Calibrated (Probabilistic) Confidence Scoring for Biometric Identification

Goal: from scores to probabilities

\( (0, .5, .5) \rightarrow (80\%, 10\%, 10\%) \)

IBPC Conference (NIST), March 2-4, 2010

Dmitry O. Gorodnichey (PhD, CS/Math)
Head of Video Surveillance and Biometrics Section (2008-)

Richard Hoshino (PhD, Math)
Head of Mathematics and Data Exploration Section (2008-)

Science and Engineering Directorate
Why iris? – Easily accepted by public, touch-less / non-intrusive

Today: for collaborative user-engaged identification of pre-approved travellers in structured/overt environment (NEXUS)

Tomorrow: for fully-automated stand-off (on-the-fly) identification of Good and Bad people as they cross the border? (3 persons crossing/sec)

Recent RFI examination (Feb 2009-Aug 2009) exposed the problems even with Today’s systems/data

With Tomorrow’s stand-off systems, these problems will be even more significant!


Problems exposed through RFI

(With over 20,000,000 CBSA iris data, several state-of-art products, and over 6 months of coding and collecting/analyzing results)

1. There exist many (>5) matching algorithms now
   - All produce single scores output only (no confidence)!
   - Binomial nature of Imposter distributions
   - Binomial nature of Genuine distribution ? - with no noise

2. High FNMR (False Rejects, False Non-Match Rate)

3. High FTA (Failure To Acquire)

4. Despite many vendor/publications claims, systems often have:
   1) more than one match below the threshold,
   2) two or more close matching scores

There is a need therefore to assign Confidence value to output!
Anonymized score distributions
Anonymized stats

Using Multi-order score analysis [Gor09,10], Order 3 have shown that:

Many systems may improve FTA, FNMR, DET (match/non-match tradeoff) at the cost of allowing more than one score below a threshold.

(With 500 enrolled travelers, each having 6 passage images)
DEFINITION [Gor10]: **Failure of Confidence Rate (FCR)** – the rate of incidences in which there are more than one match below threshold...
Goal: assign confidences to decisions

**Given**: Person X arrives at the kiosk and produces n scores: n-tuple $S = (s_1, s_2, \ldots, s_n)$, $s_i = \text{HD}(X, x_i)$

**Find**: Sequence of calibrated confidence scores:
the probability vector $C = (c_1, c_2, \ldots, c_n)$, $c_i = P\{X = x_i\} | S$

**How**: as in probabilistic weather forecasting [DeGroot1983]
1. Make use of (assume) binomial nature of Genuine and Imposter score distributions [Daugman1993,2004]:
   - $G \sim \text{Binom}(m', u')$, with $u' = 0.11$, $d' = 0.065$ ($m' \approx 115$).
   - $I \sim \text{Binom}(m, u)$, with $u = 0.5$, $m = 249$ ($d \approx 0.03$)
   - $P(\text{HD} = k/m) = (k,m) u^k (1-u)^{m-k}$
2. Bayes’s Theorem for $c_i = P\{X = x_i\} | S =$
   \[ = P\{X = x_i\} \land S \} = P\{X = x_i\} \land S \} / P(S) = \ldots \]
3. $P\{X = x_i\} \land S \} = \ldots$
Enrolled: three individuals \{x_1, x_2, x_3\}, six bits in iris string.

- Thus, \( n = 3, m = m' = 6 \).
- \( G = \text{Binom}(m', u'), I = \text{Binom}(m, u) \) with \( u' = 1/3 \) and \( u = 1/2 \).
- \( x_1 = [0, 1, 0, 1, 0, 1], x_2 = [1, 0, 0, 1, 1, 1], x_3 = [1, 0, 1, 1, 0, 1] \)

New person: \( X = [0, 1, 0, 1, 0, 1] \).

- Matching scores \( S = (0, 0.5, 0.5) \). Decision scores: \( (1, 0, 0) \).

Using the theorem (for \( q=0 \) and \( P_1=P_2=P_3 \)), we obtain:

- confidence scores \( C = (0.8, 0.1, 0.1) \).

How to apply to real system?

- Vendor should provide: \( m', u' \, m, u \)
- User knows: \( P_i, q \) (a-priory probabilities of each person / imposter)
Proposed probabilistic score calibration can be added to any system at little computation cost as post-processing filter:

- Provides more meaningful output - for risk mitigating procedures
- Improves overall recognition
- Introduces Order-3 biometric systems

EER = 5.40% $\rightarrow$ 2.84%

DETAUC (area under the DET) $\rightarrow$ 0.17
Appendices
Iris biometrics

- Image converted to 2048 binary digits \( \{0, 1\} \)
  - only small subsets of bits are mutually independent [1].

- Impostor HD scores follow binomial distribution:
  \( I \sim \text{Binom}(m, u) \),
  \( m = 249 \) and \( u = 0.5 \).

- The variable \( m \) represents the degrees-of-freedom and is a function of the mean \( u \) and the standard deviation \( d \):
  \( m = u(1 - u) / d^2 \)

- Genuine HD scores [2]:
  \( G \sim \text{Binom}(m', u') \) with \( u' = 0.11 \), \( d' = 0.065 \)
Main theorem and proof:

**Theorem 3.1** Let $G$ be the set of genuine matching scores, and $I$ be the set of impostor matching scores. Suppose $G \sim \text{Binom}(\hat{m}, \hat{u})$ and $I \sim \text{Binom}(m, u)$. 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

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

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

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

To calculate $r_i = P(\{X = x_i\} \cap S)$, 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 $G \sim \text{Binom}(\hat{m}, \hat{u})$, 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 $m_{s_i}$ of the $\hat{m}$ bits differ. We derive the analogous result for the impostor distribution $I \sim \text{Binom}(m, u)$, for all $1 \leq j \leq n$ with $j \neq i$. Therefore, we have

$$r_i = p_i \left( \frac{\hat{m}}{m_{s_i}} \right)^{\hat{m} - m_{s_i}} \cdot \prod_{j=1, j \neq i}^n \left( \frac{m}{m_{s_j}} \right)^{m_{s_j}} \cdot (1-u)^{m-m_{s_j}} \cdot (1-\hat{u})^{\hat{m} - m_{s_i}}.$$
Because \( m = m' = 6 \), and \( u = 1-u=1/2, \ 2*u'=1-u'=2/3 \) many things get cancelled out …

\[
Z_i (S_i) = (6, 6*S_i) / (6, 6*S_i) * ( (1/3 ^ 6 * 1/2 ^6) / (1/2 ^ 6 * 2/3 ^ 6) ) ^ S_i = (1/2^6)^S_i = (1/2)^(6*S_i)
\]

For \( S_2 = S_3 = 0.5 \), we have: \( Z_2 = Z_3 = (1/2)^3 = 1/8 \).

For \( S_1 = 0 \), \( Z_1 = 1 \)

Then \( C_i = ( Z_i ) / (\Sigma Z_i) = Z_i/ ( 1/8 + 1/8 + 1) \),

and \( C_2 = 4/5 * (1/8) = 1/10, \quad C_1 = 8/10 \)
Multi-order performance evaluation

Order 0:

Order 1:

Order 2:

Order 3:

Ref. [Gorodnichy2009,2010]
Order 1 (Traditional):
- Examine single-scores to report trade-off (FMR/FNMR) curves

Order 2:
- Examine all scores to report the best (smallest) score

Order 3:
- Examine all scores relationship to report Confidences

Five-score example: \{ 0.51, 0.32, 0.47, 0.34, 0.31 \}. \( T = 0.33 \)
- Order 1 \( \rightarrow \) 0.32
- Order 2 \( \rightarrow \) 0.31
- But in reality it could have been 0.34 ! (if there was noise)


