Appeared in Proc. of IEEE International Workshop on Information Forensics and Security (WIFS), (Guangzhou, China), November 2013

De-Duplication Errors in a Biometric System: An Investigative Study

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*—The biometric de-duplication problem examines whether an input biometric sample has a corresponding match in a reference database during the enrollment process. If the input biometric is deemed to have a match, then the individual is *not* enrolled as a new identity in the database in order to prevent duplicate entries; otherwise, a new identity profile is created and the individual is enrolled in the system. The goal is to insure that the biometric data of an individual is associated with a single identity or label in the database. De-duplication is necessary in applications that render services to enrolled individuals. However, little to no research has been performed to examine the errors involved in a de-duplication task, and their potential consequences. We formally introduce the types of errors that may arise in biometric de-duplication, and examine whether these errors can be modeled using traditional error measures such as the false match rate, false non-match rate, false positive identification rate, and false negative identification rate. Experimental results demonstrate that de-duplication error is impacted by the order biometric samples are tested for a duplicate and that traditional error measures are not adequate for estimating empirical de-duplication error.**

I. Introduction

Biometrics is the science of recognizing humans based on the physical or behavioral traits of an individual, such as face, fingerprint, iris, hand geometry, voice, and gait [1], [2]. A classical biometric system has two distinct operational stages: the enrollment stage, when biometric data acquired from an individual is stored in the database along with a label or an identifier denoting identity; and the recognition stage when the input biometric data of an individual is compared against the enrolled data in order to recognize an individual. During enrollment, it is possible for a single individual to be associated with multiple labels or identities. This is referred to as identity *duplication*. A duplicate identity may be created by a malicious individual who intends to derive multiple benefits from the system (e.g., a welfare disbursement system). Alternatively, duplication may be a result of unintentional oversight by the system administrator during enrollment.

The process of detecting and managing duplicate entries associated with a single individual is referred to as *de-duplication*. So during enrollment, the input biometric sample is compared against the previously enrolled data by a biometric matcher in order to determine if a duplicate entry exists. If a duplicate entry is found, then the current input sample is flagged by the system.[1] In the simplest case, the input biometric data is not stored in the system. If no duplicates exist, then the input biometric sample is associated with a new label (i.e., identity profile) and stored in the system. A simple illustration of de-duplication is given in Figure 1.

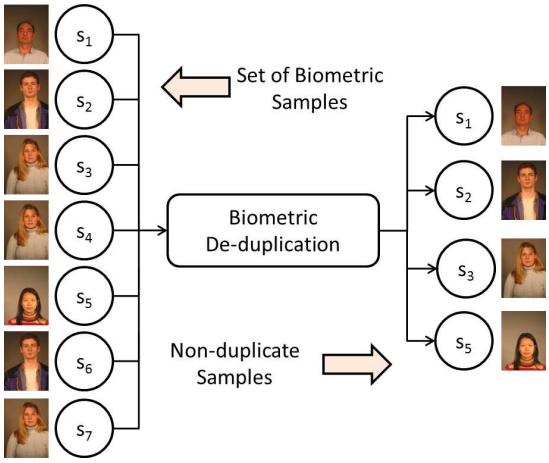

Fig. 1. Simple illustration of the input (left) and output (right) of a de-duplication task. Note that the output set contains one sample per identity (i.e., no duplicates). Face images are from the FRGC dataset [3].

The de-duplication task has gained considerable attention as of late, particularly in the context of national scale ID programs such as UID [4], and in maintenance of large scale forensic or government databases [5]. However, the application itself has not been rigorously studied in the literature. In particular, there has not been any work pertaining to the types of errors in a de-duplication task, their potential consequences, and whether they can be appropriately estimated.

Errors in classical biometric recognition are quantified using the false match rate (FMR), false non-match rate (FNMR), and receiver operating characteristic (ROC) curve (in the verification scenario); the false positive identification rate (FPIR) and false negative identification rate (FNIR) (in the open-set

[1]The response to a flag for a duplicate can vary according to system needs.

WIFS'2013, November 18-21, 2013, Guangzhou, China.
ISBN 978-1-4673-5593-3

identification scenario); or the cumulative match characteristic (CMC) curve (in the closed-set identification scenario). Each of these measures has been well-studied in the literature [1]. However, these measures may not adequately model true de-duplication error. For example, in traditional biometric verification or identification testing, the occurrence of a matching error is assumed to be a static event and cannot impact future matches. However, in the de-duplication problem, the reference database (with non-duplicate entries) has the potential to expand following each test for a duplicate (in particular, when a duplicate is not found). Consequently, the order in which biometric samples are observed during enrollment, can impact the error rate of the de-duplication task. Figure 2 presents an example demonstrating how two different sample orders can affect which samples are de-duplicated. Thus, in some sense, a de-duplication error resembles a *dynamic* matching error, similar to an Anonymous Identification system [6].

The motivation for this study is to formally introduce and analyze errors in biometric de-duplication, and determine whether these errors can be reliably estimated via traditional measures (e.g., FPIR, FNIR, etc.). Thus the goal of this study is to (a) Introduce and define the errors in biometric de-duplication (Section II); (b) Investigate whether traditional biometric error measures can first be leveraged in a *simplified* de-duplication problem space (Section III); and (c) Evaluate whether the designed measures accurately estimate constrained de-duplication error, and discuss confounding factors, if present (Section IV).

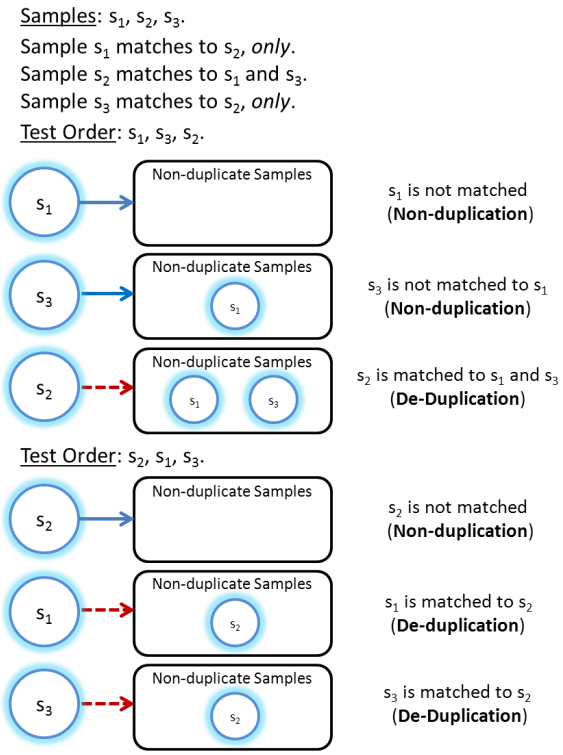

Fig. 2. Example illustrating the effect of sample order on the outcome of a de-duplication process. Note that the probability of a sample being "de-duplicated" depends on both the "non-duplicate" sample list and its position in the sequence when it is tested for a duplicate.

II. Understanding De-duplication

A. The De-duplication Task

Consider a set of N individuals, where each individual has provided N_G biometric samples. Denote the total number of ordered samples as N_T, where $N_T = N \cdot N_G$. A de-duplication task would reduce (or otherwise consolidate) the total number of samples to N, such that each individual is represented by *exactly* one sample. Note that this is done as follows:

Suppose a set of N_T biometric samples, $G_{init} = \{s_1, s_2, \ldots, s_{N_T}\}$, is to be de-duplicated. Additionally, define G_{out} as the set of non-duplicate samples remaining after the de-duplication task. Let $N_{out} = |G_{out}|$, the number of identity profiles or non-duplicate samples. Initially, G_{out} is initialized to the empty set and $N_{out} = 0$. When the first sample, s_1, is checked against G_{out} for a duplicate, there are no samples to match against and s_1 is placed in G_{out}. For all remaining samples, the k^{th} sample ($k = 2, 3, \ldots, N_T$) is matched against all the entries in G_{out}. A *de-duplication* occurs if the similarity (distance) match score generated between the k^{th} sample and the i^{th} element in G_{out} ($i = 1, 2, \ldots, N_{out}$) exceeds (is less than) a value of γ, where γ denotes a decision threshold. In the event of a de-duplication, the k^{th} sample is flagged for further action. For the purposes of this study, the sample is discarded. If a de-duplication does not occur, a *non-duplication* occurs and sample s_k is added to G_{out} and the value of N_{out} is increased by one.

Note that this scheme represents one approach towards performing the de-duplication task. In particular, the action taken following the flag for a duplicate. Other approaches may involve, for example consolidating information from the probe and its matching sample to update the stored identity profile.

B. De-duplication Errors

The de-duplication task incurs type-1 (false match) and type-2 (false non-match) errors. These errors are defined as follows:

***False de-duplication* (FDD)**: A sample incorrectly matches to an identity in the non-duplicate set, G_{out}. As a result, the input data *is not* added to G_{out} and the identity of an individual input may not be present in the non-duplicated set.

***False non-duplication* (FND)**: A sample, which has a matching identity in G_{out}, is incorrectly not matched to any sample in G_{out}. As a result, one identity is represented by multiple identity profiles.

The consequences of these errors can impact the outcome of the de-duplication task in different ways. For example, a large incidence of false de-duplication errors will result in a majority of identities not being represented in G_{out}. The operational impact of this error might be that several individuals will be unable to utilize services or receive resources, having been inadvertently deleted from the list of individuals.

The result of a false non-duplication, on the other hand, is that a single individual has *multiple* identities in the system.

Thus, a single individual may then be able to "double-dip" and procure services or resources intended for a single person.

III. Estimating De-duplication Errors

Given that matching errors persist in the de-duplication task, it is necessary to determine whether they can be estimated. In the traditional biometric literature, these errors are often measured through the false match rate (FMR), false non-match rate (FNMR), false positive identification rate (FPIR), and false negative identification rate (FNIR). In this section, we define two *simplified* de-duplication test scenarios which, on the surface, appear to enable direct usage of the FMR, FNMR, FPIR, and FNIR for estimating de-duplication error rates.

A. False de-duplication

Suppose N_T samples representing N identities are to undergo a de-duplication test (as defined in Section II-A). Additionally, suppose each identity is represented in the initial set of samples, G_{init}, *exactly* once (i.e., $N_G = 1, N_T = N$). Under these conditions, when each sample is tested for a duplicate, no genuine matches will exist in the non-duplicate set, G_{out}. Further, the probability of error cannot be confounded by a false non-match. Therefore, the probability of observing a false de-duplication error (under these conditions) depends on the probability of generating at least one of N_{out} impostor scores exceeding γ.

1) FMR-based Estimation: In the classical verification task [2], the FMR can be loosely interpreted as the probability that a generated impostor score exceeds a decision threshold (γ). Thus, an argument can be made that the FMR raised to the power m denotes the probability that m impostor match scores exceed a decision threshold. Conversely (1−FMR) raised to the power m denotes the probability that m impostor match scores are less than a decision threshold [7]. This formulates the basis for estimating false de-duplication error via the FMR. As such, the probability of observing a false de-duplication error when G_{out} contains N_{out} elements is the complement of the probability that all generated match scores are less than a decision threshold and is defined in Equation (1).

$$P(FDD|N_{out}) = 1 - (1 - FMR)^{N_{out}} \tag{1}$$

2) FPIR-based Estimation: In the classical identification task [2], a probe biometric sample is matched against a database of N labeled identities. The system computes match scores for every identity in the database and orders them from highest (similarity) to lowest. The output is a set of L identities whose match scores exceed a certain decision threshold. In open-set identification, the actual identity of the probe may or may not exist in the database (common with the de-duplication problem). Traditionally, the performance of open-set identification is measured through the FPIR and FNIR as demonstrated in the evaluation tests conducted by NIST [8]. The FPIR is defined as the proportion of probe samples that do not have a matching identity in the database but whose match scores with one or more database entries exceed γ. Thus, the FPIR (in this case) is equal to the probability that at least one generated impostor score exceeds γ, which corresponds to the probability of false de-duplication. This is expressed formally in Equation (2).

$$P(FDD|N_{out}) = FPIR(N_{out}) \tag{2}$$

B. False Non-duplication

Again, suppose N_T samples representing N identities is to undergo a de-duplication test. Here, let $N_G = 2$ and constrain γ such that the false match rate is negligible (i.e., FMR ≈ 0). Note these constraints are more stringent than in the previous section, but are introduced in order to mitigate any confounding effects from *prior* errors (i.e., errors made in the de-duplication process prior to encountering the current sample). For example, $N_G > 1$ is required to produce genuine scores (and thus the false non-duplication error), however, in the general case, $N_G - 1$ genuine scores can theoretically be generated for a given test sample, as a result of previous false non-duplication errors. Establishing $N_G = 2$ eliminates this artifact, as each test sample can *at most* generate one genuine score. Similarly, FMR ≈ 0 is introduced such that a *prior* false de-duplication error does not impact whether a genuine matching sample to the *current* test sample was erroneously discarded.[2] Thus, in the interest of simplicity, the problem is constrained to prevent this artifact and isolate the non-duplication error. Thus, the false non-duplication error (in this case) reduces to the probability a matching sample to s_k was previously observed, multiplied by the probability a generated genuine score is less than γ.

The probability that a generated genuine score is less than γ can be approximated using the FNMR and FNIR. By definition, the FNMR is defined as the proportion of genuine scores that are lower than a threshold, γ. Loosely interpreted, the FNMR denotes the probability that a generated *genuine* match score is less than a decision threshold. The FNIR is defined as the proportion of times a probe that does have a matching entry in the database generates a genuine score less than γ *or* is observed at a rank greater than R ($R = 1, 2, \ldots, N$). When $R = 1$, the FNIR denotes the probability a probe with a genuine matching identity in the database is incorrectly not matched due to one of two conditions: (a) the generation of a genuine match score less than γ *or* (b) a better match was found with another identity in the database. Note the second condition suggests that there are impostor scores greater than γ. However, with the assumption that FMR ≈ 0, this is not likely to bias the probability that a genuine score is less than γ.

1) FNMR-based Estimation: Regarding whether a genuine matching sample to s_k has been previously observed, if we assume samples are tested uniformly, this probability will simply be $\frac{k}{N_T}$. However, we require the expected value of k, given N_{out}, which may not be equal (due to true de-duplication events). Let $P(k, m)$ denote the probability N_{out} is equal to

[2]The challenge in measuring this probability is that it also depends on whether the erroneous matching samples were also observed and not subject to false de-duplication errors.

m after testing k samples. Then, the expected value of N_{out} after testing k samples, $E[N_{out}|k]$ is the sum of products of m and $P(k,m)$ for $m = 1, 2, \ldots, N_T$:

$$E[N_{out}|k] = \sum_{m=1}^{k} mP(k,m). \tag{3}$$

However, we are interested in computing $E[k|N_{out}]$, the expected value of k, given N_{out}. Define $\rho^m = \{\rho_1^m, \rho_2^m, \ldots\}$ as the set of values of k for which $P(k,m)$ is non-zero for a specific m. In other words, for $N_{out} = m$, this set denotes the range of potential sample indexes and $\sum_{k\in\rho^m} P(\rho^m, m) = 1.0$. In addition, define $|\rho^m|$ as the number of elements in this set. The expected value of k, given N_{out} can be computed as the average of ρ_i^m, for $i = 1, 2, \ldots, |\rho^m|$. This is given in Equation (4).

$$E[k|N_{out}] = \sum_{i=1}^{|\rho^m|} \frac{\rho_i^m}{|\rho^m|}, \;\; N_{out} = 1, 2, \ldots, N_T \tag{4}$$

The probability, $P(k,m)$ can be derived iteratively, for $m = 1, 2, \ldots, k$. In a de-duplication test, after the first sample is observed, it is added to G_{out} and $N_{out} = 1$. Thus, $P(k = 1, m = 1) = 1.0$. After the second sample is tested, one match score is generated, which can be either genuine or impostor. Denote the probability the match score can be classified as genuine as P_{Gen}, and the probability the match score can be classified as impostor as P_{Imp}. Note these probabilities must be "assumed" depending on the problem space. Concerning the outcome following the second sample ($k = 2$), $N_{out} = 1$ occurs if (a) the match score is genuine and correctly de-duplicated, or (b) the match score is impostor and falsely de-duplicated. The latter is assumed not to occur with FMR ≈ 0. Thus, $P(k = 2, m = 1)$ can be estimated using the FNMR, where a correct de-duplication event (i.e., a match was correctly found) is estimated as 1-FNMR (probability a genuine score exceeds γ), scaled by the probability the match score is genuine, P_{Gen}. The other outcome, $N_{out} = 2$ at $k = 2$, occurs if (a) the match score is genuine and falsely non-duplicated, or (b) the match score is impostor and correctly non-duplicated. Here, the probability of the former is defined by P_{Gen} multiplied by the FNMR, while that of the latter is defined by P_{Imp}. This process can be repeated to compute $P(k,m)$ for $k = 1, 2, \ldots, N_T$ and $m = 1, 2, \ldots, k$, enabling implementation of Equation (3). Thus, the probability of observing a false non-duplication is the product of the FNMR and $E[k|N_{out}]$, divided by N_T and is summarized in Equation (5).

$$P(FND|N_{out}) = \frac{FNMR \cdot E[k|N_{out}]}{N_T} \tag{5}$$

2) FNIR-based Estimation: The FNIR can be substituted for the FNMR in the derivation of $E[k|N_{out}]$ (Equation (4)), and $P(k,m)$, as a measure for estimating the false non-duplication rate under the stated assumptions. This is summarized in Equation (6).

$$P(FND|N_{out}) = \frac{FNIR(N_{out}) \cdot E[k|N_{out}]}{N_T} \tag{6}$$

IV. Experimental Results

A. Datasets and Evaluation

Experiments are conducted to (a) demonstrate the effect the sequential testing order has on de-duplication error and (b) evaluate whether traditional error measures can describe de-duplication error in the constrained scenarios presented in Section III-A and Section III-B. To enable this, similarity match scores were generated from a subset of the Facial Recognition Technology (FERET) database [9]. In particular, the subsets for regular frontal facial expression (code "fa") and alternative frontal facial expression (code "fb") are used. These subsets contain $N = 1009$ identities, with $N_G = 2$ samples per identity. Match scores were obtained using the commercial software VeriLook.

In our experiments, two mutually exclusive partitions of 504 identities are randomly selected for training and testing. These partitions are divided into two further subsets, denoted by the labels "A", "B", "C", or "D". In subsets "A" and "B", only one sample per identity is utilized (from FERET code "fa"). In subsets "C" and "D", both samples per identity are utilized. These partitions are summarized in Table I.

TABLE I
DATA PARTITIONS FROM THE FERET DATABASE [9].

	# Samples	# Identities	Code(s)
Partition A (Test)	504	504	"fa"
Partition B (Train)	504	504	"fa"
Partition C (Test)	1008	504	"fa" and "fb"
Partition D (Train)	1008	504	"fa" and "fb"

Samples in partitions "B" and "D" are used to generate estimates of the false match rate (FMR), the false positive identification rate (FPIR), and where applicable, the false non-match rate (FNMR), and the false negative identification rate (FNIR). Samples in partitions "A" and "C" are used to generate the empirical error rates after executing the de-duplication algorithm specified in Section II-A.

B. De-duplication Error and Testing Order

In this experiment, the observed false de-duplication error rate is computed for different sequential orders of test data. The intent of this experiment is to demonstrate that de-duplication error is *dynamic*, and can vary depending on the explicit order in which samples are tested.

To demonstrate this, a de-duplication test (as defined in Section II-A) is performed using samples from Partition "C" (Table I). In total, 10,000 tests are performed using the same test data but ordered differently. In each test, the *average* observed false de-duplication rate and false non-duplication rate is computed for a set of five decision thresholds, γ. The values of γ used in this experiment are those which correspond to a false match rate approximately equal to 0.25, 0.1, 0.01, 0.001, and 0.0001.

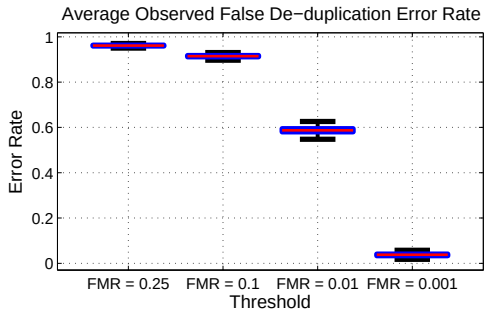

Fig. 3. Boxplot of the average false de-duplication error for selected values of γ. Note that the error rate *varies* depending on the order samples are tested.

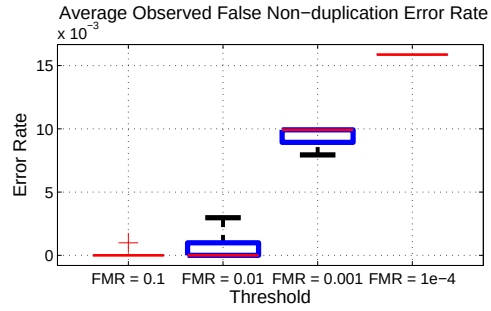

Fig. 4. Boxplot of the average false non-duplication error for selected values of γ. Note that the error rate *varies* depending on the order samples are tested.

These results are illustrated in Figures 3 and 4 in the form of a box plot. The width of each box denotes the upper and lower quartile of observed false de-duplication error. The lines extending beyond each box denote the full range of observed false de-duplication error. Outliers are designated by a "+". In order to reduce redundancy in the results, data from FMR $= 0.0001$ and FMR $= 0.25$ are neglected in Figures 3 and 4, respectively. These figures demonstrate that de-duplication error is *dynamic* and varies (between 3-10% for FDD and 0-0.3% for FND) depending on the sequential order samples are tested for a duplicate.

C. Estimating De-duplication Error

In this experiment, the ability to estimate false de-duplication and false non-duplication error under the *simplified* conditions described in Section II-B is evaluated. This is accomplished by comparing the observed false de-duplication error rate to the FMR-based and FPIR-based measures, as presented in Section III-A. Similarly, observed false non-duplication error is compared with the FNMR-based and FNIR-based measures presented in Section III-B. In addition, estimates of FMR, FNMR, and FNIR are included for additional comparison (where appropriate).

Here, parameters for the false de-duplication error models are estimated using Partition "B" and the empirically observed false de-duplication error is computed on Partition "A". In addition, observed and estimated error rates correspond to a decision threshold, γ, resulting in FMR ≈ 0.01. The parameters of the false non-duplication error models are

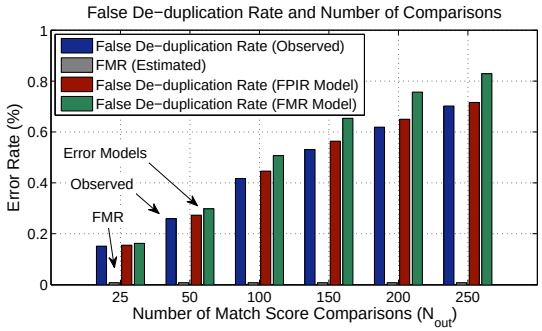

Fig. 5. Comparison of the FMR-based and FPIR-based error models to the observed false de-duplication rate. Note in this case, the error models denote a biased estimation of the false de-duplication error.

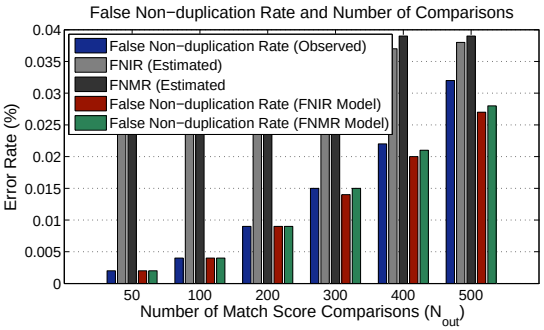

Fig. 6. Comparison of the FNMR-based and FNIR-based error models to the observed false non-duplication rate. Note in the case (with constraints), the error models appear to estimate false non-duplication error.

estimated from Partition "D" and the empirically observed false non-duplication error is computed from Partition "C". Here, observed and estimated error rates correspond to a value of γ resulting in FMR $\approx \frac{1}{N_T}$. To remove sampling bias, 100 different combinations of Partitions A, B, C, an D are computed and the resulting errors are averaged. These results are illustrated in Figures 5 and 6, where the false de-duplication error (Figure 5) and false non-duplication error (Figure 6) is shown in the form of a bargraph for set values of N_{out} at the stated value of γ. Note that since the data in Figures 5-6 is computed from different values of N_T and γ, the maximum value of N_{out} will be different.

D. Discussion

The above experiments highlight two major points. First, that de-duplication (and its errors) are *dynamic* and largely influenced by (a) the sequential order in which samples are tested for a duplicate (Figures 3-4) and (b) the number of elements in the non-duplicated sample set, G_{out} (Figures 5-6). This effect is peculiar to recognition tasks where matching outcomes can influence the future composition of the reference database as discussed previously by DeCann and Ross for an anonymous identification system [6]. Second, that de-duplication errors are complex and difficult to predict. In our defined "simple case" for the false de-duplication error ($N_T = N$), the FMR and FPIR estimators denote noticeable

bias (Figure 5). This is an interesting result, as the assumptions built into the problem appear to directly correlate with the definition of the FPIR. Therefore the logical question is: "What is the source of the bias?", for which we identify two likely sources.

The first source is related to the fact that classical error measures (in particular, the FMR and FNMR) denote *aggregated* match score statistics, which may not provide accurate representations of error on a per-identity level. In other words, these measures are based on a *global* analysis of error, while at a per-identity level, the error rate of an individual identity may differ from the FMR, FNMR, FPIR and FNIR. An example of this phenomenon is the Doddington's Zoo classification system of individuals in a biometric system based on their *individual* contributions towards the FMR and FNMR [10]. For example, in the Doddington's Zoo framework, "lambs" denote identities whose biometric feature set overlaps significantly with others. Such identities are likely to generate false de-duplication errors, and depending on *when* such identities are observed and the proportion of them that exist in G_{out}, the observed false de-duplication error rate can vary drastically. Similarly, "goats" denote identities whose biometric feature set does not match well against itself. Such users are likely to generate false non-duplication errors, which can also impact the observed error rate.

The second, and perhaps more significant source concerns a specific assumption built into the definition of FPIR and FNIR (and to a lesser extent, the FMR and FNMR). That being, these measures generally assume a probe can be compared against a database containing any combination of the other $N_T - 1$ samples. For example, the FPIR is computed by selecting some subset of samples (1 to $N_T - 1$) to define the "enrolled database", and the samples that do not have a corresponding match in the database are tested for a matching error. Note that there is *no restriction* on how the "enrolled database" is created. In other words, any possible combination of N_{gal} samples ($1 \leq N_{gal} \leq N_T - 1$) is valid. However, in de-duplication, some combinations of samples to comprise G_{out} are *outside* the set of possible outcomes (i.e., cannot occur).

To demonstrate this effect, consider the following "toy-example". Let θ denote a set of three biometric samples ($\theta = \{s_1, s_2, s_3\}$), where each sample denotes a different identity. Assume that s_1 and s_3 "match" (incorrectly) to s_2 (and vice-versa). To compute the FPIR we would choose N_{gal} ($1 \leq N_{gal} \leq 2$) samples to denote the database and test for an error with the other samples. Let $\mathcal{G}$ denote the set of hypothetical database combinations, which are: $\mathcal{G} = \{\{s_1\}, \{s_2\}, \{s_3\}, \{s_1, s_2\}, \{s_1, s_3\}, \{s_2, s_3\}\}$. However, in a de-duplication test, the gallery combinations $\{s_1, s_2\}$ and $\{s_2, s_3\}$ cannot occur, as the pair of samples match to one another and a de-duplication event will prevent these combinations from manifesting. Thus, **the sample space for estimating the FPIR *is not* the same as the sample space for estimating the false de-duplication rate**.

Although we do not see an immediate bias in the estimation of the false non-duplication error rate, this should not be interpreted as FNIR and FNMR being ideal estimators. In the general case ($N_G > 2$, FMR > 0), the false de-duplication error (which was effectively mitigated for the data in Figure 6) can reduce the probability that a test sample has a matching identity in G_{out}. Consequently, if the stated false non-duplication measures are adopted, a biasing artifact will be induced. Similarly, the false de-duplication error can be affected by prior false non-duplication errors. Therefore, given that the false de-duplication error cannot be estimated via traditional measures in the simple case, and that both de-duplication errors influence one another in the general case, it is likely traditional error measures will not provide reliable estimations of generalized de-duplication error. However, if the problem is re-defined such that the interest is in quantifying the error rate *given* N non-duplicate samples (in closed-set), then it is likely the observed error rate would converge to traditional error measures.

V. Conclusion

In this study we formally introduce the errors in the biometric de-duplication task, and describe the conditions required to generate a de-duplication error. Next, we demonstrate that under constrained conditions, the FMR and FPIR result in a biased estimation of false de-duplication error, while the FNMR and FNIR can act as an unbiased estimation of false non-duplication error. The observed bias is due to implicit assumptions present in estimating the FMR, FNMR, FPIR, and FNIR that do not hold for the de-duplication task. Therefore, traditional error measures may not be completely reliable when used to describe de-duplication error in the general case.

References

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

[2] A. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, January 2004.

[3] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Cheng, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[4] UIDAI, "Role of Biometric Technology in Aadhaar Enrollment," Government of India (GoI), Tech. Rep., January 2012.

[5] A. Jain, B. Klare, and U. Park, "Face Matching and Retrieval in Forensics Applications," *IEEE MultiMedia*, vol. 19, no. 1, 2012.

[6] B. DeCann and A. Ross, ""Has this Person Been Encountered Before?": Modeling an Anonymous Identification System," *IEEE Computer Society Workshop on Biometrics at the Computer Vision and Pattern Recognition (CVPR) Conference*, June 2012.

[7] P. Phillips, P. Grother, R. Michaels, D. Blackburn, T. Elham, and J. Bone, "FRVT 2002: Facial Recognition Vendor Test," DoD, Tech. Rep., April 2003.

[8] P. Grother, G. Quinn, J. Matey, M. Ngan, W. Salamon, G. Fiumara, and C. Watson, "IREX III Performance of Iris Identification Algorithms," National Institute of Standards and Technology (NIST), Tech. Rep. NIST Interagency Report 7836, 2012.

[9] 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*, vol. 22, no. 10, pp. 1090–1104, 2000.

[10] 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*, pp. 1351–1354, November 1998, Sydney, Australia.