Appeared in Proc. of 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2013

Relating ROC and CMC Curves via the Biometric Menagerie

Brian DeCann [#1] and Arun Ross [*2]
[#] Lane Department of Computer Science and Electrical Engineering, West Virginia University
[1] bdecann@mix.wvu.edu
[*] Department of Computer Science and Engineering, Michigan State University
[2] rossarun@cse.msu.edu

Abstract

In the academic literature, the matching accuracy of a biometric system is typically quantified through measures such as the Receiver Operating Characteristic (ROC) curve and Cumulative Match Characteristic (CMC) curve. The ROC curve, measuring verification performance, is based on aggregate statistics of match scores corresponding to all biometric samples, while the CMC curve, measuring identification performance, is based on the relative ordering of match scores corresponding to each biometric sample (in closed-set identification). In this study, we determine whether a set of genuine and impostor match scores generated from biometric data can be reassigned to virtual identities, such that the same ROC curve can be accompanied by multiple CMC curves. The reassignment is accomplished by modeling the intra- and inter-class relationships between identities based on the "Doddington Zoo" or "Biometric Menagerie" phenomenon. The outcome of the study suggests that a single ROC curve can be mapped to multiple CMC curves in closed-set identification, and that presentation of a CMC curve should be accompanied by a ROC curve when reporting biometric system performance, in order to better understand the performance of the matcher.

1. Introduction

Biometrics is the science of recognizing humans based on the physical or behavioral traits of an individual. Examples of these traits include face, fingerprint, iris, hand geometry, voice, and gait [11, 12]. A biometric system typically operates in either *verification* mode or *identification* mode [12]. In verification, the probe biometric data is submitted along with a claimed identity. To validate the identity claim, the system compares the probe data *strictly* with similarly labeled identities stored in a reference database. The output of a verification operation is a match or non-match. This sort of matching is also referred as 1:1 matching, as the probe is compared against a single (or relatively small) number of reference entities.

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 probe against *every* reference identity. The output of an identification operation is a sorted list of identities, ordered from the best match to the worst match. This type of matching operation is also referred as 1:N matching, with N being the size of the reference database. The identification operation can be either *closed-set* or *open-set*. In closed-set identification, the identity of the input probe is known to be present in the reference database. However, in open-set identification, the identity corresponding to the probe may or may not be in the reference database.

1.1. Measuring Biometric System Performance

The performance of a biometric matcher, operating in the verification or identification mode, can be evaluated based on the match scores generated from test biometric data. In a set of test data, let N be the number of identities and N_G be the number of biometric samples (e.g., face images) per identity. The total number of samples is N_T (i.e., $N_T = N \cdot N_G$). By comparing each of N_T samples against the remaining $N_T - 1$ samples and assuming a symmetric matcher, a total of $\frac{1}{2}N_T(N_T - 1)$ similarity match scores can be computed. Define this procedure as an "all-to-all" match test. In computing the match scores for an "all-to-all" match test, two classes of match scores are generated: *genuine* match scores and *impostor* match scores. Genuine match scores denote the scores generated when comparing two biometric samples belonging to the same individual. Impostor scores denote the scores generated when matching two biometric samples belonging to different individuals. The total number of genuine and impostor scores that can be computed are $N\binom{N_G}{2}$ and ${N_G}^2\binom{N}{2}$, respectively. Using the generated match scores, a pair of probability density functions regarding the likelihood of observing a genuine or impostor score with a certain value can be estimated. De-

Appeared in Proc. of 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2013

note the genuine and impostor score distributions as $f_G(s)$ and $f_I(s)$, respectively.

Verification performance is typically evaluated by assessing the *false match rate* (FMR) and the *false non-match rate* (FNMR). The FMR denotes the percentage of impostor scores that exceed a numerical threshold t and are incorrectly classified as matches. The FNMR denotes the percentage of genuine scores that are below a threshold t and are incorrectly classified as non-matches. Graphically, the FMR and FNMR are often expressed by a Receiver Operating Characteristic (ROC) curve. The ROC curve plots 1-FNMR versus FMR by varying the threshold t. As such, we refer to FMR, FNMR, and the ROC curve as aggregate-based metrics.

When evaluating identification performance, a set of N_{probe} probe samples is compared against a set of N_{ref} reference samples, resulting in N_{probe} sets of match scores, with each set containing N_{ref} match scores. The match scores in each set are ordered from highest to lowest. In open-set identification, these sets are used to assess the *false positive identification rate* (FPIR) and *true positive identification rate* (TPIR) [8]. The FPIR is defined as the proportion of times a probe that does not have a corresponding reference identity (i.e., no genuine scores were generated), generates an impostor score exceeding the value of a threshold, t. The TPIR is defined as the proportion of times a probe that does have a corresponding reference identity (i.e., genuine scores were generated), the correct identity is observed within the top K ($K \leq N$) ranks (i.e., a genuine score occurs within the top K sorted scores in the set) and whose match score exceeds the value of t.

In closed-set identification, the ordered score sets from the N_{probe} probes are used to estimate the probability that the correct matching identity pertaining to a probe is observed within the top K ($K \leq N$) ranks (i.e., compute the TPIR with $t = 0$). These probabilities are typically expressed visually through the Cumulative Match Characteristic (CMC) curve [13]. Unlike the ROC curve, which is generated by looking at genuine and impostor scores all-at-once, the data in the CMC curve is obtained based on the explicit ordering of $N_G - 1$ and $N_G - N_T$ genuine and impostor scores, respectively, for each biometric probe. As such, we refer to the CMC curve as a *rank-based* metric. An example of both a ROC and CMC curve is presented in Figure 1.

1.2. Closed-set Identification

In general, most biometric identification systems in real-world applications operate in the open-set mode [8]. However, in the literature, most performance *evaluations* are conducted in the closed-set mode [10, 17, 14]. *For the purposes of this study, we therefore focus on the closed-set problem, with the intent of pointing out that reporting only identification accuracy in closed-set evaluations may not be appropriate.*

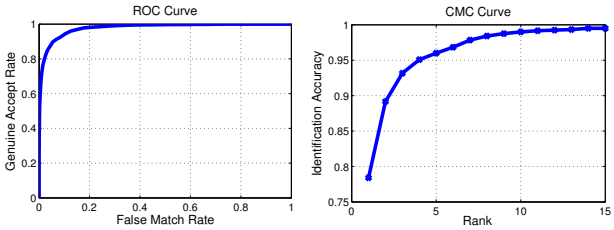

Figure 1. Example of an ROC curve (top) and CMC curve (bottom).

1.3. Relationship Between the ROC and CMC

If the ROC (aggregate-based) and CMC (rank-based) curves are estimated from the same set of match scores, it is not unreasonable to expect some degree of "correlation" between the two curves. This topic has received some attention in the literature, yielding mixed conclusions.

Phillips et. al. [13] first developed a measure for estimating the CMC curve directly from the ROC curve.[1] The measure was found to consistently underestimate the values of an experimentally derived CMC [9]. Later, Bolle et. al. [1] argued that the CMC is directly related to the ROC and can be used to deduce the performance of a 1:1 verification system. Additionally, Bolle et. al. developed a mathematical model for estimating the CMC based on the ROC when $N_G = 2$. Similarly, Hube [9] also argued in favor of a direct relationship between the ROC and CMC, developing a different model for estimating the CMC from the ROC.

In the recent past, however, the notion that the ROC and CMC are directly related has been challenged. Gorodnichy first presented an argument stating that aggregate-based metrics such as the FMR, FNMR, and ROC fail to appropriately evaluate operational systems characterized by large sample size and non-static populations, or systems performing identification at a distance (e.g., systems without a controlled biometric acquisition protocol) [6, 7]. Further, Gorodnichy argues that verification systems should be evaluated (and developed) as 1:N identification systems [7], stating that measures for identification (i.e., *ranked* statistics) reveal more information regarding the relationships between users involved in a biometric system. DeCann and Ross present a case arguing that it is theoretically possible to observe a "poor" ROC curve and a "good" CMC curve (and vice-versa) from the same set of match scores [4].

Based on the conclusions drawn from Bolle et. al. [1], Hube [9], Gordnichy [6, 7], and DeCann and Ross [4], it is clear that support in the literature for a *direct* relationship between the ROC and CMC curves is mixed. In Fig-

[1] In this article, the terms "CMC curve" and "ROC curve" will be interchangeably used with the terms "CMC" and "ROC", respectively

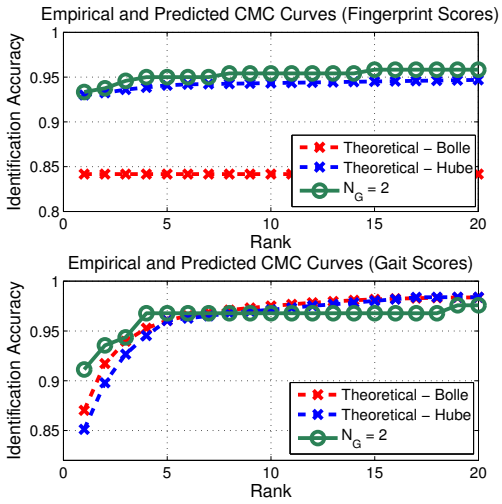

Figure 2. Output of the CMC prediction models (from ROC curves) by Bolle *et. al.* [1] and Hube [9] on match scores obtained from a fingerprint matcher (top), and a gait matcher [3] (bottom). Note that neither model perfectly predicts the CMC curve for both sets of match scores.

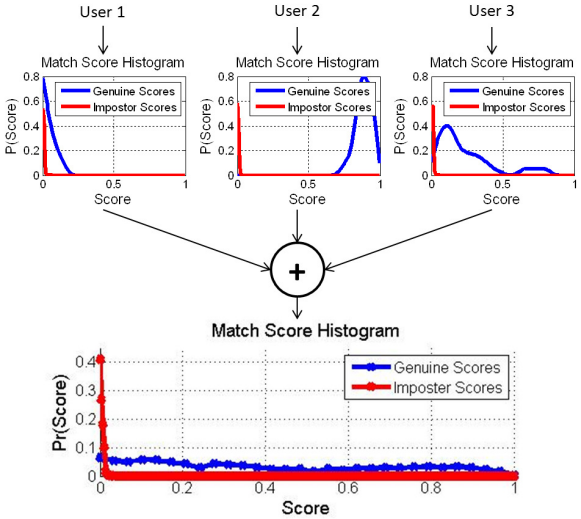

Figure 3. Visual example depicting the contribution of *individual identities* towards the overall genuine and impostor match score distributions, $f_G(s)$ and $f_I(s)$. Note that genuine and impostor score distributions corresponding to an identity may be *distinct* (above) and the *aggregation* of these individual distributions comprises the global genuine and impostor match score distributions (below). Here, the individual match score distributions are based on fingerprint scores computed on the WVU Multimodal Dataset [2].

ure 2, the CMC prediction models of Bolle et. al [1]. and Hube [9] are compared on two different sets of match scores generated by two different matching algorithms. The first set of match scores represents gait scores generated using a gait recognition algorithm [3] on the CASIA B dataset [19]. Here, $N = 124$ and $N_G = 2$. The second set of match scores are fingerprint (left-index) scores from the WVU Multimodal Dataset [2]. These scores were generated using Verifinger,[2] a commercial fingerprint matcher. Here, $N = 240$ and $N_G = 2$. Note that the intent of Figure 2 is *not* to show the performance of the matchers, but rather to analyze the ability of the two models to predict the empirically obtained CMC curve. The data in Figure 2 suggests the prediction models of Bolle et. al. and Hube do not accurately estimate the CMC curve in all cases.

Although the data in Figure 2 demonstrates that there may be some degree of "correlation" between the ROC curve and CMC curve, it is clear that neither model completely predicted the empirical CMC curve based solely on the ROC data. One reason this might be the case is that aggregate-based statistics do not account for the unique manner in which different individuals contribute towards the overall performance of a biometric system. In other words, the genuine and impostor score distributions pertaining to two different individuals can be significantly different. Such differences cannot be captured in aggregate statistics. Visually, this is depicted in Figure 3, where a subset of three individual genuine and impostor score distributions are shown using the left-index (L1) match scores from the WVU Multimodal Dataset [2]. Note that each of the three genuine and impostor distributions are different from one another, and that the accumulation of these subsets result in the aggregate distributions, $f_G(s)$ and $f_I(s)$.

Doddington et. al. [5] first discussed the notion that different identities contribute differently towards overall biometric system performance by introducing a scheme to classify identities based on their propensity to generate a false match or false non-match error in speaker recognition [5]. This observation is referred to as the *Biometric Menagerie* in the literature [18]. If each identity contributes to the performance of a biometric system differently, it may be possible that for a *single* pair of genuine and impostor match score distributions, *multiple* rank-based statistics (e.g., CMC curves) can be generated. Further, these differences in rank-based statistics may result in multiple CMC curves with large differences in cumulative rank-K accuracy.

In an earlier study, DeCann and Ross [4] demonstrated that a "poor" ROC curve can produce a "good" CMC curve; however, their analysis did not account for inter- and intra-class relationships (as manifested through the match scores) and did not demonstrate the possibility of associating multiple CMC curves with a single ROC curve. In this study, our aim is to demonstrate this while accounting for such relationships (the role of the Biometric Menagerie). By modeling the inter- and intra-class relationships, it is pos-

[2] http://www.neurotechnology.com/verifinger.html

Appeared in Proc. of 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2013

sible to demonstrate that a fixed set of match scores can be reassigned *differently* among N identities. This reassignment of existing match scores to virtual identities is accomplished by utilizing the "Doddington Zoo" user classification scheme.

Thus, the contributions of this study are as follows:

- Given a set of real match scores pertaining to multiple identities, we describe a method by which the scores can be reassigned to virtual identities such that they describe different types of intra-class and inter-class statistics based on the Doddington Zoo phenomenon.

- Based on this reassignment process, we demonstrate that match scores sharing common aggregate statistics (ROC) can have multiple ranked statistics (CMC's).

2. Match Score Relationships in a Biometric System

The model for characterizing inter- and intra-class relationships operates by assigning real match scores to *virtual* identities. Here, a *virtual identity* is defined as an identity, whose individual genuine and impostor match score distributions, $f_G^n(s)$ and $f_I^n(s)$ $(n = 1, 2, \ldots, N)$, have been sampled (without replacement) from $\mathbf{s}_{Gen}$ (with mean μ_{Gen} and variance σ^2_{Gen}) and $\mathbf{s}_{Imp}$ (with mean μ_{Imp} and variance σ^2_{Imp}). Note that $\mathbf{s}_{Gen}$ and $\mathbf{s}_{Imp}$ denote sets of genuine and impostor scores generated by a biometric matcher on a dataset of N identities. For example, $\mathbf{s}_{Gen}$ and $\mathbf{s}_{Imp}$ may be the fingerprint match scores illustrated in the bottom of Figure 3.

In defining each virtual identity, an assumption is made that the range of genuine and impostor scores for each virtual identity is smaller than the range of the overall distributions, $f_G(s)$ and $f_I(s)$. The "tightness" of these ranges can be defined by the variance in match scores on a per-identity basis. Define these per-identity variances as σ^2_{n-n} and σ^2_{n-m}, where σ^2_{n-n} denotes the average variance in genuine scores for each identity and σ^2_{n-m} denotes the average variance in impostor scores for each pair of identities. Here, we remark that the intent of this assumption is to ensure created virtual identities *do not* share the same individual genuine and impostor match score distribution as the aggregate genuine and impostor score distributions. The output following the creation of each virtual identity is $\mathbf{S}$, a matrix of size N_TxN_T, wherein each column (or row) of $\mathbf{S}$ contains match score information for one "virtual" biometric sample, matched against $N_G - 1$ samples from the same "virtual" identity and $N_T - N_G$ samples from the remaining $N - 1$ "virtual" identities. Note that this exercise preserves the aggregate score statistics; what changes is the set of match scores pertaining to every identity.

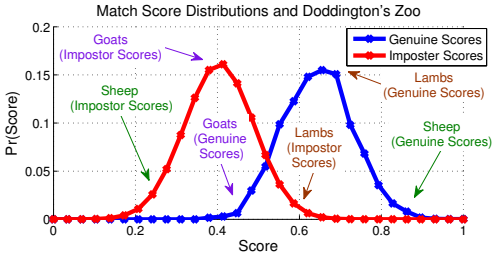

Figure 4. Visual illustrating the general concept of the proposed model for defining inter- and intra- class relationships in match scores, which creates virtual identities based on the "Doddington's Zoo" framework [5].

2.1. Modeling Inter- and Intra-class Variations

Our model for reassigning match scores to virtual identities is inspired by the "Doddington's Zoo" user-classification scheme, which characterizes identities based on their contribution towards the FMR and FNMR [5]. The Doddington's Zoo classification scheme consists of four classes: Sheep, Goats, Lambs, and Wolves. Sheep are defined as "well behaved" individuals who are easily recognized and do not incorrectly match with others. Goats are individuals who are intrinsically difficult to recognize and contribute to false non-match errors. Lambs are individuals whose biometric data can often be confused with other identities, resulting in false match errors. Finally, wolves are defined as individuals who willfully and successfully spoof the biometric data of other individuals, increasing the rate of false match errors.

In terms of match scores, sheep can be loosely characterized as having "high" genuine scores and "low" impostor scores. Meanwhile, goats can be loosely characterized as having "low" genuine scores. Finally, lambs (and wolves) can be loosely characterized as having "high" impostor scores. These simple characterizations formulate the basis of our model for reassigning scores to virtual identities, and is visually depicted in Figure 4.

The score reassignment model consists of two stages: initialization and sampling. During initialization, each of N virtual identities are assigned a label, χ_n $(n = 1, 2, \ldots, N)$, $\chi_n \in \{Sheep, Goat, Lamb\}$. The number of virtual identities corresponding to each label is pre-specified (see Section 3). Next, each identity is *assigned* match scores (from the original score set) based on the properties of a "Sheep", "Goat", or "Lamb". Sampled match scores are drawn (without replacement) from the original scores $\mathbf{s}_{Gen}$ and $\mathbf{s}_{Imp}$, and stored in $\mathbf{S}^n_{Gen}$ and $\mathbf{S}^n_{Imp}$, which are the reassigned genuine and impostor scores for the n^{th} virtual identity. Finally, a matrix of match scores of size N_TxN_T is created (denoted by $\mathbf{S}$). Each row in $\mathbf{S}$ stores the $N_G - 1$ assigned genuine scores and $N_T - N_G$ assigned impostor scores for each sample of a given virtual identity.

Appeared in Proc. of 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2013

Algorithm 1: Reassigning Genuine Scores

Input: Vector $\mathbf{s}_{Gen}$, containing the genuine scores.
Vector χ, a set containing the labels of each identity (e.g., "Sheep", "Goat", "Lamb").
Define: δ, ϵ_{Gen}: Scaling parameters.
Output: Matrix $\mathbf{S}$ populated with genuine scores.
\\ *begin algorithm*
Step 1: For each identity, note the assigned label.
Step 2a: Draw a genuine score (without replacement), ϕ, $\mathbf{s}_{Gen}$, from within subset $\mathbf{s}_{rng}$, where
$\mathbf{s}_{rng} = (\mu_{Gen} + \sigma_{Gen}, 1)$, if $\chi_n = Sheep$.
$\mathbf{s}_{rng} = (0, \mu_{Gen} - \sigma_{Gen})$, if $\chi_n = Goat$.
$\mathbf{s}_{rng} = (0, \mu_{Gen} + \sigma_{Gen})$, if $\chi_n = Lamb$.
Step 2b: If $\mathbf{s}_{rng}$ is a null set, and $\mathbf{s}_{rng} = (a, b)$, set $a = \delta \cdot a$, $b = \frac{b}{\delta}$ and repeat Step 2a.
Step 3a: Draw $\binom{N_G}{2} - 1$ scores (without replacement) from $\mathbf{s}_{Gen}$ within $\phi \pm \epsilon_{Gen}$.
Step 3b: If less than $\binom{N_G}{2} - 1$ scores can be drawn set $\epsilon_{Gen} = \frac{\epsilon_{Gen}}{\delta}$ and repeat Step 3a.
Step 4: Store the sampled genuine scores in $\mathbf{S}$.
return $\mathbf{S}$
\\ *end algorithm*

Assignment of genuine scores to each virtual identity is a relatively straightforward process. For each virtual identity, $\binom{N_G}{2}$ genuine scores are drawn without replacement[3] from $\mathbf{s}_{Gen}$ and stored in $\mathbf{S}$. Depending on the label of the virtual identity, a target range from which scores will be sampled, is first defined. This range is assumed to be between $(\mu_{Gen} + \sigma_{Gen}, 1)$, $(0, \mu_{Gen} - \sigma_{Gen})$, and $(0, \mu_{Gen} + \sigma_{Gen})$ for "Sheep", "Goats", and "Lambs", respectively. Denote the subset of genuine scores within this range as $\mathbf{s}_{rng}$. If $\mathbf{s}_{rng}$ is a null set, the target range is opened (i.e., increased) by multiplying (dividing) the lower (upper) bound of $\mathbf{s}_{rng}$ by a factor of δ ($0 < \delta < 1.0$) until $\mathbf{s}_{rng}$ contains at least one element. Next, one element (i.e., score) from $\mathbf{s}_{rng}$ is sampled and stored in $\mathbf{S}$. Denote the *value* of this score as ϕ. The remaining $\binom{N_G}{2} - 1$ scores are sampled from the range $\phi \pm \epsilon_{Gen}$, where ϵ_{Gen} is a tolerance parameter. As with the range used to sample ϕ, if no match scores are found within $\phi \pm \epsilon_{Gen}$, the range is opened by dividing ϵ_{Gen} by δ. This process for sampling genuine scores is summarized in Alg. 1. Note that this sampling method ensures that (a) sampled genuine scores for each identity are consistent, and (b) the genuine scores for a "Sheep" are distinct from those of a "Goat", and a "Lamb" (when possible).

Assignment of impostor scores to each virtual identity captures the inter-class relationships between identities. As such, assignment of impostor scores is viewed as being between pairs of identities (and therefore labels), rather than for a single identity. This results in six possible scenarios, viz. "Sheep-Sheep", "Sheep-Goat", "Sheep-Lamb", "Goat-Goat", "Goat-Lamb", and "Lamb-Lamb".

[3]equiprobable sampling

When sampling impostor scores between a pair of identities, ${N_G}^2$ impostor scores are sampled from $\mathbf{s}_{Imp}$, of which a single score, ϕ, is first drawn from a target range, $\mathbf{s}_{rng}$. $\mathbf{s}_{rng}$ is dependent on the labels denoting the pair of identities. Denote $\mathbf{S}^n_{Gen}$ and $\mathbf{S}^m_{Gen}$ as the set of *assigned* genuine scores for the n^{th} and m^{th} identities (i.e., the genuine scores assigned following implementation of Alg. 1). When both of the labels are a "Sheep" or "Goat", $\mathbf{s}_{rng}$ is limited to $(0, min\{max\{\mathbf{S}^n_{Gen}\}, max\{\mathbf{S}^m_{Gen}\}\})$, the *minimum* of the maximum genuine score observed for both identities. This constraint attempts to ensure that sampled impostor scores for a "Sheep" or a "Goat" will always be less than their corresponding genuine scores, preventing the occurrence of a false match error.

When one of the labels is a "Lamb", the only constraint emplaced is that $\mathbf{s}_{rng}$ is below the maximum genuine score for the paired identity. That is, if $\chi_n = Lamb$, and $\chi_m = Sheep$, $\mathbf{s}_{rng} = (0, max\{\mathbf{S}^m_{Gen}\})$. When $max\{\mathbf{S}^m_{Gen}\} > max\{\mathbf{S}^n_{Gen}\}$, this enables (but does not guarantee) the possibility of drawing an impostor score which can generate a false match (at rank-1) for the identity denoted as a "Lamb", but not the "Sheep". If $\chi_n = \chi_m = Lamb$, no constraints are emplaced on $\mathbf{s}_{rng}$, enabling (but not guaranteeing) the possibility of a false match (at rank-1) to occur for both identities.

As with the sampling of genuine scores, if $\mathbf{s}_{rng}$ is a null set, $\mathbf{s}_{rng}$ is opened fully to $(0, 1)$. Once a valid range of $\mathbf{s}_{rng}$ is identified, one impostor score is drawn from $\mathbf{s}_{Imp}$ and stored in $\mathbf{S}$. The remaining ${N_G}^2 - 1$ impostor scores are sampled from a range of $\phi \pm \epsilon_{Imp}$, where ϵ_{Imp} is a tolerance parameter. If no match scores are found within $\phi \pm \epsilon_{Imp}$, the range is opened by dividing ϵ_{Imp} by δ. This process for drawing impostor scores is summarized in Alg. 2. Note that this drawing method ensures that (a) the impostor scores between pairs of identities are consistent, and (b) the error dynamics for a "Sheep", "Goat", and "Lamb" are upheld (when possible).

3. Experimental Results

3.1. Datasets and Evaluation Criteria

The match scores included in our analysis correspond to the scores generated from face and gait modalities. Face scores were extracted from the WVU Multimodal Dataset [2] using the commercial software VeriFace. The face subset consists of $N_G = 5$ frontal face images for $N = 240$ unique individuals. Gait match scores were collected from the CASIA B [19]. The CASIA B dataset is a multi-camera dataset containing $N = 124$ individuals walking normally (6 sequences), with a coat (2 sequences), and with a backpack (2 sequences) from 11 different viewpoints. Here, we consider only those biometric samples where a subject is walking at a normal pace perpendicular to the optical axis

Appeared in Proc. of 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2013

Algorithm 2: Reassigning Impostor Scores

Input: Vector $\mathbf{s}_{Imp}$, containing the impostor scores.
Matrix $\mathbf{S}$, where sampled genuine scores are stored (from Alg. 1) and sampled impostor scores will be stored.
Vector χ, containing the labels of each identity (e.g., "Sheep", "Goat", "Lamb").
$\mathbf{S}^n_{Gen}, \mathbf{S}^m_{Gen}$, *Assigned* genuine scores for identities n, m.
Define: δ, ϵ_{Imp}: Scaling parameters.
Output: Matrix $\mathbf{S}$ populated with genuine and impostor scores.
\\ *begin algorithm*
Step 1: For all combinations of n and m ($n = 1, \ldots, N$, $m = n+1, \ldots, N$), note χ_n and χ_m.
Step 2: Draw an impostor score, ϕ from $\mathbf{s}_{Imp}$, within interval $\mathbf{s}_{rng}$, where
$\mathbf{s}_{rng} = (0, min\{max\{\mathbf{S}^n_{Gen}\}, max\{\mathbf{S}^m_{Gen}\}\})$, if $\chi_n = Sheep$ or $Goat$, $\chi_m = Sheep$ or $Goat$.
$\mathbf{s}_{rng} = (0, max\{\mathbf{S}^n_{Gen}\})$, if $\chi_n = Sheep$ or $Goat$, $\chi_m = Lamb$.
$\mathbf{s}_{rng} = (0, max\{\mathbf{S}^m_{Gen}\})$, if $\chi_n = Lamb$, $\chi_m = Sheep$ or $Goat$.
$\mathbf{s}_{rng} = (0, 1)$, if $\chi_n = \chi_m = Lamb$.
Step 3: If $\mathbf{s}_{rng}$ is a null set, $\mathbf{s}_{rng} = (0, 1)$.
Step 4a: Draw $N_G^2 - 1$ scores from $\mathbf{s}_{Imp}$ within $\phi \pm \epsilon_{Imp}$.
Step 4b: If less than $N_G^2 - 1$ scores can be drawn set $\epsilon_{Imp} = \frac{\epsilon_{Imp}}{\delta}$, and repeat Step 4a.
Step 5: Store the sampled impostor scores in $\mathbf{S}$.
return $\mathbf{S}$
\\ *end algorithm*

of the camera (i.e., $N_G = 6$). Match scores were extracted using the gait curves matching algorithm [3].

In our analysis, *aggregate* statistics are expressed by the area underneath the ROC curve (denoted by AUC). *Rank* statistics are expressed via the Weighted Rank-M identification accuracy, which is a weighted sum of the identification accuracies corresponding to the first M ranks in the CMC curve. Here, M is defined as 5% of the number of identities, N. The weight of the i^{th} rank, w_i ($i = 1, 2, \ldots, M$), is defined by $\frac{1}{i}$, and normalized such that $\|\mathbf{w}\|_2 = 1$. In Figure 5, a visualization of $f_G(s)$ and $f_I(s)$ is presented for the face and gait scores. A baseline evaluation consisting of AUC, Weighted Rank-M accuracy, predicted Weighted Rank-M accuracy (via the models of Bolle et. al. [1] and Hube [9]) and the empirically obtained proportions of "Sheep", "Goats", and "Lambs" is provided in Table 1. The strategy used to obtain empirical proportions of "Sheep", "Goats", and "Lambs" is the same as the one defined by Ross et. al. [15]. However, it should be noted that this scheme will always classify at least 30% of identities as having properties of a "Goat", or "Lamb", regardless of whether these identities contribute to adverse recognition performance. Note that in Figure 5 and Table 1, the performance values for both modalities are similar, but the genuine and impostor distributions, $f_G(s)$ and $f_I(s)$, for the gait scores share a larger range of nonzero values.

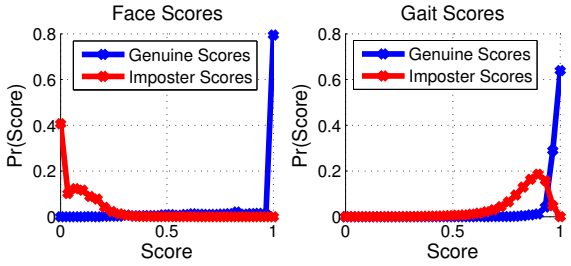

Figure 5. Genuine and impostor score distributions, $f_G(s)$ and $f_I(s)$, for the datasets used in our evaluation.

Table 1. Baseline AUC, Weighted Rank-M, estimated Weighted Rank-M, and the empirically obtained proportion of "Sheep", "Goats", and "Lambs" for the face and gait datasets.

	Face Scores	Gait Scores
Actual AUC	0.999	0.980
Actual Weighted Rank-M	1.0	0.978
Est. Weighted Rank-M (Bolle et. al.[1])	0.991	0.895
Est. Weighted Rank-M (Hube [9])	0.991	0.878
Proportion of {Sheep, Goat, Lamb} (%) (Ross et. al. [15])	{62, 28, 10}	{66, 24, 10}

3.2. Generating Multiple Ranked Statistics

Here, the Doddington-based model for creating virtual identities is implemented to create *alternative* realizations of inter- and intra-class relationships from the same set of scores. The intent of this experiment is to demonstrate that two sets of match scores sharing the same aggregate statistics can result in different ranked statistics. To enable this, the model is run with multiple proportions of "Sheep", "Goats", and "Lambs". Parameters for δ, ϵ_{Gen}, and ϵ_{Imp}, are set to 0.98, $0.25\sigma_{Gen}$, and $0.25\sigma_{Imp}$, respectively, for both face and gait scores. These results are tabulated in Tables 2 and 3.

Table 2. AUC and Weighted Rank-M values after reassignment of face match scores for different proportions of "Sheep", "Goats", and "Lambs". Note that in this case, the Weighted Rank-M accuracy does not change much.

Sheep (%)	Goats (%)	Lambs (%)	AUC	Rank-M
100	0	0	0.999	1.0
82	10	8	0.999	1.0
50	26	24	0.999	0.997
15	10	75	0.999	0.997

In addition, we highlight one such proportion that might result in a *lower* Rank-M performance. That is, the labels of χ_n are altered such that the number of "Sheep" or "well-

Appeared in Proc. of 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2013

Table 3. AUC and Weighted Rank-M values after reassignment of gait match scores for different proportions of "Sheep", "Goats", and "Lambs". Note that in this case, the Weighted Rank-M accuracy changes significantly.

Sheep (%)	Goats (%)	Lambs (%)	AUC	Rank-M
100	0	0	0.980	1.0
82	10	8	0.980	0.966
50	26	24	0.980	0.915
15	10	75	0.980	0.800

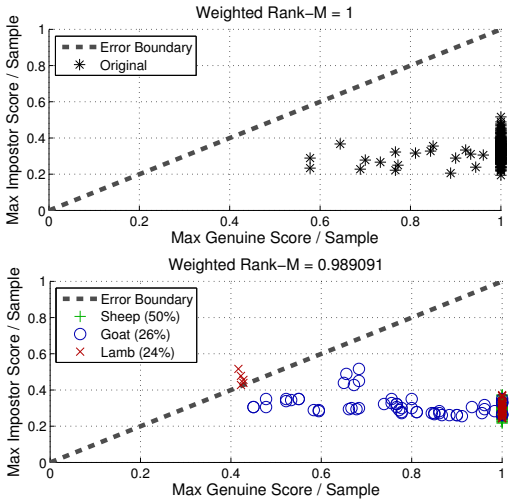

Figure 6. Comparing Weighted Rank-M accuracies before (above) and after (below) the score reassignment process for the face dataset. Note that here, although it is possible to generate a different realization of ranked match scores, the resulting Rank-M accuracy does not significantly vary (1 and 0.989).

behaved" virtual identities is reduced. The selected proportions for face and gait modalities are $\{50\%, 26\%, 24\%\}$ and $\{15\%, 10\%, 75\%\}$ for "Sheep", "Goats", and "Lambs", respectively. These highlighted proportions are illustrated visually in Figures 6 and 7, which plot the maximum impostor score against the maximum genuine score for each biometric sample,[4] for both reassigned and original face and gait scores. Visualization in this way illustrates how changing the proportion of "Sheep", "Goats", and "Lambs" can alter Rank-1 matching statistics. In addition, Figure 8 illustrates the actual ROC and CMC curves generated from the original and reassigned match score data for both face and gait scores.

4. Discussion and Future Work

In our experiment (Section 3.2), the score reassignment model is used to generate virtual identities with differing

[4]For a biometric sample, when its maximum impostor score exceeds its maximum genuine score, then a Rank-1 identification error will occur

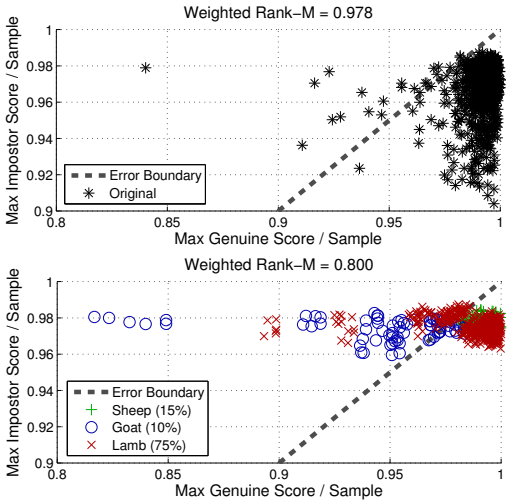

Figure 7. Comparing Weighted Rank-M accuracies before (above) and after (below) the score reassignment process for the gait dataset. Note that here, it is possible to generate a different realization of ranked match scores with a significantly different Weighted Rank-M accuracy (0.978 and 0.8). This suggests that multiple CMC curves can accompany the same ROC curve.

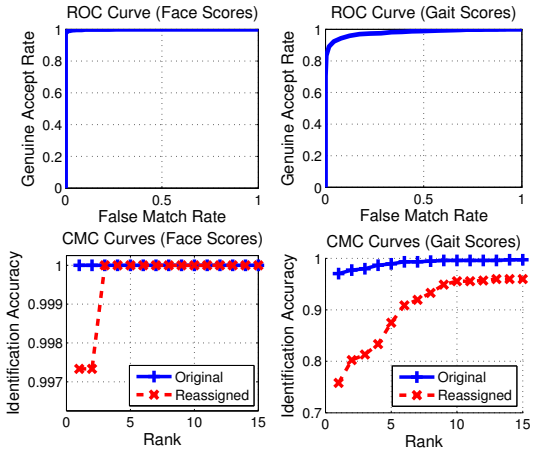

Figure 8. ROC and CMC curves for the original and reassigned face (left) and gait (right) match scores. Note that for both sets of match scores, the ROC data is the same, while the CMC data is different for the original and reassigned scores.

proportions of "Sheep", "Goats", and "Lambs" than the baseline values found in Table 1. In Figure 6 (involving face scores), when varying the proportion of "Sheep", "Goats", and "Lambs", while a difference in ranked statistics can be observed, the resulting Weighted Rank-M accuracy was not significantly different than that of the original data nor the predicted values (Table 1). However, in Figure 7 (involving gait scores), while varying the proportion of "Sheep", "Goats", and "Lambs", a realization with a significantly lower Weighted Rank-M accuracy (Weighted Rank-M =

0.8) was discovered. These observations are also evident in the CMC curves from Figure 8. Why then was this phenomena observed with the gait scores and not the face scores? The answer has to do with the extent of overlap between $f_G(s)$ and $f_I(s)$ (i.e., the range of s for which both $f_G(s)$ and $f_I(s)$ are non-zero). If $f_G(s)$ and $f_I(s)$ have less overlap, although match scores can be arranged differently between identities, this is unlikely to change the ordered ranking of match scores in the CMC curve. However, if $f_G(s)$ and $f_I(s)$ are reasonably overlapped, then **it cannot be guaranteed that aggregate-based statistics will correlate with rank-based statistics**. This property becomes particularly important when biometric systems increase in scale, as $f_G(s)$ and $f_I(s)$ may not conform to any specific distribution [16], and also in unconstrained biometric systems, which may yield larger inter- and intra-class variances as a consequence of uncontrolled biometric acquisition. This clearly suggests that CMC curve should not be presented without the associated ROC curve in closed-set evaluation scenarios.

5. Summary

The goal of this work was to study the impact of the Biometric Menagerie on the relationship between a ROC curve and the associated CMC curve generated from a set of genuine and impostor match scores (in a closed-set identification mode). In this regard, we designed a sampling scheme that reassigns the match scores generated by a matcher to virtual identities whose intra-class and inter-class relationships are defined based on the "Doddington Zoo" phenomenon. Experiments using fingerprint and gait scores suggest that multiple CMC curves can indeed be associated with a single ROC curve. This means, it is important that researchers report both ROC and CMC curves when comparing and/or analyzing the performance of biometric matchers.

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

[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.

[4] B. DeCann and A. Ross. Can a "Poor" Verification System be a "Good" Identification System? A Preliminary Study. *IEEE Workshop on Information Forensics and Security (WIFS)*, 1:31–36, 2012.

[5] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance. *IEEE International Conference on Language and Speech Processing*, pages 1351–1354, November 1998. Sydney, Australia.

[6] D. Gorodnichy. Multi-order Analysis Framework for Comprehensive Biometric Performance Evaluation. *SPIE Conference on Defense, Security and Sensing. DS108: Biometric Technology for Human Identification*, April 2010.

[7] D. Gorodnichy. Multi-Order Biometric Score Analysis Framework and Its Application to Designing and Evaluating Biometric Systems for Access and Border Control. *IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 44–53, 2011.

[8] P. Grother, G. Quinn, and P. Phillips. Report on the evaluation of 2D still-image face recognition algorithms. Interagency/Internal Report (NISTIR) 7709, National Institute of Standards and Technology (NIST), 2010.

[9] J. Hube. Using Biometric Verification to Estimate Identification Performance. *Biometrics Symposium*, pages 1–6, September 2006.

[10] A. Jain and J. Feng. Latent Fingerprint Matching. *IEEE Transactions on Patt*, 33(1):88–100, 2011.

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

[12] 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.

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

[14] 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.

[15] A. Ross, A. Rattani, and M. Tistarelli. Exploiting the Doddington Zoo Effect in Biometric Fusion. *3rd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–7, September 2009. Washington DC, USA.

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

[17] T. C. Y. Huang, D. Xu. Face and Human Gait Recognition Using Image-to-Class Distance. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(3):431–438, 2010.

[18] N. Yager and T. Dunstone. The Biometric Menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):220–230, 2010.

[19] 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.