathrm{space} = \mathrm{Euclidean}$, $\mathrm{method} = \mathrm{Cosine}$, $\gamma = 0.372$, and $\mathrm{f} = 1.1417$. (b) $\mathrm{dim} = 100$, $\mathrm{space} = \mathrm{Euclidean}$, $\mathrm{method} = \mathrm{L2}$, $\gamma = 0.732$, and $f = 1.6785$. (c) and (d) Plots of f versus γ. Each point at curves represents a result obtained under a measurement parameter set. Curves are fitting results using GLM models. (c) Results using FRGC V1.0 Experiment 1 data. (d) Results using FERET FC set.

subspace feature vector can be normalized by its corresponding standard deviation in the PCA subspace such that the similarity is measured in the Mahalanobis space [2]. Therefore, the parameters can be the dimension of subspace (dim), the measurement methods (L1, L2, or Cosine measurements), and the measurement space ("Euclidean" or "Mahalanobis"). In the following experiments, parameters of the recognition algorithm are varied and f is calculated for each set of parameters using only gallery data. The actual recognition rates (γ) using each set of parameters are computed using different query sets in FERET and FRGC, as shown in Fig. 2.

Figs. 2a and 2b show the PRSS curves of all the gallery images using two different parameter sets. Each curve represents PRSS of a gallery image. It shows that the system with better recognition accuracy has a larger difference between matching and nonmatching similarity scores; therefore, the corresponding f and γ are larger. More examples of the relationship between f and actual recognition rates are shown in Figs. 2c and 2d, where each point at the curves represents a recognition rate obtained from a specific parameter set. It is observed that the recognition rate almost monotonically increases with f. Such a relationship can be fitted with a generalized linear model (GLM) [6]. We call the generalized linear model characterizing the relationship between f and γ "performance characteristic curve."

The monotonic relationship between the actual recognition rate γ and f can be used to tune system parameters offline by selecting the parameter corresponding to the largest f, thus achieving improved system performance. However, it is also observed that the slopes of performance characteristic curves may vary, depending on the measurement methods used. For example, the performance characteristic curve obtained using the Cosine measurement method is different from that obtained using L1 or L2 measurement methods because the Cosine measurement method scales the similarity scores in a different manner. To effectively compare performance metrics of different measurement methods, we need to unify all the performance characteristic curves into one unified performance characteristic curve. To achieve this, an assumption is made that all the performance characteristic curves achieve similar mean and lower performance bound although their upper performance bound could be different. Our empirical observations indicate that this assumption is basically true for different measurement methods. Given this assumption, we can then perform the unification. Let the mean and the lowest recognition rates of $i\mathrm{th}$ performance characteristic curve be denoted as $\gamma^m(i)$ and $\gamma^l(i)$. Since all of the curves are near-linear, the average slope of $i\mathrm{th}$ curve is approximated as

$$\frac{\gamma^m(i) - \gamma^l(i)}{f^m(i) - f^l(i)},$$

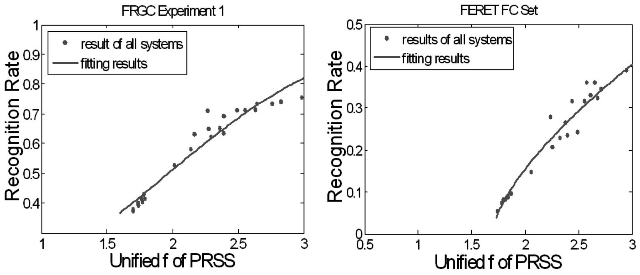
where $f^m(i)$ and $f^l(i)$ are the performance metrics corresponding to $\gamma^m(i)$ and $\gamma^l(i)$, respectively. Based on the assumption that $\gamma^m(i) \approx \gamma^m(j)$ and $\gamma^l(i) \approx \gamma^l(j)$ for the $i\mathrm{th}$ and $j\mathrm{th}$ curves, we have:

$$f(j) \approx \frac{f^m(i) - f^l(i)}{f^m(j) - f^l(j)}(f(i) - f^l(i)) + f^l(j). \tag{2}$$

Equation (2) unifies a metric $f(i)$ on the $i\mathrm{th}$ curve to its corresponding metric $f(j)$ on the $j\mathrm{th}$ curve. Fig. 3 shows unified performance characteristic curves from different measurement methods. Please note that (2) does not involve $\gamma^m(i)$ and $\gamma^l(i)$ since they have been canceled out. Since only PRSS performance metrics computed from intrinsic factors are needed in (2), it is possible to compute the unified characteristic curve without additional data. In Section 5, we demonstrate how f can be used to improve recognition performance.

4 Performance Prediction

As discussed, the closed-set recognition results can be simply categorized into two cases: correct match (CM) and incorrect match (IM). Given an individual query image, performance prediction involves predicting whether the recognition result of the query image is a correct match or an incorrect match. To perform such prediction, training data in addition to gallery images is needed. Given a query image x_i, we define the Actual Recognition Similarity Scores (ARSS) as the similarity scores $S(i,j)$ between the x_i and the gallery g_j. The performance metric of ARSS, calculated in the same way as PRSS, is also denoted as f_i for x_i. In



(a) (b)

Fig. 3. Relationship between unified f and actual recognition rates of different systems: (a) Systems using FRGC V1.0 Experiment 1 data set. (b) Systems using FERET FC data set.

TABLE 1
Performance Prediction Algorithm

- Training stage: given an FR method, its parameters, the gallery data, and some training data
  - Calculate the perfect recognition similarity scores (PRSS) using the equations given in Section III-B
  - Adjust the parameters of the FR method to maximize the performance metric *f* as discussed in Section III-C
  - Perform face recognition using the training images and obtain their actual recognition similarity scores (ARSS). Label the correctly recognized images as positive data while the rest as negative data
  - Calculate the difference vectors using ARSS and the corresponding PRSS, as in Eqn. (4)
  - Train a performance predictor (e.g., a SVM) using the extracted difference vectors and their labels
- Prediction stage: given a query image, predict its recognition result:
  - For the given query image, calculate its ARSS
  - Calculate the difference vector using the ARSS of the query image and the corresponding PRSS, and input the difference vector to the trained SVM to predict the recognition

fact, PRSS can be seen as a special case of ARSS since the query set in perfect recognition is the duplication of the gallery set.

In our method, the differences between the ARSS and its corresponding PRSS are used as features for individual recognition result prediction. Mathematically, the similarity score difference vector $\mathbf{D}_{x_i}^1$ of the rank 1 recognition is defined as:

$$\begin{aligned} d_k^1(x_i) &= S(i, j_k) - S'(j_1, j_k) \\ \mathbf{D}_{x_i}^1 &= \big(d_1^1(x_i)w_1, \dots, d_n^1(x_i)w_n\big), \end{aligned} \tag{3}$$

where $S(i, j_k)$ is the kth score of ARSS. To prevent confusion between PRSS and ARSS, we use $S'(j_1, j_k)$ to denote the kth score of PRSS corresponding to a rank 1 recognition result, where j_k is the index of the gallery image corresponding to the kth sorted similarity score and n is the total number of gallery images. The difference of kth similarity score $d_k^1(x_i)$ is smoothed by a weight w_k to emphasize the scores of the first several ranks since they are more important for recognition. In our work, w_k is defined as

$$w_k = exp\left\{\frac{-(k-1)^2}{2\sigma_r^2}\right\},$$

where σ_r is set as 20.

Besides $\mathbf{D}_{x_i}^1$, we also include the differences at more recognition ranks to expand a feature vector for performance prediction. At the rank m, the difference vector is $\mathbf{D}_{x_i}^m = \{d_1^m(x_i)w_1, \dots, d_n^m(x_i)w_n\}$, where $d_k^m(x_i) = S(i, j_k) - S'(j_m, j_k)$. $S'(j_m, j_k)$ is the corresponding perfect recognition similarity score at the rank m. The extracted feature vector $\mathbf{V}$ is then defined as in (4):

$$\begin{aligned} \mathbf{V} &= (d_1^1(x_i)w_1, \dots, d_K^1(x_i)w_K, d_1^2(x_i)w_1, \dots, \\ &\quad d_K^2(x_i)w_K, \dots, d_1^m(x_i)w_1, \dots, d_K^m(x_i)w_K). \end{aligned} \tag{4}$$

In the feature vector $\mathbf{V}$, the differences between ARSS and PRSS at the first M ranks are used, and within each rank, only the difference of the first K scores are used. So, totally, there are $M * K$ elements in the feature vector. Given such defined features, a Support Vector Machine (SVM) is trained using collected training data to predict the face recognition result of the individual query image as either a correct match or an incorrect match.

The algorithm of performance prediction is summarized in Table 1. It needs to be noted that when predicting incorrectly, the performance predictor could commit two kinds of mistakes: false positive and false negative. False positive is defined as an incorrect prediction that classifies an incorrect match as a correct match. The false negative, on the other hand, is defined as an incorrect prediction that classifies a correct match as an incorrect match.

5 Experiments

Two face databases, FERET [8] and FRGC V1.0 [7], are used in our experiments. FERET uses a fixed gallery image set and different query image sets to study recognition performance under changes of facial expressions (FB), illumination (FC), and age (Dup1). In FRGC Experiment 1, both gallery and query images are taken under controlled environments while query images in Experiment 4 are taken under much less controlled conditions of lighting, pose angle, and facial expression. We apply a PCA-based recognition method, in which each face is normalized to the size of 45×30 pixels. The pixel intensity is further normalized by histogram equalization. In the following experiments, we show, respectively, the results of offline parameter tuning using performance modeling, recognition result prediction, and online face alignment refinement.

5.1 Offline Parameter Selection

For the first experiment, the performance characteristic curves for different measurement methods are first unified into one curve by the linear correction describ ed in Section 3.3, and the parameter set corresponding to the largest unified f is selected as the optimal parameter set. As a result, the selected parameter set for FERET is [200, *Cosine*, *Mahalanobis*], which means that the system uses 200 PCA coefficients, the Cosine measurement method, and the Mahalanobis space. The parameter set selected for FRGC V1.0 Experiments 1 and 4 is [120, *Cosine*, *Mahalanobis*]. To verify the optimality of the selected parameters, we exhaustively test the recognition rates of all possible parameters, and compare them with the recognition rates of selected parameter sets, as summarized in Table 2, which also shows the parameters corresponding to maximal actual recognition rate. From the table, we can observe that different query sets need different parameters to achieve the best recognition rate. However, the offline selected parameters can achieve near-optimal accuracy for different query sets acquired under different environments. This is even true for the data sets with large accuracy ranges, such as FERET FC. This experiment demonstrates the optimality of the selected parameter set only using gallery images, as well as its capability of generalization to different query sets.

TABLE 2
Summary of Offline Parameter Selection

Query Set	Accuracy of selected parameter	Offline selected parameters	Accuracy range	Parameters of maximal actual accuracy
FERET FB	80.0%	[200, Cos., Maha.]	[70.2% , 82.0%]	[160, L1, Eucli.]
FERET FC	49.4%	[200, Cos., Maha.]	[5.2% , 50.7%]	[180, Cos., Maha.]
FERET Dup1	34.7%	[200, Cos., Maha.]	[22.6% , 38.8%]	[100, Cos., Maha.]
FRGC Exp. 1	75.1%	[120, Cos., Maha.]	[32.7% , 75.5%]	[100, Cos., Maha.]
FRGC Exp. 4	23.4%	[120, Cos., Maha.]	[4.9% , 27.0%]	[100, Cos., Maha.]

TABLE 3
Performance Prediction Accuracy with Intraset and Interset Cross-Validation on FERET and FRGC

Data Set	Prediction accuracy ([false positive rate, false negative rate])	
	Intra-set validation	Inter-set validation
FERET FB	[10.6%, 15.6%]	[11.7%, 20.8%]
FERET FC	[10.8%, 21.2%]	[12.2%, 21.0%]
FERET Dup1	[9.6%, 31.1%]	[8.4%, 36.3%]
FRGC Exp. 1	[9.0%, 25.7%]	[10.5%, 31.1%]
FRGC Exp. 4	[13.0%, 56.3%]	[23.7%, 50.0%]

5.2 Recognition Result Prediction

The following experiments demonstrate how our method predicts individual recognition results, and improves recognition performance. The difference values between ARSS and PRSS are extracted to train a performance predictor (e.g., SVM) to classify individual recognition results as either a correct match or an incorrect match. We validate the performance predictor through a cross-validation, in which 50 percent data is used for training, and the remaining 50 percent is used for validation. To assess the generalization capability of the performance predictor, two types of cross-validation methods, i.e., intraset and interset validation methods, are applied. In the intraset validation method, all the query sets are merged into one query set. The training data is uniformly sampled from the merged set, and then the predictor is validated on the remaining data of the merged set. In the interset validation method, the predictor is trained using data selected from only some of the sets, and is validated on the other sets. For example, when using FERET data sets, the predictor is trained with data from the FB (or FC and Dup1) set, and validated on the FC and Dup1 (or FB) sets. When using FRGC V1.0 data sets, the predictor is trained on the Experiment 1 (or the Experiment 4) set, and validated on the Experiment 4 (or the Experiment 1) set. The intraset validation method assumes that the testing data is obtained from the same environment as the training data while the interset validation method simulates the situation where the training and testing data are acquired from different environments.

The false positive rates and false negative rates of the performance predictor are summarized in Table 3 for both intraset and interset validation. The overall error rate of the performance predictor is between 15 percent and 25 percent for FERET sets and FRGC Experiment 1 while FRGC Experiment 4 shows worse accuracy. Table 3 shows that the accuracy of the interset validation is only slightly worse than the accuracy of the intraset validation, which demonstrates that this prediction method is not constrained in a specific environment, but can be applied in various environments after the predictor has been trained.

Since the performance predictor can identify the query images that are most likely mismatched by an FR system, such actions can be taken with respect to these images as reacquiring them or fusing multiple recognition modalities. For this study, only the images predicted to result in correct matches will be preserved for recognition; therefore, improving the overall system performance. Experimental recognition results with and without applying performance prediction are compared in Fig. 4, where P is the percentage of preserved "good" query images, i.e., the images that are predicted to be correctly matched by the FR system. The threshold corresponding to each P is obtained from training sets. Fig. 4 shows the recognition results when preserving all images (i.e., the curves corresponding to "All"), and the results when only preserving a certain percentage of "good" query images (i.e., the curves corresponding to different P values in Fig. 4). It is shown that the recognition accuracy is largely improved with the use of performance prediction.

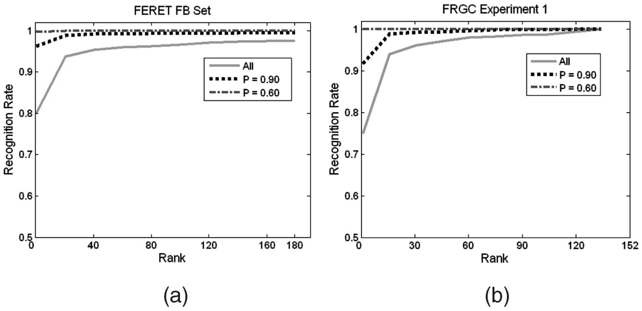


Fig. 4. CMC curves of face recognition with and without performance prediction: (a) FERET FB set. (b) FRGC V1.0 Experiment 4 set.

TABLE 4
Summary of Rank 1 Recognition Rates with and without Performance Prediction

Data Set	All	P = 90%	P = 60%
FERET FB	γ_0 = 80.0%	γ' = 96.2% P' = 74.8%	γ' = 99.7% P' = 48.1%
FERET FC	γ_0 = 49.3%	γ' = 93.7% P' = 47.4%	γ' = 96.4% P' = 30.7%
FERET Dup1	γ_0 = 34.7%	γ' = 82.9% P' = 37.7%	γ' = 93.1% P' = 22.4%
FRGC Exp. 1	γ_0 = 75.0%	γ' = 91.8% P' = 73.5%	γ' = 100.0% P' = 45.0%
FRGC Exp. 4	γ_0 = 23.9%	γ' = 57.2% P' = 37.6%	γ' = 64.3% P' = 22.3%

The experimental results are further summarized in Table 4. In the table, γ_0 denotes the original recognition rate at rank 1 using all the available data without performance prediction. γ' is the recognition rate when using only the data predicted to result in correct matches. The percentage of total data used for recognition, P', can thus be calculated as $P' = \frac{\gamma_0 * P}{\gamma'}$. It shows that performance is greatly improved by applying performance prediction. For example, the recognition rate of FERET FB set is increased to 96.2 percent from 80 percent when about 90 percent of query images that can be correctly matched are preserved (which also means that 74.8 percent of total query images are used for recognition). For the query sets that usually have low recognition rates, such as FERET FC, FERET Dup1, and FRGC Experiment 4, the performance improvement is also obvious. The error rate goes down to near zero when preserving fewer query images. But, the price paid is that much query data that can be correctly recognized is also discarded. However, the P value can be adjusted to achieve the desired trade-off between the recognition rate and the amount of data loss. Another strategy for performance improvement is to fuse multiple FR systems based on performance prediction. The images predicted to be incorrect matches for one system can be picked up by another system that might be able to recognize them correctly. The performance prediction based on multiple system fusion represents a natural extension of this work.

5.3 Online Adjusting Face Alignment

All of the above experiments use manually marked eyes for face alignment. However, most real-world applications require automatic eye localization. Although some accurate eye localization methods have been developed, there still exist localization errors (or at least inconsistencies), such that results using automatic eye localizations are consistently lower than those using manually marked eyes [8], [12]. Since the performance metric has a monotonic relationship with the actual recognition rate, it can be used to guide

TABLE 5
Summary of Rank 1 Recognition Rate with Adjusted Eyes

Data Set	Manual eyes	Adjusted on manual eyes	Automatic eyes	Adjusted on automatic eyes
FERET FB	79.8%	85.1%	74.8%	84.8%
FERET FC	49.3%	59.8%	43.3%	57.2%
FERET Dup1	34.8%	44.6%	30.6%	42.9%

the adjustment of the initial eye positions, whether automatically or manually marked, for better recognition. Assuming that there are k-pairs of eye position candidates, the performance metric f_i is calculated for the query face image which is aligned using the ith eye-pair candidate. The eye-pair candidate producing the highest f_i is selected for face alignment prior to providing recognition results.

In the following experiments, for each eye, nine candidate eye positions (the initial eye position plus its eight neighbors) are searched. Table 5 compares the recognition rates of using the original eye and adjusted eye positions, where the automatic eye detection method described in [12] is used. It is observed that the face recognitions of using adjusted eyes outperform those of using initial eye position estimates. Also, the eye locations adjusted from automatic eye locations provide better recognition rates than those of manually marked eyes. These experiments demonstrate that the manually marked eyes do not always provide the best face alignment for recognition, and that face recognition performance can be improved with the use of our presented face alignment adjustment method.

6 Summary

In this paper, we present our work on performance modeling and prediction of face recognition systems based on the analysis of similarity scores. We introduce the concept of "perfect recognition" and derive a performance metric based on the perfect recognition similarly scores to model the relationships between the intrinsic factors of an FR system and its performance. The performance metric is subsequently used to tune the FR system parameters offline. The performance metric calculated from query images can be used to adjust face alignment online for improved recognition accuracy. By comparing the perfect recognition similarity scores with the actual recognition similarity scores, we present a method to predict the recognition results of individual query data. Experimental results demonstrate that the methods provide various ways to improve the performance of face recognition systems. While our method is not specifically designed for the PCA method, its applicability to other FR methods need be studied. One future work will be further validation of our method with additional data sets and with different recognition methodologies. Another future work is to extend the proposed methods to databases that contain multiple gallery and query images for each face. We also plan to apply the methods to other biometric systems, such as fingerprint and iris recognition.

Acknowledgments

The research described in this paper is supported in part by a grant (N41756-03-C-4028) to Rensselaer Polytechnic Institute from the Task Support Working Group (TSWG) of the United States.

References

[1] Anon, "Face Recognition Vendor Test," 2006.

[2] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, "The CSU Face Identification Evaluation System User's Guide: Version 5.0," Computer Science Dept., Colorado State Univ., May 2003.

[3] G. Givens, J.R. Beveridge, B.A. Draper, P. Grother, and P.J. Phillips, "How Features of the Human Face Affect Recognition: A Statistical Comparison of Three Face Recognition Algorithms," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 2, 2004.

[4] A.Y. Johnson, J. Sun, and A.F. Bobick, "Using Similarity Scores from a Small Gallery to Estimate Recognition Performance for Larger Galleries," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures,* pp. 100-103, 2003.

[5] W. Li, X. Gao, and T.E. Boult, "Predicting Biometric System Failure," *Proc. IEEE Int'l Conf. Computational Intelligence for Homeland Security and Personal Safety,* pp. 57-64, 2005.

[6] P. McCullagh and J.A. Nelder, *Generalized Linear Models,* second ed. Chapman & Hall/CRC, 1989.

[7] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 947-954, 2005.

[8] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1090-1104, Oct. 2000.

[9] N.A. Schmid, B. Cukic, M. Ketkar, and H. Singh, "Performance Analysis of Iris Based Identification System at the Matching Score Level," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 2, pp. 93-96, 2005.

[10] N.A. Schmid and J.A. O'Sullivan, "Performance Prediction Methodology for Biometric Systems Using a Large Deviations Approach," *IEEE Trans. Signal Processing,* vol. 52, no. 10, pp. 3036-3045, 2004.

[11] E. Tabassi, C.L. Wilson, and C.L. Watson, "Fingerprint Image Quality," Technical Report NISTIR 7151, Nat'l Inst. of Standards and Technology, 2004.

[12] P. Wang, M.B. Green, Q. Ji, and J. Wayman, "Automatic Eye Detection and its Validation," *Proc. IEEE Workshop Face Recognition Grand Challenge Experiments,* 2005.

[13] R. Wang and B. Bhanu, "Learning Models for Predicting Recognition Performance," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[14] *Biometric Systems Technology, Design and Performance Evaluation,* J. Wayman, A. Jain, D. Maltoni, and D. Maio, eds. Springer, 2005.

[15] W. Zhao, R. Chellappa, P.J. Philips, and A. Rosenfeld, "Face Recognition: A Literature Survery," *ACM Computing Survey,* vol. 35, pp. 399-458, 2003.