Measuring coherence of computer-assisted likelihood ratio methods

Rudolf Haraksim[a], DanielRamos[b], DidierMeuwly[a], Charles E.H.Berger[a]

Measuring the performance of forensic evaluation methods that compute likelihood ratios (LRs) is relevant for both the development and the validation of such methods. A framework of performance characteristics categorized as primary and secondary is introduced in this study to help achieve such development and validation. Ground-truth labelled fingerprint data is used to assess the performance of an example likelihood ratio method in terms of those performance characteristics. Discrimination, calibration, and especially the coherence of this LR method are assessed as a function of the quantity and quality of the trace fingerprint specimen. Assessment of the coherence revealed a weakness of the comparison algorithm in the computer-assisted likelihood ratio method used.

Keywords: Coherence, Forensic evidence, Fingermark, Fingerprint, Likelihood ratio, Validation

1 Introduction

Forensic research makes progress in the field of evaluation of forensic findings. An increasingly adopted approach [1] uses a logical framework based on Bayes' Theorem to report forensic evidence in terms of likelihood ratios [1,2]. Computer-assisted LR methods (also referred to simply as *LR methods*) have been developed to assist the forensic practitioner in his role of forensic evaluator [3-9]. In these methods pattern recognition algorithms are often used for the feature extraction, the feature comparison, and statistical models are used for the evaluation of the forensic findings.

In this article the term validation refers to a series of experiments, and the application of a set of performance metrics and validation criteria to demonstrate validity – a LR method is valid if it is appropriate for a given use according to given criteria. This is different from its use in [10], where the term validity was defined as a single metric and equated to accuracy. The specific performance characteristics[1], performance metrics[2] and validation criteria[3] are used to describe the performance of methods computing LRs and to assess the limits of their validity when used in casework. The LR describes the strength of the evidence, and does not imply a decision by itself. Therefore, the validation of LRs is not the validation of a decision process, but of a description process. We define *coherence* as a performance characteristic, understood as the ability of a LR method to perform better and to maintain low rates of misleading evidence as the quantity and quality of the features in the trace specimen improves. A concrete example is provided by studying and assessing the coherence of a forensic fingermark evaluation method, based on a comparison algorithm of an AFIS (Automated Fingerprint Identification System). When analysing the coherence of the method we hope to observe strength of a LR value increasing with the intrinsic quantity and quality of the information present in the trace data (such as the length of a speech fragment or the number of minutiae in a fingermark).

Forensic service delivery makes progress in the field of quality assurance. Initiatives in the European Network of Forensic Science Institutes (ENFSI) focus on best practices,

---

[1] Performance characteristic is the characteristic of LR methods that is thought to contribute positively towards the validation of one given method. For instance, LR values should be discriminating in order to be valid, clearly distinguishing between comparisons under different propositions. In this case, discriminating power is a performance characteristic.

[2] Performance metric is the variable whose numeric value measures the performance characteristic. For instance, the rates of misleading evidence are known to measure discriminating power (among other properties), and therefore they can be a performance metric of the discriminating power.

[3] Validation criterion presents a condition related to the performance characteristic that has to be met in order for the LR method to be valid. For instance, a validation criterion can be as follows: only methods with having rates of misleading evidence less than 1% can be considered as valid. Note that several validation criteria can be applied in order to consider a method valid, not only one.

method validation and service accreditation [11]. But because LR methods for forensic evaluation are still very new, the question of their validation has not been addressed in the context of quality assurance yet. Currently, performance characteristics, performance measures, and validation criteria exist to assess analytical forensic methods [12] and human-based methods used for forensic evaluation [13,14]. These approaches are however not suitable for the validation of LR methods developed for forensic evaluation. Such a validation requires specific performance characteristics, performance measures and validation criteria related to the nature of the LRs and the computation methods involved.

Studying the coherence contributes to describing the performance of LR methods using datasets in which some measurable parameters influencing the strength of the evidence vary. The variation of the length of utterances in forensic automatic speaker recognition and the variation of the number of minutiae in fingermarks are examples of such parameters. Coherence is a highly desirable property of a LR method. In this article, we propose a framework for the measurement of performance characteristics towards the establishment of validation protocols for LR methods in forensic science. We particularly focus on the *coherence* performance characteristic, illustrating its importance with an example in AFIS-based fingermark evidence evaluation.

The remainder of this article is structured as follows. The definition of coherence in a set of performance characteristics is presented in Section 2. Section 3 introduces the experimental example for assessment of the coherence of LRs assigned using computer-assisted methods. The different datasets used to measure the performance characteristics are described in Section 4, while the relevance of the use of the datasets is described in Section 5. The performance metrics related to the performance characteristics used are introduced in Section 6. Results in terms of coherence of the LR method are presented in Section 7, followed by general discussion and conclusions in Section 8.

Throughout this article we frequently use the terms performance characteristic and performance metrics. These definitions are ours and the terms may have different meanings in other related works.

2 Performance characteristics

Several performance characteristics have been defined to assess the performance of computer-assisted LR methods developed for forensic evaluation. We propose to structure them into primary and secondary performance characteristics. Primary performance characteristics directly measure desirable properties of the LRs. The secondary performance characteristics measure how sensitive primary performance characteristics are to factors like

the quantity of information in the data and to the forensic casework circumstances, such as degraded quality, different technical and temporal conditions related for example to the acquisition of trace and test[4] specimens, representativeness of the data, etc.

2.1 Primary performance characteristics

To assess the performance of computer-assisted LR methods, several performance characteristics have been defined recently in forensic evaluation [15]. A very important one is accuracy, defined as the combination of discrimination (discriminating power) and calibration [15-17].

Accuracy is defined as the closeness of agreement between the decision – driven by a LR computed by a given method – and the ground truth. With ground-truth we understand the proposition that is actually true in a given case. The LR is accurate if it helps to lead to a decision that is correct.[5] In case of source level inference, the ground truth relates to the following pair of propositions:

H_p: The pair of specimens compared come from the same source (SS)
H_d: The pair of specimens compared come from different sources (DS)

Ground-truth labels are defined as SS (same source) when the LR was calculated for specimens originating from the same source, and as DS (different source) when the LR was calculated for specimens originating from the different sources. If an experimental set of LR values is to be evaluated, and the corresponding ground-truth label of each of the LR values is known, then a given LR value is evaluated as more accurate if it supports the true (known) proposition to a higher degree, and vice versa.

Discrimination (or discriminating power) is a property of a set of LRs that allows distinguishing between the propositions involved. See [15,16] for details.

Calibration[6] is another property of a set of LRs. Perfect calibration of a set of LRs means that those LRs can be probabilistically interpreted as the evidential value of the comparison result for either proposition in a Bayesian evaluation framework. Finding a LR = x will be x

[4] In the fingerprint modality the trace usually refers to the fingermark recovered from the crime scene and the test specimen usually refers to the rolled, inked fingerprint of a suspected individual.

[5] The LR does not imply a decision, but the accuracy measurement is inserted in a decision-theoretical process as explained in [15,16]. The accuracy of the LR is defined as a measure of how close one gets to the true proposition (also dubbed as goodness of the LR) rather than how close one gets to the "true value of the LR".

[6] The term calibration is throughout this article understood as a property of a set of LRs and not as the activity aimed at improving the LR.

times more probable under H_p than under H_d (in other words, the LR of the LR is the LR [18], [19]). Under those conditions the LR is exactly as big or small as is warranted by the data. Well-calibrated LRs tend to yield stronger support with better discrimination of a given method [15].

2.2 Example factors influencing the primary performance characteristics

Quality[7] of the data is a measurable parameter that has no information about the proposition, but impacts the performance of that comparison. In other words, specimens of high quality to be compared in a forensic case lead to better performance, while comparisons with low quality samples lead to worse performance of a LR method. For example a quality of the ridge flow in a fingermark / fingerprint image.

Quantity or amount of data, e.g., the length of a speech fragment, the number of minutiae in a fingermark, etc.

Representativeness of the data used to train the LR method with respect to the data used in operational conditions. The smaller the dataset shift [20] between the two, the more representative the training data is for those in operational conditions.[8]

2.3 Secondary performance characteristics

Coherence is defined as the ability of the method to yield LRs with better performance with an increase of the quantity and/or quality of the information present in the data.

Generalization is defined as the property of a given method to maintain its performance under dataset shift. LR method 1 generalizes better than LR method 2 if, under similar conditions of dataset shift in both methods, the performance of method 1 decreases less than the performance of method 2.

Robustness is the ability of the method to maintain performance when the quantity or quality of the data decreases. For instance, method 1 is more robust to data sparsity than method 2 if, with decreasing amount of data, the performance of method 1 decreases less than the performance of method 2.

---

7 Quality is not an intrinsic property, but influences the ability of a system to extract features from the specimens, and to compare and evaluate this information.

8 A simple definition of dataset shift can be the amount of the difference in descriptive statistics of two datasets. For instance, the bigger the difference in means in the position of two minutiae datasets, the bigger their dataset shift. The dataset shift can be also measured with metrics of distance between their probability distributions, such as the Kullback–Leibler divergence used below.

In the next section we present an experimental example to illustrate the measurement of coherence, discuss the datasets used in the LR method development and the performance metrics used to establish the coherence of LRs produced by the method.

3 Measuring coherence: experimental example with LRs inferred from fingermarks

The comparison of the minutiae of a fingermark and fingerprint using an AFIS comparison algorithm results in a comparison score. The strength of evidence of this score can be assessed in terms of a LR. Since the LR method in our case consists of modelling the SS and DS score distributions, it is referred to as a LR model from here on. A detailed description of the LR model used – derived from [6] – is beyond the scope of this article, since the aim is to present the validation methodology with the focus on the analysis of coherence.

Recall the set of propositions from the Section 2.1. Without loss of generality we can rephrase them to fit our fingerprint example:

H_p: The fingermark and fingerprint come from the same source (SS)
H_d: The fingermark and fingerprint come from different sources (DS)

Having defined the set of propositions with respect to which the comparison scores are evaluated, we proceed to build the LR model [6]:

- Use the minutiae comparison algorithm to compare the fingermarks of a suspect with the fingerprint of a suspect to produce a same source score distribution (SS).
- Use the minutiae comparison algorithm to compare the crime scene fingermark to the fingerprint of a suspect to produce the evidence score (E).
- Use the minutiae comparison algorithm to compare the crime scene fingermark to a database of fingerprints of individuals other than the suspect to produce a different source score distribution (DS).
- Model the SS and DS score distributions using probability density functions or a discriminative approach, e.g., using logistic regression [17].
- Compute the strength of the evidence given by the likelihood ratio[9]:

$$\mathrm{LR} = \frac{p\left(E \mid H_p\right)}{p\left(E \mid H_d\right)}. \qquad (1)$$

[9] While there are several different options to obtain the denominator of the LR, the LR in this work is obtained in a similar way as described in [4] using option 2: “the item of unknown source with randomly generated items from a relevant database”.

The comparison algorithm applied in this work to generate scores is a commercial product Motorola BIS 9.1, used as a black-box. The minutiae extraction and comparison technology remains outside the scope of this work, but we still present some of its functionality. The algorithm used is speed-optimised and outputs comparison scores in three separate score ranges. The comparison algorithm considers two different comparison methods depending on the number of minutiae in the mark: one for 5–10 minutiae configurations and another one for configurations of 11 and more minutiae. Within each of these two methods, the maximum score is directly proportional to the number of features in agreement. We get back to the two methods of the comparison algorithm in Section 7.

4 Datasets used

We have two datasets at our disposal. First we use a training dataset consisting of pseudo fingermarks to obtain the values of the parameters of the model. Second we use a relatively small dataset consisting of forensic fingermarks to determine validity of the LR model for forensic casework. In the following sections we present the two datasets used in more detail. We justify the dataset shift both numerically using the Kullback–Leiber (KL) divergence, a measure commonly used in probability and information theory [21], and visually by comparing the histograms of selected score distributions.

4.1 Forensic dataset

The forensic dataset consists of data from real forensic cases: 58 identified fingermarks in 12-minutiae configuration and their corresponding fingerprints. The ground-truth labels of the dataset, indicating whether a fingermark/print pair originates from the same source is denoted as "ground-truth by proxy" because of the nature of the pairing between fingermarks and fingerprints: they have been assigned after examination by human examiners, taking into account not only the 12 minutiae, but also other minutiae, ridge pattern, etc. The minutiae feature vectors[10][10] of the fingermarks have been manually extracted by examiners while the minutiae feature vectors of the fingerprints have been automatically extracted using a feature extraction algorithm and manually checked by examiners.

In order to obtain multiple minutiae configurations for the validation of the LR method, the minutiae extracted from the fingermarks have been clustered into configurations of 5–12 minutiae, according to the method described in [22]. Following the clustering procedure we obtain 481 minutiae clusters in a 5-minutiae configuration from the 58

[10] Minutiae feature vectors of a fingermark or fingerprint in our case consist of feature type, position, and orientation (parallel to the ridge flow).

fingermarks with 12 minutiae. For each cluster in the marks, a same-source (SS) score is obtained by comparing each minutiae cluster from a fingermark with the corresponding reference print. Similarly, a different-source (DS) score distribution is obtained by comparing a fingermark to a subset of a police fingerprint database. This dataset consists of roughly eight million prints. The higher the number of minutiae in each cluster, the lower the number of clusters, as can be seen in Table 1. An example of a forensic fingermark is presented in Fig. 1.

4.2 Pseudo-fingermarks dataset

Pseudo-fingermarks were obtained by capturing a video sequence of a finger of a known individual moving on a glass plate in different directions in order to capture as much distortion as possible. Reference print(s) of the same finger of the same individual were recorded on a 10-print card. This dataset consists of 200 individuals (100 male and 100 female) times 10 video sequences (1 per finger). The process of obtaining the pseudo marks dataset is described in detail in [22].

The dataset of pseudo fingermarks consists of 25,000 fingermarks of known origin, from which we produce the SS and DS score distributions. In this dataset the pseudo fingermarks were clustered into different minutiae configurations according to the procedure described in [22] and different numbers of fingermarks per number of minutiae were obtained as shown in Table 2. An example of a pseudo-fingermark on a forensic background is presented in Fig. 1.

The main advantage of using the pseudo-fingermarks over the real forensic ones is that it is relatively easy and cost-efficient to scale up the experiment and produce these in high quantities.

5 Measuring dataset shift between the datasets of pseudo and forensic fingermarks

Since the two datasets (forensic and pseudo-fingermarks) were acquired under different conditions, it is appropriate to establish the degree of similarity between the distributions of the scores generated by them. We use the KL (Kullback–Leiber) divergence to quantitatively express the shift between the DS score distributions of the two datasets.[11] We convert the

[11] The feature comparison algorithm outputs the similarity scores in three different mutually exclusive regions see reference [23] – page 77. The KL divergence was computed on a "per-region" basis. It is statistically not robust to talk about score distributions with very few occurrences of SS scores in particular regions; therefore the KL divergence was shown for the DS distribution only where the density of the scores is much higher.

score distributions into normalized histograms representing relative frequencies of observations of comparison scores in each of the two datasets – forensic (*F*) and pseudo (*S*) – and compute the KL divergence as follows:

$$\mathrm{KL} = \sum_i F(i) \cdot \ln\left(\frac{F(i)}{S(i)}\right). \tag{2}$$

where the index *i* in Eq. (2) refers to the *i*-th bin in the histogram. Note that if the two distributions *F* and *S* are identical the KL divergence is equal to zero, and the more similar the histograms are the smaller is the divergence and the smaller the dataset shift.

Since the KL divergence is a non-commutative distance between the two distributions *F* and *S*, we propose to calculate the distance between *F* and *S* and *S* and *F*. The final, symmetric KL divergence is represented as the average of those two distances:

$$\mathrm{KL}_{\mathrm{sym}} = \frac{\sum_i F(i) \cdot \ln\left(\frac{F(i)}{S(i)}\right) + \sum_i S(i) \cdot \ln\left(\frac{S(i)}{F(i)}\right)}{2}, \tag{3}$$

where index *i*, as in Eq. (2) refers to *i*-th bin in the histogram.

The KL divergence of the two datasets, calculated using Eq. (3), is presented in Table 3. Recall from Eq. (2) that the more similar the two score distributions is, the closer to zero is the resulting $\mathrm{KL}_{\mathrm{sym}}$. The highest degree of similarity between the pseudo and the forensic dataset is found for the fingermarks clustered in 6-minutiae configuration, while the lowest degree of similarity is found for the fingermarks in 5-minutiae configuration.

For better understanding the KL divergence, the similarity of the two score distributions can also be visually assessed in Fig. 2, Fig. 3. We compare the normalized histograms of the scores for the pseudo and the forensic datasets, presenting as an example the results for the 5-minutiae configurations (lowest degree of similarity $\mathrm{KL}_{\mathrm{sym}} = 0.033$) and the 6-minutiae configurations (highest degree of similarity $\mathrm{KL}_{\mathrm{sym}} = 0.007$). The difference between these most similar and least similar score distributions appears negligible in Fig. 2 and Fig. 3.

Establishing a degree of similarity between the two datasets acquired under different conditions is a very important step in LR method development, especially when using probability density functions to produce LRs. We conclude that the dataset of pseudo fingermarks is a representative approximation of the forensic dataset.

6 Performance measures used

In this part we introduce a set of plots and performance measures used to evaluate the performance of the model for different minutiae configurations. Although alternative measures can be used to illustrate the coherence of the LR method, we think that visual representations and measures proposed are sufficient.

6.1 Detection error trade-off (DET) plot and equal error rate (EER)

The DET plot [24] presents the false acceptance rate (FAR) as a function of the false rejection rate (FRR). The error rates are plotted on a Gaussian-warped scale. This makes the DET curves linear when the log(LR) values are normally distributed. The closer the curve is to the origin, the better the discrimination of the method. The intersection of a DET curve with the diagonal of the DET plot marks the Equal Error Rate (EER). The EER is used as a performance measure to show the coherent behaviour of the LR method. For example, when comparing forensic fingermarks in different minutiae configurations the EER should be larger for configurations with fewer minutiae (see Fig. 4). Even if a DET plot is meant to characterize a system that makes decisions, it is informative about the coherence of the LR method when evaluating datasets with different quantities of information, since the discriminating power is an important contributor of LR performance [15,16].

6.2 Tippett plots

Tippett plots [25] are representations of cumulative distributions of LRs. The curves in it represent the proportion of comparisons resulting in a log(LR) greater than t versus that value t, when either proposition H_p or H_d is true. In a Tippett plot, the rates of misleading evidence for either proposition can be observed at the intersection of each of the curves supporting either the prosecution or the defence proposition and the vertical line at $t = 0$. The log(LR) value zero corresponds to a LR value of 1. A LR is misleading if it is greater than one though H_d is actually true, or less than one though H_p is actually true. Using Tippett plots it is relatively easy to distinguish the performance of an LR method when presented with different quantities of evidential information.

Examples of Tippett plots are shown in Fig. 5 for the 5 and 10-minutiae configurations. The decrease in misleading evidence due to the five additional minutiae can clearly be seen.

6.3 Empirical cross-entropy (ECE) plot and the log likelihood ratio cost (C_{llr})

The Empirical Cross-Entropy or ECE plot [15,16] is a representation of the performance and calibration of the LR values and complements other already established methods such as those discussed above [16]. The C_{llr} is a closely related cost function of the log(LR) defined in reference [17]. ECE and C_{llr} are both lower when the likelihood ratio better supports the ground-truth proposition. The difference between them lies in the interpretation of both measures. The C_{llr} is interpreted as an average decision cost for all prior probabilities in a given set of LR values. On the other hand, the ECE has an information-theoretical interpretation as the amount of information lacking compared to full knowledge of the ground-truth, on average in a given set of LR values. The C_{llr} is an average over costs and priors, and therefore is not giving the performance for a given value of the prior, but for an average of all possible priors. An ECE-plot shows the ECE for a certain range of priors [15], [16]. It can be easily shown that the C_{llr} is the ECE at prior log-odds of 0 (i.e. a prior probability of 0.5). In this sense, the ECE is a more general and interpretable performance metric than the C_{llr} in a forensic context, where no decision is to be made by the forensic examiner and where the value of the prior changes very much from one case to another. It also appears to be more suitable to show the validity of a method over a relevant set of priors that are generally unknown. On the other hand, the C_{llr} is a summary of the ECE in a single number, useful for comparing and ranking methods.

We use the C_{llr} as a measure of accuracy, consisting of two components: discrimination C_{llr}^{min} and calibration C_{llr}^{cal} [17]. In an ECE plot the posterior probabilities are computed for each prior in a range of <0,1> using the LR values. The resulting value of the ECE is then represented as a function of the prior probability. The solid curve in the ECE plot also represents accuracy: the lower it is, the better the accuracy of the method. The dashed curve represents the discrimination, and is sometimes referred to as “accuracy after PAV”, because it is the ECE after applying the Pool Adjacent Violators algorithm (PAV). It is an algorithm that improves the calibration of a set of LRs while not affecting their discrimination; see [17] for details. The difference between these two curves represents calibration losses: the smaller the distance, the better the LR method's calibration.

Besides the information-theoretical aspect, the ECE provides the “range of application” of the LR method under evaluation. A LR method should perform better than a reference method producing LR = 1 for the whole range of prior probabilities. In a range of prior probabilities where this is not the case, using the LR method would be worse than not using any method at all.

Fig. 6 presents an example for the sake of illustration, showing the ECE plots of the LR method evaluating the fingermarks in 5-minutiae configuration in two different settings: LRs with lesser calibration and LRs with better calibration (following PAV calibration). Calibrating the LR method not only improves the accuracy of the LR method (here measured by the C_{llr} and ECE), it also extends the applicable range of this method. The LR method with lesser calibration presents an ECE larger than that of the reference method for prior log-odds above 0.5, which does not happen for the well-calibrated LR method.

7 Results

We use the same LR method to produce LR values for 5–12-minutiae configuration comparisons. For each forensic *n*-minutiae configuration dataset the LR method is trained with the corresponding *n*-minutiae pseudo fingermark dataset.

In order to establish the coherence of the LRs, we measure the primary performance characteristics: accuracy (using C_{llr} and ECE as a measure), discrimination (using $C_{\text{llr}}^{\text{min}}$ and ECE-after-PAV as a measure) and calibration (using $C_{\text{llr}}^{\text{cal}}$ and the difference between ECE and ECE-after-PAV as a measure). Recall that the coherence is not a primary but a secondary performance measure: it describes the variation of the performance of the LR method when varying quality or quantity of the information (in our case the number of minutiae).

The performance as a function of the number of minutiae is presented using ECE, Tippett and DET plots. The C_{llr}, $C_{\text{llr}}^{\text{min}}$, and EER are determined for all minutiae configurations and presented in Table 4.

The ECE plots in Fig. 7 show a decreasing trend (solid curves), which corresponds to increased accuracy and discrimination (dashed curves) when increasing the number of minutiae from 5 to 10. The values for the accuracy and discrimination show the same trend and are summarized in Table 4. The sudden increase of these plots and values for the 11-minutiae configurations are related to the comparison algorithm, which changes its method from 11 minutiae onwards.

The Tippett plots in Fig. 8 also show coherence of the method with the increasing distance between the curves based on LRs supporting either proposition as the number of minutiae increases. In an ideal system the rates of misleading evidence would be equal to zero, and both curves in the Tippett plots would present a maximal vertical separation. The coherence is shown in the Tippett plots if with an increasing number of minutiae a decrease in the rates of misleading evidence and an increase in the vertical separation of the curves is observed. The rate of misleading evidence in favour of H_{d} (RMED [25], [26]) decreases from 31% for 5-minutiae configurations to 3.5% for 12-minutiae configurations, while the rate of

misleading evidence in favour of H_p (RMEP [25], [26]) decreases from 1.2% for 5-minutiae configurations to 0.06% for 12-minutiae configurations. In the Tippett plots the largest separation is shown for the 10-minutiae configuration and a performance decrease in separation is observed for 11 and 12-minutiae configurations.

The DET curves in Fig. 9 capture the discrimination in a lot more detail, complementing the Tippett and ECE plots. Coherent behaviour of the LR method used can be observed in the decreasing values of the EER for an increasing number of minutiae. The highest performance in terms of EER was achieved for the 9-minutiae configuration dataset (EER = 1.6%). The lowest performance of the LR method was observed for the 5-minutiae configuration dataset (EER = 15.7%). Table 4 lists the EER values and apart from the overall decreasing trend shows increases for 10 and 11 minutiae. Not too much meaning can be attached to this because of the overlap and irregular behaviour of the DET curves for the highest number of minutiae. This is mainly because the EER is difficult to compute when the dataset sparsity increases.

8 Discussion and conclusions

The purpose of this article is to introduce coherence as a secondary performance characteristic for LR methods developed for forensic evaluation, and to demonstrate its use with an experimental example. In Section 2 we have split various performance characteristics into primary and secondary ones with examples of factors influencing the primary performance characteristics. We then focused on one performance characteristic in particular – the coherence – by giving an experimental example from the area of forensic fingerprint examination. Coherence has been defined as the property of a given method to perform better when the quality or quantity of information increases, which in our experimental example has been simulated by varying the number of minutiae present in fingermarks from 5 to 12.

The performance of the LR method was evaluated using different performance measures (Rates of Misleading Evidence, C_{llr} and EER) and their corresponding graphical representations: Tippett, ECE, and DET plots. The LR method used showed coherent behaviour: performance increased with the number of minutiae increasing from 5 to 10. It also showed somewhat incoherent behaviour and a small decrease in performance when moving from 10 to 11 minutiae.

This incoherent behaviour of the comparison algorithm's performance is believed to be caused by a switch of the method it uses when more than 10 minutiae are mainly due to the core comparison algorithm used by the AFIS technology. The experimental example

therefore reveals the importance of coherence in order to detect points of improvement in computer-assisted LR methods.

Acknowledgements

This research was conducted in the scope of the BBfor2–European Commission Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803) at the Netherlands Forensic Institute, and in collaboration with the ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid and the National Police Services Agency of the Netherlands.

References

[1] I.W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Sci. Justice, 38 (1998), pp. 198-202.

[2] D.V. Lindley, A problem in forensic science, Biometrika, 64 (1977), pp. 207-213.

[3] C. Neumann, I. Evett, Quantitative assessment of evidential weight for a fingerprint comparison I. Generalisation to the comparison of a mark with set of ten prints from a suspect, Forensic Sci. Int., 207 (2011), pp. 101-105.

[4] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int., 219 (2012), pp. 129-140.

[5] D. Meuwly, Forensic individualization from biometric data, Sci. Justice, 46 (2006), pp. 205-213.

[6] N. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems–modelling within finger variability, Forensic Sci. Int., 167 (2007), pp. 189-195.

[7] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, J. Ortega-Garcia, Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems, Forensic Sci. Int., 155 (2005), pp. 126-140.

[8] G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, John Wiley and Sons (2014).

[9] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process., 15 (2007), pp. 2104-2115.

[10] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice, 51 (2011), pp. 91-98.

[11] European Network of Forensic Science Institutes, Annual Report 2012: http://enfsi.eu/wp-content/uploads/2016/09/enfsi_report2012.pdf

[12] ILAC-G19: 2002, Guidelines for forensic science laboratories.

[13] RvA-T015, 2010, Explanation of NRN-RN ISO/IEC 17025:2005.

[14] ISO/IEC 17025: 2005, General requirements for the competence of testing and calibration laboratories.

[15] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int., 230 (2013), pp. 156-169.

[16] D. Ramos, J. Gonzales-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, J. Forensic Sci., 58 (2013), pp. 1503-1518.

[17] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang., 20 (2006), pp. 230-275.

[18] I.J. Good, Weight of evidence: a brief survey, Bayesian Stat., 2 (1985), pp. 249-270.

[19] D.A. van Leeuwen, N. Brümmer, The distribution of calibrated likelihood-ratios in speaker recognition, arXiv:1304.1199, Interspeech 2013.

[20] J. Quiñonero-Candela, Dataset Shift in Machine Learning, The MIT Press, Cambridge, Massachusetts (2009).

[21] T. Cover, J. Thomas, Elements of Information Theory, (second ed.), Wiley & Sons, Hoboken, New Jersey (2006).

[22] C.M. Rodriguez, A. de Jongh, D. Meuwly, Introducing a semi-automated method to simulate a large number of forensic fingermarks for research on fingerprint identification, J. Forensic Sci., 57 (2012), pp. 334-342.

[23] R. Haraksim, Validation of likelihood ratio methods used for forensic evidence evaluation: application in forensic fingerprints, (PhD thesis, 2014).

[24] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, Proc. EuroSpeech (1997), pp. 1895-1898.

[25] D. Meuwly, Reconnaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approche Automatique, (PhD thesis, 2001).

[26] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, D.Meuwly, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of three minutiae, J. Forensic Sci., 51 (2006), pp. 1255-1266.

Tables

Minutiae	Number of SS comparisons	Number of DS comparisons
5	481	10,283,780
6	432	9,236,160
7	426	9,107,880
8	387	8,274,060
9	342	7,311,960
10	286	6,114,680
11	190	4,062,200
12	58	1,240,040

Table 1 Forensic dataset sizes, for SS and DS scores. Note that the number of SS scores is the same as the number of clusters for a given number of minutiae.

Minutiae	Number of SS comparisons	Number of DS comparisons
5	16,653	33,306,000
6	25,058	50,116,000
7	24,876	49,752,000
8	25,015	50,030,000
9	25,036	50,072,000
10	24,994	49,988,000
11	24,658	49,316,000
12	24,443	48,886,000

Table 2 Pseudo fingermarks dataset sizes for SS and DS scores.

Minutiae	**KL_{sym}**
5	0.034
6	0.007
7	0.011
8	0.019
9	0.013
10	0.010
11	0.014
12	0.011

Table 3 KL_{sym} divergence of the DS comparison scores (simulated and forensic dataset).

Minutiae	**Accuracy C_{llr}**	**Discrimination C_{llr}^{min}**	**DET-EER (%)**	**RMEP (%)**	**RMED (%)**
5	0.50	0.43	15.69	31.39	1.18
6	0.28	0.26	6.91	19.68	0.89
7	0.16	0.14	3.95	11.74	0.69
8	0.13	0.11	2.42	7.75	0.68
9	0.075	0.063	1.56	3.80	0.63
10	0.074	0.063	2.19	3.86	0.48
11	0.100	0.081	2.73	5.26	0.19
12	0.084	0.057	1.82	3.45	0.06

Table 4 Performance of the LR method for different number of minutiae.

Figures

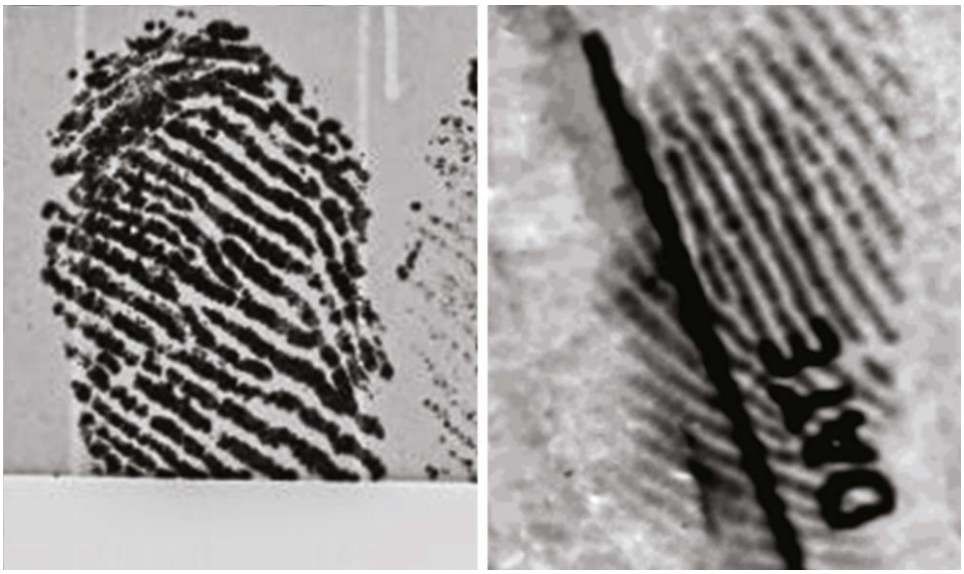

Figure 1 Forensic (left) vs. pseudo (right) fingermark.

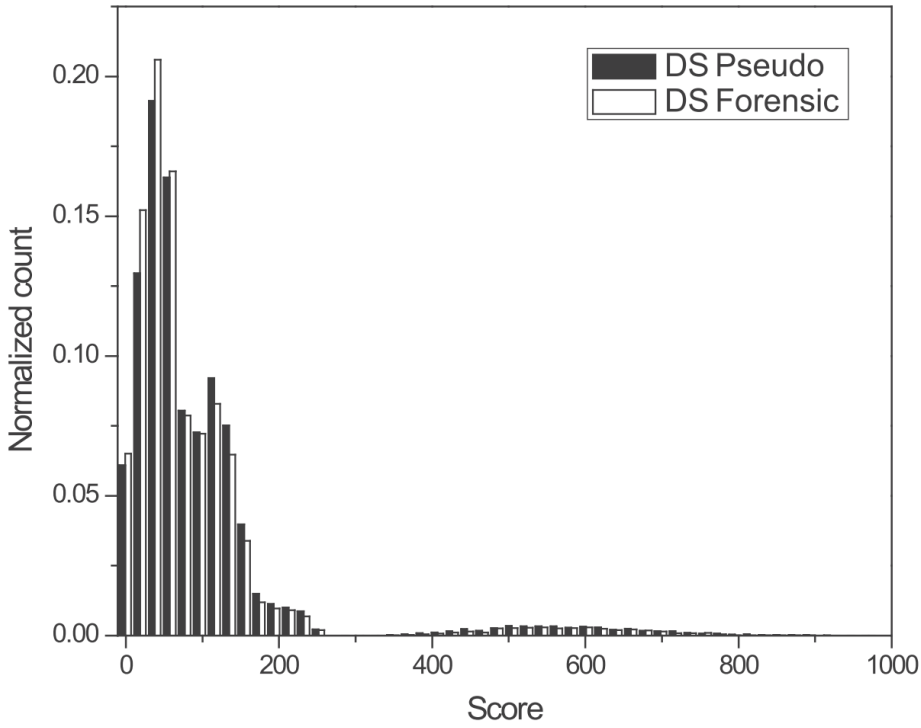

Figure 2 Normalized score distribution for 5-minutiae configurations of forensic versus pseudo fingermark datasets, showing low degree of similarity according the KL_{sym} measure.

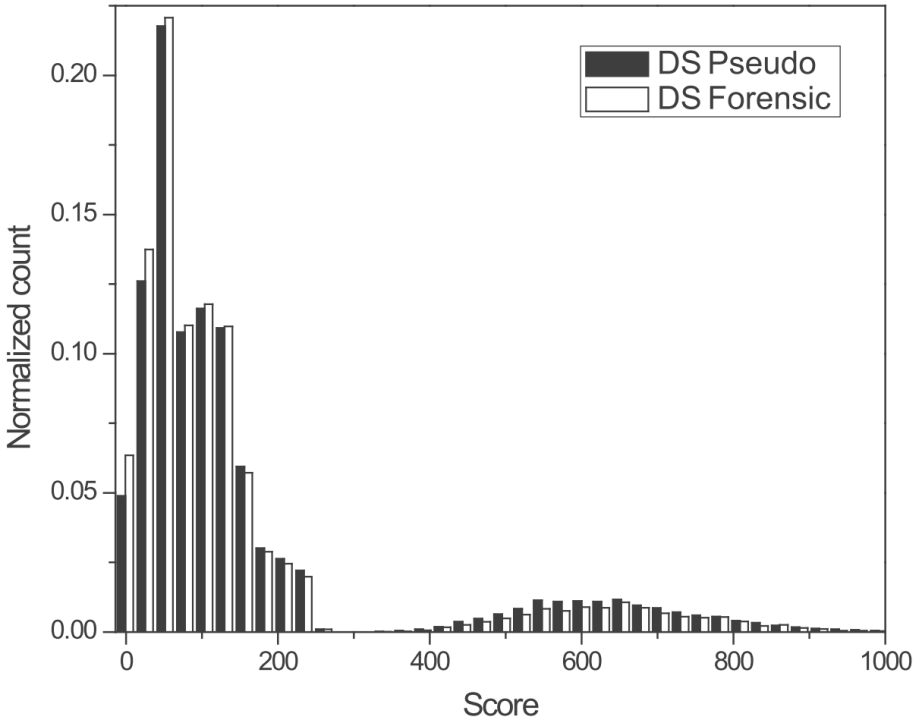

Figure 3 Normalized score distribution for 6-minutiae configurations of forensic versus pseudo fingermarks datasets, showing high degree of similarity according on the KL_{sym} measure.

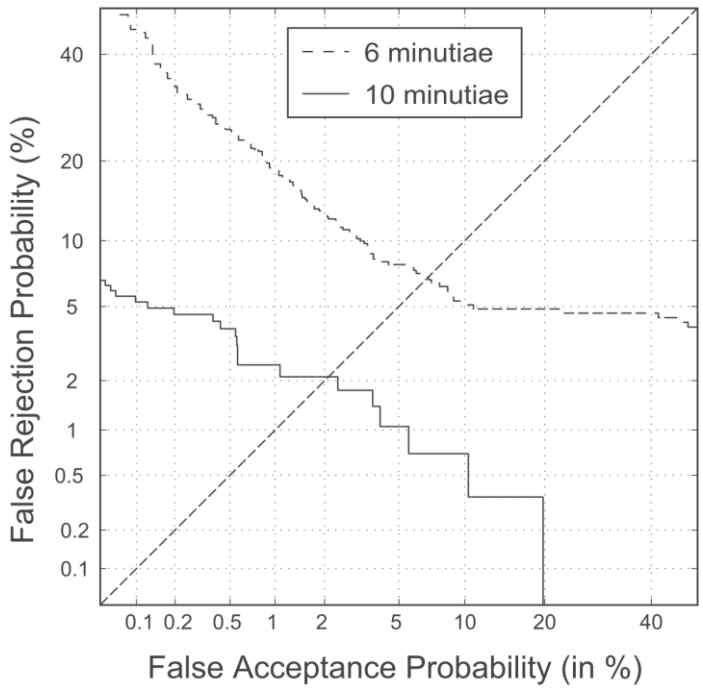

Figure 4 DET curves showing the performance of the same LR method with different quantities of information. The dashed curve shows worse discrimination in the LRs of comparisons for 6-minutiae configurations, while the solid line shows better discrimination in the LRs of comparisons for 10-minutiae configurations. The equal error rates are given by the intersection of the curves with the diagonal of the plot, and are 6.9% and 2.2%, respectively.

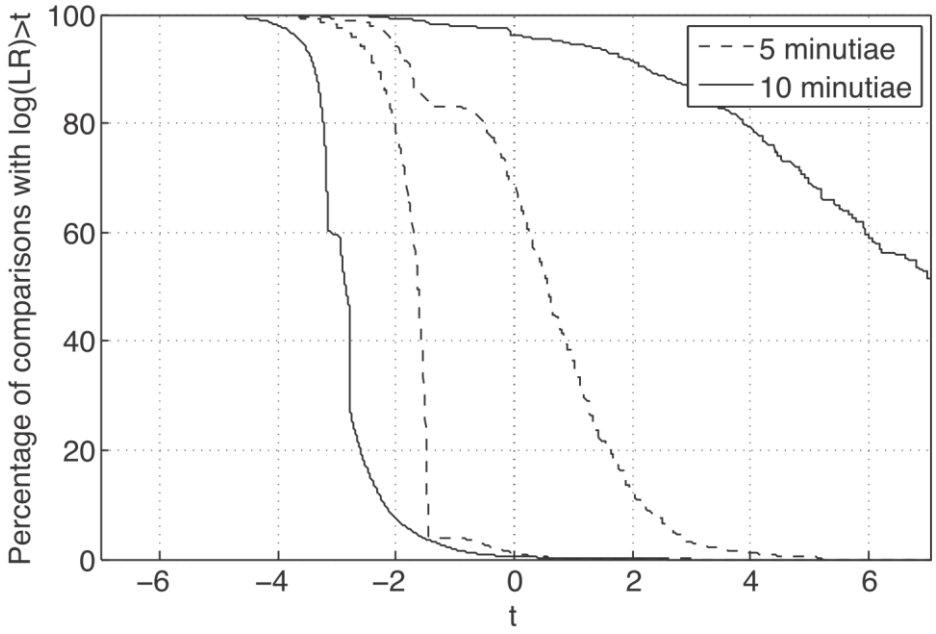

Figure 5 Tippett plots (before calibration) showing the performance of the same LR method with different quantities of information. For the 10 minutiae the solid curves are furthest vertically separated – they capture more evidential information and present better discrimination.

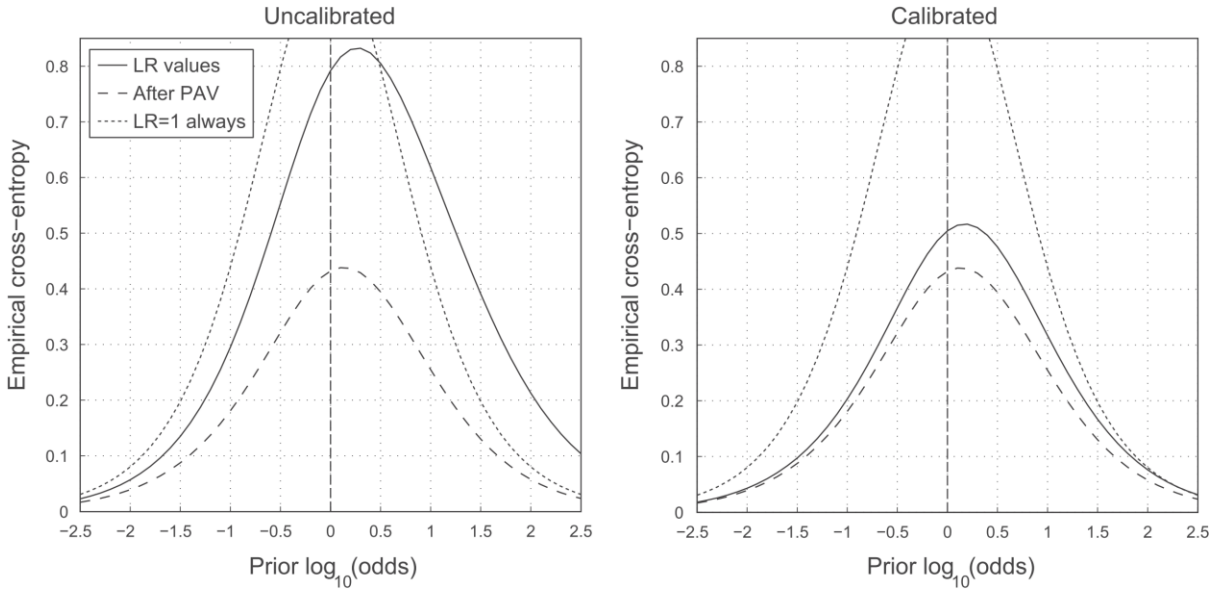

Figure 6 ECE plots for the same LR method (same set of LR values) showing different calibration performance. The left-hand side plot represents LR values with worse calibration than the right-hand side plot. The lack of calibration is visible in the left plot by the fact that above prior-log(odds) = 0.5 the ECE exceeds that of the reference method which always gives LR = 1. For that range of prior odds the worse calibrated method performs worse than a method that always returns the "I don't know" answer (i.e., always yielding LR = 1).

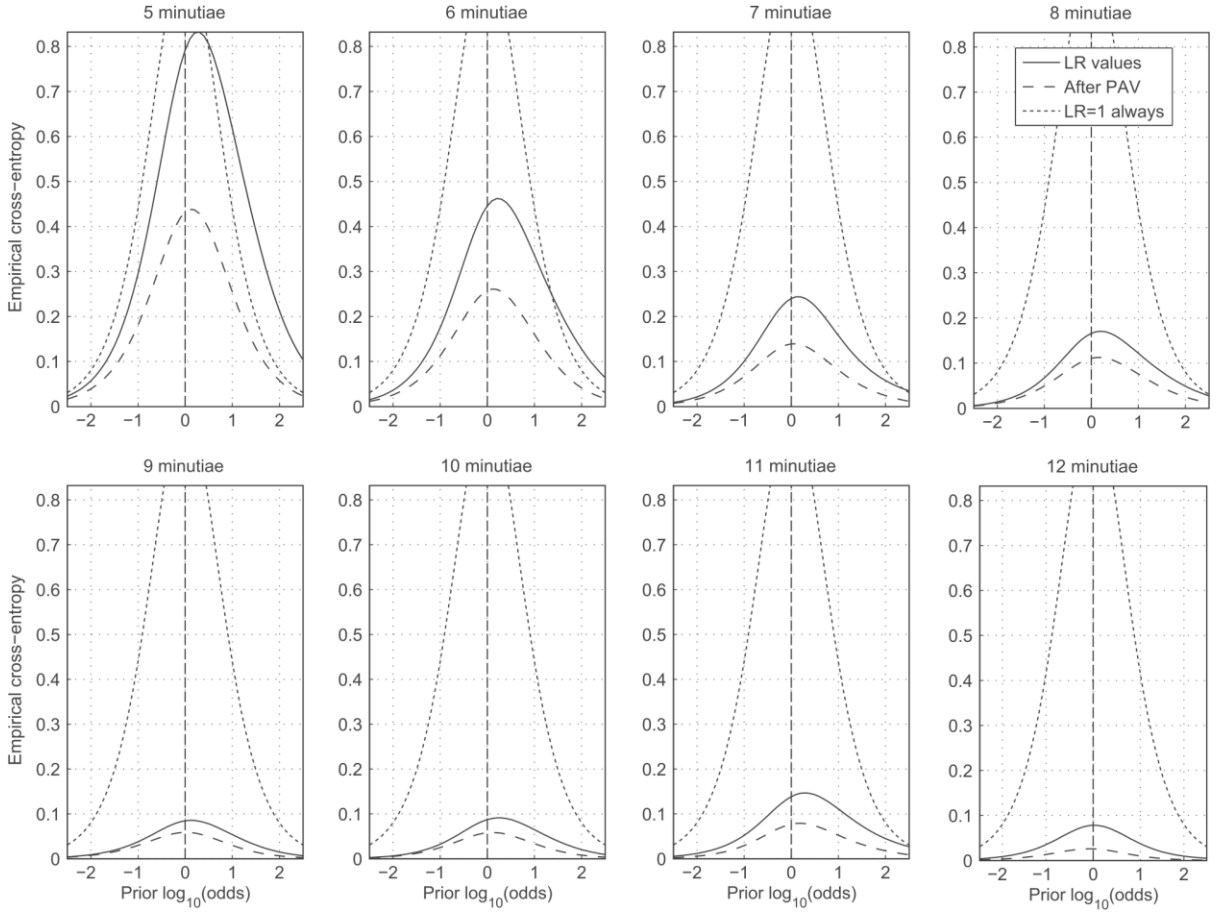

Figure 7 ECE plots for LRs generated for forensic marks with 5–12-minutiae configurations. Note the different scaling of the *y*-axis in the upper and lower row of plots.

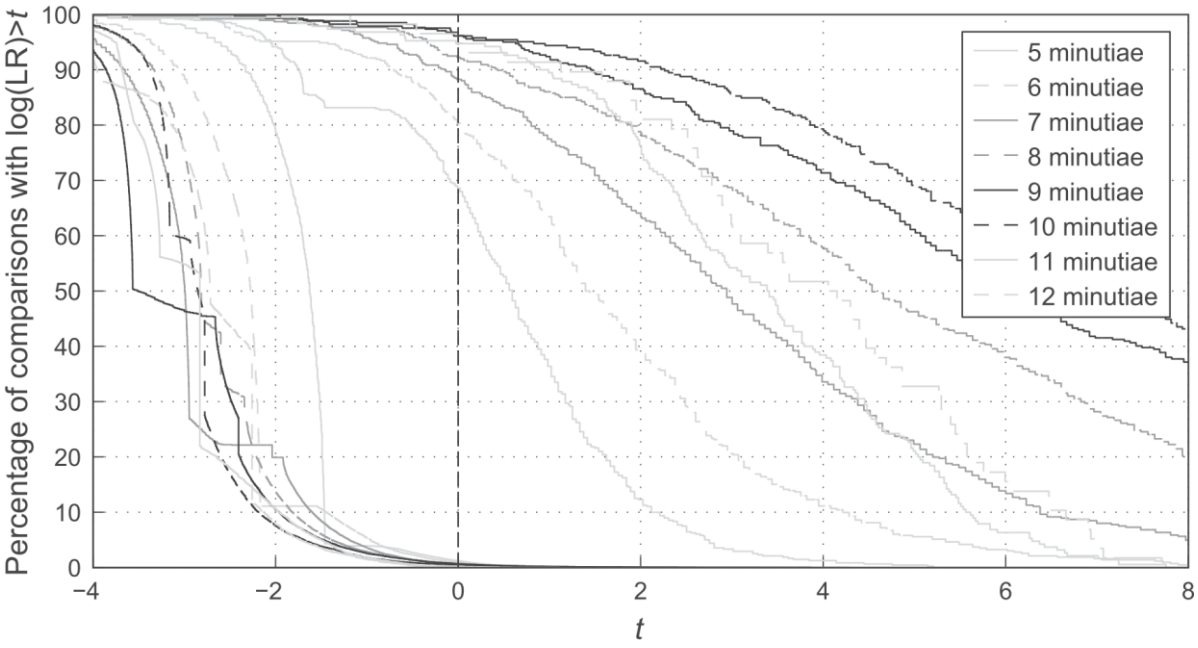

Figure 8 Tippett plots for LRs generated for forensic marks with 5–12-minutiae configurations.

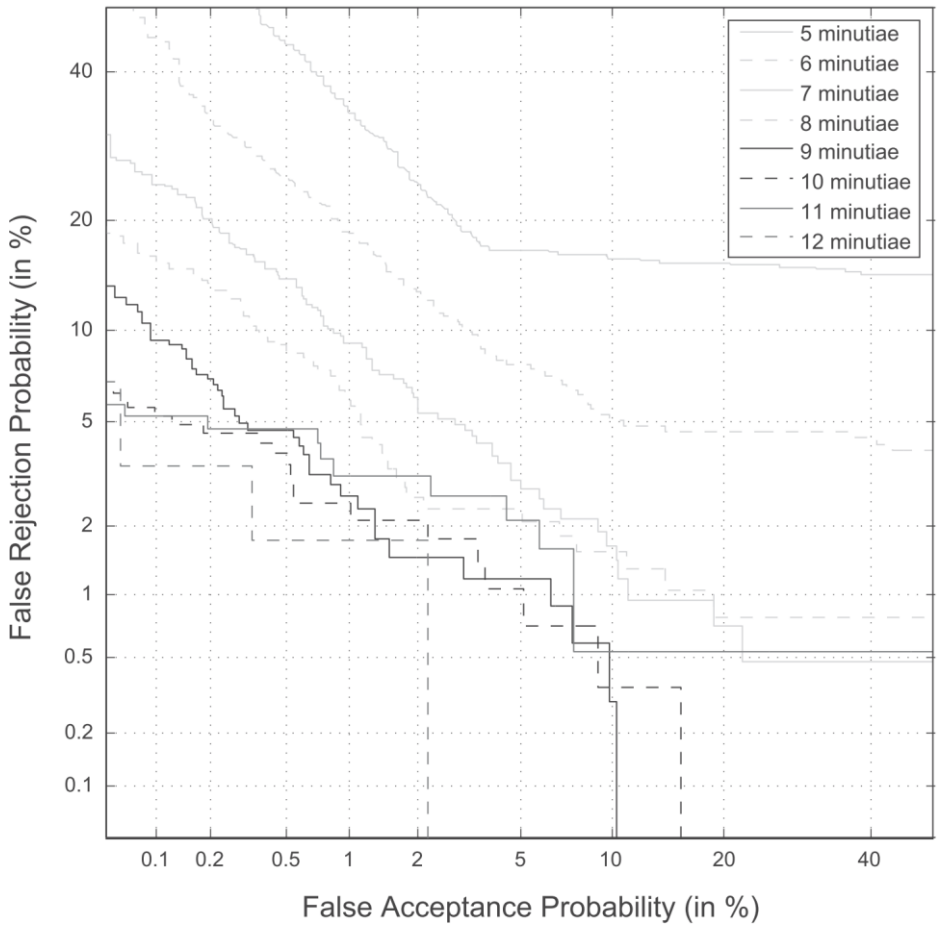

Figure 9 DET plots for LRs generated for forensic marks with 5–12-minutiae configurations.