Evaluation of AFIS-ranked latent fingerprint matched templates

Ram P. Krish[1], Julian Fierrez[1], Daniel Ramos[1], Raymond Veldhuis[2], and Ruifang Wang[1]

[1] Biometric Recognition Group - ATVS, EPS - Univ. Autonoma de Madrid
C/ Francisco Tomas y Valiente, 11 - Campus de Cantoblanco - 28049 Madrid, Spain
{ram.krish,julian.fierrez,daniel.ramos,ruifang.wang}@uam.es
[2] Chair of Biometric Pattern Recognition, Faculty EEMS,University of Twente,
PO Box 217, 7500 AE Enschede, The Netherlands
{r.n.j.veldhuis}@utwente.nl

Abstract. The methodology currently practiced in latent print examination (known as ACE-V) yields only a decision as result, namely individualization, exclusion or inconclusive. From such a decision, it is not possible to express the strength of opinion of a forensic examiner quantitatively with a scientific basis to the criminal justice system. In this paper, we propose a framework to generate a score from the matched template generated by the forensic examiner. Such a score can be viewed as a measure of confidence of a forensic examiner quantitatively, which in turn can be used in statistics-based evidence evaluation framework, for e.g, likelihood ratio. Together with the description and evaluation of new realistic forensic case driven score computation, we also exploit the developed experimental framework to understand more about matched templates in forensic fingerprint databases.

Keywords: ACE-V methodology, criminology, forensics, latent fingerprint, likelihood ratio, quantification of evidence.

1 Introduction

Latent fingerprints (or partial fingermarks) lifted from the crime scenes have been used for identification for more than a century. The use of fingerprints in *criminology* was popularized by Sir Francis Galton in late 19^{th} century. The philosophy surrounding its use in criminology is the fact that fingerprint of an individual is unique, and the friction ridge pattern being persistent over time [1]. Starting from early 20^{th} century, several methods were proposed to formalize a standard friction ridge analysis for forensic examination. Eventually the forensic examination methodology matured to a standard procedure called ACE-V (analysis, comparison, evaluation and verification) which is currently followed in the forensic community. But recently this kind of procedure has been criticized, arguing for a robust quantitative evaluation of the weight of evidence [4] [2].

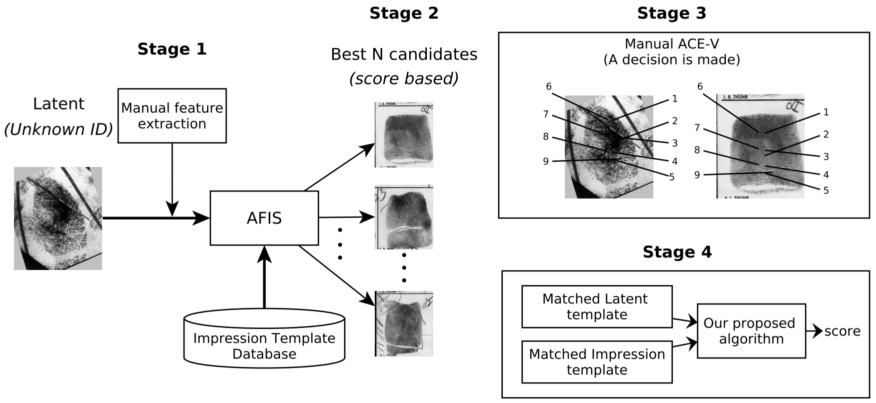


Fig. 1. Stage 1 to Stage 3 captures the Latent fingerprint examination methodology currently practiced. In Stage 4, we propose our framework to generate a score from matched template obtained from Stage 3.

The various stages involved in current forensic examination methodology are summarized in Stage 1, Stage 2 and Stage 3 in Fig.1. In Stage 1, a forensic examiner manually extracts the minutiae features from a latent fingerprint, and these feature are then converted into a digital template format used by an Automated Fingerprint Identification System (AFIS). In Stage 2, the AFIS compares the manually extracted latent template with all the templates in the forensic database and shortlists a set of possible suspects ranked based on a similarity score. In Stage 3, the forensic examiner manually follows the ACE-V methodology for the given latent print against all the templates shorlisted by the AFIS, and yields a decision whether the prints match, do not match or the comparison is inconclusive.

ACE-V methodology comprises of the following four phases [6] [12]:

1. *Analysis :* The examiner looks for sufficiency of the details present in the given latent print. This comprises of checking for ridge clarity, quantity of Level 1, Level 2 and Level 3 details, and determining the anatomical sources - whether the print came from finger, palm, toe or foot.
2. *Comparison :* Once the latent print passes the analysis phase, many useful friction ridge details together with the minutiae feature are extracted manually and are compared against one or more exemplar/reference fingerprints shorlisted by an AFIS to determine whether they are in agreement.
3. *Evaluation :* Based on the conclusions derived from the analysis and comparison phases, the forensic examiner yields a decision as *individualization (identification or match), exclusion (non-match) or inconclusive.*
4. *Verification :* In this phase, another qualified forensic examiner reexamines the previous decision by following the above three phases.

The latent fingerprints which are the unintentional traces left behind by the perpetrator or by the victim are of poor quality in nature [9] [7] [3]. So, a reliable manual feature extraction is mainly influenced by the perception and decision making ability of a forensic examiner, which eventually affects the final decision. One of the more popularly cited examples where an erroneous individualization was made is the Brandon Mayfield case. Other similar cases of erroneous individualization have been reported in [8].

There is no established protocol to characterize any uncertainty involved in the ACE-V procedure. Also, there is no scientific framework currently in use at the criminal justice system to express the strength of opinion of a forensic examiner quantitatively. The new paradigm coming forward in this regard [10] avoids hard identification decisions by considering evidence reporting methods that incorporate uncertainty and statistics. Amongst all the methods of evidence evaluation, the likelihood ratio is receiving greater attention [5] [12]. To use likelihood ratio incorporating the ACE-V level, scores are required in place of decisions. This was the motivation for the current work, where we developed a framework, Stage 4 in Fig.1 to take the matched templates from the ACE-V stage and generate a score as a measure of confidence for forensic examiner. This score can be used in statistics-based evidence evaluation framework to derive quantitative weight to express the strength of opinion of the examiner with adequate scientific basis.

The remainder of the paper is organized as follows. We first explain in detail about the real forensic casework databases used in this study, then the method developed to generate a score as a measure of confidence for a forensic examiner. We then present the experimental protocol and results, followed by a discussion on the usefulness of the technique developed in quantifying the evidence of fingerprints.

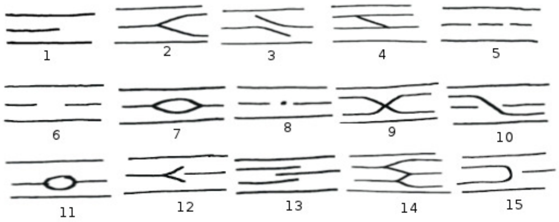


Fig. 2. Minutiae types used in Guardia Civil Database. Names corresponding to individual type numbers can be found in Table 1.

2 Database

The forensic fingerprint database largely used in academic domain is the NIST Special Database (SD) 27, which was made publicly available by NIST. This

No	TypeName	No	TypeName	No	TypeName
1	Ridge Ending	6	Interruption	11	Circle
2	Bifurcation	7	Enclosure	12	Delta
3	Deviation	8	Point	13	Assemble
4	Bridge	9	Ridge Crossing	14	M-structure
5	Fragment	10	Transversal	15	Return

Table 1. List of typical and rare minutiae in Guardia Civil Database. Numbering with respect to Fig.2.

database is used by researchers working in forensic domain to understand about the challenges, as well as to have a transparent and benchmark database to evaluate the approaches developed by the researchers. This database has both images as well as minutiae validated by forensic examiners.

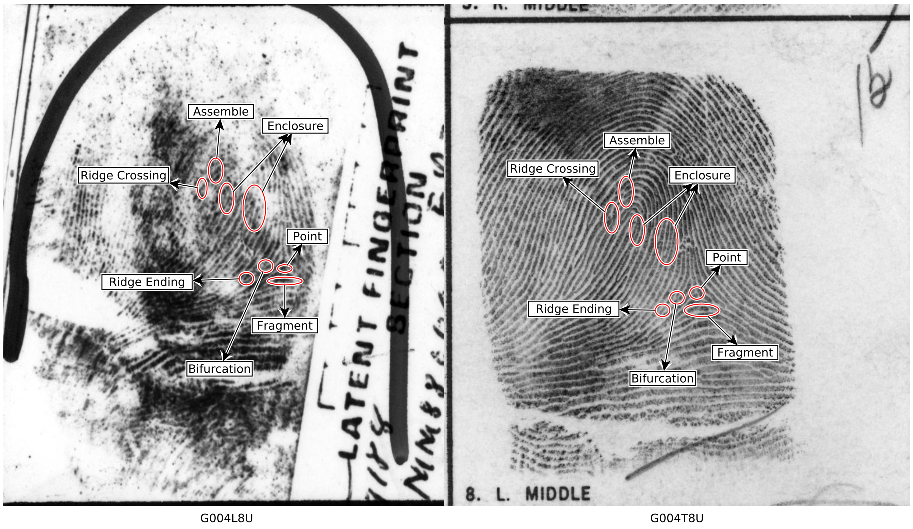
NIST SD27 minutiae template database is broadly classified into two: 1) *ideal*, and 2) *matched* minutiae database. The *ideal* minutiae set for latents was manually extracted by a forensic examiner without any prior knowledge of its corresponding impression image. The *ideal* minutiae for impressions was initially extracted using an AFIS, and then these minutiae were manually validated by at least two forensic examiners. The *matched* minutiae templates contains those minutiae which are in common between the latent and its mated impression image. There is a one-to-one correspondence in the minutiae between the latent and its mate in the matched template. This ground truth was established by a forensic examiner looking at the images and the *ideal* minutiae following an ACE-V procedure. For NIST SD27 database, only the ideal-latent templates had type information for each minutiae in addition to location and orientation attributes. The other three datasets (ideal-impression, matched-latent and matched-impression) do not have type information but only location and orientation attributes.

In this study, we also acquired the forensic fingerprint database from Guardia Civil, the law enforcement agency of the Government of Spain. The Guardia Civil database (GCDB) is similar to the NIST SD27 database except that all the templates in ideal and matched sets in GCDB have type information. Apart from having typical minutiae types (*ridge-endings, bifurcations*), GCDB also comprises rare minutiae types like *fragments*, *enclosures*, *points/dots*, *interruptions, etc.* Please refer to Fig.2 and Table 1 for a comprehensive list of all minutiae feature types present in GCDB. Table 2 shows the statistics of various types of minutiae features present in the 258 template pairs available in GCDB. Rest of the minutiae types were not observed so far in this collection of GCDB.

We will follow the notation GCDB-M and NIST-SD27-M to denote the matched template database of GCDB and NIST SD27 respectively. In Fig.3, we show a latent fingerprint and its corresponding impression with typical features and some of the rare features annotated with their correspondences. The latent and impression images used here were taken from NIST SD27 database, and the typical and rare minutiae features were manually annotated on them.

No	Contribution	No	Contribution	No	Contribution
1	56%	4	0.265%	7	2.058%
2	36.38%	5	4.515%	8	0.332%
3	0.166%	6	0.232%	10	0.0332%

Table 2. Statistics of typical and rare minutiae present in Guardia Civil Database. Numbering with respect to Fig.2.



G004L8U

G004T8U

Fig. 3. Typical and some rare minutiae features on a latent and its mated impression fingerprint. The latent image G004L8U and the impression image G004T8U were taken from NIST SD27 database.

3 Algorithm

We propose an algorithm to generate a score for the templates in GCDB-M. This algorithm can be adapted to templates from NIST-SD27-M by discarding the weights for type information when calculating *typeError* explained in the algorithm. The various stages involved in the computation of the score are as follows:

3.1 Alignment and correspondence

Since the framework is developed to deal with matched databases, we expect that for genuine matches, superimposing the centroids of both latent and impression minutiae points with appropriate rotation alignment would lead to an approximate fitting of point patterns based on mated pairs with minimum overall fitting error, and for impostors it would lead to a high fitting error.

As typical minutiae features are the majority with 92% (see Table 2), we only use typical features to estimate the rotation parameters. By rotating the latent template over the impression template w.r.t centroid in a range of $[-45°, +45°]$, we find the closest matching minutiae pairs, and add their distance. The rotation for which the average sum of closest pairs is the minimum is considered to be the best rotation alignment for their approximate pattern fitting.

After the alignment, all those minutiae pairs which are within a threshold distance are considered to be mated pairs, and their correspondences are established.

3.2 Fitting and Orientation errors

Once the correspondences are established for all the typical minutiae features, the scores are computed hierarchically looking at each of minutiae attributes, namely *location, orientation* and *type* information. Scores based on *type* information are discussed in the next subsection.

For all the typical minutiae which established correspondences based on optimal rotation, we find a fitting error using an affine transformation for the mated minutiae patterns by least square fitting. This score is denoted as *fittingError*, which is averaged w.r.t total number of mated minutiae pairs.

Again for all the mated minutiae pairs, we sum up all the orientation differences of corresponding minutiae and average this sum of degrees w.r.t total number of mated pairs. When averaging the orientations, the circularity of degrees are taken care of. This score is denoted as *orientationError.*

3.3 Type errors

If the mated pairs disagree w.r.t *type* information, which otherwise are mated based on only location and orientation attributes, we associate a penalty for such type of mismatches. The penalty for each typical minutiae type is a constant factor estimated from Table 2. This score is denoted as *typeError.* This is possible because the *type* information for both latent and impression are estimated manually by a forensic examiner, and we assume type information is available here.

Based on the alignment estimated using *typical* minutiae, we also look for the presence of *rare* minutiae correspondences. If they are within a location and orientation threshold, then they constitute mated pairs, and thus correspondence is established. As the percentage of occurrence of rare minutiae is very small, around 8%, we only estimate *typeError* for rare minutiae. The penalty for each rare minutiae type is a constant factor estimated from Table 2.

3.4 Final Score

Since all the individual scores we have generated are of different nature, namely *fittingError* in distance, *orientationError* in degrees, *typeError* in probability based cost, these scores are combined using logistic regression to generate the final score [11]:

$$\begin{aligned} finalScore = &(\alpha \times fittingError) \\ &+ (\beta \times orientationError) \\ &+ (\gamma \times typeError) \end{aligned} \tag{1}$$

where α, β, γ are the logistic regression coefficients for each classifier respectively.

This final score can be viewed as a measure of confidence of the forensic examiner numerically, otherwise the forensic examiner only have a logical decision at the stage of ACE-V. Note that the $finalScore$ is a dissimilarity score, so the higher the score the higher the distance between a match and non-match.

4 Experiments

4.1 Experimental protocol

The total number of latent fingerprint templates in GCDB-M is 258, with their corresponding matched impression fingerprint templates. This size of GCDB-M is equivalent to the publicly available NIST-SD27-M. This way, we could do some performance comparisons between databases, unbiased in terms of partitioning for train and test dataset sizes. For training the logistic regression coefficients, we used 129 template pairs and 129 for testing.

4.2 Results

We performed a 2-fold cross validation to study the performance of the developed approach by comparing the degree of overlap between matching and non-matching scores in the matched databases. Various parameters like the distance and orientation thresholds were finetuned to minimize this degree of the overlap.

We also tested the performance of a commercial SDK from Neurotechnology Verifinger 4.3which is a general purpose matcher, on both GCDB-M and NIST-SD27-M. When using Verifinger for the analysis, all 258 template pairs of GCDB-M and NIST-SD27-M were used for testing.

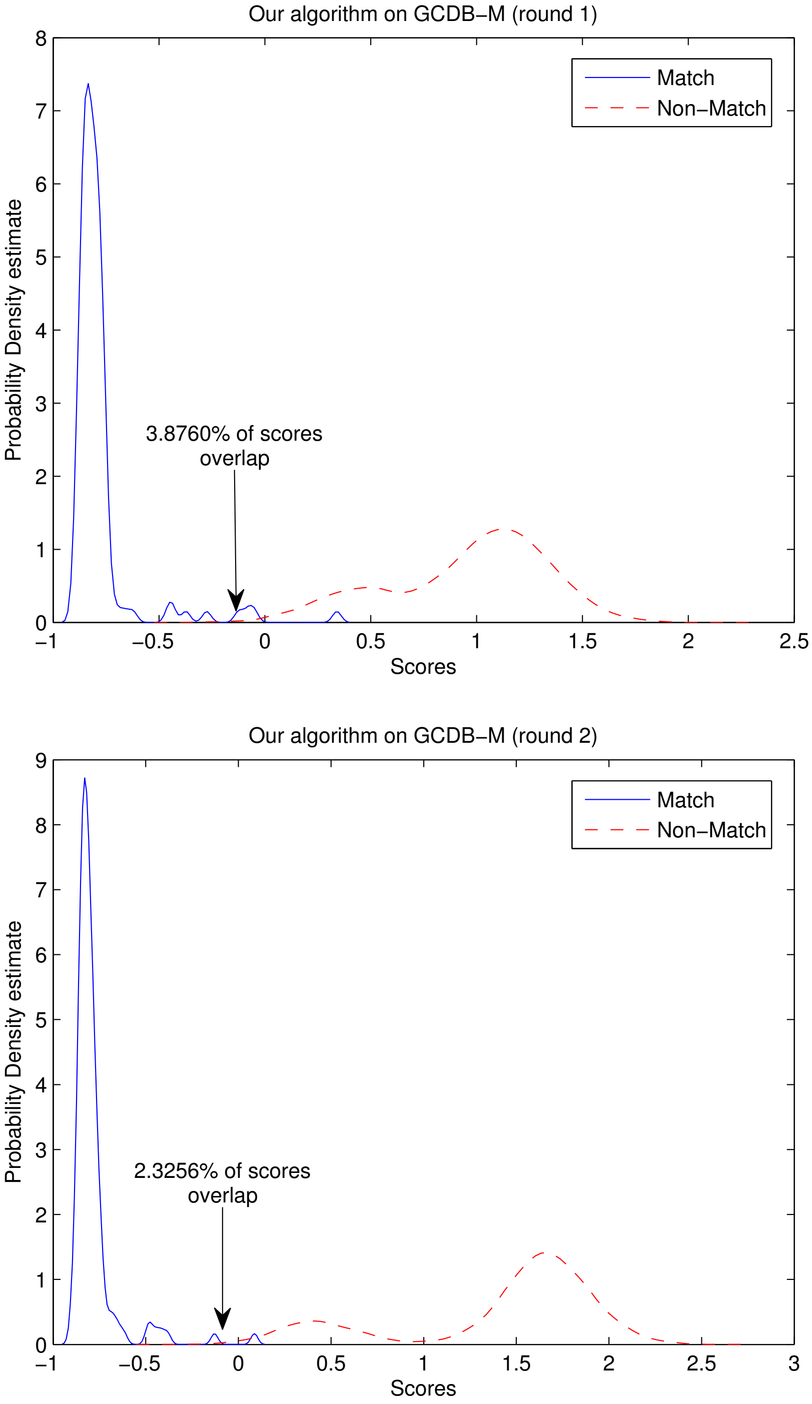


Fig. 4. 2-fold cross validation performance of our algorithm on GCDB-M.

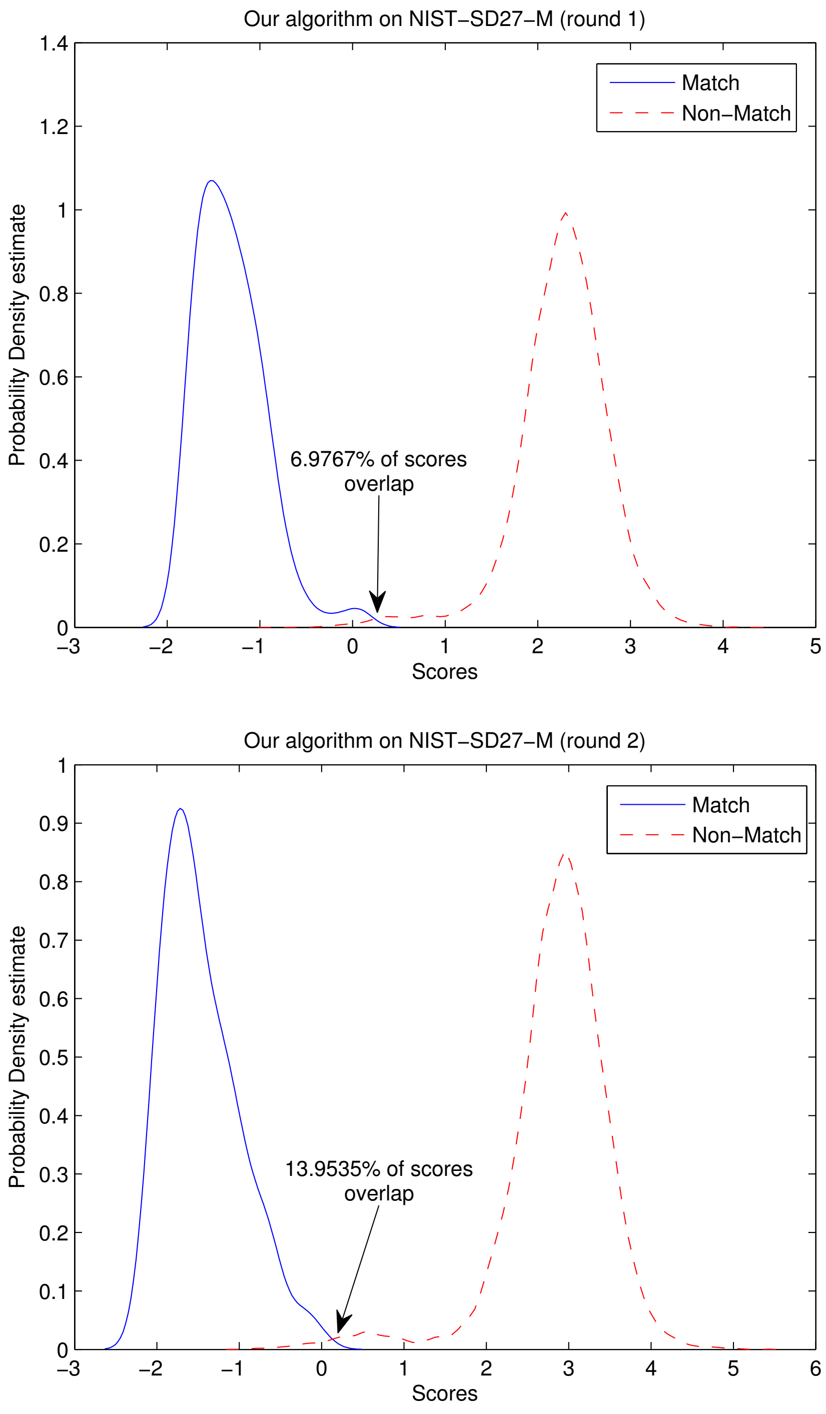


Fig. 5. 2-fold cross validation performance of our algorithm on NIST-SD27-M.

	GCDB-M	NIST-SD27-M
Our algorithm	3.1008%	10.4651%
Verifinger	-	7.7519%

Table 3. Degree of overlap of scores generated by our algorithm and Verifinger. Verifinger on GCDB-M is not reported here because the results were too inconsistent.

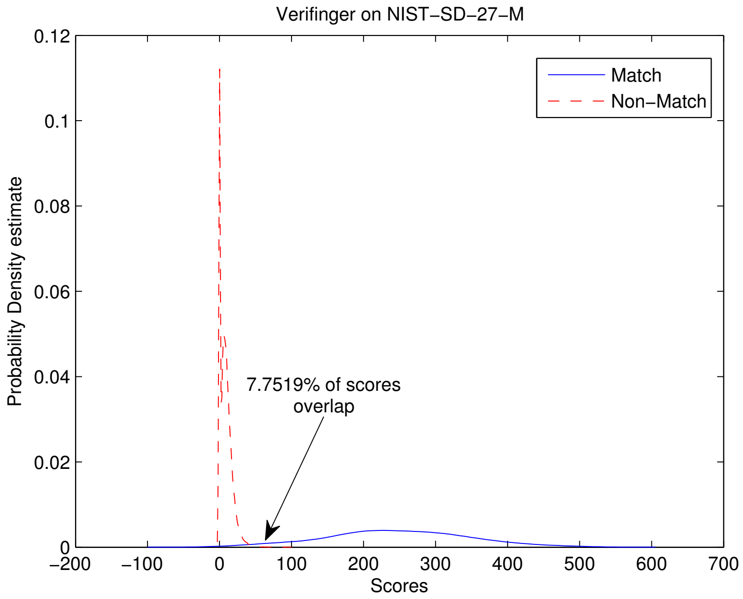


Fig. 6. Performance of Verifinger on NIST-SD27-M.

Fig.4 and Fig.5 shows the degree of overlap of scores on both GCDB-M and NIST-SD27-M using our proposed algorithm respectively. In Table 3, we summarize the degree of overlap of scores in percentage. Since we performed a 2-fold cross validation, the average degree of overlap is shown in the table for our algorithm on both GCDB-M and NIST-SD27-M, but in the figures, the individual overlaps are shown.

Using our algorithm, the average degree of score overlap was 3.10% for GCDB-M, and for each round of cross validation, the degree of score overlaps were 3.87% and 2.32% respectively. Similarly, for NIST-SD27-M, the average degree of score overlap was 10.46%, and for each round, the degree of score overlaps were 6.97% and 13.95% respectively. Fig.6 shows the performance of Verifinger on NIST-SD27-M with a score overlap of 7.75%.

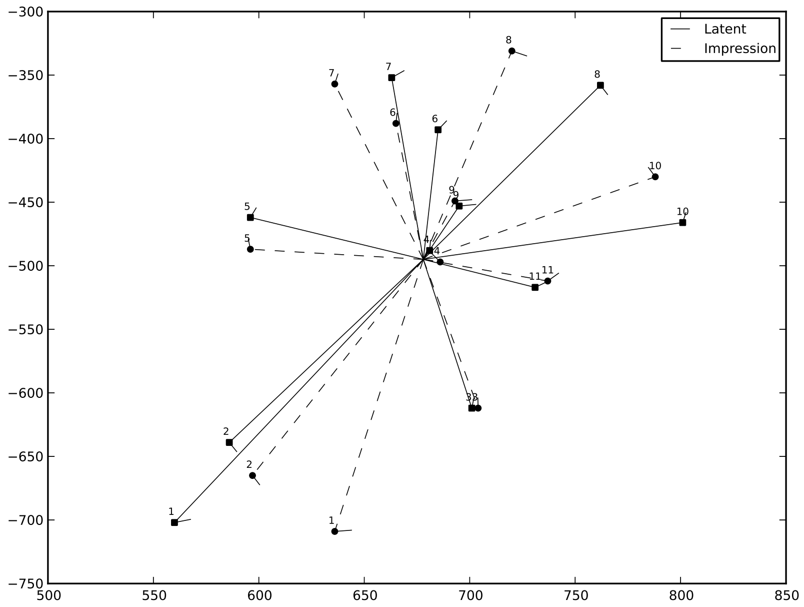


Fig. 7. An example from GCDB-M where a genuine match having high dissimilarity score. The solid lines represent the latent minutia pattern with respect to its centroid, and dashed lines represent the impression minutia pattern with respect to its centroid.

Fig.7 shows a latent (*solid lines*) and its mated impression (*dashed lines*) minutiae (*location* and *orientation*) from GCDB-M superimposed over their centroids before estimating alignment parameter (*rotation*) which leads to high dissimilarity score. Ideally, the mated templates are supposed to show less dissimilarity, but in this case, nonlinear deformations leads to wrong orientation parameter estimation. For e.g, the pairs (1,1), (8,8), (3,3), (11,11) show high nonlinear deformations compared against the pairs (2,2), (5,5), and (6,6). Such disparities makes it difficult to optimally estimate the best alignment parameter, and as a consequence leads to high fitting error among the mated minutiae pairs. This shows the limitation of the Affine transformation model used in the algorithm for estimating fitting error to capture the nonlinear deformations of the pattern globally.

5 Discussions

We developed a framework to generate a score from matched templates generated by the forensic examiner as a result of ACE-V. This score can be viewed as a measure of confidence of the forensic examiner numerically in place of a logical decision. Such a score can be exploited by any statistics based evidence evaluation framework to include the ACE-V stage. We tested the algorithm on GCDB-M as well as on the publicly available NIST-SD27-M database. We also exploited this framework to understand more about the nature of matched template forensic fingerprint databases.

A deeper analysis about the importance of rare minutiae features, as well as an implementation of the likelihood ratio approach based on these scores to express the strength of opinion of forensic examiners quantitatively is in order.

Acknowledgment

R.K. and R.W. are supported by Marie Curie Fellowships under project BB-for2 (FP7-ITN-238803). This work has also been partially supported by Spanish Guardia Civil, Cátedra UAM-Telefónica, and projects Bio-Shield (TEC2012-34881) and Contexts (S2009/TIC-1485).

References

1. J.G. Barnes, Chapter 1, History, *The Fingerprint Sourcebook*, U.S Department of Justice, 2011.
2. A. Nagar and H.-S. Choi and A. K. Jain, Evidential Value of Automated Latent Fingerprint Comparison: An Empirical Approach, *IEEE Transactions on Information Forensics and Security*, 2012.
3. A.K. Jain and J. Feng, Latent Fingerprint Matching, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 88-100, 2011.
4. C. Neumann and I. Evett and James Skerrett, Quantifying the weight of evidence assigned to a forensic fingerprint comparison: a new paradigm, *J. R. Statist. Soc. A*, 175, 371-415, 2012.
5. I. Evett, et al, Expressing evaluative opinions: A position statement, *Science and justice*, 2011.
6. Expert Working Group on Human Factors in Latent Print Analysis. *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*, NIST, 2012.
7. C. Champod, C.J. Lennard, P. Margot and M. Stoilovic, Fingerprints and other ridge skin impressions, *CRC*, 2004.
8. S:A. Cole, More than zero: Accounting for error in latent fingerprint identification, *Journal of Criminal Law and Criminology*, 985-1078, 2005.
9. F. Alonso-Fernandez, J. Fierrez and J. Ortega-Garcia, Quality Measures in Biometric Systems, *Security Privacy, IEEE*, 2012.
10. M.J. Saks and J.J. Koehler, The Coming Paradigm Shift in Forensic Identification Science, *Science*, 892-895, 2005.
11. F. Alonso-Fernandez, J. Fierrez, D. Ramos and J. Gonzalez-Rodriguez, Quality-Based Conditional Processing in Multi-Biometrics: application to Sensor Interoperability, *IEEE Transactions on Systems, Man and Cybernetics Part A*, 2010.
12. S.N. Srihari, *Quantitative Measures in Support of Latent Print Comparison: Final Technical Report: NIJ Award Number: 2009-DN-BX-K208*, University at Buffalo, SUNY, 2013.