IET Biometrics

Research Article

IET Journals
The Institution of Engineering and Technology

Validation of likelihood ratio methods for forensic evidence evaluation handling multimodal score distributions

ISSN 2047-4938
Received on 4th August 2015
Revised 8th June 2016
Accepted on 20th September 2016
E-First on 9th November 2016
doi: 10.1049/iet-bmt.2015.0059
www.ietdl.org

Rudolf Haraksim[1] ✉, Daniel Ramos[2], Didier Meuwly[3,4]

[1]Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland
[2]ATVS – Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, C/Francisco Tomas y Valiente 11. 28049 Madrid, Spain
[3]Netherlands Forensic Institute, P.O. Box 24044, 2490AA The Hague, The Netherlands
[4]Faculty of EEMCS, University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands
✉ *E-mail: haraksim@gmail.com*

Abstract: This study presents a method for computing likelihood ratios (LRs) from multimodal score distributions, as the ones produced by some commercial off-the-shelf automated fingerprint identification systems (AFISs). The AFIS algorithms used to compare fingermarks and fingerprints were primarily developed for forensic investigation rather than for forensic evaluation purposes. Thus, in some of those algorithms, the computation of discriminating scores is speed-optimised. In the case of the AFIS algorithm used in this work, the speed-optimisation is achieved by performing the comparison in three different stages, each of which outputs scores of different magnitudes. As a consequence, all scores together present a multimodal distribution, even though each fingermark-to-fingerprint comparison generates one single score. This multimodal distribution of scores might be typical for other biometric systems or other algorithms, and the method proposed in this work can be also applied to those cases. As a result, the authors propose a probabilistic model for LR computation that presents more robustness to overfitting and data sparsity than other traditional approaches, like the use of models based on kernel density functions.

1 Introduction

Score-based biometric systems have been recently proposed as a source of information for evidence evaluation in forensic cases [1, 2]. The aim is to use the automated fingerprint identification system (AFIS) technology primarily to exploit the distinctiveness of the fingermarks, and subsequently to evaluate the strength of evidence in the form of score-based likelihood ratios (LRs). This approach has been used for example in [1, 2] and was proposed as an alternative and complementary way of extracting information from fingermark-to-fingerprint comparisons in forensic cases, aiding the forensic examiners conducting their analysis. It also makes the AFIS scores interpretable in a Bayesian probabilistic framework. Typical commercial ‘off-the-shelf’ (COTS) AFIS algorithms producing scores are primarily developed to support the process of selection of candidates for forensic investigation and not intended for the use in forensic evidence evaluation [3]. In any case, the ultimate aim is to discriminate the comparisons where the fingermark and the fingerprint come from the same source or from different sources, and therefore the information extracted by these systems can be used to evaluate the strength of evidence by means of LRs. In this work, we use an AFIS algorithm, which was optimised to perform a large number of comparisons in the shortest time possible (speed-optimisation). Due to this optimisation, each fingermark-to-fingerprint comparison is performed in one of three different output stages, depending on the quantity of information in accordance, which results in scores of different magnitudes, together forming a multimodal score distribution. Thus, the scores generated by the AFIS algorithm are structured in three regions (R_1, R_2 and R_3). The rationale of this comparison process is sketched as follows:

- Region 1 (R_1). This is the first stage of the comparison. Here, a quick comparison is performed. If the AFIS algorithm finds very few minutiae in agreement between the mark and the print, a score of −1 is produced, and the process ends. If this is not the case, the comparison proceeds to the second stage.
- Region 2 (R_2). This is the second stage of the comparison. All the comparisons performed in this stage are assigned score values in the range between 0 and 300. This value of 300 is not chosen or tuned, but fixed by the system. A computationally light comparison is performed. If low similarity is observed between the mark and the print, the score is produced and the comparison ends. If this is not the case, the comparison proceeds to the third stage.
- Region 3 (R_3). This is the third stage of the comparison. All the comparisons performed in this stage are assigned score values greater than 300. Again, this value of 300 is not chosen or tuned, but fixed by the system. At this stage a full comparison with higher computational burden is performed to produce the score.

As described above, each fingermark-to-fingerprint comparison finally outputs one single score. Depending on the stage in which the score is produced, it will respectively project into regions R_1, R_2 or R_3. Scores projected into regions R_1 and R_2 will be referred to as *early-outs*.

The score outputs of the three different stages of the AFIS algorithm used are illustrated in Fig. 1, where the multimodal distribution of the scores is illustrated.

Table 1 summarises possible values of the scores in the three stages of the AFIS algorithm. As the contribution of this paper is the LR model from the AFIS scores as in [1, 2], the description of the score computation algorithm used by this particular AFIS algorithm is out of the scope of this paper, and therefore the system will be used as a black-box. In this paper, fingermarks with 8-minutiae configurations are used, and we have not observed a dramatic difference in behaviour of the distribution of scores for fingermarks with configurations down to 5 minutiae and up to 12 minutiae [4]. It is true that different AFIS (or other biometric systems) might of course provide scores of different magnitudes, but the AFIS used here illustrates the problem of multimodal score distributions, and the proposed method is general for any multimodal distribution of scores coming from any biometric system.

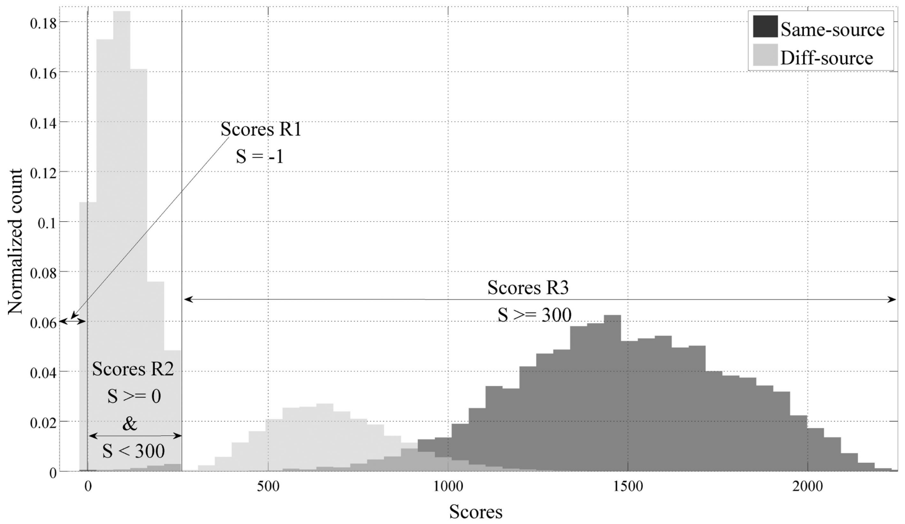

Fig. 1 *Histogram of scores produced by the AFIS algorithm used in this paper*

The *early-outs* process of computing scores has been used in the past by some COTS AFIS systems in order to improve computational efficiency in large database searches. Consequently, it is expected that researchers and forensic scientists performing evidence evaluation from AFIS scores will typically face multimodal score distributions. Moreover, from our experience, it has been observed that some of the freely available fingerprint systems sometimes present a high concentration of scores having the same, characteristically lowest value. In this context, typical baseline methods for computing LRs from multimodal score distributions rely on kernel density functions (KDFs) [1, 5]. However, as it will be explained below, this approach is prone to overfitting, not robust to the lack of data, and not well suited in situations where many scores present the same single value (like in region 1, in our case).

In this work, we propose a LR model which handles multimodal score distributions, presents robustness to the lack of data and to the observation of high concentrations of scores with the same value. The method proposed is based on splitting of the score range into three regions, and relies on the subsequent application of the rules of probability to combine independent evidence evaluation methods for the scores in each region. Also, a Bayesian solution is proposed to provide robustness to the lack of scores in different regions.

This paper is structured in the following way. In Section 2, we introduce the datasets used. Section 3 is dedicated to the definition of the evidence evaluation problem when multimodal score distributions are present, and to the introduction of the baseline LR method. In Section 4, we propose a new method to compute LR values from multimodal score distributions. In Section 5, we discuss the performance measures used and present the results in Section 6. Finally, Section 7 is dedicated to the discussion and conclusion.

Table 1 Different stages of the AFIS algorithm

Algorithm stage	Score range
1 – early outs 1	−1
2 – early outs 2	[0, 300]
3 – full comparison	>300

Table 2 SS and DS training scores

Individual	SS scores	DS scores
person 1	8455 marks 1 print	8455 marks 200,000 prints
person 2	2751 marks 1 print	2751 marks 200,000 prints
person 3	4666 marks 1 print	4666 marks 200,000 prints
person 4	2206 marks 1 print	2206 marks 200,000 prints
person 5	3179 marks 1 print	3179 marks 200,000 prints
person 6	3758 marks 1 print	3758 marks 200,000 prints

2 Datasets used

Since it is notoriously difficult to find forensically relevant, sufficiently large datasets with a reliable known ground-truth regarding the origin of the samples, we decided to use a set of simulated fingermarks [6]. Simulated fingermarks in this case refer to series of image captures of a finger moving on a glass plate of a fingerprint scanner [The fingerprint scanner used was the Smiths Heimann Biometrics ACCO 1394S live scanner.] (the procedure is described in detail in [6]), subsequently processed by degrading the quality, adding noise and a background image simulating the fingermarks recovered in a typical forensic scenario. 8-minutiae configuration fingermarks from six individuals paired with their corresponding fingerprints were used as our experimental dataset.

In order to train the LR models used for evidence evaluation, AFIS *training scores* are needed [1]. The training scores, resulting from the comparison of simulated fingermarks with the corresponding ground-truth reference fingerprint captured from the same individual in controlled conditions, are used for modelling a same-source (*SS*) score distribution, which represents the distribution of scores in comparisons where the fingermark and fingerprint originate from the same individual (i.e. the same source). AFIS training scores resulting from the comparison of simulated fingermarks with a 200,000 fingerprint subset of a Dutch police database are used to model the different sources (*DS*) score distribution, which represents the distribution of scores in comparisons where the fingermark and fingerprint originate from different individuals. Therefore, for all of these training scores (both SS and DS) the ground truth is known, because they have been generated using fingermarks and fingerprints of known origin. The number of fingermarks and fingerprints used for the LR model training are summarised in Table 2.

The aforementioned probability distributions (SS and DS) are used to compute the LR. In particular, they are used in combination with the *score* in a case, as given by the biometric system. This so-called *evidence score*, denoted as *S*, is the result of the comparison between the fingermark in the case (of unknown origin) and the fingerprint in the case (of known origin). We define evidence same source ($S_{\rm SS}$) as the evidence scores when the mark and the print come from the same source, and evidence different source ($S_{\rm DS}$) as the evidence scores when the mark and the print come from different sources. Thus, the evidence score, regardless of whether it is a $S_{\rm SS}$ or an $S_{\rm DS}$, is interpreted as a LR by the application of the previously trained LR model. $S_{\rm SS}$ scores will be transformed into SS LR values, and $S_{\rm DS}$ will be transformed into DS LR values for further performance evaluation. Therefore, this process is sometimes referred to as a score-to-LR transformation. Details of this process and its implementation for this article will be given below.

In our experiment, as the database size is limited, the evidence scores will be obtained on a leave-one-out basis from the entire training database (as in e.g. [7]). Thus, for Person 1, scores for the evidence same source (S_{SS}) are obtained on a 'leave-one-out' basis from the total available SS scores by comparing fingermarks of Person 1 with the fingerprints of Person 1; and scores for the evidence different source (S_{DS}) are obtained from the AFIS scores by comparing the fingermarks of Person 1 with the fingerprints from the aforementioned Dutch Police database (200,000 fingerprint subset). This process is repeated for all the individuals in the database. Using the 'leave-out' approach, in a given comparison one of the fingermarks is chosen to play the role of the crime-scene mark, and one of the fingerprints is chosen to play the role of the fingerprint in the case. Then, the rest of fingermarks are used to form training SS and DS score distributions for the LR model training. This process is iteratively repeated for all possible fingermark–fingerprint comparisons in the database.

3 LR computation from multimodal AFIS scores

In order to describe multimodal distributions of scores for evidence evaluation, KDFs have been explored in different previous works [1, 5]. The use of this method can be justified if the scoring mechanism (in our case, the AFIS comparison algorithm) produces the scores in a continuous range, with a smooth probabilistic distribution. However, our situation is different. The selected AFIS comparison algorithm outputs scores in three different ranges, one of it presents a single value, which causes poor performance of the KDF method in our case, as we will show below.

The regions in which the score range is divided by the AFIS algorithm are not overlapping. Therefore, a score cannot be observed in two regions simultaneously. We exploit this fact in our proposed method. It is based on splitting the score range in a way that a LR can be computed for each of the three score regions independently.

As a consequence of the strategy of splitting the score range, there are some of the regions where the number of scores reduces significantly. In particular, the majority of the SS scores are expected to be into region 3, leaving a few SS observations for regions 1 and 2. A similar situation occurs with the DS scores, where the majority of the scores project are expected to be into regions 1 and 2, leaving a few observations for region 3. The situation where no SS or no DS score is observed creates an unstable behaviour, resulting in LR values of unreasonable magnitudes, including the values of 0 or ∞. We propose a solution to this instability below.

3.1 LR computation

The LR approach, firstly considers the definition of the relevant hypotheses, or propositions, in a forensic case. These propositions are typically two, and referred to as the prosecution proposition (supporting that the suspect has a relationship with the crime) and the defence proposition (supporting the opposite). We can shape the prosecution and defence propositions at different inference levels depending on the investigation scenario [8] – the source level (where we inquire regarding the source of origin of the fingermark), the activity level (where we inquire regarding the activity that led to the transfer of the fingermark onto the crime scene) or at the offence level (the highest level, not commonly addressed by forensic examiners, but by the trier-of-fact, as it usually implies direct association between the offence and the fingermark).

At the source level, for the SS proposition we can further inquire whether the fingermark found on the crime scene is coming from a particular finger of the suspected individual or any finger of the suspected individual (finger/person level propositions). For the DS proposition we can inquire, whether the mark is coming from a different finger of a suspected individual, from a particular finger of any other individual in the available database (conditioning on a particular finger is not common for the DS proposition), or any finger of any other individual in the database.

For simplicity, we address simpler (and less informative) propositions at the source level [A change at the level or specificity of the propositions induces a change in the LR model.], namely:

H_p (SS): The fingermark and the fingerprint originate from the same finger.
H_d (DS): The fingermark and the fingerprint originate from different fingers.

Under these propositions, the SS and DS training score distributions can be obtained from a probabilistic model. These distributions will be used to compute the LR for the given evidence score S in a case.

For the AFIS used in this work, the distributions of SS and DS scores are not only multimodal, but the score distributions observed in each region also vary in their shape, mainly due to the three-different-stages of the scoring process of the AFIS. Moreover, in most of the cases, the majority of the SS scores projects into R_3 region, because the comparison showing a higher score tends to be the SS comparison resulting in a score >300. Conversely, the majority of the DS scores projects into R_1 and R_2 regions, because a comparison showing a lower score tends to be a DS comparison and results in a score <300.

In the forensic literature, different strategies have been proposed for calculating LRs from continuously distributed AFIS scores. In the field of score-based biometric recognition [1, 2, 9–12], the following general LR model has been defined (example shown in Fig. 2):

$$LR_i = \frac{f(S|H_p)}{f(S|H_d)} = \frac{f(\Delta(mx, py)|H_p, \boldsymbol{S}_d)}{f(\Delta(mx, py)|H_d, \boldsymbol{S}_p)} \quad (1)$$

where for the fingerprint evidence evaluation datasets are defined in the following way:

$f()$ – in the equation stands for the probability density function of continuous scores.
$S = \Delta(mx,\ py)$ – an evidence score between the fingermark mx found on the crime scene and the fingerprint py of the suspect.
$\boldsymbol{S_p}$ – a set of training scores obtained from comparing the training set of simulated fingermarks of the suspect with the reference fingerprint of the suspect.
$\boldsymbol{S_d}$ – a set of training scores obtained from comparing the crime scene fingermark and a subset of the fingerprints from the population database used in the model (in this case a subset of operational 10-print card database of individuals not related to the marks, and provided by the Dutch Police). This subset, counting 20k individuals (200,000 fingerprints), was chosen based on the fact that the features extracted from the fingerprints in this subset have been manually (human) verified.

3.2 Baseline method – KDF

KDF is a common strategy used for handling univariate, multimodal distributions, as it can be seen in [1, 5]. KDF is non-parametric and has been shown to be prone to overfitting (see Fig. 3), since the number of degrees of freedom of the model increases with the data. Thus, dataset shift between the training and the testing data can lead to a loss of performance of the LR method, because the inference model overfits the training data. Overfitting can even lead to so-called strongly misleading evidence [13], i.e. LR values of high magnitudes that support the wrong proposition in a case. These huge values are mainly due to the poor description of the tails of the score distribution, an effect which is aggravated with the lack of data.

In Fig. 3, we present two examples of erroneous behaviour of KDF due to overfitting. The plot at the left-hand-side shows a LR for an evidence SS score ($S_{\text{SS}} = 940$ with LRE_{SS}) of a numerically infinite magnitude. The plot at the right-hand-side illustrates an

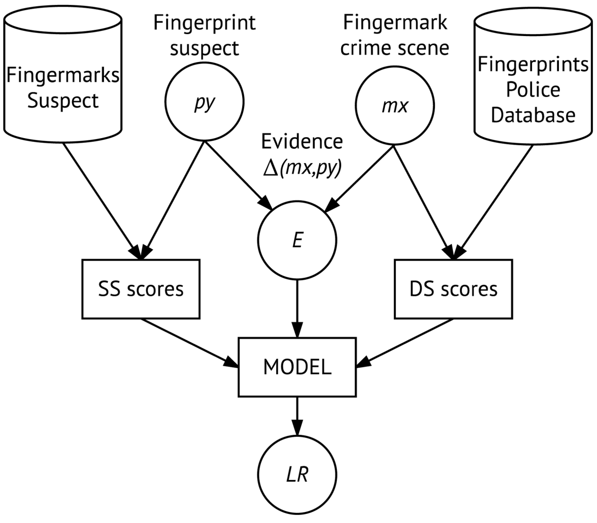

Fig. 2 *Illustration of the general LR model as used in this paper*

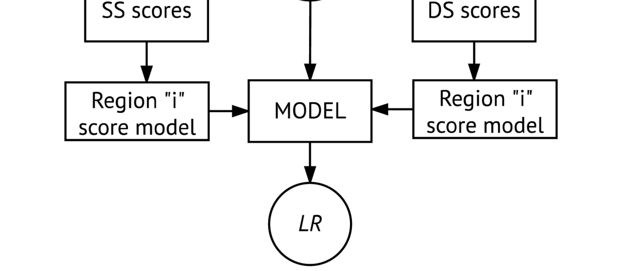

Fig. 4 *Proposed LR model for multimodal distributions*

example of a LR = 10^{91} for an evidence DS score ($S_{\mathrm{DS}} = 1080$ with $\mathrm{LRE_{DS}}$), which is supporting the wrong proposition in a very strong way. The main reason for the huge score magnitudes is the overfitting of the DS KDF density computed from the DS scores, which assumes that the probability of finding a score outside the training score range is numerically close to zero.

4 Solution proposed

Although the KDF method can be appropriate in other scenarios, the examples of LRs of extreme magnitudes and strongly misleading evidence when applying the KDF to the AFIS score distributions, as shown in Fig. 3, advocate for the use of an alternative method in our case.

4.1 Multimodal LR method

Each fingermark-to-fingerprint comparison performed by the AFIS system used results in one single score. This score can either project to region 1, region 2 or region 3. The events of observing the scores in a particular region (R_1, R_2 or R_3) are therefore mutually exclusive and exhaustive. Then, from (1), the following expression can be derived: (see (2)) The density of a score will be zero outside the region R_i if the score is conditioned to be in region R_i. Thus, assuming that the evidence score S belongs to the region R_i, we obtain the following compact representation of (2), where we replace LR with LR_i to stress that the score for which we compute the LR belongs to region R_i:

$$LR_i = \frac{f(S|R_i, H_p) \times P(R_i|H_p)}{f(S|R_i, H_d) \times P(R_i|H_d)} \tag{3}$$

This expression considers that the rest of probability densities for all regions will be zero outside R_i. Moreover, $(P(R_i|H_p))/(P(R_i|H_d))$ is the ratio of probabilities of observing R_i scores given that the fingermark and the fingerprint originate from the same finger over the probability of observing R_i scores given that the fingermark and the fingerprint originate from different fingers. This way, different LR models can be defined for each region, because we actually know in which region is located each observed score (see Fig. 4).

We can now proceed with describing the models chosen for each of the three regions. It should be noted that a variety of models proposed in the literature were tested and a brief overview of the models considered is summarised in Section 5.1.

4.1.1 Scores in region 3:

$$LR_3 = \frac{f(S|R_3, H_p) \times P(R_3|H_p)}{f(S|R_3, H_d) \times P(R_3|H_d)} \tag{4}$$

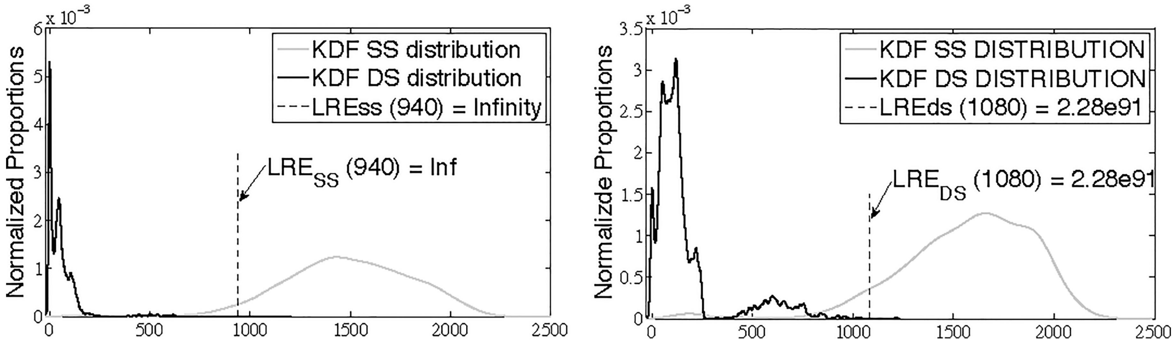

Fig. 3 *KDF fit to the SS and the DS score distributions – examples of LR values with extremely high magnitudes: LRE_{SS} (left) and LRE_{DS} (right)*

$$\begin{aligned} LR &= \frac{f(S|H_p)}{f(S|H_d)} \\ &= \frac{f(S|R_1, H_p) \times P(R_1|H_p) + f(S|R_2, H_p) \times P(R_2|H_p) + f(S|R_3, H_p) \times P(R_3|H_p)}{f(S|R_1, H_d) \times P(R_1|H_d) + f(S|R_2, H_d) \times P(R_2|H_d) + f(S|R_3, H_d) \times P(R_3|H_d)} \end{aligned} \tag{2}$$

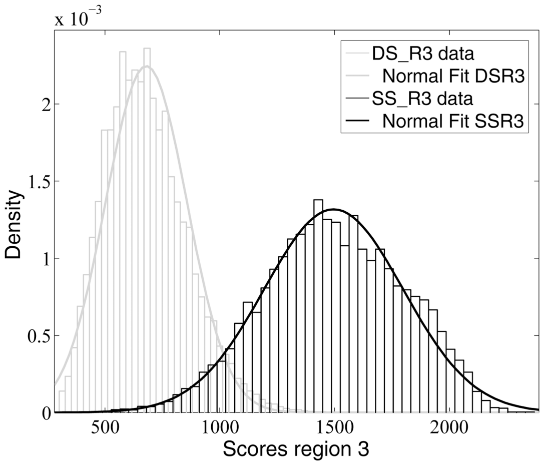

Fig. 5 *Score distributions in R_3 region, with a Gaussian fit*

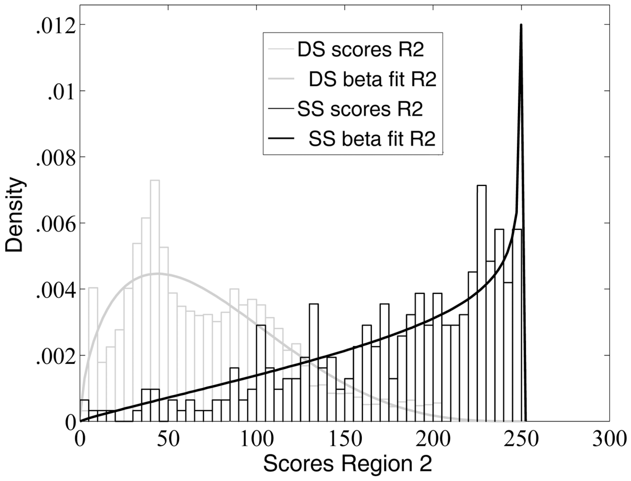

Fig. 6 *Scores in R_2 region, with a Beta fit*

From the histograms of the SS and the DS score distributions in Fig. 5, we consider as a reasonable initial assumption that the scores in R_3 region can be well fitted to a Gaussian distribution. The Gaussian model in region 3 yields our best performance amongst the models considered.

$$LR_2 = \frac{f(S|R_2, H_p) \times P(R_2|H_p)}{f(S|R_2, H_d) \times P(R_2|H_d)} \qquad (5)$$

The DS score distribution in R_2 region appears to be skewed, and the SS score distribution appears to be monotonically rising in this region. Although different parametric and non-parametric data fits have been tested for R_2 region scores [14], the Beta function was chosen due to its computational simplicity, its limited range and its flexibility; because R_2 region is limited, other distributions such as Gaussian or Gamma with unlimited score ranges were rejected (see Fig. 6). The Beta model in region 2 yields our best performance amongst the models considered.

4.1.3 Scores in region 1: From (3), and for region 1, we have

$$LR_1 = \frac{f(S|R_1, H_p) \times P(R_1|H_p)}{f(S|R_1, H_d) \times P(R_1|H_d)} \qquad (6)$$

As mentioned in Section 1, all of the scores observed in the R_1 region present one particular value of the score (−1) assigned by the AFIS algorithm. Nevertheless, the scores in this region also carry evidential information, and as such should not be excluded from the evaluation. For example, if the majority of training scores in region 1 are DS scores, an evidence score belonging to region 1 should provide support towards the DS proposition. The proposed model implicitly considers this, as it will be shown.

Equation (6) considers the ratio of densities $f(S|R_1, H_p)$ and $f(S|R_1, H_d)$. However, as the only possible value of the score is −1 for both densities, it can be shown that, by integrating over a limiting small region around the −1 value in both numerator and denominator densities, their ratio can be accurately approximated to one. Therefore, (6) further simplifies to the ratio of probabilities of observing a score in R_1 region under either of the propositions $(P(R_1|H_p))/(P(R_1|H_d))$. Hence, the scores in region R_1 possess certain evidential value, despite the fact that all of them share the same value.

With this model, let's assume that a very few number of SS training scores are observed in R_1 region, and that they are mostly DS training scores. If we observe an evidence score of −1 (in R_1 region), the LR should then support the defence hypothesis. This happens if $LR = (P(R_1|H_p)/P(R_1|H_d))$. In addition, the apparent solution to the −1 scores of ignoring them because all of them have the same value appears to be a waste of the discriminating information, given by the fact that in R_1 there are mostly DS scores.

4.2 Robustness to the lack of scores in a region

In our proposed model, we have to consider the following probability ratio for each region R_i:

$$\frac{P(R_i|H_p)}{P(R_i|H_d)}$$

These are the probabilities of observing a score in the ith region, respectively, under the prosecution and the defence propositions. Even if a wealth of training data is available, we can encounter a situation when we have very limited number of scores in a region under one of the propositions (or no scores at all). This effect may cause instabilities in the LRs obtained by the model due to the lack of training scores. We exemplify this below.

4.2.1 Effect of the lack of scores in a region: Assigning $P(R_i|H_p)$ and $P(R_i|H_d)$ to each of the different regions R_i, $i = 1,2,3$ needs to address robustness to the sparsity of the scores in the training dataset. In order to illustrate this, we start with a simplified example, where we divide the score axis into two regions R_1 and R_2 – a *binary* division. We consider for illustration the scores under the assumption that H_d is true, but this example can be analogously applied to the scores under the assumption that H_p is true.

In order to assign the probability $P(R_i|H_d)$ that a given score will be observed in the region R_i, we need previous knowledge regarding the observations of scores in each of the regions, i.e. some training observations. Those observations are taken from the training scores $\boldsymbol{S_d}$, being N_d the number of scores observed under the defence proposition, in the following way. Let $R_d = \left\{R_d^1, \ldots, R_d^{N_d}\right\}$ be a sample of random variables, where R_d^j represents the region in which the jth DS training score was observed. In this *binary* example, the possible outcomes of each R_d^j are R_1 and R_2. Then, the outcome of R_d^j will be the region in which the jth score in $\boldsymbol{S_d}$ is observed. Thus, the training observations are the particular values of each of those random variables. We assume that variables $R_d = \left\{R_d^1, \ldots, R_d^{N_d}\right\}$ are identically distributed according to a Bernoulli distribution, where the probability of observing a score in R_i is precisely $P(R_i|H_d)$. Moreover, we assume that the variables are *conditionally independent given the model*. Then, it can be shown that the *maximum likelihood* rule to assign the probability that a score will be observed in R_i is as follows:

$$P(R_i|H_d) = \frac{M_i}{N_d} \qquad (7)$$

where M_i is the number of DS scores in the training set observed in R_i and the $N_d$is the number of observations of the scores under the defence proposition. If the training scores become sparse, and the training scores under H_d contains zero score observations in R_i, i.e. $M_i = 0$, we get the following:

$$\frac{P(R_i|H_p)}{P(R_i|H_d)} = \frac{P(R_i|H_p)}{M_i/N_d} = \frac{P(R_i|H_d)}{0/N_d} = \infty \tag{8}$$

In this particular case, it results in the undesirable effect of producing an unrealistic strength of evidence: a LR of infinite value. An analogous derivation results in a null LR for SS scores observed in a region where no SS scores have been observed before.

An outcome of $LR = 0$ or ∞ is very likely to occur if no score, either SS or DS, is observed in one of the regions. The problem arises particularly in R_1 region, where the SS scores are quite rare, but can likewise occur in R_2 or R_3 regions.

4.2.2 Bayesian solution: In forensic science, the apparent problem of assigning probabilities when no observations are made in the training data has been studied for example in [15]. We propose a Bayesian solution to assign $P(R_i|H_p)$ and $P(R_i|H_d)$. We start from the above binary example, where a maximum-likelihood rule was considered. Under the same assumptions, if we instead consider that the parameter of the Bernoulli distribution has a uniform prior distribution (in the [0, 1] range), it can be shown that the solution inferred is the *predictive distribution*, which takes the following form:

$$P(R_i|H_d) = \frac{M_i + 1}{N_d + 2} \tag{9}$$

A full derivation is tractable, and can be found in [16] [Equations (6.66) to (6.73)]. The result is known as the *Laplace's rule of succession* [17]. For the sake of simplicity the application of this rule on our dataset will be demonstrated in R_1 region, where all the scores reach the value $S = -1$. Recall the binary example, where in R_1 region we obtained $LR = \infty$ because there were no scores observed in the training data in this particular region. Suppose a number of DS training scores $N_d = 20$ and that none of these scores are observed in region 1, thus $M_1 = 0$. Then, according to the previously proposed maximum-likelihood rule, we would obtain $P(R_1|H_d) = M/N = 0/20 = 0$ and the LR would be infinite. However, with the Bayesian uniform prior on the Bernoulli's parameter (Laplace rule of succession), we get the $P(R_1|H_d) = (M + 1)/(N + 2) = 1/22 \simeq 0.05$, which with an increasing number of training scores N_d will be asymptotically approaching zero, but will provide a non-zero numerical value. The interpretation of this result is that, additionally to the training data, a uniform prior for the model parameters forces to consider always at least an observation of one score in each of the regions under each of the propositions. Therefore, if H_d is true, we have to consider $N_d + 2$ scores for the two regions, and the scores observed in each region will be at least one. An analogous derivation provides equivalent interpretation for the case when H_p is true.

4.2.3 Generalisation to more than two regions: The problem addressed in this work requires a generalisation with respect to the rule of succession for the binary example, because we are dividing the score range into more than two regions. That means that the variables $\left\{R_d^1, \ldots, R_d^{N_d}\right\}$ will now have more than two possible outcomes. Therefore, their distribution cannot be a Bernoulli distribution. The generalisation to more than two possible outcomes, say Q possible regions, involves the assumption that the variables $\left\{R_d^1, \ldots, R_d^{N_d}\right\}$ follow a multinomial distribution. Moreover, since there are now Q parameters for this multinomial model, the prior distribution of the model parameters will be a Dirichlet distribution. Under these conditions, the derivation of the predictive distribution $P(R_i|H_d)$ for each of the regions can be found in [18]. It generalises the rule for more than two regions and provides the following result for the predictive distribution:

$$P(R_i|H_d) = \frac{M_i + 1}{N_d + Q} \tag{10}$$

or in the case of three regions as in the model proposed in this paper, we have

$$P(R_i|H_d) = \frac{M_i + 1}{N_d + 3} \tag{11}$$

Again, the analogous derivation produces a similar result for the case where H_p is true.

In our method, (11) will be used in all three regions to assign all the probabilities $P\left(R_i|H_p\right)$ and $P(R_i|H_d)$. This is because in case where both SS and DS scores are present, the probabilities do not change significantly with respect to the maximum-likelihood solution. On the other hand, in the cases where there are zero scores of either SS or DS, it will make the model more robust and will help to avoid, results of $LR = 0$ or ∞.

One can argue that a system providing $LR = 0$ or ∞ is the best that can be achieved if always correct. However, when the quantity or quality of the data is limited, a system providing $LR = 0$ or ∞ is not desirable, since such an output underestimates or overestimates the strength of evidence. In practice, those extreme *LR* values are the consequence of artefacts of the score modelling method and do not provide a realistic strength of evidence.

The motivations for the use of the Laplace rule of succession and its generalisation are thoroughly justified in [16, 17].

5 Performance measures

We will measure the performance of the KDF baseline and the proposed multimodal LR method, mainly focusing on their accuracy and their discriminating power. The accuracy was defined in [4, 19] as the closeness of agreement between a LR computed by a given method and the ground truth status of the proposition in a decision-theoretical inference model. The discriminating power was defined in [4, 19] as the performance property representing the capability of a given LR method to distinguish amongst forensic comparisons where different propositions are true. As it can be seen in [7, 19–21], empirical cross-entropy (ECE) curves and log LR cost (Cllr) are increasingly accepted measures of accuracy, the latter being a summarising measure of the former.

To measure the discriminating power, it is increasingly popular to use detection error trade-off (DET) curves [22] and its summarising measure, the equal error rate (EER). The DET curve, a Gaussian-wrapped version of the receiving operation characteristic (ROC) curve, is a two-dimensional plot of false acceptance versus false rejection rates. The linearity of the DET curve increases as the log-LR distributions become more normal. The closer the curve to the coordinate origin, the better the discrimination capabilities of the model [22]. Other methods increasingly used are Cllr$^{\text{min}}$ (minimum Cllr) and ECE$^{\text{min}}$ (minimum ECE) curves [6, 19].

Alongside the ECE and DET curves, a Tippet plot [13, 19, 23] showing the rates of misleading evidence for the prosecution and defence propositions (RMEP/RMED) will be presented, since they are also popular for LR-based evidence evaluation methods [24, 25].

5.1 Comparison and selection of LR models for each region

In the proposed multimodal LR method, the choice of the model for region 1 was straightforward and the calculation was based on the proportion of the SS and the DS scores in this region [∝ (SS/DS)]. The choice of Beta and Gaussian functions to describe the score distributions in regions 2 and 3, respectively, is presented in

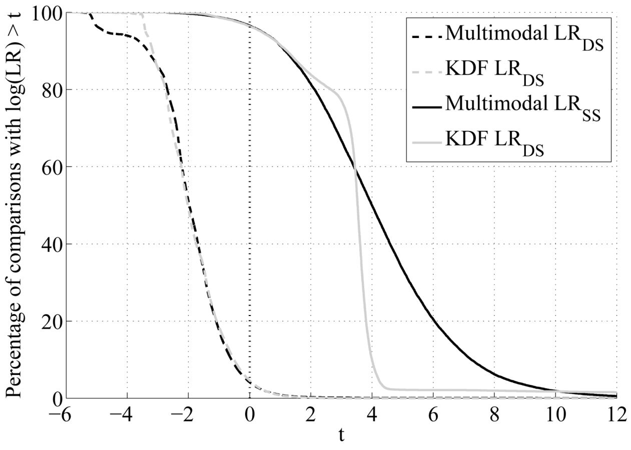

Fig. 7 *Tippett plots of the baseline KDF versus multimodal LR method*

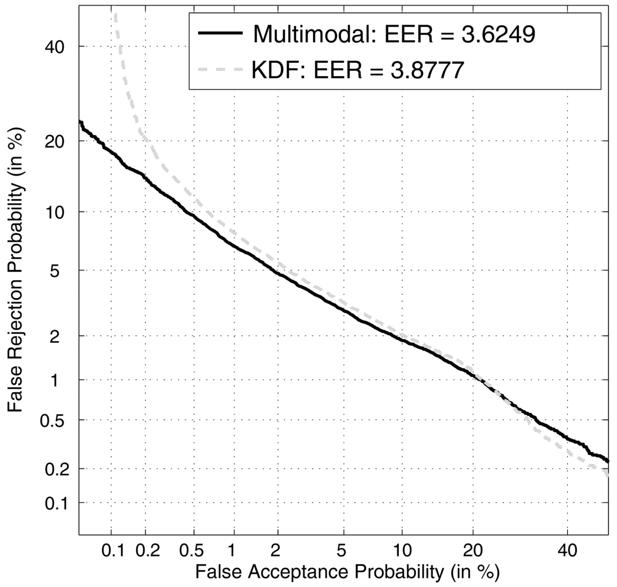

Fig. 8 *DET plots of the baseline KDF versus multimodal LR method*

this section by analysing the results of the different models considered.

The methods considered for regions 2 and 3 include popular approaches to score-to-LR transformation found in the literature. Apart from modelling the scores with Gaussian and Beta functions, linear logistic regression (LLR) and pool adjacent violators (PAV) were applied to the scores of regions 2 and 3 [21]. The combination of methods for this article was selected on the basis of the best performance, evaluated using EER, Cllr and Cllr^{min}. The complete list of all methods tested is summarised in Table 3.

The choice of Beta distribution for region 2 and Gaussian distribution for region 3 was based on the results presented in Table 3 and the best performance of the multimodal model in this configuration. Table 4 summarises the chosen model.

6 Results

Fig. 7 shows the Tippett plots for the baseline and for the proposed multimodal LR method. Tippett plots present one minus empirical cumulative distributions of LR values in the experiment. The intersection of both curves with the $\log LR = 0$ vertical line defines the *rates of misleading evidence* (RME) for the DS and SS propositions. The corresponding rates of misleading evidence (as observed in Fig. 7) are $\text{RMEP}_{\text{MULTIMODAL}} = 3{,}45$; $\text{RMEP}_{\text{KDF}} = 3{,}6$; $\text{RMED}_{\text{MULTIMODAL}} = 4{,}6$; $\text{RMED}_{\text{KDF}} = 4{,}16$. However, in Fig. 7, we observe LR_{SS} values of huge magnitude for the baseline KDF method. This is due to the fact that the inverse cumulative density function of the LR_{SS} fails to converge in the bottom right corner. The curve never converges, indicating that roughly 3% of LR values are enormously high. In extreme cases the LR values reach infinity. Please note that the $\log_{10}(\text{LR})$ values represented in the curves (x-axis) have been limited for illustration purposes for the baseline method. This instability in the strength of the LR values is observed despite reporting similar rates of misleading evidence for both models. LR values of similarly high magnitudes have not been observed for the multimodal method, where the LR values show more moderate strength for the SS proposition.

In Fig. 8, we present the DET plot of both methods compared, in which the EER can be observed at the intersection of the DET curves [22] and the $x = y$ line in the plot. Another undesirable effect, different from the one observed in the Tippett plots shown earlier, is seen in DET plots for the KDF method in the form of a clear deviation of the dashed curve from linearity in the top left corner, which causes the false-acceptance rates to never approach 0%. This happens because some of the DS evidence scores (roughly 0.1% of the total DS scores) yield extremely large LR values, strongly supporting the *wrong* proposition. This again is a highly undesirable effect, which has consequences on the reliability of the LR values using the KDF method. Apart from this effect, we observe improvement in the EER for the multimodal method, indicating slightly better discrimination performance ($\text{EER}_{\text{KDF}} = 3878$ and $\text{EER}_{\text{MULTIMODAL}} = 3{,}625$, see Table 5). However, as reported in previous work [21], it is apparent that measuring the performance of a LR method solely using the EER becomes insufficient.

Table 3 Comparison of different models for LR calculation

	Region 1	Region 2	Region 3	Discriminating power		Accuracy
				EER	Cllr^{min}	Cllr
method combination	∝(SS/DS)	Beta	Gauss	3.62	0.14	0.15
method combination	∝(SS/DS)	Beta	PAV	3.69	0.14	0.15
method combination	∝(SS/DS)	Beta	LLR	3.84	0.15	0.16
method combination	∝(SS/DS)	PAV	Gauss	3.67	0.14	0.15
method combination	∝(SS/DS)	PAV	PAV	3.77	0.14	0.15
method combination	∝(SS/DS)	PAV	LLR	3.92	0.15	0.16
method combination	∝(SS/DS)	LLR	Gauss	3.70	0.14	0.15
method combination	∝(SS/DS)	LLR	PAV	3.79	0.14	0.15
method combination	∝(SS/DS)	LLR	LLR	3.95	0.15	0.16

Table 4 Choice of the multimodal and the baseline methods for the LR calculation

Region 1	Region 2	Region 3
MULTIMODAL method		
$\propto(\text{SS/DS})_{\text{Bayesian}}$	Beta	Gaussian
BASELINE method		
KDF baseline for the entire SS and DS score distributions in all regions		

In the ECE plots in Fig. 9, the performance of the two methods is presented in terms of accuracy (solid line, the smaller the better).

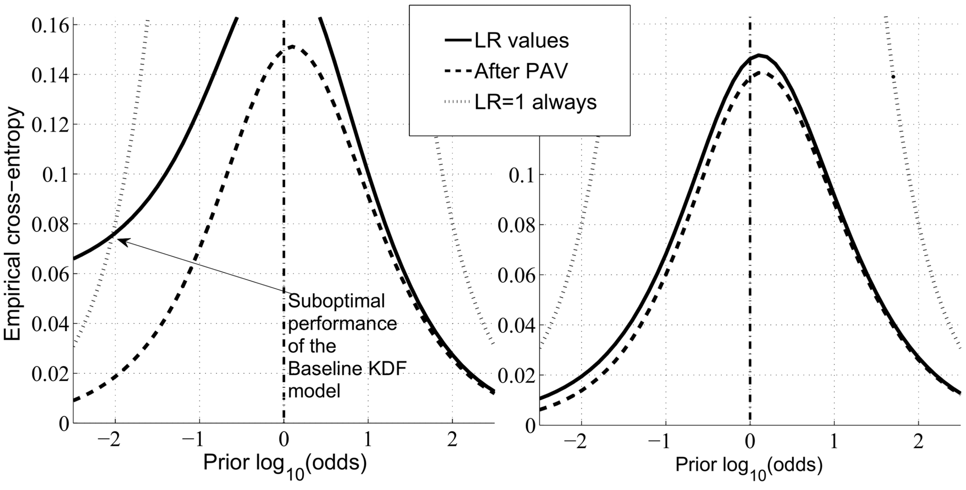

Fig. 9 *ECE plots of the baseline KDF (left) versus multimodal LR method (right)*

The value of Cllr is found at the intersection of the solid line and the vertical line at prior log-odds = 0. Moreover, the discriminating power can be seen in the so-called ECE$^{\text{min}}$ curve (dashed line, the lower the better), and by Cllr$^{\text{min}}$ – found at the intersection of the dashed line and the line at prior log-odds = 0. The smaller the distance between the solid and dashed curves, the better the calibration of the method. Ideally, both solid and dashed lines should be below the black dotted curve, which represents a *neutral* reference system, which always assigns $LR = 1$ to every evidence score.

A clearly undesired behaviour of the baseline KDF method is also shown in the ECE plots in Fig. 9 for the prior log odds smaller than −2. In that particular range, the baseline model performance is worse than that of the neutral system (which constantly returns $LR = 1$), essentially resulting in much larger calibration loss than the multimodal method. It also warns about the low reliability of the baseline LRs if they are used in cases where the prior-log-odds are below −2. This means that the baseline LR method is not reliable for all possible forensic cases and should be used with caution. For the baseline method we measured Cllr = 0.19 and Cllr$^{\text{min}}$ = 0.15; while for the multimodal method Cllr = 0.15 and Cllr$^{\text{min}}$ = 0.14 (Table 5).

Fig. 9 also shows that the multimodal method presents much better performance figures (ECE, ECE$^{\text{min}}$ and calibration) than the baseline method for the entire range of prior odds. Moreover, the ECE curve of the multimodal method is always below that of the neutral reference, as opposed to KDF. This means that the multimodal method can be used for any range of the prior-odds, and it will give useful information for decision-making.

Table 5 summarises the results presented graphically above. The relative improvement of the multimodal method compared to the baseline KDF achieved was approximately 21% for the Cllr and 6.5% for the EER. The performance improvement is relevant and stable for all the performance measures presented.

To summarise the results, we can say that the proposed multimodal LR method based on Gaussian and Beta distributions outperformed the baseline KDF method. It must be noted that the objective of this paper was not to find the best score distribution descriptions for the scores in each region, which can be achieved by further optimisation or choice of different methods for describing the score distributions. The objective of this paper was to address the multimodal nature of the scores by the partitioning of the score range. The resulting model has been shown to be more robust than the baseline to overfitting and data sparsity. In this sense, the splitting of the score range and the application of the Bayesian solution proposed is the main contribution of this work, since it can be used for any score-based biometric system, or any comparison algorithm outputting one score per comparison. The distributions selected to represent the data in different regions may vary in different biometric modalities or systems (e.g. in [25] the AFIS scores are reported to be better modelled by a log-normal distribution), but if the scores present a multimodal distribution with high concentration of scores in single-score ranges, this model represents an attractive alternative.

Table 5 Comparative performance of the baseline KDF and multimodal methods

LR method	Performance		
	Discriminating power		Accuracy
	EER	Cllr$^{\text{min}}$	Cllr
multimodal	3.62	0.14	0.15
baseline KDF[a] all regions	3.87	0.15	0.19

[a] The performance of the baseline KDF method was only possible to be measured after removing the extreme outliers (LR = infinity) and after setting a hard threshold at log(LR) = 30. As such, the reader is required to treat the over-optimistic results produced by the KDF baseline method with a certain amount of moderation in mind.

7 Discussion and conclusion

The main drawback of the traditionally used KDF method for modelling multimodal score distributions is its poor description of the tails of the training score distributions, together with the tendency to overfit them. The latter is mainly due to the high number of degrees of freedom of the non-parametric model, which increases with the increasing number and complexity of the training data. Thus, when a dataset shift is present between the training and testing data, as it typically happens in forensic science, the performance seriously degrades. This effect aggravates when the training data becomes sparse. Using a model based on KDF distributions in our problem we observed LRs of enormous magnitude supporting the correct proposition (e.g. $LRE_{\text{SS}} = 10^{130}$, $LRE_{\text{SS}} = \infty$), and even supporting the wrong proposition (e.g. $LRE_{\text{DS}} = 10^{91}$). This not only indicates a poor or unrealistic performance, in the context of forensic casework it also provides an illusion of certainty that transcends reality and leads to a misleading interpretation of the forensic evidence. In the ECE plots, we observed a poor calibration of the baseline KDF method, even worse than the neutral method in some cases.

In the method proposed, the SS and the DS score distributions were split into three different regions (R_1, R_2 and R_3) and modelled independently. Thus, the entire score distribution is divided into several simpler distributions that can be fit to simpler sub-models, dramatically reducing the degrees of freedom of the resulting model. As a consequence, the model generalises better to new, previously unseen data. We used ECE plots and Cllr values to evaluate the performance of the LR models, and observed a 21% relative improvement in the accuracy of the multimodal method in terms of Cllr with respect to the KDF baseline. We also observed much better calibration of the multimodal method than for the baseline KDF method for the whole range of the prior $\log_{10}$ odds. The multimodal method improved the discriminating capabilities of the system in terms of EER (6.5% relative improvement of the multimodal method over the baseline KDF).

Based on its improved performance and computational simplicity, the multimodal LR method was used in [4] to evaluate the coherence of the scores produced by an AFIS algorithm. The proposed approach can be used in cases where multimodal score distributions are observed, even in cases in which discrete score regions are presented. Finally, it can be also used for any score-based discipline, biometric or not.

8 Acknowledgments

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

9 References

[1] Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., *et al.*: 'Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems', *Forensic Sci. Int.*, 2005, **155**, (2–3), pp. 126–140

[2] Egli, N., Champod, C., Margot, P., *et al.*: 'Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – modelling within finger variability', *Forensic Sci. Int.*, 2007, **167**, pp. 189–195

[3] Meuwly, D., Veldhuis, R.G.F.: 'Forensic biometrics: from two communities to one discipline'. Int. Conf. of the Biometrics Special Interest Group, Proc. of the BIOSIG, Darmstadt, Germany, 2012, pp. 207–218

[4] Haraksim, R., Meuwly, D., Ramos, D., *et al.*: 'Measuring coherence of computer-assisted likelihood ratio methods', *Forensic Sci. Int.*, 2015, **249**, pp. 123–132

[5] Meuwly, D.: 'Reconnaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approche Automatique'. *PhD Thesis*, University of Lausanne, 2001

[6] Rodriguez, C.M., Jongh, A.de, Meuwly, D.: 'Introducing a semi-automated method to simulate a large number of forensic fingermarks for research on fingerprint identification', *J. Forensic Sci.*, 2012, **57**, (2), pp. 334–342

[7] Ramos, D., Gonzalez-Rodriguez, J., Zadora, G., *et al.*: 'Information-theoretical assessment of the performance of likelihood ratio methods', *J. Forensic Sci.*, 2012, **58**, (6), pp. 1503–1518

[8] Cook, R., Evett, I.W., Jackson, G., *et al.*: 'A hierarchy of propositions: deciding which level to address in casework', *Sci. Justice*, 1998, **38**, (4), pp. 231–239

[9] Hepler, A.B., Saunders, C.P., Davis, L.J., *et al.*: 'Score-based likelihood ratios for handwriting evidence', *Forensic Sci. Int.*, 2012, **219**, (1–3), pp. 129–140

[10] Ramos, D.: 'Forensic Evaluation of the Evidence using Automatic Speaker Identification System'. *PhD Thesis*, Univeridad Autonoma de Madrid, 2007

[11] Meuwly, D.: 'Forensic individualization from biometric data', *Sci. Justice*, 2006, **46**, pp. 205–213

[12] Gonzalez-Rodriguez, J., Rose, P., Ramos, D., *et al.*: 'Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (7), pp. 2104–2115

[13] Lucena-Molina, J.J., Ramos-Castro, D., Gonzalez-Rodriguez, J.: 'Performance of likelihood ratios considering bounds on the probability of observing misleading evidence', *Law, Probab. Risk*, 2015, **14**, (3), pp. 175–192

[14] Haraksim, R.: 'Validation of likelihood ratio methods used for forensic evidence evaluation: application in forensic fingerprints'. *PhD thesis*, University of Twente, Enschede, 2014

[15] Johnson, W.E.: '*Logic part III: the logical foundations of science*' (Cambridge University Press, 1924)

[16] Brenner, Ch.: 'Fundamental problem of forensic mathematics-The evidential value of a rare halotype', *Forensic Sci. Int., Genetics*, 2010, **4**, pp. 281–291

[17] Jaynes, E.T.: '*Probability theory: the logic of science*' (Wiley, 1994), ch. 18

[18] Zabell, S.L.: 'The rule of succession', *Erkenntnis*, 1989, **31**, (2/3), pp. 283–321

[19] Meuwly, D., Ramos, D., Haraksim, R.: 'A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation', *Forensic Sci. Int.*, 2016, in press DOI: http://dx.doi.org/10.1016/j.forsciint.2016.03.048

[20] Ramos, D., Gonzales-Rodriguez, J.: 'Reliable support: measuring calibration of likelihood ratios', *Forensic Sci. Int.*, 2013, **230**, (1–3), pp. 156–169

[21] Brümmer, N., du Preez, J.: 'Application independent evaluation of speaker detection', *Comput. Speech Lang.*, 2006, **20**, (2–3), pp. 230–275

[22] Martin, A., Doddington, G., Kamm, T., *et al.*: 'The DET curve in assessment of detection task performance'. Proc. Eurospeech '97, Rhodes, Greece, September 1997, vol. **4**, pp. 1899–1903

[23] Tippett, C., Emerson, V., Fereday, M., *et al.*: 'The evidential value of the comparison of paint flakes from sources other than vehicles', *J. Forensic Sci. Soc.*, 1968, **8**, (2–3), pp. 61–65

[24] Neumann, C., Champod, C., Puch-Solis, R., *et al.*: 'Computation of likelihood ratios in fingerprint identification for configurations of three minutiae', *J. Forensic Sci.*, 2006, **51**, (6), pp. 1255–1266

[25] Egli, N.: 'Interpretation of partial fingermarks using an automated fingerprint identification system'. *PhD Thesis*, University of Lausanne, 2009