Biometric Performance and Its Optimal Calibration

Talk previously presented in IJCB2014 Tutorial
Additional topics

Norman Poh
norman.poh@bjss.com
Data Scientist, BJSS London
Biometric performance

- Performance metrics
- Biblio note
- Procedures
- DET confidence intervals
- Performance synthesis
- User-specific score calibration
- User-specific fusion
- Tricks of the trade

- Variability vs Repeatability
- Biometric score theory
- Performance under attacks
- Performance reporting

- Doddington’s classification
- Yager and Dunston’s
- Biometric Menagerie Index
- Menagerie persistence
- User-ranking
- User-ranking on unseen data
- User-specific fusion with modality selection

- Tutorial: Exhaustive pairwise comparison

- Performance reporting
Thank you

n.poh@surey.ac.uk
TERMS AND PERFORMANCE METRICS
Some basic terms

- Biometric matcher
- Query or probe
- Score
- Template or reference
- Gallery
Genuine score

Template

Biometric matcher

Query or probe

Match score

Genuine score
(Mated score)
Zero-effort impostor score

Zero-effort impostor score
(Non-mated score)
Nonzero-effort impostor score

Template

Biometric matcher

Query or probe

Fake/spoofed sample

Spoof score

Reject

Nonzero-effort impostor score
Biometric Menagerie – scope

Future research

- Biometric antispooﬁng
- biometric presentation attack and countermeasures
Different-pair (nonmatch/impostor) scores

Same-pair (match/genuine) scores

\[ p(y|\omega = I) \]

\[ p(y|\omega = G) \]

pdf: probabilistic density function
Score, $y$

$\Pr(y|\omega = I)$

$\Pr(y|\omega = G)$
The likelihood is given by:

\[ p(y|\omega = I) \]

and

\[ p(y|\omega = G) \]

The False match rate, FMR (False Acceptance Rate, FAR) is defined as:

\[ FMR(\Delta) = P(y > \Delta|\omega = I) \]

\[ = 1 - P(y \leq \Delta|\omega = I) \]

The False non-match Rate, FNMR (False Rejection Rate, FRR) is defined as:

\[ FNMR(\Delta) = P(y \leq \Delta|\omega = G) \]

\[ P(y \leq \Delta|\omega) = \int_{-\infty}^{\Delta} p(y|\omega) \, dy \]
Note: $\Phi$ is the PDF of a normal distribution; $\Phi^{-1}$ is its inverse.
ROC versus DET

Receiver’s Operating Characteristic (ROC) curve

Detection Error Trade-off (DET) curve

EER: Equal Error Rate or Cross-over error rate

Further reading: Biometric Testing and Statistics
Side note: Similarity vs dissimilarity scores

Distance (dissimilarity) scores

Similarity scores

Same-pair

Diff-pair

Same-pair

Diff-pair
TUTORIAL: EXHAUSTIVE PAIRWISE COMPARISON EXPERIMENT
Generating an exhaustive pairwise comparison experiments

Probes

Gallery (templates)
What’s the difference between a wolf and a lamb???

Each subject may have multiple query images

(1 template each)
Exhaustive pair-wise comparison

https://normanpoh.github.io/blog/2017/12/29/generate-pairwise-fprint-scores.html
Same-pair (match) scores

Different-pair (nonmatch) scores

(a) Density

(b) FAR and FRR

(c) WER

(d) DET
Binary classification

Cross-over operating point
Equal error rate (EER)
PERFORMANCE REPORTING
Performance Reporting

Security applications:
Fix FAR to 1 in a million, report performance on FRR (security application)

Convenient applications:
Fix FRR to 1 in 100K, report performance on FAR

Equal:
FAR and FRR

\[
HTER(\Delta) = \frac{1}{2}FAR(\Delta) + \frac{1}{2}FRR(\Delta)
\]

Weighted:
\[
WER(\Delta) = \beta FAR(\Delta) + (1 - \beta)FRR(\Delta)
\]
Organising your data to get the scores

- **Training** set
  - Training
  - Development
- **Test** set
  - Evaluation

Parameter tuning/Training -> Model

Model -> Scoring (inference) -> Dev scores, Eva scores
Reporting performance

$$\Delta_\beta = \arg \min_{\Delta} \text{cost}_\beta(\Delta | \text{dev})$$

Where $\Delta$ is tuned \textit{a priori} on a \textbf{dev set} according to the criterion:

$$\text{cost}^{\text{wer}}_\beta(\Delta | \text{dev}) = \beta \text{FAR}(\Delta | \text{dev}) + (1 - \beta) \text{FRR}(\Delta | \text{dev})$$

$$\beta \in [0, 1]$$

Final performance on \textbf{eva set}:

$$\text{HTER} = \frac{\text{FAR} + \text{FRR}}{2}$$
Expected Performance Curve (EPC)

Low cost of FAR

High cost of FAR
Cumulative Match Characteristic (CMC) curve – Closed set identification

- Rank-k accuracy: Classification accuracy when the correct candidate is found within top k
VARIABILITY AND REPEATABILITY OF BIOMETRIC PERFORMANCE
Ask the right questions

Vendor: “Our false rejection is 2% when operating at a false acceptance rate of 1 in a million”
What questions should you ask?

1/50 FRR

1/1M FAR

Questions?
Sources of variability

- Biometric matcher
- Query or probe
- Biometric features of the design population
- Environment
  - Lighting
  - Temperature
- Application scenario
  - Covert vs overt
  - Level of user’s cooperation
  - Mounted or mobile

We are trying to collect the scores and measure the system performance

Three types of biometric tests

Technology test
- Use a (sequestered) database to test/compare algorithms
- **Repeatable** because many factors are ‘fixed’ once data is collected

Scenario testing
- Measure the overall system (sensor + algorithm) performance in a typical real-world application domain
- It is not repeatable because another field test will not give exactly the same result

Operational testing
- Model the performance in a specific application using a specific biometric system
- Not repeatable


Repeatability

Fix the conditions

- Same SOP for enrolment, same operators
- Same biometric system & sensor
- Same application scenario
- Similar (controlled) environment
- Comparable user size but different subjects

Will the two curves be the same?
How much do they vary?
Clients-based bootstrapping

Variability due to the user composition is observed under technology test

Questions:
• How much do they vary?
• Why do they vary?
• Can we extrapolate (generalize) to another scenario?
Session variability – why it matters?

Queries of genuine samples are from the same day (dev) or different days (eva)

Would the performance be similar?
With vs without session variability

<table>
<thead>
<tr>
<th>Label</th>
<th>template ID ( {n} )</th>
<th>Modality</th>
<th>Sensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>fa</td>
<td>1</td>
<td>Still Face</td>
<td>web cam</td>
</tr>
<tr>
<td>ft</td>
<td>1–6</td>
<td>Fingertip</td>
<td>Thermal</td>
</tr>
<tr>
<td>ir</td>
<td>1</td>
<td>Left iris</td>
<td>LG</td>
</tr>
</tbody>
</table>

Biosecure DS2 score database

http://personal.ee.surrey.ac.uk/Personal/Norman.Poh
Revisit the questions

1/50 FRR

1/1M FAR

Your answers...
Quiz

Should we consider session variability when testing a biometric system?

A: No, we should not because we want to fix all parameters in favour of experimental repeatability

B: Yes, we must do so in every biometric testing

Need a hint? Do you know of any biometric system that is designed to operate only on the same day as enrolment?

Explanation We must acknowledge that we can never fix all parameters and fixing all is not always desirable
What assumptions?

The demographic composition of the zero-effort impostors

Nonzero-effort impostor

Adverse conditions

The quality of the environment may vary

User interaction

A myriad # of unmeasurable factors

Samples are independently and identically distributed (i.i.d.)

Score, \( y \)
Formation of nonmatch scores

Zero effort impostors

Client templates/models

Impostor scores

\[ Y_1^I \]

\[ Y_2^I \]

\[ Y_3^I \]
Formation of match scores

Client accesses

Client template/model

Client scores

Scores of 3 clients

\[ Y_1^G \]

\[ Y_2^G \]

\[ Y_3^G \]
An Observed Experimental Outcome

Impostor scores

Client scores

Strong correlation exists among scores generated by the common client model

Samples are independently and identically distributed (i.i.d.) given the claimant ID

likelihood

score
How would i.i.d. samples look like?

Impostor scores

Client scores

$Y^I$

$Y^G$

Samples are independently and identically distributed (i.i.d.)

likelihood

score
Which is correct?

Samples are i.i.d.

Samples are i.i.d. given the claimed ID
• XM2VTS face system (DCTmod2, GMM)
• 3 genuine scores per user (blue curve)
• 200 users/clients
• 400 impostor scores per user (red curve)
User-specific fingerprint score distribution

Biosecure DS2
Fingerprint matcher (Bozorth3)
How are the local and global models related?

\[ p(y|\omega) = \sum_{j=1}^{J} p(y|\omega, j)P(j|\omega) \quad \text{for } \omega \in \{G, I\} \]

- The class-conditional pdf associated to claimant \( j \)
- The prior probability of claimant \( j \) given the class label \( \omega \)

There is no reason why one claimant is more important than another; so, \( P(j|\omega) \) should be uniform: \( P(j|\omega) = \frac{1}{J} \)

The system-level class-conditional pdf score is simply an average of the pdf of the claimant-specific pdf.
Consequence

The user composition has direct influence on the system performance even when all experimental conditions are fixed.

\[ P(y \leq \Delta | \omega = G) = \frac{1}{J} \sum_{j=1}^{J} P(y \leq \Delta | \omega = G, j) \]

\[ P(y > \Delta | \omega = I) = \frac{1}{J} \sum_{j=1}^{J} P(y > \Delta | \omega = I, j) \]
PERFORMANCE UNDER PRESENTATION ATTACKS
Nonzero effort attacks

Zero effort attacks

Nonzero effort attacks

likelihood

Score, $y$
ISO/IEC 30107-3 defined metrics for assessing the performance of the PAD methods

**Attack Presentation Classification Error Rate (APCER)**

\[
APCER_{PAI} = \frac{1}{N_{PAI}} \sum_{i=1}^{N_{PAI}} (1 - Res_i)
\]

*APCER_{PAI} is computed once for each PAI*

**Bona Fide Presentation Classification Error Rate (BPCER)**

\[
BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}}
\]

*Res_i = 1, declare presentation attack for sample i; 0 otherwise*
A single metric summarising all PAIs

- Average Classification Error Rate (ACER)

\[
ACER = \frac{\max_{PAI=1...S} \left( APCER_{PAI} \right) + BPCER}{2}
\]

\( APCER_{PAI} \) is computed once for each PAI
Performance under attack

Zero-effort FMR

Spoof FMR

(FNMR)
MODEL-BASED PERFORMANCE EVALUATION
Model-based assessment
Other Parameteric Point-based Assessment

\[ \text{EER} = \frac{1}{2} - \frac{1}{2} \text{erf} \left( \frac{F\text{-ratio}}{\sqrt{2}} \right) \]

where
\[ F\text{-ratio} = \frac{\mu^c - \mu^I}{\sigma^c + \sigma^I} \]

and
\[ \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp \left(-x^2\right) \, dx \]

\[ \text{Fisher-ratio} = \frac{(\mu^c - \mu^I)^2}{(\sigma^c)^2 + (\sigma^I)^2} \]

\[ d' = \frac{|\mu^c - \mu^I|}{\sqrt{\frac{1}{2} (\sigma^c)^2 + \frac{1}{2} (\sigma^I)^2}} \]
F-ratio and EER
F-ratio and EER

180 experiments from XM2VTS, NIST2001 and BANCA
Application of parametric analyses

User-specific performance analysis

Performance trend estimation over time

When no empirical error is observed

Multimodal fusion diagnosis
  • How correlation affect the system performance?
  • Which combination of biometric traits are optimal?
User-specific performance analysis

- XM2VTS face system (DCTmod2, GMM)
- 200 users/clients
- 3 genuine scores per user (blue curve)
- 400 impostor scores per user (red curve)

[Doddington et al, 1998]
Error estimation over time

1. Estimate score density over time
2. Estimate the performance over time

Step 1

Step 2
The homomorphic framework

1. Regression analysis
2. F-ratio calculation
3. EER calculation
4. User classification by trend
5. Trend grouping
6. Performance assessment
Degree of polynomial (1)
Degree of polynomial (2)
Degree of polynomial (3)
Group into 5 groups
EER of 256 Users

FRGC
DET curve evolution

http://epubs.surrey.ac.uk/812522/1/norman_eer_trend_mmbio_journal_v1.pdf
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Source</th>
<th>DOI</th>
<th>Type</th>
<th>Show details</th>
</tr>
</thead>
</table>

https://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2014.0107;jsessionid=xe60qfewntgh.x-iet-live-01
Classification of literature

BIBLIOGRAPHY NOTES
Classification of research in Biometric Menagerie

Studies of biometric menagerie

- Measuring the phenomenon
  - Looking at the extremes of distribution
  - Relative ranking
  - Directly characterising the variability

- Reducing the impact
  - Score normalisation
  - Fusion

- Collecting the impact
  - Doddington’s zoo
  - Yager and Dunstone’s classification
  - User discrimination
  - Biometric menagerie index
  - Parametric: Z-norm, F-norm, Group-based F-norm
  - Training-based: MS-LLR, logistic regression
Studies of biometric menagerie

Measuring the phenomenon
- Looking at the extremes of distribution
  - Doddington’s zoo
  - Yager and Dunstone’s classification
- Relative ranking
  - User discrimination
- Directly characterising the variability
  - Biometric menagerie index
  - Parametric: Z-norm, F-norm, Group-based F-norm
- Score normalisation
  - Training-based: MS-LLR, logistic regression

Reducing the impact
- Fusion
Some references

Doddington’s zoo


Yager and Dunstone’s classification


User discrimination


Biometric menagerie index

N. Poh and J. Kittler, A Biometric Menagerie Index for Characterising Template/Modelspecific Variation, in Int’l Conf. on Biometrics (ICB’09), 2009.

Parametric: Z-norm, F-norm, Group-based F-norm


Training-based: MS-LLR, logistic regression

Doddington’s Zoo

Strong impostor

High FAR

High FRR

Good clients (the majority of the claimants)

\[ p(y|\omega = I, j) \]

\[ p(y|\omega = G, j) \]

\[ p(y|\omega = I) \]

\[ p(y|\omega = G) \]

---

Characterising claimant-specific pdf

Need more and more data

- First order: $\mu^I_j$
- Second order: $(\sigma^I_j)^2$
- Skewness: $m^I_{3j}$
- Kurtosis: $m^I_{4j}$

- First order: $\mu^G_j$
- Second order: $(\sigma^G_j)^2$
- Skewness: $m^G_{3j}$
- Kurtosis: $m^G_{4j}$

$p(y|\omega = I, j)$

$p(y|\omega = G, j)$

Score, $y$
Mean
\[ \mu = E_y[y] \]

Variance
\[ \sigma^2 = E_y[(\mu - y)^2] \]

Skewness
\[ \frac{E_y[y-\mu]^3}{\sigma^3} \]

Kurtosis
\[ \frac{E_y[y-\mu]^4}{\sigma^4} \]


‘Peakedness’
What’s the difference between a wolf and a lamb???

Each subject may have multiple query images
What’s the difference between a wolf and a lamb??
What’s the difference between a wolf and a lamb???

Wolves are those who can generate very high ZE impostor scores

Claimants can be wolves, too!

Claimants (1 template each)

Queries
Open-set experiment

<table>
<thead>
<tr>
<th>Claimants (1 template each)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1^G$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>$Y_2^G$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>$Y_3^G$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

queries
Wolves are lambs and lambs are wolves when ...

We use a symmetric matcher and consider a closed set

A symmetric matcher is one that satisfies: $M(a,b) = M(b,a)$
Why do wolves matter?

\[ p(y|\omega = I, j \text{ is a wolf}) \]

\[ p(y|\omega = I) \]

\[ p(y|\omega = G) \]

Summary: Goats and lamb detection

Each subject may have multiple query images

<table>
<thead>
<tr>
<th>Claimants</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen</td>
<td>1000</td>
<td>1001</td>
<td>...</td>
<td>4999</td>
<td>5000</td>
</tr>
<tr>
<td>1</td>
<td>$Y^G_1$</td>
<td></td>
<td></td>
<td>$Y^I_1$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>$Y^G_{100}$</td>
<td></td>
<td></td>
<td>$Y^I_{100}$</td>
<td></td>
</tr>
</tbody>
</table>
Summary: Goats and lamb detection
The original Goat test was implemented using nonparametric Kruskal-Wallis rank sum test; we use a simplified test here.

Summary: Wolves detection

Each subject may have multiple query images

<table>
<thead>
<tr>
<th>Gen</th>
<th>1000</th>
<th>1001</th>
<th>...</th>
<th>4999</th>
<th>5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y^G_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>$Y^G_{100}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Wolves
Discussion: How were the 1M nonmatch scores chosen?

Vendor: “Our false rejection is 2% when operating at a false acceptance rate of 1 in a million”
Wolf diff-pair

- How will the performance look like using difficult samples chosen based on the wolf?
- Why would you do that? Computational complexity
YAGER AND DUNSTONE’S CLASSIFICATION
Biometric Menagerie

Group-based F-norm

Error rates
ON THE PERSISTENCE OF BIOMETRIC MENAGERIE
Key issues

• Are subjects inherently “hard to match”?  
  – Match score analysis: This question depends on how well we understand the environment in which biometric operates, thus requiring some generalization capability to unseen data  
  – Nonmatch score analysis: This can be answered!

• Why biometric menagerie: a conjecture
Are subjects inherently hard to match?

**Same-pair score analysis**

\[ \{U_j^1, U_j^2, U_j^3, U_j^4\} \]

\[ M(U_j^1, U_j^2) \]

---

**Scenario**

- Template
- Queries

**User-specific score calibration:** to reduce the menagerie effect


---

**Template selection:** to find a ‘representative’ template


\[ (n\choose 2) = \frac{n(n-1)}{2} \text{ scores} \]
Are subjects inherently hard to match?

**Diff-pair score analysis**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Set A queries</th>
<th>Set B queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1^I$</td>
<td>1000</td>
<td>2001</td>
</tr>
<tr>
<td>$\mu_{100}^I$</td>
<td>2000</td>
<td>4000</td>
</tr>
</tbody>
</table>

Impostor-based Uniqueness Measure

**Claimants**

Compare
Non-match scores are very stable

XM2VTS score-level benchmark database

\[ \mu_j^{\omega=I} \]

\[ \mu_j^{\omega=C} \]

\[ \sigma_j^{\omega=I} \]

\[ \sigma_j^{\omega=C} \]

One data set

\( \times 200 \) claimants
\( \times 13 \) experiments

Download the database here:
https://gitlab.com/normanpoh/xm2vts_fusion

Generalisation ability of biometric menagerie

Can we predict the statistics?

One of the 13 experiments

\[ \mu_j^{\omega \mid \text{dev}} \]

\[ \sigma_j^{\omega \mid \text{dev}} \]

corr: \( l=0.946, C=0.555 \)

corr: \( l=0.920, C=0.173 \)
How predictable are the statistics?

One of the 13 experiments
Grouping users by their performance
Conjectures

Biometric menagerie is system-dependent

- Feature space

It is related to subject representativeness

- Poor quality enrolment

Even with well-controlled enrolment, the menagerie still exists

User-specific schemes can reduce the effect

Changing menagerie membership


Wittman et al, Empirical Studies of the Existence of the Biometric Menagerie in the FRGC 2.0 Color Image Corpus, CVPRW 2006


Should the menagerie classify the users or their templates?

Classify users

Classify templates

Algorithm

Scenario

Users’ score characteristics persist!

Number of goats overlapped out of 50 trials


Is biometric menagerie persistent?

Assumption: Is the template fixed? No

- If you update your template, you need to recompute the separability criterion

Will the system operate in the same environment as the development set? No

- You need to compute the separability criterion for the expected operating condition

Is the system expected to operate for less than 1 year? No

- At some point, you may need to update the template (adaptive biometric system)

The menagerie is likely to persist

Menu
BIOMETRIC MENAGERIE INDEX
Desirable properties of a menagerie

index = 0

0 ≤ index ≤ 1

index = 1
Biometric Menagerie Index

N. Poh and J. Kittler, A Biometric Menagerie Index for Characterising Template/Model-specific Variation, in Int’l Conf. on Biometrics (ICB’09), 2009.
BMI = Mean of between client variance

Total variance
Global mean = client mean

Within-client variance dominates

Between client variance dominates
XM2VTS database

- N. Poh and J. Kittler, A Biometric Menagerie Index for Characterising Template/Model Specific Variation, in Int’l Conf. on Biometrics (ICB'09), 2009.
- Get the data set: http://goo.gl/CdXw9Z
Findings

Impostor BMI is insensitive to the choice of (zero-effort) impostor

client BMI > Impostor BMI

High client BMI values suggest that client-specific threshold or normalization is essential
USER RANKING
# User ranking

<table>
<thead>
<tr>
<th></th>
<th>Genuine</th>
<th>ZE Impostor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1^G$</td>
<td>$Y_1^I$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>100</td>
<td>$Y_{100}^G$</td>
<td>$Y_{100}^I$</td>
</tr>
</tbody>
</table>

Calculate the statistic:

- $\mu_1^G$
- $\mu_1^I$, $\sigma_1^I$
- $\mu_{100}^G$
- $\mu_{100}^I$, $\sigma_{100}^I$

Sort

Rank the users
Why rank the subjects?


Quality control during enrollment
- If the newly acquired template is not representative of the person, acquire a better (more representative) sample.

A tool to assess the worst-scenario DET curve
- How bad could a system perform for a group of weak (the weakest) users?
- The experience of each user in interacting with a biometric device matters!

A modality selection criterion in fusion
- Use only the more representative biometric modality instead.

Novel group specific decision
- An optimal decision threshold for each group of users.


Criteria to rank the users

• Calculate the performance of each user template and rank them somehow!

\[ EER_j = \frac{1}{2} - \frac{1}{2} \text{erf} \left( \frac{\text{F-ratio}}{\sqrt{2}} \right) \]

\[ \text{F-ratio} = \frac{\mu^c - \mu^I}{\sigma^c + \sigma^I} \]

\[ \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp \left[ -x^2 \right] \, dx \]

For each user/template!

[IEEE Trans. SP, 2006]
Many parametric discrimination criteria available

F-ratio = \frac{\mu^c - \mu^i}{\sigma^c + \sigma^i}

d' = \frac{|\mu^c - \mu^i|}{\sqrt{\frac{1}{2}(\sigma^c)^2 + \frac{1}{2}(\sigma^i)^2}}

J_1 = \frac{\mu^c}{\mu^i}, J_2 = \frac{(\mu^c - \mu^i)^2}{\mu^c \mu^i}

J_3 = \frac{(\mu^c - \mu^i)^2}{(\sigma^c)^2 + (\sigma^i)^2}

Multivariate statistics

Bhattacharyya distance

Chernoff Bound

Fisher-ratio = \frac{(\mu^c - \mu^i)^2}{(\sigma^c)^2 + (\sigma^i)^2}
User-ranking algorithm

Calculate the statistics

\[ \mu_1^G, \mu_1^I, \sigma_1^I \]
\[ \vdots \]
\[ \mu_{100}^G, \mu_{100}^I, \sigma_{10}^I \]

Apply separability criterion

\[ C_1 \]
\[ \vdots \]
\[ C_{100} \]

Sort

Rank the users

Group their scores

Plot performance

Worst perf

Best perf

Results on Face, Fingerprints & Iris

Biosecure database – a face example

F-ratio

\[ F\text{-ratio} = \frac{\mu^c - \mu^I}{\sigma^c + \sigma^I} \]

d-prime

\[ d' = \frac{|\mu^c - \mu^I|}{\sqrt{\frac{1}{2}(\sigma^c)^2 + \frac{1}{2}(\sigma^I)^2}} \]

Performance disparity across users

<table>
<thead>
<tr>
<th>Label</th>
<th>template ID {n}</th>
<th>Modality</th>
<th>Sensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>fa</td>
<td>1</td>
<td>Still Face</td>
<td>web cam</td>
</tr>
<tr>
<td>ft</td>
<td>1–6</td>
<td>Fingerprints</td>
<td>Thermal</td>
</tr>
<tr>
<td>ir</td>
<td>1</td>
<td>Left iris image</td>
<td>LG</td>
</tr>
</tbody>
</table>
Findings

Very few weak users; but they dominate the errors

F-ratio and d-prime can be used to rank the users

Can we rank the subject with maximal stability (on unseen data)?

- “B-ratio” (PRJ 2013) ranks users on unseen data
USER RANKING ON UNSEEN DATA
Rank the users on unseen data!

Independent data sets (different sessions)

Can we predict the statistics?

We designed 6 criteria and searched the best one

Approach

Calculate the statistics

\[ \mu_1 \quad \mu_1, \sigma_1 \quad \ldots \quad \mu_{100} \quad \mu_{100}, \sigma_{10} \quad \mu_{101} \quad \mu_{101}, \sigma_{101} \]

Apply separability criterion

\[ C_1 \quad \ldots \quad C_{100} \quad C_{101} \]

Rank the users

Sort

We should be able to predict problematic users before they use the technology

Re-enrol with new samples

Use a different biometric modality

Recommend another authentication mechanism

We should be able to predict problematic users before they use the technology

Re-enrol with new samples

Use a different biometric modality

Recommend another authentication mechanism

We should be able to predict problematic users before they use the technology

Re-enrol with new samples

Use a different biometric modality

Recommend another authentication mechanism
Predictability of user ranking on unseen data

F-ratio

B-ratio (in F-norm domain)

XM2VTS 13 systems
BIOMETRIC MENAGERIE
“PROCEDURE”
Biometric Menagerie “Procedures”

Definition 1 – Performance finetuning/calibration: Interventions that allow us to exploit Biometric Menagerie to improve the system’s future operation

1. User ranking
2. User tailoring

Definition 2 – Performance estimation / generalization / prediction: Algorithms that exploit Biometric Menagerie to estimate the system’s future operational performance

1. Performance intervals estimation
2. Performance synthesis
PERFORMANCE CONFIDENCE ESTIMATION
DET confidence intervals

Different subjects, different demographics

How well is the predicted curve?
Measure coverage
Coverage

![Graph showing FRR vs FAR with different markers and curves representing different categories: Inside, Outside, test curve, interval, and median. The graph has a grid and labeled axes for FRR and FAR.](image-url)
Conventional bootstrap

Impostor scores

Client scores

likelihood

score

135
A two-level bootstrap

Handle **between model variance** by bootstrapping the enrolled identities

Handle **within model variance** by bootstrapping the within-model scores

Subset Bootstrap

Bolle, Ratha & Pakanti, Error analysis of pattern recognition systems: the subsets bootstrap, CVIU 2014
Two-level Bootstrap

Users/Models

Level 1

Level 2

Bootstrap samples within a model: $S=4$ times

Plot DET
DET confidence via bootstrapping
Fix the users (claimants), bootstrap only the samples

Bootstrap the claimants, and keep all the scores due to the chosen claimants
Effect of subject samples

Increased subject

Better precision
Findings

User-induced variability is more important than the sample variability

Increasing the subject population reduces the uncertainty

Coverage achieved 70-90% on NIST 2005 Speaker Evaluation data set
Average coverage of 24 NIST 2005 systems

Configuration: Trained on 31 users; tested on another 62 users
How to best generate 10,000 genuine score samples in order to conduct a biometric test?

A. Recruit 10 subjects each contribute 1000+1 samples

B. Recruit 100 subjects each contribute 100+1 samples

C. Recruit 1000 subjects each contribute 10+1 samples
On-going research

PERFORMANCE SYNTHESIS
Synthesizing DET curves to other operating conditions

Data with recorded conditions

Operational data that we haven’t collected

Previous operational data

Application specs

A new installation

Performance synthesis

Compare prediction quality

Data with recorded conditions

Previous operational data

Application specs

A new installation

Performance synthesis

Compare prediction quality
Benefits of performance synthesis

• Liberates performance curves from its assessment data set
• Provides a framework for test reporting that promotes documentation and measurement of experimental conditions
• Potentially reduces cost in testing
Negative (impostor) scores

Negative scores (cond 1)

Estimate cdf

Bootstrap

Combine

Positive (genuine) scores

Positive scores (cond |Q|)

Estimate cdf

Bootstrap

Combine

Set the priors of $P(Q|\omega)$
The underpinning theory

\[ p(y|\omega) = \sum_Q p(y|\omega, Q) P(Q|\omega) \]

The system-level (class-conditional) score pdf is a mixture of factor-dependent pdfs.

\[ P(y \leq \Delta|\omega) = \sum_Q P(y \leq \Delta|\omega, Q) P(Q|\omega) \]

For performance estimating, we don’t need to estimate the pdfs, but only their cdfs, which is monotonic function.
\[ P(y \leq \Delta|\omega = G) = \sum_Q P(y \leq \Delta|\omega = G, Q)P(Q|\omega = G) \]

\[ P(y > \Delta|\omega = I) = \sum_Q P(y > \Delta|\omega = I, Q)P(Q|\omega = I) \]
Fingerprint with NFIQ

4&5 (lowest quality)

3(ML)  2 (MU)

1 (highest quality)

False nonmatch rate

Zero-effort impostor scores (False match Rate)
$FMR(\tau|\omega_1, Q = 4 & 5)$

$FMR(\tau|\omega_1, Q = 2)$

$FMR(\tau|\omega_1, Q = 3)$

$FMR(\tau|\omega_1, Q = 1)$

$FMR(\tau|\omega_0, Q)$ for $Q = \{1, 2, 3, 4 & 5\}$

$FMR(\tau|\omega_1, Q) = 1 - P(y < \tau|\omega_1|Q)$
Simulated DET curves of varying quality

High quality tendency
[8 4 2 1]

Low quality tendency
[1 2 4 8]

Equal prior quality
[1 1 1 1]

Relative weights of the 4 curves:
[Q1 Q2 Q3 Q4&5] or [H MU ML L]
Study II: Cross-sensor performance

Cross-sensor

Sensor 1
(Biometrika)

Sensor 2 (Ital)
Simulated DET curves with mixture of data

The ratio shows [ S1 S2 cross-sensor]
Study III: Spoof attack

Zero-effort FMR

Spoof FMR

(FNMR)
Performance under various attacks

Assessment with DET angles
Achievable coverage

Coverage achieved between 60% and 80%
USER-SPECIFIC SCORE NORMALISATION
Score normalisation

Original matching scores

\[ y \]

Z-norm

\[ y^Z = \frac{y - \mu^I_j}{\sigma^I_j} \]

F-norm

\[ y^F = \frac{y - \mu^I_j}{\bar{G}_j - \mu^I_j} \]

Bayesian classifier

\[ y^B = \log \frac{p(y|I,j)}{p(y|G,j)} \]

[IEEE TASLP’08]
<table>
<thead>
<tr>
<th>Procedures</th>
<th>Formulas</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-norm</td>
<td>$y_j^Z = \frac{y - \mu_j^I}{\sigma_j^I}$</td>
<td>$E_j[y_j^Z</td>
</tr>
<tr>
<td>F-norm</td>
<td>$y_j^F = \frac{y - \mu_j^I}{\gamma \mu_j^C + (1 - \gamma) \mu_j^c - \mu_j^I}$</td>
<td>$E_j[y_j^F</td>
</tr>
<tr>
<td>EER-norm</td>
<td>$y_{EER}^j = y - \Delta_j$</td>
<td>$y_{EER}^j &gt; 0$ is an optimal decision function (at EER) for all $j$</td>
</tr>
<tr>
<td>MS-LLR norm</td>
<td>$y_{\text{LLR}}^j = \log \frac{p(y</td>
<td>c, j)}{p(y</td>
</tr>
</tbody>
</table>

$$
\mu_j^{F,c} = E[y_j^F | C] = \frac{E[y_j^C] - \mu_j^I}{\mu_j^C - \mu_j^I} = 1, \text{ for all } j
$$

$$
\mu_j^{F,I} = E[y_j^F | I] = \frac{E[y_j^I] - \mu_j^I}{\mu_j^C - \mu_j^I} = 0, \text{ for all } j
$$
USER-SPECIFIC FUSION
Conventional multimodal fusion
User-specific multimodal fusion

Claimed ID → User-specific parameters

Face expert

Claimed ID → User-specific parameters

Speaker verif. expert

User-specific Normalisation

User-specific Normalisation

Fusion

\[ u_F(j) \]

\[ u_S(j) \]
Client-specific fusion

Original fusion problem

Fusion problem after applying F-norm
Some results

(a) relative change of EER

\[
\frac{EER_{\text{norm}} - EER_{\text{bline}}}{EER_{\text{bline}}}
\]
USER-SPECIFIC & SELECTIVE FUSION
System architecture

Criterion = “B-norm of Fratio”

Modality 1 → F-norm → Compute criterion → Recommend usage

Modality 2 → F-norm → Compute criterion → Recommend usage

Fusion with selection
Rationale: The B-ratio of F-norm

F-ratio = \frac{\mu_j^G - \mu_j^I}{\sigma_j^G + \sigma_j^I}

B-ratio = \frac{\mu_j^G - \mu_j^I}{\sigma_j^I}

\sigma_F^G \text{ unknown}

The discrimination power of the user is fully captured by

\sigma_F^I

\mu_F^{I, j} = 0 \quad \mu_F^{G, j} = 1
Cost savings

computational saving = 1 - \frac{\sum_{j\in J} \sum_{i=1}^{N} I(\text{system}_{i,j})}{N \times J}

= \frac{r}{N} \times 100\%.

XM2VTS database.
# of users J = 200
# of system = 2
A user-specific and selective multimodal biometric fusion strategy by ranking subjects, PRJ 2013
Performance on test set

10 of 15 fusion experiments shown here
HETEROGENEOUS INFORMATION FUSION
Subject dependency

Sample quality

Spoof attacks

Performance measurement

System design (mitigation strategies)
Three sources of information

Biometric sample quality

Cohort information

Client-specific information

Feature \rightarrow Classifier \rightarrow Normalisation

Quality assessment

Claimed ID \rightarrow Client-specific parameters

Cohort analysis

y \rightarrow q

c \rightarrow u
One system

Feature → Classifier → y → Heterogeneous Information Fusion

Quality assessment

Cohort classifiers

Classifier

Classifier

Cohort analysis

Claimed ID

Client-specific parameters
Representing the information sources

\[
\begin{align*}
    s_Q & \equiv [q^\text{tmplt}, q^\text{qry}] \\
    & \text{Quality of template} \\
    & \text{Quality of probe} \\
    s_C & \equiv [\mu^C, \sigma^C] \\
    & \text{Mean of cohort scores} \\
    & \text{Std. dev. of cohort scores} \\
    s_{US} & \equiv [\mu^\text{US}, \mu^\text{US}, \sigma^\text{US}] \\
    & \text{Mean of US impostor scores} \\
    & \text{Std. dev. of US impostor scores} \\
    & \text{Mean of US genuine scores} \\
    s_y & = [y] \\
    & \text{Matching score}
\end{align*}
\]
Cohort-based approach

Claimants (1 template each)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1^G$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>$Y_2^G$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>$Y_3^G$</td>
</tr>
</tbody>
</table>

Cohort users

queries

$\mu, \sigma$
Using cohort scores for normalisation

Extract features

Closest cohort  Runner up  3rd closest

gradient
Representing the information sources

- **Biometric sample quality**
  \[ s_q \equiv [q_{\text{tmpl}}, q_{\text{qry}}] \]
  - Quality of template
  - Quality of probe

- **Cohort information**
  \[ s_c \equiv [\mu^c, \sigma^c] \]
  - Mean of cohort scores
  - Std. dev. of cohort scores

- **Client-specific information**
  \[ s_{US} \equiv [\mu_{US}^{1}, \mu_{US}^{0}, \sigma_{US}^{0}] \]
  - Mean of US impostor scores
  - Std. dev. of US impostor scores
  - Mean of US genuine scores

- **Matching score**
  \[ s_y = [y] \]
Experimental setting

- Biosecure DS2 database
- Fingerprint systems:
  - NIST fingerprint matcher “Bozorth3”
  - Quality assessment module “NFIQ”
- Face systems:
  - Omnipercception SDK for face recognition and quality assessment
- Each subject provides 4 impressions per device and per finger
- Total of 4 impressions $\times 6$ fingers $\times 2$ devices = 48 impressions per subject
- 333 subjects are available
Results

There are 30 experiments. Do not try to read this.

Summarize the results, instead:

\[ \text{RPG} = \frac{\text{metric}_{\text{fusion}} - \text{metric}_{\text{baseline}}}{\text{metric}_{\text{baseline}}} \times 100\% \]
Preliminary results

- Multiple classifiers fusion
  - Naïve Bayes
  - Heterogeneous fusion
- Heterogeneous fusion
- Biometric sample quality
- Client-specific information
- Cohort information
- Baseline (score)

rel. change of EER(%)
Setting
Biosecure database
Face modality
Zero-effort attacks only

Heterogeneous information fusion

TRICKS OF THE TRADE
Generalized logit transform

Biometric matcher outputs are not normally distributed

\[ y' = \log \left( \frac{y - a}{b - y} \right) \]

Example:

\( y \) is the posterior probability of a genuine user given feature \( x \), i.e., \( y = P(G|x) \)

\( y = [a, b] \)

\[
\begin{align*}
y^{llr} & = \log \left( \frac{y}{1 - y} \right) = \log \left( \frac{P(G|x)}{P(I|x)} \right) \\
& = \log \left( \frac{p(x|G)}{p(x|I)} \right) + \log \left( \frac{P(G)}{P(I)} \right) \\
& = \log \left( \frac{p(x|G)}{p(x|I)} \right) + \text{const},
\end{align*}
\]
The effect of logit transform

<table>
<thead>
<tr>
<th>Original scores</th>
<th>Probabilistic-inversed scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Because some classifiers are MLPs with tanh function, these scores are not normally distributed.</td>
<td>Contrary to the left, these scores fit better using a single Gaussian distribution. Note the average correlation values in the title.</td>
</tr>
</tbody>
</table>