Calibrated Confidence Scoring for Biometric Identification*

Dmitry O. Gorodnichy† and Richard Hoshino
Science and Engineering Directorate, Canada Border Services Agency
14 Colonnade Road, Ottawa, Ontario, Canada, K2E 7M6

Abstract

Existing biometric identification systems, such as those used in trusted traveler programs, attempt to identify an individual's identity from an enrollment database of n people. The output is either the name of an enrolled person, or a rejection message indicating that no match was found. Traditionally, no measure of confidence is given to the output; an individual is either granted or denied access. In this paper, we propose an extension to existing biometric systems by applying a calibration function to the n matching scores. We introduce a computationally-light calculation that can be applied either as a post-processing filter or embedded directly into an algorithm to yield perfectly calibrated probability-based scores. In addition to attaching a meaningful confidence measure to the output, the proposed methodology is also shown to improve the overall performance of a biometric system. The theoretical proof of the calibration formula is followed by its application to iris biometrics, on a data set consisting of nearly $60,000$ iris images. By comparing the detection error trade-off (DET) curves, we show that our calibrated post-processing filter reduces the area under the DET curve (AUC) by nearly 40% and reduces the equal error rate (EER) by nearly 50%.

1 Motivation

Biometric systems have evolved significantly over the past two decades, from single-sample non-automated verification systems to multi-sample fully-automated systems used for person identification and intelligence gathering [8, 9]. Despite the evolution in biometric system complexity, including its ability to handle multiple modalities (e.g. face, iris, fingerprint), the methodology for biometric performance evaluation has remained essentially static, still largely limited to graphing detection error trade-off (DET) curves, and reporting rates of false matches (FMR) and false non-matches ($FNMR$) [7, 8, 9, 10, 11, 12].

In conventional biometric systems, the output score is not associated to a probability. Instead, these biometric systems produce the same output (i.e., a match decision) regardless of whether there exist other close matching scores. Ideally however, biometric systems should be able to distinguish a confident output from a less-confident one. Specifically, it would be helpful to have a means of differentiating between a system output of "I am 100% sure that this person is George" and "I am 50% sure that this person is George, 45% sure that this person is Paul, and 5% sure that it is someone else." Even though both scenarios would lead to the system deciding that the individual is George, the decision for the first scenario should always be correct, while the decision for the second scenario should only be correct half the time.

Our goal is to introduce a *calibrated* confidence measure so that the output of a biometric system reflects the actual probability that a person is identified correctly. This concept is motivated by the work of DeGroot and Fienberg [3] who applied calibration to weather prediction, so that the statement "there is an 80% chance of rain tomorrow" is correct exactly 80% of the time. In this paper, we show how the same concept can be applied to biometrics to produce a meaningful confidence measure to each output. We introduce an algorithm that replaces traditional *matching* scores with perfectly calibrated *confidence* scores, and apply the theory to iris biometrics.

*In Proceedings of the International Biometric Performance Conference (IBPC 2010), NIST, Gaithersburg, March 2-4, 2010.
†Corresponding author (email: dmitry.gorodnichy@cbsa-asfc.gc.ca).

1.1 Multi-Order Analysis of Iris Biometrics

For iris recognition, the conventional method of identifying an individual is based on an adaptable parameter known as a *matching threshold*, determined from the pairwise Hamming Distances (HD) of the binary 0-1 strings that correspond to iris patterns. While there have been several recent exceptions, the most widely deployed iris biometric systems use Hamming Distances as matching scores, by calculating the measure of dissimilarity between two pairs of iris images.

Following the *Multi-Order Biometric Analysis* framework defined in [4, 5], consider an iris recognition system with the matching threshold set at $T = 0.33$. Suppose that an individual's iris image is compared against the images of five (different) people in an enrollment database. Suppose the five matching scores are 0.51, 0.32, 0.47, 0.34, and 0.31. If the algorithm selected the *first* person whose score was lower than the threshold – defined as an Order-1 system – then the system would have identified the individual as the second person in the database. If the algorithm computed all the matching scores and selected the person most below the threshold – defined as an Order-2 system – then the system would have identified the individual as the fifth person. In reality however, since there were three similar matching scores, it could have easily been the fourth person!

Expanding on our recent work [6], we introduce a calibrated scoring algorithm for biometric recognition that is a function of all n matching scores. Our calibrated confidence function is an example of an Order-3 system, a concept first presented in [4, 5]. While multi-order evaluation is not a built-in feature of traditional iris biometric systems, the potential and value of this approach is realized in the following three ways:

(a) This algorithm ensures the confidence scores are perfectly calibrated, regardless of the size of the enrollment database or the nature of the distributions of the genuine and impostor matching scores. Thus, a meaningful probabilistic confidence measure can always be assigned.

(b) This algorithm produces a convex DET curve that dominates the DET curve produced by *any* other algorithm. Therefore, this approach of turning matching scores into calibrated confidence scores maximizes the overall accuracy of the biometric system, and cannot be improved any further.

(c) The algorithm effectively separates the genuine confidence scores from the impostor confidence scores, with the overwhelming majority of genuine comparisons receiving the maximum confidence score of $c = 100\%$ and nearly every impostor comparison receiving the minimum confidence score of $c = 0\%$.

This paper proceeds as follows. In Section 2, we provide a brief explanation of iris biometrics to establish context. In Section 3, we prove our calibration result and illustrate it with a simple example. In Section 4, we apply the theory to a data set of $59,500$ iris images and demonstrate the effectiveness of this calibration algorithm. In Section 5, we discuss how these ideas can be implemented in practice, and conclude the paper in Section 6.

2 Brief Theory Behind Iris Biometrics

In traditional iris biometrics, an algorithm [1] based on Gabor wavelets turns an iris image into a 2048-digit binary string where each bit is either 0 or 1. When comparing two images, we either have a genuine comparison (iris images belonging to the same person) or an impostor comparison (iris images belonging to two different people).

The expected proportion of differing bits between impostor comparisons is $HD = 0.5$. Based on the analysis of Daugman [2], it is known that the histogram of impostor Hamming Distance scores follows a nearly perfect binomial distribution $Binom(m, u)$ with $m = 249$ and $u = 0.5$. The variable m represents the degrees-of-freedom and is a function of the mean u and the standard deviation σ:

$$m = \frac{u(1-u)}{\sigma^2}.$$

While each digit is assumed equally likely to be 0 or 1, only small subsets of bits are mutually independent due to internal correlations within an iris. That is why $m = 249$ rather than $m = 2048$. The frequency distribution of the impostor matching scores follows a binomial curve, analogous to a Bernoulli trial with

$u = 0.5$ and $m = 249$. Then, the probability that the Hamming Distance of two different iris images is $\frac{k}{m}$ is $p(HD = \frac{k}{m}) = \binom{m}{k} u^k (1-u)^{m-k}$. Since m is large, the majority of impostor matching scores is very close to 0.5.

Even comparing iris images of identical twins, the expected matching score is 0.5. However, when comparing two iris images belonging to the *same* individual, the matching score is considerably lower. Based on an analysis of $7,070$ genuine comparisons [2], the average HD was found to be $\hat{u} = 0.11$ with a standard deviation of $\hat{\sigma} = 0.065$. That is why iris biometric recognition has proven to be very effective, due to miniscule *intra-class* variability and large *inter-class* variability.

In the case of Hamming Distances, a lower score indicates a more confident match. Given a fixed threshold t, a *false match* occurs whenever the HD of an impostor comparison is less than or equal to t. Conversely, a *false non-match* occurs whenever the HD of a genuine comparison lies above t. Similarly, we can define a true match and a true non-match. For each t, we define the False Match Rate (FMR) and the False Non-Match Rate ($FNMR$) to be the percentage of false matches and false non-matches, respectively. Note that in the case of confidence scores, the same definitions apply but need to be flipped, as a higher score indicates a more confident match.

The detection error trade-off (DET) curve plots FMR versus $FNMR$ for a variable threshold. As the threshold decreases (towards 0), FMR decreases while $FNMR$ increases. Conversely, as the threshold increases (towards 1), FMR increases while $FNMR$ decreases. Thus, the DET curve measures the overall performance of a biometric system over all possible thresholds. The equal error rate (EER) is determined by finding the intersection of the DET curve with the line $y = x$. At this point of intersection, we have $FMR = FNMR$, and this is the value of EER. Two commonly-used performance metrics are to determine the EER, and to measure the area under the DET curve (AUC). A perfect algorithm will have $EER = AUC = 0$.

3 The Main Theorem

Let $\{x_1, x_2, \ldots, x_n\}$ be the set of enrolled people. Each of these n people have had their irises digitally photographed, and converted into a 2048-digit binary string. Let G be the set of genuine matching scores, and I be the set of impostor matching scores. We will assume that G and I follow binomial distributions, with $G \sim Binom(\hat{m}, \hat{u})$ and $I \sim Binom(m, u)$.

Suppose person X arrives at the kiosk. For each $1 \leq i \leq n$, define $s_i = HD(X, x_i)$ to be the matching score of x_i. Thus, person X produces the n-tuple $S = (s_1, s_2, \ldots, s_n)$, the vector of matching scores. We wish to determine $c_i = P(\{X = x_i\} \mid S)$, i.e., the probability that X is passenger x_i, given the n-tuple S. The probability vector $C = (c_1, c_2, \ldots, c_n)$ is the desired sequence of calibrated confidence scores.

Let $p_i = P(X = x_i)$ be the probability that an individual arriving at the kiosk is person x_i. Furthermore, let q be the probability that an individual arriving at the kiosk is unenrolled.

We now state and prove the main result of this paper.

Theorem 3.1 *Let G be the set of genuine matching scores, and I be the set of impostor matching scores. Suppose $G \sim Binom(\hat{m}, \hat{u})$ and $I \sim Binom(m, u)$. Let $p_i = P(X = x_i)$ and $q = 1 - \sum_{i=1}^{n} p_i$. Let $S = (s_1, s_2, \ldots, s_n)$ be the n-tuple of matching scores produced by person X. Then for each $1 \leq i \leq n$, we have*

$$c_i = P(X = x_i \mid S) = \frac{p_i z_i}{\displaystyle\sum_{i=1}^{n} p_i z_i + q \cdot \frac{(1-u)^m}{(1-\hat{u})^{\hat{m}}}}, \quad \text{where } z_i = \frac{\binom{\hat{m}}{\hat{m}s_i}}{\binom{m}{ms_i}} \cdot \left(\frac{\hat{u}^{\hat{m}}(1-u)^m}{u^m(1-\hat{u})^{\hat{m}}} \right)^{s_i}.$$

<u>Proof:</u> For each $1 \leq i \leq n$, define $r_i = P(\{X = x_i\} \wedge S)$. Also define $r_{imp} = P(\{X \notin \{x_1, x_2, \ldots, x_n\}\} \wedge S)$. By definition, $r_{imp} = P(S) - \displaystyle\sum_{i=1}^{n} r_i$. By Bayes' Theorem, we have

$$c_i = P(\{X = x_i\} \mid S) = \frac{P(\{X = x_i\} \wedge S)}{P(S)} = \frac{r_i}{r_1 + r_2 + \ldots + r_n + r_{imp}}.$$

To calculate $r_i = P(\{X = x_i\} \wedge S)$, we multiply the probabilities of the following $n+1$ independent events: it is x_i who comes to the kiosk; the genuine matching score $HD(X, x_i)$ is s_i; and the impostor matching score $HD(X, x_j)$ is s_j for all $1 \leq j \leq n$ with $j \neq i$.

Since $G \sim Binom(\hat{m}, \hat{u})$, there are $\hat{m}$ degrees-of-freedom, and the probability that any of these $\hat{m}$ bits differ is $\hat{u}$. So if $HD(X, x_i) = s_i$, then $\hat{m}s_i$ of the $\hat{m}$ bits differ. We derive the analogous result for the impostor distribution $I \sim Binom(m, u)$, for all $1 \leq j \leq n$ with $j \neq i$. Therefore, we have

$$\begin{aligned}
r_i &= p_i \binom{\hat{m}}{\hat{m}s_i} \hat{u}^{\hat{m}s_i}(1-\hat{u})^{\hat{m}-\hat{m}s_i} \cdot \prod_{j=1, j\neq i}^{n} \binom{m}{ms_j} u^{ms_j}(1-u)^{m-ms_j} \\
&= p_i \frac{\binom{\hat{m}}{\hat{m}s_i}}{\binom{m}{ms_i}} \cdot \frac{\hat{u}^{\hat{m}s_i}(1-\hat{u})^{\hat{m}-\hat{m}s_i}}{u^{ms_i}(1-u)^{m-ms_i}} \cdot \prod_{j=1}^{n} \binom{m}{ms_j} u^{ms_j}(1-u)^{m-ms_j} \\
&= p_i \frac{\binom{\hat{m}}{\hat{m}s_i}}{\binom{m}{ms_i}} \cdot \left(\frac{\hat{u}^{\hat{m}}(1-u)^m}{u^m(1-\hat{u})^{\hat{m}}}\right)^{s_i} \cdot \frac{(1-\hat{u})^{\hat{m}}}{(1-u)^m} \cdot \prod_{j=1}^{n} \binom{m}{ms_j} u^{ms_j}(1-u)^{m-ms_j} \\
&= p_i z_i \cdot \frac{(1-\hat{u})^{\hat{m}}}{(1-u)^m} \cdot \prod_{j=1}^{n} \binom{m}{ms_j} u^{ms_j}(1-u)^{m-ms_j}.
\end{aligned}$$

To calculate r_{imp}, we multiply the probabilities of the following $n+1$ independent events: it is an impostor who comes to the kiosk; and the impostor score $HD(X, x_j)$ is s_j for all $1 \leq j \leq n$. This yields

$$r_{imp} = q \cdot \prod_{j=1}^{n} \binom{m}{ms_j} u^{ms_j}(1-u)^{m-ms_j}.$$

Therefore, $c_i = \dfrac{r_i}{r_1 + r_2 + \ldots + r_n + r_{imp}} = \dfrac{p_i z_i}{\sum_{i=1}^{n} p_i z_i + q \cdot \dfrac{(1-u)^m}{(1-\hat{u})^{\hat{m}}}}$, and our proof is complete. ∎

3.1 Example

Below we illustrate the calibration theorem with a simple example. Let the enrollment gallery consist of three individuals, and suppose that each iris string has just six bits which are mutually independent. Thus, $n = 3$, $m = 6$, and $\hat{m} = 6$. Further, assume that G and I are binomially distributed with $\hat{u} = \frac{1}{3}$ and $u = \frac{1}{2}$.

Let $\{x_1, x_2, x_3\}$ be the three individuals in the gallery, and suppose their iris strings are $[0, 1, 0, 1, 0, 1]$, $[1, 0, 0, 1, 1, 1]$ and $[1, 0, 1, 1, 0, 1]$, respectively. Let X be one of these three individuals, chosen at random (i.e., $p_1 = p_2 = p_3 = \frac{1}{3}$). We wish to determine the identity of X, given that the iris string of X is $[0, 1, 0, 1, 0, 1]$.

We have $HD(X, x_1) = 0$, $HD(X, x_2) = \frac{3}{6}$, and $HD(X, x_3) = \frac{3}{6}$. Thus, person X generates the triplet of matching scores $S = (s_1, s_2, s_3) = (0, 0.5, 0.5)$. We wish to determine $C = (c_1, c_2, c_3)$, the vector of confidence scores where each $c_i = P(\{X = x_i\} \mid S)$ represents the probability that $X = x_i$, given S. Note that we do *not* have $C = (1, 0, 0)$, since it is possible that $X = x_2$ or $X = x_3$, which occurs when three of the six incorrect bits happen to match identically to produce the iris string of x_1.

By substituting the above values into Theorem 3.1, we determine that $z_1 = 1$ and $z_2 = z_3 = \frac{1}{8}$, from which we derive $C = (c_1, c_2, c_3) = (0.8, 0.1, 0.1)$. In other words, given that the vector of matching scores is $S = (0, 0.5, 0.5)$, eighty percent of the time the individual will be the first person in the gallery, and each of the other people ten percent of the time. Hence, the correct confidence score attached to x_1 must be 80%, with a confidence score of 10% assigned to each of x_2 and x_3. The output $C = (0.8, 0.1, 0.1)$ is preferable to a decision algorithm based solely on HD scores, which will always identify X to be individual x_1, i.e., $C = (1, 0, 0)$.

Note that this calibration formula preserves monotonicity, i.e., if x_i has a lower matching (HD) score than x_j, then x_i will have a higher confidence score than x_j. This is reflected in the above example. Then, a natural question is why this formula would be an improvement over the existing paradigm. To answer this question, we will show that the two approaches produce different DET curves, and that the DET

curve produced by the calibrated score function dominates the curve generated by the matching scores. By selecting the right threshold, we can reduce both false matches *and* false non-matches by applying this calibration algorithm as a post-processing filter.

In [6], we take this one step further and show that this algorithm produces a convex DET curve that dominates the DET curve produced by *any* other algorithm, implying optimality. Our optimality result is a corollary of the fact that the post-processing filter produces calibrated scores. Note that if the distributions are not binomial (e.g. they are Gaussian), then the formula in Theorem 3.1 will need to be replaced by something more complicated, or approximated by some discrete function. But once the correct calibration formula is established, this will produce an algorithm with an optimal DET curve [6].

In the following section, we illustrate the application of the theoretical result with real biometric data.

4 Application

Theorem 3.1 was applied to an actual data set consisting of matching scores for $59,500$ comparisons obtained with state-of-the-art iris recognition software. There was one iris image for each of 100 individuals, representing the enrollment gallery. Then a probe set of 595 people was matched against each of the 100 individuals in the gallery, producing $59,500$ comparisons. There were no unenrolled people in the probe set, i.e., each of the 595 iris images belonged to exactly one of the 100 enrolled individuals. Thus, there were 595 genuine comparisons and $58,905$ impostor comparisons.

Before applying Theorem 3.1, we need to know the values of $p_1, p_2, \ldots, p_{100}, q$. Since there were no impostors, we set $q = 0$. Not knowing how frequently each of the 100 enrolled individuals appeared among the sample of 595 individuals, we simply assume that $p_1 = p_2 = \ldots = p_{100} = 0.01$.

The mean and standard deviation of the two sets are $\hat{u} = 0.074$, $\hat{\sigma} = 0.088$, $u = 0.39$ and $\sigma = 0.0456$. As we are assuming that both distributions are binomial, we can determine the values for which $G \sim Binom(\hat{m}, \hat{u})$ and $I \sim Binom(m, u)$. We have $\hat{u} = 0.074$, $\hat{m} = \frac{\hat{u}(1-\hat{u})}{\hat{\sigma}^2} = 9$, $u = 0.39$ and $m = \frac{u(1-u)}{\sigma^2} = 114$.

For each of the 595 people in the probe set, we determine the matching score vector $S = (s_1, s_2, \ldots, s_{100})$. Then we apply Theorem 3.1 to transform S into the calibrated confidence score vector $C = (c_1, c_2, \ldots, c_{100})$, where $\sum c_i = 1$. Each c_i score is rounded to six decimal places.

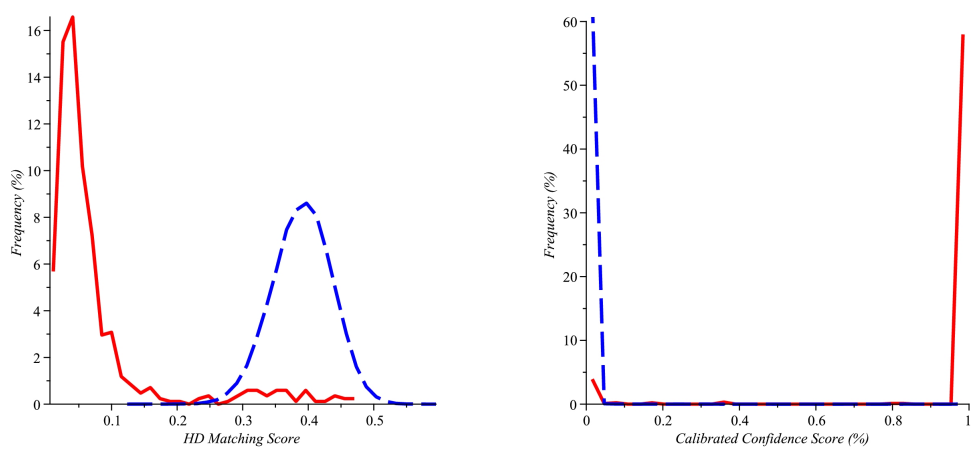

Figure 1: Genuine (solid line) and Impostor (dashed line) distributions of the matching scores and the calibrated confidence scores.

Figure 1 shows the resulting frequency distributions of both the genuine and impostor scores, with the graph on the left representing matching scores, and the graph on the right representing calibrated confidence scores. In the graph on the left, we note that the impostor matching scores follow a near-perfect binomial distribution, centered at $u = 0.39$.

When we apply the calibrated scoring function, we find that there is a significant separation in the genuine and impostor confidence scores. We no longer have two intersecting binomial curves. In fact, 90.1%

of genuine comparisons receive the maximum confidence score of $c = 1$, while 94.8% of impostor comparisons receive the minimum confidence score of $c = 0$.

To see how well the algorithm is calibrated, we tabulate the number of true matches and false matches at each threshold. For the sake of readability, Table 1 groups the thresholds into seven intervals.

Confidence Score	False Match	True Match	Accuracy
$c = 1$	0	536	100.00%
$0.9 < c < 1$	10	12	54.55%
$0.5 < c \leq 0.9$	10	2	16.66%
$0.1 < c \leq 0.5$	40	5	11.11%
$0.01 < c \leq 0.1$	319	9	2.74%
$0 < c \leq 0.01$	2710	15	0.55%
$c = 0$	55816	16	0.03%
TOTAL	58905	595	1.00%

Table 1: True and False Matches of confidence scores for this real data set

By the definition of calibration, the accuracy of each scenario should be the value of the corresponding c score. The theory is nicely confirmed for $c = 1$ and $c = 0$, as in the former case, 536 of the 536 comparisons are true matches (100%) and in the latter case, 16 of the 55832 comparisons are true matches (0.03%). The other scenarios need to be grouped together as there are so few instances. Nonetheless, had we been able to perform the experiment with $50,000$ individuals rather than 595, we would have had a larger sample to draw from, and the experimental results would have shown the confidence scores to be well-calibrated.

The real value of Theorem 3.1 is not just the improved separation of genuine and impostor scores; it is the creation of a better DET curve that implies fewer false matches and false non-matches, as shown in Figure 2. Note that score calibration produces a DET curve that completely dominates the curve produced by the original matching score algorithm.

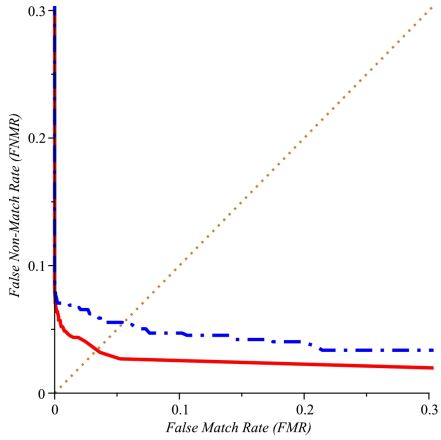

Figure 2: DET curves of the matching scores (dashed line) and calibrated confidence scores (solid line).

A DET curve achieves perfection as the curve approaches the origin. One way to measure the performance of a scoring algorithm is to calculate its *equal error rate* (EER), found by obtaining the intersection point of the DET curve with the line $y = x$. The lower the EER, the better the algorithm is. The status quo algorithm based on matching scores produces a DET curve with $EER = 5.40\%$, compared to $EER = 2.84\%$ for the calibrated algorithm. This represents an improvement of nearly 50%.

Another performance metric is calculating the area under the DET curve (AUC), which is in the interval $[0, 1]$. The smaller the area, the better the algorithm is. The status quo algorithm based on matching scores produces an AUC of 0.0244, compared to 0.0148 for the calibrated algorithm. This represents an improvement of nearly 40%. The results are summarized below in Table 2.

	EER	*AUC*
Status Quo Matching Scores	5.40%	0.0244
Calibrated Confidence Scores	2.84%	0.0148
Improvement	47.4%	39.3%

Table 2: Table of Results

5 Implementation

We close by describing some practical steps for actual implementation.

First, one should investigate the score distributions of genuine and impostor scores prior to deployment. Ideally, these distributions would be provided by the vendor. If this info is not available from a vendor, it can be computed empirically from a set of sample data by obtaining the values of $m, u, \hat{m}, \hat{u}$, and applying these values into the formula given in Theorem 3.1. Note that we can determine m (and $\hat{m}$) from the standard deviation σ (and $\hat{\sigma}$), since $m = \frac{u(1-u)}{\sigma^2}$.

Second, we recommend that instead of applying Theorem 3.1 to all n matching scores, one could take a smaller subset (e.g. the best 100 scores) and restrict the formula to this subset, since the remaining scores would almost certainly all have a confidence score close to 0. This would reduce the required computational costs and enable real-time implementation of this calibration function as a post-processing filter to existing conventional biometric systems.

As we saw in Section 4, when Theorem 3.1 was applied to a real data set with an enrollment gallery of $N = 100$, over 90% of genuine comparisons received the maximum confidence score of $c = 1$, while over 94% of impostor comparisons received the minimum confidence score of $c = 0$. These figures were attained when rounding each confidence score to *six* decimal places. If N is larger, these percentages will be even higher at these two endpoints. Thus, there is no need to perform all of the combinatorial computations when N is large, especially as the confidence score will be zero for all but a handful of cases.

We therefore propose applying Theorem 3.1 to the top 100 matching scores only, which would then assign a confidence score for 100 passengers in the gallery, while giving a confidence score of $c = 0$ to the other $N - 100$ passengers. This procedure will ensure that the post-processing filter is not computationally laborious, while preserving the accuracy and significance of the output. As for the actual display to the end user, one could provide a further restriction by outputing the names of only those individuals who have $c > 0.1$. For this particular threshold, the system will display only a few names along with their respective confidence scores.

Finally, we could consider a range of options for the variable q in our calculations for the confidence score. Since q is unknown, Theorem 3.1 could be applied for different values of q to obtain a range of possible outputs. As a result, we could end up with an output such as "This person is Rachel, with a confidence score between 97.5% and 99.2%."

6 Conclusion

It is not uncommon for contemporary biometric systems to have more than one match below the matching threshold, or to have two or more matches having close matching scores. This is especially true for the systems that store large quantities of identities and are applied to measure loosely-constrained biometric traits, as in stand-off biometrics. It is therefore important for such systems that their biometric recognition decision be accompanied by a meaningful confidence measure.

In this paper we have shown how a confidence score can be assigned to the output of a biometric system using probability-based scores. The proposed calibrated confidence scoring, which can be used either as a post-processing filter or embedded directly into a matching algorithm, is demonstrated to improve the overall performance of a biometric system.

The derived theoretical proof for the performance improvement is well supported by the actual data obtained from real-life testing of biometric systems. In our analysis of this real data set, we were able to

decrease the EER by nearly 50% and the AUC by nearly 40%, by simply applying the proposed calibration function to the default matching score outputs. Our approach promotes the multi-order performance analysis introduced in [4, 5] and establishes a concrete example of an Order-3 biometric system.

Acknowledgments

This research was partially funded by the Public Security Technical Program led by Defence Research and Development Canada (DRDC), within their Centre for Security Science (CSS). We gratefully acknowledge the work of Michael Lin, Anthony Tulai, David Xu, and Elan Dubrofsky in conducting the experiments and preparing the data used in Section 4.

References

[1] Daugman, J. (1993). High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15, 1148-1161.

[2] Daugman, J. (2004). How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology, 14(1) 21-30.

[3] DeGroot, M. H. and Fienberg, S. E. (1983), The comparison and evaluation of forecasters. Statistician, 12, 12-22.

[4] Gorodnichy, D. O. (2009). Evolution and evaluation of biometric systems. Proceedings of the IEEE Workshop on Applied Computational Intelligence in Biometrics, IEEE Symposium: Computational Intelligence for Security and Defence Applications (CISDA).

[5] Gorodnichy, D. O. (2010). Multi-order analysis framework for comprehensive biometric performance evaluation. Proceedings of SPIE Conference on Defense, Security and Sensing: track on Biometric Technology for Human Identification. Orlando, 5 - 9 April.

[6] Gorodnichy, D. O., Hoshino, R. (2010). Score calibration for optimal biometric identification. Proceedings of the Canadian conference on Artificial Intelligence. Ottawa, May 31 - June 2.

[7] Grother, P., Micheals, R.J., Phillips, P.J. (2003). Face recognition vendor test - 2002 performance metrics. Proceedings of the Fourth International Conference on Audio-Visual Based Person Authentication, 937-945.

[8] Jain, A.K., Flynn, P., Ross, A. (2007). Handbook of Biometrics. Springer.

[9] Li, S., (2009). Encyclopedia of Biometrics. Elsevier.

[10] Mansfield, N., Wayman, J.L. (2002). U.K. biometric working group best practices document. Teddington, UK. National Physical Laboratory.

[11] Schuckers, M.E., Hawley, A.M., Mramba, T.N., Livingstone, K.A., Knickerbocker, C.J. (2004), A Comparison of Statistical Methods for Evaluating Matching Performance of a Biometric Identification Device - A Preliminary Report. Proceedings of the Biometric Technology for Human Identification Conference.

[12] Wayman, J.L., Jain, A.K., Maltoni, D., Maio, D. (2005). Biometric Systems: Technology, Design and Performance Evaluation. Springer.