Kotzerke *et al. EURASIP Journal on Information Security* (2016) 2016:24
DOI 10.1186/s13635-016-0050-3

EURASIP Journal on Information Security

RESEARCH Open Access

# Identification performance of evidential value estimation for ridge-based biometrics

Johannes Kotzerke[1,3*], Hao Hao[1], Stephen A. Davis[1], Robert Hayes[2], L. J. Spreeuwers[3], R. N. J. Veldhuis[3] and K. J. Horadam[1]

**Abstract**

Law enforcement agencies around the world use ridge-based biometrics, especially fingerprints, to fight crime. Fingermarks that are left at a crime scene and identified as potentially having evidential value (EV) in a court of law are recorded for further forensic analysis. Here, we test our evidential value algorithm (EVA) which uses image features trained on forensic expert decisions for 1428 fingermarks to produce an EV score for an image. First, we study the relationship between whether a fingermark is assessed as having EV, either by a human expert or by EVA, and its correct and confident identification by an automatic identification system. In particular, how often does an automatic system achieve identification when the mark is assessed as not having evidential value? We show that when the marks are captured by a mobile phone, correct and confident automatic matching occurs for 257 of the 1428. Of these, 236 were marked as having sufficient EV by experts and 242 by EVA thresholded on equal error rate. Second, we test four relatively challenging ridge-based biometric databases and show that EVA can be successfully applied to give an EV score to all images. Using EV score as an image quality value, we show that in all databases, thresholding on EV improves performance in closed set identification. Our results suggest an EVA application that filters fingermarks meeting a minimum EV score could aid forensic experts at the point of collection, or by flagging difficult latents objectively, or by pre-filtering specimens before submission to an AFIS.

**Keywords:** Ballprint, Evidential value, Fingermarks, Identification, Image quality

## 1 Introduction

Fingermarks or other marks found at crime scenes are often highly distorted because they are accidentally left behind, rather than purposely recorded within a specified environment and under controlled conditions. A mark's quality may range from excellent to poor and determines its further use. Experts evaluate the forensic quality of a mark (or its digital representation): that is, the quantity of information available in the mark and the relevance of the mark at the crime scene. There is no benefit attached to collecting all marks found at a crime scene and submitting them for further analysis regardless of their forensic quality. This leads to additional workload for forensic specialists who use the traditional ACE-V system (rather than the likelihood ratio approach [1]) to analyse the low-quality marks and disregard them afterwards. Hence, it is desirable to limit the collection to those marks that at least meet some minimal condition of sufficient evidential value to be of use in an ongoing police investigation.

Our goal here is to assess whether our evidential value algorithm (EVA; which estimates a mark's evidential value (EV) automatically), introduced in Kotzerke et al. [2], has the capacity to reduce workload and streamline the capture process by entrusting non-expert police with mark selection and collection. We test the performance of EVA in two ways. First, in the worst case, EVA might score a mark to be of no evidential value but it nonetheless will identify its donor in an automatic identification. We test if there are any marks that can be automatically and with confidence identified (against a reference database) but are not of EV according to either EVA or expert assessment. Secondly, we investigate if EVA is sufficiently general that its EV score for an image can be used to infer an image quality value for specimens from several relatively

*Correspondence: johannes.kotzerke@rmit.edu.au
[1]School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia
[3]Services, Cybersecurity and Safety, University of Twente, Enschede, The Netherlands
Full list of author information is available at the end of the article

challenging ridge-based biometric databases and, if so, how this image quality affects a closed set identification.

This article extends and updates the conference paper [3]. The experimental setup in [3] for the worst case test has been updated and the experiments repeated. Our updated results appear in Section 3.1. All the work in Section 3.2 using EVA to infer an image quality value for other databases is new, as are the corresponding identification experiments.

### 1.1 Background

Law enforcement agencies rely heavily on fingermarks to exclude or to identify persons of interest using automatic systems (AFIS) and human fingerprint experts [4]. Fingerprint examiners follow the analysis, comparison, evaluation, and verification (ACE-V) protocol [5]. During analysis, they decide if the mark is of value for individualisation (VID), value for exclusion only (VEO) or no value (NV). Those with VID or VEO have evidential value (EV); those with NV do not.

Fingermarks often suffer from low quality (a small amount of relevant information present or a low degree of distinctiveness) due to being smudged or partial, overlapped with other marks [6] or distorted by the surface pattern of the object on which they are found [7]. For fingerprints captured under controlled conditions, fingerprint quality metrics abound (for a recent review, see Yao et al. [8]). However, the forensic value of a fingermark is quite different and is difficult to grasp for non-experts. Ulery et al. show that accuracy and repeatability vary even amongst forensic experts and mostly depend on the print quality [9, 10], especially for borderline decisions. Consequently, Kellman et al. use image features to predict "expert performance and subjective assessment of difficulty in fingerprint comparisons" [11].

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 [12]. Therefore, they are suited to operational law enforcement agency setups and are optimised and tested for contact scanners [13–16] but not fingermarks. This has resulted in algorithms tuned to a capture resolution of 500 ppi. If the input image deviates from the assumed resolution, the algorithm usually falls short.

On the other hand, fingermarks require robust methods to estimate their quality because all the above factors will vary and influence the quality and its estimate. Yoon et al. demonstrated in [17] that the NIST quality estimator reference implementation NFIQ1 is outdated because an AFIS was able to return a print's mate although it was classified as of lowest quality. They proposed a feature-based quality measure for latents, LFIQ, which they show is a good predictor of AFIS decisions. The NIFQ1 replacement NFIQ2 [16] has recently been officially released. However, it is still primarily developed for fingerprints captured at a known resolution.

If the capture resolution is unknown, an estimate based on image features can improve the performance significantly. For instance, the RLAPS algorithm of Kotzerke et al. [2] estimates the inter-ridge spacing of a fingerprint or fingermark image. The power spectrum is computed and its radial average is determined only around its maximum peak within a certain frequency range. Assuming an average inter-ridge spacing of 9 px for an adult leads to a capture resolution estimate.

Kotzerke et al. [2] include capture resolution estimation (CRE) as a component of an algorithm which computes an evidential value score of an image based on image features and trained on expert decisions. We refer to this specific estimation algorithm as EVA. In [2], it is established that mobile phone images are suitable for EVA to estimate if a fingermark has EV and can achieve results close to an expert assessment, based on the image features.

### 1.2 Outline

In the following, Section 2, we describe EVA and the methodology used in the two experiments and detail the databases tested (three of adult fingerprints and one of newborn ballprints). In Section 3, we describe our experiments and report their results. First, we perform verification experiments using EVA to demonstrate the interplay between a mark's correct and confident identification (ccID) and if it is of EV according to either experts or EVA (Section 3.1). This is followed by application of EVA to the four databases to estimate their specimens' quality and the role this image quality estimate plays in a mate retrieval scenario (Section 3.2). In Section 4, we discuss our results, summarise their implications and point out directions for future research.

## 2 Methods

We begin by describing EVA, the ridge-based biometric evidential value estimation algorithm that we use in this paper. We introduce the idea of correct and confident identification (ccID) that is applied in the first experiment. Then, we present our argument for using the evidential value score output by EVA as an estimate of image quality, as we do in our second experiment. Lastly, we detail the databases we use to test EVA.

### 2.1 Evidential value algorithm

The ridge-based biometric evidential value estimation algorithm EVA is introduced in [2] (see Fig. 1). An image is scaled if necessary using some CRE algorithm to 500 ppi, and a mask is built using optimal thresholding to remove background. The CRE options available are none (image assumed to be 500 ppi), RLAPS and Global (a scaling factor estimated for the capture device). Then, NFIQ2

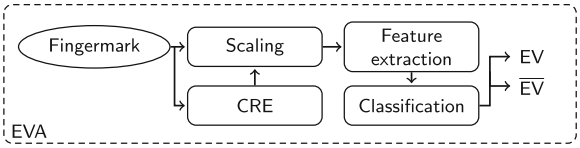


**Fig. 1** [2] Diagram of the evidential value algorithm (EVA). A fingermark is captured with an imaging device and scaled based on the capture resolution estimation (CRE) to 500 ppi; image features are extracted and classified; resulting after thresholding in the binary decision if this mark is of sufficient evidential value (EV) or not ($\overline{EV}$)

and Verifinger features are extracted from the masked image. NFIQ2 features are as specified in the preliminary definition guide [16]: that is, Frequency Domain Analysis, Gabor Segment, Gabor Shen, Gabor, Local Clarity Score, Mu Mu Block, Mu Mu Sigma Block, Mu Sigma Block, Mu, Orientation Flow, Orientation Certainty Level, Radial Power Spectrum, Ridge Valley Uniformity, Sigma Mu Block and Sigma. Verifinger features, extracted using Neurotechnology Verifinger 6.7 [18], are its quality value and the number of minutiae. Their Fusion (concatenation of the NFIQ2 and Verifinger feature vectors) is also used. Classification algorithms available are support vector machine (SVM), discriminant analysis (DA) and $k$-nearest neighbour ($k$-NN). After classification, a raw evidential value estimate $q \in [0, 1]$, the EV score, is output by EVA. If a binary decision is required, $q$ is compared to a threshold $t$ to conclude if there is sufficient evidential value (EV) or not ($\overline{\text{EV}}$).

In [2], the Victoria Police fingermark database $\boldsymbol{T}_{\text{VP}}$ and its ground truth (see Section 2.4.1) is used to train and optimise EVA and estimate evidential value using these feature sets and CREs for three capture devices (scanner, DSLR, phone). For instance, the best equal error rates (EERs) for evidential value from EVA versus ground truth are obtained for the Fusion feature set and Global CRE [2, Table 1], using DA for scanner and phone images and SVM for DSLR.

We use these optimised parameter settings and EV scores in our experiments.

### 2.2 Correct and confident identification

Assuming that a fingermark is compared against a reference database containing $N$ distinct fingerprints, a verification score $S_j$ is returned for every comparison. We define a decision as correct and confident (ccID) if and only if the mark and the print with the highest score $S_i$ are from the same subject (correct) and if the highest score is larger by factor $d > 1$ than any other score (confident):

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

This would lead to a correct and confident 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 confident but is in fact due to low verification scores derived from poor-quality images.

### 2.3 Quality implications of evidential value

In real situations, there are often marks or prints with questionable quality (and hence questionable distinctiveness). Those marks are prone to produce false matches. The matching accuracy and repeatability varies even for forensic experts and mostly depends on the mark's quality [9, 10].

The question is how the identification performance changes if these low-quality images are removed from the set. For our second experiment, we use the forensic experts' decisions for fingermarks to train a classifier to classify the quality of images showing ridge-based biometric features according to their image features.

This approach is based on the assumption that evidential value and image quality are correlated. This hypothesis is reasonable because in [2] it is shown that for pseudo fingermarks the evidential value can be derived from a set of image features. Therefore, the contrary argument should hold true as well and the EV score $q$ should serve as a basic image quality estimate.

We test four image sets of ridge-based biometrics. Each image set consists of a reference set $\boldsymbol{R}$ and a test set $\boldsymbol{T}$, which do not intersect. All elements of both sets are pairwise compared against each other using a commercial matcher and a matching score is computed for each comparison.

We simulate an identification scenario on a closed set and investigate the rank of the correct mate ($\in \boldsymbol{R}$) of the query print ($\in \boldsymbol{T}$). We measure the percentage of query specimens having their mate found within rank $k$ and how this percentage changes as we remove images with low quality $q$.

### 2.4 Databases

We employ six different databases. They are the Victoria Police pseudo fingermark database with its ground truth and its reference database (Section 2.4.1), the IIIT-D latent fingerprint database and its reference database (Section 2.4.2), an imposter database based on various FVC databases (Section 2.4.3) and ballprints of newborns from the Happy Feet database (Section 2.4.4).

#### *2.4.1 Victoria Police pseudo fingermark database*

In [2], Kotzerke et al. introduced a pseudo fingermark database $\boldsymbol{T}_{\text{VP}}$ ; it consists of 1428 normal and deliberately distorted fingermarks from two males and two females. In order to create the distorted marks, they defined six different distortion categories. There are 168 marks per distortion category; the other 420 "normal" marks are not deliberately distorted. Details are in columns 1 and 2 of Table 1.

**Table 1** A breakdown of the 1428 marks into the categories of distortion (including no deliberate distortion), and the percentage of marks found to be EV by each of three experts, with decision on ground truth made by majority vote [3]

| 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 | 1428 | 34.1 | 34.5 | 34.7 | 34.5 | 38.66 |

The EV distribution for EVA has been calculated for the mobile phone images scaled using CRE Global and the Fusion feature set at the decision threshold corresponding to the EER

All fingermarks were left on a sheet of paper, brushed with magnetic black powder then laminated, under supervision by a Victoria Police fingerprint expert. All sheets were digitised with three different capture devices: a flatbed scanner (scanner); a high-quality camera (DSLR); and a mobile phone (phone). The 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 DSLR was attached to an operational stand setup usually used for police work, which led to a lower capture resolution than 500 ppi. The estimated Global capture resolutions are 1200 ppi (scanner), 460 ppi (DSLR) and 890 ppi (phone).

All laminated marks have been assessed by three Victoria Police experts who decided for each mark if it is of EV by performing at least a partial markup process. The ground truth is the majority vote of their assessments. Of the 1428 marks, 492 have ground truth EV. Details appear in columns 3–6 of Table 1. Further details can be found in [2].

For this work, we created a reference database $\boldsymbol{R}_{\mathrm{VP}}$ of 40 genuine prints against which to match the pseudo fingermarks. We collected all ten fingerprints of the same four subjects with a Digital Persona U.are.U 4000 fingerprint scanner, without any deliberate distortion, to imitate a reference scenario. We captured the central finger tip area for consistency with the reference database in Section 2.4.2. A reference image appears next to its pseudo fingermarks in Fig. 2.

#### *2.4.2 IIIT-D latent fingerprint database*

The IIIT-D latent fingerprint database was introduced by Sankaran et al. [19]. It consists of 1046 images of which 1025 contain fingermarks. These marks have been captured from each finger of 15 subjects under semi-controlled conditions. The latent marks were dusted using black powder and captured with a high-quality camera (Canon EOS 500D) at a resolution of $4752 \times 3168$ px. The authors captured the marks over the course of multiple sessions on two different backgrounds (card and tile); the fingers producing the marks showed different levels of "dryness, wetness and moisture". We cropped the images manually in order to contain only the fingermarks. After cropping, we rescaled the images to 500 ppi to ensure Verifinger would recognise them before we separated them into different test databases according to the background used: $\boldsymbol{T}^{C}_{\mathrm{IIIT-D}}$ (card, 383 images) and $\boldsymbol{T}^{T}_{\mathrm{IIIT-D}}$ (tile, 642 images).

The same authors supply IIIT-D Latent Mated 500 ppi Fingerprint Database [20]. This reference database $\boldsymbol{R}_{\mathrm{IIIT-D}}$ contains one capture for each finger of the 15 subjects from the latent fingerprint database (150 images in total). They were captured using a Crossmatch L Scan fingerprint scanner. Again, we cropped the images manually in order to contain only the fingerprints and scaled them to 500 ppi. Example specimens are shown in Fig. 2.

#### *2.4.3 FVC imposter database*

In order to extend the fingerprint reference databases for our experiments, we set up the imposter database $\boldsymbol{R}_{\mathrm{FVC}}$. The images were specifically chosen to exhibit alike characteristics (no deliberate distortion) and were all captured with optical fingerprint scanners. Specifically, we selected all third prints of FVC2000 DB3 [21] and FVC2004 DB2 [22] and all sixth prints of FVC2002 DB1 [23]. This leads to a database consisting of 330 imposter prints. We verified via the cross verification scores that no duplicate of any imposter is included.

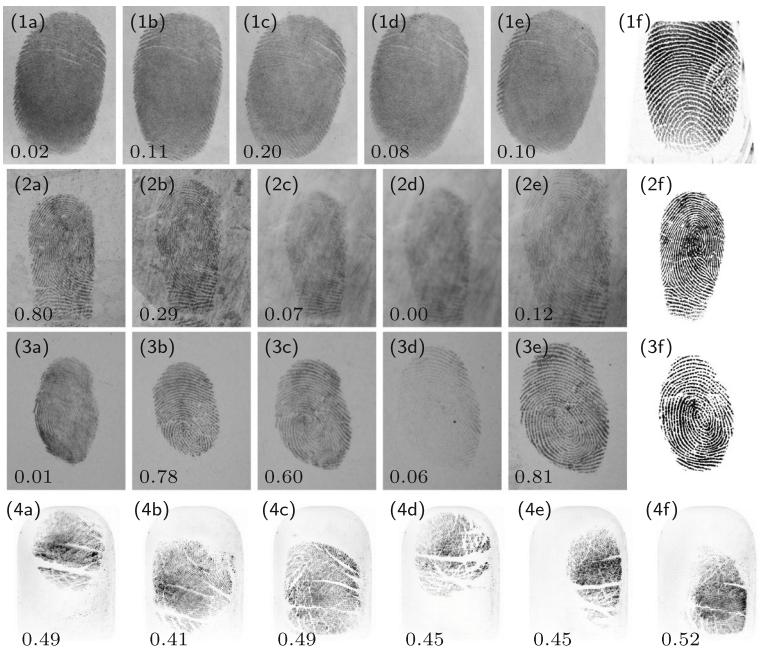


**Fig. 2** Example images of the test databases and their image quality value estimated by EVA. Example images for one subject from each database: Victoria Police pseudo fingermark database (1a–1f), IIIT-D lifted from a card (2a–2f), IIIT-D lifted from a tile (3a–3f) and Happy Feet (4a–4f). Columns a–e are specimens taken from the test set and column f is the corresponding reference image. The EV scores $q$ are stated in the bottom left corner of each individual image. This does not apply to the reference images except for the ballprints where we selected the best of the specimens as reference

#### *2.4.4 Happy Feet*

The Happy Feet database [24, 25] is a 2-year longitudinal footprint and ballprint database from newborn through infant. It contains captures of both feet and balls (right and left) for each of the 54 participants in the study "Happy Feet", which was funded by the Bill and Melinda Gates Foundation to investigate the potential of the footprint and ballprint as infant biometrics. We limit our experiment to the ballprints of the newborns (who ranged from the age of 11 to 683 h with an average age of 67 h) because this is the most challenging set.

Ballprints were captured with an adult single fingerprint scanner, the NEC PU900-10, typically six ballprint impressions per individual foot. The scanner's native sensor resolution was 1000 ppi, but its output was downsampled in hardware to 500 ppi. There are 589 images in total.

We split the database into reference set $\boldsymbol{R}_{\mathrm{BP}}$ and test set $\boldsymbol{T}_{\mathrm{BP}}$. The reference set contains the ballprint image with the highest quality per foot (101 images in total), the remaining 488 images are in $\boldsymbol{T}_{\mathrm{BP}}$. The quality was estimated by EVA (see Section 2.3). There are fewer than 108 feet because there are no prints available for some of the newborns due to their uncooperative nature and playful attitude. Examples of ballprints taken from a newborn are shown in Fig. 2.

## 3 Experiments and results

Because we are interested in the possibility of mobile phone capture of fingermarks in the field, in the last column of Table 1 is listed the distribution of phone captures in $\boldsymbol{T}_{\mathrm{VP}}$ which have sufficient EV using EVA for the parameters giving best EER for phone (Fusion, Global, DA, $t =$ EER). We observe that they are highly correlated with the ground truth (correlation coefficient $r = 0.995$) with EVA performing a little more conservatively in general, so that a higher proportion of marks would be passed by EVA as of EV for further analysis than the experts will pass. Three types of distortion (smeared, light twist and strong twist) have very low numbers of marks with sufficient evidential value assessed by either experts or EVA.

In Section 3.1, we make a more careful analysis of $\boldsymbol{T}_{\mathrm{VP}}$, to determine if any marks which can be verified by an automatic system are not passed either by EVA or experts: the worst case scenario.

In Section 3.2, we use $\boldsymbol{T}_{\mathrm{VP}}$ and its ground truth to estimate the image quality in a subset $\boldsymbol{T}^{*}_{\mathrm{VP}}$ of higher evidential value and in the three other ridge-based biometric test sets. We then investigate how the exclusion of specimens estimated to have lowest quality influences the retrieval of the correct mate in closed set identification.

### 3.1 Comparing evidential value with automatic verification score

This experiment aims to investigate the relationship between a fingermark which can be automatically identified with high confidence and the evidential value assigned to it by experts or EVA. The proposed experiment is shown in Fig. 3.

First, we calculate the EV score $q$ for each fingermark in $\boldsymbol{T}_{\mathrm{VP}}$ using the trained and optimised EVA, for the three feature sets (NFIQ2, Verifinger, Fusion), three CREs (None, Global, RLAPS) and three capture devices (scanner, DSLR, phone), as described above. For the optimal Fusion parameter choice, the receiver operating characteristics (ROCs) are given in Fig. 4a–c.

It is clear from Fig. 4a–c that capture resolution cannot be ignored and must be estimated and allowed for and that in general the Global estimate performs at least as well as the RLAPS estimate (it is also faster). With Global scaling, the distribution of EV across the distortion classes according to experts, and to EVA as it varies according to decision threshold, is shown in Fig. 4d–f. This shows that for all capture devices, the marks labelled as "smeared", "twisted lightly" and "strong twist" have markedly lower EV than the other distortion classes across all decision thresholds for EVA, in agreement with the ground truth.

Next, each fingermark image, scaled according to CRE, is submitted to a verification process performed by Neurotechnology Verifinger 7.0. For every fingermark, a verification score for each print in the reference database $\boldsymbol{R}_{\mathrm{VP}} \cup \boldsymbol{R}_{\mathrm{FVC}}$ (containing 370 prints) is computed and the ccID fingermarks are identified. We use $d = 1.5$ in Eq. 1. We chose $d$ experimentally, so that it delivers a compromise between rejecting questionable matches while retaining clear ones. Verifinger failed to compare 20 query fingermarks to the database because of their very high image resolution; this was only the case for (Scanner, None) images. These 20 images were regarded as incorrectly classified.

Then, we check if the ccID fingermarks are considered to be of EV by either the experts or EVA. In case of EVA, initially, the binary decision threshold $t = \mathrm{EER}$ has been chosen. Table 2 presents our results. For the mobile phone with Global scaling, correct and confident automatic identification occurs for 257 marks. Of these, 236 had been marked as having sufficient EV by experts and 242 by EVA using the Fusion feature set. A detailed check shows 248 with EV according to either EVA or the experts, with only 6 of the 257 marks having EV according to the experts but not according to EVA.

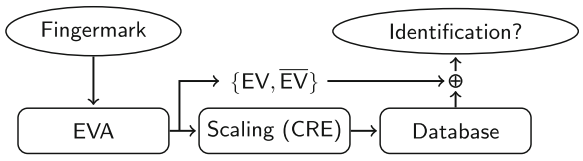


**Fig. 3** Diagram of the ccID experiment. A fingermark is captured; its binary evidential value {EV, $\overline{\mathrm{EV}}$} is estimated by EVA and scaled in the same way as in EVA. The number of correct and confident identifications (ccID) of the mark matched to a reference database is measured w.r.t. the image capture device and EVA parameters [3]

Finally, we test the effect of varying the decision threshold $t$ for the EVA scores $q$. 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 classified as having EV according to EVA which is a superset of the experts' EV set. Results appear in Fig. 4g–i. We see that while this holds for scanner and DSLR images, it does not hold for phone images.

### 3.2 Thresholding on quality score for identification

For this experiment, all images submitted to EVA are scaled to 500 ppi using Global scaling if necessary and their Fusion feature set is extracted because these were determined in the first experiment to be the best parameters overall (see Fig. 1). For classification in EVA, we used the Victoria Police fingermark database $\boldsymbol{T}_{\mathrm{VP}}$ and its ground truth to train and optimise classifiers. Each classifier is trained on the scanner images, and its parameters are chosen via the lowest error at a fixed false match rate (FMR) for the DSLR and phone images.

We removed the fingermarks labelled as "smeared", "twisted lightly" and "strong twist" from $\boldsymbol{T}_{\mathrm{VP}}$ to give the test set $\boldsymbol{T}^{*}_{\mathrm{VP}}$ because most of them had no evidential value according to the experts (see Table 1).

For $\boldsymbol{T}^{*}_{\mathrm{VP}}$, the trained classifier is DA optimised at FMR100, and for the other three test sets, the trained classifier is $k$-NN optimised at FMR5. We made these choices experimentally. For each test set $\boldsymbol{T} \in \{\boldsymbol{T}^{*}_{\mathrm{VP}}, \boldsymbol{T}^{C}_{\mathrm{IIIT-D}}, \boldsymbol{T}^{T}_{\mathrm{IIIT-D}}, \boldsymbol{T}_{\mathrm{BP}}\}$, the trained classifier is applied to each image's feature set and its EV score $q \in [0, 1]$ is obtained. For $\boldsymbol{T}^{*}_{\mathrm{VP}}$, we restrict to the scanner images. Now, each image in a test set $\in \boldsymbol{T}$ has a quality value $q$. Basic statistics for the quality value distributions appear in Table 3. All test sets have similar image quality on average, with the exception of $\boldsymbol{T}^{T}_{\mathrm{IIIT-D}}$ (IIIT-D, marks lifted off a tile), which is lower.

We compare a query image ($\in \boldsymbol{T}$) against the corresponding reference set $\boldsymbol{R} \in \{\boldsymbol{R}_{\mathrm{VP}} \cup \boldsymbol{R}_{\mathrm{FVC}}, \boldsymbol{R}_{\mathrm{IIIT-D}} \cup \boldsymbol{R}_{\mathrm{FVC}}, \boldsymbol{R}_{\mathrm{BP}}\}$ and sort the resulting match scores from high to low. This is the retrieval order. The actual rank is the position at which the corresponding print ($\in \boldsymbol{R}$) is found.

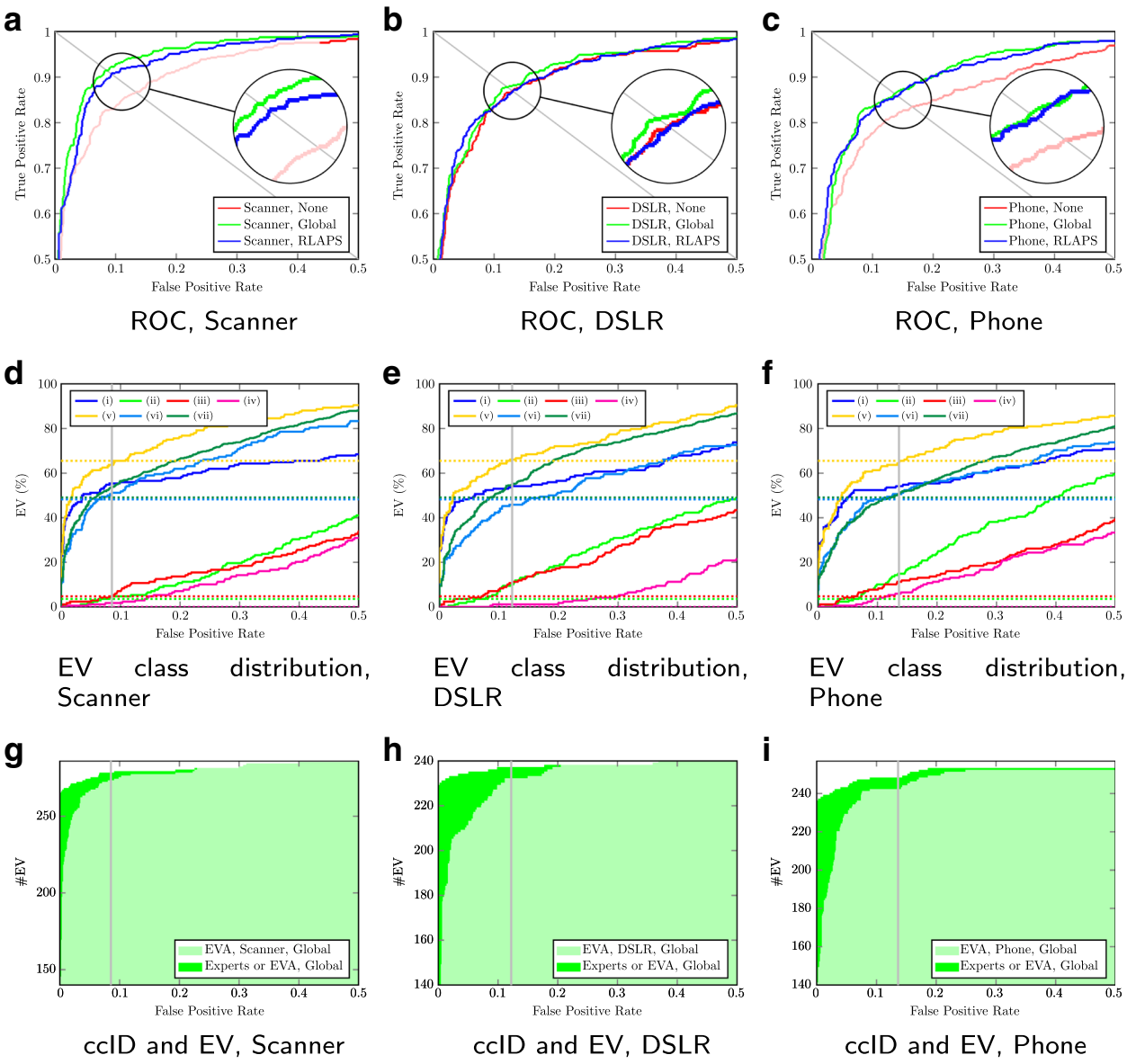


**Fig. 4** ROCs (**a**–**c**), EV distribution according to EVA and experts (**d**–**f**) and ccID results (**g**–**i**). All results are for the Fusion feature set. The *first row* (**a**–**c**) shows the top left corner of the receiver operating characteristics (ROCs) for all capture devices. The colour varies according to the fraction of $^{\text{EV and ccID}}/_{\text{ccID}}$ 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 1 across the whole range in all other cases). With Global scaling, the *second row* (**d**–**f**) shows the EV distribution according to the mark's distortion class, as computed by EVA (*solid*) and the experts (*dashed*). The latter is unaffected by the decision threshold and remains constant. The *third row* (**g**–**i**) shows the number of ccID marks classified as EV by EVA (*light colour*), any additional ones by the experts (*dark colour*) and those not classified as EV by either (*white*). The *gray line* is the threshold corresponding to the EER when the operating point moves along the ROC

Finally, for each $\boldsymbol{T}$, we threshold on quality value $t$ to create a reduced test set $\boldsymbol{T}(t)$ of increasingly higher quality specimens and recompute retrieval order and rank. We calculate the percentage of all query images that have retrieved their mate within rank $k$ w.r.t. the quality threshold $t$. The results are shown in Fig. 5 and are summarised in Table 4. The correct retrieval rate at rank $k$ improves across all databases if low-quality specimens, as determined by EVA, are removed.

## 4 Conclusions

The first experiment shows a strong correlation between the EV score output by EVA and if a particular fingermark can be confidently identified by a commercial matcher

**Table 2** Number of fingermarks which have been correctly and with confidence identified (ccID) and the number of ccID marks which have been classified by experts or EVA to be of sufficient EV w.r.t. capture device (Scanner, DSLR, Phone), CRE Global and feature set (NFIQ2, Verifinger, Fusion) if applicable

| | | Capture device | | |
|---|---|---|---|---|
| | CRE | Scanner | DSLR | Phone |
| ccID | None | 1 | 244 | 10 |
| | Global | 286 | 240 | 257 |
| | RLAPS | 115 | 122 | 119 |
| Experts | Global | 263 | 228 | 236 |
| $EVA_{NFIQ2}$ | Global | 268 | 228 | 246 |
| $EVA_{Verifinger}$ | | 269 | 224 | 242 |
| $EVA_{Fusion}$ | | 273 | 232 | 242 |

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 the ccID rows above)

(see Fig. 4). 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 [2]. 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 scaling factor and the DSLR images without any scaling 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) scaling are used. Nevertheless, the image quality due to the use of different capture devices is not a major drawback. The phone performs more strongly than the DSLR but falls shy of the scanner, under the condition that the capture resolution is adjusted properly.

Table 1 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 or relevance) than just image quality. Nevertheless, an expert's accuracy and repeatability regarding borderline decisions can vary, mostly due to the print quality [9, 10].

We note there are ccID marks which have no sufficient evidential value according to the experts' assessment. This might be again due to experimental setup that heavily favours image processing algorithms or to the limited size of the test population and database. Some of the ccID fingermarks are only considered to be of evidential value by the experts or EVA, but not both. For instance, in Table 2, the 21 out of 257 ccID fingermarks not passed by the experts, and the 15 missed by EVA, may not have evidential value but they do represent a loss of criminal intelligence if they were not collected. Encouragingly, only 6 marks have EV according to the experts but not according to EVA. However, for phone images, there is no false acceptance rate for EVA at which criminal intelligence is not lost, since there are always marks passed by experts but not by EVA (Fig. 4i).

The difference between the image feature sets extracted is rather small but should be considered in a real-world framework.

The second experiment shows, perhaps surprisingly, that EVA can be used to estimate the EV and infer the image quality for specimens taken from other databases, even if they are from a different ridge-based biometric.

Furthermore, it seems that $\boldsymbol{T}^{C}_{\mathrm{IIIT-D}}$ (IIIT-D, marks lifted off a card) contains marks which are less distinctive and more challenging to match successfully. Interestingly enough, $\boldsymbol{T}^{T}_{\mathrm{IIIT-D}}$ seems to perform worse than $\boldsymbol{T}^{C}_{\mathrm{IIIT-D}}$ when the entire test set is used and no specimens are excluded but the situation reverses when the specimens with the worst quality as determined by EVA are excluded.

The correct retrieval rates at rank $k$ for ballprints seem low when compared to the other test sets which exhibit similar image quality statistics. Reasons are that the distinctiveness of a newborn's ballprint suffers from the small and fragile ridge-lines which are incredibly difficult to capture properly [25, 26]. Secondly, the playful or uncooperative nature of the child increases the difficulty to capture the same area of the ball reliably. This leads to prints which have only a small area in common and hence are difficult to match correctly. Nevertheless, the rank 1 retrieval rate of about 17 % is comparable with other newborn studies [25].

**Table 3** Image quality value distribution (mean, standard deviation, first and third quartile) for the different image sets as inferred by EVA

| | $\boldsymbol{T}_{\mathrm{VP}}$ | $\boldsymbol{T}^{*}_{\mathrm{VP}}$ | $\boldsymbol{T}^{C}_{\mathrm{IIIT-D}}$ | $\boldsymbol{T}^{T}_{\mathrm{IIIT-D}}$ | $\boldsymbol{R}_{\mathrm{BP}} \cup \boldsymbol{T}_{\mathrm{BP}}$ | $\boldsymbol{T}_{\mathrm{BP}}$ |
|---|---|---|---|---|---|---|
| Mean $\pm SD$ | $0.32 \pm 0.41$ | $0.48 \pm 0.43$ | $0.49 \pm 0.32$ | $0.39 \pm 0.32$ | $0.48 \pm 0.12$ | $0.46 \pm 0.10$ |
| $Q_1$ | 0.00 | 0.02 | 0.14 | 0.07 | 0.41 | 0.39 |
| $Q_3$ | 0.82 | 0.97 | 0.79 | 0.76 | 0.55 | 0.53 |

For $\boldsymbol{T}_{\mathrm{VP}}$ and $\boldsymbol{T}^{*}_{\mathrm{VP}}$, only the scanned images are used

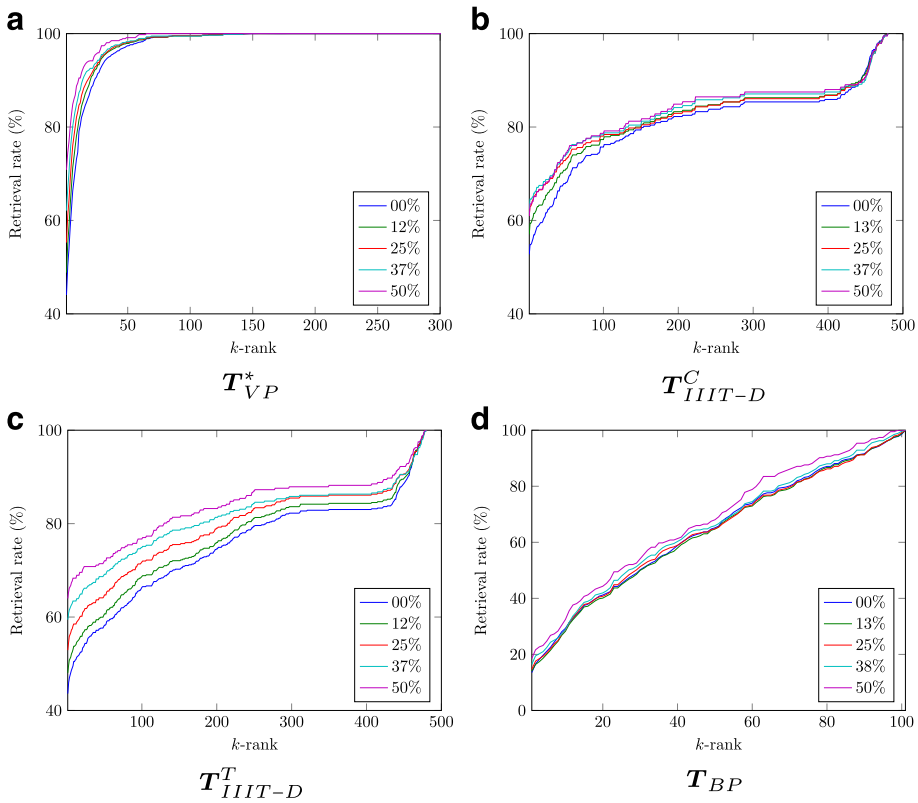


**Fig. 5** Retrieval of the correct mate within rank *k*. Results for the retrieval of the correct mate within rank *k* for different databases when one thresholds on the quality score for identification. The legends indicate the percentage of images removed from the test set. **a** $T^*_{\text{VP}}$, **b** $T^C_{\text{IIIT-D}}$, **c** $T^T_{\text{IIIT-D}}$, and **d** $T_{\text{BP}}$

The EVA shows similar behaviour on all databases, even for newborn ballprints, for which the algorithm has been neither trained nor optimised, and which suffer from severe quality issues due to small and fragile ridges, and the uncooperative nature of the newborn. This emphasises the algorithm's robustness.

We have shown that there is a strong correlation between the fact that a fingermark can be automatically identified with confidence and its EVA-inferred evidential value. We gain confidence that EVA could be used as a pre-filter for submitting marks to systems such as IAFIS for matching. Since EVA is trained on expert decisions, the correlation of EV score $q$ with expert decisions means EVA is a candidate for objectively flagging difficult latents (with low $q$) before or in parallel with forensic examiner analysis. This would have procedural consequences for forensic services.

**Table 4** Percentage of mates retrieved within rank *k* for various databases from the corresponding reference database

| | $T^*_{\text{VP}}$ | $T^C_{\text{IIIT-D}}$ | $T^T_{\text{IIIT-D}}$ | $T_{\text{BP}}$ |
|---|---|---|---|---|
| Rank 1 | 44.05 %(@0 %) | 52.74 %(@0 %) | 43.61 %(@0 %) | 13.52 %(@0 %) |
| Rank 5 | 62.55 %(@0 %) | 55.61 %(@0 %) | 47.98 %(@0 %) | 21.11 %(@0 %) |
| Rank 10 | 74.57 %(@0 %) | 57.96 %(@0 %) | 50.47 %(@0 %) | 29.30 %(@0 %) |
| Rank 20 | 87.66 %(@0 %) | 60.31 %(@0 %) | 52.65 %(@0 %) | 41.19 %(@0 %) |
| Rank 1 | 71.00 %(@50 %) | 62.35 %(@33 %) | 64.17 %(@50 %) | 17.27 %(@49 %) |
| Rank 5 | 82.90 %(@50 %) | 65.52 %(@47 %) | 66.87 %(@50 %) | 23.29 %(@49 %) |
| Rank 10 | 89.46 %(@50 %) | 66.67 %(@38 %) | 68.42 %(@50 %) | 29.72 %(@49 %) |
| Rank 20 | 94.19 %(@50 %) | 68.97 %(@47 %) | 69.66 %(@50 %) | 42.50 %(@43 %) |

The upper half of the table indicates the results without the removal of low-quality specimens; the lower half states the highest retrieval rate when up to 50% of the images are removed based on their quality estimated by EVA

Our findings also indicate that it is feasible to use an automatic mobile phone EVA application to determine if a fingermark at a crime scene is of sufficient evidential value and could be collected by non-experts. In the case that the capture conditions are unknown, it is sensible to use a capture resolution estimator to improve performance.

Furthermore, our results suggest that EVA can be applied to other ridge-based biometrics such as ballprint to derive a specimen's image quality despite being trained and optimised on fingermarks. This underlines the algorithm's robustness.

In the future, we would like to perform more exhaustive testing of EVA on additional and considerably larger databases and with different matching systems. 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 and apply EVA there, even to marks of other ridge-based biometrics.

**Acknowledgements**
We thankfully acknowledge the Victoria Police fingerprint examiners who kindly assessed all fingermarks. We thank Image Analysis and Biometrics Lab @ IIIT Delhi for providing the IIIT-D Latent Fingerprint Database. The research project is funded by the Victoria Police. We would like to thank Dr. Arathi Arakala for her valuable input and suggestions. Most of this work forms part of the PhD thesis of the first author.

**Authors' contributions**
JK, LJS and RNJV developed EVA. RH helped supplying the Victoria Police database and provided his expertise in the area of forensics. JK, SAD and KJH conceived and designed the experiments; JK performed them. HH prepared the IIIT-D feature sets. JK, HH and KJH contributed to the writing of the final manuscript, and all read and approved it.

**Competing interests**
The authors declare that they have no competing interests.

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

Received: 27 April 2016 Accepted: 14 October 2016
Published online: 26 October 2016

**References**

1. D Meuwly, in *Encyclopedia of Biometrics*, ed. by SZ Li, AK Jain. Forensic use of fingerprints and fingermarks (Springer, New York City, 2014), pp. 1–15
2. J Kotzerke, SA Davis, R Hayes, LJ Spreeuwers, RNJ Veldhuis, KJ Horadam, in *Biometrics and Forensics (IWBF), 2015 International Workshop On*. Discriminating fingermarks with evidential value for forensic comparison (IEEE, New York City, 2015), pp. 1–6
3. J Kotzerke, S Davis, R Hayes, L Spreeuwers, R Veldhuis, K Horadam, in *Biometrics Special Interest Group (BIOSIG), 2015 International Conference of The*. Identification performance of evidential value estimation for fingermarks (IEEE, New York City, 2015), pp. 1–6. doi:10.1109/BIOSIG.2015.7314607
4. D Maltoni, D Maio, AK Jain, S Prabhakar, *Handbook of Fingerprint Recognition,* 2nd edn. (Springer, New York City, 2009)
5. DR Ashbaugh, *Quantitative-Qualitative Friction Ridge Analysis: an Introduction to Basic and Advanced Ridgeology*. (CRC press, Boca Raton, Florida, 1999)
6. J Feng, Y Shi, J Zhou, Robust and efficient algorithms for separating latent overlapped fingerprints. IEEE Trans. Inf. Forensic Secur. **7**(5), 1498–1510 (2012). doi:10.1109/TIFS.2012.2204254
7. NJ Short, MS Hsiao, AL Abbott, EA Fox, in *Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th International Conference On*. Latent fingerprint segmentation using ridge template correlation (IET, Stevenage, 2011), pp. 1–6. doi:10.1049/ic.2011.0125
8. Z Yao, JML Bars, C Charrier, C Rosenberger, Literature review of fingerprint quality assessment and its evaluation. IET Biom. **5**(3), 243–251 (2016). doi:10.1049/iet-bmt.2015.0027
9. BT Ulery, RA Hicklin, J Buscaglia, MA Roberts, Accuracy and reliability of forensic latent fingerprint decisions. Proc. Natl. Acad. Sci. **108**(19), 7733–7738 (2011)
10. BT Ulery, RA Hicklin, J Buscaglia, MA Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners. PloS ONE. **7**(3), 32800 (2012)
11. PJ Kellman, JL Mnookin, G Erlikhman, P Garrigan, T Ghose, E Mettler, D Charlton, IE Dror, Forensic comparison and matching of fingerprints: using quantitative image measures for estimating error rates through understanding and predicting difficulty. PloS ONE. **9**(5), 94617 (2014)
12. F Alonso-Fernandez, J Fierrez, J Ortega-Garcia, J Gonzalez-Rodriguez, H Fronthaler, K Kollreider, J Bigun, A comparative study of fingerprint image-quality estimation methods. IEEE Tran. Inf. Forensic Secur. **2**(4), 734–743 (2007)
13. Y Chen, SC Dass, AK Jain, in *Audio-and Video-Based Biometric Person Authentication*. Fingerprint quality indices for predicting authentication performance (Springer, New York City, 2005), pp. 160–170
14. H Fronthaler, K Kollreider, J Bigun, in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference On*. Automatic image quality assessment with application in biometrics (IEEE, New York City, 2006), pp. 30–30
15. S Lee, H Choi, K Choi, J Kim, Fingerprint-quality index using gradient components. IEEE Trans. Inf. Forensic Secur. **3**(4), 792–800 (2008)
16. The National Institute of Standards and Technology, NFIQ2 Feature Definitions Document (v0.5) (2013). http://biometrics.nist.gov/cs_links/quality/NFIQ_2/NFIQ-2_Quality_Feature_Defin-Ver05.pdf. Accessed 23 Oct 2016
17. S Yoon, K Cao, E Liu, AK Jain, in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference On*. Lfiq: Latent fingerprint image quality (IEEE, New York City, 2013), pp. 1–8
18. Neurotechnology: VeriFinger SDK (2015). http://www.neurotechnology.com/verifinger.html. Accessed 23 Oct 2016
19. A Sankaran, TI Dhamecha, M Vatsa, R Singh, in *Biometrics (IJCB), 2011 International Joint Conference On*. On matching latent to latent fingerprints (IEEE, New York City, 2011), pp. 1–6
20. I Analysis, BLI Delhi, Fingerprint resources (2016). http://iab-rubric.org/resources.html#finger. Accessed 23 Oct 2016
21. D Maio, D Maltoni, R Cappelli, JL Wayman, AK Jain, Fvc2000: Fingerprint verification competition. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 402–412 (2002)
22. D Maio, D Maltoni, R Cappelli, JL Wayman, AK Jain, in *Biometric Authentication*. Fvc2004: Third fingerprint verification competition (Springer, New York City, 2004), pp. 1–7
23. D Maio, D Maltoni, R Cappelli, JL Wayman, AK Jain, in *Pattern Recognition, 2002. Proceedings. 16th International Conference On, vol. 3*. Fvc2002: Second fingerprint verification competition (IEEE, 2002), pp. 811–814
24. J Kotzerke, S Davis, K Horadam, J McVernon, in *Proceedings of 2013 IEEE 20th International Conference on Image Processing (ICIP 2013)*, ed. by IEEE. Newborn and infant footprint crease pattern extraction (IEEE, New York City, 2013)
25. J Kotzerke, A Arakala, S Davis, K Horadam, J McVernon, in *Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings, 2014 IEEE Workshop On*. Ballprints as an infant biometric: A first approach (IEEE, New York City, 2014), pp. 36–43. doi:10.1109/BIOMS.2014.6951533
26. D Weingaertner, ORP Bellon, L Silva, MNL Cat, Newborn's biometric identification: can it be done. Proc. VISAPP. **1**, 200–205 (2008)