Interoperability in Fingerprint Recognition: A Large-Scale Empirical Study

Luca Lugini, Emanuela Marasco, Bojan Cukic
Lane Department of
Computer Science and Electrical Engineering
West Virginia University
Morgantown, WVU (USA)
{emanuela.marasco, bojan.cukic}@mail.wvu.edu
lulugini@mix.wvu.edu

Ilir Gashi
Centre for Software Reliability
City University London
London, United Kingdom
i.gashi@csr.city.ac.uk

***Abstract*—Biometric systems are widely deployed in governmental, military and commercial/civilian applications. There are a multitude of sensors and matching algorithms available from different vendors. This creates a competitive market for these products, which is good for the consumers but emphasizes the importance of interoperability. Interoperability is the ability of a biometric system to handle variations introduced in the biometric data due to the deployment of different capture devices. The use of different biometric devices may increase error rates. In this paper, we perform a large-scale empirical study of the status of interoperability between fingerprint sensors and assess the performance consequence when interoperability is lacking.**

I. INTRODUCTION

Fingerprint based user authentication is one of the most prolific commercial branches of biometrics. Since authentication process needs two samples from each user, most systems need to anticipate that the device used for a user's enrollment (creation of the so called gallery image or template) may not be the same as the device used at the time of identification or identity verification (so called probe image or template). Fingerprints can be acquired through different Live-scan sensing technologies belonging to three main families: *optical*, *solid-state* and *ultrasound* [1]. In optical sensors, the finger is placed on the surface of a transparent prism which is typically illuminated trough the left side and the image is taken through a camera. The light entering the prism is reflected at the *valleys* and absorbed at the *ridges* of a fingerprint. In *solid-state* devices, the finger is modeled as the upper electrode of the capacitor, while the metal plate is modeled as the lower electrode. The variation in capacity between valleys and ridges can be measured when the finger is placed on the sensor. In the case of swipe *solid-state* sensors, impressions are obtained by swiping the finger on the surface of the sensor. *Ultrasound* sensors exploit the difference of acoustic impedance between the skin of the ridges and the air in the valleys of a finger.

Even within the specific sensing technology, the acquisition may vary across sensors [1]. Different arrangements of sensing elements in each device may introduce variations and distortions in the biometric data. In particular, differences in resolution and scanning area impact the *feature set* [1] extracted from the acquired image. A biometric matching system is required to handle variations introduced in the biometric data due to the deployment of different devices [2]. When the acquisition of the gallery and the probe samples is done using different biometric devices, the reliability of the biometric matcher may be reduced [4]. While such diversity is to be expected, commercial fingerprint matchers typically show a decrease in inter-device performance. A realistic scenario where the sensor interoperability is important is the US VISIT [2] program, deployed at US international airports. In this application, fingerprints are currently enrolled using a 500 dpi optical sensor with a sensing area of 1.2" x 1.2". As different devices may be used for enrollment and then verification, the lack of interoperability between the devices is a significant concern. Interoperability grows in importance as the scale of adoption of biometric devices and the pace of innovation increase: older biometric devices get replaced with newer designs, but the samples enrolled with older devices remain in operational use.

In this paper we report early results from a large scale study of the interoperability of fingerprint devices. We captured fingerprints of 494 participants using 4 different two-dimensional biometric devices[3]. This is a large sample because we are dealing with human subjects and follow a properly approved collection protocol which requires volunteers to dedicate 1 hour of time for which they are adequately compensated.

We found that the genuine matching scores, the scores that reflect a similarity between two different samples collected from the same person, were generally higher when both images where captured using the same device, compared with cases where different devices are used for capture of the two samples. We also found that *false-non-match-rates,* the failures to determine that two samples come from one user, were affected by capture device diversity. Conversely the *false-match-rates*, representing instances in which fingerprints from two individuals are found to be sufficiently similar to declare them a match, were not. We

[1] A *feature set* is composed by characteristics describing the object to be classified. It is expected to be representative with respect to the classes of the problem.

[2] http://www.dhs.gov/us-visit-office

[3] We also captured data using three 3D devices, but have not yet analyzed the data.

also found that the similarity scores are in general much more sensitive to the quality of the fingerprint image when different devices are used than in cases when images come from the same device.

While most of the preliminary findings have not been a big surprise, the analysis we have done allows us to precisely quantify the effects from the (lack of) interoperability between fingerprint sensors. To the best of our knowledge, this is the first such systematic study able to arrive at statistically significant results due to a sufficiently large number of participants and a variety of fingerprint scanners.

The rest of the paper is organized as follows: in Section 2 we summarize the related work in interoperability of fingerprint devices; Section 3 describes the experimental methodology; Section 4 contains an analysis of the results; Section 5 contains a discussion of the results, conclusions and provisions for further work.

II. Related Work

Recent works have pointed out the importance of investigating the impact on the error rates when capturing fingerprints with a new device [7].

Poh *et al.* proposed methods to mitigate effects due to a device acquisition mismatch scenario [10]. They investigated the problem of comparing a biometric template to a query that is generated from a different or unknown biometric device. The problem was modeled in terms of a Bayesian Network used to estimate the posterior probability of the device d given quality measures q, referred to as $p(d|q)$. The device is represented by a discrete variable whose values depend upon how many devices are available for training and it is observed during the training. The term $p(d|q)$ of the network is estimated using the Gaussian mix model (GMM) based on training data. During testing, the device is unknown and it can be inferred based on the quality measures extracted from the images. They demonstrated that their approach improves the performance of a unimodal biometric system by estimating a more accurate decision threshold.

Jain and Ross analyzed the problem of the interoperability of a biometric system in terms of the variability introduced in the feature set by different sensor technologies (e.g. optical vs. capacitive) [6]. They reported an Equal Error Rate (EER) of 23.13% when matching images acquired with Digital Biometrics and Veridicom sensors, and EER of 6.14% and EER 10.39% when using only Digital Biometrics and Veridicom, respectively. It is important to note that the sensors in our study are significantly higher in quality than those in [6].

Ross and Nadgir highlighted the importance of comparing feature sets obtained from different sensors [2]. Features extracted from fingerprint images (e.g., minutiae points) are impacted by resolution, scanning area, sensing technology, etc.; subsequently, the template stored in the database is affected too. They identified two possible approaches for addressing the problem of interoperability in the context of fingerprints: *i)* distortion compensation model based on the sensing technology of a specific biometric device; *ii)* inter-sensor compensation model which computes the relative distortion between images acquired using different devices. In their approach, the inter-sensor distortion is modeled by a thin-plate spline in which parameters rely on control points manually selected in order to cover representative areas where distortions can occur in the fingerprint image.

Campbell and Madden conducted a study to understand the causes of the lack of interoperability. The objective was to determine both *native* (enrollment and verification using the same device) and *non-native* or interoperable (enrollment and verification using different devices) False Match and False Non-Match rates. 60,902 fingerprint images over 10 products were used for the evaluation. The main goal was to test which products could interoperate at levels of 1% False Accept Rate (FAR) and 1% False Reject Rate (FRR). Results demonstrated that only 2 products out of 10 were able to interoperate at the specified levels [12].

III. Experimental Setup

A. Dataset

The dataset we use was collected in 2012 in West Virginia University. Data were collected from each participant using multiple devices, all based on optical sensors. The order of use of fingerprint scanners was the same for all participants. 494 participants were randomly selected. They provided information on age (53% varying between 20 and 29 years old) and ethnicity (57.2% of the population is Caucasian). This is summarized in Figure 1.

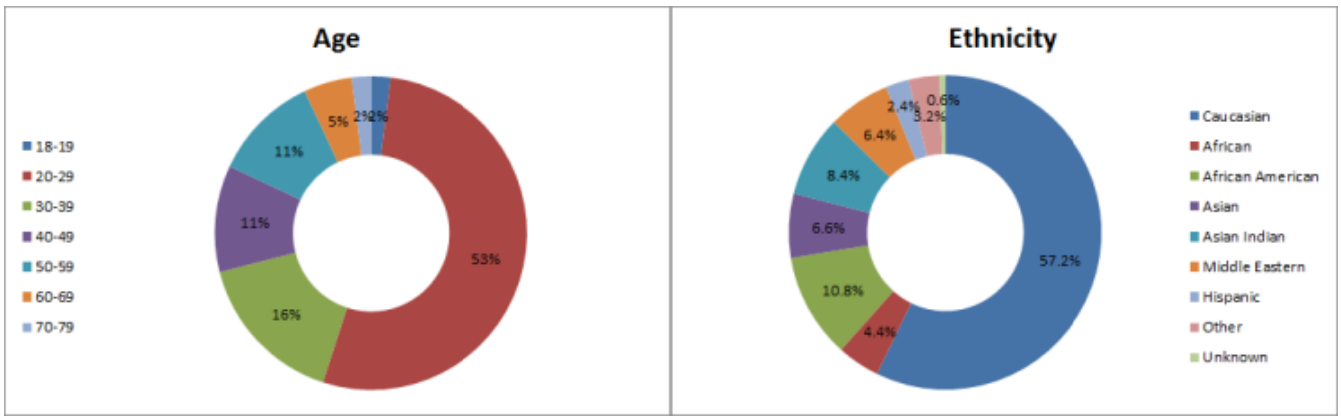


Figure 1: Age and ethnicity groups of the participants to the data collection.

Fingerprints were acquired using four Live-scan devices (D0 – D3, see Table 1) and ink-based ten-print cards (D4). Ten-print cards were scanned at resolutions of 500 dpi using a flat-bed scanner. Ink-based fingerprints were acquired at the end, to not affect the quality of Live-scan fingerprints. For each Live-scan device users provided two *sets* of fingerprints subsequently, each consisting of: rolled individual fingers on both hands, left slap (i.e. slapping the four (non-thumb) fingers on the device), right slap, and thumbs slap. The optical technique utilizes a glass platen, a laser light-source and a Charge-Coupled Device (CCD) or a Complementary Metal–Oxide–Semiconductor (CMOS) camera for constructing fingerprint images. The finger is placed on the glass plate, a laser light is reflected through the prism and facilitates the imaging. Fingerprints were collected without controlling the quality and the centering of the finger.

	Device	Model	Resolution (dpi)	Image size (pixels)	Capture area (mm)
D0	Cross Match	Guardian R2	500	800 x 750	81 x 76
D1	i3	digID Mini	500	752 x 750	81 x 76
D2	L1 Identity Solutions	TouchPrint 5300	500	800 x 750	81 x 76
D3	Cross Match	Seek II	500	800 x 750	40.6 x 38.1

Table 1: Characteristics of the Live-scan devices used for the fingerprint acquisition carried out in this study.

Image quality was assessed using the NIST (National Institute for Standard Technology) Fingerprint Image Quality (NFIQ) algorithm [4]. NFIQ is an open source tool developed by NIST, and has become the industry standard for fingerprint image quality assessment. Fingerprint quality is classified into five levels, 1 (highest) to 5 (lowest).

Match scores were generated using the Identix BioEngine Software Development Kit [9] matching algorithm. A matching algorithm compares two fingerprint images and returns a *score* based on how similar it thinks the two templates are. The higher the score the more likely it is that the two images / templates come from the same finger. The main aim of our study is then to compare these scores in two matching scenarios: *i)* comparing two fingerprints captured with the same device, and *ii)* comparing two fingerprints captured with two different devices. The notation reflecting the types of similarity match scores is given in Figure 2. Since the total number of impostor scores could be very large, we limited it to a random subset which is still sufficient for statistical confidence. For the DMG case, we only consider the four Live-scan devices because only 1 *set* of fingerprints was collected from each subject on ten-print cards, making the matching between ink-based prints impossible. Table 3 reveals the number of scores in each category.

Device Match Genuine (DMG): Genuine match scores are generated when we match the same user's right point fingers. The image captured in the first user's interaction with a sensor is stored in the *gallery* (the database of fingerprint images in which we search). The image acquired using the same device the second time is called the *probe* (the set of images we submit for identification or verification). Since we have *494* participants and *4* devices (for ink-based ten print cards we only have one image) the total number of DMG scores is *1,976*.
Device Match Impostor (DMI): Impostor match scores are generated by matching the fingerprint image / template of a participant against those of all the other participants. DMI scores include only those in which both the gallery and probe images are acquired using the same device. The number of imposter scores is potentially very large. We limit our analysis to randomly obtained 120,855 DMI match scores.
Diverse Device Match Genuine (DDMG): Genuine match scores generated when gallery and probe images acquired using different devices. For each subject, having 5 collection sensors, we have 10 possible combinations with two match scores for each probe, resulting in *9,880* match scores.
Diverse Device Match Impostor (DDMI): Impostor match scores generated using images from different devices.

Table 2: Notation table for similarity score computations.

Matching	Subjects	Number of devices	Samples	Similarity scores
DMG	*494*	*4*	*2*	1,976
DDMG	*494*	*5*	*2*	9,880
DMI	*494*	*5*	*2*	*120,855*
DDMI	*494*	*5*	*2*	*483,420*

Table 3: Match score for different match scenarios.

IV. Analysis of Results

A. Overview

In this section we present preliminary results of our interoperability analysis. Figure 2 shows the distribution of DMG scores (in blue) and DMI scores (in red) for the Cross Match Guardian R2 device. As expected, most of the DMG scores are high (appearing on the right-hand side of the

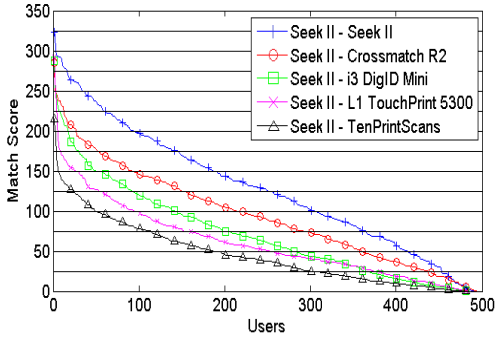


Figure 2: Genuine match scores (DDMG), ordered by the magnitude, for different sensor probe images vs. Seek II gallery fingerprints.

distribution) and most of the DMI scores are low. No DMI scores are higher than 7, but there are some DMG scores below 7. Hence for a given system the decision on where to place the threshold between genuine and impostor scores will depend on the relative costs difference between *false match* and *false non-match*.

Figure 3 shows the distribution of DDMG scores (in blue) and DDMI scores (in red) when matching fingerprint images acquired with the Cross Match Guardian R2 for enrollment, and the i3 digID Mini for verification. The overlap of genuine and impostor score distributions is greater when they were acquired from diverse sensors. Reader may note that a substantially higher number of genuine scores is less than 7, though very few imposter scores are high too. This implies that the use of diverse devices may result in a higher number of false non-matches. This observation is quite consistent across all the diverse pairs we analyzed. The impostor scores never go higher than 7, but the number of genuine scores with values of less than 7 is higher in diverse vs. non-diverse sensor choices.

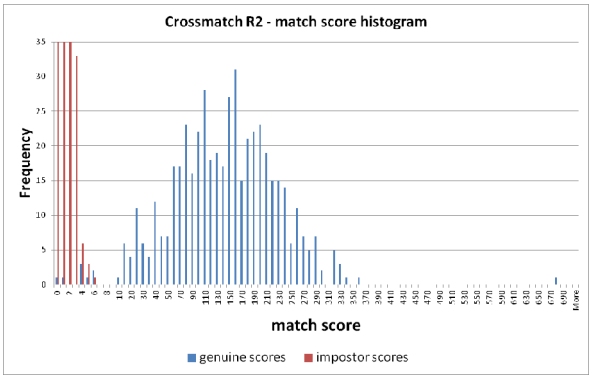


Figure 3 – Histogram of the DMG and DMI match scores for the Cross Match Guardian R2. The frequency of the DMI scores for the range 0-1 is *18,721*, for 1-2 is *5,121* and for 2-3 is *296*.

B. *Impact of the sensor interoperability*

Figure 4 shows the genuine match score distribution when matching probe fingerprints acquired using all devices against the gallery of fingerprints acquired using the Cross Match Seek II sensor (i.e. DDMG). The figure confirms that the match scores are the highest when measuring the similarity between images acquired by the same sensor. For all other sensor combinations the scores are lower, with the lowest match scores representing the similarity with the ink-based ten-print scans as probes. Matching scores of any Live-scan devices are higher than those obtained from ten-prints. We observed the same trends when using other fingerprint sensors for gallery images.

C. *Statistical Analysis of Sensor Interoperability*

In order to estimate the degree of change in genuine match scores across different sensors, we carried out the Kendall's rank correlation statistical test. We compare the

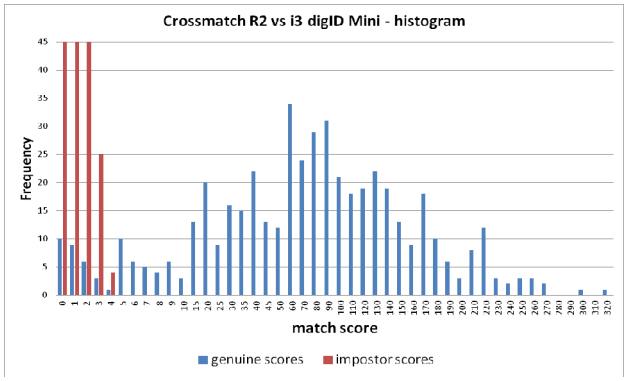


Figure 4: The histogram of the DDMG and DDMI scores for the Cross Match Guardian R2 vs. i3digID Mini. The frequency of the DDMI scores in the 0-1 range is 19,889, for 1-2 is 4,024 and for 2-3 is 229.

scenario in which the gallery and probe are acquired using the same device (DX vs. DX) to the scenario where gallery and probe images are acquired using different devices (DX vs. DY). Table 4 shows the p-values under the null hypothesis of interoperability scenarios. If the p-value is further from zero, the DDMG genuine scores are significantly different. If the p-value is close to zero, the DMG and DDMG scores do not differ significantly.

	DX-D0	**DX-D1**	**DX-D2**	**DX-D3**	**DX-D4**
D0	5.42 e-242	5.32 e-93	1.24 e-84	1.29 e-66	1.04 e-07
D1	2.72 e-68	6.19 e-242	2.99 e-65	2.35 e-59	2.59 e-06
D2	7.11 e-69	6.02 e-01	5.47 e-242	7.79 e-55	2.41 e-08
D3	2.14 e-76	6.28 e-01	5.62 e-01	5.47 e-242	3.03 e-08

Table 4: p-values from Kendall's rank correlation statistical test.

Results shown in Table 4 indicate a statistically significant difference for sensor pairs {D2,D1}, {D3,D1}, {D3D2}. Further, genuine match scores generated from the matching of a ten-print probe against a gallery acquired by any of the four Live-scan devices are very distant from those generated in any of the scenarios where the optical devices are used for fingerprint acquisition. It is interesting and surprising, however, that the results of Kendall's rank test are not symmetric.

The interoperability related to the False Non-Match-Rate (FNMR) matrix is shown in Table 5. Rows list the device used for enrollment and columns list the devices used for the capture of probe images. The values along the diagonal indicate performance when enrollment and verification fingerprint images are taken from the same device. The values off the diagonal refer to the system performance when gallery and probe fingerprint images are acquired by different sensors. The FNMR in intra-device match scenarios were found to be lower than those in inter-device matching. The exceptions are data sets {D1,D1} and {D3,D3}, for which the FNMR are 0.0024% and 0.0018% respectively. In particular D2 presents a larger image size with respect to D1; the capture area of D3 was smaller compared to the capture area of the other devices, resulting in anomalies.

FNMR at fixed FMR of 0.01%					
	D0	**D1**	**D2**	**D3**	**D4**
D0	2.70E-04	1.89E-03	1.62E-03	6.07E-04	2.90E-03
D1	1.89E-03	2.43E-03	2.36E-03	2.16E-03	4.93E-03
D2	1.75E-03	2.16E-03	1.62E-03	1.82E-03	4.05E-03
D3	6.75E-04	2.16E-03	1.89E-03	1.75E-03	3.31E-03
D4	2.70E-03	4.72E-03	4.05E-03	2.97E-03	1.35E-04

Table 5: Interoperability FNMR matrix.

D. Effect of Image Quality on the Scores

US National Institute for Standards and Technology (NIST) provides recommendations regarding quality control for fingerprint image acquisition [4]. The agency developed NIST Fingerprint Image Quality (NFIQ) software. It generates a number, in a range between 1 and 5, which predicts fingerprint matcher's performance as a function of image quality. The quality number reflects the predictive positive or negative contribution of an individual sample to the overall performance of a fingerprint matching system. NFIQ level 1 indicates a high quality fingerprint image, while level 5 indicates the poorest quality. The agency recommends that fingerprints be reacquired from the user up to three times, if the NFIQ quality of thumbs and index fingers is greater than three. Table 6 indicates the FNMR rates when the image quality is four or less. These FNMR rates are much worse than those reported for the entire experiment in Table 5. With respect to the differences in FNMR for intra and inter sensor scenarios, they simply appear unpredictable.

Figure 5 shows the frequency of matching scores lower than 10 for a given pair of image quality scores. Figure 5 (a) depicts DMG scores, while 5 (b) shows DDMG scores. When comparing probe and gallery images acquired by the same device, it seems that as long as one of the images has a quality score between 1 and 3, the frequencies of low matching scores are negligible. When acquisition reflects the use of diverse devices, to reduce the chance of getting a low genuine match score, there needs to be a more stringent control on image quality. Both the gallery and probe images need to be in the range 1-2 to reduce the incidence of genuine low match scores.

FNMR at fixed FMR of 0.1% for images with NFIQ quality < 3					
	D0	**D1**	**D2**	**D3**	**D4**
D0	0.000135	0.00027	6.75E-05	0	0.00054
D1	0.000202	0.000405	0.000135	0.00027	0.00027
D2	6.75E-05	0.000135	0.00027	0.000135	0.000472
D3	6.75E-05	0.000202	0.000135	0.000405	0.000405
D4	0.000405	0.000337	0.000472	0.000337	0.000135

Table 6: Interoperability FNMR matrix for fingerprint images with NFIQ quality below 3.

V. DISCUSSION, CONCLUSIONS AND FURTHER WORK

In this paper we presented initial results from a large scale study of interoperability between optical fingerprint acquisition sensors. Using fingerprint images collected from 494 participants with 4 different devices, plus the scanned versions of ink-based fingerprint imprints on ten-print cards, were able to study the match score distributions, false-match and false-non-match-rates in various scenarios.

Our preliminary findings show that the genuine match rates are always higher if the gallery and the probe image are acquired by the same sensor. The false-non-match-rates are affected by the use of different devices, indicating the impact of limited interoperability between biometric sensors. The false-match-rates do not seem to be affected by interoperability.

We also studied the effect of image quality on the FMR and FNMR. We observed that a significant number of low match scores appear when the system attempts to match the genuine fingerprint pairs, acquired either by the same device or different devices, in which at least one of them or both have a low NFIQ quality score. Nevertheless, when images are acquired by different devices, their quality scores become more important if we are to reduce the instances of low genuine match scores.

Most of the findings presented in this study are not surprising. The exception is the lack of similarity in the match scores and error rates when the sensor sources of gallery and probe images are swapped. Nevertheless, to the best of our knowledge, this is the first study that has systematically gathered fingerprint data from multiple top-of-the-line commercial sensors. Such data collection allowed us to obtain detailed measurements on the effects of the lack of interoperability.

The research is on-going and current and future plans for further work include:

- More detailed analysis on the effects of diverse matchers on interoperability. We especially want to explore examples where diverse matchers improve the detection rates even if the average FNMR and FMR rates may deteriorate when using different sensors.
- What advice can we prescribe for an overall architecture of fingerprint recognition that:
  - Employs diverse sensors, and/or
  - Improves interoperability.
- The effect of user habituation on the quality of the fingerprint samples obtained, and the effect they have on FMR and FNMR. In other words, do the quality of the images obtained improve when we compare, say, the first sample obtained from a participant with the last one.
- Statistical and probabilistic modeling to help us conceptualize the phenomena observed and allow for better predictive behavior. For example, being able to answer questions such as "*what is the probability that I will have a False Non-Match pertaining to a user enrolled using the Device X and verified using the Device Y?*".

- Using more than one fingerprint image from a given participant to improve the FMR and FNMR rates and overall Decision Making

ACKNOWLEDGMENT

This material is based upon work at West Virginia University partially supported by the National Science Foundation award number 1066197, and the National Institute of Justice award number 2010-DD-BX-K037. Ilir Gashi is partially supported by a Pump Priming Grant from City University London, and an EU Artemis initiative / UK TSB funded project SESAMO. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the sponsoring organizations.

REFERENCES

[1] A. Jain, D. Maltoni, D. Maio and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer, 2003.

[2] R. Nadgir, "Facilitating sensor interoperability and incorporating quality in fingerprint matching systems". Dissertation, West Virginia University, 2006.

[3] A. Ross, and R. Nadgir. "A calibration model for fingerprint sensor interoperability." *Proceedings of SPIE*. Vol. 6202. 2006.

[4] SP800-76, NIST Special Publication 800-76, Biometric Data Specification for Personal Identity Verification, February 2005.

[5] S. Modi., S. Elliott, and K. Hale, "Statistical analysis of fingerprint sensor interoperability performance." IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS), 2009.

[6] A. Ross and A. Jain. "Biometric sensor interoperability: A case study in fingerprints." Biometric Authentication (2004): 134-145.

[7] Vielhauer, C., Yanikoğlu, B., Garcia-Salicetti, S., Guest, R. M., & Elliott, S. J. (2008). Special section on biometrics: Advances in security, usability, and interoperability. Journal of Electronic Imaging, 17(1).

[8] S. Modi, "Analysis of fingerprint sensor interoperability on system performance", Diss. Purdue University, 2008.

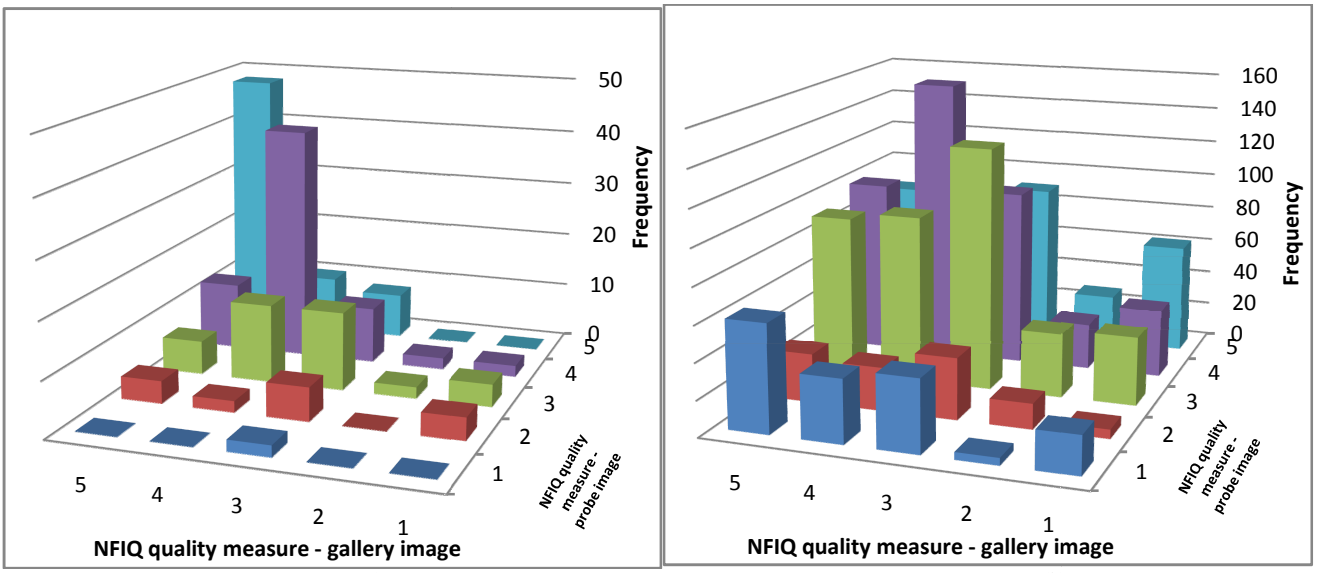
[9] http://www.l1id.com/pages/100-bioengine-sdk.

[10] P. Grother, et al; MINEX – Performance and Interoperability of the INCITS 378 Fingerprint Template; NISTIR 7296; National Institute of Standards and Technology; March 21, 2006.

[11] N. Poh, J. Kittler, and T. Bourlai, "Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization." First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), 2007.

[12] J. Campbell and M. Madden, "International Labor Organization (ILO) Seafarers' Identity Documents Biometric Interoperability Testing Report Number 3", 2009.

[13] B. Carterette. "On rank correlation and the distance between rankings". In Proc. 32nd SIGIR, pages 436-443, 2009.



(a) (b)

Figure 5: Histograms of genuine match scores below 10, grouped by the qualities of gallery and probe images. (a) indicates match scores obtained when using gallery images and probe images acquired using the same device; (b) indicates match scores obtained when using gallery image and probe images acquired using two different devices. When the device used for verification is different than that one used for ennrollment, the number of genuine match score <10 significantly increases. This lack of interoperability leads to an increase of the FNMR. This impact is higher when in the presence of a low quality gallery (quality value ranging from 3 to 5).