DISCRIMINATING FINGERMARKS WITH EVIDENTIAL VALUE FOR FORENSIC COMPARISON

J. Kotzerke¹², S.A. Davis¹, R. Hayes³, L.J. Spreeuwers², R.N.J. Veldhuis² and K.J. Horadam¹

¹School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia ²Services, Cybersecurity and Safety, University of Twente, Enschede, The Netherlands ³Forensic Services Department, Victoria Police, Melbourne, Australia

ABSTRACT

Law enforcement agencies all around the world are using biometrics and especially fingerprints to solve and fight crime. Often forensic experts are needed to record fingermarks at crime scenes and to ensure that those captured are of forensic value. In times of increased demand for forensic services, this process needs to be automated and streamlined as much as possible to improve efficiency and reduce workload.

Hence, we investigate if the forensic evidential value (suitability for forensic analysis and/or examination) of fingermark images can be determined at an early stage automatically without any expert involvement, especially when using a mobile phone camera. We explore the interplay of different factors such as the capture device and the constraints inferred, image feature sets and classifiers used, and their interplay.

A database of 1,428 pseudo fingermarks has been collected and its ground truth, whether a mark is of forensic value or not, has been determined by 3 experts. The lowest equal error rate achieved, when using a mobile phone to capture the marks, is 13.62%.

These promising results suggest that it might be possible to streamline forensic procedures by the application of an independent automated tool to assist with certain tasks.

Index Terms— Fingermark, Mobile phone capture, Quality estimation, Sufficient forensic evidential value

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 [1]. Forensic experts must travel in many cases to the crime scene and collect the evidence themselves, spending a lot of time travelling to the scene. 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 propose 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 cases' evidence collection the



Fig. 1. Diagram of the investigated problem. A fingermark is captured with an imaging device and rescaled based on the Capture Resolution Estimation (CRE) to 500 ppi, image features are extracted and classified; resulting in the binary decision if this mark is of sufficient evidential value (EV) or not (\overline{EV}) .

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 chemical mark development. 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.

We investigate if the binary evidential value of a fingermark can be derived from its captured image in general, and from its image taken with a mobile phone in particular. Hence, we look into 4 factors and their interplay: different imaging devices and their constraints (i), quality feature sets for fingerprints and the necessity to rescale the image to a fixed resolution (ii), algorithms for Capture Resolution Estimation (CRE) (iii), and different classifiers to derive a binary decision if a certain mark is of sufficient evidential value (EV) or not (\overline{EV}) (iv). The proposed framework is shown in Fig. 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 [2]. These examiners are expected to follow the Analysis, Comparison, Evaluation, and Verification (ACE-V) protocol [3]. 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).

However, fingermarks suffer often from low quality due to being smudged or partial, overlap with other marks [4], or distorted by the surface pattern of the object they are found on [5]. Their forensic value is difficult to grasp for nonexperts. Ulery *et al.* show that accuracy and repeatability varies even for forensic experts and mostly depends on the print quality [6, 7], especially for borderline decisions. Consequently, Kellman *et al.* use image features to predict "expert performance and subjective assessment of difficulty in fingerprint comparisons" [8].

Most quality measures are used to prevent low quality images from being automatically matched because they tend to produce false minutiae and hence false matches [9]. Therefore, they are suited to operational law enforcement agency setups and only optimised and tested for contact scanners [10–13] 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 [14] 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 [13] 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.

Unfortunately, in the proposed scenario, capturing fingermarks with a mobile phone, the capture resolution isn't either fixed or known for two main reasons. Firstly, the phone is hand-held by the operative, most likely at varying distances for each capture. Secondly, the phone model and maker can differ and so does the camera module and its resolution. In order to ensure consistent results and more importantly to use already existing quality features and algorithms, the approach has to be capture resolution independent.

One way to achieve this independence is to measure image characteristics, infer the capture resolution (cf. CRE) and to rescale the image based on this estimate to a standard resolution in order to use existing quality features and algorithms.

1.2. Outline

We investigate if the evidential value of a fingermark can be determined from its image w.r.t. the capturing device (scanner, high-quality camera, phone), CRE algorithm, image quality feature set (NFIQ2, Neurotechnology Verifinger, their fusion) and classifier used (cf. Fig.1).

In the following three sections we introduce our own CRE algorithm RLAPS to rescale the fingermark images to 500 ppi (Section 2), elaborate on the database collection, perform ex-

periments to demonstrate the interplay between CRE algorithm, quality feature set and classifier used and discuss the results obtained (Section 3). We conclude our findings and present the direction of our future research (Section 4).

2. METHODOLOGY

The main idea behind our CRE algorithm is to measure the inter-ridge spacing (IRS) and to rescale the image based on the ratio of the measured value to a known reference. Hence, we make two major assumption regarding the captured image, its quality and fingermarks: first, we presume that an adult fingerprint has been captured and infer an average IRS of irs' = 9 px at c' = 500 ppi; commercial products make similar assumptions as a child's fingerprint and hence its IRS is much smaller [15]. Second, this implies also that there must be visible ridge pattern present.

Our algorithm is very similar to other approaches using the radially averaged power spectrum to either estimate a fingerprint' quality [13] or its inter-ridge spacing [2]. The difference is that it uses only a part of the spectrum, the Radially Limited Averaged Power Spectrum (RLAPS). Furthermore, it takes into account that the IRS varies depending on where and how it is determined. The IRS increases around singular points [2] so it is often measured in areas where none are present [16]. Hence, we choose the spectral peak corresponding to the highest frequency but with an amplitude close to the maximum.

Therefore, RLAPS features increase robustness against fingerprint distortions, which lead to an uneven and noncircular energy distribution or certain fingerprint characteristics as singular points. This results in accurate estimates and little variation (see Section 3.2.1).

The algorithm consists of three main stages: (i) the computation of the power spectrum and its filtered version, (ii) the computation of the radially averaged power spectrum for a limited area according to the highest energy distribution and (iii) finding the "last large peak" and converting the interridge spacing to a CRE.

Now, let's assume that a gray-value image I has the horizontal and vertical resolution of M px and N px, respectively. The pixel at the position (x, y) is denoted by $I_{(x,y)}$ and accesses its assigned (gray) value where $I_{(x,y)} \in \{1, \ldots, 255\}$, $x \in \{1, \ldots, M\}$ and $y \in \{1, \ldots, N\}$.

The image is 2 D-DF transformed into the frequency domain, maintaining the same size, and its quadrants are rearranged that the zero-frequency component is located at the centre, resulting in F. The power spectrum A is computed from the complex modulus of the re-arranged spectrum; $A = |F|^2$. There might be straight lines contained within the spectrum due to spectral distortions. The Hough transform is employed to find and mask those; additionally A is smoothed by convolving it with a square matrix ($w \times w$) consisting only of ones, resulting in A'.



Fig. 2. The radially limited averaged power spectrum (RLAPS, black curve) for the same mark captured by a scanner (a), a high-quality camera (b) and a mobile phone (c) as well as the algorithm (and manual) inter-ridge spacing estimates. The ridge pattern introduces a local increase in power, the last high peak (green circle) indicates the ridge frequency; other individual peaks are highlighted by blue circles and the maximum by a red one. RLAPS is calculated from the spectral part enclosed by the white borders. Lines contained within the spectrum have not been masked from the images for demonstration purposes but during the calculation.

The values of A and A' are accessed via their polar coordinates ρ and θ rather than their Cartesian coordinates x and y. Finally, the maximum energy peak in AF is located w.r.t. to the constraints that $\theta \ge 0$ disregarding the third and fourth quadrants due to duplicity and $r_{min} \le \rho \le r_{max}$ limiting the search range and ignoring the main energy located at the centre. The corresponding angle is stored in α .

$$\alpha = \underset{A'_{(\theta,\rho)}}{\operatorname{argmax}}, \ \theta > 0, \ r_{max} \ge \rho \ge r_{min}$$
$$r_{max} = \lfloor \frac{1}{2} \min(N, M) \rfloor, \ r_{min} = \lfloor \gamma r_{max} \rfloor$$
(1)

The factor γ and the angle β need to be chosen roughly depending on the expected frequency range and distortion strength, respectively (cf. Section 3.2.1).

Next, the RLAPS is obtained from A by averaging w.r.t the energy values over the angle $\theta = [\alpha - \beta, \alpha + \beta]$ for $\rho = \{r_{min}, \dots, r_{max}\}$ and results in the vector J containing l elements where $l = r_{max} - r_{min} + 1$.

We define that all high peaks have at least a certain fraction of the global maximum:

$$\lambda \operatorname{argmax}_{J_{(i)}}, \text{ with } i = \{1 \dots l\}.$$
⁽²⁾

The candidate with the highest frequency and hence the greatest index (say r') is chosen and adjusted to $r = r_{min} - 1 + r'$. Therefore the dominant IRS contained in the image is $irs = \frac{2r_{max}}{r}$.

Finally, the capture resolution c is obtained by rescaling the fraction using the assumed values for adult IRS irs' at a given resolution c' as previously discussed. This leads to the equation

$$c = \frac{c'}{irs'}irs = \frac{500}{9}\frac{2r_{max}}{r}.$$
 (3)

There are two kind of uncertainties involved. Firstly, the assumed average IRS of 9 px doesn't account for individual differences due to e.g. gender, ethnicity or age (cf. 3) and secondly, measurement inaccuracies.

3. EXPERIMENTS AND DISCUSSION

In this section we evaluate how different image quality feature sets perform for estimating if a certain fingermark is of sufficient evidential value and if mobile phone image captures of those marks are a suitable source for evidential value estimation compared to established capturing methods such as a flatbed scanner or a high-quality camera and their interplay with our CRE algorithm RLAPS (cf. Fig.1).

First, we introduce our new database (Section 3.1), then we determine the fingermark value prediction performance for NFIQ2 image features, the features of a commercial fingerprint SDK (Neurotechnology Verifinger) and their fusion w.r.t the use of different CREs and image capture devices (Section 3.2). Finally we discuss our findings (Section 3.3).

3.1. Database

Two males (subjects 2 and 4) and two females (subjects 1 and 3) volunteered to create a large database of pseudo fingermarks, including normal and deliberately distorted marks (1, 428 in total). Six categories of deliberate distortion, each containing 168 marks, were defined: (i) finger placed "lightly" on the page, (ii) mark smeared, (iii) finger twisted lightly, (iv) finger twisted strongly, (v) finger placed "heavily" on the page, and (vi) part of the finger placed "heavily" on the page (cf. Fig. 2).



Fig. 3. The capture resolution estimate is plotted for each individual subject and capture device to highlight the intra-class variation; the different distortion types are colour encoded and have been spread horizontally for visualisation purposes. We have manually estimated the capture resolution for the different devices; scanner: 1200 ppi (a), high-quality camera: 460 ppi (b), mobile phone: 890 ppi (c). As the algorithm uses a generalised scaling factor, the individual subject resolutions differ.

After leaving a fingermark on a paper sheet in predefined marked areas according to the specified distortion, the sheets were brushed using magnetic black powder to make the latent mark visible and conserve it. Finally, each individual sheet was laminated to ensure its permanent integrity. A Victoria Police fingerprint expert supervised the whole process and performed a closer visual inspection of the marks before and after the lamination process to reassure that no major distortions had been added.

Subsequently, all sheets have been individually digitised using 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). The scanner and camera captured the whole sheet at once, whereas the mobile phone photographed each mark individually at an approximate distance of 10 cm. The phone was held mostly parallel to the sheet to avoid light reflections induced by the laminate. Additionally, a light source was present at all times perpendicular to both the sheet and the phone; and a remote control was used to release the shutter to minimise any movement due to touching the phone during capture.

Finally, three Victoria Police experts individually assessed the laminated marks and decided for each if it is of sufficient evidential value (to be more specific: VID at the analysis stage of ACE-V). The overall agreement between the assessors is high; they reached the same conclusion in 97.8% (assessors 1&2), 95.7% (assessors 1&3) and 99.4% (assessors 2&3) of the marks.

The overall ground truth is obtained via a majority vote; a certain mark is considered to be of sufficient evidential value if and only if at least two assessors agreed that it is indeed of evidential value. According to this methodology, 34.5% of the 1,428 marks are of sufficient evidential value.

3.2. Fingermark evidential value prediction

The evidential value prediction is subject to two major variables: (i) the image quality feature set and subsequently the CRE method used as it influences the feature extraction and (ii) the image capture device. We employ the NFIQ2 features specified in the preliminary definition guide, Neurotechnology Verifinger 6.7 [17] and its quality value and the number of found minutiae (Verifinger) and all features together (Fusion). We have implemented the NFIQ2 features according to their preliminary specification [13].

3.2.1. Capture resolution estimation

All feature sets are computed for all images of all capture devices with three different CREs performed: none at all (None) and either global (Global) or individual (RLAPS) estimates.

For this purpose, we have manually approximated a Global capture resolution of 1200 ppi (scanner), 460 ppi (high-quality camera) and 890 ppi (mobile phone). The estimates are based on the average of samples randomly chosen for which we compared the pixel size of the mark to the one of the image scanned as it has been captured at the known capture resolution of 1200 ppi.

RLAPS has been applied to all images using the following parameters experimentally obtained: $\beta = 15^{\circ}$, w = 7, $\gamma = 0.06$ (scanner), $\gamma = 0.17$ (DSLR), $\gamma = 0.8$ (phone) and $\lambda = 0.8$. Its results are visualised in Fig. 3 w.r.t. the capture device, individual subject and type of distortion.

Their CRE distribution shows a significant difference related to the different IRS between the sexes [16]. Therefore, the consistency of the CRE or IRS ratio (cf. Equation 3) between males and females across different devices indicates the algorithms stability on a wide range of different images: 1.21 (Scanner), 1.20 (DSLR), 1.22 (Phone).



Fig. 4. The top left corner of the receiver operating characteristics (ROCs) for all capture devices and different CREs calculated on the fused image quality feature set of NFIQ2 and Verifinger.

3.2.2. Database partition

The database creation and the ground truth acquisition have been discussed in Section 3.1. In order to run unbiased experiments, the database D needs to be partitioned into a training set T_k , a validation set V_k and a test set S_k . We use a K-fold with K = 5, $k \in \{1, \ldots, K\}$. We choose all sets randomly but within two main constraints: firstly, the union of all test sets must be the database and the pairwise intersection of the test sets must be the empty set. Secondly, the number of samples per set is roughly 60% (training), 20% (validation) and 20% (test). All samples are used and there is no overlap between the sets during each fold.

3.2.3. Experiment

This experiment is performed for three image quality feature sets (NFIQ2, Verifinger, Fusion) subject to different CREs (None, Global, RLAPS) and three capture devices (Scanner, DSLR, Phone). All images are rescaled by the factor $\frac{500}{c}$ (see Section 1.1 and cf. Equation 3) using the nearestneighbor interpolation, if applicable. The features obtained are classified by a Support Vector Machine (SVM), the Discriminant Analysis (DA) and the k-Nearest Neighbors algorithm (kNN) trained according to the database partition and different parameter sets (cf. Fig.1). For each k the classifier of choice $C \in \{$ SVM, kNN, DA $\}$ is trained on T_k with different sets of parameters $P_i \in \mathbf{P}$ and then applied to the validation set V_k . Then the results are scored using the true positive rate for a fixed false match rate of 0.01 (TPR@FMR100) and averaged over k and the capture method (scanner, high-quality camera, mobile phone) in order to find the parameter set P_i that leads on average to the best performance.

Now each classifier, trained on T_k and using the parameter set P_i , is applied to each test set S_k individually, resulting in S'_k . The final score is calculated on the concatenation of all test set results $\{S'_1, \ldots, S'_K\}$. The constraints ensure that there is only one unique result for every mark captured. See

		Capture device		
Feature set	CRE	Scanner	DSLR	Phone
NFIQ2	None	15.65%	16.24%	16.56%
	Global	11.75%	15.17%	14.23%
	RLAPS	11.18%	16.13%	15.60%
Verifinger	None	32.10%	17.07%	30.75%
	Global	13.25%	15.85%	14.43%
	RLAPS	13.82%	16.56%	15.87%
Fusion	None	13.35%	13.01%	16.46%
	Global	8.54%	12.20%	13.62%
	RLAPS	9.29%	13.03%	14.21%

Table 1. EER w.r.t. capture device (Scanner, DSLR, Phone), quality feature set (NFIQ2, Verifinger, their fusion) and CRE algorithm (None, Global, RLAPS).

Fig. 4 and Table 1 for the corresponding receiver operating characteristics (ROCs) and for their equal error rate (EER) values, respectively.

There are two major remarks regarding the experimental procedure to keep in mind. Firstly, we are basically evaluating K different classifiers instead of one as it is trained differently during each fold, despite using the same set of parameters. Secondly, this methodology uses a relatively small training set and large test set if compared to other approaches (e.g. leave-one-out cross-validation). Therefore we are establishing a lower performance boundary.

3.3. Discussion

The experiments indicate that it is possible to determine if a query mark is of sufficient evidential value based on its image quality features as long as the capture resolution is normalised to 500 ppi. In case this is unknown, it can be estimated by the introduced algorithm without any significant performance setback. Also, the algorithm's accuracy isn't overly important but its consistency is. NFIQ2 performs surprisingly well for unadjusted images; whereas this is only true for DSLR captures when using Verifinger because their capture resolution is close to 500 ppi.

Furthermore, the mobile phone image quality is not a major limitation as long as the capture process is done properly. If so, the performance is on a par with a the high-quality camera but shy of the flatbed scanner.

The NFIQ2 feature set performs better than the Verifinger features but their fusion shows the most promising and consistent results.

The reader has to keep in mind that firstly the fingermarks vary in quality without any strong background patterns or surface distortions and secondly that the image quality feature algorithms used have all been developed for images acquired by contact scanners. Furthermore, we are aware of the database's limited size and that inter-age and inter-racial cross validation are needed to draw any further conclusions.

4. CONCLUSION AND FUTURE WORK

The experiments performed demonstrate that it is possible to use images, captured with a mobile phone, to derive if a fingermark is of sufficient evidential value. Already today, the mobile phone's image quality is not a limiting constraint. As technology development progresses, mobile phone cameras will improve further and will be able to deliver consistently high image quality even under more difficult circumstances than we have evaluated in this case study.

The results are encouraging, particularly as they establish a lower performance bound. One major drawback is that the image capture resolution needs to be estimated in order to be scaled to 500 ppi for optimal performance as most fingerprint image quality features are optimised for this particular resolution. However, the experiment emphasises that a rough approximation is reasonable if no setup with a known resolution and fixed distance is available.

We are looking forward to the official release of NFIQ2 to ensure the correctness of our work and benefit from its fast reference implementation.

Acknowledgment

We thankfully acknowledge the Victoria Police fingerprint examiners who kindly assessed all fingermarks. The research project is funded by the Victoria Police.

5. REFERENCES

- [1] Victoria Police, "Crime statistics 2012/2013," http: //www.police.vic.gov.au/retrievemedia. asp?Media_ID=72176, August 2013.
- [2] D. Maltoni, *Handbook of fingerprint recognition*, Springer, 2009.

- [3] D. R. Ashbaugh, Quantitative-qualitative friction ridge analysis: an introduction to basic and advanced ridgeology, CRC press Boca Raton, 1999.
- [4] J. Feng, Y. Shi, and J. Zhou, "Robust and efficient algorithms for separating latent overlapped fingerprints," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 5, pp. 1498–1510, 2012.
- [5] N. J. Short, M. Hsiao, A. Abbott, and E. Fox, "Latent fingerprint segmentation using ridge template correlation," in *Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, 4th International Conference on, 2011, pp. 1–6.
- [6] B. T. Ulery, R. A. Hicklin, J. Buscaglia, and M. A. Roberts, "Accuracy and reliability of forensic latent fingerprint decisions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 19, pp. 7733–7738, 2011.
- [7] B. T. Ulery, R. A. Hicklin, J. Buscaglia, and M. A. Roberts, "Repeatability and reproducibility of decisions by latent fingerprint examiners," *PloS one*, vol. 7, no. 3, pp. e32800, 2012.
- [8] P. J. Kellman, J. L. Mnookin, G. Erlikhman, P. Garrigan, T. Ghose, E. Mettler, D. Charlton, and I. E. Dror, "Forensic comparison and matching of fingerprints: Using quantitative image measures for estimating error rates through understanding and predicting difficulty," *PloS one*, vol. 9, no. 5, pp. e94617, 2014.
- [9] F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, J. Gonzalez-Rodriguez, H. Fronthaler, K. Kollreider, and J. Bigun, "A comparative study of fingerprint image-quality estimation methods," *Information Forensics and Security, IEEE Transactions* on, vol. 2, no. 4, pp. 734–743, 2007.
- [10] Y. Chen, S. C. Dass, and A. K. Jain, "Fingerprint quality indices for predicting authentication performance," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2005, pp. 160–170.
- [11] H. Fronthaler, K. Kollreider, and J. Bigun, "Automatic image quality assessment with application in biometrics," in *Computer Vision and Pattern Recognition Workshop*, 2006. *CVPRW'06. Conference on.* IEEE, 2006, pp. 30–30.
- [12] S. Lee, H. Choi, K. Choi, and J. Kim, "Fingerprint-quality index using gradient components," *Information Forensics and Security, IEEE Transactions on*, vol. 3, no. 4, pp. 792–800, 2008.
- [13] 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.
- [14] S. Yoon and A. 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.
- [15] A. K. Jain, K. Cao, and S. S. Arora, "Recognizing infants and toddlers using fingerprints: Increasing the vaccination coverage," in *Biometrics (IJCB)*, 2014 International Joint Conference on, October 2014.
- [16] M. A. Acree, "Is there a gender difference in fingerprint ridge density?," *Forensic science international*, vol. 102, no. 1, pp. 35–44, 1999.
- [17] Neurotechnology, "VeriFinger SDK 6.7," http://www. neurotechnology.com/verifinger.html, 2014.