Identification performance of evidential value estimation for fingermarks

J. Kotzerke[*†], S.A. Davis[*], R. Hayes[‡],
L.J. Spreeuwers[†], R.N.J. Veldhuis[†], K.J. Horadam[*]

[*] School of Mathematical and Geospatial Sciences,
RMIT University, Melbourne, Australia

[†] Services, Cybersecurity and Safety,
University of Twente, Enschede, The Netherlands

[‡] Forensic Services Department, Victoria Police, Melbourne, Australia

{johannes.kotzerke, stephen.davis, kathy.horadam}@rmit.edu.au
robert.hayes@police.vic.gov.au
{l.j.spreeuwers, r.n.j.veldhuis}@utwente.nl

Abstract: Law enforcement agencies around the world use biometrics and fingerprints to solve and fight crime. Forensic experts are needed to record fingermarks at crime scenes and to ensure those captured are of evidential value. This process needs to be automated and streamlined as much as possible to improve efficiency and reduce workload.

It has previously been demonstrated that is possible to estimate a fingermark's evidential value automatically for image captures taken with a mobile phone or other devices, such as a scanner or a high-quality camera.

Here we study the relationship between a fingermark being of evidential value and its correct and certain identification and if it is possible to achieve identification despite the mark not having sufficient evidential value. Subsequently, we also investigate the influence the capture device used makes and if a mobile phone is an option worth considering.

Our results show that automatic identification is possible for 126 of the 1,428 fingermarks captured by a mobile phone, of which 116 were marked as having evidential value by experts and 123 by an automated algorithm.

1 Introduction

Increases in the rate of reported crime are evident in Victoria. Official recorded offences for the year 2012/13 have risen by 3.4% to 406,497, compared to 2011/12 [Vic13]. Forensic experts must travel in many cases to the crime scene and collect the evidence themselves, spending a lot of time travelling. Highly trained specialists such as fingerprint examiners are valuable resources, making streamlining of processes and the search for tools to assist both experts and non-experts in the field a priority. Therefore, we want to determine if fingermarks are of insufficient evidential value as early as possible to ensure the marks collected are of sufficient evidential value and to assist in case evidence collection for the specialists. This can be achieved by using mobile phones to capture fingermarks, determine their binary evidential value and transmit the valuable ones directly to the forensics unit; all done automatically either at the scene or at the lab after mark development/enhancement. This task can be performed by regular police officers or professionals with a different area of expertise, thus allowing the fingerprint experts to focus on the analysis of the fingermarks.

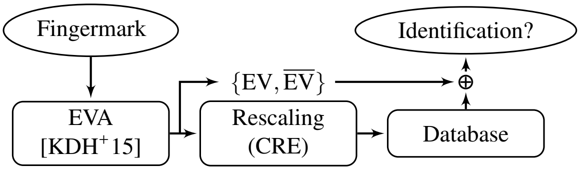


Figure 1: Diagram of the experiment performed. A fingermark is captured, its evidential value $\{\mathrm{EV}, \overline{\mathrm{EV}}\}$ is estimated by the Evidential Value Algorithm (EVA) of [KDH+15]) and rescaled in the same way as in EVA. The number of correct and certain identifications (ccID) of the mark matched to a reference database is measured w.r.t. the image capture device and evidential value estimation method.

Previously, Kotzerke *et al.* have established that e.g. mobile phone images are suitable to estimate if a fingermark is of sufficient evidential value (EV) and that an automated algorithm (EVA) can achieve results close to an expert assessment, based on the image quality [KDH+15]. Now, we extend this work and investigate the following worst case scenario. Are there any marks, which can be automatically and with certainty identified (against a reference database we collected) but are not of EV according to either the algorithm or the expert assessment from [KDH+15]? The proposed experiment is shown in Figure 1.

1.1 Background

Fingermarks are of essential value in order to exclude or to identify suspects. Nowadays, law enforcement agencies rely heavily on the fingermark via automatic systems such as IAFIS and forensic experts [Mal09]. These examiners are expected to follow the Analysis, Comparison, Evaluation, and Verification (ACE-V) protocol [Ash99]. During the analysis phase, they decide if the mark at hand is of value for individualisation (VID), value for exclusion only (VEO) or no value (NV). Those with VID or VEO are EV; those with NV are $\overline{\mathrm{EV}}$.

However, fingermarks suffer often from low quality due to being smudged or partial, overlap with other marks [FSZ12], or distorted by the surface pattern of the object they are found on [SHAF11]. Their forensic value is difficult to grasp for non-experts. Ulery *et al.* show that accuracy and repeatability varies even for forensic experts and mostly depends on the print quality [UHBR11, UHBR12], especially for borderline decisions. Consequently, Kellman *et al.* use image features to predict “expert performance and subjective assessment of difficulty in fingerprint comparisons" [KME+14].

Most quality measures are used to prevent low quality images from being automatically matched because they tend to produce false minutiae and consequently false matches [AFFOG+07]. Therefore, they are suited to operational law enforcement agency setups and only optimised and tested for contact scanners [CDJ05, FKB06, LCCK08, The13] but not fingermarks. This has resulted in various algorithms tuned to a capture resolution of 500 ppi.

On the other hand, fingermarks require robust methods to estimate their quality because all factors mentioned above will vary and influence the quality and its estimate. Yoon and Jain demonstrated in [YJ13] that the current NIST quality estimator reference implementation NFIQ1 is outdated because IAFIS was able to return the print's mate although it has been classified to have the lowest possible quality. Currently, NFIQ2 [The13] is under development and closing this gap; it is scheduled to be released soon. However, it is still primarily developed for fingerprints captured at a known resolution. In a scenario where the capture resolution is unknown, an estimate based on image features can improve the performance significantly [KDH+15].

Despite the need to reject low quality fingermarks for matching to avoid false identifications, the proposed scenario takes place much earlier. It includes the danger of missing a potentially valuable mark, which could solve a criminal case, because according to some algorithm the mark isn't of EV. Naturally, there is a trade-off involved between the likelihood of missing some important marks and capturing as few marks as possible.

1.2 Outline

We investigate how the (estimated) EV of a fingermark influences an identification scenario, the performance achieved w.r.t. the capturing device (scanner, high-quality camera, phone) and capture resolution estimation (CRE) algorithm used (cf. Figure 1) and if any certain identifications would be missed if only marks of EV were to be analysed.

In the following sections we set up the methodology used during the identification scenario (Section 2), elaborate on the databases employed, perform identification experiments to demonstrate the interplay between a mark's correct and certain identification (ccID) and if it is of EV, and discuss the results (Section 3). Finally, we summarise our findings and their implications and point out the direction for our future research (Section 4).

2 Methodology

The main idea behind our experiment is to evaluate if fingermarks, which are not of EV and hence wouldn't be collected in a crime scene scenario, can be correctly identified with certainty (worst case). We now recap some important concepts relevant to the experiment, which have been introduced in [KDH+15].

As already motivated, there are scenarios when the capture resolution for an image is unknown because of an unconstrained setup. Most quality or feature extraction algorithms are optimised towards a certain resolution, most commonly 500 ppi and if the input image deviates from the assumed resolution, the applied algorithm usually falls short. Therefore it is sensible to perform a CRE. The RLAPS algorithm introduced in [KDH+15] estimates the inter-ridge spacing of a fingerprint or fingermark image and infers the capture resolution used. The power spectrum is computed and its radial average is determined only around its maximum peak within a certain frequency range. Finally, the assumption of an average inter-ridge spacing of 9 px for an adult is applied and leads to a capture resolution estimate.

Type of distortion	Number of marks taken	Prints of sufficient evidential value				
		Assessor 1	Assessor 2	Assessor 3	Ground truth	EVA
(i) light placement	168	48.2%	48.2%	48.2%	48.2%	54.2%
(ii) smeared	168	3.6%	4.2%	3.6%	3.6%	14.9%
(iii) finger twisted lightly	168	4.2%	4.8%	4.8%	4.8%	11.3%
(iv) strong twist	168	0.0%	0.0%	0.0%	0.0%	6.0%
(v) heavy placement	168	69.6%	65.5%	65.5%	65.5%	64.9%
(vi) partial, heavy placement	168	45.8%	48.2%	48.2%	48.2%	50.6%
(vii) normal	420	47.4%	49.0%	50.0%	49.0%	50.7%
Total	1,428	34.1%	34.5%	34.7%	34.5%	38.66%

Table 1: A breakdown of the 1,428 marks into the categories of distortion (including no deliberate distortion), and the final status of the assessment of the 3 experts in terms of the proportion of marks found to be of EV. Assessor 1's opinion regarding the marks of categories (iii) and (vii) are respectively 9 and 21 decisions short of the total number. However, the other assessors agree on those marks and therefore a clear decision on ground truth can be made via majority vote. The EV distribution for EVA has been calculated for the mobile phone images which have been rescaled using CRE Global and the fused quality feature set at the decision threshold corresponding to the EER because of its performance (cf. Figure 3 and [KDH+15]).

Furthermore in [KDH+15], the idea of sufficient evidential value has been introduced and an algorithm to compute it based on image features has been presented. The feature sets of NFIQ2 as specified in the preliminary definition guide [The13], Neurotechnology Verifinger 6.7 [Neu15] and its quality value and the number of minutiae and their Fusion (concatenation of their feature vectors) have been investigated. We refer to this specific estimation algorithm as EVA. For details see [KDH+15].

Finally, we would like to clarify the concept of ccID. Assuming that a fingermark is compared against a reference database containing N unique fingerprints, a verification score S_i is returned for every comparison. We define a decision as correct and certain if and only if the mark and the print with the highest score are from the same subject *and* if the largest score is larger by factor $d > 1$ than any other score:

$$\nexists S_j : S_i \leq dS_j, \quad i, j \in [1, N], \quad i \neq j. \tag{1}$$

This would lead to a correct and certain identification. One has to keep in mind that the smaller d is chosen, the greater the likelihood becomes that a decision is considered to be certain but is in fact due to low verification scores derived from poor quality images.

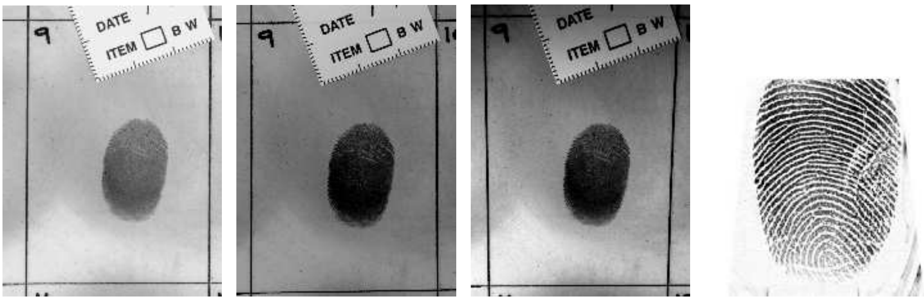
3 Experiments and discussion

In this section we evaluate the ability of an automatic system to perform a certain identification and how many of the fingermark images are considered to be of EV in the context of different capture devices such as a flatbed scanner, a high-quality camera and a mobile phone and their interplay with different CRE methods (cf. Figure 1).

First, we recapitulate the properties of the fingermark database and its ground truth from [KDH+15] and introduce our own reference database (Section 3.1), then we determine the identification performance and look at those images' evidential value determined by the experts and an algorithm w.r.t. the use of different CREs and image capture devices (Section 3.2). Finally we discuss our findings and their implications (Section 3.3).

3.1 Databases

Recently, Kotzerke *et al.* have introduced a pseudo fingermark database [KDH+15]; it consists of 1,428 normal and deliberately distorted fingermarks from two males and two females. In order to create the distorted marks, they defined six different distortion categories listed in the first column of Table 1. There are 168 marks per distortion category, the other 420 marks are "normal" and don't suffer from deliberate distortions (cf. Table 1).



(a) Scanner (b) High-quality camera (c) Mobile phone (d) Fingerprint scanner

Figure 2: A subject's right middle finger "heavily" placed on the sheet captured by different devices: scanner (a), high-quality camera (b), mobile phone (c) and fingerprint scanner (d). The first three images (a – c) have been cropped more closely before entering them into the database, the image captured with the fingerprint scanner (d) is used in the reference database and only shown here for reference purposes.

All fingermarks were left on a sheet of paper, brushed with magnetic black powder and laminated afterwards, under the supervision of a fingerprint expert. All sheets were digitised with 3 different capture devices: (i) a flatbed scanner (HP Scanjet G4010, abbr: Scanner), (ii) a high-quality camera (Nikon D3S with a Nikkor $f/2.8$ 60 mm-macro lens attached, abbr: DSLR) and (iii) a mobile phone (Apple iPhone 4S, abbr: Phone). It has to be noted that capture resolution for the mobile phone varies as it has been used in an unconstrained setup. However, its captures were taken perpendicular to the fingermark sheets and both capturing and lighting conditions were kept as consistent as possible. The high-quality camera was attached to an operational stand setup, which is usually used for police work. The estimated capture resolutions are 1200 ppi (Scanner), 460 ppi (DSLR) and 890 ppi (Phone). More details can be found in [KDH+15].

All laminated marks have been assessed by three Victoria Police experts who decided for each mark if it is of EV by undergoing at least a partial markup process. The ground truth is the majority vote of their assessment. The EV distribution can be found in Table 1.

In this research, we created a reference database to match the marks against. For this purpose, we collected all ten fingerprints of the same subjects as found in the fingermark database with a Digital Persona U.are.U 4000 fingerprint scanner. We captured one image per print without any deliberate distortion to imitate a reference scenario (cf. Figure 3.1). Also, we added imposter images with alike characteristics (no deliberate distortion) which were all captured with optical fingerprint scanners similar to the one we employed. Specifically, we used all third prints of FVC2000 DB3 [MMC+02a] and FVC2004 DB2 [MMC+04] and all sixth prints of FVC2002 DB1 [MMC+02b]. This leads to a reference database consisting of 40 genuine and 330 imposter prints. We verified via the cross verification scores that there are no duplicates included.

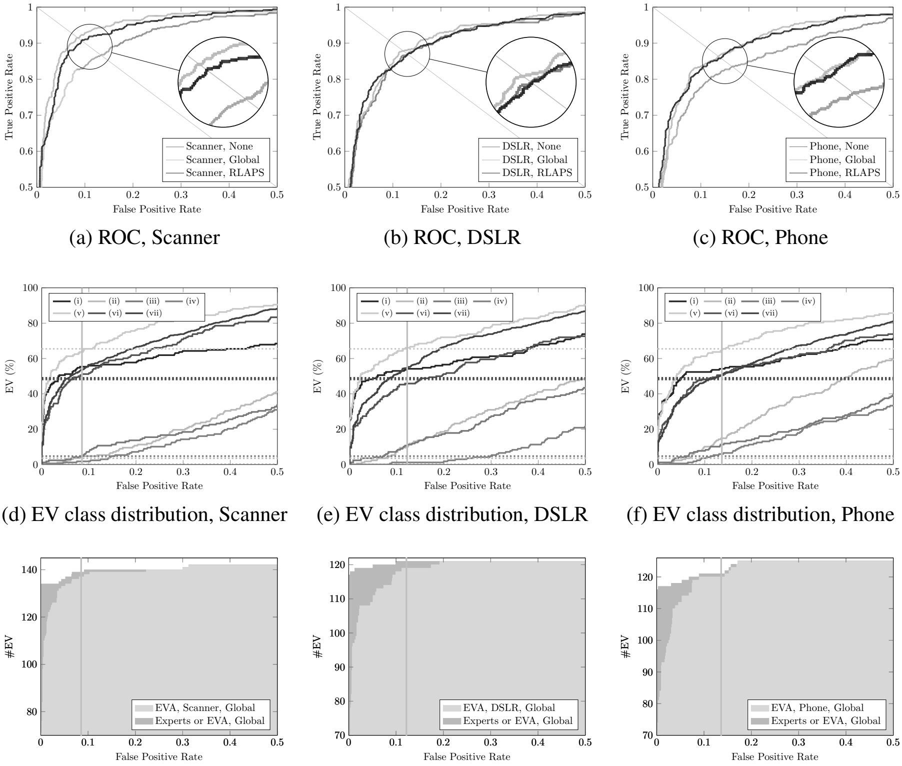
3.2 Experiment

This experiment aims to investigate the relationship between a fingermark, which can be automatically identified with a high certainty and the evidential value assigned to it by experts or EVA (cf. Figure 1).

The EVA is predominantly influenced by (i) the image properties such as capture device used and CRE and hence (ii) the quality features extracted as briefly discussed in Section 2. The feature sets of NFIQ2, Verifinger and their fusion (concatenation of their feature vectors) have been investigated. In this context, we use EVA from [KDH+15] to obtain the estimated evidential values for three different CREs (None, Global, RLAPS) and three capture devices (Scanner, DSLR, Phone) for all 1,428 fingermarks. This process is extensively described in [KDH+15].

Next, a verification score for every fingermark matched against every print in the reference database is computed. This verification process is performed by a commercial fingerprint extractor and matcher, Neurotechnology Verifinger 7.0. We consider it to be ccID if and only if the highest score returned for one particular fingermark is from the comparison to the same subject's finger and the second highest score multiplied by $d = 1.5$ is still smaller than the highest one (see Equation 1). The factor d has been empirically chosen in order to ensure a reliable and certain estimation due to the actual fingermark and fingerprint similarity rather than poor image quality (cf. Section 2).

This experiment is carried out on all fingermarks, which have been digitised and rescaled using different CREs such as no rescaling, a global rescaling factor for each device, or an individual estimate based on image characteristics (RLAPS).



(a) ROC, Scanner (b) ROC, DSLR (c) ROC, Phone

(d) EV class distribution, Scanner (e) EV class distribution, DSLR (f) EV class distribution, Phone

(g) ccID and EV, Scanner (h) ccID and EV, DSLR (i) ccID and EV, ccID, Phone

Figure 3: The first row ((a) to (c)) shows the top left corner of the receiver operating characteristics (ROCs) for all capture devices calculated on the Fusion image quality feature set with global rescaling. The colour varies according to the fraction of $^{\text{EV \& ccID}}/_{\text{ccID}}$ as the classification threshold moves along the ROC; the smaller the fraction, the lighter the colour (only applicable to "Scanner, None" and "Phone, None"; the fraction equals one across the whole range in all other cases). The second row ((d) to (f)) shows the EV distribution according to the mark's distortion class which has been computed by EVA (solid) and the experts (dashed). The latter isn't affected by the decision threshold and hence remains constant. The third row ((g) to (i)) shows the number of ccIDs classified as EV by EVA (light colour), the additional ones by the experts (dark colour) and the ones not classified as EV by either but are ccID (white). The gray line is the threshold corresponding to the EER when the operating point moves along the ROC.

Additionally, we check if the ccID fingermarks are considered to be of EV by either the experts or EVA. In case of the algorithm, the decision threshold corresponding to the equal error rate (EER) has been chosen (cf. Figure 3g to 3i).

The step to determine if a fingermark is of EV is performed first and subject to capture device and CRE. Therefore, the verification scores used for identification are calculated afterwards on the already rescaled image (see Figure 1). Verifinger failed to compare 20 query fingermarks to the database because of their very high image resolution; this was only the case for unrescaled scanner images.

Finally, we applied different decision thresholds (instead of just the one corresponding to the EER) to the evidential value raw scores. The aim is to observe if allowing more false positive errors (and hence collecting more marks in a real world scenario) would lead to a set of marks of being EV according to EVA which is a superset of the experts' decision. The results are shown in Figure 3 and Table 2.

	CRE	Capture device		
		Scanner	DSLR	Phone
ccID	None	4	118	6
	Global	145	122	126
	RLAPS	29	46	36
Experts	Global	133	117	116
$\text{EVA}_{\text{NFIQ2}}$	Global	137	118	123
$\text{EVA}_{\text{Verifinger}}$		134	118	119
$\text{EVA}_{\text{Fusion}}$		137	119	120

Table 2: Number of fingermarks which have been correctly and with certainty identified (ccID) and the amount of those marks which have been classified by experts or EVA to be of sufficient evidential value (EV) w.r.t. capture device (Scanner, DSLR, Phone), CRE Global and quality feature set (NFIQ2, Verifinger, Fusion) if applicable. EVA uses the threshold corresponding to the EER. The EV results for the CREs None and RLAPS are not reported separately due to their much smaller numbers compared to Global (see *ccID*). Please refer to Table 1 for the total number of EV decisions or their distribution amongst the different distortion classes.

3.3 Discussion

The experiment shows a strong correlation between the automatically estimated evidential value and if a certain identification is possible to be performed for a particular fingermark. This is partially due to the setup used because both the matching score computation and EV estimation are based on image features.

Further limitations of the matching system used became evident and confirm the findings in [KDH+15]. Verifinger is very resolution dependent and requires marks or prints to be in a very narrow capture resolution window (around 500 ppi) with as little variation as possible to perform properly. This is the reason that a global rescaling factor and the high-quality camera images without any rescaling work well. It also explains why there are very few ccIDs when images with very high resolution without (CRE None) or with individual (CRE RLAPS) rescaling are used. Nevertheless, the image quality due to the use of different capture devices is not a major drawback. The mobile phone performs more strongly than the DSLR but falls shy of the scanner, under the condition that the capture resolution is adjusted properly. The difference between the quality feature sets is rather small but should be considered in a real world framework.

Further, we note there are prints which can be automatically identified with certainty but haven't sufficient evidential value according to the experts' assessment. This might be again due to experimental setup that heavily favours image processing algorithms or the limited size of the test population and database. Additionally, it is worth pointing out that some of the identified fingermarks are only considered to be of evidential value by the experts or the algorithm, but not both.

Encouragingly, only *one* of the 116 marks being identified with certainty and having EV according to the experts was missed by the algorithm in a mobile phone scenario with global rescaling and the NFIQ2 feature set.

Table 1 also indicates that EVA works rather conservatively and tends to flag a fingermark as being of sufficient evidential value slightly more often than an expert who applies other considerations (such as court eligibility) than just image quality. Nevertheless, an expert's accuracy and repeatability can vary mostly due to the print quality [UHBR11, UHBR12] regarding borderline decisions and experts "tend to free the guilty rather than to convict the innocent" [TTM11].

Our results underscore the importance of capturing all fingermarks of sufficient evidential value in the field. They should have VEO if they don't have VID.

4 Conclusion and future work

The experiment performed indicates a strong correlation between the fact that a fingermark can be automatically identified with certainty and its inferred evidential value. Therefore, it is sensible to run fingermarks with sufficient evidential value against a reference database to potentially obtain an identification.

Also, our findings indicate that an automatic mobile phone setup is suitable to determine if a fingermark at a crime scene is of sufficient evidential value and could be operated by non-experts. In the case that the capture conditions are unknown, it is sensible to use a capture resolution estimator to improve performance.

In the future, we would like to perform more exhaustive testing on additional and considerably larger databases with different matching systems as well. Also, a fingermark determined to be of EV needs to be evaluated as either VEO or VID. Eventually we would like to test performance in the field.

Acknowledgment

We thankfully acknowledge the Victoria Police fingerprint examiners who kindly assessed all fingermarks. The research project is funded by the Victoria Police. Also we would like to thank Dr. Arathi Arakala for her valuable input and suggestions. This work forms part of the PhD thesis of the first author.

References

[AFFOG+07] Fernando Alonso-Fernandez, Julian Fierrez, Javier Ortega-Garcia, Joaquin Gonzalez-Rodriguez, Hartwig Fronthaler, Klaus Kollreider, and Josef Bigun. A comparative study of fingerprint image-quality estimation methods. *Information Forensics and Security, IEEE Transactions on*, 2(4):734–743, 2007.

[Ash99] David R Ashbaugh. *Quantitative-qualitative friction ridge analysis: an introduction to basic and advanced ridgeology*. CRC press Boca Raton, 1999.

[CDJ05] Yi Chen, Sarat C Dass, and Anil K Jain. Fingerprint quality indices for predicting authentication performance. In *Audio-and Video-Based Biometric Person Authentication*, pages 160–170. Springer, 2005.

[FKB06] Hartwig Fronthaler, Klaus Kollreider, and Joseph Bigun. Automatic image quality assessment with application in biometrics. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 30–30. IEEE, 2006.

[FSZ12] Jianjiang Feng, Yuan Shi, and Jie Zhou. Robust and Efficient Algorithms for Separating Latent Overlapped Fingerprints. *Information Forensics and Security, IEEE Transactions on*, 7(5):1498–1510, 2012.

[KDH+15] J. Kotzerke, S. A. Davis, R. Hayes, L. J. Spreeuwers, R. N. J. Veldhuis, and K. J. Horadam. Discriminating fingermarks with evidential value for forensic comparison. In *Biometrics and Forensics (IWBF), 2015 International Workshop on*, pages 1–6. IEEE, March 2015.

[KME+14] Philip J Kellman, Jennifer L Mnookin, Gennady Erlikhman, Patrick Garrigan, Tandra Ghose, Everett Mettler, David Charlton, and Itiel E Dror. Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates through Understanding and Predicting Difficulty. *PloS one*, 9(5):e94617, 2014.

[LCCK08] Sanghoon Lee, Heeseung Choi, Kyoungtaek Choi, and Jaihie Kim. Fingerprint-quality index using gradient components. *Information Forensics and Security, IEEE Transactions on*, 3(4):792–800, 2008.

[Mal09] Davide Maltoni. *Handbook of fingerprint recognition*. Springer, 2009.

[MMC+02a] Dario Maio, Davide Maltoni, Raffaele Cappelli, James L. Wayman, and Anil K. Jain. FVC2000: Fingerprint verification competition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):402–412, 2002.

[MMC+02b] Dario Maio, Davide Maltoni, Raffaele Cappelli, James L Wayman, and Anil K Jain. FVC2002: Second fingerprint verification competition. In *Pattern recognition, 2002. Proceedings. 16th international conference on*, volume 3, pages 811–814. IEEE, 2002.

[MMC+04] Dario Maio, Davide Maltoni, Raffaele Cappelli, Jim L Wayman, and Anil K Jain. FVC2004: Third fingerprint verification competition. In *Biometric Authentication*, pages 1–7. Springer, 2004.

[Neu15] Neurotechnology. VeriFinger SDK. `http://www.neurotechnology.com/verifinger.html`, 2015.

[SHAF11] Nathan J. Short, M.S. Hsiao, A.L. Abbott, and E.A. Fox. Latent fingerprint segmentation using ridge template correlation. In *Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th International Conference on*, pages 1–6, 2011.

[The13] The National Institute of Standards and Technology. NFIQ2 Feature Definitions Document (v0.5). http://biometrics.nist.gov/cs_links/quality/NFIQ_2/NFIQ-2_Quality_Feature_Defin-Ver05.pdf, April 2013.

[TTM11] Jason M Tangen, Matthew B Thompson, and Duncan J McCarthy. Identifying fingerprint expertise. *Psychological science*, 22(8):995–997, 2011.

[UHBR11] Bradford T Ulery, R Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences*, 108(19):7733–7738, 2011.

[UHBR12] Bradford T Ulery, R Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. Repeatability and reproducibility of decisions by latent fingerprint examiners. *PloS one*, 7(3):e32800, 2012.

[Vic13] Victoria Police. Crime Statistics 2012/2013. http://www.police.vic.gov.au/retrievemedia.asp?Media_ID=72176, August 2013.

[YJ13] Soweon Yoon and Anil Jain. Quality assessment of latent fingerprints. http://biometrics.nist.gov/cs_links/quality/NFIQ_2/presentations_4-26/nfiq2_yoon-2013-04-26.pdf, April 2013.