

Article

Biometric Performance as a Function of Gallery Size

Lee Friedman ^{1,*} , Hal Stern ² , Vladyslav Prokopenko ¹, Shagen Djanian ³, Henry Griffith ⁴  and Oleg Komogortsev ¹

¹ Department of Computer Science, Texas State University, 601 University Dr, San Marcos, TX 78666, USA

² Department of Statistics, University of California–Irvine, Irvine, CA 92697, USA

³ Department of Computer Sciences, Aalborg University, 9220 Ålborg, Denmark

⁴ Engineering, San Antonio College, San Antonio, TX 78212, USA

* Correspondence: lfriedman10@gmail.com

Abstract: Many developers of biometric systems start with modest samples before general deployment. However, they are interested in how their systems will work with much larger samples. To assist them, we evaluated the effect of gallery size on biometric performance. Identification rates describe the performance of biometric identification, whereas ROC-based measures describe the performance of biometric authentication (verification). Therefore, we examined how increases in gallery size affected identification rates (i.e., Rank-1 Identification Rate, or Rank-1 IR) and ROC-based measures such as equal error rate (EER). We studied these phenomena with synthetic data as well as real data from a face recognition study. It is well known that the Rank-1 IR declines with increasing gallery size, and that the relationship is linear against $\log(\text{gallery size})$. We have confirmed this with synthetic and real data. We have shown that this decline can be counteracted with the inclusion of additional information (features) for larger gallery sizes. We have also described the curves which can be used to predict how much additional information would be required to stabilize the Rank-1 IR as a function of gallery size. These equations are also linear in $\log(\text{gallery size})$. We have also shown that the entire ROC-curve was not systematically affected by gallery size, and so ROC-based scalar performance metrics such as EER are also stable across gallery size. Unsurprisingly, as additional uncorrelated features are added to the model, EER decreases. We were interested in determining the impact of adding more features on the median, spread and shape of similarity score distributions. We present evidence that these decreases in EER are driven primarily by decreases in the spread of the impostor similarity score distribution.

Keywords: biometrics; gallery size; synthetic data; rank-1 identification rate; equal error rate (EER)



Citation: Friedman, L.; Stern, H.; Prokopenko, V.; Djanian, S.; Griffith, H.; Komogortsev, O. Biometric Performance as a Function of Gallery Size. *Appl. Sci.* **2022**, *12*, 11144. <https://doi.org/10.3390/app122111144>

Academic Editors: Andrey Makrushin and Jana Dittmann

Received: 11 September 2022

Accepted: 24 October 2022

Published: 3 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Many developers of biometric systems start with modest samples before general deployment. However, they are interested in how their systems will work with much larger samples. To assist them, we evaluated the effect of gallery size on biometric performance.

We start with the hypothesis that as gallery size increases, more information is required to achieve any particular level of performance. We will be addressing performance in terms of identification rate, specifically Rank-1 Identification Rate (Rank-1 IR), and also in terms of ROC-based measures such as EER. (All error rates rates are expressed as percent in the present study). It is well known from empirical studies that the Rank-1 IR declines with increasing gallery size and that this relationship is linear in $\log(\text{gallery size})$ (see Figure 1) [1–3]. We will attempt to further describe this relationship, and also show that the decrease in Rank-1 IR can be reversed with the addition of new information.

Baveja et al. 2010 [2] conducted a theoretical analysis of this issue using extreme-value theory. They develop large sample approximations to illustrate how Rank-1 IR should change with gallery size. Their basic result (assuming independence of the probe scores against different imposters) gives a linear relationship in $\log(\text{gallery size})$ to the first order.

The authors show that their theoretical result does not match up well with empirical results. However, they indicate that this was likely due to dependence among the probe-imposter scores due to image quality; a modification using a categorization of image quality provides a much better fit to the empirical results. The end of their paper talks about the “open”-set case where ROC-curves tend to be utilized. They indicate that asymptotic theory suggests ROC metrics should depend on the tail of the imposter score distribution; no empirical comparisons are provided.

Our work complements that of Baveja et al. (2010) [2] in that we provide extensive empirical evidence regarding the effect of gallery size. Our results support their theoretical suggestion of a log-linear relationship. We also probe the importance of feature reliability (i.e., temporal persistence) among the features used to define the score. We also perform an extensive empirical analysis of the effect of gallery size on ROC-curves. Although Baveja et al. (2010) [2] emphasize the important role of the tails of the similarity score distributions, in the present study we emphasize the role of the spread (i.e., inter-quartile range (IQR)) of the similarity score distributions.

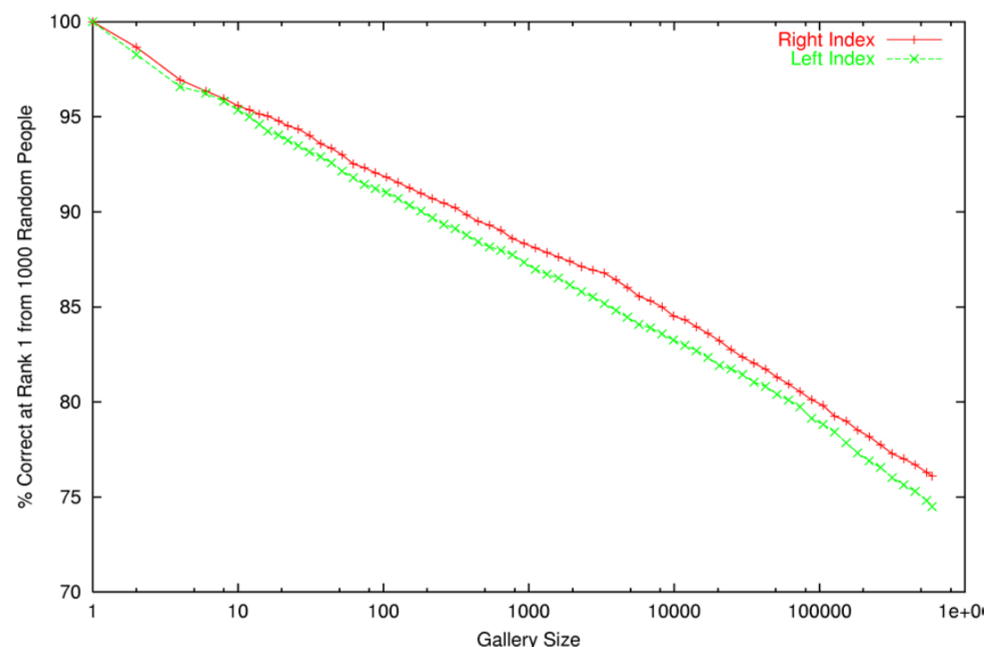


Figure 1. Probability of detection at Rank-1 IR as a function of more than 600,000 plain index fingers [1] (See Figure 19). Note the log scale for the x-axis. Probe set size was 1000 subjects. Fingerprint source: Department of Homeland Security, US Federal Government.

There is little published research on the impact of increasing gallery size on ROC-based measures such as EER, aside from the work of Baveja et al. 2010 [2], noted above. Although it is well established that increasing the number of subjects will decrease the confidence limits on any ROC-curves produced [4,5], this evidence does not imply anything about the central tendency of any estimated error rate.

We were interested in evaluating the influence of changes in gallery size on a large scale (up to 100,000 subjects). For reasons of convenience, availability and control, we employed synthetic data sets. However, at each step in the analysis, we provide comparison analyses for a real face-recognition data set (MORPH-II) [6] with $N = 13,930$ subjects. The real face recognition data allowed us to evaluate if the substantive findings with synthetic data were replicated in real biometric data.

In Section 2 of the manuscript, we present a method for creating synthetic data sets with a number of properties that are helpful for studying biometric performance. Because the data are synthetic, we are able to control the degree of temporal persistence of the features while also ensuring that features are approximately independent of each other and thus

provide unique pieces of information for biometric verification. (The concept of “temporal persistence” and the method for its measurement are covered in [7]. In other contexts, this measure is used to assess the inter-rater reliability of a feature). We think that having unique pieces of information will allow us to address several theoretical notions relevant to biometric analysis in this and subsequent studies. Additionally, in Section 2, we present our methods for biometric performance assessment of the synthetic data. In Section 3 we describe the MORPH-II face recognition data set and our face recognition analyses, including biometric performance assessment. In Section 4, we discuss our results for Rank-1 IR. In Section 5 we describe our results for ROC-based metrics. We end with a discussion (Section 6).

2. Creation and Analysis of Synthetic Data Sets

2.1. Creation of Synthetic Data

Recall that the intraclass correlation coefficient (ICC) is a measure of the correlation expected for repeated measurements of the same feature on different occasions. Unlike the Pearson r correlation coefficient, which is typically applied as an interclass measure of relative agreement (i.e., two series can be correlated even if they differ substantially in level and spread), the ICC is an intraclass measure of absolute agreement [7]. Measures from the same set of subjects at two different times are intraclass measurements (same metric and variance). ICC ranges from 0.0 to 1.0 with the latter corresponding to perfect temporal persistence. Our goal is to create synthetic features with a specified target ICC (denoted ICC_{Target}). Let X_{ijs} denote the measurement of feature j ($j = 1, \dots, K$) on session (occasion) s ($s = 1, \dots, S$) for individual i ($i = 1, \dots, N$). Although the ICC can be calculated based on many sessions, in our experience, biometric assessment is typically performed comparing only two points in time. Therefore, henceforth we will set $S = 2$. We generate normally distributed features such that the theoretical intraclass correlation of repeated measurements of the same feature on the same subject is ICC_{Target} while the theoretical correlation of measurements of different features on the same individual and the theoretical correlation of measurements from different individuals are zero. In practice when data are simulated there are small variations in the empirical ICCs and there are small intercorrelations between features (and individuals) due to chance.

The R code to create our datasets is included in [8]. Here we describe the method briefly. A more formal presentation is spelled out in Algorithm 1. The starting point is to populate the full set of session one measurements X_{ij1} with random draws from a standard normal distribution (mean zero and variance one). Then the measurements for the second session are set equal to the value of the given feature from the first session, $X_{ij2} = X_{ij1}$ for ($i = 1, \dots, N, j = 1, \dots, K$). At this point both sessions have the same data and each feature has ICC equal to 1.0 (perfect persistence). We obtain the desired ICC by adding a draw from a normal distribution with mean = 0 and variance = $(1 - ICC_{Target}) / ICC_{Target}$ to each of the measurements. At the end we apply a z-score transform to each feature (with all sessions concatenated together) so that they all have mean 0 and standard deviation one. It can be shown that the resulting measurements have the desired ICC (up to simulation noise).

Using this method, we can create features which are normally distributed, that have specified ICCs, with as many subjects and sessions as we desire. These features all have mean = 0 and SD = 1. These features are generally independent, but there are some small intercorrelations between features due to chance. To illustrate the approach, we generated data for 10,000 subjects, 1000 features and 2 occasions with $ICC_{Target} = 0.7$. Figure 2A shows a histogram of the resulting empirical ICCs. Figure 2B shows a histogram of the resulting inter-feature correlations.

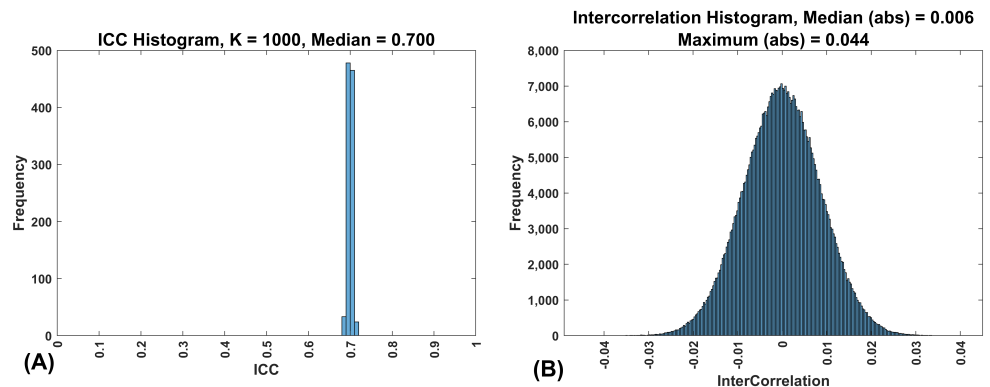
Algorithm 1: Creating Synthetic Features**Input** : N (subjects), K (features), ICC_{Target} **Output**: 3-dimensional ($N \times K \times 2$) feature matrix X_{ijs} with desired correlation structurefor $j = 1, \dots, K$ for $i = 1, \dots, N$ Set $X_{ij1} = Z$ where Z is a random standard normal deviate. Set $X_{ij2} = X_{ij1}$ for $j = 1, \dots, K$ for $i = 1, \dots, N$ for $s = 1, 2$ Set $X_{ijs} = X_{ijs} + W$; where W is a random normal deviate with mean = 0
 and standard deviation = $\sqrt{(1 - ICC_{Target}) / ICC_{Target}}$ For each feature j , treat X_j as a single vector of length $N \cdot S$ and apply a z-score transform to ensure mean = 0 and standard deviation = 1

Figure 2. (A) Frequency histogram of ICCs for 1000 features with an $ICC_{Target} = 0.7$. This was from a synthetic data set with 10,000 subjects. (B) Frequency histogram of correlations between 1000 features for 10,000 subjects, two sessions, with an $ICC_{Target} = 0.7$. Note that the median and maximum are of the absolute value of the correlations.

2.2. Creation of Sets of Features with Varying Degrees of Persistence

To match what was found in our face recognition data set (see below), we generated a synthetic data set with varying ICCs in the range of 0.8 to 0.9 ($0.8 < ICC \leq 0.9$). We refer to this data set as the Band 8 data set. Within the ICC band, the ICCs are evenly distributed across the 0.8 to 0.9 range.

2.3. Biometric Performance Assessment for Synthetic Features

All of the synthetic analyses for this report are based on Band 8 ($0.8 < ICCs \leq 0.9$). This was chosen because in our face recognition data set, described below, all of the features had ICCs in this range as well. A synthetic dataset with 50 features, and 100,000 subjects studied on 2 sessions with ICC in Band 8 was constructed.

Biometric performance was evaluated using various numbers of features. Given very low or non-existent inter-correlations between features, the number of features is a proxy for biometric performance. Higher numbers of features always produce better biometric performance. The number of features ranged from 2 to 20 for both synthetic and FaceNet data, depending on the goal of the analysis. The precise number is always provided for each analysis. These features were randomly selected from the full set of 50 features.

For one analysis of the synthetic data, we employed gallery sizes of 1000, 2000, 4000, 8000, 16,000, 32,000 or 64,000 subjects. For other types of analysis, gallery sizes were 1000,

10,000 or 100,000 subjects. We employed the cosine distance metric, since we have shown in an earlier (unpublished) report that the best biometric performance was produced with this choice (Link to unpublished report: <https://www.doi.org/10.13140/RG.2.2.17510.06727>, (accessed on 20 October 2022)). The resulting distance measures were scaled to go from 0 to 1 and then they were reflected ($1 - \text{distance}$) to compute similarity scores. The Rank-1 IR for a data set represents how often (what percent of the time) the greatest similarity score for a probe subject was for the same subject in the gallery. A “genuine” distribution of similarity scores was constructed from the similarity scores for each subject and their self. All other similarity scores were considered impostors. Applying different decision thresholds to the genuine and impostor similarity scores yields false acceptance and false rejection rates. These can be plotted as a receiver operating characteristic (ROC) curve. The EER is the point on this curve at which the false acceptance rate (FAR) and the false rejection rate (FRR) are equal.

The number of similarity scores is equal to $(\text{gallery size})^2$. However, we can only do exact ROC analyses in computer memory for up to 20,000 subjects. Therefore we created software which estimates ROC-based measures for much larger sample sizes, and which does not hold all of the similarity scores in memory. We will call the large scale procedure for estimating ROC-based measures BIG_ROC. Python code is available at <https://digital.library.txstate.edu/handle/10877/16141>, (accessed on 20 October 2022). Download `big_roc-master_12_04_2019.zip`.

The core of this analysis is the computation of the genuine and impostor distribution frequency histograms. All the ROC-based statistics are computed from these two histograms. First, both frequency histograms are initialized to zero. We use 1000,000 bins of equal size. Cosine similarity scores can assume values from 0 to 1, so each bin has a width of $1/1000,000$. Then both histograms are computed from genuine and impostor similarity scores of subjects from session 1 and session 2. The similarity scores are calculated iteratively, in batches of 1000 subjects from each session per batch (1000,000 similarity scores at a time).

False rejection rate (FRR) and false acceptance rate (FAR) values are computed based on the genuine and impostor distributions frequency histograms: one FRR and FAR value per histogram bin. Let $\text{FRR}(\text{Similarity score})$ and $\text{FAR}(\text{Similarity score})$ functions be piecewise linear functions based on these FRR and FAR values. EER was computed as a value at the intersection of $\text{FAR}(\text{Similarity score})$ and $\text{FRR}(\text{Similarity score})$ functions.

The statistics of genuine and impostor distributions are calculated based on a random sample from the respective PDFs, approximated by the relative frequency histograms (calculated from the aforementioned frequency histograms)

3. Face Recognition Data and Methods

3.1. Data Set and Image Preparation

We would have preferred to find a publicly available data set with 100,000 or more subjects. We were not successful. However, we did find the MORPH Craniofacial Longitudinal Morphological Face Database (MORPH-II) [6] https://uncw.edu/oic/tech/morph_academic.html (accessed October 30, 2022)). The MORPH-II data set contains mug shots for 13,930 subjects. Since biometric performance assessment requires at least 2 images per subject, subjects with only one image were excluded ($N = 857$). The images are colored (RGB), have various dimensions, include more than just the face, and are not spatially registered. We employed the Viola-Jones algorithm for face detection. The Viola-Jones algorithm failed in 836 of 53,404 total images. Viola-Jones algorithm failures accounted for 69 subjects being lost. This left 13,004 subjects with 2 or more images for further analysis. Prior to processing these images for facial recognition, the steps in Algorithm 2 were applied.

Algorithm 2: Steps in the preparation of images for facial recognition.

1. Detect faces in the images using the Viola-Jones algorithm and save the face-only images.
2. Register the face-only images using an affine transformation (translation, rotation, scale, and shear).
3. Save registered images as 120×100 pixels.
4. For each subject, correlate each image (as gray scale) with each other image.
5. For each subject, choose the 2 most highly correlated images, and discard all other images.

Note that our goal was not to provide a fair assessment of our face recognition approach to the MORPH-II data set. Rather, it was to perform a reasonable face recognition analysis which could then be used to evaluate gallery size effects. For this purpose, we wanted a full range of performance, and our choice to use only the most highly correlated pairs of images was designed to obtain excellent performance under optimal conditions.

3.2. Face Recognition Approach—FaceNet

For face recognition, we employed the set of features supplied by the FaceNet algorithm [9]. FaceNet is a deep convolutional network designed by Google, trained to solve face verification, recognition and clustering problems with efficiency at scale. It is highly accurate and robust to occlusion, blur, illumination, and steering. It directly maps face images to a compact Euclidean space, where distances directly correspond to a measure of face similarity. Once this space has been produced, tasks such as face recognition can be easily implemented using standard techniques. It achieved accuracy of 99.63% on the Labeled Faces in the Wild (LFW) data set, and 95.12% on the YouTube Faces Database.

To create the FaceNet features, we applied the python Keras implementation (Available at: <https://github.com/nyoki-mtl/keras-facenet>, (accessed 30 October 2022)). We employed the pretrained Keras model (trained using the Microsoft-Celeb-1M data set, which as of 6 June 2019 is no longer available). For this, the images needed to be resized to 160×160 pixels and globally rescaled (compute the mean and SD across all intensities from each color channel and, for each channel, subtract the mean intensity and divide by the SD intensity). The algorithm produced 128 numerical features per subject.

3.3. Checking Distributions for Normality of FaceNet Features

To assess the normality of the FaceNet features, we computed the skewness and kurtosis of each feature. The normal distribution has a skewness of 0 and a kurtosis of 3.0. The 128 FaceNet features had a skewness range of -0.23 to 0.34 and a kurtosis range of 2.76 to 3.33 . On the basis of these ranges, we considered it reasonable to treat all of the FaceNet features as normal. For comparison, the skewness of a random uniform distribution ($k = 10,000$) was -0.015 and the kurtosis was 1.82 . A log-normal distribution with $\text{mean}(\log(x)) = 0$ and $\text{sd}(\log(x)) = 1$ had a skewness of 5.35 and a kurtosis of 59.16 .

3.4. Obtaining the ICC and the Feature Intercorrelation for the FaceNet Features

To further characterize the FaceNet features, we were interested in determining the temporal persistence of the features. For this we computed the ICC of each feature and present a frequency histogram of these ICCs in Figure 3A. All of the features fall between $\text{ICC} = 0.8$ and 0.9 . This means that the features are all highly reliable and are very similar to our synthetic features for Band 8.

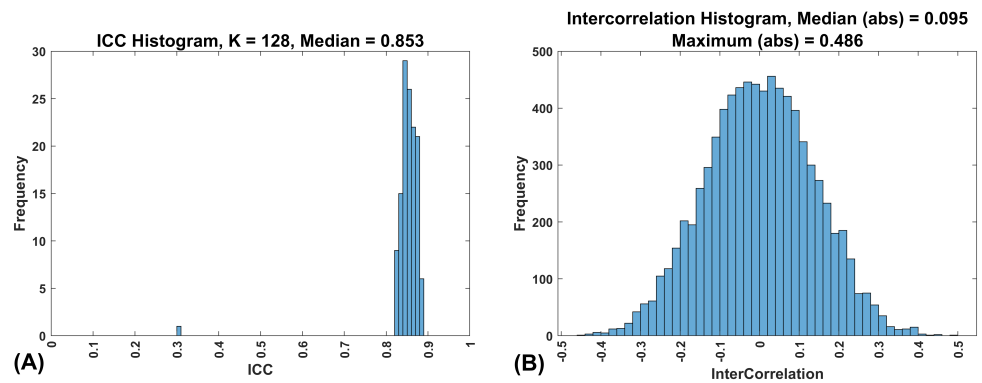


Figure 3. Characteristics of the FaceNet features. (A) This is a frequency histogram of the ICCs for the 128 FaceNet features. All of the features are between 0.8 and 0.9, which corresponds exactly with our synthetic Band 8 features. This indicates that these features are very reliable over time. (B) This is a frequency histogram of the intercorrelations of the 128 FaceNet features.

The FaceNet features were substantially inter-correlated (Figure 3B) For this reason, we decided to perform a PCA on these features.

3.5. PCA Analysis of FaceNet Features

In order to create a set of uncorrelated features and to reduce the dimensionality of the data set, we performed a PCA analysis on the FaceNet features. This analysis included the data for all subjects. However, during biometric assessment of subsets of the entire dataset (see below), PCA was independently performed on each subset. As is clear from Figure 4, only approximately 60 uncorrelated features were required to explain 100% of the variance in the 128 FaceNet features.

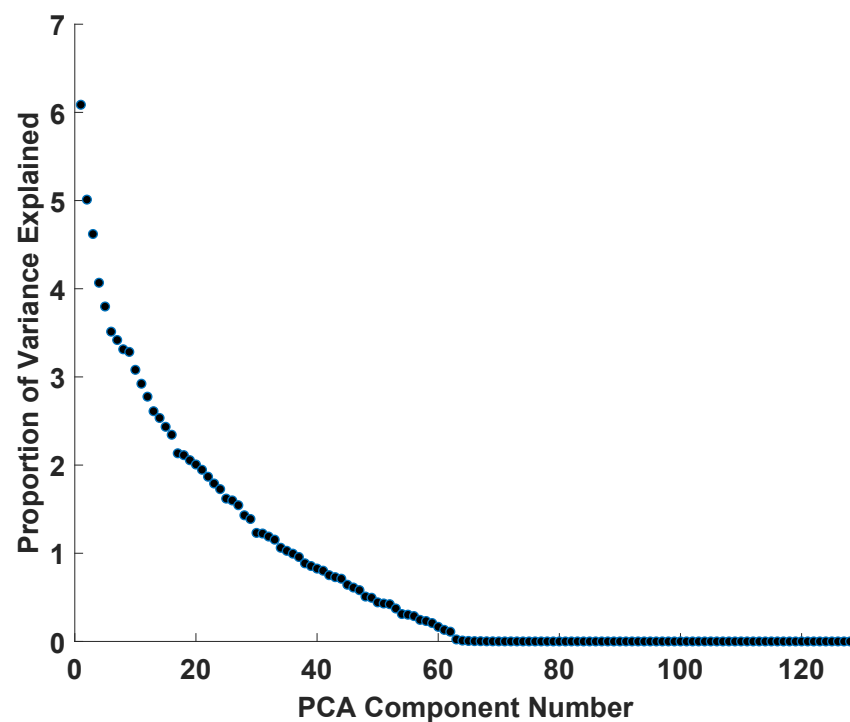


Figure 4. After PCA of the 128 FaceNet Features, we plot the variance explained by each PCA component against component number. Essentially all of the variance is accounted for by approximately 60 completely uncorrelated PCA components.

3.6. Biometric Performance Assessment for PCA Components from FaceNet Features

Biometric performance was evaluated for various numbers of PCA components. The number of PCA components is a proxy for biometric performance. Higher numbers of components always produce better biometric performance. The number of components ranged from 4 to 20, but the precise number was chosen to illustrate each particular analysis. Gallery sizes ranged from 1000 to 9000 in steps of 1000. After a subset of the 13,004 subjects was selected as a data set, PCA was computed on all the first images of each pair. The PCA coefficients computed from the first images were then used to calculate the PCA components for the second images. Biometric performance was evaluated in memory (i.e., the BIG_ROC was not needed). Cosine distances were computed, converted to similarity scores, and subjected to a conventional ROC analysis.

4. Results: Rank-1 IR

The Rank-1 IR performance for synthetic Band 8, for 10 features, evaluated at various gallery sizes (1000, 2000, 4000, 8000, 16,000, 32,000, 64,000) is presented in Figure 5A. Every dot in this figure represents the mean across 30 random repetitions (random subset of features and subjects). The Rank-1 IR for 1000 subjects was approximately 58%. This rate declines steadily as the gallery size increases. At a gallery size of 64000 the Rank-1 IR has dropped below 15%. The decrease is described by a linear function of $\log(\text{Gallery Size})$ (Figure 5B).

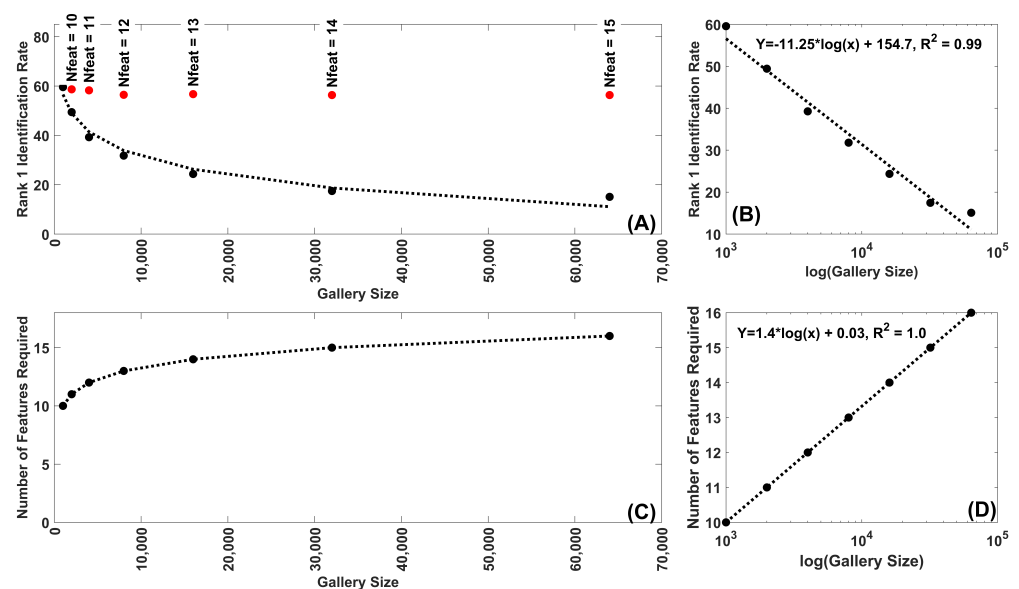


Figure 5. (A) Rank-1 IR as a function of gallery size for synthetic Band 8 features. Each dot represents the mean of 30 repetitions. The black dots are the Rank-1 IR for 10 features for the following gallery sizes: 1000, 2000, 4000, 8000, 16,000, 32,000, 64,000 subjects. The red dots represent the Rank-1 IR for feature numbers greater than 10 which are chosen to produce a Rank-1 IR most similar to that for 10 features and 1000 subjects. (B) Same data as (A) plotted on a $\log(\text{Gallery size})$ scale. Note the fit of the decline to a linear function of $\log(\text{Gallery Size})$. (C) Plot of the number of features required for gallery sizes greater than 1000 to match the Rank-1 IR for 10 features, 1000 subjects. (D) Same data as (C) plotted on a $\log(\text{Gallery size})$ scale. Note the fit of the increase to a linear function of $\log(\text{Gallery Size})$. In this case a linear equation in $\log(\text{Gallery Size})$ was able to match the results perfectly (r-squared = 1.0).

This drop can be prevented if additional information is added in the form of additional features. We understand that, in any real world application, investigators would likely use all of the information available in the first instance. Nonetheless, we believe this additional analysis provides some insight and guidance for improving Rank-1 IR performance. The red dots in Figure 5A represent the Rank-1 IR that can be achieved by adding additional

features. The number of features needed to achieve this performance was also indicated. Note that this analysis is limited in accuracy by the discrete nature features.

In Figure 5C, we plot the number of features required to stabilize the Rank-1 IR as gallery size increases. In this case, the best fitting function for these feature numbers was also a linear function in terms of $\log(\text{Gallery Size})$ (Figure 5D).

In Figure 6, we present a comparable analysis for our face recognition data set. In this case the Rank-1 IR was calculated for the first 10 PCA components, as a function of gallery size from 1000 to 9000 in steps of 1000 (Figure 6A). Once again, we see a decrease in Rank-1 IR as gallery size increases. The decrease was linear in $\log(\text{gallery size})$ (Figure 6B). The number of PCA components required to stabilize the Rank-1 IR performance was also a linear function of $\log(\text{gallery size})$ (Figure 6D).

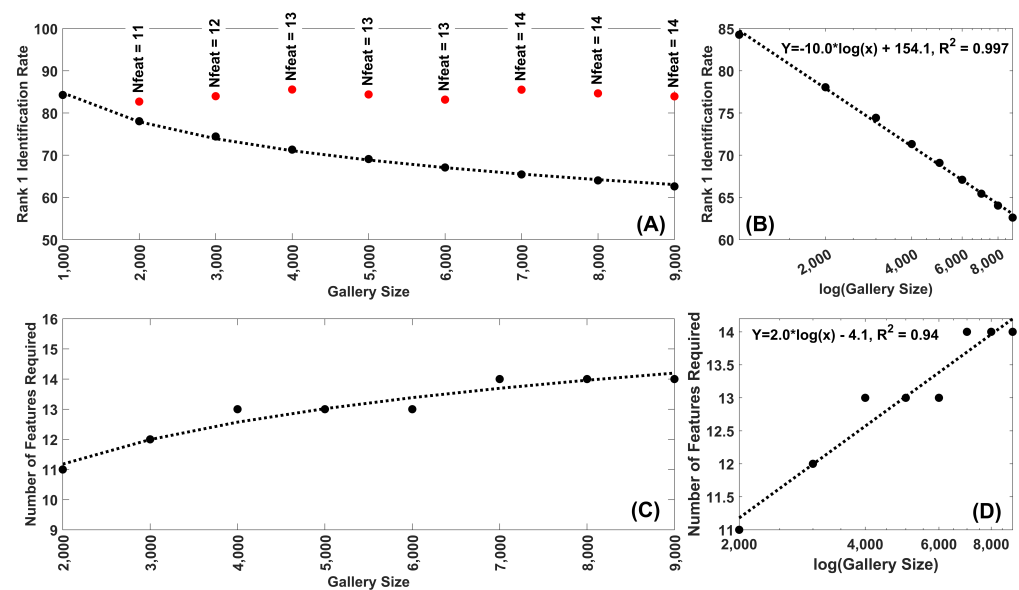


Figure 6. (A) Rank-1 IR as a function of gallery size for FaceNet features. Each dot represents the mean of 30 repetitions. The black dots are the Rank-1 IR for 10 features for gallery sizes from 1000 to 9000 in 1000 size steps. The red dots represent the Rank-1 IR for feature numbers greater than 10 which are chosen to produce a Rank-1 IR most similar to that for 10 features and 1000 subjects. (B) Same data as (A) plotted on a $\log(\text{Gallery size})$ scale. Note the fit of the decline to a linear function of $\log(\text{Gallery Size})$, with an $r^2 = 0.997$. (C) Plot of the number of features required for gallery sizes greater than 1000 to match the Rank-1 IR for 10 features, 1000 subjects. (D) Same data as (C) plotted on a $\log(\text{Gallery size})$ scale. Note the fit of the increase to a linear function of $\log(\text{Gallery Size})$. In this case a linear equation in $\log(\text{Gallery Size})$ was able to match the results quite well ($r^2 = 0.94$).

5. Results: ROC-Based Measures

5.1. Equal Error Rater (EER)

Figure 7 (Top) illustrates the EER across gallery size for synthetic data evaluated at 5, 10, 15 and 20 features. Each point is the mean of 30 repetitions (random subset of features and subjects). Mean EER was apparently very stable across gallery size. Figure 7 (Bottom) illustrates the EER across gallery size for our face recognition data evaluated at the first 5, 10, 15 and 20 PCA components. Each point is the mean of 30 repetitions (over randomly chosen subjects). Note the stability across gallery size here.

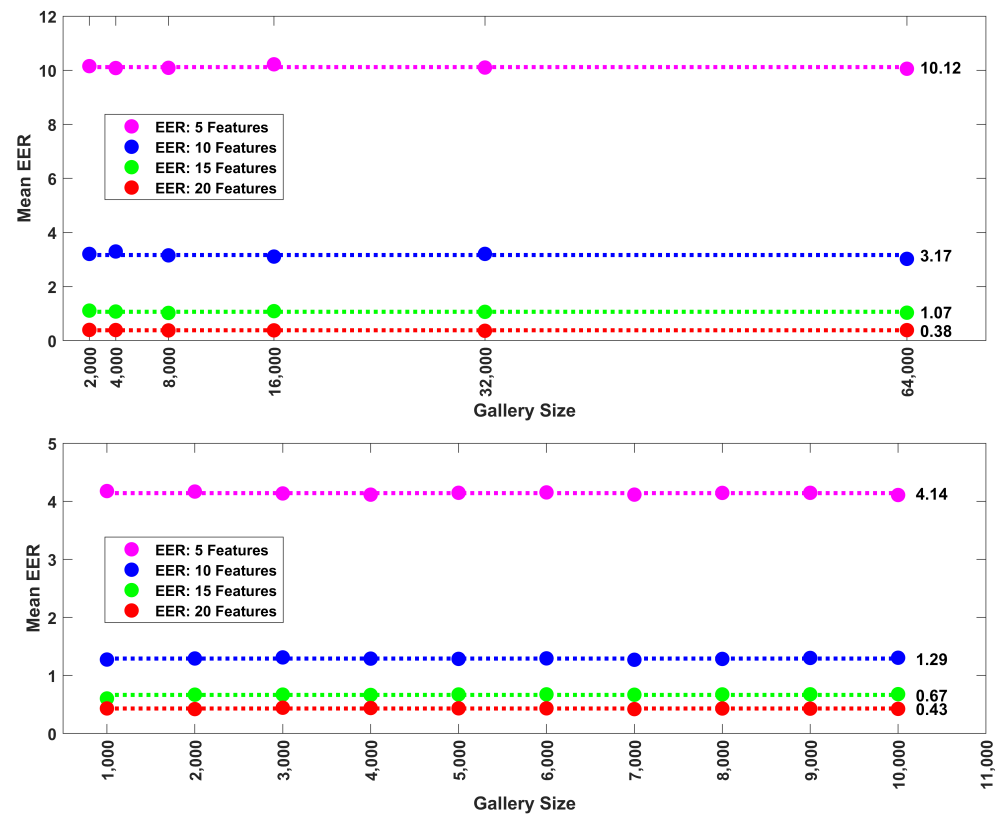


Figure 7. (Top): Relationship between mean EER (30 repetitions) and gallery size for synthetic data (Band 8). Data shown for 5, 10, 15 and 20 features. The average EERs (shown to the right of the last dot) across gallery size are illustrated with dotted lines. Mean EER was highly stable with increases in gallery size. **(Bottom):** Relationship between mean EER (30 repetitions) and gallery size (1000 to 10,000 subjects in 1000 subject steps) for FaceNet PCA components. Data shown for the first 5, 10, 15 and 20 PCA components. The average EERs (shown to the right of the last dot) across gallery size are illustrated with dotted lines. Mean EER does not change as a function of gallery size.

5.2. Other Points on the ROC-curve

Since the EER was so stable across gallery size, we thought it important to check other points on the ROC-curve. In Figure 8, (Top), for synthetic data, we present the false rejection rater (FRR) when the false acceptance rate (FAR) = 0.01% for 5, 10, 15 and 20 features. Note the stability across gallery size.

In Figure 8, (Bottom), for face recognition data, we present similar results for $FRR@FAR = 0.001\%$ for our face recognition data set. We analyzed 5, 10, 15, and 20 PCA components. This error rate was quite stable across levels of gallery size.

Neither EER or $FRR@FAR = x\%$ metrics appear to change systematically with gallery size. Therefore, we hypothesized that the entire ROC-curve was also not changing with gallery size. This was tested below.

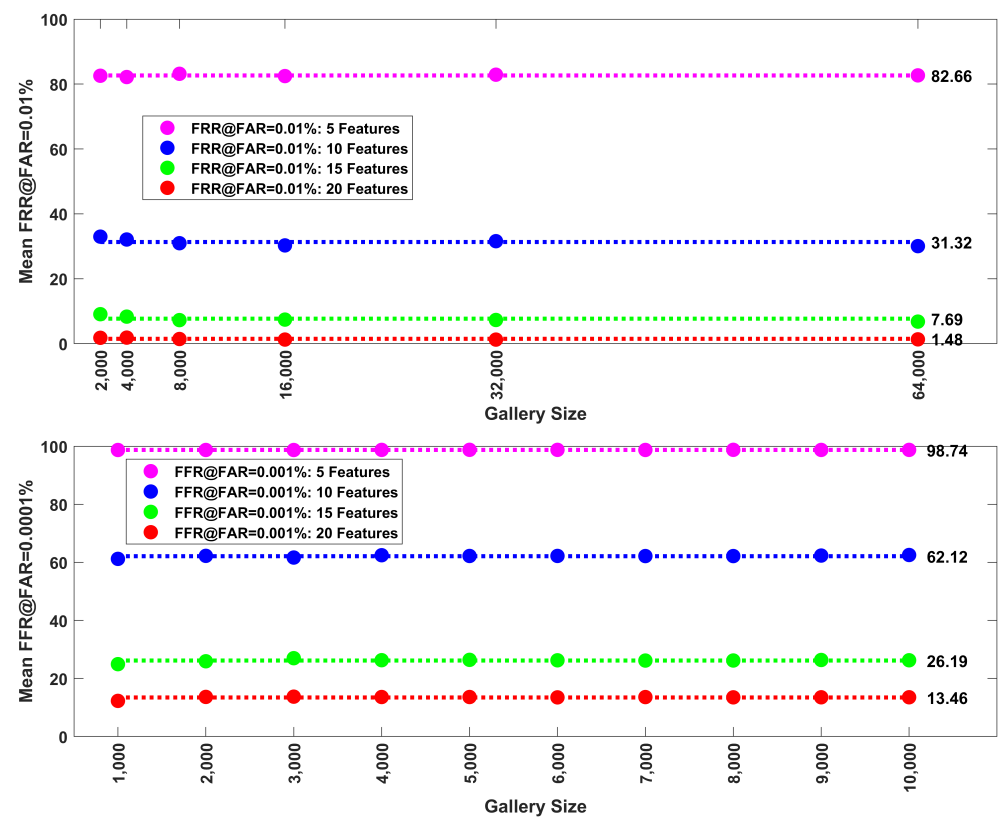


Figure 8. (Top): In addition to mean EER, we also evaluated the mean false rejection rate (FRR) at a false positive rate of 0.01%. Here we plot these error rates (mean of 30 repetitions) versus gallery size (1000, 2000, 4000, 8000, 16,000, 32,000, 64,000 subjects) for synthetic data (Band 8). Data shown for 5, 10, 15 and 20 features. The average error rates (shown to the right of the last dot) across gallery size are illustrated with dotted lines. There was some instability at small gallery sizes, but otherwise, the mean error rate did not change as a function of gallery size. **(Bottom):** FRR at FAR = 0.0001%. Here we plot these error rates (mean of 30 repetitions) versus gallery size (1000 to 10,000 in steps of 1000 subjects) for FaceNet PCA components. Data shown for 5, 10, 15 and 20 PCA components. The average error rates (shown to the right of the last dot) across gallery size are illustrated with dotted lines. The mean error rate did not change as a function of gallery size.

5.3. ROC-curves

Here we calculate and plot entire ROC-curves for several numbers of features and gallery sizes. In Figure 9 we plot the ROC-curves for the synthetic data. In Figure 10 we plot the ROC-curves for the FaceNet data. These ROC-curves facilitate the comparison between entire curves as a function of gallery size.

As a general matter, entire ROC-curves do not appear to change as a function of gallery size. Since the ROC-curves are based on the characteristics of the genuine and similarity score distributions, it makes sense that the ROC-curves do not change. With increases in gallery size, the distributions will become more well defined, and the median and inter-quartile range (IQR) should become increasingly stable. However, there is no basis for predicting a change in the central tendency or spread of these distributions as gallery size increases.

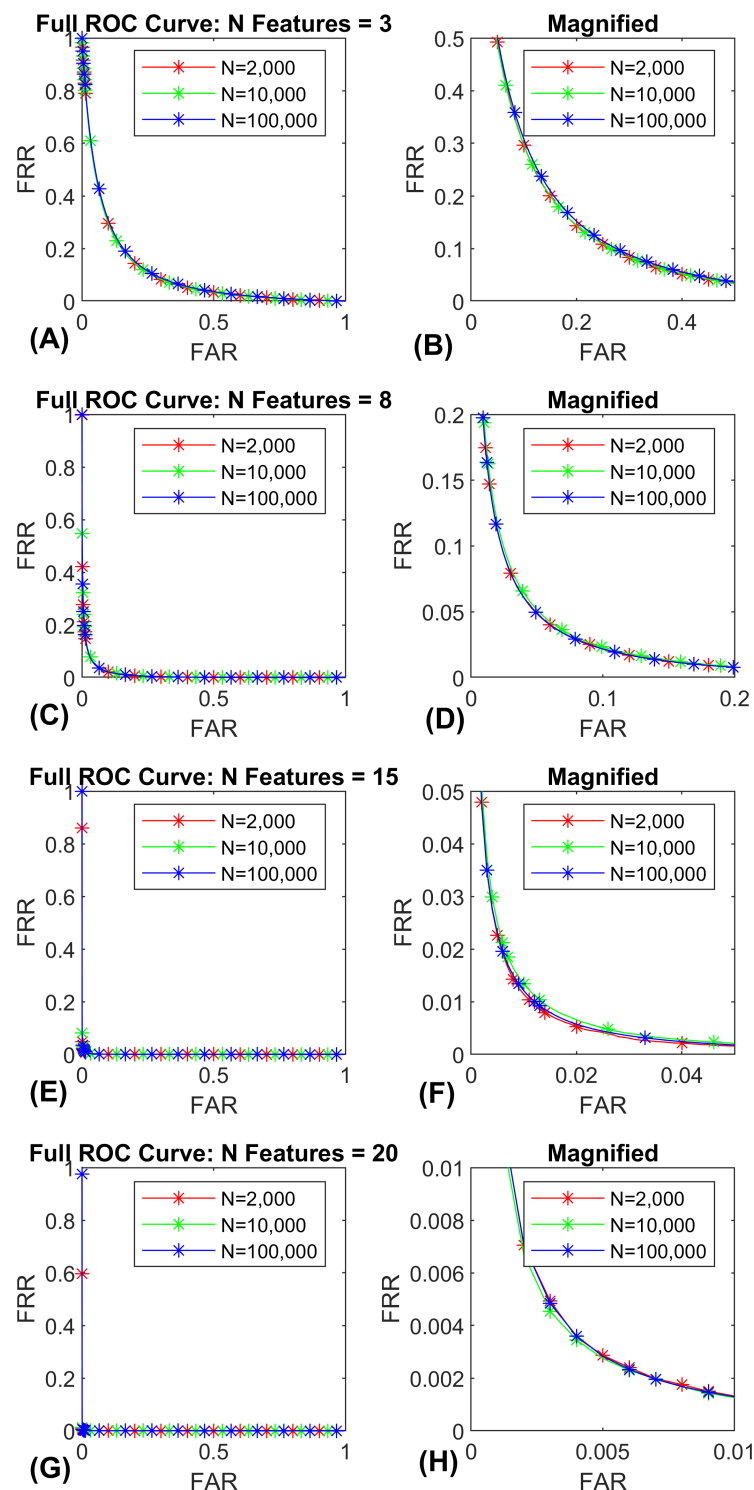


Figure 9. ROC-curves for synthetic data (Band 8). Curves for 3, 8, 15 and 20 features. Each curve was evaluated at gallery sizes of 2000, 10,000 and 100,000 subjects. Each curve is the average over 30 repetitions. (A) illustrates the entire 3 ROC-curves for 3 features. (B) illustrates the same data in (A) zoomed in on the lower error rates for enhanced visibility. (C) illustrates the entire 3 ROC-curves for 8 features. (D) illustrates the same data in (C) zoomed in on the lower error rates for enhanced visibility. (E) illustrates the entire 3 ROC-curves for 15 features. (F) illustrates the same data in (E) zoomed in on the lower error rates for enhanced visibility. (G) illustrates the entire 3 ROC-curves for 20 features. (H) illustrates the same data in (G) zoomed in on the lower error rates for enhanced visibility. Note that the ROC-curves are essentially overlapping across gallery size.

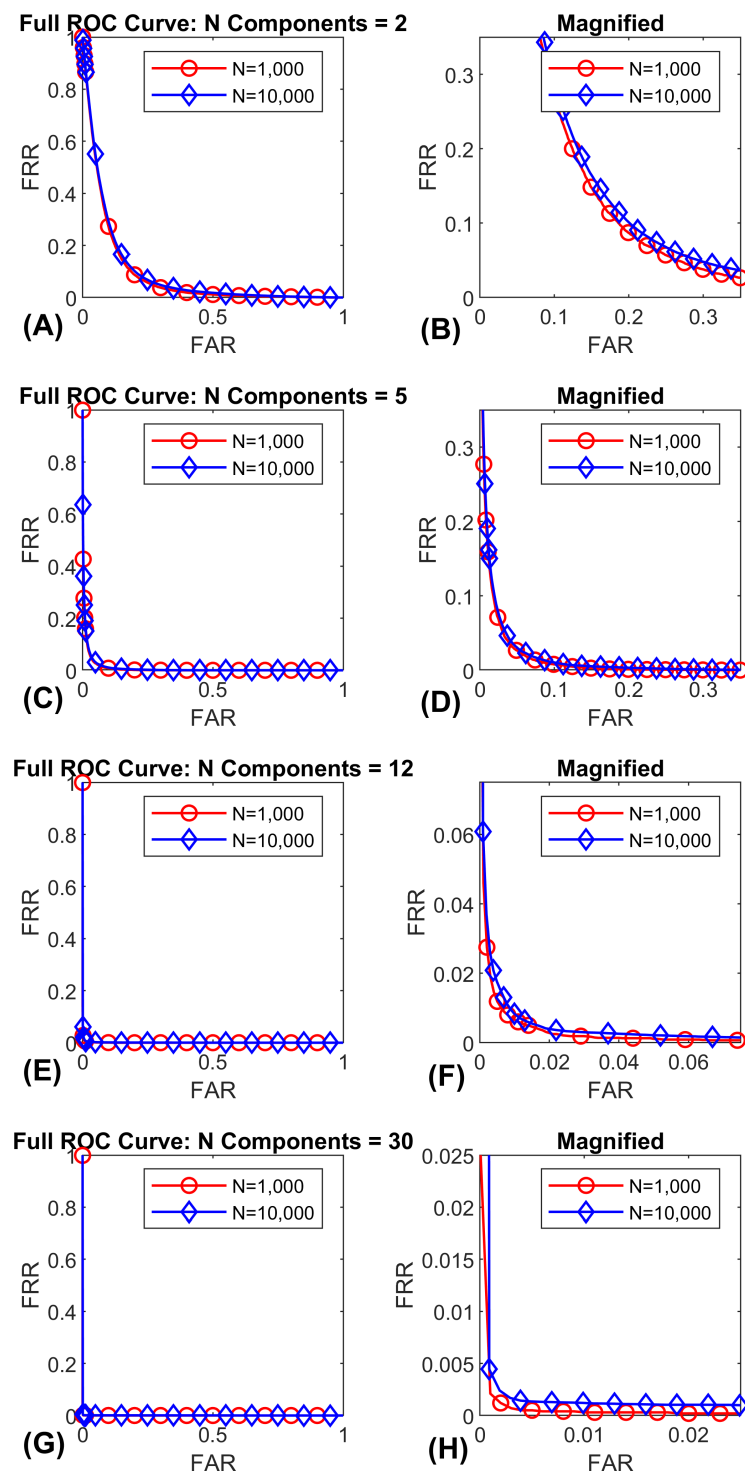


Figure 10. ROC-curves for FaceNet PCA components. Plots for 2, 5, 12 and 30 PCA components are displayed. Each curve was evaluated at 1000 and 10,000 subjects. Each curve is the average over 30 repetitions. (A) illustrates the entire 2 ROC-curves for 2 features. (B) illustrates the same data in (A) zoomed in on the lower error rates for enhanced visibility. (C) illustrates the entire 2 ROC-curves for 5 features. (D) illustrates the same data in (C) zoomed in on the lower error rates for enhanced visibility. (E) illustrates the entire 2 ROC-curves for 12 features. (F) illustrates the same data in (E) zoomed in on the lower error rates for enhanced visibility. (G) illustrates the entire 2 ROC-curves for 30 features. (H) illustrates the same data in (G) zoomed in on the lower error rates for enhanced visibility. Note that the ROC-curves are essentially overlapping across gallery size.

5.4. Similarity Score Distribution Metrics

ROC-curves are based on genuine and impostor similarity score distributions. We were interested in evaluating changes in similarity score distributions as a function of gallery size. To this end, we evaluated the median and IQR of both genuine and impostor distributions as a function of number of features and gallery size. The results for synthetic data are presented in Figure 11 and for FaceNet data in Figure 12. All dots are means and each mean is associated with an SD error bar. Often, the SD was so small that it was not visible. For synthetic data, this was especially the case for the impostor IQR (Figure 11D). The effect of number of features was extremely strong with no overlap. Of these 4 plots, this statement was only true for Figure 11D. In Figure 11A, note the very narrow range, and the large degree of overlap of medians from the genuine distributions. In Figure 11C, again note the very narrow range, and the large degree of overlap of medians from the impostor distributions. The pattern in Figure 11B resembles that in Figure 11D, but the effect was relatively weak. From this analysis we conclude that, as features are added and biometric performance improves, it was mostly the change in the spread of the impostor distribution that was driving the biometric results, at least for synthetic data.

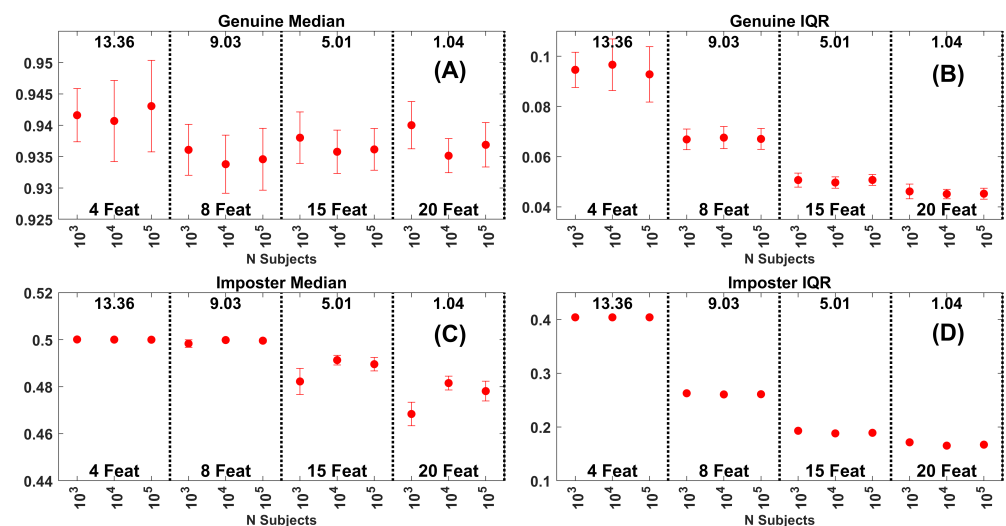


Figure 11. (A) Median of the genuine similarity score distributions for synthetic data. Data are for the Band 8 data set with 4, 8, 15 and 20 features. Each dot is based on 30 repetitions. The error bars are at ± 1 SD. The numbers at the top of each plot are the average EERs across the gallery sizes represented. (B) IQR of the genuine similarity score distributions for the same data in (A). (C) Median of the impostor similarity score distributions. (D) IQR of the impostor similarity score distributions.

The results for the FaceNet data are presented in Figure 12. For each plot, the dot represents the mean and the SD error bars are also plotted. Often, the SD was so small that it was not visible. There are many more invisible error bars in Figure 12 than in Figure 11. Although there was very little overlap in all the plots, notice the range in Figure 11D and compare it to the other ranges. Clearly the effect of number of features on the impostor IQR was remarkable compared to all the other metrics. From this analysis we conclude that, as features are added and biometric performance improves, it was mostly the change in the spread of the impostor distribution that is driving the biometric results for the FaceNet data.

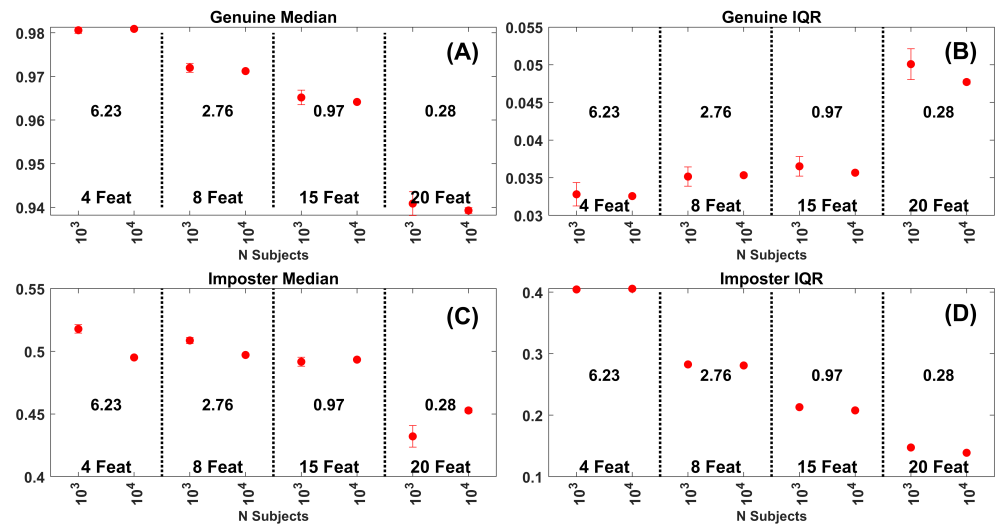


Figure 12. (A) Median of the genuine similarity score distributions for the FaceNet PCA Components. Results for 4, 8, 15 and 20 PCA components. Each dot is based on 30 repetitions. The error bars are at ± 1 SD. The numbers in the middle of each plot are the average EERs across the gallery sizes represented. (B) IQR of the genuine similarity score distributions for the same data in (A). (C) Median of the impostor similarity score distributions. (D) IQR of the impostor similarity score distributions.

To compare the contribution of each distribution metric (median and IQR) to EER, we performed a stepwise linear regression. The dependent variable was the mean EER for all features sets. The independent variables were the medians and IQRs of the two distributions. We did this analysis for both synthetic and face recognition data sets. For both data sets, the first independent variable entered into the model was for the impostor IQR (synthetic: $p = 2 \times 10^{-15}$, $r^2 = 0.998$; real: $p = 1.8 \times 10^{-6}$, $r^2 = 0.979$). This is consistent with the idea that the decrease in impostor IQR with increasing features was the driving force behind the lower EERs obtained.

Note that the particular patterns we see in Figures 11 and 12 may be related to the distance metric (cosine) that we have employed. In the future, we hope to evaluate the role of distance metric and other design elements on the pattern of these similarity score medians and IQRs as performance improves.

5.5. Similarity Score Distributions

In Figure 13 (synthetic) and Figure 14, (FaceNet) we present histograms of both the impostor and genuine distributions based on synthetic data with either 4, 8, 15 or 20 features and 10,000 subjects. Focusing on the spread of the impostor distributions in both figures one can clearly notice a marked decrease in the spread of the impostor distributions as more features are added and biometric performance improves. As features were added, the spread of the genuine distributions also decreased somewhat for synthetic data, but changed very slightly for the FaceNet data. The medians of the two distributions vary only slightly. This is consistent with our conclusion from the previous section that the spread of the similarity score distributions was the driving factor leading to improved biometric performance, especially for the spread of the impostor distribution.

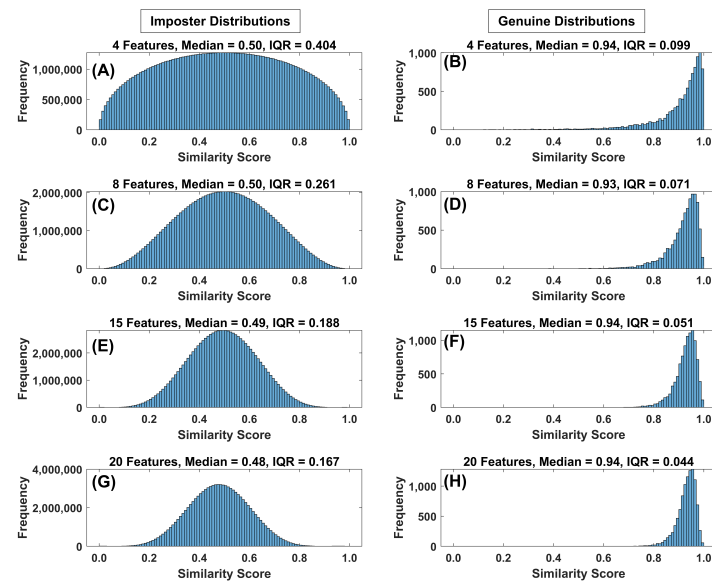


Figure 13. (A) Histogram of the impostor similarity score distribution for synthetic data (4 features, Band 8 (all histograms), 10,000 subjects (all histograms)). Note the median and IQR value just above the histogram. (B) Histogram of the genuine similarity score distribution for 4 features. (C) Histogram of the impostor similarity score distribution for 8 features. (D) Histogram of the genuine similarity score distribution for 8 features. (E) Histogram of the impostor similarity score distribution for 15 features. (F) Histogram of the genuine similarity score distribution for 15 features. (G) Histogram of the impostor similarity score distribution for 20 features. (H) Histogram of the genuine similarity score distribution for 20 features.

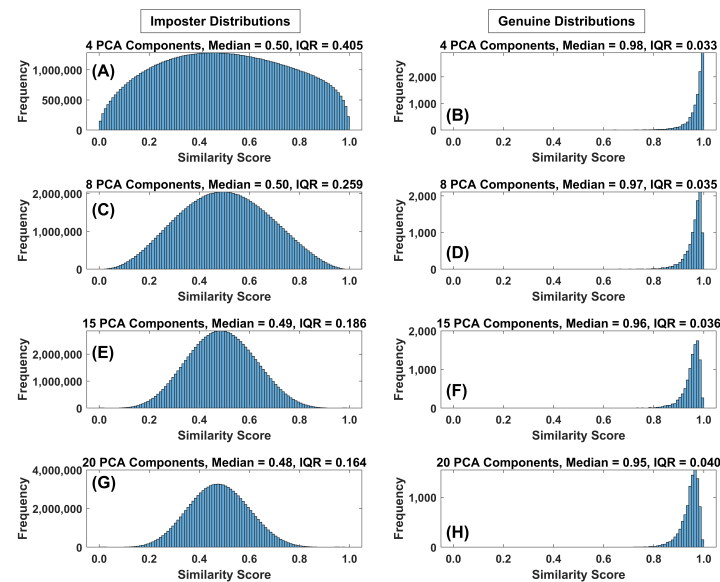


Figure 14. (A) Histogram of the impostor similarity score distribution for FaceNet PCA components (4 PCA components, 10,000 subjects (all histograms)). Note the median and IQR value just above the histogram. (B) Histogram of the genuine similarity score distribution for 4 PCA components. (C) Histogram of the impostor similarity score distribution for 8 PCA components. (D) Histogram of the genuine similarity score distribution for 8 PCA components. (E) Histogram of the impostor similarity score distribution for 15 PCA components. (F) Histogram of the genuine similarity score distribution for 15 PCA components. (G) Histogram of the impostor similarity score distribution for 20 PCA components. (H) Histogram of the genuine similarity score distribution for 20 PCA components.

5.6. Summary of Results

See Table 1.

Table 1. Summary of Major Results.

Rank-1 IR decreases in a log-linear fashion with increasing gallery sizes
EER was stable across gallery sizes.
The false rejection rate at several false positive rates was also stable across gallery sizes.
The shape of the ROC-curves was not affected by the gallery size.
The largest change among the 4 metrics (median genuine, IQR genuine, median impostor and IQR impostor), as biometric performance improves, was for the IQR of the impostor similarity score distribution.

6. Discussion

In this report, we have replicated and extended findings regarding the effects of gallery size on biometric performance. We have confirmed the finding that Rank-1 IR declines as a function of gallery size. We have also confirmed that this relationship was linear in $\log(\text{Gallery Size})$. We have also shown that this decline can be counteracted with the inclusion of additional information (features) for larger gallery sizes. We have also described the curves which can be used to predict how much additional information is required to stabilize the Rank-1 IR as a function of gallery size. These equations are also linear in $\log(\text{gallery size})$.

It is important to note that our findings are based on datasets with certain characteristics. In particular we are using either features or PCA components that are reasonably normally distributed. Furthermore, we are using features or PCA components that are either completely uncorrelated or nearly completely uncorrelated. Any data set with approximately normal features that are more strongly intercorrelated can be transformed into a completely uncorrelated data set using the inverse Cholesky transformation [10] (See “Decorrelation of n random variables” at https://en.wikipedia.org/wiki/Pearson_correlation_coefficient (accessed on 20 October 2022)).

We have also shown that ROC-curves are not systematically affected by gallery size, and so ROC-based scalar performance metrics such as EER are also stable across gallery size. We have illustrated how changes in similarity score distribution characteristics (median and IQR) change as additional features are added to the analysis. The most important predictor of change in EER as additional features are added was in the spread of the impostor similarity score distribution, which becomes narrower as additional features are added. This particular pattern of changes may be related to the distance metric chosen (cosine) or other aspects of the design. The topic of changes in similarity score distributions that accompany changes in biometric performance, for different approaches, will be addressed in future work. These findings should be of interest in the abstract theoretical sense, and should be of real practical value to the biometric community, when planning biometric studies with various gallery sizes.

The ROC-based measures are based on the central tendency and spread of the genuine and similarity score distributions. Although measures of central tendency and spread will become more stable with increasing gallery size, there is no basis for predicting a systematic change in the central tendency or spread of these distributions with increasing gallery size.

7. Conclusions

Gallery size strongly impacts the Rank-1 IR. The decline in Rank-1 IR is linear in $\log(\text{gallery size})$. Biometric performance characteristics based on a ROC analysis (EER, FRR @ FPR = 0.01, ROC shape) are not affected by gallery size. Increases in gallery size compress the confidence limits around an ROC-curve but do not alter its shape. For a given

analysis, as gallery size increases, the imposter and genuine similarity score distributions become smoother. However, there is no theoretical reason to expect, or empirical evidence to demonstrate, that their central tendency changes with increasing gallery size.

Author Contributions: Conceptualization, L.F.; Data curation, L.F. and S.D.; Formal analysis, H.S.; Methodology, L.F., H.S., V.P., S.D. and H.G.; Project administration, O.K.; Resources, O.K.; Software, L.F.; Supervision, O.K. and H.S.; Writing—original draft, L.F.; Writing—review & editing, L.F., H.S., V.P., S.D., H.G. and O.K. All authors have read and agreed to the published version of the manuscript.

Funding: The study was funded by 3 grants to Komogortsev: (1) National Science Foundation, CNS-1250718 and CNS-1714623, www.NSF.gov; (2) National Institute of Standards and Technology, 60NANB15D325, www.NIST.gov; (3) National Institute of Standards and Technology, 60NANB16D293. Stern's work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

Acknowledgments: We wish to acknowledge the assistance of Dillon J Lohr, a doctoral student in our group. He was immensely helpful in preparing the L^AT_EX version of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wilson, C.; Watson, C.; Garris, M.; Hicklin, R. *Studies of Fingerprint Matching Using the NIST Verification Test Bed (VTB)*; NIST Interagency/Internal Report (NISTIR); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2003. Available online: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=50783 (accessed on 20 October 2022).
2. Baveja, M.; Yuan, H.; Wein, L.M. Asymptotic Biometric Analysis for Large Gallery Sizes. *IEEE Trans. Inf. Forensics Secur.* **2010**, *5*, 955–964. [[CrossRef](#)]
3. Grother, P.P.; Phillips, P.J. Models of large population recognition performance. In Proceedings of the 2004 IEEE Computer Society Conference Computer Vision and Pattern Recognition (CVPR'04), Washington, DC, USA, 27 June–2 July 2004; pp. 68–77. [[CrossRef](#)]
4. Dass, S.C.; Yongfang, Z.; Jain, A.K. Validating a Biometric Authentication System: Sample Size Requirements. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1902–319. [[CrossRef](#)] [[PubMed](#)]
5. Schuckers, M.E. *Computational Methods in Biometric Authentication*, 1st ed.; Springer: London, UK, 2010; p. 317. [[CrossRef](#)]
6. Ricanek, K.; Tesafaye, T. Morph: A longitudinal image database of normal adult age-progression. In Proceedings of the IEEE 7th International Conference on Automatic Face and Gesture Recognition, Southampton, UK, 10–12 April 2006; pp. 341–345.
7. Friedman, L.; Nixon, M.S.; Komogortsev, O.V. Method to assess the temporal persistence of potential biometric features: Application to oculomotor, gait, face and brain structure databases. *PLoS ONE* **2017**, *12*, e0178501. doi: 10.1371/journal.pone.0178501. [[CrossRef](#)]
8. Friedman, L.; Stern, H.S.; Price, L.R.; Komogortsev, O.V. Why Temporal Persistence of Biometric Features, as Assessed by the Intraclass Correlation Coefficient, Is So Valuable for Classification Performance. *Sensors* **2020**, *16*, 4555. [[CrossRef](#)]
9. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [[CrossRef](#)]
10. Kessy, A.; Lewin, A.; Strimmer, K. Optimal Whitening and Decorrelation. *Am. Stat.* **2018**, *72*, 309–314. [[CrossRef](#)]