Appeared in Proc. of IEEE International Workshop on Information Forensics and Security (WIFS), (Tenerife, Spain), December 2012.

Can a "Poor" Verification System be a "Good" Identification System? A Preliminary Study

Brian DeCann and Arun Ross
West Virginia University
bdecann@mix.wvu.edu arun.ross@mail.wvu.edu

Abstract

The matching accuracy of a biometric system is typically quantified through measures such as the False Match Rate (FMR), False Non-match Rate (FNMR), Equal Error Rate (EER), Receiver Operating Characteristic (ROC) curve and Cumulative Match Characteristic (CMC) curve. In this work, we analyze the relationship between the ROC and CMC curves, which are two measures commonly used to describe the performance of verification and identification systems, respectively. We establish that it is possible for a biometric system to exhibit "good" verification performance and "poor" identification performance (and vice versa) by demonstrating the conditions required to produce such outcomes. Experimental analysis using synthetically generated match scores confirms our hypothesis that the ROC or CMC alone cannot completely characterize biometric system performance.

1. Introduction

Biometrics is the science of establishing human identity based on the physical or behavioral traits of an individual. Examples of these traits include face, fingerprint, iris, hand geometry, voice, and gait [8, 9]. A biometric system typically operates in either *verification* mode or *identification* mode [9]. In verification, the probe biometric data is labeled with a claimed identity. The system, therefore, compares the probe data *strictly* with similarly labeled templates in the gallery. In essence, the system is *verifying* the users claim of identity. This sort of matching is also referred as 1:1 matching, as a probe is compared against a single (or relatively small) number of gallery entries.

In identification, the probe biometric data is *not* labeled with any identity. Therefore, in order to determine the identity of the probe, the system compares the features extracted from the input biometric data to *every* template in the gallery. In doing so, the system is attempting to *identify* the identity of the probe. This type of matching operation is also referred as $1{:}N$ matching, as N comparisons, equal to the size of the gallery are performed during the matching operation.

1.1. Measuring Biometric System Performance

In general, a biometric system incurs error by either incorrectly matching biometric data from two different individuals, or incorrectly not matching biometric data from the same individual. These errors can be defined as a *false match* and *false non-match*, respectively and can manifest into a biometric system operating in either verification or identification mode. Since a biometric system is prone to errors, it is necessary to develop techniques capable of quantifying the performance of a biometric system. In order to facilitate this, a test database of biometric templates is required. Assume that N_G templates per identity are available for N individuals. Denote the total number of templates as N_T (i.e., $N_T = N \cdot N_G$). By comparing each of the N_T templates against the remaining $N_T - 1$ templates, a total of $N_T(N_T - 1)$ match scores can be generated. Assuming the match score between the i^{th} and j^{th} templates is equal when either template represents the probe or gallery element, the number of distinct match scores is reduced to $N_T*(N_T-1)/2$. Such a procedure is defined as performing an "all-to-all" match test.

A distribution of match scores, $F(x)$, consists of two subsets: the genuine score distribution and the impostor score distribution. The genuine match score distribution, denoted by $F_G(x)$, represents the scores generated when matching two templates belonging to the same identity. The impostor match score distribution, denoted by $F_I(x)$, represents the scores generated when matching templates belong to different identities. In addition to serving as a visual aid of the separability of match scores, the distributions $F_G(x)$ and $F_I(x)$ can be used to derive the false match rate (FMR) and false non-match rate (FNMR). Mathematically, FMR is defined as the integral of $F_I(x)$ for $x \in [t, \infty)$. Similarly, FNMR is defined as the integral of $F_G(x)$ for $x \in (-\infty, t - \epsilon]$.

Traditionally, verification performance is assessed via

the Receiver Operating Characteristic (ROC) curve. The ROC is defined as a graphical plot of FMR versus 1-FNMR (defined as GAR, the genuine accept rate) for the range of threshold t. The ROC itself has been extensively studied in the literature. Hanley and McNeil demonstrated that for a two-class problem, the area underneath the ROC (denoted by AUC) represents the probability that randomly selected data from both classes can be correctly classified. [6]. Martin *et. al.* defined the Detection Error Tradeoff (DET) curve, as a variant of the ROC curve [10]. The DET curve plots the false non-match rate versus the false match rate, visualizing the tradeoff between observing both types of errors. Green and Swets also define the d' metric, which similar to the AUC, attempts to qualitatively measure the ROC using a single number [4].

The performance of an identification system is typically summarized through the Cumulative Match Characteristic (CMC) curve [5, 12, 13]. The CMC presents the probability that the matching algorithm will return the correct identity within the top K ($K \leq N$) ranks. As with the ROC, the CMC is estimated from the *same* set of match scores. Unlike the ROC, the CMC is not estimated from the distribution of match scores per se, but rather through a sorting process wherein the best matching identities for each input probe is determined from the gallery templates.

1.2. Relationship Between the ROC and CMC

If the ROC and CMC curves are estimated from the same set of match scores, it is not unreasonable to expect some degree of correlation between the two curves. One could argue that if the ROC curve demonstrates reasonable matching performance, it is likely that the CMC curve should as well. However, few researchers have fully investigated this relationship. Phillips *et. al.* first developed a measure for estimating the CMC curve from the ROC curve [12]. The measure was found to consistently underestimate the values of an experimentally derived CMC. Later, Bolle *et. al.* [1] explored this possibility and argued that the CMC is directly related to the ROC and can be used to deduce the performance of a $1:1$ verification system. Bolle *et. al.* also argued that for the special case of $N_G = 2$, the CMC curve can be estimated directly from the ROC curve. Similarly, Hube [7] also argued that the ROC curve can be used to provide an estimate of the CMC curve and provides a method for its estimation. Both of these measures estimated the CMC curve using direct knowledge of the ROC curve. However, these techniques carried an implicit assumption that the gallery consisted of one entry per identity (i.e. $N_G = 2$).

In reality, the assumption that $N_G = 2$ may be overly strict. For example, while enrolling new users into a biometric system, it is common to extract multiple templates. Alternatively, individuals may be enrolled into the system multiple times. In Figure 1, the models from Bolle *et. al.* and Hube are compared against two different sets of match scores computed from different matching algorithms. In addition, the parameter N_G is also varied between $N_G = 2$ and $N_G > 2$, wherein the additional genuine templates are stored in the gallery. The first set of match scores represents gait scores extracted from a gait recognition algorithm [3] on the CASIA B dataset [15]. Gait scores were obtained with $N = 115$ and $N_G = 2, 3, 4, 6$. The second set of match scores are fingerprint (left-index) scores from the WVU Multimodal Dataset [2]. These scores were generated using Verifinger, a commercial fingerprint algorithm. Fingerprint scores were obtained with $N = 240$ and $N_G = 2, 3, 4$. Note that the intent of Figure 1 is *not* to show the performance values of the CMC curves, but rather the ability of the CMC prediction models to predict the experimentally obtained CMC curve. The data in Figure 1 suggests the measures from Bolle *et. al.* or Hube may not necessarily deduce accurate estimations of the CMC curve, even when $N_G = 2$.

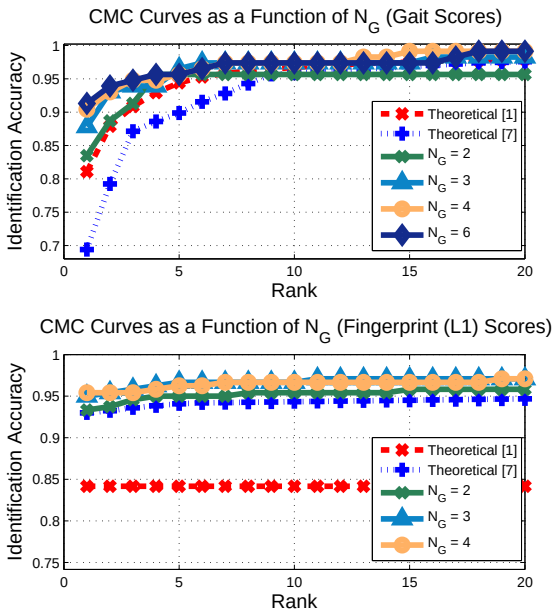

Figure 1. Example demonstrating the CMC prediction models by Bolle *et. al.* [1] and Hube [7] on match scores obtained from a gait matcher [3] (top), and match scores obtained from Verifinger, a fingerprint matcher (bottom). Note that neither model perfectly predicts the CMC curve for both sets of match scores.

Although these models demonstrate there may be some degree of correlation between the ROC curve and CMC curve, it may be prudent to suggest that an ROC curve exhibiting good verification performance (assessed using EER, d', AUC, etc.) infers a corresponding CMC curve with a high rank-1 accuracy. Similarly, it is not been demon-

Appeared in Proc. of IEEE International Workshop on Information Forensics and Security (WIFS), (Tenerife, Spain), December 2012.

strated in the literature that an ROC curve demonstrating poor verification performance infers a CMC curve with a poor rank-1 accuracy. *Thus, we raise the following question: Is it possible that a biometric system performing well at the verification task may not be adept at the identification task, and vice versa?* In this work we aim to further explore the relationship between the ROC and CMC curves, and demonstrate that the aforementioned scenarios are possible, while determining the conditions under which they can occur. We expand upon the current understanding by illustrating how the *global* statistics of a set of match scores may contribute to the relationship between the CMC and ROC curves.

2. Outcomes of a Performance Test

While the ROC and CMC curves in their entirety can be used to assess the matching performance of a biometric system, it is not uncommon to use single-value measures to summarize each of these curves. For example, the rank-1 accuracy exemplifies such a measure from the CMC curve. Since the rank-1 accuracy refers to the probability the maximum match score corresponds to the gallery entry with the correct identity, a high rank-1 accuracy is an implication of "good" identification performance. However, rank-1 accuracy does not explain how the system performs from ranks 2 to N, wherein a significant portion of information is also contained. It may be possible that a low rank-1 accuracy is followed by exceptionally high accuracies in ranks 2 to N. Thus, it may be more beneficial to consider a weighted rank-M accuracy, where M is some percentage of N (5-10%).

With regard to the ROC curve, identifying a single-valued measure is slightly more tricky. Since the ROC curve visualizes the tradeoff between the FMR and GAR, it is less clear how to define what constitutes "good" performance. Arguably the best metric in this regard is the area under the ROC curve [6], which measures the probability that a pair of randomly drawn genuine and impostor match scores are classified correctly. If the performance of the ROC and CMC curves can be summarized as being "good" or "poor", a performance test can result in either good or poor verification, and good or poor identification. These outcomes can be summarized as follows:

Good Verification Good Identification (GVGI): A performance test is classified as good verification good identification (GVGI) when the properties of the ROC and CMC indicate excellent performance. Here, the system is adequately able to perform both verification and identification tasks. Such an outcome is perhaps the most desirable for a biometric system.

Good Verification Poor Identification (GVPI): A performance test is classified as good verification poor identification (GVPI) when the properties of the ROC indicate good performance, while the CMC demonstrates a poor rank-M identification accuracy. Such is system is adept at verification tasks, but is generally unreliable at performing identification.

Poor Verification Good Identification (PVGI): A performance test is classified as poor verification good identification (PVGI) when the properties of the ROC indicate poor performance, while the CMC demonstrates a good rank-M identification accuracy. Such a system is capable of performing identification, but not verification.

Poor Verification Poor Identification (PVPI): A performance test is classified as poor verification poor identification (PVPI) when the properties of both the ROC and CMC indicate poor performance. Such a system is not capable of performing verification or identification tasks adequately.

In this work, the performance of a verification system is defined to be "good" if the AUC is above 98% and "poor" if the AUC is below 75%. Identification performance is defined using a weighted rank-M strategy. Here, M is equal to 5% of the number of identities, N. The weight of the i^{th} rank, w_i, $i = 1, 2, \ldots, M$, is defined by $1/i$, and normalized such that $\|\mathbf{w}\|_2 = 1$. The weighted rank-M accuracy is then obtained by the scalar product of $\mathbf{w}$ and $\mathbf{c}$, where c_i denotes the identification accuracy at the i^{th} rank. Identification is then classified as "good" if the weighted rank-M identification accuracy (denoted by CMC(M)) is above 90% and "poor" if the rank-M identification accuracy is below 50%. Note this definition neglects the situation where either the AUC or rank-M identification accuracy is between the ranges defined as "good" and "poor". This two tiered threshold is necessary to suggest a "poor" outcome is sufficiently poor and not "almost good" (or vice versa). These ranges are summarized in Table 1.

Table 1. Range of AUC and rank-M identification rate resulting in a PVPI, PVGI, GVPI and GVGI outcome. Outcomes outside these definitions are denoted by "***".

AUC / Rank-M	0.00-0.50	0.50-0.90	0.90-1.00
0.00-0.75	PVPI	***	PVGI
0.75-0.98	***	***	***
0.98-1.00	GVPI	***	GVGI

3. Conditions Affecting the ROC-CMC Relationship

Consider a genuine score distribution, denoted as $F_G(x)$ and an impostor score distribution, denoted as $F_I(x)$. In the interest of simplicity, assume $F_G(x)$ and $F_I(y)$ can be modeled as a normal distribution with parameters μ_{Gen}, σ^2_{Gen}, μ_{Imp}, and σ^2_{imp}. **Note, this assumption is made *strictly***

to define a means for which hypothetical distributions of match scores can be understood. In general, the probability distributions of $F_G(x)$ and $F_I(x)$ can take on any function and can vary on an algorithmic basis [14].

3.1. Properties of $F_G(x)$ and $F_I(x)$

Our primary interest lies in identifying the parameter space for $F_G(x)$ and $F_I(x)$ that contribute to good and poor verification and identification performances. Define ρ as the set of all possible combinations of μ_{Gen}, σ^2_{Gen}, μ_{Imp}, and σ^2_{imp} such that the probability of generating an impostor score within the range $\mu_{Gen} \pm 2\sigma_{Gen}$ is less than α, where $\alpha \leq 1\%$. For each of these combinations, $F_G(x)$ and $F_I(x)$ will appear visually distinct and the probability of discriminating between the two distributions with a single match threshold is high. By contrast, define θ as the set of all possible combinations of μ_{Gen}, σ^2_{Gen}, μ_{Imp}, and σ^2_{imp} such that the probability of generating an impostor score within the range $\mu_{Gen} \pm 2\sigma_{Gen}$ is β, where $\beta \geq 50\%$. The set β represents the case when $\mu_{Gen} \approx \mu_{Imp}$ and $\sigma^2_{Gen} \approx \sigma^2_{Imp}$, resulting in difficulty in differentiating between $F_G(x)$ and $F_I(x)$.

In a verification task, the user claims an identity, resulting in the retrieval of N_G templates from the gallery ($N_G \ll N$). By our definition in Section 2, verification performance is defined by the AUC. Recall the AUC depends on the probability that a randomly chosen pair of genuine and impostor scores will be correctly classified as genuine and impostor, respectively. Based on this distinction, it is reasonable to suggest that distributions of $F_G(x)$ and $F_I(x)$ in ρ represent "good" verification performance, while distributions in θ represent "poor" verification performance.

Determining the parameters of an identification task is a little more complex as $N_T - 1$ match score comparisons need to be made, wherein the number of genuine and impostor comparisons are $N_G - 1$ and $N_T - N_G$, respectively. By definition, the probability of observing a correct identification (and thus the value of CMC(1)) depends on the probability that *exactly* zero impostor scores are generated with a value exceeding the maximum genuine score. Assuming that each impostor score generated is independent, the probability of misclassification increases with N_T. Thus, while combinations of μ_{Gen}, σ^2_{Gen}, μ_{Imp}, and σ^2_{imp} in ρ might suggest "good" identification performance, for this to occur $N_T \ll 1/\alpha$. As N_T approaches $1/\alpha$, it is increasingly likely *at least* one generated impostor score exceeds the value of a genuine score and misclassification occurs. This condition, therefore, results in a GVPI outcome, and is often cited as a concern for open-set identification in large-scale biometric systems [11].

As previously mentioned, θ defines the set of possible distributions of $F_G(x)$ and $F_I(x)$ exhibiting significant overlap. Now, consider a variant of θ wherein the probability of generating an impostor score within the range $(\mu_{gen} - \sigma_{gen}, \mu_{gen})$ is approximately 1.0. In this case, the probability a randomly generated impostor score is greater than a genuine score approaches 50%. It is easy to see that such a scenario would be undesirable for verification purposes, as the decision threshold would need to be set above μ_{Gen} in order to mitigate false matches, while incurring a high rate of false non-matches. However, this type of distribution may be beneficial for identification purposes. Since the probability of generating an impostor score above μ_{Gen} is $\approx 0.00\%$, approximately half of all genuine scores generated will always exceed the maximum impostor score. This implies that with one genuine gallery entry per identification event, the corresponding rank-1 accuracy will meet or exceed 50%. With additional enrollments into the gallery, the probability of at least one genuine score above the maximum impostor score is approximated by summation of Bernoulli trials (Equation (1)), with $p = 0.5$, and $k = 1$ and $n = N_G$, the number of genuine gallery entities.

$$P(n) = \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

With $N_G = 2$, the probability of a genuine score above μ_{Gen} becomes $\approx 75\%$. At $N_G = 5$, this increases to $\approx 96.9\%$. Thus, should the range of $F_I(x)$ occur below μ_{Gen} the resulting outcome may be a PVGI system.

4. Experimental Results

4.1. Generating Match Scores

Since this work is concerned with eliciting specific conditions which affect the CMC and ROC curves, experiments are conducted primarily using synthetically generated scores. Synthetic scores are preferred in this case as many different distributions for $F_G(x)$ and $F_I(x)$ can be rapidly generated, which is more conducive for identifying trends in match score data. To be consistent with the analysis in Section 3.1, $F_G(x)$ and $F_I(x)$ are sampled from a parametric normal distribution with parameters μ_{Gen}, σ^2_{Gen}, μ_{Imp}, and σ^2_{imp} and normalized between $[0, 1]$. The range for which these parameters are sampled is given in Table 2. Note in generation of these distributions, the intraclass relationships between match scores of individuals are ignored. That is, match scores across synthetic individuals share the same probabilistic function.

Table 2. Parameters used for generating synthetic match scores.

	μ_{Gen}	μ_{Imp}	σ^2_{Gen}	σ^2_{Imp}
Minimum Value	0.5	0.1	0.0001	0.0001
Maximum Value	0.5	0.5	0.2	0.2

4.2. Outcomes From Synthetic Scores

In this experiment, we aim to identify the parametric regions of $F_G(x)$ and $F_I(x)$ that result in each of the four outcomes discussed in Section 2. To enable this, a Monte Carlo analysis is performed wherein $300,000$ distributions of $F_G(x)$ and $F_I(x)$ are randomly generated, as described in Section 4.1. The analysis is performed with $N = 240$ synthetic identities, with $N_G = 2, 3, 4, 5$. For each sampled pair of genuine and impostor distributions, the values for AUC and CMC(M) are obtained. Using these values, the pair of distributions are classified as either GVGI, GVPI, PVGI or "else" (denoting PVPI and the "***" cases) as outlined in Table 1 in Section 2. These results are provided in Table 3. Examples of a GVGI, GVPI, and PVGI outcome are provided in Figures 2-4.

Table 3. Results of synthetic match score generation. Here the probabilities of generating either a GVGI, GVPI, PVGI or "else" outcome are listed. Note that increasing N_G increases the probability of observing a GVGI and PVGI outcome.

	GVGI	GVPI	PVGI	"else"
$N_G = 2$	0.15%	0.04%	0%	99.82%
$N_G = 3$	0.18%	0.04%	0%	99.77%
$N_G = 4$	0.21%	0.04%	0.09%	99.67%
$N_G = 5$	0.25%	0.03%	0.84%	98.89%

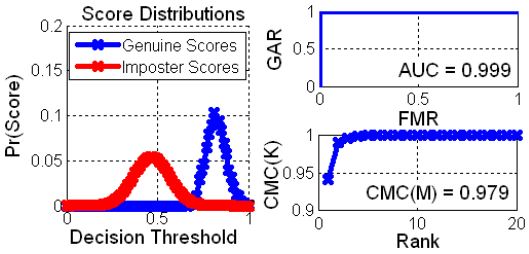

Figure 2. Example of a GVGI result where $N_G = 5$.

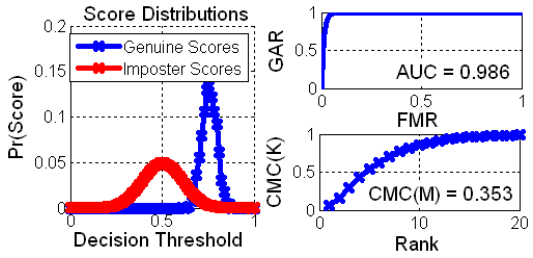

Figure 3. Example of a GVPI result where $N_G = 5$. Here, μ_{Imp} and σ^2_{Gen} are unchanged from Figure 2, while σ^2_{Imp} is increased by a factor of 2.4. The result is a significant decrease in rank-1 identification accuracy.

5. Discussion

In aggregating the results of the monte carlo analysis, the data in Table 3 suggests each of the four performance outcomes can be generated. In addition, the probability of generating a GVGI and PVGI result appears to be directly related to N_G. This suggests that for certain match score distributions, a good or poor result may be observed depending on the number of genuine gallery entries. However, it is difficult to deduce this relationship further from Table 3. Figure 5 visualizes the distribution parameters resulting in GVGI, GVPI and PVGI outcomes for $N_G = 2$ and $N_G = 5$. Note that while the parameters of μ_{Imp}, σ^2_{Gen} and σ^2_{Imp} are generally outcome specific, at $N_G = 5$, GVGI can occur at slightly higher values of σ^2_{Imp} and σ^2_{Gen}. In addition, at $N_G = 2$, PVGI does not occur. At a minimum, Table 3 and Figure 5 verify that each of these outcomes are theoretically possible.

A limitation of the experimental analysis is that each synthetic identity is equally likely to generate both high and low impostor scores with respect to other synthetic identities. In reality, this may not be true since the match scores between groups of identities may be higher or lower on a whole. Additionally, experimentally derived match scores may not necessarily be distributed normally. Although these limitations are acknowledged, they do not undermine the analysis. Figures 2 and 3 highlight that identification performance can be drastically impacted by only a slight change in σ^2_{Imp}. This example indirectly reinforces the notion that as the size of a watchlist grows, the probability of misclassification increases [11]. In the context of a PVGI scenario, while it may be atypical associate good performance with overlapping genuine and impostor score distributions (Figure 4), such distributions are not unrealistic and do occur. For example, soft biometric traits (e.g., gait, ethnicity, height, etc.) are generally more difficult to reliably extract compared to traditional biometric traits (e.g., face, fingerprint, iris). As a result, genuine match scores of a soft biometric system may contain enough discriminative information to separate groups of identities, but not individual identities themselves, resulting in increased visual similarity between the genuine and impostor score distributions. An example of this is visualized in Figure 6 with match scores extracted from a gait recognition algorithm [3], wherein the values of AUC and CMC(M) loosely resemble a PVGI outcome.

5.1. Future Work

Future work will take into consideration the effect of intraclass and interclass relationships between users of a biometric system and its effect on the verification and identification tasks. In particular, it may be possible to model the assignment of match scores to synthetic identities such that both good and poor identification performance can be realized from *the same* global distributions.

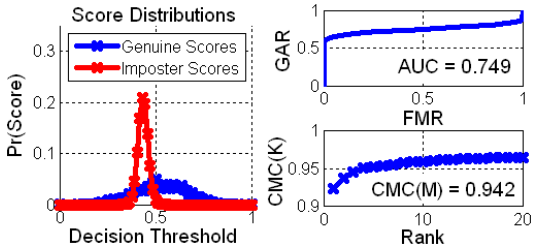

Figure 4. Example of a PVGI result where $N_G = 5$. Note the similarity of the genuine and impostor distributions to those hypothesized to represent this outcome in Section 3.1.

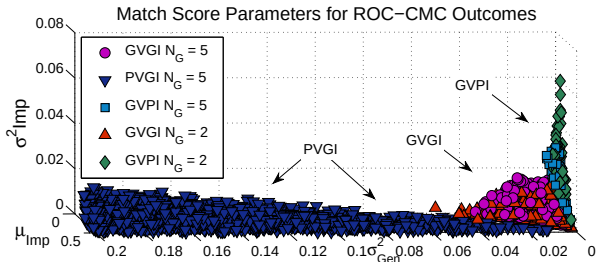

Figure 5. Scatter of the parameters μ_{Imp}, σ^2_{Gen}, and σ^2_{Imp} contributing to GVGI, GVPI, and PVGI outcomes for $N_G = 2$ and $N_G = 5$.

6. Summary

In this study, a preliminary exploration on the relationship between the ROC and CMC curves as they pertain to verification and identification performance in a biometric system is provided. An analysis of previous literature relating the CMC curve to the ROC curve demonstrated that while there may be some relationship between the two curves, it may be possible that a biometric system performing well at verification may not be adept at the identification task. As such, four possible performance outcomes are identified with respect to either good or poor verification and identification performance. Conditions for which each of these outcomes were defined in terms of match scores following a normal distribution, and verified using a Monte Carlo analysis. Finally, an example from the literature is cited which does resemble a poor verification, good identification system, emphasizing that usage of the ROC or CMC alone does not characterize biometric system performance.

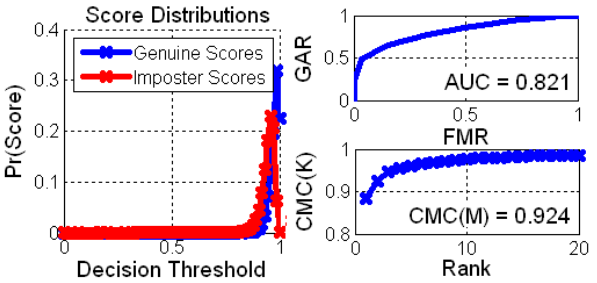

Figure 6. Performance test of a gait recognition algorithm [3] resembling a PVGI outcome.

References

[1] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. The Relation Between the ROC Curve and the CMC. *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 15–20, 2005. 2

[2] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak. A Protocol for Multibiometric Data Acquisition, Storage and Dissemination. Technical report, West Virginia University, 2007. 2

[3] B. DeCann and A. Ross. Gait Curves for Human Identification, Backpack Detection, and Silhouette Correction in a Nighttime Enviornment. *SPIE Conference on Biometric Technology for Human Identification VII*, April 2010. 2, 5, 6

[4] D. Green and J. Swets. *Signal Detection Theory and Psychophysics*. Wiley, 1966. 2

[5] P. Grother and P. Phillips. Models of Large Population Recognition Performance. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:68–75, 2004. 2

[6] J. Hanley and B. McNeil. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143:29–36, 1982. 2, 3

[7] J. Hube. Using Biometric Verification to Estimate Identification Performance. *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research*, pages 1–6, September 2006. 2

[8] A. Jain, P. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2008. 1

[9] A. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, January 2004. 1

[10] A. Martin, G. Doggington, T. Kamm, M. Ordowski, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. *EUROSPEECH*, pages 1895–1898, 1997. 2

[11] J. Pato and L. Millett. *Biometric Recognition: Challenges and Opportunities*. National Academic Press, 2010. 4, 5

[12] P. Phillips, P. Grother, R. Michaels, D. Blackburn, T. Elham, and J. Bone. FRVT 2002: Facial Recognition Vendor Test. Technical report, DoD, April 2003. 2

[13] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation Methodology for Face-recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000. 2

[14] J. Wu and C. Wilson. Nonparametric Analysis of Fingerprint Data on Large Data Sets. *Pattern Recognition*, 40(9):2574–2584, 2007. 4

[15] S. Yu, D. Tan, and T. Tan. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. *Proc. 18th International Conference on Pattern Recognition (ICPR06)*, pages 441–444, August 2006. 2