Forensic Science International 249 (2015) 123–132

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Measuring coherence of computer-assisted likelihood ratio methods

Rudolf Haraksim [a,*], Daniel Ramos [b], Didier Meuwly [a], Charles E.H. Berger [a]

[a] Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague, Netherlands
[b] ATVS – Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, C/Francisco Tomas y Valiente 11, 28049 Madrid, Spain

ARTICLE INFO

Article history:
Received 28 April 2014
Received in revised form 3 September 2014
Accepted 27 January 2015
Available online 4 February 2015

Keywords:
Coherence
Forensic evidence
Fingermark
Fingerprint
Likelihood ratio
Validation

ABSTRACT

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.

© 2015 Elsevier Ireland Ltd. All rights reserved.

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

[*] Corresponding author. Tel.: +41 787110899.
E-mail address: haraksim@gmail.com (R. Haraksim).

[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.

http://dx.doi.org/10.1016/j.forsciint.2015.01.033
0379-0738/© 2015 Elsevier Ireland Ltd. All rights reserved.

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, 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 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.

[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.

[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.

Download English Version:

https://daneshyari.com/en/article/95415

Download Persian Version:

https://daneshyari.com/article/95415

Daneshyari.com