In Proceedings of Canadian Conference on Artificial Intelligence (AI 2010)
31 May-2 June, Ottawa, 2010.

Score calibration for optimal 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. We present a calibration algorithm that converts biometric matching scores into probability-based confidence scores. Using the context of iris biometrics, we show – theoretically and by experiments – that in addition to attaching a meaningful confidence measure to the output, this calibration technique yields the best possible detection error trade-off (DET) curves, both at the score level and at the decision level, thus maximizing the overall performance of the biometric system.

1 Motivation, definitions, theoretical proof of optimality

Let X be an individual arriving at a biometric kiosk, and let $\{x_1, x_2, \ldots, x_n\}$ represent the gallery of enrolled passengers. By comparing X with enrollee x_i, a matching score s_i is generated, representing the degree of similarity between the biometric feature(s) of X and x_i. Provided one of these matching scores is above or below a certain pre-determined threshold, the decision is made to grant or deny access to an individual.

In making this decision, no consideration is given to the fact that there might be other individuals in the enrollment gallery having similar matching scores. As a result, the performance of the biometric system is sub-optimal, leading to false accepts and false rejects. This decreases the reliability of the system, which is especially problematic in security-sensitive environments, as exposed in [4].

Our objective is to attach a probabilistic measure to a biometric system, by converting the n-tuple of matching scores $S = (s_1, s_2, \ldots, s_n)$ into the n-tuple $C = (c_1, c_2, \ldots, c_n)$ of confidence scores, where each $c_i = P(\{X = x_i\} \mid S)$ represents the probability that the identity of X is x_i, given the n-tuple S. Our formula then yields a scoring algorithm that is both *normalized* (the sum of the c_i's is 1), and *calibrated* (e.g., the statement "I am 60% sure that this person is Alice" is correct exactly 60% of the time.)

In [5], this confidence measure is presented in the context of iris biometrics, where an iris image is converted into a binary string in which each bit is equally likely to be 0 or 1, for which the histograms of impostor and genuine matching scores, as measured by the Hamming Distance (HD), are known [1] to follow binomial distributions. Let G be the set of genuine matching scores, and I be the set of impostor matching scores. Let $G \sim Binom(\hat{m}, \hat{u})$ and $I \sim Binom(m, u)$, where $(\hat{m}, \hat{u})$ and (m, u) are the degrees-of-freedom and mean of the two distributions. The following *Score Calibration Function (SCF)* is proven in [5]:

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

where $p_i = P(X = x_i)$ is the a-priori probability that an individual arriving at the kiosk is person x_i, and q is the probability that the individual is unenrolled.

This SCF function replaces matching scores with meaningful confidence scores that are perfectly calibrated and normalized.

We now prove that this calibration algorithm also produces a convex *Detection Error Trade-off (DET)* curve with the minimum *Area Under the Curve (AUC)*, both at the score level and at the decision level. By definition, the DET curve at the *score level* graphs the false match rate (FMR) against the false non-match rate ($FNMR$) over all possible thresholds, which is done by examining the scores given to genuine and impostor comparisons. On the other hand, the DET curve at the *decision level* graphs the false accept rate (FAR) against the false reject rate (FRR) over all possible thresholds, which is done by comparing all n scores and seeing if the highest score lies above the threshold.

Theorem: *If G and I are both binomially distributed, then the algorithm whose scores and match decisions are based on the calibrated confidence function (Eq. 1), rather than on the matching scores, produces the biometric system's best possible DET curve both at the score level and at the decision level.*

Proof. By Eq. 1, each vector $S = (s_1, s_2, \ldots, s_n)$ of matching scores gives rise to a vector $C = (c_1, c_2, \ldots, c_n)$ of confidence scores. Since there are only finitely many values for each $s_i = HD(X, x_i)$, there are only finitely many n-tuples S that can arise. Assuming there are t possible matching score vectors S, there are at most tn confidence scores. Suppose there are k unique confidence scores, where $k < tn$. Rank these k scores from highest to lowest, labeling them $r_1, r_2, \ldots, r_k$.

Taken over all genuine and impostor comparisons, let f_i and t_i represent the number of false matches and true matches with score r_i. Letting a_i be the accuracy of matches with score r_i, we have $a_i = \frac{t_i}{f_i + t_i}$ for all $1 \le i \le k$. Let $F = \sum f_i$ be the total number of impostor comparisons, and $T = \sum t_i$ be the total number of genuine comparisons. For each possible threshold, we now determine the values of FMR and $FNMR$, and the corresponding $k+1$ points of the DET curve. This is shown in Table 1.

Each point on the DET curve is represented by $(FMR_j, FNMR_j)$, where $FMR_j = x_j = \left(\sum_{i=1}^{j} f_i\right)/F$ and $FNMR_j = y_j = \left(T - \sum_{i=1}^{j} t_i\right)/T$, and by definition, $(x_0, y_0) = (0, 1)$ and $(x_k, y_k) = (1, 0)$. These k classes should be ordered so that the resulting DET curve becomes convex, as this minimizes the AUC. This important observation has been cited in previous papers [2, 3]. The DET curve of any scoring algorithm can be made convex by arranging the resulting classes in the proper order. As we will see, arranging the classes by calibrated confidence score achieves convexity.

For convexity, we require the slopes $g_j = \frac{y_j - y_{j-1}}{x_j - x_{j-1}}$ to increase as j increases. We have $g_j = \frac{-t_j/T}{f_j/F} = -\frac{F}{T} \cdot \frac{t_j}{f_j}$. Since F and T are constant, if g_j is an increasing function, we require $\frac{t_j}{f_j}$ to be a decreasing function. Since $\frac{1}{a_j} = \frac{f_j + t_j}{t_j} = 1 + \frac{f_j}{t_j}$, if g_j is increasing, this

[t] Threshold	Included Indices	False Match Rate	False Non-Match Rate
$r_1 + \epsilon$	None	0	$(t_1 + t_2 + \ldots + t_k)/T$
r_1	1	f_1/F	$(t_2 + \ldots + t_k)/T$
r_2	$1, 2$	$(f_1 + f_2)/F$	$(t_3 + \ldots + t_k)/T$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r_{k-1}	$1, 2, \ldots, k-1$	$(f_1 + f_2 + \ldots + f_{k-1})/F$	t_k/T
r_k	$1, 2, \ldots, k$	$(f_1 + f_2 + \ldots + f_k)/F$	0

Table 1. False Match and False Non-Match Rates for each threshold.

implies that a_j must be decreasing. By definition, $r_1 > r_2 > \cdots > r_k$, and by design the scores are perfectly calibrated. This implies that $a_j = r_j$, which shows that a_j is indeed a decreasing function. Thus, we have shown that by transforming matching scores into calibrated confidence scores, we ensure that the resulting DET curve becomes convex, thus implying optimality.

The exact same technique shows the optimality of the DET curve produced at the decision level, by replacing false matches with false accepts in the above proof, as well as false non-matches with false rejects. By the definition of calibration, a score of X is correct exactly $X\%$ of the time, both at the score level and at the decision level. This completes our proof. □

2 Practical use, experimental proof, conclusions

To apply the Score Calibration Function (Eq. 1), one would obtain $m, u, \hat{m}, \hat{u}$ values from the vendor or find them experimentally, where m is found from the standard deviation σ as $m = \frac{u(1-u)}{\sigma^2}$, as in a binomial distribution. Instead of applying the SCF to all n matching scores, one could take a smaller subset (e.g. the best 10 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 the real-time implementation of this calibration function as a post-processing filter to existing conventional biometric systems. Since q would normally be unknown, the formula can be applied for different values of q to obtain a range of possible outputs for the vector C. We now describe how the SCF was applied and tested with an actual iris biometric system and real iris data.

The data set consisted of 100 enrollee and 595 probe iris images (six for each enrollee minus five that failed to acquire), producing 595 genuine comparisons and $58,905$ impostor comparisons. For simplicity, we assumed that every enrollee used the system equally often, and that unenrolled people did not use it. Thus, we set $p_1 = p_2 = \ldots = p_{100} = 0.01$ and $q = 0$. Having computed all matching scores, the mean and degrees-of-freedom of the genuine and impostor score distributions were then obtained: $\hat{u} = 0.074$, $\hat{m} = 9$ ($\sigma = 0.0456$) and $u = 0.39$, $m = 114$ ($\hat{\sigma} = 0.088$).

For each of the 595 people in the probe set, we determined the matching score vector $S = (s_1, s_2, \ldots, s_{100})$, and then applied Eq. 1 to transform S into

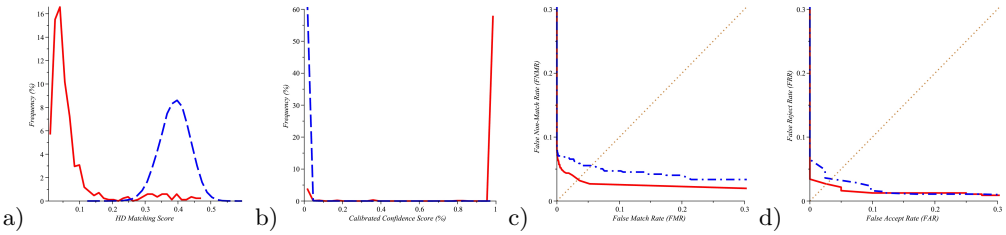

Fig. 1. Genuine and impostor distributions of the matching scores (a) and the calibrated scores (b), and their corresponding *DET* curves (dashed and solid lines, respectively) at the score level (c) and at the decision level (d).

the calibrated confidence score vector $C = (c_1, c_2, \ldots, c_{100})$, where $\sum c_i = 1$. Each c_i score was rounded to six decimal places.

Figure 1 presents the measured score distributions and the *DET* curves, showing that the *DET* curves of this calibration algorithm (solid lines) completely dominate the *DET* curves of the algorithm based on the matching scores (dashed lines), thus confirming our theoretical analysis.

To see how well the algorithm calibrates the scores, we tabulate the number of true matches and false matches for different threshold values. By the design, the calibrated confidence score c should be equal to the corresponding true match/accept rate. The experiments show this to be almost the case. – At the *score level*, of the 536 comparisons with the maximum score of $c = 1$, all 536 were true matches (100%), while of the 55832 comparisons with the minimum score of $c = 0$, only 16 were true matches (0.03%). At the *decision level*, there were 536 instances where the highest-scoring individual was given a confidence score of $c = 1$, and each was a true accept (100%), while the remaining 59 instances were insufficient to draw any statistically-significant conclusion.

Despite the small sample size, this analysis demonstrates the viability of the SCF in introducing meaningful confidence measures to the output of a biometric system. This approach may become particularly important in applications such as fully-automated Trusted Traveler identification, where the decision of the system is final and non-confident outputs may not be allowed.

References

1. Daugman, J. (2004). How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology, 14(1) 21-30.
2. Boström, H. (2005). Maximizing the area under the ROC curve using incremental reduced error pruning, ICML 2005 Workshop on ROC Analysis in Machine Learning.
3. Flach, P.A. (2003). The geometry of ROC space: Understanding machine learning metrics through ROC isometrics, Intern. Conference on Machine Learning, 194-201.
4. Gorodnichy, D.O. (2010). Multi-order analysis framework for comprehensive biometric performance evaluation, SPIE Conference on Defense, Security, and Sensing.
5. Gorodnichy, D.O., Hoshino, R. (2010). Calibrated confidence scoring for biometric identification. NIST International Biometric Performance Conference (IBPC 2010).