Title:

Automated face recognition in forensic science: Review and perspectives

Abstract:

With recent technological innovations, the multiplication of captured images of criminal events has brought the comparison of faces to the forefront of the judicial scene. Forensic face recognition has become a ubiquitous tool to guide investigations, gather intelligence and provide evidence in court. However, its reliability in court still suffers from the lack of methodological standardization and empirical validation, notably when using automatic systems, which compare images and generate a matching score. Although the use of such systems increases drastically, it still requires more empirical studies based on adequate forensic data (surveillance footages and identity documents) to become a reliable method to present evidence in court. In this paper, we propose a review of the literature leading to the establishment of a methodological workflow to develop a score-based likelihood-ratio computation model using a Bayesian framework. Different approaches are proposed in the literature regarding the within-source and between-source variability distributions modelling. Depending on the data available, the modelling approach can be specific to the case or generic. Generic approaches allow interpreting the score without any available images of the suspect. Such model is henceforth harder to defend in court because the results are not anchored to the suspect. To make sure the computed score-based LR is robust, we must assess the performance of the model with two main characteristics: the discriminating power and the calibration state of the model. We hence describe the main metrics (Equal Error Rate and Cost of log likelihood-ratio), and graphical representations (Tippett plots, Detection Error Trade-off plot and Empirical Cross-Entropy plot) used to quantify and visualize the performance characteristics.

Keywords: Facial image, automatic system, likelihood ratio, score, calibration

Introduction

Although the use of the face as a means of forensic identification had been displaced by the use of fingerprints and DNA, the multiplication of captured video and photographs of criminal events has gradually brought the comparison of faces to the forefront of the investigative and judicial scene [1, 2]. Forensic face comparison is performed manually or using an automatic biometric system. Either way, it has become a ubiquitous and essential tool to guide investigations but is still not trusted as evidence in court in the same way as fingerprints or DNA are. The main reason for that is the lack of standardization and performance measurements. Automatic systems show increasingly higher accuracy and are an unquestionable means of time gain on fastidious comparison tasks, which is the first reason to focus research on their use. Although these systems have been researched for several years, there is still no method to evaluate evidence derived from them that fully meets both forensic and judicial needs as it exists in the fields of speaker recognition [3, 4] and fingerprints comparison [5], for instance. Automated face recognition is being studied more widely through technical and statistical issues [6, 7] as well as in the field of forensic intelligence [8], but not in the context of its evidential weight. In 2018, Zeinstra and colleagues offered a survey that covered well human-based face comparison and initiated discussions about automatic processes [9]. From that point on, the main purpose of this paper is to present a critical review of evaluative models using biometric automatic systems, and to propose an evaluative framework for forensic face recognition. First, we summarize the current forensic use of facial images, from investigation to court. Then, we focus on the use of face recognition systems for the evaluation of the forensic evidence according to the Bayesian framework. Finally, we summarize the main issues that ought to be addressed in future work.

Forensic framework: applications of face recognition

Forensic science is the discipline studying the trace defined as the physical remnant of a litigious activity [10]. The overall importance and scope of use of images in forensic science are thoroughly addressed in [11]. In the field of face recognition, the trace is the query image recorded at the time of the facts under investigation, which may show the face of the offender, or other persons of interest (POI[1]). These images vary in terms of type and quality depending on the scenarios of the litigious activity (a), and their processing can serve three purposes in the forensic framework (b). We detail both of them, (a) and (b), further below.

a. Images involved in forensic case scenarios

Witness images

"Witness images" are defined in [11] as ambiguous images somehow related to an event, recorded by chance or deliberately by any individual or device. It thus concerns, for example, images recorded on mobile devices by actual "eyewitnesses" during an event. In most cases, these images are of average or poor quality, due to the operator and/or subjects movements and overall uncontrolled shooting conditions. For instance, such recordings are immediately sought by authorities after bombings to try identifying persons of interest and reconstruct the chain of events. Witness images can also refer to images extracted from social networks as well as pedo-pornography material from online sources or devices belonging to a POI. These can help to find an individual not only using face recognition but also any information from the background of the image to locate where (and when) the photo was taken. Another type of witness images are surveillance footages, recorded at all time by stationary cameras at hotspots. We describe them separately later on.

[1] Persons of interest can be any individuals related to the event under investigation, such as victims, suspects, and witnesses.

Surveillance footages

With the constant increase in the use of surveillance cameras in both private (e.g. shops, banks) and public (e.g. transport, public roads) sectors, CCTV (Closed-Circuit Television) images are now among the traces most often available in an investigation. A recurrent issue is that the images provided are not the raw recorded footages but screenshots or images in different formats. That complicates the comparison by adding distortions or reducing the quality of the image. In addition, these cameras are most often set high up on a wall and therefore record images from the top, which can lead to distortion of the feature proportions. A special case of CCTV images concerns the images recorded at ATMs (Automatic Machine Teller). These are taken at a very short distance from the subject with a hypergon lens, which allows a wide shooting angle despite the camera close proximity to the face but distorts the images by widening all in the centre compared to the edges. ATM images can be provided in cases such as thefts where a victim's credit card and/or bank notes are stolen during a withdrawal as well as for fraudulent uses of credit cards. A last type of surveillance images comes from the investigative surveillance teams who will record individuals of interest at a distance.

Official documents

Requests for face comparisons that does not imply CCTV images mostly concerns the use of false identity documents, particularly during customs controls, applications for residence permits or other state privilege [8]. In most cases, the documents provided are authenticated passports or identity cards. In this scenario, the comparison is facilitated by the fact that the trace and reference images are two identity documents with photographs taken under standardized conditions.

b. Forensic purposes

The forensic use of images go into three main steps, common to all types of forensic traces: investigative, intelligence and evaluative. A recurrent dichotomy, first defined in [12], differentiates between investigative and intelligence purposes on the one hand and evaluative process on the other hand. Indeed, these steps are parts of independent phases in the processing of forensic data, but investigation and intelligence can serve one another whereas the evaluation stage traditionally closes the investigation by presenting the relevant pieces of evidence in court.

Investigative and intelligence purposes

For every case coming to police notice, the first phase is the investigation. The aim is to gather information (e.g. traces), to exploit them to potentially point towards persons of interest or help understand the event. For a given case in which the trace is a facial image, two operations may be performed (Figure 1). First, an automatic system may allow comparing the trace to a whole database of images and generating a list of potential suspects. Secondly, when the investigation points towards a person of interest (named 'S' in Figure 1) absent from any database, the operator can compare the query image with reference photographs of that individual. All of these results can serve as effective leads to the investigators.

In parallel with the investigation, the monitoring phase consist of gathering intelligence to contribute not only to the investigation but above all to crime control and security model [13]. In a forensic intelligence process, data are collected in each case and then are analysed to extract information such as the *modus operandi* and information about the offender from the query image, along with links between cases by comparing the query image to traces from other cases. This information is then used in the current investigation and/or in a further intelligence process, to detect, extend or confirm series [2]. See [13-15] for an overview of the definition, applications and potential of forensic intelligence. In [8], the authors describe a concrete example of the current processing of images for investigative and intelligence purposes in Switzerland, *via* a platform used in the Latin part of Switzerland for the monitoring of crime series. The aim is to feed a centralized database with all casework data for every type of events and traces, to help detect and analyse criminal series. This allowed to observe that the number of links detected between cases using the images has increased drastically since 2009 along with the number of events with query images, while those made with other types of traces (DNA, fingerprint and shoemarks) remain steady [2]. However, the authors highlight the fact that in 2013, at the time of data collection, the percentage of images suitable for the use of automatic systems still was very low (3.2%). But, given the most recent development of such systems, both in terms of performance and dissemination, we expect that the number of images suitable for an automatic process would have significantly increased by now.

Evaluative purpose

When forensic experts are instructed further, the ultimate aim is to interpret the result of the comparison between trace and suspect images in order to present it as evidence in court (Figure 1). The issue in the field is that no standard method is available for that task. The comparison technique, whether it is performed manually or using an automatic system, must meet different legal requirements depending on the country to be used in court. In continental Europe, there is no legal constraint on the admissibility of scientific evidence. It is only specified that it is the judges who evaluate the relevance of the evidence according to the state of scientific knowledge. In the United States, the Daubert[2] case law (revision of Frye[3], still in use in several states) first defines that the judge retains the right to judge whether the method used to provide evidence is accepted in the scientific community (Federal Rule of Evidence 702). In addition, FRE 702 most recent revisions require that the method be based on a reliable, refutable scientific basis, that it has been verified and error rates are known, and that it is available for peer review and publication. Regardless of the legal constraints or the image comparison technique, it is necessary to develop and validate models to standardize the evaluative process and empower, or strengthen, the use of face comparison in court.

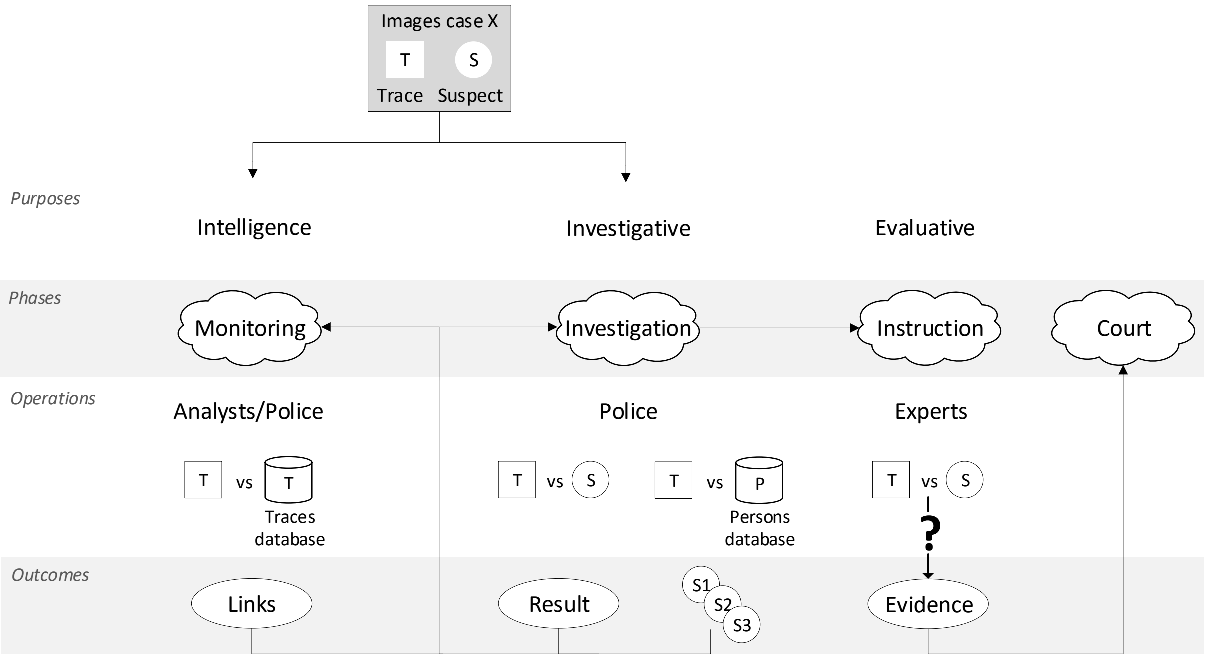

Figure 1 : Overview of the current forensic processing of images for a case with a trace and a suspect[4]

[2] Daubert et al. v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993)

[3] Frye v. United States, 293 F. 1013 D.C. Cir. (1923)

[4] We use the term “suspect” here, instead of “person of interest”, to simplify the case at hand.

Although the use of automatic systems is increasingly studied in the field of face comparison, it still lacks standardization and validation to be used for legal purposes. This is why the cases of face comparison where results may be presented in court are currently carried out only manually by specialized face comparison experts [8]. Four methods are typically used to analyse and compare faces: holistic, morphologic and photo-anthropometric processes, along with direct superposition of the images [16]. These methods are not exclusive and can be combined in order to carry out the most exhaustive analysis with regard to the information available on the image. We are not going into details about manual comparison in this paper. The reader is invited to refer to [8, 16-20] for details and recommendations about these processes. However, manual comparison also needs improvement in terms of standardization and validation [19-21], hence does not appear as a viable solution. In [8], the authors state that the low quality of CCTV images and the issues regarding the modelling of within-source variability (see Section 4 for more details) simply make trace images not suitable for being used in court. On the contrary, in [22], studies suggest that performance of face recognition tasks highly depends on the quality of training data on which a method is built. Therefore, comparing low-quality images is possible and requires using an adequate model developed from data of similar quality. All of these aspects are thoroughly discussed in this paper to assess the actual potential of facial images in an evaluative purpose, more specifically using automatic systems, as well as the shortcomings and limitations of these traces and methods. The evaluative process that should be carried out by the in order to present the evidence in court is addressed thoroughly in this paper and summarized in Figure 6.

We described in the above paragraphs the three main purposes of the forensic use of facial images. We made a rather strict demarcation between investigation and evaluation. We will discuss that further in the paper and submit that the latter should be more useful to the former, instead of only finalizing it. It is thus part of our future work to propose a method that includes early considerations of the evaluation process within the investigation stage.

c. Automatic face recognition systems

Application

Commercial systems currently used in forensic science show high and increasing performances [23, 24]. Corporations like Facebook and Google have developed their own proprietary algorithms, which they have claimed to be outperforming others [25, 26], and Microsoft have recently launched their systems, which perform better than the best commercial ones on several tasks [24]. This suggests that this gap may narrow in the coming years and that is why it is important to explore already opportunities and perspectives offered by commercial systems. A face comparison algorithm can have two distinct applications as for any biometric system: verification and authentication. Verification consists of the direct comparison of the image of an unknown person and a reference image of a known person (1-*vs*-1 comparison). This process is increasingly present in our daily lives, to unlock a computer or airport gates, for instance. The system compares the two images and calculates a score that reflects the degree of similarity between the two faces. If this score is above a predetermined threshold, the system concludes that both photographs represent the same person and grant access. Conversely, if the score is below this threshold, the system refuses access as the two are considered too distant to belong to the same person, according to it [27, 28]. Thresholds are typically set so that the false acceptance rate is as close as possible to zero, and the false rejection rate would have to be acceptable as to ensure fast and optimized use. On the other hand, authentication consists in comparing the image of an unknown person with all the images in a database of known individuals (1-*vs*-N comparisons). In this case, the system sorts the database to find the closest candidates. This application corresponds to the use already described for the investigation phase of a forensic case (Figure 1). Note that the systems used in forensic science allow ranking among potential sources and do not return any decision. It is based on this ranked list that a forensic expert will manually compare faces and take a decision.

So, unlike in face verification applications, automatic systems used for forensic cases do not generate a binary result based on the comparison of a score with a set tolerance threshold. It is the operators' responsibility to interpret the results reflected by the similarity score and manually inspect the image without their judgement being biased by the knowledge of circumstantial investigative information that may lead them to lack impartiality [29].

In addition, commercial systems, no matter how efficient they may be, operate using proprietary algorithms. The black-box effect that this imposes represents a potential barrier to the use of such systems in a court of law, where the need for transparency in scientific expertise is omnipresent. These systems, when operated as black boxes, remain controversial, as already debated in relation to fingerprints and DNA [30]. Furthermore, although the exact structure of the algorithms is unknown, the majority of them do not take into account soft biometrics, essential information for the expert such as gender, ethnicity, and the presence of distinctive signs (scars, freckles, moles, etc.). But the algorithm takes into account other elements that the expert does not exploit and allows to process large databases. It is important to note that, despite the emergence of face recognition algorithms based on deep learning and artificial intelligence [31], the black-box effect remains. The structure of these neural networks indeed allows generating impressive results by letting the algorithm learn to detect and sort relevant information from large databases (see [32] for general structure and applications of deep learning, and [33, 34] as an example of the most recent studies applied to face recognition). However, while the overall structure of these neural networks is theoretically known, the processing of features is not transparent, even for the developers. The issue raised from the lack of transparency hence remains topical, despite significant computational developments. The black-box effect should be contained by the publication of empirical studies attesting to the performance of these systems in order to validate their use in a forensic setting.

In the investigative purpose of authenticating a face against a database, it represents a complementary tool to the expert's work and not an alternative to it, insofar as the operator examines the list of results generated by the system. There is no standardized and validated method dedicated to the use of automatic systems in forensic applications [35]. For investigative purposes, this has currently less impact since results are discussed by the investigators to guide the investigation, whereas for evaluative purposes the result may become evidence in court. In this case, the result and the entire scientific procedure followed by the expert must meet legal requirements. In that context, the use of automatic systems remains little discussed, particularly because of the lack of validation of the methods, as alluded to previously.

Evaluation of evidence for automated forensic face recognitionScore-based likelihood ratio in the Bayesian framework

In forensic science, the scientist's purpose is to evaluate the weight given to alternative propositions by the available evidence. In this perspective, the Bayesian framework described in [36, 37] is recommended and used to interpret various types forensic evidence, including fingerprints, shoeprints, voices, DNA, handwritings or firearms [38-43]. More specifically, likelihood-ratios (LR) computation based on biometric similarity scores are being studied for speaker recognition [3, 4, 44], for fingerprints, using the AFIS system in [5], and more recently for face recognition [45, 46]. Within these fields, two LR computation methods can be used: the first exploits matching scores generated by automatic systems (score-based LR, SLR), whereas the second relies on the comparison of the features extracted from the facial characteristics (features-based LR). Nevertheless, the score-based computation method, is the most commonly used for forensic biometry [35]. For forensic face recognition, the evidence to consider is the score, which numerically translates the distance (or proximity) between two compared images. In this case, the LR expressed in equation (a) is related to the score s(S,T), distance between the vectors of image S (the known reference) and T (the unknown trace)

[47, 48]. Henceforth, the SLR, for a given score (s), is described as the ratio of two probability densities, as follows:

$$\mathrm{SLR}(s) = \frac{\mathrm{f}(s(\mathrm{S},\mathrm{T})|\mathrm{H}_1,\mathrm{I})}{\mathrm{f}(s(\mathrm{S},\mathrm{T})|\mathrm{H}_2,\mathrm{I})} \rightarrow \frac{\text{Probability to get this score when } \mathrm{H}_1 \text{ is true}}{\text{Probability to get this score when } \mathrm{H}_2 \text{ is true}} \qquad (a)$$

With

- s(S,T), the score of the comparison between trace (query image 'T') and reference (picture of the POI 'S').

- H_1 and H_2, the mutually exclusive alternative propositions, namely:

- H_1 : "The reference image and the trace image represent the same person"
- H_2 : "The reference image and the query image represent two different persons"

- I, the circumstantial information provided by the investigation, such as the ethnic group of the POI, or the biometric system used.

Here, we present a general and simplified phrasing as an example. In fact, the formulation of alternatives proposition and its conditioning information I is one of the most crucial steps and they have to be adapted to each case depending on the data and information at hand. The numerator of the SLR is conditioned by the within-source variability (WSV) and the denominator depends on the between-source variability (BSV), which in turns depends on I. The WSV represents the variations in score values when comparing several pictures of the same person, and the BSV is the variations in such values that occur when comparing pictures of different individuals.

In general, it would be considered as an ideal method to compute SLR directly from the data extracted from the characteristics, *i.e.* features-based model, using an open-source automatic system [47]. It would allow exploiting multivariate data without loss of information, while avoiding the black-box effect by knowing the algorithm details, such as the exact type of features analysed by the system, the frequencies at which they appear and the databases used. However, there is not such system (yet) in the field of face comparison. Current automatic systems, whether based on deep learning or not, compare two images and generate the result in the form of a similarity score. Currently, feature-based evaluation based carried out by the expert's based on manual comparisons while all automatic systems directly provide a score and not the numerical characteristics they used to compute it. It is therefore not yet possible to compare feature-based and score-based approaches in the field of automated face recognition.

b. Computing SLR

In order to compute a SLR from a score value for evaluation purposes, several questions have to be addressed first. Does the quality of the query image allow using an automatic system? Which database to use? What are the propositions according to the prosecution and defence? How to model the WSV and BSV distributions given the available data? All of these questions are essential for the development of the SLR-computation model, which we will further detail in Figure 6.

Regarding the quality of the image, several points are to be addressed. First, the systems may use a frontalization process to adjust non-frontal images, but for angles over 45°, a lot of information is lost and the rate of incorrect results rises [46]. Secondly, resolution must be high enough to distinguish the eyes along with the main features and the face should not be obstructed by sunglasses, scarf or other attributes. As introduced previously, studies have highlighted the correlation of face recognition performance with the nature of training data used to develop a model [22]. For example, low-resolution trace images require using a model developed with training data of similarly low quality, to ensure more efficient comparisons. Globally, the author suggests that models should be built with images that meet the real-life constraints of trace images.

Finally, optical distortion such as in ATM recordings does not prevent the system from analysing the image but this can affect the comparison outcomes as it changes the features proportions. See Figure 2 for a striking demonstration of morphological variations due to focal length and shooting distance changes [29].

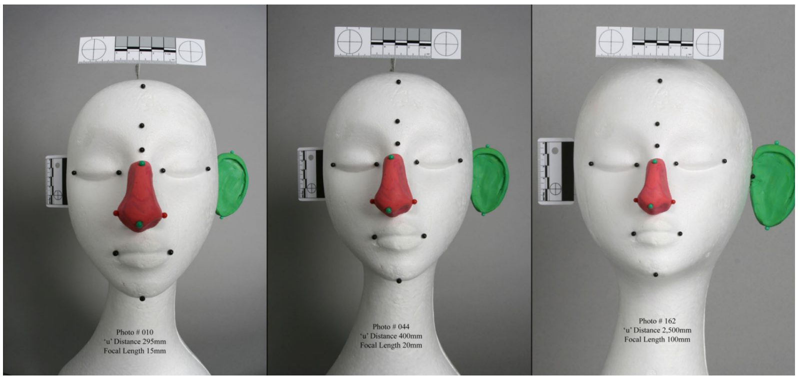

Figure 2: Illustration of apparent morphological face variations from three different perspectives. The subject does not change, only the focal length and shooting distances vary (from left to right, distance (mm): 295, 400 and 2'500, focal length (mm): 15, 20 and 100) [29]

If the case data (images for the trace T and the suspect S) are judged suitable for an automatic process, trace(s) and suspect reference(s) are compared automatically[5]. Then, the resulting score is used to assign a SLR regarding the appropriate WSV and BSV. These data allow assigning a SLR to a given score as in Equation a (see Figure 3 for an example of the assignation of a SLR to a given score value). To generate the WSV and BSV scores distributions, the expert has first to determine the propositions H_1 and H_2 along with the relevant population adequate to the case data and information at hand. The relevant population is determined according to the proposition H_2 under which the reference image and the query image are considered to represent two different persons. This population, specific to each case, includes individuals who are not the suspect but whose faces could be the one visible on the question photograph. In most fields, witnesses would be the ones to provide potential information about the author such as gender, skin colour or age range, for example. In face recognition cases, this kind of information may be directly observed on the query image. But it has to be noted that the relevant population is based first and foremost on the assumption of the defendant party. Therefore, if H_2 is: "The query image represents a different person, male or female, than the POI", even though the experts assumed from the image details that the person in the query image was a woman, the relevant population must contain individuals of both genders. The point is to ensure that the SLR weighs the evidence against the propositions set by the parties' allegations, and not by the expert. The role of the expert is to adapt the phrasing of the proposition according to the method chosen to model the BSV and WSV with the available data. These methods are detailed in the following section along with their use of the relevant population.

[5] When several traces and/or references are provided for the same event, each trace must be compared to each reference to generate as many scores (hence evidence) as comparisons performed. The SLR must then be generated combining all pieces of evidence (if they have been formally linked to the same event). See [49-51] for extensive researches and discussions about the general problematic of combining evidence, and [52, 53] for methods focused on combining biometric scores.

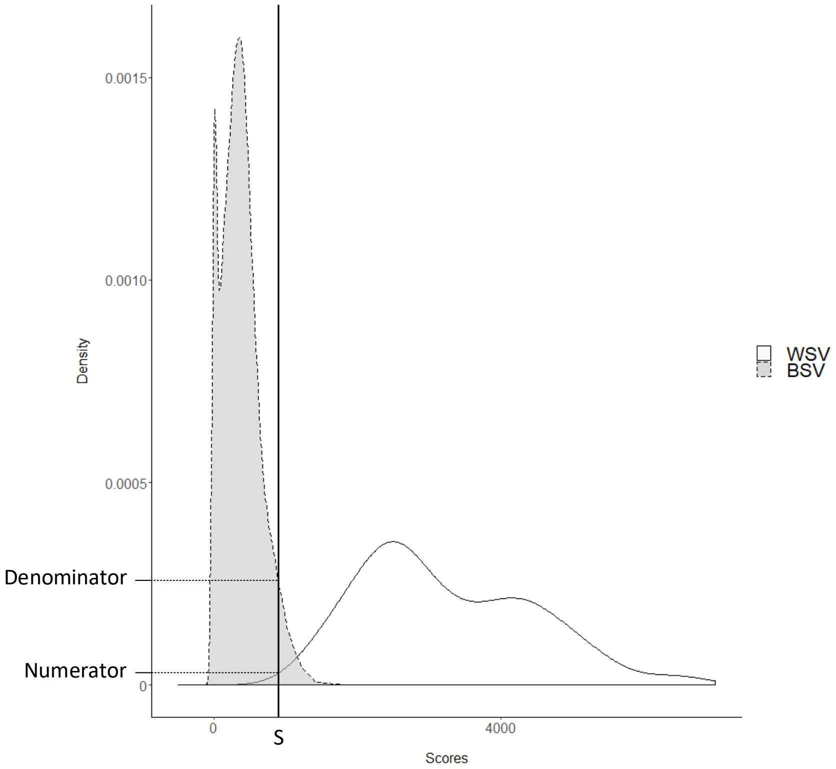

Figure 3: Method for the assignation of SLR numerator and denominator values from a score 's' and WSV and BSV modelling

c. Between-sources and within-source variability modelling

Different approaches are proposed in the literature regarding the BSV and WSV distribution modelling. Depending on the data available, the modelling approach can be specific to the case or generic. This has a direct impact on the SLR numerator and denominator conditioning and therefore on the work propositions to communicate in court. Details of all denominators and numerators conditioning as well as the corresponding H_1 and H_2 formulations are summarized in Table 1.

Between-sources variability

Between-source variability is modelled with scores generated from comparisons matching the alternative proposition, according to which the query image and the reference does not show the same person. This allows estimating the SLR denominator. The aim of this distribution is to represent the variations in score values when comparing photographs of different individuals. We can highlight three main approaches in the literature to grasp the BSV: anchoring the scores to the trace, to its source, or computing generic scores.

Trace-anchored approach

The trace-anchored approach described in [54] consists in calculating BSV scores by comparing the query image (T) with each picture of the relevant population database, as illustrated in Figure 4a. The alternative proposition H_2 thus states that “the individual at the source of ‘T’ is not the suspect, but someone else from the relevant population”. The relevant population will be defined by several apparent characteristics the expert may note during the trace analysis phase, as described previously. This population therefore should include all individuals from the database whose faces may be the one visible on the trace, according to the expert. This population therefore would include all individuals from the database whose faces may be the one visible on the trace, according to the expert (e.g. all males or all females, all with or all without apparent tattoo). However, we think that, for face recognition, the use of observable features to compose the relevant population may suffer from multiple restrictions. First, skin marks like tattoos, moles, scars, etc., are features too complicated to be used, because this would mean either the expert would have to sort the database manually to find individuals with similar features (which would be highly time-consuming), or the expert could sort individuals automatically using tags/attributes initially register to describe the visible soft features. Only the latter solution may make trace-anchored approach applicable, but such database with tagged images has to be built beforehand. Secondly, ethnic origin is mostly assumed on the basis of skin and hair colours. But not only is this prone to errors, notably due to genetic mixing in most populations, but such features can appear differently depending on the image quality and shooting conditions. Therefore, assuming individual's ethnicity should also be avoided. At last, age range could be used for composing relevant population, as long as it is determined by combined observations of several pertinent features. In light of all discussed points, it appears that the trace-anchored approach could not be applied for facial images if based on the visual analysis of soft biometrics by the expert.

Nevertheless, we see two methods to consider for composing the relevant population by comparing the trace to an unsorted database, automatically or manually. In the first case, the trace is compared automatically to all images of the database, and the individuals in the higher ranks of the matching scores list will composed the relevant population. Doing so, the downside would be to underestimate the LR by modelling BSV from scores closest to the WSV scores, according to the system. In the second case, the expert manually composes the relevant population by selecting potential sources with the most resembling faces compared to the face on T, according to him. This option would not suffer from the same constraint encountered using soft biometrics, as the expert would compare faces holistically (i.e. even if the expert assumes the suspect is a man, any resembling female would be considered as a potential source). However, depending on the size of the database, this solution can be utterly time consuming, and depends strongly on the expert's performance at sorting unfamiliar faces. But no study has yet shown the impact of variations in the choice of the relevant population for automatic face recognition, neither in investigative nor evaluative purpose.

Source (suspect)-anchored approach

The "source" here, as described in [55] for the evaluation of handwriting evidence, refers to the person allegedly at the source of the trace, *i.e.* the suspect, or POI in our discussion. This second approach thus consists in comparing the reference image of the POI with individuals in the relevant population (Figure 4b). In this case, the alternative proposition H_2 considered is: "The suspect is not the person in the trace image". The BSV thus depicts the range of score values when considering that the reference image and an image randomly selected from the population represent different persons. The aim is for the SLR denominator to reflect how rare/frequent it is to find, in the relevant population, a face similar to the suspect's face. This notion is defined as "typicality" in [56]. This is a crucial criterion to be included in any SLR computation, but it should not be considered as a whole approach but to be a part of it. The typicality has to be integrated through the selection of the training sets of images used to model the BSV, *i.e.* the relevant population chosen to best fit the case, as described previously for the trace-anchored approach. In this approach, the generated scores highly depend on the quality of case images. In particular, comparing query images from low quality CCTV with standard frontal reference pictures would generate lower similarity scores than when comparing two good-quality frontal images as both trace and reference.

Generic approach

In a generic approach, we compare all images of the relevant population pairwise (Figure 4c). The similarity scores generated for each comparison are used to model a generic BSV, *i.e.* not specific to the case at hand. Here, the BSV represents the range of score values when considering that two images randomly selected from the population represent different persons. As no information from neither the trace nor the suspect's reference image are considered to compute the SLR, this approach is much less informative than the previous two [57].

The range of similarity scores values are expected to be higher - respectively lower for distance scores - than when comparing a lower quality CCTV trace image with the reference images in the database, for most images of the database come from high-quality identity documents and police mugshots. For this approach, the proposition that is conditioning the SLR denominator is formulated in a more generic way: "Two different persons are visible in the two compared images".

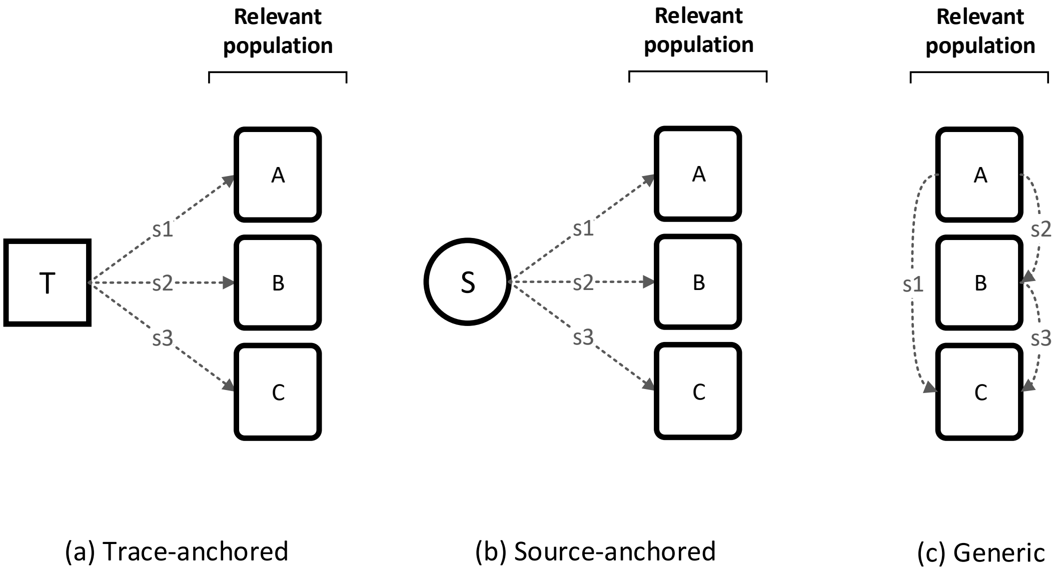

Figure 4: Scores 's$_i$' generated by the automatic system when comparing (a) the trace with images of the relevant population for the modelling of the trace-anchored BSV (b) the reference image of the suspect with images of the relevant population for the source-anchored BSV (c) images of each individual in the relevant population with each other for the generic BSV

Within-source variability

Within-source variability (WSV) modelling provides a representation of the variations that may exist between images when comparing several photographs of the same person. It informs the calculation of the probability of observing the result of the comparison between the trace image and the reference under the H_1 proposition, according to which the same person is visible on both compared images. The two main methods used consist in modelling the suspect-anchored or the suspect-independent WSV. These approaches have been discussed for face recognition in [45], as discussed later in this section. In [56], in addition to the description of the "typicality" to take into account in the SLR denominator computation, the author defines the "similarity" between the query sample and the know sample (the reference) as the main criteria of the numerator computation.

Suspect-anchored approach

In cases requiring images comparison, an expert may request additional photographs of the POI in order to model the within-source variability. If available, the person may be photographed directly to provide expert with both reference descriptive images (illustrated as circles in Figure 4 and 5) and control images (hexagons in Figure 5a). Control images are taken under conditions comparable to those of the query image [54] to be used as special references. They are also described as "pseudo-traces" in [54, 58, 59] in biometrics in general and more specifically for voices and fingerprints. When comparing facial images, the time between the recording of the trace and the references or control images often become problematic because of the ageing of the person. See [24, 60, 61] for examples of studies addressing the impact of ageing on the performance of face recognition systems. For most algorithms tested, the larger the age gap between two compared images, the less efficient the system is. However, it has been noted that the FaceNet algorithm, first described in [26], still performs honourably over very variable age ranges, when trained on adequate datasets [61].

Having both high-quality reference and control images is the ideal situation for modelling the within-source variability of the POI. This method consists in comparing the control images, on the one hand, to the reference images on the other (Figure 5a1). The purpose of comparing the control images with the references is to obtain a WSV score distribution more representative and specific to the case at hand, taking into account all information from the query image, since the control images are taken under conditions comparable to those of the trace. For instance, if the query image comes from CCTV footage, the control images will be recorded with a surveillance camera, ideally on the actual "crime scene", or elsewhere in similar conditions. The reference images, against which the control images will be compared, would still be standard mugshot photographs. These control images are ideal material but are rarely taken in operational cases. In most cases, the suspect is not available and even when he/she is, the recording of such images may be too expensive and time consuming to be considered. In casework in which only reference images are available, the WSV scores are calculated by comparing all references to each other, as illustrated in Figure 5a2. For the suspect-anchored approach, whether the expert is provided with both control and reference images or only the latter, the H_1 proposition can be formulated as follows: "The suspect is the person in the trace image".

Suspect-independent/Generic approach

In most cases, the expert has only one or even no reference image. Then, a generic suspect-independent approach is to be used [4], to overcome the lack of reference data using only photographs of "potential suspects" of the relevant population. First, this method requires selecting all individuals from the relevant population with at least two pictures available. Then, we obtain the WSV from the similarity scores for each potential suspect, by comparing their photographs in pairs (Figure 5b). The scores obtained for all potential suspect are then combined to model a single generic distribution. Proposition H_1 hence can no longer be centred on the suspect since the WSV is modelled with images of people not directly related to the case. Similarly to the generic BSV, the proposition is: "The same person is visible in the two compared images". Another generic approach is described for speaker recognition in [4]. The data available are the trace and only one reference recording of a suspect. Hence, the authors created two databases, one of them containing the recordings of 10 individuals taken under conditions comparable to those of the trace (five records per individual), and the second containing three reference recordings for each person under conditions comparable to those of the suspect's reference record. By transposing these conditions to a face recognition case, this means creating two databases of potential sources with, respectively, images matching the trace shooting conditions (ID documents, CCTV footages, etc.), and images matching the suspect's picture description (mostly ID documents or other portraits). Such an approach remains generic because neither the trace nor the suspect reference are used by the system. However, it requires either already having a large database with images taken under multiple different conditions, with several pictures for each individual, or finding several volunteers to take adequate pictures for each case in hand. Therefore, this approach does not seem easily applicable for operational forensic purposes.

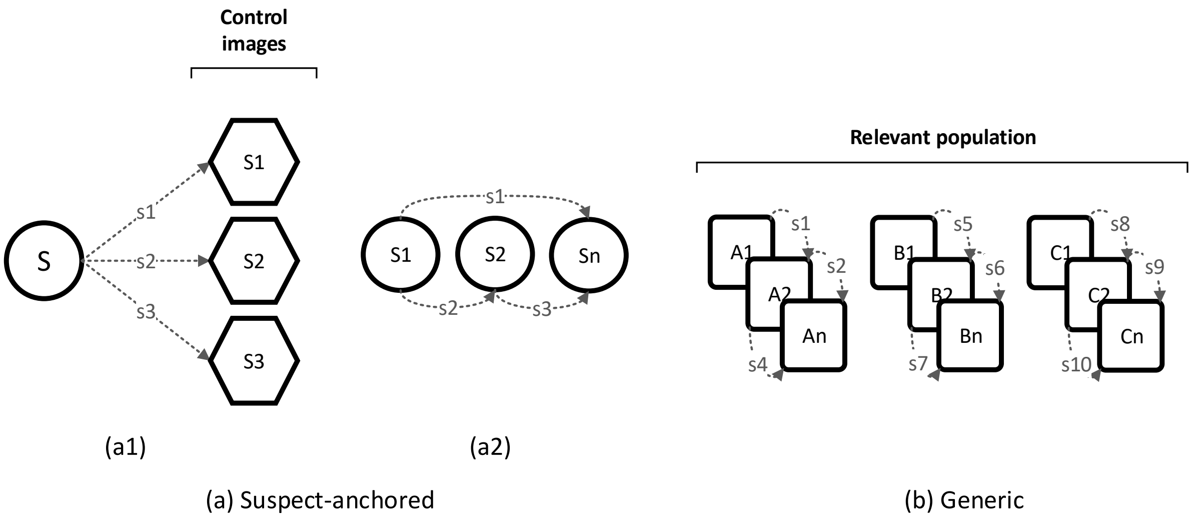

Figure 5: Scores 's_i' generated for comparisons of (a1) the suspect reference with their control set (a2) all the suspect reference images, both for modelling the suspect-anchored WSV (b) all reference images of each potential suspect in the relevant population for modelling the suspect-independent WSV

Among these studies describing and contrasting the various BSV and WSV modelling approaches [4, 45, 54-57], very few use forensic data to compare them in an operational context. It is thus essential to conduct such a study by applying each approach to a corpus of forensic cases to assess the resulting variations in SLR values for each of them. In the field of face recognition, the impact of a WSV suspect-independent approach on SLR values compared to a suspect-anchored approach has been studied in [62]. To do so, the authors use photographs of five individuals randomly selected from the FRGC (Face Recognition Grand Challenge) database [63]. For each of them, 36 images are considered, half of which is used as a reference database, and the other half is degraded (reduction of definition and resolution) to better match the uncontrolled conditions under which CCTV trace images are taken. These degraded photographs constitute the control database of each individual. From these data, the authors model the WSV and BSV for both suspect-independent and suspect-anchored approach. The computed scores allow computing Logarithmic SLR (LLR), and the orders of magnitude of their values for both approaches are compared using the corresponding verbal scale, adapted from [64]. In 59.2% of cases on the average, the LLR obtained by the suspect-independent approach correspond to those of the suspect-anchored approach. These results support the use of such an approach in the field of face recognition. However, for two out of five subjects, the suspect-independent approach was more effective and the percentage of LLR corresponding to the same verbal level for both approaches dropped to 28.5%. In order to more accurately assess the impact of the two approaches on SLR values, it would therefore be necessary to confirm these results with more subjects and genuine low-quality control images, instead of down-sampling high-quality images, as also suggested in [22].

Variability	Approach	Part of SLR(s) formula concerned	Proposition
BSV	Trace-anchored	$= \frac{\text{num}}{f(\mathbf{s}(\mathbf{S},\mathbf{T})\lvert\mathbf{T},\mathbf{H_2},\mathbf{I})}$	H$_2$: “The individual at the source of 'T' is not the suspect but someone else from the relevant population”
	Suspect-anchored	$= \frac{\text{num}}{f(\mathbf{s}(\mathbf{S},\mathbf{T})\lvert\mathbf{S},\mathbf{H_2},\mathbf{I})}$	H$_2$: “The suspect is not the person in the trace image”
	Generic	$= \frac{\text{num}}{f(\mathbf{s}(\mathbf{S},\mathbf{T})\lvert\mathbf{H_2},\mathbf{I})}$	H$_2$: “Two different persons are visible in the two compared images”
WSV	Suspect-anchored	$= \frac{f(\mathbf{s}(\mathbf{S},\mathbf{T})\lvert\mathbf{S},\mathbf{H_1},\mathbf{I})}{\text{denom}}$	H$_1$: “The suspect is the person in the trace image”
	Generic	$= \frac{f(\mathbf{s}(\mathbf{S},\mathbf{T})\lvert\mathbf{H_1},\mathbf{I})}{\text{denom}}$	H$_1$: “The same person is visible in the two compared images”

Table 1: Summary of the different approaches for the modelling of WSV and BSV along with the inherent work propositions with ‘s(S,T)’ the resulting score of the comparison between ‘S’, image of a suspect (source), and ‘T’, the trace image.

In Table 1, we propose a summary of each SLR formula conditioning and propositions phrasing depending on the chosen modelling methods. For the SLR denominator, the probability function of obtaining the score s(S,T) is conditioned by the data used for BSV modelling, respectively the query image T or the reference of source S, for anchored approaches. Meanwhile, for the WSV anchored approach, the score s(S,T) at the numerator can only be conditioned on the source S. In generic approaches for both BSV and WSV, scores s(S,T) are conditioned on neither of the trace T nor the source S, as they are modelled comparing images of individuals of the relevant population and henceforth only depend on I.

The formulation of the propositions is crucial as they condition the results that the expert will present in court. They must best represent the statement of both parties of a trial using the data available. For instance, the lack of any reference material leads to the use of a generic approach, using exclusively pictures of individuals of a database, not involved in the case in hand. The expert’s conclusion then cannot be directly anchored to the suspect, as none of their pictures has been used. That is why the propositions only refer to “distinct persons” or “same person” in “two compared images” without mentioning the suspect (Table 1). At this point, the limitation is that, if a party asks “May the person on the query image be the suspect or not?”, the experts could not comment further. Their generic model does not allow them to include specifically the suspect in their conclusion, contrary to a specific approach in which they use the suspect’s facial pictures. This is a brief introduction to the issues inherent to the propositions formulation. See for example [65] for a thorough and more practical explanation, applicable to any type of forensic evidence.

Once generated the WSV and BSV scores using the adequate approach, the corresponding probability density distributions, as illustrated by the example in Figure 3, have to be modelled in a way that they best fit the empirical score values. We will not discuss further this statistical aspect, we refer to [5, 46, 66] for examples and more details about the choice of parametric and non-parametric fitting methods. Figure 6 sums up the complete workflow of SLR assignation model development carried out by the expert for evaluative purposes. Starting from a hypothetical forensic case - if the material quality fits the requirements for automated use - the expert sets the model specifications concerning WSV and BSV modelling, formulation of the propositions and determination of the relevant population in regard to the information provided by the trace T and reference S. All comparisons performed afterwards using the biometric system as described in Figure 4 and 5 are summarized here in the operations block. The preliminary outcomes of this process are the computed scores, which allow fitting probability densities using the most adequate method, and then assign the reported SLR to present as evidence in court.

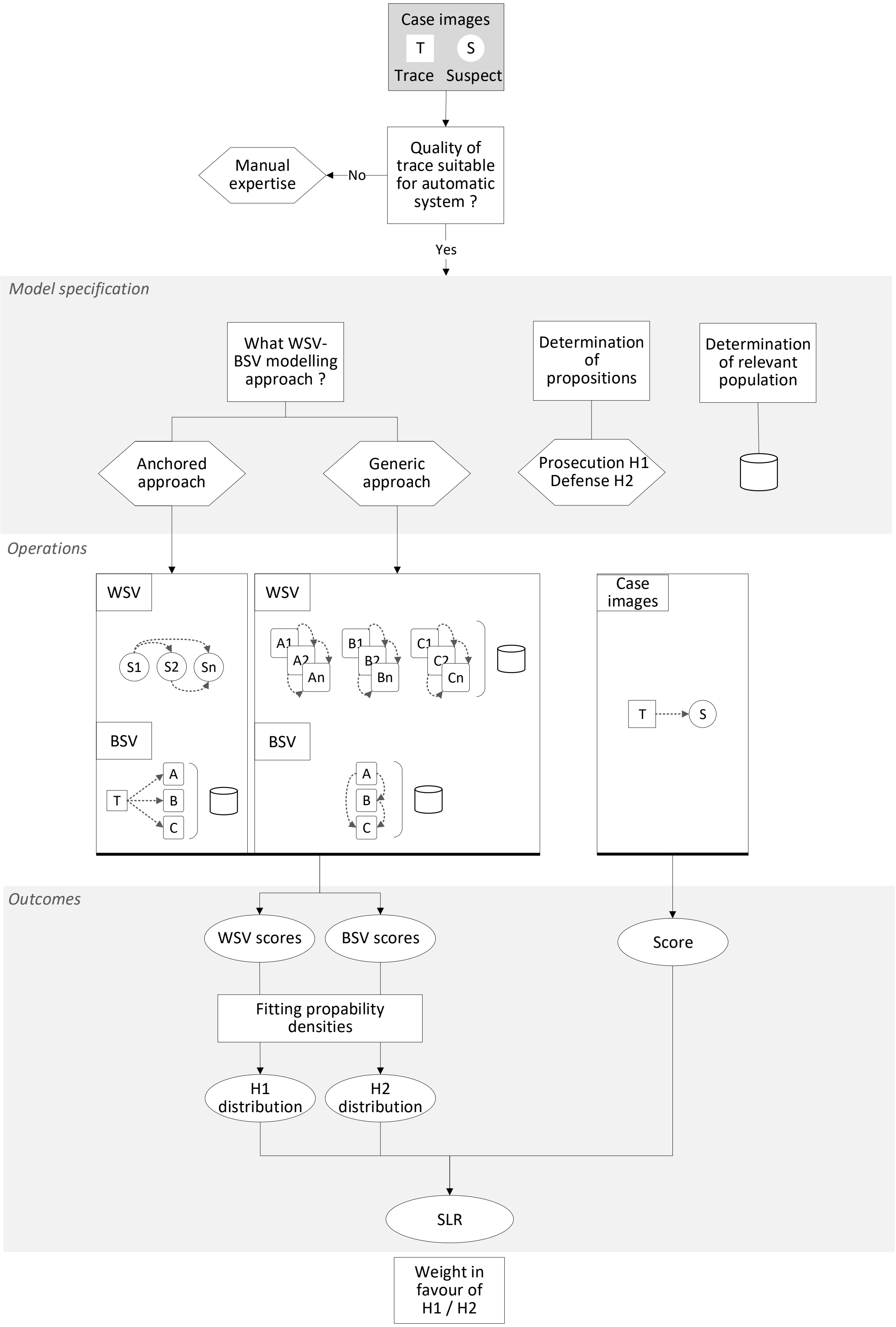

Figure 6: Workflow of the SLR assignation model development by the expert in an evaluative purpose.

d. Assessing the robustness of the model

Once the previous step achieved, the expert will have a model to assign an SLR automatically to any comparison for a query case. To make sure this SLR is reliable[6], one has to assess the performances of the model. Two main characteristics have to be addressed: the discriminating power and the calibration state of the model [35, 37]. In this section, we successively describe these two performance characteristics as well as several metrics and graphical representations most often used to visualize them. Figure 7 below gathers examples of four simulated models with distinct performances.

Discrimination

The discriminating power of a model illustrates its capacity to separate the two competing propositions clearly. See the graphical representations of the distributions in Figure 7. This characteristic is inherent to the algorithm structure, and therefore can only be improved by modifying the system code [37]. The discriminating power can be measured with metrics such as the EER (Equal Error Rate) and the minCllr (minimum Log-Likelihood-Ratio Cost) [35] as well as the False Acceptance and False Rejection rates (respectively, FAR and FRR) [44]. In the Bayesian framework, the FAR and FRR may be referred to as the Rate of Misleading Evidence in favour of the Prosecution (RMEP) and the Rate of Misleading Evidence in favour of the Defence (RMED), respectively [67]. These misleading evidence rates illustrate how often the SLR gives weight to one proposition when the second one is known to be true. These metrics are described in the following section on performance representations.

Calibration

Unlike discrimination, calibration is carried out typically at the end of the method, without having to access the system algorithm. In general terms, the calibration state of a model refers to the closeness of the computed value to the known value. In our case, however, there is no known value to aim. Therefore, the calibration measures the extent to which the SLR points towards the correct proposition. For example, the SLR computed for model 3 (Figure 7) are properly discriminating between the two propositions, but almost all of them points to H_1, even when comparing pictures of different individuals (RMEP = 96.6%). A calibration process will then rebalance error rates and minimize the costs associated with the errors of the model. The SLR values can either all be adjusted at once by applying the same shift or be modified separately, as long as their ranks in the resulting SLR list do not change [37].

Calibration has been described in general terms in [68] and addressed in the Bayesian framework subsequently (see [69] for example). The most broadly used and studied calibration method is the Pool-Adjacent Violators (PAV) algorithm [70-74], especially in forensic speech recognition. We will not detail this method in this paper; although, it is important to note that it possesses some weaknesses that may turn out to be actual limitations for use in court. Indeed, the PAV algorithm generates finite SLR values only from the data where the distributions overlap. All SLR values beyond this zone, *i.e.* below the lowest value under H_1 and above the highest value under H_2, will be equal to zero and infinity, respectively. A solution has been proposed to overcome this issue [70]. Briefly, it consists in adding dummy SLR to trick the algorithm into generated finite values even outside the overlap area. According to [70], the more data used to model the distributions the smaller the effect of the dummy scores, but this remains an additional critical point to address in a forensic expertise. Consequently, another calibration method, the Logistic Regression [44, 71, 75], may be preferred, as it produces calibrated SLR exclusively on a finite interval.

[6] Reliable is used for a legal appreciation of overall robustness of a model, combining correctness, precision and reproducibility.

It is important to note that the term "calibration" is used in the literature to describe two (not so) distinct processes. It often refers to "score-to-SLR calibration", *i.e.* the SLR computation step, as review in [66], or it can more specifically refer to the subsequent process, the SLR calibration, as in [76], when needed for models with high discriminating power but poorly calibrated (e.g. Figure 7 model 3). As this second step is essential to enhance the overall performance of a model, we think that the term "calibration" should not differentiate between these two steps (score-to-SLR then SLR-to-calibrated SLR) but should cover every computation used from the initial score to the final reported SLR, regardless of the number of treatment needed.

Representation of performance

This section gives an overview of the main metrics and graphical representations used to assess the robustness of a SLR computation model. These tools are more thoroughly described in [73] through examples of application for LR methods from physicochemical data, and in [35] for the validation process of automatic-system-based models.

Detection Error Trade-off plot and Equal Error Rate

The Detection Error Trade-off (DET) is sufficient to compare the discriminating power of several models, but does not provide information on their calibration state. The thresholds of LLR values are illustrated as a function of the model FAR and FRR [77]. Therefore, the closer the curve is to the origin of the axis, the higher the discriminating power is. We can read the value of the Equal Error Rate (EER) at the intersection of the model curve with the diagonal line passing through the origin (see Figure 7).

Tippett plot

A Tippett plot [78-80] represents the curves of the cumulative LLR distributions under each proposition. The RMEP and RMED associated with a model are read at the intersection of the curves (one for each proposition) with the vertical axis at LLR=0 (see Figure 7b). A model is the more accurate when the two distributions are clearly discriminated, *i.e.* the cumulative curves are far apart from each other. In addition, a model calibration state may be implicitly observed by the separation of both curves on either side of the LLR=0 axis. For instance, in Figure 7, models 1 and 3 both efficiently separate the two propositions although the model 3 mostly points to H_1, even when comparing images of different persons, which shows the need for a subsequent calibration for this model. In comparison, model 2 is less efficient than the models 1 and 3, but applying calibration would not increase its accuracy without distorting the model results. It is to be noted that these models are used here because they illustrate easily the main purposes of the representation. Models with more complex results may complicate the reading of some information. See [80] for more details of the limitations of Tippett plots.

Empirical Cross-Entropy and Cost of log-likelihood-ratio

Proposed in [44], the Empirical Cross-Entropy (ECE) curves represent the costs associated with the SLR values generated by a system. The higher the maximum value, the more information needed for the SLR to point to the correct proposition. Henceforth, a high-performance model has minimal losses when making errors. On an ECE curve, the Cllr metric [70], representing the average cost associated with a model, is represented as the ECE value at the intersection of Logit($P(H_1)$)=0 and the empirical model curve. We can see this in Figure 7, where the dashed curve represents the empirical model (before calibration) and the dotted curve shows the calibrated model. The minCllr is the cost associated with the calibrated model and directly quantify its discriminative power. Table 2 below summarizes the use and adequacy (shown by '-', '+' or '++') of the different metrics and graphical representations in the assessment of a model performance.

Performance characteristics	**Metric**	**Graphical representation**			
		Distributions	DET curve	Tippett plot	ECE plot
Discrimination	RMEP-RMED, EER, minCllr	+	++	++	++
Calibration	calCllr	+	-	+	++

Table 2: Metrics and graphical representations used in the validation process of SLR computation models. The degree of adaptability of each representation method to visualize both performance characteristic is described as such: (-) not adapted, (+) limitations depending on the case, (++) adapted representation

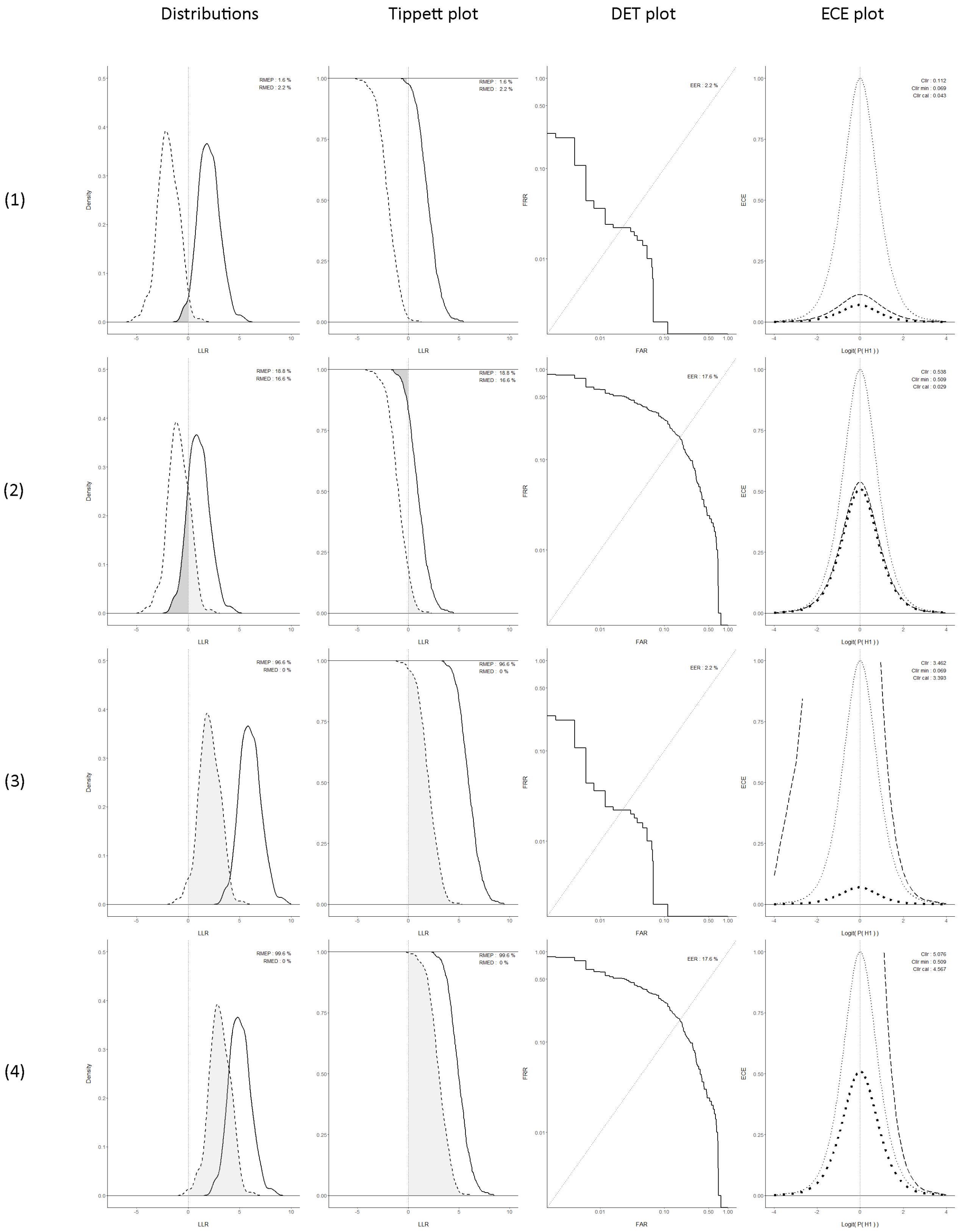

Figure 7: Representations of the LLR distribution, Tippett plots, DET plots and ECE plots for four hypothetical models: (1) a high-performance and calibrated model (2) a well calibrated but with lower performance model (3) a high-performance but poorly-calibrated model and (4) a model with both low performance and calibration state.

Conclusion and perspectives

In this paper, we have assessed from the current literature the state of the art of forensic face recognition using automatic systems. Zeinstra and colleagues in 2018 [9] offered a previous survey that covered well human-based face recognition for forensic purposes and available datasets, and we wanted to concentrate on automatic techniques such as the use of biometrics systems. We identified the framework of applications depending on the forensic scenarios (CCTV, witness images, official documents) and the usage of such information either for intelligence, investigative or evaluative purposes (Figure 1). We focused then on the various ways to compute a score-based likelihood ratio (SLR). Through that process, we clarified the concept of anchoring (Figure 4, Figure 5) and the impact on the phrasing of the propositions considered in a forensic case (Table 1). We conclude that a suspect-anchored is more relevant but also identified that the respective benefits and limitations of each approach remain unknown. That review allowed us to present a global workflow that illustrates the steps that ought to be taken to compute a SLR (Figure 6). The robustness of such SLR should be measured. The main performance indicators have been reviewed and their metrics explained in the context of measuring discriminative power and calibration of the model (Table 2).

Another research avenue that we introduced in this paper relates to the demarcation between investigation and evaluation purposes. Currently, during a police investigation, the preliminary results are mostly communicated in an informal manner to guide the investigators. These results may be a list of potential suspects or the brief comparison of the trace with a POI picture, leading to an informal "match" or "non-match" guiding the investigation. However, we think that it would be beneficial to provide investigators with probabilistic and weighted results during the investigation phase, as a "preliminary evaluation". This evaluation will be based on the Bayesian framework, through the computation of a SLR from a similarity score. During the investigation phase, the only data available are the query image and pictures of one or more potential suspects. From that, each "trace *vs* potential suspect" score is generated to compute a preliminary "investigative SLR", along with an estimation of what SLR value should be expected after a thorough evaluation with all the needed data in the instruction phase (*i.e.* complete reference material and an adequate database as relevant population). Obviously, this estimation has to be supported by empirical data. A study has then to simulate several cases where experts are provided with 1) a trace and a list of potential suspects from a database (or from an eyewitness) during the investigation phase, and 2) the adequate training datasets and propositions defined in the instruction phase. The questions to be answered in the investigation step are "What is the weight of the preliminary result? How and how much would this weight vary with adequate supplementary data? What if more reference images from this suspect cannot be provided?". Both preliminary-SLR and final SLR are compared in each case to assess the variation of their values as well as the factors of these variations depending on the provided data. The generic WSV approach may be a good asset for the preliminary-evaluation. It would allow computing an investigative preliminary-SLR from a generic denominator for a given population, and that would be usable for every case that shares the same population. This is part of our future work and should be beneficial not only for face recognition but also to many other forensic fields. In such purpose, it is essential to improve the discussion between forensic expert, investigators and legal practitioners to best develop this method with respect to the needs and constraints of each.

Bibliography

[1] C. Peacock, A. Goode, and A. Brett, "Automatic forensic face recognition from digital images", *Science & Justice,* 44(1), pp. 29-34, 2004.

[2] Q. Rossy, S. Ioset, D. Dessimoz, and O. Ribaux, "Integrating forensic information in a crime intelligence database", *Forensic Science International,* 230(1-3), pp. 137-146, 2013.

[3] D. Meuwly, "Reconnaissance de locuteurs en sciences forensiques: L'apport d'une approche automatique", PhD Thesis, Université de Lausanne, Ecole des Sciences Criminelles, Suisse, 2001.

[4] F. Botti, A. Alexander, and A. Drygajlo, "An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data", *ODYSSEY04 - The Speaker and Language Recognition Workshop Toledo, Spain*, pp. 63-68, 2004.

[5] N. Egli, "Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System", PhD Thesis, Université de Lausanne, Ecole des Sciences Criminelles, Suisse, 2009.

[6] S. Z. Li and A. K. Jain, Eds. "Encyclopedia of Biometric". New-York, USA: Springer Science+Business Media, LLC, 2009.

[7] M. Tistarelli and C. Champod, Eds. "Handbook of Biometrics for Forensic Science", Advances in Computer Vision and Pattern Recognition. Cham, Switzerland: Springer International Publishing AG, 2017.

[8] D. Dessimoz and C. Champod, "A dedicated framework for weak biometrics in forensic science for investigation and intelligence purposes: The case of facial information", *Security Journal,* 29(4), pp. 603-617, 2015.

[9] C. G. Zeinstra, D. Meuwly, A. C. C. Ruifrok, R. N. J. Veldhuis, and L. J. Spreeuwers, "Forensic face recognition as a means to determine strength of evidence: a survey", *Forensic Science Review,* 30(23), pp. 23-34, 2018.

[10] O. Ribaux, *Police scientifique: le renseignement par la trace*, Lausanne: PPUR Presses polytechniques et Universitaires Romandes, 2014.

[11] Q. Milliet, O. Delemont, and P. Margot, "A forensic science perspective on the role of images in crime investigation and reconstruction", *Science & Justice,* 54(6), pp. 470-80, 2014.

[12] G. Jackson, S. Jones, G. Booth, C. Champod, and I. W. Evett, "The nature of forensic science opinion - a possible framework to guide thinking and practicce in investigation and in court proceedings", *Science & Justice,* 46(1), pp. 33-44, 2006.

[13] O. Ribaux *et al.*, "Intelligence-led crime scene processing. Part I: Forensic intelligence", *Forensic Science International,* 195(1-3), pp. 10-16, 2010.

[14] O. Ribaux, A. Girod, S. J. Walsh, P. Margot, S. Mizrahi, and V. Clivaz, "Forensic intelligence and crime analysis", *Law, Probability and Risk,* 2, pp. 47-60, 2003.

[15] O. Ribaux, S. J. Walsh, and P. Margot, "The contribution of forensic science to crime analysis and investigation: forensic intelligence", *Forensic Science International,* 156(2-3), pp. 171-181, 2006.

[16] T. Ali, R. Veldhuis, and L. Spreeuwers, "Forensic face recognition: A survey", in *Face Recognition: Methods, Applications and Technology*, A. Quaglia and C. M. Epifano, Eds. (Computer Science, Technology and Applications), Nova Publishers, pp. 9-28, 2012.

[17] G. Porter and G. Doran, "An anatomical and photographic technique for forensic facial identification", *Forensic Science International,* 114(2), pp. 97-105, 2000.

[18] R. Moreton and J. Morley, "Investigation into the use of photoanthropometry in facial image comparison", *Forensic Sci Int,* 212(1-3), pp. 231-7, 2011.

[19] European Network of Forensic Science Institutes, "Best practice manual for facial image comparison (ENFSI-BPM-DI-01)", 2018. Available: http://enfsi.eu/documents/best-practice-manuals/.

[20] Facial Identification Scientific Working Group, "Guidelines for Facial Comparison Methods", 2012. Available: https://fiswg.org/documents.html.

[21] Forensic Science Regulator, "Forensic image comparison and interpretation evidence : guidance for prosecutors and investigators", issue 2, 2016. Available: https://www.gov.uk/government/publications/.

[22] Y. Peng, "Face recognition at a distance: Low-resolution and alignment problems", PhD Thesis, Digital Society Institute, University of Twente, Enschede, The Netherlands, 2019.

[23] P. Grother and M. Ngan, "Face Recognition Vendor Test (FRVT): Performance of face identification algorithms", National Institute of Standards and Technology (NIST) Interagency Report 8009, 2014.

[24] P. Grother, M. Ngan, and K. Hanaoka, "Face Recognition Vendor Test (FRVT) Part 2: Identification", National Institute of Standards and Technology - NISTIR 8238, 2019.

[25] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification", *Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815-823, 2015.

[27] L. Best-Rowden, S. Bisht, J. C. Klontz, and A. K. Jain, "Unconstrained face recognition: establishing baseline human performance via crowdsourcing", *IEEE Sixth International Conference on Biometrics : Theory, Applications and Systems (BTAS)*, pp. 1-8, 2013.

[28] A. Jain, P. Flynn, and A. Ross, Eds. "Handbook of Biometrics". New-York, USA: Springer Science+Business Media, LLC, 2007.

[29] G. Edmond, K. Biber, R. Kemp, and G. Porter, "Law's Looking Glass: Expert Identification Evidence Derived from Photographic and Video Images", *Current Issues in Criminal Justice,* 20(3), pp. 337-377, 2009.

[30] E. J. Imwinkelried, "Computer source code: A source of the growing controversy over the reliability of automated forensic techniques", *66 DePaul Law Review,* 66(1), pp. 97-132, 2017.

[31] Z. Geradts, "Digital, big data and computational forensics", *Forensic Sciences Research,* 3(3), pp. 179-182, 2018.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature,* 521(7553), pp. 436-44, 2015.

[33] A. J. O'Toole, C. D. Castillo, C. J. Parde, M. Q. Hill, and R. Chellappa, "Face Space Representations in Deep Convolutional Neural Networks", *Trends in Cognitive Sciences,* 22(9), pp. 794-809, 2018.

[34] C. J. Parde, Y. Hu, C. Castillo, S. Sankaranarayanan, and A. J. O'Toole, "Social Trait Information in Deep Convolutional Neural Networks Trained for Face Identification", *Cogn Sci,* 43(6), p. e12729, 2019.

[35] D. Meuwly, D. Ramos, and R. Haraksim, "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation", *Forensic Science International,* 276, pp. 142-153, 2017.

[36] C. G. Aitken, F. Taroni, and J. Wiley, *Statistics and the evaluation of evidence for forensic scientists*, 2nd ed. (Statistics in Practice), Chichester: John Wiley & Sons, Ltd., 2004.

[37] B. Robertson, G. A. Vignaux, and C. E. H. Berger, *Interpreting Evidence – Evaluating Forensic Science in the Courtroom*, 2nd ed., Chichester: John Wiley & Sons, Ltd, 2016.

[38] C. Champod and I. W. Evett, "A probabilistic approach to fingerprint evidence", *Journal of Forensic Identification,* 51(2), pp. 101-122, 2001.

[39] I. W. Evett, J. A. Lambert, and J. S. Buckleton, "A Bayesian approach to interpreting footwear marks in forensic casework", *Science & Justice,* 38(4), pp. 241-247, 1998.

[40] S. R. Lewis, "Philosophy of Speaker Identification", *Police Applications of Speech and Tape Recording Analysis – Proceeding of the Institute of Acoustics,* 6(1), pp. 69-77, 1984.

[41] I. W. Evett and B. S. Weir, *Interpreting DNA Evidence – Statistical Genetics for Forensic Scientists*, Sunderland: Sinauer Associates, Inc., 1998.

[42] I. W. Evett, "Forensic Handwriting Comparison, Probability and the Nature of the Science", *Joint Meeting of the European Conferences for Police and Government Handwriting Experts and Document Experts*, Kincardine-on-Forth (Scotland), 1998.

[43] S. Bunch and G. Wevers, "Application of likelihood ratios for firearm and toolmark analysis", *Science & Justice,* 53(2), pp. 223-229, 2013.

[44] D. Ramos-Castro, "Forensic evaluation of the evidence using automatic speaker recognition systems", PhD Thesis, Universidad Autónoma de Madrid, Escuela Politécnica Superior, España, 2007.

[45] T. Ali, "Biometric Score Calibration for Forensic Face Recognition", PhD Thesis, University of Twente, The Netherlands, 2014.

[46] A. Macarulla Rodriguez, Z. Geradts, and M. Worring, "Validation of Score-based Likelihood Ratio Estimation for Automated Face Recognition", *20th Irish Machine Vision and Image Processing conference*, Belfast, Northern Ireland, pp. 145-153, 2018.

[47] A. Bolck, H. Ni, and M. Lopatka, "Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison", *Law, Probability and Risk,* 14(3), pp. 243-266, 2015.

[48] L. J. Davis, C. P. Saunders, A. Hepler, and J. Buscaglia, "Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios", *Forensic Science International,* 216(1-3), pp. 146-157, 2012.

[49] D. A. Schum, *Evidential Foundations of Probabilistic Reasoning* (Wiley Series in Systems Engineering), New York: John Wiley and Sons, 1994.

[50] I. De March and F. Taroni, "Bayesian networks and dissonant items of evidence: A case study", *Forensic Science International: Genetics,* 44, 2020.

[51] P. Juchli, "Combining evidence", PhD thesis, School of Criminal Justice, University of Lausanne, Switzerland, 2016.

[52] N. Susyanto, R. Veldhuis, L. Spreeuwers, and C. Klaassen, "Semiparametric likelihood-ratio-based biometric score level fusion via parametric copula", *IET Biometrics,* 8(4), pp. 277-283, 2019.

[53] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 30(2), pp. 342-347, 2008.

[54] D. Meuwly, "Forensic Individualisation from Biometric Data", *Science & Justice,* 46(4), pp. 205-213, 2006.

[55] A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia, "Score-based likelihood ratios for handwriting evidence", *Forensic Science International,* 219(1-3), pp. 129-140, 2012.

[56] G. S. Morrison, "Calculation of forensic likelihood ratios: use of Monte Carlo simulations to compare the output of scrore-based approaches with true likelihood-ratio values", arXiv e-prints, 2016. Available: http://geoff-morrison.net.

[57] C. Neumann and P. Margot, "New perspectives in the use of ink evidence in forensic science: Part III: Operational applications and evaluation", *Forensic Science International,* 192(1-3), pp. 29-42, 2009.

[58] C. Neumann, C. Champod, M. Yoo, T. Genessay, and G. Langenburg, "Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks", *Forensic Science International,* 248, pp. 154-171, 2015.

[59] D. Ramos, J. Maroñas-Molano, and A. Lozano-Diez, "Bayesian Strategies for Likelihood Ratio Computation in Forensic Voice Comparison with Automatic Systems", *Subsidia 2017: Tools and Resources for Speech Sciences*, Malaga, Spain, 2018.

[60] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications", *IEEE Multimedia,* 19(1), pp. 2-10, 2012.

[61] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4873-4882, 2016.

[62] T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly, "Effect of calibration data on forensic likelihood ratio from a face recognition system", *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1-8, 2013.

[63] P. J. Phillips *et al.*, "Overview of the Face Recognition Grand Challenge", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 947-954, 2005.

[64] I. W. Evett, G. Jackson, J. A. Lambert, and S. McCrossan, "The Impact of the Principles of Evidence Interpretation on the Structure and Content of Statements", *Science & Justice,* 40(4), pp. 233-239, 2000.

[65] C. Champod and I. W. Evett, "Evidence Interpretation: a Logical Approach", in *Wiley Encyclopedia of Forensic Science*, vol. 2, A. Moenssens and A. Jamieson, Eds.), Chichester, UK: John Wiley & Sons, pp. 968- 976, 2009.

[66] T. Ali, L. Spreeuwers, and R. Veldhuis, "A review of calibration methods for biometric systems in forensic applications", *33rd WIC Symposium on Information Theory in the Benelux*, 2012.

[67] C. Neumann *et al.*, "Computation of likelihood ratios in fingerprint identification for configurations of three minutiae", *Journal of Forensic Sciences,* 51(6), pp. 1255-1266, 2006.

[68] D. V. Lindley, A. Tvesky, and R. V. Brown, "On the Reconciliation of Probability Assessments", *Journal of the Royal Statistical Society. Series A,* 142(2), pp. 146-180, 1979.

[69] M. H. DeGroot and S. E. Fienberg, "The Comparison and Evaluation of Forecasters", *The Statistician,* 32, pp. 12-22, 1983.

[70] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection", *Computer Speech & Language,* 20(2-3), pp. 230-275, 2006.

[71] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech", PhD Thesis, Department of Electrical and Electronic Engineering, University of Stellenbosch, South Africa, Matieland, 2010.

[72] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*, Frankfurt: Verlag für Polizeiwissenshaften, 2016.

[73] G. Zadora, A. Martyna, D. Ramos, and C. Aitken, J. Wiley, Ed. *Statistical analysis in forensic science: Evidential value of multivariate physicochemical data*, United Kingdom: John Wiley & Sons, Ltd, 2014.

[74] D. Ramos and J. Gonzalez-Rodriguez, "Reliable support: Measuring calibration of likelihood ratios", *Forensic Science International,* 230(1-3), pp. 156-169, 2013.

[75] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio", *Australian Journal of Forensic Sciences,* 45(2), pp. 173-197, 2013.

[76] D. Ramos, J. Gonzales-Rodriguez, G. Zadora, J. Zieba-Palus, and C. Aitken, "Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation", *Third International Symposium on Information Assurance and Security*, Manchester, UK: IEEE-CS Press, 2007.

[77] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance", *Eurospeech 1997*, 4, pp. 1895-1898, 1997.

[78] C. F. Tippett *et al.*, "The Evidential Value of the Comparison of Paint Flakes from Sources other than Vehicles", *Journal of the Forensic Science Society,* 8(2-4), pp. 61-65, 1968.

[79] I. W. Evett and J. S. Buckleton, "Statistical analysis of STR data", *16th Congress of the International Society for Forensic Haemogenetics*, 6, pp. 79-86, 1996.

[80] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken, "Information-theoretical assessment of the performance of likelihood ratio computation methods", *Journal of Forensic Sciences,* 58(6), pp. 1503-1518, 2013.