ICE Mining: Quality and Demographic Investigations of ICE 2006 Performance Results

P. Jonathon Phillips
NIST

Patrick J. Flynn
University of Notre Dame
(NIST Associate, 2007-2008)
Outline

• Motivation
• Quality measure correlation
• Quality measure effects on performance
  – Nested quality intervals
  – Disjoint quality quartiles
• Performance variations by demography
• Conclusions and comments

• Note: a report on this work is in preparation
Motivation

• Iris image acquisition typically expects highly controlled environment
  – Cooperative subject (minimize iris occlusion)
  – Active lighting
  – Active focusing
  – Standoff manipulation

• Strong texture contrast & focus yield subjective “good quality”
  – Strong texture filter responses
  – Reliable phase estimates
Iris Quality in the Literature

• Common biometric sample quality concepts
  – Fidelity vs. application-specific criteria for quality
  – Methodology for quality based performance analysis
    (Grother and Tabassi, PAMI 2007)
  – Subject and sensor effects on quality

• Iris-specific aspects
  – Focus (spectral content)
  – Occlusion (e.g., % iris), frontality, motion blur
Sample ICE 2006 iris subject session

LG EOU 2200 was industry recommended at the inception of data collection.
ICE 2006 data acquisition method

- Take a shot of 3 iris images
- If one or more is of sufficient quality, save all three
ICE 2006 Image Quality Reporting

Input Iris Image

Performers Image Quality Module

Integer {0..100}
ICE2006 Quality data

- Three competitive ICE 2006 performers (Sagem-Iridian, Cambridge and Iritech) (de-identified henceforth)
- 59,558 iris images
- Each image has three quality scores (one per performer)
Mining Quality: Generic properties
Should quality measures produced by different algorithms be correlated?
Quality measure scatter plots
Correlation of Quality scores table

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Pearson’s r</th>
<th>Spearman’s ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>A vs. B</td>
<td>0.122</td>
<td>0.131</td>
</tr>
<tr>
<td>A vs. C</td>
<td>0.349</td>
<td>0.348</td>
</tr>
<tr>
<td>B vs. C</td>
<td>0.120</td>
<td>0.108</td>
</tr>
</tbody>
</table>
Quality Score Correlation Between Eyes

Quality measure A
Cor Coeff = 0.362

Quality measure B
Cor Coeff = 0.896

Quality measure C
Cor Coeff = 0.790
Quality effects on matching performance
Step 1: Compute Image Quality

Input Iris Image → Performer’s Image Quality Module → Integer {0..100}
Step 2: Compute Quality Matrix

Contains quality scores for all possible comparisons

Quality score for a match pair is the minimum of the quality scores of its two signatures
Step 3: Compute global threshold on matching score

Compute threshold $\lambda$ that yields FAR = 0.001
Step 4: Prune matching scores by quality

Complete Similarity matrix

Subset by quality threshold

Fused quality threshold values: 5, 10, ..., 90, 95, 100
- 20 sub-experiments with nested sets of matching scores
- Compute FAR, FRR from global threshold $\lambda$
Calculation of FAR and FRR

• From unpruned set, compute threshold $\lambda$ that yields FAR = 0.001 (ICE 2006 operating point)

• Let $Q_F(g)$ and $Q_F(p)$ be the qualities of target and query samples $g$ and $p$

• Using $\lambda$, calculate FAR and FRR from all match pairs $(g', p')$ with $\min\{Q_F(g'), Q_F(p')\} \geq q$
Performance by Quality

ICE2006 B – false reject rate

ICE2006 B – false accept rate
Performance by Quality

ICE2006 A – false reject rate

- False reject rate (FRR)
- Fraction of samples discarded

ICE2006 A – false accept rate

- False accept rate (FAR)
- Fraction of samples discarded

Legend:
- A Quality
- B Quality
- C Quality
Covariate Analyses

• Response of quality algorithms to demographic subsets

• Response of vendor matchers to demographic subsets
Quality Covariate Study

- Race and eye color

- Three covariates
  - East Asian
  - Caucasian w/Light Eyes
  - Caucasian w/Dark Eyes

- Quality scores normalized
  - Empirical CDF
Quality Covariate Study
Classic Scoring of Results

Matching Scoring FRR = 0.020
FAR = 0.001

Matching Scoring FRR = 0.015
FAR = 0.001

Is the difference significant?
Re-sampling Techniques

Global Threshold $\lambda$
@ FAR = 0.001

Ellipse
Error Ellipse
Ellipse

- Is the level of difference significant?

Case 1: Yes

Case 2: No
Performance Covariate Study

- Race, eye color, eye

- Covariates
  - East Asian
  - Caucasian w/Light Eyes
  - Left eye
  - Right eye

- Measure effect
  - FAR
  - FRR
Performance variations by combination of matcher and demographic

• For each matcher
  – Compute similarity threshold that yields FAR = 0.001 for entire data set
  – For each demographic category in {East Asian, Caucasian Light eyes}
    • Divide match pairs with target and query in demographic category into 60 equal-sized subsets of matches
    • For each subset
      – Compute and plot FAR, FRR for each subset using global threshold
Performance Covariate Study

• First look
  – East Asian
  – Caucasian w/Light Eyes

• Four groupings
  – Left eye -- East Asian
  – Right eye -- East Asian
  – Left eye -- Caucasian w/Light Eyes
  – Right eye -- Caucasian w/Light Eyes
Performance Covariate Study

• Next look
  – Left eye
  – Right eye

• Four groupings
  – Left eye -- East Asian
  – Right eye -- East Asian
  – Left eye -- Caucasian w/Light Eyes
  – Right eye -- Caucasian w/Light Eyes
Quartile Quality Study

• Effect of Quality
  – FAR
  – FRR

• Bracket Quality by Quartile
  – High Quality Quartile
    • 25% highest quality samples
  – Low Quality Quartile
    • 25% lowest quality samples
  – Disjoint quality intervals; no matches in common
Algorithm B and Quality Measure B

![Graph showing the relationship between False reject rate and False accept rate for High Quality Quartile and Low Quality Quartile. The graph indicates a trend where higher False reject rates correspond with lower False accept rates for High Quality Quartile, whereas the Low Quality Quartile shows a more scattered distribution.](image-url)
Algorithm A and Quality Measure A

![Graph showing the relationship between False reject rate and False accept rate for High Quality Quartile and Low Quality Quartile.](image)
Error estimation: Data-imposed limitations

- Number of Non-Matches (impostors) in ICE 2006: **562,301,273**

<table>
<thead>
<tr>
<th>False accept rate</th>
<th>1:1000</th>
<th>1:1,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of false accepts</td>
<td>562,301</td>
<td>562</td>
</tr>
<tr>
<td>60 partitions</td>
<td>9400</td>
<td>9.4</td>
</tr>
<tr>
<td>Eye, race, eye color</td>
<td>1000</td>
<td>1</td>
</tr>
</tbody>
</table>
Observations and Conclusions

• Initial examination of ICE 2006 quality data
• Iris image quality affects performance (general trends, from aggregated ICE 2006 performance data)
  – FAR decays with quality @ fixed FRR
  – FRR nearly invariant for a range of quality ranks after an initial drop, at fixed FAR
• Also:
  – Demographic effects for quality measures
  – Demographic effects on FAR
• Non-match distribution affected by quality and demographics (not presented here)
Conclusions (contd.)

• Iris image quality measurement needs more research and thorough testing
  – Lack of correlation between three ICE2006 responders suggests that they were measuring different aspects of quality, or measuring them with different degrees of accuracy
  – Opportunities:
    • for further research
    • Fusion

• Quality is not in the eye of the beholder; it is in the recognition performance figures!
ICE Mining

• Should enable development of formal structural models, with specialized analyses
  – e.g., Generalized Linear Mixed Models
Thank You

Some say the world will end in fire;
Some say in ice.
From what I've tasted of desire
I hold with those who favor fire.
But if it had to perish twice,
I think I know enough of hate
To know that for destruction ice
Is also great
And would suffice.