To appear in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York City, June, 2006

Performance Modeling and Prediction of Face Recognition Systems

Peng Wang
Section of Biomedical Image Analysis
Department of Radiology
University of Pennsylvania
wpeng@ieee.org

Qiang Ji
Department of Electrical, Computer
and Systems Engineering
Rensselaer Polytechnic Institute
qji@ecse.rpi.edu

Abstract

It is a challenging task to accurately model the performance of a face recognition system, and to predict its individual recognition results under various environments. This paper presents generic methods to model and predict the face recognition performance based on analysis of similarity measurement. We first introduce a concept of "perfect recognition", which only depends on the intrinsic structure of a recognition system. A metric extracted from perfect recognition similarity scores (PRSS) allows modeling the face recognition performance without empirical testing. This paper also presents an EM algorithm to predict the recognition rate of a query set. Furthermore, features are extracted from similarity scores to predict recognition results of individual queries. The presented methods can select algorithm parameters offline, predict recognition performance online, and adjust face alignment online for better recognition. The experimental results show that the performance of recognition systems can be greatly improved using presented methods.

1. Introduction

How to evaluate the performance of an algorithm has been studied for many years in the computer vision community. Especially with the intensive research and application of biometric systems, the performance modeling and prediction receives a lot of attention since it involves the great concerns of security and privacy [14]. Face recognition is one of the most popular biometric systems. However, current face recognition systems always have errors, and their performance varies under different environments. This paper presents generic methods to model and predict the system performance based on analysis of similarity scores.

In our work, the "performance" of a recognition system means its accuracy in correctly matching face images. We do not consider other aspects of performance, such as speed, cost, availability and maintainability. We also use "failure recognition" to refer to the misclassification of a given input image. We first introduce a concept of "perfect recognition" and a statistical analysis of similarity scores from "perfect recognition". Such analysis only depends on the intrinsic structure of a recognition system, and provides a metric that can characterize the recognition performance under different environments without empirical testing. The performance metric is further assumed to be Gaussian distributions under the cases of success and failure recognition, and is used to predict the recognition accuracy of a query set via an EM algorithm. To predict individual recognition results, we extract features by comparing actual recognition results with their corresponding perfect recognition results, and train a performance predictor with the extracted features.

Our methods can select optimal or near-optimal algorithm parameters offline without using additional training data, predict face recognition result online, and adjust the face alignment online for better recognition. Experiment results demonstrate that our methods can significantly improve the performance of a face recognition system. In this paper, our methods are validated on PCA based face recognition systems [7]. However, the methods can be easily generalized to any other recognition systems using similarity scores.

The paper is organized as following. Related work is reviewed in Section 2. In Section 3, we introduce the modeling method of a face recognition system. The face recognition prediction methods are introduced in Section 4. Experiments results are presented in Section 5. We conclude in Section 6.

2. Related Work

Sampling methods are the most popular methods to empirically evaluate a recognition system. In these methods, the training and testing is conducted separately on different sets which are randomly sampled or specially designed. The typical random sampling methods include cross validation method and Bootstrap method [3]. To study the system per-

formance under specific environments, special experiments are designed, such as the face recognition vendor test sets of FERET [9] and FRGC [8]. Although such specifically designed experiments can directly assess the performance of a system under typical circumstance, they cannot perform online performance prediction, and they need to acquire training data for different environments.

There are already some work on performance modeling and prediction of biometric systems, such as fingerprint recognition[11], iris recognition[10], and face recognition [9, 8, 4]. In [11], the quality of a fingerprint image is defined as the normalized distance between matching and non-matching similarity scores. A 11-dimensional feature vector is extracted from image analysis algorithms to identify the existence of feature points, e.g., minutia, and outliers. Then a Neural Network is trained using the feature vectors to predict the image quality. The experiments show that the images with higher predicted quality will achieve better recognition accuracy. The feature extraction method for fingerprint image quality prediction cannot be directly used to face recognition since most face recognition methods use holistic appearance instead of feature points.

Schmid et. al. provide a probabilistic estimation of lower bound of Iris recognition algorithms based on analysis of the hamming distance between query and gallery iris images [10]. The distance is assumed to be a single Gaussian distribution under both genuine and imposter hypothesis, and the likelihood ratio is used to identify the pattern best matching the query iris. With learned parameters, the method estimates ROC of iris recognition by applying the Chernoff bound theory and the Large Deviation theory. However, both of the lower bounds only provide approximate error orders. They cannot be used to predict either an individual recognition result, or the performance of systems which do not use likelihood ratio method for recognition.

Givens and Beveridge et. al. apply statistical tools to analyze how the human face features, such as age, race, gender, skin, glasses, and expression, affect face recognition accuracy [4]. A generalized linear model is built to regress the relationship between the affecting factors and recognition accuracy. The analysis of variance (ANOVA) is conducted to study how significantly each factor affects the recognition accuracy. To model the performance, the statistical model needs to explicitly identify each affecting factor, which is an extremely difficult task in practical implementation. Also, such factors cannot totally model the face recognition performance. It is shown that about 34% of variance cannot be explained by the generalized liner model.

Some other work uses similarity scores to predict system performance. Li et. al. cluster the similarity scores into different sets, and then use the distance among the sets as features which are selected and combined with AdaBoost to detect failure recognition [6]. A problem is that AdaBoost usually needs a large pool of features and many training samples, so over 10,000 samples are used in [6]. Such a large number of training samples are usually difficult to collect for practical systems. The similarity scores are also used to predict CMC curves with a small set of gallery data [5, 13]. In their methods, the rank k recognition results are modeled using parametric models. With model parameters estimated from a small gallery set, their methods can predict CMC when more gallery data are applied. Their method can only work well for the case that gallery data and query data are under the same condition, and cannot predict individual recognition results online either.

3. Performance Modeling

In this section, we analyze face recognition systems, and then introduce a concept of "perfect recognition". By analyzing the similarity scores output from perfect recognition, we present a metric that can model system performance without using additional training data.

3.1. Model of Face Recognition Systems

There is no shortage of algorithmic approaches to face recognition [15]. The function of a face recognition system is to map a query (also called probe) image to a label that represents its identification. Usually a face recognition system consists of at least two *intrinsic* components, i.e., a set of *gallery* images and a face recognition *algorithm*. The gallery set, denoted as $G = \{g_1, g_2, ..., g_n\}$, includes n exemplars of known identification to be used for the comparison with query data. A face recognition algorithm maps query data to a feature space, measures the similarity between query data and gallery data, and outputs the identification of query data . For a query image, a recognition algorithm usually outputs n similarity scores corresponding to n gallery images respectively. For rank k recognition, the system outputs labels of the gallery images corresponding to the k largest similarity scores.

The similarity score plays an important role in face recognition because it relates query images with both the recognition algorithm and each gallery image available in the 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. There are many type of similarity measurements [1], and larger similarity scores mean better recognition. In our work, all the similarity scores are sorted in a descending order, and are further normalized to the range $[0, 1]$. So the set of similarity scores of data x_i are represented as $S_i = \{S(i, j_1) = 1, S(i, j_2), ..., S(i, j_n) = 0\}$, where j_k indicates the label of gallery data corresponding to the k-th sorted similarity score. The largest similarity score is called "matching" score since it represents the best matching between query and gallery while the remaining

similarity scores are called "non-matching" scores.

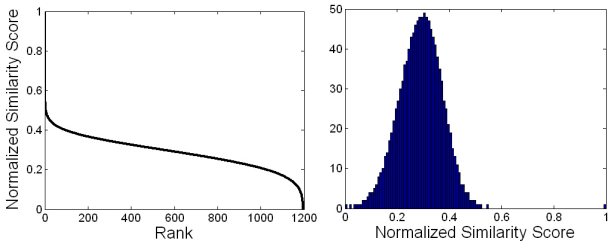
3.2. Perfect Recognition

When all the intrinsic components of a recognition system are given, we believe that its performance is actually fixed, but unknown to users. To empirically measurement its performance, we need large sets of query data with ground truth. The resulting performance analysis will apply only to the particularly query images and will not be extendable to images, even from the same people, if taken in unknown environments. In this work, we utilize statistical analysis of similarity scores to discover the relationship between the intrinsic structure of a recognition system and its performance under various environments. To systematically analyze the intrinsic components of a face recognition system, we introduce the concept of "perfect recognition".

The definition of "perfect recognition" is simple and straightforward. A query set Q is duplicated from the gallery set, i.e., $Q = G = \{g_1, ..., g_n\}$. The "perfect recognition" uses the duplicated set as the query set for recognition, and obtains the similarity scores: $S_i = \{s(g_i, g_1), s(g_i, g_2), ..., s(g_i, g_n)\}, i = 1, ..., n$. We call such similarity scores as "Perfect Recognition Similarity Scores" (PRSS).

The defined perfect recognition has two characteristics. First, it can achieve 100% recognition accuracy. Second, the perfect recognition encodes information of all the components in a recognition system: it uses all the gallery data, and the similarity scores encode both the recognition algorithm and its parameters. So by analyzing PRSS, it is possible to model the performance of a recognition system without using additional query images.

3.3. Performance Metric from Similarity Scores



(a) (b)

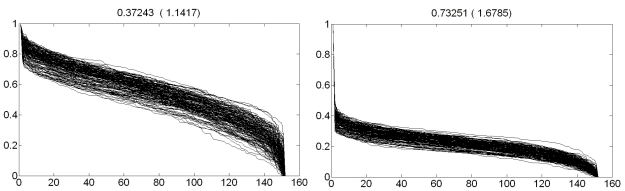
Figure 1. Normalized similarity scores of a single query data. (a) normalized similarity scores sorted in a descending order. (b) histogram of normalized similarity scores

An example of PRSS is shown in Figure 1, where the system uses FERET gallery data and PCA based recognition algorithm. It shows that non-matching scores (less than 1) are much smaller than the matching score (equal to 1), and non-matching scores can be modeled using a single Gaussian model. To quantitatively characterize the difference between matching and non-matching scores for data x_i, a measurement q_i is calculated as Equation (1).

$$q_i = \frac{S(i, j_1) - \mu_i^{nm}}{\sigma_i^{nm}} \quad (1)$$

where μ_i^{nm} and σ_i^{nm} are the mean and standard deviation of non-matching scores $S(i, j_k)$, $k = 2, ..., n$. Such defined q_i has also been used to represent image quality in fingerprint recognition [11]. Based on the distance between matching and non-matching scores, we define a metric from similarity scores as $f_i = exp\{\frac{q_i}{\lambda}\}$, where λ is a constant to scale the performance metric. It is set as 20 in this paper. However, our method is insensitive to the value of λ. For a recognition system, the mean of all f_i's, i.e. $f = \frac{\sum_i f_i}{n}$, is used to describe the whole set of PRSS.

Intuitively, a system with good performance should also be able to well discriminate the gallery data, regardless of query data. Since PRSS represent the similarity measurement among gallery data, the metric f extracted from PRSS is able to model the system performance. To quantitatively demonstrate the intuition, the following experiments are designed. Firstly, parameters of a recognition algorithm are changed to get different recognition systems. In a PCA based recognition system, the parameters can be the dimension of subspace, the measurement methods (L1, L2 or Cosine measurements), and the measurement space ("Euclidean" or "Mahalanobis" space) [1]. Then f of each system is calculated, and the actual recognition accuracy is also validated with a query set.

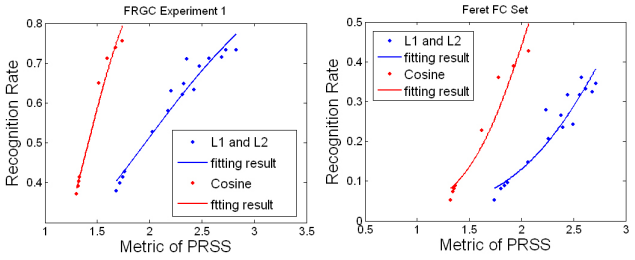


(a) (b)

Figure 2. PRSS of two systems using FRGC V1.0 Experiment 1 data set. In each graph, the horizontal axis is the rank, and the vertical axis is the corresponding PRSS values. (a) dim = 40, space = Euclidean, method = Cosine. The recognition rate is 37.2%, $f = 1.1417$. (b) dim = 100, space = Euclidean, method = L2. The recognition rate is 73.3%. $f = 1.6785$

Figure 2 shows PRSS of two PCA based recognition systems with different parameters. As observed from the figure, the system with better performance has larger difference between matching similarity scores and non-matching similarity scores, so its f is larger. More relationship between f and actual recognition rates under different query sets is shown in Figure 3, from which we observe that the recognition rate almost monotonically increases with f. Such relationship can be fitted with a generalized linear

model(GLM). The generalized linear model that characterizes the relationship between the recognition accuracy and f is called "performance characteristic curve" in this paper.



(a) (b)

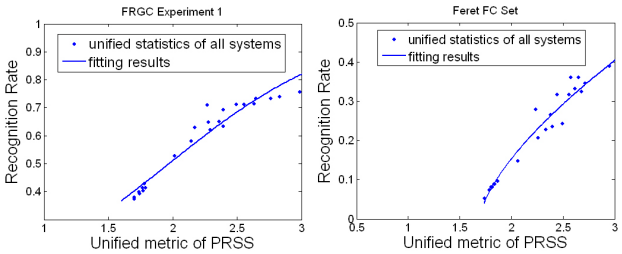
Figure 3. Relationship between f and actual recognition performance. Blue and red points represent the results of using different measurement methods. The lines are the fitting results using GLM model. (a) systems using FRGC V1.0 experiment 1 data. (b) systems using FERET FC set

It is also observed from the graphs that the performance characteristic curve of using Cosine measurement method is different from that of using L1 and L2 measurement methods since Cosine measurement method scales the similarity in a different way from other methods. To evaluate the systems using various measurement methods, a linear correction algorithm is presented to unify all the performance characteristics curves into one curve. We assume that all the performance characteristic curves achieve the similar mean and lower bound of performance although their performance upper bound could be different. The mean and lower recognition rates of i-th performance characteristic curve are denoted as $P^m(i)$ and $P^d(i)$. Since all the curves are near linear, the average gradient of i-th curve is approximated as $\frac{P^m(i)-P^d(i)}{f^m(i)-f^d(i)}$ where $f^m(i)$ and $f^d(i)$ are the metric corresponding to $P(i)^m$ and $P(i)^d$ respectively. Based on the assumption that $f^m(i) \approx f^m(j)$ and $f^d(i) \approx f^d(j)$ for i-th and j-th curves, we have

$$f(j) \approx \frac{f^m(i) - f^d(i)}{f^m(j) - f^d(j)}(f(i) - f^d(i)) + f^d(j) \qquad (2)$$

Equation (2) unifies a metric $f(i)$ on the i-th curve to the j-th curve, and only PRSS performance metrics are needed. The actual recognition rate is eliminated by assuming similar mean and low performance bounds. This assumption, of course, is only very approximate and may introduce some errors during parameter selection. However, it allows us to avoid using training data for offline parameter tuning and can be generalized to other measurement methods due to its simplicity.

Figure 4 shows the unified performance characteristic curve. The monotonic relationship between unified f and system performance can be used to select system parameters offline to achieve optimal or near-optimal performance. More experiment results are presented in Section 5.



(a) (b)

Figure 4. Relationship between unified metrics and actual recognition rates of different systems. (a): systems using FRGC V1.0 experiment 1 data set. (b) systems using FERET FC data set

4. Performance Prediction

Actual face recognition results are categorized into two cases: success recognition (SR) and failure recognition (FR). A variable $R(x)$ is introduced to indicate the recognition accuracy of query x. $R(x)$ is defined as Equation (3), where $\Phi(x)$ is the label output from a face recognition system, and $I(x)$ is the true label of query x.

$$\begin{aligned} R(x) = 1 \quad &: \quad I(x) = \Phi(x) \\ -1 \quad &: \quad I(x) \neq \Phi(x) \end{aligned} \qquad (3)$$

$R(x) = 1$ indicates a success recognition, and $R(x) = -1$ otherwise.

In this section, we study how to predict recognition accuracy of a query set or an individual query data. The performance prediction methods are also based on the analysis of actual similarity scores. Assuming the metric f_i of actual recognition similarity scores as a single Gaussian under SR and FR, an EM algorithm is applied to estimate recognition rates of a query set at different ranks, i.e., to predict its cumulative matching curve (CMC). Then, our method predicts an individual query to be a success or failure recognition. For this purpose, a predictor (e.g., a Support Vector Machine) is trained with features extracted from similarity scores.

4.1. Predicting CMC of a Query Set

Actual Recognition Similarity Scores (ARSS), which are the similarity scores between query data and gallery data, are used to predict the actual recognition performance given a query set. The performance metric of ARSS is defined in the same way as PRSS, and is also denoted as f_i for query data x_i. In fact, PRSS can be seen as a special case of ARSS since the query set in perfect recognition is the duplication of gallery data.

We model f_i using a single Gaussian distribution under SR or FR , i.e., $P(f_i|SR) = N(f_i; \mu_s, \sigma_s)$ and $P(f_i|FR) = N(f_i; \mu_f, \sigma_f)$. Given a query set, the distribution of f_i is actually a mixture of Gaussian, as Equation (4), where its two components correspond to success and

failure recognition respectively.

$$P(f_i) = \pi_s P(f_i|SR) + \pi_f P(f_i|FR) \quad (4)$$

In (4), π_s and π_f are the percentages of success and failure recognition, and $\pi_s + \pi_f = 1$. Therefore π_s is actually the recognition rate of rank 1. Given a data set, we can apply an EM algorithm to estimate the model parameters, therefore to predict the recognition rate.

The previously defined f_i only characterizes recognition quality of rank 1 . Following the same principle, the metric f_i^k are defined to characterize the recognition quality of rank k, as Equation (5):

$$\begin{aligned} q_i^k &= \frac{S(i,j_1) - \mu_i^{nm}(k)}{\sigma_i^{nm}(k)} \\ f_i^k &= exp\{\frac{q_i^k}{\lambda}\} \end{aligned} \quad (5)$$

where $\mu_i^{nm}(k)$ and $\sigma_i^{nm}(k)$ are the mean and standard deviation of rank k non-matching scores $\{S(i,j_{k+1}), ..., S(i,j_N)\}$. Compared with the previously defined f_i for rank 1 recognition, the non-matching scores in Equation (5) are limited to the scores after k-th rank. The reason behind is that for rank k recognition, the first k similarity scores are all matching scores, and the maximum of matching scores is $S(i,j_1)$.

To estimate the parameters in the mixture of Gaussian model, an EM learning algorithm is applied. It needs an initialization of mean and standard deviation at rank 1, which can be learned from a small set of data. We assume that the parameters of the mixture model $P(f_i^k)$ smoothly change with increasing rank k, so the estimation results from rank k can be used as the initialization of rank $k+1$. Also recognition rate of rank $k+1$ is not less than the recognition rate of rank k, which can help smooth the prediction result of CMC. The EM algorithm to predict CMC is summarized in Table 1.

4.2. Predicting Individual Recognition

To predict each individual recognition result as success or failure recognition, the relationship of ARSS and PRSS are further studied. If an actual recognition is closer to its corresponding perfect recognition, it is more likely to get a success recognition result. The difference between an actual recognition and its corresponding perfect recognition can be quantitatively represented by the difference between ARSS and PRSS. Mathematically, the similarity score difference vector D_x^1 of rank 1 is defined as:

$$\begin{aligned} d_k^1(x) &= s(x,j_k) - s(j_1,j_k) \\ D_x^1 &= \{d_1^1(x)w_1, ..., d_n^1(x)w_n\} \end{aligned} \quad (6)$$

where $s(x,j_k)$ is k-th score of ARSS, and $s(j_1,j_k)$ is the k-th score of PRSS corresponding to rank 1 recognition result.

- Given a query set. Initialize model with parameters learned from a small set of query data. $C_0 = 0$;
- For $k = 1...T$, estimate the recognition rate of rank k as the follows.
  1. Initialize the mixture model $P(f_i^k) = \pi_s^k N(f_i^k; \mu_s^k, \sigma_s^k) + \pi_f^k N(f_i^k; \mu_f^k, \sigma_f^k)$ with the parameters of rank $k-1$, i.e. $\mu^k = \mu^{k-1}$, $\sigma^k = \sigma^{k-1}$, and $\pi^k = \pi^{k-1}$.
  2. For each data x_i in the query set, calculate f_i^k as Equation (5). The set of f_i^k is applied to learn the parameters of $P(f_i^k)$ using the standard EM algorithm.
  3. The weight corresponding to the component with larger mean in the mixture model is π_s, and $C_k = max(\pi_s, C_{k-1})$.
- Output CMC curve, $C_k, k = 1, ..., T$.

Table 1. Algorithm of predicting CMC curve

The difference of k-th similarity score $d_k^1(x)$ is smoothed by a weight w_k to emphasize the scores of first several ranks since they are more important for recognition. In this paper, w_k is defined as $w_k = exp\{\frac{-(k-1)^2}{2\sigma_r^2}\}$ where σ_r is set as 20.

Based on our experiments, the first difference $d_1^1(x)$ can separate about 50% of success recognition results from the failure recognition results. However, it is still not enough to predict all the success and failure recognition cases. So the difference vectors of more ranks are included as features. For rank m, the difference vector is $D_x^m = \{d_1^m(x)w_1, ..., d_n^m(x)w_n\}$ where $d_k^m(x) = s(x,j_k) - s(j_m,j_k)$. The extracted feature vectors V_x is as:

$$\begin{aligned} V_x = \{ & d_1^1(x)w_1, ..., d_K^1(x)w_K, \\ & ..., \\ & d_1^M(x)w_1, ..., d_K^M(x)w_K\} \end{aligned}$$

where the difference between ARSS and FRSS of the first M ranks are used. For each rank, only difference of the first K scores are used. Totally there are $M * K$ elements in the feature vector (some elements may be redundant because $s(x,j_1) = s(j_1,j_1) = 1$ due to normalization).

A Support Vector Machine (SVM) [2] is trained with extracted features to predict face recognition results. Usually, a SVM outputs a continuous value $dis(V_x)$, which represents a distance of input data V_x to the class boundary in a high dimensional feature space. By thresholding the con-

tinuous output $dis(V_x)$, the SVM gives the prediction results $R'(x)$ as failure recognition ($R'(x) = -1$) or success recognition ($R'(x) = 1$), as in Equation (7).

$$\begin{aligned} R'(x) &= 1 : dis(V_x) >= dis_h \\ &= -1 : dis(V_x) < dis_h \end{aligned} \quad (7)$$

$R'(x)$ is the predicted value of $R(x)$ in Equation (3). The performance predictor also has misclassification error itself. By adjusting the threshold dis_h, the predictor shows different false alarm rate and positive error rate. The false alarm of performance predictor means that the data causing failure recognition is predicted to cause successful recognition, i.e., $R'(x) = 1$ and $R(x) = -1$. The positive error rate is the case where $R'(x) = -1$ and $R(x) = 1$.

5. Experiments

Two face databases, FERET [9] and FRGC V1.0 [8], are used in our experiments. FERET provides a fixed gallery set and some query sets to study recognition performance under changes of facial expression (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 uncontrolled environments. We implement the PCA-based recognition method, in which each face is normalized to the size of 45 by 30, and the pixels at image corners are removed with an ellipse mask. Pixel intensity is normalized by histogram equalization. The following experiments show the results of offline selection of system parameters, recognition performance prediction, and online adjusting face alignment for better recognition.

5.1. Offline Parameter Selection

In the previous sections, we have shown that $f = \frac{\sum_i f_i}{n}$ can be used to offline select system parameters since f has near linear relationship with recognition accuracy without using training data. In this experiment, we try to find the optimal parameter out of all possible parameters based on f of PRSS. The parameters include the dimension of subspace, measurement method and measurement space, as stated in Section 3.3. The performance characteristic curves for different measurement methods are unified into one curve by linear correction, and the parameter corresponding to the largest unified f is selected as the optimal parameter. As a result, the selected parameter for FERET is $[200, Cosine, Mahalanobios]$, which means that the system uses 200 PCA features, Cosine measurement methods, and Mahalanobios space. The parameter selected for FRGC V1.0 Experiment 1 and 4 is $[120, Cosine, Mahalanobios]$. We test the recognition rates of all the possible parameters, and compare them with the recognition rate of selected parameter, as Table 2. From the table, we can observe that different query sets actually need different parameters to achieve the maximal accuracy. However, the offline selected parameters consistently achieve near-optimal accuracy under different environments even the accuracy range is large for some sets, such as FERET FC.

Table 2. Summary of parameter selection and actual recognition accuracy

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

5.2. Recognition Performance Prediction

We apply the algorithm shown in Table 1 to predict the recognition accuracy of a query set. The prediction results are summarized, and compared with actual recognition results in Table 3. In this experiment, the initial parameters of the Gaussian models are learned from a small set (20% of the whole query set), and are used to predict the performance on the remaining data. Due to model error, the algorithm usually underestimates the recognition rate. However, the method provides a rough estimation of the error range in the case that only a small portion of ground truth is provided.

Table 3. Summary of predicting recognition rate (actual recognition rate vs. predicted recognition rate)

Data Set	Rank=1	Rank=5	Rank=15
FERET FB	80% vs. 71%	90% vs. 76%	96% vs. 82%
FERET FC	49% vs. 42%	82% vs. 47%	90% vs. 57%
FERET Dup1	35% vs. 31%	46% vs. 36%	55% vs. 48%
FRGC Exp. 1	76% vs. 62%	90% vs. 75%	96% vs. 78%
FRGC Exp. 4	23% vs. 19%	45% vs. 32%	63% vs. 42%

To predict individual recognition results, the difference values between ARSS and FRSS are extracted as features to train a SVM to classify individual recognition results into two cases: success and failure cases. In the following experiments, we firstly validate the accuracy of trained predictor on FERET and FRGC data sets, and then apply the predictor to improve face recognition performance. The performance predictor is validated using cross-validation methods, in which 50% data is used for training, and the remaining 50% data is used for validation. To validate the generalization capability of trained predictor, two types of cross-validation methods, intra-set and inter-set validation methods, are applied. In the intra-set validation method, the training data is uniformly sampled from all the data sets, and then the predictor is validated on the remaining data

of each set. In the inter-set validation method, the predictor is trained with data selected from only some of the sets, and validated on the other sets. From example, when using FERET data sets, the predictor is trained with data from FB (or FC and Dup1) set, and validated on FC and Dup1 (or FB) sets. When using FRGC V1.0 data sets, the predictor is trained with experiment 1 (or experiment 4) set, and validated on experiment 4 (or experiment 1). The intra-set validation method assumes that we can obtain training data from different environments while inter-set validate method simulates the situation that we can only obtain training data of limited environments.

Figure 5 shows the intra-set validation results on FERET data sets, and the false alarm rate and positive error rate are further summarized in Table 4 for both intra-set and inter-set validation. The overall error rate of the performance predictor is between 15% and 25% for FERET sets and FRGC experiment 1 while FRGC experiment 4 shows worse accuracy. From the table, we can see that the accuracy of inter-set validation is only slightly worse than intra-set validation, which demonstrates that the presented prediction method is not constrained in a specific environment, but can be applied in various environments after the predictor has been trained.

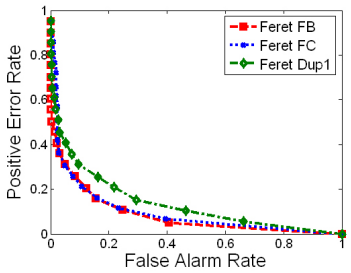


Figure 5. ROC curves of predictor on FERET (intra-set validation)

Table 4. Summary of performance prediction accuracy with intra-set and inter-set cross-validation on FERET and FRGC

Data Set	Prediction accuracy ([false alarm rate, positive error rate])	
	Intra-set validation	Inter-set validation
FERET FB	[0.1063, 0.1563] [0.1568, 0.1239]	[0.1174, 0.2079] [0.1973, 0.1530]
FERET FC	[0.1079, 0.2121] [0.1601, 0.1623]	[0.1218, 0.2096] [0.1921, 0.1622]
FERET Dup1	[0.0961, 0.3106] [0.1630, 0.2545]	[0.0843, 0.3629] [0.1783, 0.2555]
FRGC Exp. 1	[0.0896, 0.2574] [0.1642, 0.2025]	[0.1053, 0.3114] [0.1447, 0.2500]
FRGC Exp. 4	[0.1295, 0.5625] [0.2634, 0.3625]	[0.2366, 0.5000] [0.3259, 0.3500]

The predictor is applied on validation sets to improve the recognition results. To improve the recognition performance, the data predicted to cause success recognition will be preserved while the data predicted to cause failure recognition will be discarded. The experiments comparing the recognition results with or without applying performance prediction are shown in Figure 6. In the experiments, the data can actually be successfully recognized is called "good" data, and a threshold is adjusted to preserve a certain percentage of good data for recognition. The percentage is denoted as P, and the threshold corresponding to each P is obtained from training sets. The experiments results summarized in Table 5 show that performance are greatly improved by applying performance prediction. For example, the recognition rate of FERET FB set is increased to 96.2% from 80.0% when only 10% good data is discarded. For the query sets that usually have low recognition rate, such as FERET FC, FERET Dup1, and FRGC experiment 4, the performance improvement is also obvious. It is shown the error rate is near zero when only 10% of good data are preserved. But, the price paid is that many useful data is also discarded. It remains our future research to preserve all the data while still improving face recognition performance.

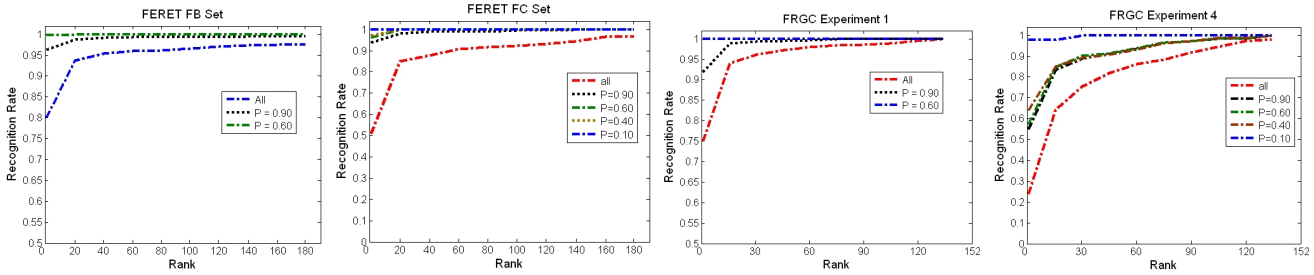
Table 5. Summary of rank 1 recognition rate with and without performance prediction

Data Set	All	P = 90%	P = 60%	P = 10%
FERET FB	80.0%	96.2%	99.7%	100.0%
FERET FC	49.3%	93.7%	96.4%	100.0%
FERET Dup1	34.7%	82.9%	93.1%	100.0%
FRGC Exp. 1	75.0%	91.8%	100.0%	100.0%
FRGC Exp. 4	23.9%	57.2%	64.3%	97.9%

5.3. Adjusting Face Alignment Online

All the above experiments use the manually marked eye positions for face alignment. However, real world applications require automatic eye localization. Although some eye localization methods have been developed, there still exist localization errors, so the results of using automatic eye localizations are consistently lower than those of using manually marked eyes [9, 12]. In addition, the problem if the manually marked eye positions can provide the optimal face alignment for recognition has not been answered in the face recognition community.

Our method automatically adjusts the eye position around an initial eye position, which is automatically or manually marked, for better recognition. In our experiments, 9 candidates are searched around each initial eye. The distance between neighbor candidates is 2 pixel. There are totally 81 eye-pair candidates to be evaluated. We calculate the f_i of each eye-pair candidate to represent its recognition quality. The eye-pair candidate corresponding to the maximal f_i is selected as the adjusted eyes for alignment. Table 6 compares the recognition rates of using the original eyes and adjusted eyes. In this experiment, the automatic eye localization method in [12] is used. It is observed that the adjusted eyes not only outperform the automatically



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

Figure 6. CMC curves of face recognition with and without performance prediction: (a) FERET FB set; (b) FERET FC set; (c) FERET Dup1 set; (d) FRGC V1.0 Experiment 1 set; (e) FRGC V1.0 Experiment 4 set

detected eyes, but also provide better recognition accuracy than the manually marked eyes.

Table 6. 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%

6. Conclusion

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 a concept of "perfect recognition," and analyze the output from "perfect recognition" to model the intrinsic system performance without training data. Based on the analysis of actual recognition similarity scores, we present methods to predict recognition results of individual or a set of query data. The presented methods provide various ways to improve the performance of recognition systems. The future work will apply our methods to other similarity measurement based biometric systems.

Acknowledgement

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] Ross Beveridge, David Bolme, Marcio Teixeira, and Bruce Draper, *The csu face identification evaluation system users guide:version 5.0*, Computer Science Department, Colorado State University, May 2003. 2, 3

[2] Corinna Cortes and Vladimir Vapnik, *Support-vector networks*, Machine Learning **20** (1995), no. 3, 273–297. 5

[3] Richard O. Duda, P.E.Hart, and David G. Stork, *Pattern classification*, second ed., John Willey Sons, 2000. 1

[4] 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*, CVPR, vol. 2. 2

[5] A.Y. Johnson, J. Sun, and A.F. Bobick, *Using similarity scores from a small gallery to estimate recognition performance for larger galleries*, IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 100–103. 2

[6] Weiliang Li, Xiang Gao, and T.E. Boult, *Predicting biometric system failure*, IEEE Intl. Conf. on Computational Intelligence for Homeland Security and Personal Safety, 2005, pp. 57–64. 2

[7] Baback Moghaddam and Alex Pentland, *Probabilistic visual learning for object representation*, IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997), no. 7, 696–710. 1

[8] 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*, CVPR, 2005. 2, 6

[9] P. J. Phillips, Hyeonjoon Moon, S.A. Rizvi, and P.J. Rauss, *The FERET evaluation methodology for face-recognition algorithms*, IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000), no. 10, 1090–1104. 2, 6, 7

[10] N.A. Schmid, B. Cukic, M. Ketkar, and H. Singh, *Performance analysis of iris based identification system at the matching score level*, ICASSP **2** (2005), 93–96. 2

[11] Elham Tabassi, Charles L. Wilson, and Craig L. Watson, *Fingerprint image quality*, Technical Report NISTIR 7151, National Institute of Standards and Technology, 2004. 2, 3

[12] Peng Wang, Matthew B. Green, Qiang Ji, and James Wayman, *Automatic eye detection and its validation*, IEEE Workshop on Face Recognition Grand Challenge Experiments, 2005. 7

[13] Rong Wang and Bir Bhanu, *Learning models for predicting recognition performance*, ICCV, 2005. 2

[14] J. Wayman, A. Jain, D. Maltoni, and D. Maio (eds.), *Biometric systems technology, design and performance evaluation*, 1 ed., Springer Publisher, 2005. 1

[15] W. Zhao, R. Chellappa, P.J. Philips, and A. Rosenfeld, *Face recognition: A literature survery*, ACM Ciomputing Survey (2003). 2