Using similarity scores from a small gallery to estimate recognition performance for larger galleries *

Amos Y. Johnson **Jie Sun** **Aaron F. Bobick**

GVU Center / College of Computing
Georgia Tech
Atlanta, GA 30332
[amos | sun | afb]@cc.gatech.edu

Abstract

We present a method to estimate recognition performance for large galleries of individuals using data from a significantly smaller gallery. This is achieved by mathematically modelling a cumulative match characteristic (CMC) curve. The similarity scores of the smaller gallery are used to estimate the parameters of the model. After the parameters are estimated, the rank 1 point of the modelled CMC curve is used as our measure of recognition performance. The rank 1 point (i.e.; nearest-neighbor) represents the probability of correctly identifying an individual from a gallery of a particular size; however, as gallery size increases, the rank 1 performance decays. Our model, without making any assumptions about the gallery distribution, replicates this effect, and allows us to estimate recognition performance as gallery size increases without needing to physically add more individuals to the gallery. This model is evaluated on face recognition techniques using a set of faces from the FERET database.

1. Introduction

In general, an identification technique measures some property of an individual and stores that information as a template. This process is repeated for a number of individuals to form a gallery of templates. Once the gallery is formed, a method is needed to ascertain the ability of the identification technique to correctly discriminate between individuals.

A widely used, yet simple, method to determine recognition performance is to extract another set of measurements from the same set of individuals (this is considered to be a probe set) and to compute the percentage of time the new measurements are correctly matched to their corresponding template. This is the nearest-neighbor method; however, this method alone does not give a way to determine recognition performance as the number of individuals in the gallery increases. As the gallery size increases, the recognition performance decreases. The goal of this paper is to be able to estimate nearest-neighbor recognition-performance of an identification technique as the number of individuals in the gallery increases using only the information provided by the original, smaller gallery.

In our pervious work [2], we achieved this by using the feature-space produced from the identification technique to mathematically model a cumulative match characteristic (CMC) curve. A CMC curve [4] shows various probabilities of recognizing an individual depending on how similar their measurements are to other individuals' measurements in the gallery. The rank 1 point on the CMC curve is the nearest-neighbor recognition-performance.

This approach, while effective, is predicated on knowing and using the feature space of the identification technique and the L_2norm or Mahalanobis distance to compute similarity within the feature space. However, there are a wide range of identification techniques that do not have an actual feature space, and similarity is computed with a variety of techniques.

In this paper, we create a CMC model using only similarity scores. This makes the CMC model more general, and removes the need to accurately model the feature-space. To determine recognition performance we focus on the rank 1 point of the CMC curve. In the following sections, related work for using similarity scores to estimate recognition performance is presented, and our model of the CMC curve is discussed and evaluated on two face recognition techniques using a set of faces from the FERET database.

2. Related Work

Estimating and/or predicting recognition performance is not a new problem. Wayman [6] and Daugman [1] present binomial models that only used the non-matching similarity scores to estimate the probability that a false match never occurs. However, [5] shows that these models dramatically underestimate recognition performance for large

* In IEEE International Workshop on Analysis and Modeling of Faces and Gestures held in conjunction with ICCV 2003 in Nice, France

gallery sizes.

In the same work [5], Phillips et al create another binomial model designed to estimate a CMC curve. The model uses both the matching and non-matching similarity scores. The match and non-match scores were used to estimate a single match and single non-match distribution as input to the model. They are able to predict recognition performance as the gallery size increases better than [6] and [1], but they too experience underestimation of recognition performance.

Our model of the CMC curve is also a binomial model most similar to [5]; however, our model differs in two ways. First, instead of estimating match and non-match distributions, we use a counting method. We found that estimating a distribution from the similarity scores is difficult given a limited amount of data using non-parametric techniques like parzen windows or k-nearest-neighbor. Second, instead of combining all non-match scores from each individual into a single non-match score set, we keep the non-match scores from each individual separate to form multiple non-match score sets. Using all the non-matching scores as a single set will underestimate recognition performance, as we will discuss in the next section.

3. Recognition Model

Before the recognition model is presented, we begin with some terminology. A gallery contains a set of templates extracted from N individuals. A probe, p^i, is another measurement of an individual, i, represented in the gallery. p^i is a random variable that maybe represented by M instances (i.e.; there are M other measurements of individual i).

After computing the similarity[1] of the M measurements for individual i to the gallery, we have k^i that is the set of all non-matching scores (size $M(N-1)$), and y^i is the set of all matching scores (size M).

3.1. CMC Model

To develop the CMC model, we first define the probability that an individual i's match score is exactly rank n to be a binomial probability distribution [3],

$$B(R, y_j^i, k^i) = C_{R-1}^{N-1} (A(y_j^i, k^i))^{R-1} (1 - A(y_j^i, k^i))^{N-R}, \quad (1)$$

where

$$C_{R-1}^{N-1} = \frac{(N-1)!}{(R-1)!(N-R)!}.$$

R is rank, y_j^i is the j^{th} match score, and $A(y_j^i, k^i)$ is the probability that y_j^i is greater than or equal to any non-match score from k^i. Next, we average over all match scores

$$\sum_{j=1}^{M} B(R, y_j^i, k^i) \frac{1}{M}.$$

The probability that an individual i's match score is within rank n, which is the definition of a CMC curve, is

$$CMC(R) = P(R \leq n) = \sum_{R=1}^{n} \sum_{j=1}^{M} B(R, y_j^i, k^i) \frac{1}{M}. \quad (2)$$

Finally, the expected value of the CMC curve over all individuals, is

$$\overline{CMC(R=n)} = \sum_{i=1}^{N} \sum_{R=1}^{n} \sum_{j=1}^{M} B(R_i, y_j^i, k^i) \frac{1}{M} \frac{1}{N}. \quad (3)$$

3.2. Nearest-Neighbor Model

Equation 3 is the expected value of the CMC curve for any rank n; however, here we focus on the rank 1 point. The rank 1 point is probably the most important point because it answers the question *how well does the identification technique recognize individuals* in one number. If we evaluate Equation 3 at rank 1 it becomes

$$\overline{CMC(R=1)} = \sum_{i=1}^{N} \sum_{j=1}^{M} (1 - A(y_j^i, k^i))^{N-1} \frac{1}{M} \frac{1}{N}. \quad (4)$$

Before we discuss evaluating this equation, we make a necessary and practical assumption that all individuals have a similar distribution of *match* scores. This assumption is necessary because some identification experiments only extract a small number of measurements per individual and usually only 1 (i.e.; M = 1). With such a small M, the average over the match scores in Equation 4 would not be accurate. The assumption is also practical because if we assume that measurement variation for an individual due to senor noise or condition is independent of the identity of the individual and is greater than the inherent individual variation, then measurement noise should dominate the cause of variation.

Using this assumption we combine the match scores, y^i, for all individuals into a single match-score set, $x = [y_{1..M}^1, ..., y_{1..M}^N]$, so Equation 4 becomes

$$\sum_{i=1}^{N} \sum_{j=1}^{MN} (1 - A(x_j, k^i))^{N-1} \frac{1}{MN} \frac{1}{N}. \quad (5)$$

We note that Equation 5 is the expected value of the recognition performance; however, if we combined all the

[1] In this paper, the closer the match the higher the similarity score.

non-match scores into one set, $T = [k^1 ... k^N]$, and substituted T for k^i in Equation 5,

$$\sum_{j=1}^{MN} (1 - A(x_j, T))^{N-1} \frac{1}{MN},$$

this result would not give the expected value of the recognition performance. It would give an underestimate of the expected value of the recognition performance because an average over the sets of non-match scores (per individual) would not have been taken.

Returning our attention to Equation 5, we make a distinction between the number of individuals, N_p, in the set of probes, and the number of templates, N_g, in the gallery. Every individual i in the set of probes has a corresponding template in the gallery; however, the number of templates in the gallery maybe larger than the number of probes. Also the variable N is redefined to be the size of the gallery being estimated, and Equation 5 becomes

$$\sum_{i=1}^{N_p} \sum_{j=1}^{MN_p} (1 - A(x_j, k^i))^{N-1} \frac{1}{MN_P} \frac{1}{N_p}. \qquad (6)$$

We now need to find $1 - A(x_j, k^i)$, which is the probability that the match score x_j is less than any non-match score from k^i. If we had enough non-match scores, a density of the non-match scores could be estimated non-parametrically. However, the most important part of the match density to estimate is the region where match scores exist. If the match score region is very small, then of all the non-match scores only a few are used to estimate the density over that region. With only a few scores, the estimate of that region will not be representative of the true density. Since the number of non-match scores in the region of overlap maybe limited, we choose to use the non-match scores *as is* and do simple counting. Using counting, our nearest-neighbor recognition performance model is

$$\sum_{i=1}^{N_p} \sum_{j=1}^{MN_p} (\frac{v(k^i, x_j)}{M(N_g - 1)})^{N-1} \frac{1}{MN_P} \frac{1}{N_p}. \qquad (7)$$

where $v(k^i, x_j)$ is the number non-match scores from k^i that are less than match score x_j and $M(N_g - 1)$ is the total number of non-match scores per individual.

3.3. Error Analysis

Computing the probability $\frac{v(k^i, x_j)}{M(N_g - 1)}$ is equivalent to computing a CDF for the non-match scores (see Figure 1). Because of a finite sample of non-match scores, the CDF is discontinuous where as it should be continuous. This problem is magnified for lower mathch-scores where there are not any samples, at these points the CDF has a zero probability.

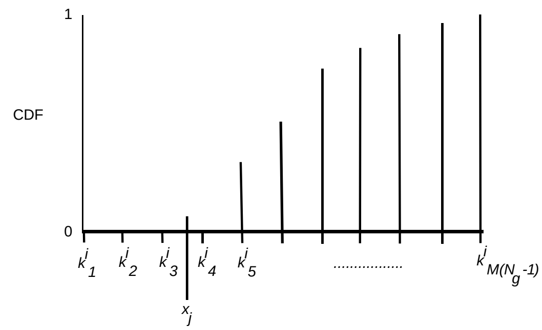

Figure 1: CDF for the non-match scores k^i. x_j (match score).

Table 1: Recognition Performance: The estimated recognition performances were estimated from a gallery of size 100. Technique #1.

Gallery Size (N)	Actual	Estimate	Error
100	91.0%	90.6%	-0.4%
500	86.6%	85.1%	-1.5%
1005	81.4%	80.8%	-0.6%

To combat this problem the gallery size, N_g, could be increase to create a denser sampling of the non-match scores. However, since the goal of this paper is to estimate recognition performance for a large gallery from a smaller gallery, increasing N_g is not desired. Instead, linear interpolation my be employed between sample points on the CDF.

3.4. Model Limitations

Our model is limited by the available gallery and probe sets, which are used as input to the model. The model is designed to estimate/predict recognition performance for larger galleries with similar characteristics of the smaller gallery and probe sets that were used to make the prediction. For example, in a face-recognition test, if the gallery and probe set have frontal face images, our model will estimate recognition performance for larger galleries of frontal images. However, it is not designed to estimate recognition performance of a new gallery or probe set containing images significantly differing from the frontal condition.

4. Experimental Evaluation

To demonstrate our model's (Equation 7) ability to estimate recognition performance, we use face images from the FERET database provided to us by Phillips et al at NIST. From this data set, we used the similarity scores of 1005 individuals generated by two different face recognition techniques[2]. The gallery contains 1005 individuals from condition "FA", and the probe set contains condition "FB" of

[2] Multiple face recognition techniques were evaluated and compared at NIST using the face images from the FERET data set. We randomly chose two of the techniques to evaluate our recognition model.

Table 2: Recognition Performance: The estimated recognition performances were estimated from a gallery of size 100. Technique #2.

Gallery Size (N)	Actual	Estimate	Error
100	91.0%	90.8%	-0.2%
500	88.0%	85.8%	-2.2%
1005	82.1%	81.5%	-0.6%

Table 3: Recognition Performance: The estimated recognition performances were estimated from a gallery of size 100. Different sets of 100 and 500 (than Table 2) for the Actual calculated recognition performance were used. Technique #2.

Gallery Size (N)	Actual	Estimate	Error
100	88.0%	90.8%	+2.0%
500	84.6%	85.8%	+1.2%

those 1005 individuals. Of the 1005 individuals, we randomly chose 100 individuals to train our model. Therefore we have $N_p = 100, N_g = 100, M = 1$.

Table 1 and 2 show the results of estimating gallery sizes of 100, 500, and 1005, just using a gallery size of 100 to make the estimation. The galleries of 100 and 500 where chosen randomly from the set of 1005. The estimates of the larger gallery sizes are very close to the actual calculated ones, which demonstrate the ability of our model to estimate recognition performance.

Table 3 uses a different set of 100 and 500 individuals than Table 2 (for the same Technique). The difference in the actual calculated recognition performance simply illustrates that recognition performance for a technique is a random variable with a mean and variance.

5. Conclusion and Future Work

In this paper we have presented a mathematical model to estimate recognition performance for larger galleries using similarity scores of a significantly smaller gallery. The model doesn't make any assumptions about the distribution of the data. The result of experimentation shows that our model is able to estimate recognition performance.

In future work, we would like to determine how large N_p and N_g need to be to accurately estimate a gallery of size N, determine what error should be expected if N_p and N_g are not large enough, and determine which parameter N_p or N_g is most important to estimating recognition performance.

Acknowledgments

We would like to thank Jonathan Phillips and Patrick Grother for providing us with face data and useful insight into CMC behavior. Funding for this research was supported in part by the DARPA HID Program, contract #F49620-00-1-0376.

References

[1] J. Daugman. The importance of being random: statistical principals of iris recognition. *Pattern Recognition*, 36(2):279–291, 203.

[2] A. Y. Johnson, J. Sun, and A. F. Bobick. Predicting large population data cumulative match characteristic performance from small population data. *4th International Conference on Audio- and Video Based Biometric Person Authentication*, University of Surrey, Guildford, UK, June 2003.

[3] W. Mendenhall and R. J. Beaver. *Introduction to Probability and Statistics*. Duxbury Press, Belmont, California, 9 edition, 1994.

[4] H. Moon and P. J. Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *In Perception*, 30(3):303–321, 2001.

[5] P. J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone. Face recognition vendor test 2002: Evaluation report. Technical report, NIST, March 2003. http://www.frvt.org/FRVT2002/documents.htm.

[6] J. L. Wayman. Error-rate equations for the general biometric system. *IEEE Robotics & Automation Magaizine*, 6(1):35–48, 1999.