Systematic Evaluation Methodology for Fingerprint-Image Quality Assessment Techniques

J. Hämmerle-Uhl, M. Pober, A. Uhl
Presenter: Christof Kauba

Department of Computer Sciences
University of Salzburg

May 30th, 2014
Overview

1. Introduction & Motivation
2. The Stirmark Toolkit
3. Experiments
4. Conclusion
1 Introduction & Motivation

2 The Stirmark Toolkit

3 Experiments

4 Conclusion
Fingerprint recognition robustness and sample quality

- Sample image quality impacts on recognition accuracy
- Skin conditions (e.g., dryness, moisture, dirt, cuts and bruises, ageing), sensor type and conditions (e.g., dirt, noise, size), user cooperation, crime scene preservation
- Benchmarking frameworks: FVC, BioSecure, SFinGe, StirMark [1]

**Essential:** Indices to reliably determine fingerprint image quality in various circumstances, but how to assess such proposed indices?

**Here:** Propose standardised tool to assess correlation between image quality indices and recognition accuracy on fingerprint image data representing a wide range of real-world acquisition conditions and quality levels (which are simulated by StirMark [1]).

Outline

1. Introduction & Motivation
2. The Stirmark Toolkit
3. Experiments
4. Conclusion
### Basic idea

The StirMark Benchmark is a generic benchmark test for evaluating the robustness of digital image watermarking methods. Many of the systematic errors introduced into data can be interpreted as specific fingerprint acquisition conditions.

- **Additive noise**: Actual dust on the fingerprint contact area, sensor noise, grainy surface a latent fingerprint has been taken off.
- **Median Cut filtering**: Simulates blur in fingerprint images, e.g. smudgy fingerprints (too much moist) etc.
- **Remove Lines and Columns**: Sensor errors, esp. sweep sensors can be affected by line removal (examples are shown).
- **Rotation**: Omnipresent challenge in fingerprint recognition.
- **Stretching**: A higher force applied when pressing the finger onto the contact area, in forensics a soft or flexible surface.
- **Shearing**: Simulates a setting where the applied pressing force is not perpendicular to the contact area.
- **Random distortions**: Modelling e.g. unevenly distributed pressure or a latent fingerprint scanned from an uneven surface.
StirMark Examples

(a) Noise (level 15)
(b) Median Cut Filter (size 9)
(c) Rotation of $-15^\circ$
(d) Stretching $(d = 1.350)$
(e) Shearing $(b = c = 0.20)$
(f) Random Dist. (lrnddist 4.2)

Figure: Examples for the application of StirMark tests.

Andreas Uhl: Systematic Evaluation Methodology for Fingerprint-Image Quality Assessment Techniques
Real Examples

Figure: Examples for distortions from actual acquisition problems.

(a) Missing lines

(b) Warping effects
Fingerprint quality indices assessment strategy

- Generate a large corpus of test data exhibiting various quality levels
  - Start with an available dataset of the target sensor (or multiple sensors of interest)
  - Apply StirMark image manipulations of different types in various intensities
- Conduct fingerprint recognition experiments on these data with different types of feature extraction / matching algorithms
- Correlate recognition result parameters (e.g. EER) to the quality index values of different strength intensities
- In this manner, specific strengths and weaknesses of candidate quality indices can be identified
Types of Fingerprint Matchers

**Correlation-Based Matcher** Use the fingerprint images in their entirety, the global ridge and furrow structure of a fingerprint is decisive. Images are correlated at different rotational and translational alignments.

**Ridge Feature-Based Matcher** Deal with the overall ridge and furrow structure in the fingerprint, yet in a localised manner. Characteristics like local ridge orientation or local ridge frequency are used.

**Minutiae-Based Matcher** The set of minutiae within each fingerprint is determined and stored as list, each minutia being represented (at least) by its location and direction. The matching process then basically tries to establish an optimal alignment between the minutiae sets.
Experimental Settings: Data & Recognition Software

- DB1 - DB3 from the FVC2004 data are used in a verification setting employing the evaluation protocol as specified by FVC.

- Fingerprint matching software
  - 3 minutiae-based schemes [2], including “NIST Biometric Image Software” (NBIS) package (mindtct and bozorth3). Due to the similarity of the results, average EER is given only.
  - **Phase only correlation (POC)**, custom implementation. First, the normalised cross spectrum (or cross-phase spectrum) of the DFT of the two images is computed. The POC is then obtained by taking the inverse DFT of the normalised cross spectrum.
  - **Fingercode (FC)**, custom implementation. A Gabor filter bank is applied to the orientation image resulting in a “Ridge Feature Map” which is translationally and rotationally aligned for matching.

- EER is used to assess recognition accuracy

---

Experimental Settings: StirMark Settings & Quality Indices

- StirMark Settings: We use 12 types of StirMark manipulations, each of which with 3 to 10 intensity levels, overall 91 different manipulations per image.

- Quality Indices

  - *nfiq*: Part of the *NIST Biometric Image Software* (NBIS) package. It relies on information produced by the minutiae detector *mindtct* and basically conducts a neural net-based classification of the minutiae-vector into one of five overall fingerprint quality classes.

  - *SpatDom*: Based on determining the block-wise clarity of the ridges and furrow orientation in the spatial domain. Per foreground-block, the gradient vectors of the gray level intensities are used to build a covariance matrix, based on which the *normalized coherence measure* is computed, which is then combined for all foreground-blocks in a weighed sum.

  - *FreqDom*: Image quality is defined in terms of energy concentration within a specific frequency band containing ridge frequency, which is measured in terms of entropy.
With respect to matching StirMark manipulated fingerprint images, the enrolled gallery image is taken in its original (non-manipulated) version, while the probe image involved in matching is the manipulated one.

Thus, for each matching scheme, for each manipulation type, and for each manipulation strength we are able to compute the corresponding EER, overall 91 EER values per database per matching scheme, arranged in lists of increasing manipulation strength.

We generate quality indices for all involved images and generate mean values per manipulation type and intensity level. From these data, we generate lists of quality mean values, ordered by manipulation intensity as well.

Finally, we compute *Spearman’s rank order correlation* per manipulation type, per quality measure, per database, and per fingerprint matcher.
We screen the results for quality indices where correlation is close to zero (low) for some data sets or matcher types, while it is clearly positive or negative (high) for other conditions.

Existence of such results does indicate that quality indices need to be assessed for each data set and matching scheme separately, while the absence of such results allows to draw data and matcher-independent conclusions.
Results: Minutiae-based schemes vs. FC on DB1

(a) Minutiae-based matchers

(b) Fingercode (FC)

Figure: Mean correlation for DB1.

→ significant inter-matcher variability is observed for \textit{nfiq} (for rml1)
Results: FC vs. POC on DB1

(a) Fingercode (FC)

(b) Phase-only Correlation (POC)

Figure: Mean correlation for DB1.

→ *FreqDom* exhibits inter-matcher variability (for *noise*)
Results: Minutiae-based vs. FC on DB2

(a) Minutiae-based  
(b) FC

Figure: Mean correlation for DB2.

→ obvious inter-matcher variability for \textit{nfiq}  
→ comparing minutiae-based and FC schemes on DB1 & DB2, significant inter-data variability is found for all three matching types
Results: Minutiae-based vs. FC on DB3

(a) minutiae-based

(b) FC

Figure: Mean correlation for DB3.

→ inter-data set variability: nfiq results correspond better to DB2, the behaviour on DB1 is much different.

→ inter-matcher variability: For SpatDom, low correlation values are seen with FC matching for medianCut, convMean, convGauss, while high values are obtained for the minutiae-based matchers.
Outline

1. Introduction & Motivation
2. The Stirmark Toolkit
3. Experiments
4. Conclusion
Lessons learnt:

- We observe significant inter-data set variability as well as cases of significant inter-matcher variability with respect to rank order correlation between recognition accuracy and fingerprint image quality.

- Such effects are found for all three example fingerprint quality indices.

We reveal that fingerprint image quality indices need to be related to different data sets (or sensors) and different fingerprint matching schemes separately, which can be done efficiently with the proposed methodology.

A general purpose fingerprint quality index reliably applicable to any sensor / matching scheme combination does not seem to have come into existence so far.
Questions?