Information Fusion 111 (2024) 102489

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Evidence evaluation in craniofacial superimposition using likelihood ratios

Práxedes Martínez-Moreno [a,b,*], Andrea Valsecchi [c], Pablo Mesejo [a,b], Óscar Ibáñez [d], Sergio Damas [e,b]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, 18071, Spain
[b] Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, Granada, 18016, Spain
[c] Panacea Cooperative Research S. Coop., Ponferrada, 24402, Spain
[d] Faculty of Computer Science, CITIC, University of A Coruña, A Coruña, 15008, Spain
[e] Department of Software Engineering, University of Granada, Granada, 18071, Spain

ARTICLE INFO

Keywords:
Skeleton-based forensic human identification
Craniofacial superimposition
Likelihood ratio
Decision support system

ABSTRACT

Craniofacial Superimposition is a forensic identification technique that supports decision-making when skeletal remains are involved. It is based on the analysis of the overlapping of a post-mortem skull with ante-mortem facial photographs. Despite its importance and wide applicability, the process remains complex and challenging. To address this, computerized methods have been proposed, but subjectivity and qualitative reporting persist in decision-making. This study introduces an evidence evaluation system proposal based on Likelihood Ratios, previously used in other forensic fields, such as DNA, voice, fingerprint, and facial comparison. We present a novel application of this framework to Craniofacial Superimposition. Our work comprises three experiments in which our LR system is trained and tested under distinct conditions concerning facial images: the first utilizes frontal facial photographs; the second employs lateral facial photographs; and the last one integrates both frontal and lateral facial photographs. In the three experiments, the proposed LR system stands out in terms of calibration and discriminating power, providing practitioners with a quantitative tool for evidence evaluation and integration. However, the lack of massive actual data obliged us to focus our study on synthetic data only. Therefore, it should be considered a proof of concept. Nevertheless, the resulting likelihood-ratio system offers objective decision support in Craniofacial Superimposition. Further studies are required to validate in a real scenario the conclusions achieved.

1. Introduction

Post-mortem human identification through the analysis of skeletal remains is a significant challenge in the field of Forensic Anthropology [1]. One of the techniques that assist in this identification process is Craniofacial Superimposition [2]. This technique involves the comparison of an ante-mortem facial photograph with a recovered post-mortem skull, by projecting the latter onto the former. This provides forensic practitioners with information to determine whether the skull and face correspond to the same individual. Three stages can be distinguished in the Craniofacial Superimposition process [3,4]:

1. Acquisition and processing of the post-mortem skull and ante-mortem facial photographs, together with the localization of anatomical landmarks that usually guide Craniofacial Superimposition. There are two types of landmarks: craniometric, which are specific points of relevance in the morphology of the skull; and cephalometric, which are homologous points but located on the surface of the subject's face. The correspondence between a cranial and a facial landmark is not exact due to the existence of soft tissue, which surrounds and protects structures and organs of the body. Fig. 1 shows the location of the landmarks used in this study, which are described in Section 4.1.
2. Skull-Face Overlay, which focuses on achieving the best possible projection of a skull (or a 3D model of it) onto a single ante-mortem image of the subject, by matching the corresponding cranial and facial landmarks. If more than one photograph is available, several independent Skull-Face Overlay processes will be applied to obtain different overlays.
3. Decision-making, carried out by the practitioner, who is responsible for determining the degree of anatomical correspondence according to the Skull-Face Overlay(s) obtained in the previous stage, in order to conclude whether the skull and facial photograph belong to the same person. A scale for craniofacial matching evaluation was established by leading experts in

* Corresponding author at: Department of Computer Science and Artificial Intelligence, University of Granada, Granada, 18071, Spain.
E-mail addresses: praxedesmm@ugr.es (P. Martínez-Moreno), valsecchi.andrea@gmail.com (A. Valsecchi), pmesejo@ugr.es (P. Mesejo), oscar.ibanez@udc.es (Ó. Ibáñez), sdamas@ugr.es (S. Damas).

https://doi.org/10.1016/j.inffus.2024.102489
Received 5 October 2023; Received in revised form 17 May 2024; Accepted 18 May 2024
Available online 22 May 2024

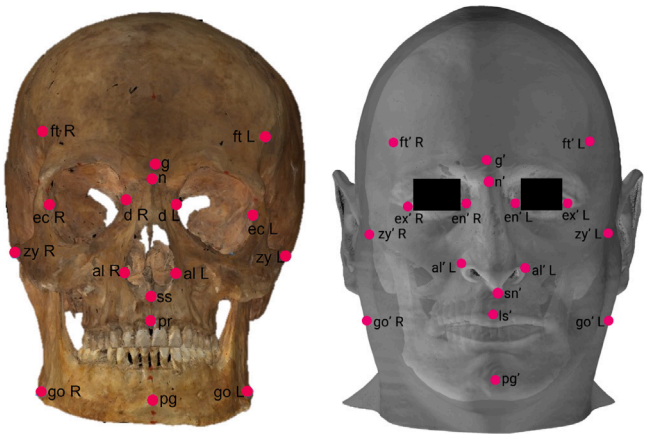

Fig. 1. Localization of the cranial (left) and facial (right) landmarks involved in this study.

craniofacial identification within the MEPROCS project [5,6]. The final decision is presented in terms of strong, moderate, or limited support.

Traditionally, Craniofacial Superimposition has lacked a standardized methodology and the MEPROCS[1] European project represented an attempt to tackle this issue. [The main achievements and best practice recommendations of MEPROCS were published in [5]. Furthermore, the three stages are performed manually in most forensic laboratories. This is a difficult and time-consuming task, even for a trained professional, and the results are often considered subjectively biased to a non-negligible degree. The decision-making process is subject to variability, hinges on the proficiency of the forensic examiner, and is influenced by the quality and quantity of the materials used, including the photographs and the skull. Moreover, it is worth noting that the resultant decision is communicated in a qualitative manner by assigning one of the three degrees of support, stated by the practitioner. Therefore, there is no possibility of statistical interpretation, leading to a subjective inference process. For these reasons, automatic systems supporting the Craniofacial Superimposition technique have become relevant during the last decade.

The computerized systems developed for the first stage of Craniofacial Superimposition are mainly related to cephalometric and craniometric landmark location. This problem has been addressed in the Computer Vision and Deep Learning communities. However, the sets of landmarks used in Deep-Learning-based works are not the same as those used in Anthropology. We can only find two publications that focus on the automatic localization of cephalometric landmarks in photographs [7,8], and one tackling craniometric landmarks in skull 3D models [9].

Several proposals have been presented in literature to perform the second stage, the Skull-Face Overlay task. The most natural way to deal with the Skull-Face Overlay problem is to replicate the original scenario of the ante-mortem photograph in which the living person was in a given pose somewhere inside the camera's field of view. The first computer-aided approach for the Skull-Face Overlay task was proposed by Nickerson et al. [10]. In the last decade, most of the works commonly tackle Skull-Face Overlay automation using evolutionary algorithms and fuzzy sets [11,12]. These methods solved the projection problem by an iterative optimization process, where multiple solutions are evaluated at every step, eventually converging to a high-quality solution. Unlike these methods, Posest-SFO [13] solves a system of polynomial equations relating the distances between the points before and after the projection. This approach solves two problems that arise during Craniofacial Superimposition: "camera calibration" and "perspective from N points". The former involves estimating the internal parameters of the camera that captured the input facial photograph(s), while the latter involves determining the pose of the calibrated camera given the positions of n 3D points and their corresponding 2D positions in the photograph(s). This method has been demonstrated to be both extremely fast and significantly accurate. However, contrary to previous publications [11,14], it does not address the sources of uncertainty, i.e., the articulation of the mandible, the estimation of the soft tissue thickness, and the intra- and inter-error in landmark location. It is clear then that there is still a large margin for improvement. Nevertheless, Posest-SFO is the state-of-the-art and thus the one employed in this work.

Finally, there are just a few works tackling the automation of the analysis of craniofacial correspondences within the framework of Craniofacial Superimposition identification [15,16]. Most of the existing literature was published more than 20 years ago and the works are extremely basic and limited. Recently, in [17–19], the authors presented a hierarchical system for the decision-making stage and Computer Vision algorithms to evaluate the anatomical consistency of morphological criteria between the face and the skull. This means that from a series of Skull-Face Overlays of the same individual, the decision support system provides the forensic expert with a quantitative output value indicative of the morphological matching consistency of a given Craniofacial Superimposition problem. However, there are some limitations concerning the proposed decision support system. The final degree of craniofacial correspondence is determined by aggregating multiple matching degrees obtained from various aggregations across the hierarchical structure's different levels (criterion evaluation, Skull-Face Overlay evaluation, and Craniofacial Superimposition evaluation). Therefore, this output value does not provide direct and interpretable conclusions about the case under discussion. In addition, a threshold needs to be selected by the forensic practitioner in order to analyze the mentioned output value. As a consequence, this process gives rise to a decision-making stage that is subjective, lacks robustness, and yields less meaningful outcomes. To address these limitations, it is imperative to develop a decision support system capable of providing quantitative, objective, and informative assistance to practitioners in their assessment of evidence. Furthermore, such a system could facilitate the presentation of findings in a court of law in a more robust and defensible manner.

A commonly accepted manner for conveying the strength of evidence in forensic science and legal proceedings is through Likelihood Ratios (LRs) [20]. This was recommended by the European Network of Forensic Science Institutes (ENFSI[2]) in a guideline for the expression of conclusions in forensic evaluation reports [21], where an LR framework is [suggested] for all forensic disciplines and laboratories. The LR framework has been widely applied and validated in the field of forensic-voice comparison [22,23], and it has also been employed in other forensic disciplines, such as facial image comparison [24], fingerprint analysis [25], handwriting [26], glass evidence [27], DNA [28–30], among others. Even though there was a lack of proposals to apply LR in Forensic Anthropology, there are recent applications of LR to both age and sex estimation problems [31–37].

The main contributions of our study are related to the application, implementation, and validation of the LR framework in the identification of skeletal remains using the Craniofacial Superimposition technique. This application involves an innovative methodology and utilizes a method to simulate Skull-Face Overlay data. In addition, our work encompasses three distinct experiments, each implies training and testing the proposed LR system under specific scenarios with facial images: the first experiment involves frontal facial photographs; the

[1] https://cordis.europa.eu/project/id/285624/es

[2] https://enfsi.eu

second makes use of lateral facial photographs; and the third one combines both frontal and lateral facial photographs. Consequently, the LR system is presented as a prototype demonstration since the data employed to train and validate them were synthetically generated. The primary goal is to implement a decision-making support system to provide practitioners with an objective and quantitative means of drawing conclusions that can be presented in court. Moreover, an LR-based system may be utilized to evaluate and validate the effectiveness of the identification method employed (Posest-SFO in this proposal). This can be accomplished by transforming the information extracted by the method into corresponding LR values and subjecting them to evaluation. Through this process, the reliability and accuracy of the identification method can be assessed, thereby enhancing confidence in the conclusions drawn from the results obtained.

The paper is organized as follows: the LR fundamentals and the state of the art are summarized in Section 2, including different methodologies for applying and validating the LR framework. The application of the latter into the Craniofacial Superimposition technique is proposed in Section 3. Section 4 explains the process followed to evaluate the LR system through three different experiments. For this purpose, the methodology, performance characteristics and metrics, and dataset used to do so are detailed. Finally, the conclusions are comprised in Section 5.

2. LR fundamentals and state of the art

This section provides an overview of the LR framework application in forensic sciences (Section 2.1). Finally, different performance characteristics and metrics to validate an LR system are detailed in Section 2.2.

2.1. Evidence evaluation using likelihood ratios

An LR is characterized as the ratio of two probabilities,[3] associated with the observation of some evidence (E) under competing hypotheses (H_0 and H_1). Specifically, it represents the probability of E if H_0 were true divided by the probability of E if H_1 were true, both conditioned on Background Information (I):

$$LR = \frac{p\left(E \mid H_0, I\right)}{p\left(E \mid H_1, I\right)} \tag{1}$$

E is defined as the information extracted from the available material in a legal procedure, which involves the objects of comparison in a forensic case. Background Information I may include information such as police investigations, witness testimony of the case, or the analysis of other forensic evidence, including speech, glass fragments, or fingerprints, among others [20,38,39]. The forensic practitioner should not, however, be exposed to task-irrelevant information [40,41]. Note that I is usually removed from the notation for simplification purposes.

The unobserved variable of interest in a forensic case is the true hypothesis (H_0 or H_1), which is the information the fact finder wants to know. H_0 and H_1 propositions are mutually exclusive. The LR enables updating the initial belief about the relative validity of the two hypotheses based on the observed evidence. This is accomplished using the odds form of Bayes' Theorem [42,43]:

$$\underbrace{\frac{p\left(H_0 \mid E\right)}{p\left(H_1 \mid E\right)}}_{\text{posterior odds}} = \underbrace{\frac{p\left(E \mid H_0\right)}{p\left(E \mid H_1\right)}}_{\text{LR}} \times \underbrace{\frac{p\left(H_0\right)}{p\left(H_1\right)}}_{\text{prior odds}} \tag{2}$$

The prior odds are the province of the fact finder, while the calculation of the LR is the province of the forensic scientist. The posterior odds are then used by the fact finder to make a decision. Note that these statements are purely illustrative, demonstrating the theoretical application of the LR framework.

Once the LR is calculated, it can be interpreted as follows: if $LR > 1$, then the evidence will support H_0; if $LR < 1$, then the evidence will support H_1; and if $LR = 1$, the evidence will support both hypotheses. The magnitude of the LR is paramount in the interpretation of evidence. A larger LR, when it is over 1, indicates stronger support for H_0. Conversely, when the LR is below 1 and approaches 0, it signifies stronger support for H_1. The closer the LR is to 1, the more limited the support is for either hypothesis.

2.2. Assessment and validation of an LR system

Aiming at using an LR system in casework, the forensic scientist should evaluate its performance. Different performance characteristics and metrics have been proposed in literature to assess the soundness of an LR system [23], such as the C_{llr} (log-likelihood-ratio cost) [44]. This metric is computed as follows:

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{H_0}} \sum_{i=1}^{N_{H_0}} \log_2\left(1 + \frac{1}{LR_{H_{0i}}}\right) + \frac{1}{N_{H_1}} \sum_{j=1}^{N_{H_1}} \log_2\left(1 + LR_{H_{1j}}\right)\right) \tag{3}$$

where N_{H_0} and N_{H_1} are respectively the number of cases where H_0 and H_1 are true in the validation set; and $LR_{H_{0i}}$ and $LR_{H_{1j}}$ are the LR values calculated respectively in those cases. The C_{llr} is a single value summary of the system performance [45]. The C_{llr} can be decomposed into *discrimination loss* or the average cost due to a lack of discrimination (C_{llr}^{disc}), and *calibration loss* or the average cost due to a lack of calibration (C_{llr}^{cal}) [22,44,45]:

$$C_{\text{llr}} = C_{\text{llr}}^{disc} + C_{\text{llr}}^{cal} \tag{4}$$

Thus, the C_{llr} is the average cost due to the lack of accuracy. For more detail, the reader is referred to [46–49]. Note that achieving the aforementioned decomposition is not trivial. The Pool Adjacent Violators (PAV) algorithm [44,47,50] is often used to address it.

Another metric also used to evaluate an LR system is the Empirical Cross-Entropy (ECE) [38,45,46,51]. The ECE is an information-theoretical measure proposed for the validation of a set of LR values, and is stated as follows:

$$\begin{aligned} ECE = {} & \frac{P(H_0)}{N_{H_0}} \sum_{i=1}^{N_{H_0}} \log_2\left(1 + \frac{1}{LR_{H_{0i}} \times O(H_0)}\right) \\ & + \frac{P(H_1)}{N_{H_1}} \sum_{j=1}^{N_{H_1}} \log_2\left(1 + LR_{H_{1j}} \times O(H_0)\right) \end{aligned} \tag{5}$$

where $P(H_0)$ and $P(H_1)$ are respectively the prior probabilities for H_0 and H_1, and $O(H_0)$ is the prior odds for H_0. The ECE is interpreted as the *information* needed on average to determine the true hypothesis for a set of LR values. The ECE is a generalization of the C_{llr}. The C_{llr} value is the same as the ECE value when both prior probabilities are fixed at 0.5, i.e., uncertainty is maximum. Therefore, the decomposition into discrimination and calibration also applies to ECE. Moreover, the ECE has another interesting interpretation in evidence evaluation: the *range of application* of the system [46]. This insight refers to the range of prior odds where the LR system is valid. Hence, the ECE is calculated for a range of prior odds and represented in an *ECE plot*, which is further detailed in Section 4.

For an LR system under evaluation, the ECE values obtained should desirably be lower than that of a *neutral method*, which always yields LR values equal to 1 [46,51]. Otherwise, the LR system performance will not be better than that of a method that does not extract any relevant information from the evidence. The C_{llr} values for well-calibrated systems range from 0 to approximately 1 [22].

Both the C_{llr} and the ECE are crucial performance metrics. However, it is recommended to complement such indicators with graphical

[3] For continuously-valued data, probability-density is assessed.

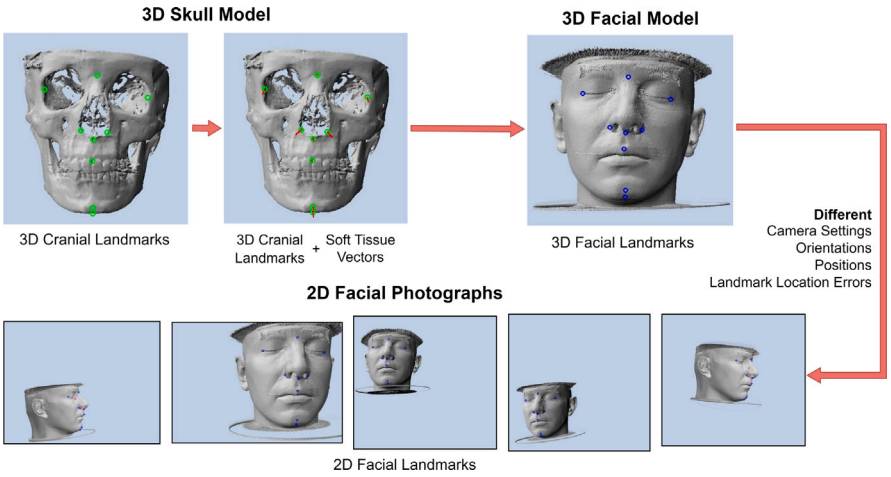

Fig. 2. Example of the Skull-Face Overlay data generation process for a specific subject.

representations, such as ECE plots or *Tippett plots* [22,23,51]. The Tippett plots are cumulative probability distribution plots expressing the proportion of LRs greater than a given value for both H_0-true and H_1-true cases. In this plot, it can be observed the rate of *misleading evidence* that leads to the calculation of LR values that support the wrong proposition [23,51]. In addition, it can also reveal problems such as bias in the output: when all the LR values are too large or too small (shift), or when the LR values are too far from the neutral value (1) or too close (scale) [22]. For more details on Tippett plots, the interested reader is referred to [23], where best practice guidelines for case evaluation and interpretation in forensic automatic and semi-automatic speaker recognition can be found.

3. Our LR methodological approach

The present section details how the LR framework is applied to Craniofacial Superimposition by making use of the state-of-the-art method for automatically solving the Skull-Face Overlay stage, Posest-SFO [13].

3.1. Materials and skull-face overlay data generation

The materials that our LR system works with are those involved in a Craniofacial Superimposition case: a 3D model of the skull to be identified (the trace object) and one or two facial photographs (one frontal and/or one lateral) of a known individual suspected to be the same of the trace skull (the reference object). In order to improve the reliability of the Craniofacial Superimposition technique, multiple photographs showing different poses were recommended [5,6,52]. However, there are scenarios where only two photos are available (which is the recommended minimum to apply Craniofacial Superimposition). It is thus proposed to improve the reliability of the first two experiments using more than one photograph but limiting that number to two so that it can be used in the worst-case scenarios. These two photographs are supposed to show different poses of the subject of the case under discussion. However, this is not a mandatory condition.

Unlike several well-established forensic methodologies, the Craniofacial Superimposition technique lacks ground-truth data, that is, a procedure that could provide a perfect Skull-Face Overlay in an objective and unquestionable manner [5]. Furthermore, due to its scarcity and incompleteness, real-world forensic data is not always available to researchers. As an alternative, controlled environments can be utilized to acquire data and generate artificial identification scenarios [53]. In [13], a method that simulates Skull-Face Overlay data was proposed, which replaces real facial photographs with generated ones. This results in a broader array of possible scenarios due to the possibility of precise manipulation of the subject's pose and the camera settings. In addition, the method allows the reproduction of inter- and intra-expert errors in the processing of input data, leading to more comprehensive and rigorous testing under realistic conditions.

The aim of this data generation process is to provide artificial identification cases that have complete and accurate information for input and solution. To create a case, the Craniofacial Superimposition materials needed are: a 3D skull model, the locations of 3D cranial landmarks, a facial photo, and the locations of the 2D facial landmarks. To know the ground-truth solution, it is also required the camera's location, orientation, and settings, as well as the soft tissue vectors at each landmark. Such soft tissue vectors are set along the skull's normal directions with a length corresponding to the mean soft tissue depth. This information at each skull landmark enables the location of the 3D facial landmarks to be known. In order to simulate facial images from a 3D face, the position, orientation, and camera settings are selected, and the ground-truth projection is calculated. The 2D facial landmarks are then computed as the 3D facial landmarks projected onto the photograph. Once the accurate location of 2D facial landmarks has been achieved, random noise can be added to simulate errors in pinpointing the 2D landmarks in the photograph in an actual casework. Fig. 2 illustrates an example of this data generation process for a specific subject, which would be repeated for each individual involved in a case.

This data generation process is used in the present study to achieve two objectives: to generate the facial photographs of the subject involved in each Craniofacial Superimposition case using different poses, and to increase the number of Skull-Face Overlays from which the scores for training and validating an LR system are drawn. The variability of such scores is enhanced by simulating inter-expert errors, as described in Section 3.3.

3.2. Definition of evidence and hypothesis proposal

One of the design decisions that should be made when building an LR system is to determine the type of evidence to be interpreted [20]. The output of commercial biometric systems, which are commonly treated as a black box and of commercial secret, may also be used as evidence [24,26,54–56]. This is our case, with the difference that the automatic Skull-Face Overlay system generating the biometric scores in question is publicly available in [13]. Hence, our proposal could be classified as a similarity-score-based[4] design. However, unlike typical

[4] Note that some deem LRs derived from scores inappropriate when the scores just reflect the similarity between the measurements on a trace and a reference object, and lack a measure of typicality [57]. Such scores are in

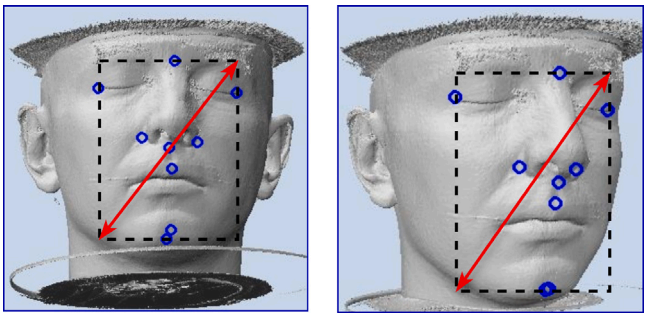

Fig. 3. Bounding boxes of the frontal and lateral facial images of Fig. 4.

score-based LR systems, the scores we obtained are not based on the measurement comparison of the trace and reference objects. They cannot be directly compared as they are of a different *nature* (i.e., skull and face), which is not the usual arrangement in literature. Instead, a biometric score (our evidence) is computed by overlapping a trace skull model with one or two reference facial photographs and then calculating a Root Mean Squared Error (RMSE) value from the resulting Skull-Face Overlay(s). For the experiment combining frontal and lateral facial images, the frontal and lateral RMSE values are added together. The RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{G}_i - G_i\right)^2}{n}} \tag{6}$$

where $\hat{G}$ is the set of 2D facial landmarks calculated in the Skull-Face Overlay stage by projecting the 3D cranial landmarks onto the facial image taking into account the soft tissue; G is the original set of 2D landmarks of the facial image, and n is the number of available 2D facial landmarks in this image. Furthermore, the biometric scores (RMSE values) are normalized using the diagonal length of the bounding box surrounding facial landmarks (see Fig. 3). This adjustment is essential because the RMSE is calculated based on pixel distances between actual facial landmarks and those estimated by the Skull-Face Overlay automatic method. Hence, when subjects are closer to the camera, there is a greater pixel distance between landmarks, leading to higher RMSE values. More detail on our biometric score can be found in Fig. 4.

The next decision to be made when building an LR system is related to the hypotheses proposal. The similarity-score-based hypotheses defined in our system are the following:

- H_0: "The subject of the reference facial photographs and the subject of the trace 3D skull model are the same".
- H_1: "The subject of the reference facial photographs and the subject of the trace 3D skull model are different. The trace skull originates from another source within the relevant population".

Depending on the specific conditions of the case, the definition of the relevant population may vary. This population typically consists of potential sources from which trace and reference objects are drawn, reflecting the forensic needs dictated by the particular circumstances of the case.

principle an incomplete representation of the evidence since, given the same score, both common and rare references may yield the same LR. However, in the absence of LR systems that model all features (and therefore rarity) well, score-based LR systems can be a good alternative since, just like feature-based LR systems, they do help improve decisions on average [20].

3.3. LR computation process

In our proposal, the LR computation process can be divided into two stages: training and LR computation. The former is conducted to adjust the system's parameters and the latter aims to compute an LR value from evidence.

Two sets of scores are obtained according to each competing hypothesis in the training stage: the S_{H_0} or *same-source* score set, which consists of scores calculated assuming that H_0 is satisfied; and the S_{H_1} or *different-source* score set, which is made up of scores calculated assuming that H_1 is satisfied. Both sets should be generated with the appropriate conditions for further comparisons: S_{H_0} and S_{H_1} should represent, as far as possible, all the sources of variability in the set in order to compute representative LR values. The probability density function (pdf) of the numerator of the LR (see Eq. (1)) would thus represent the S_{H_0} or same-source distribution, and the pdf of the denominator represents the S_{H_1} or different-source distribution [54].

The scores of S_{H_0} are calculated from Skull-Face Overlays involving pairs of facial photographs and skull models originating from the same source subject; whereas the scores of S_{H_1} are obtained from the overlays of facial photographs and skull models originating from different source subjects. Once the S_{H_0} and S_{H_1} sets of scores are obtained, their corresponding statistical distributions are found. In our case, 10 pairs of Skull-Face Overlays for each subject involved are carried out to extend the aforementioned sets of scores. This is performed by applying the Skull-Face Overlay data generation methodology explained in Section 3.1. Each Skull-Face Overlay process is repeated using different sets of 2D facial landmarks, which are simulated by introducing noise in the location of the original ones. In particular, similarly to the data generation process of [13], the noise over the 2D facial landmarks was drawn at random between −5 and +5 pixels using a uniform distribution.[5] The main purpose for this is that the sample of scores should have enough variability to represent the most significant possible number of situations that may occur. It is thus crucial to consider that the location of the landmarks involved could not be as precise as desired, and it could vary depending on the method used to perform it or even the practitioner who addresses it [9,58,59]. This methodology aims to simulate the inter-dispersion of the localization of facial landmarks by different practitioners. Note that the performance of the LR system will be dependent on the added noise.

In the LR computation stage, a biometric score is calculated using a trace skull model and one or two reference facial photographs. Then, it is transformed into an LR value. The LR computation method used in our system is KDE (Kernel Density Estimation) [60], also used in [24,25]. For more details, refer to [61]. It calculates the LR value by first modeling the pdfs of S_{H_0} and S_{H_1}. The H_0 likelihood ($p\left(E \mid H_0\right)$) and the H_1 likelihood ($p\left(E \mid H_1\right)$) are then computed for the evidence score, and the ratio of both values is the LR. We used *Silverman's rule of thumb* [62] for choosing the bandwidth of a Gaussian KDE.

4. Experiments and results

This section explains how our proposal to apply the LR framework to Craniofacial Superimposition has been validated. For this purpose, the methodology, metrics, and dataset used to perform this validation are detailed, and an analysis of the results is provided.

[5] There are publications that study and analyze the dispersion of facial landmark localization, such as [9,58]. However, there is currently no Craniofacial Superimposition method that takes this into account. Therefore, our noise modeling then serves as a first approximation to this problem.

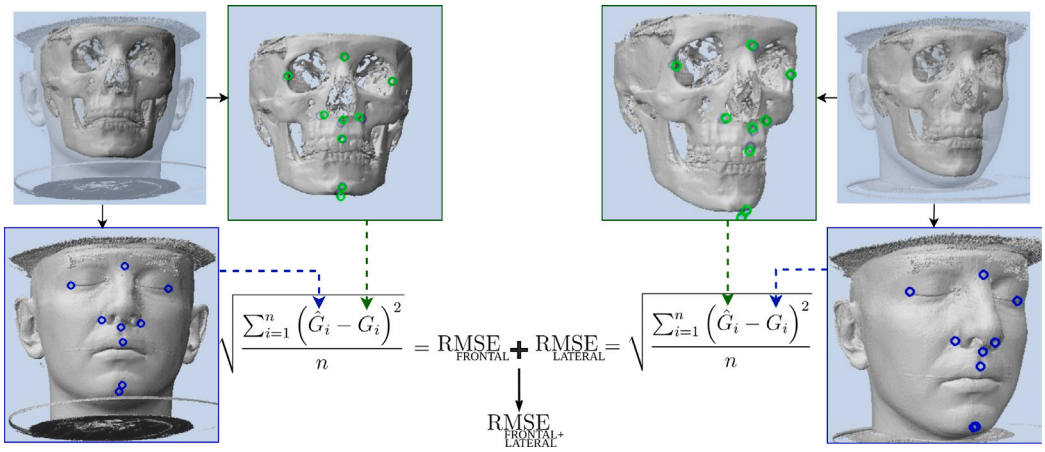

Fig. 4. The biometric scores of the frontal and lateral experiments are computed by overlapping the trace skull model with a reference facial photograph. Then, the RMSE values are calculated from the resulting Skull-Face Overlays and normalized using the diagonal length of the bounding box surrounding facial landmarks. For the frontal+lateral experiment, both RMSE values are added together.

4.1. Dataset and experimental methodology

Regarding our similarity-score-based approach, under H_0, the trace and reference objects are drawn from the same source of the relevant population; whereas, under H_1, the trace and reference objects must come from different sources of the relevant population. For more detail on the sampling process and the definition of the relevant population, the interested reader is referred to [20,22,63].

In our work, the relevant population has been defined to build a similarity-score-based LR system focused on identifying subjects from a European population. For this reason, the data was directly sourced from archives of French hospitals and medical centers during the period of 2008–2009. The subjects are of different ages and sexes, with a total of 50 males and 28 females aged between 18 and over 80 years. Further details regarding the materials can be found in [64]. Specifically, the database used consists of 78 head CT scans of the same number of subjects, each featuring no less than 6 visible landmarks. The craniometric landmarks involved are: Glabella (g), Nasion (n), Prosthion (pr), Subspinale (ss), Pogonion (pg), Left and Right Alare (al), Left and Right Dacryon (d), Left and Right Ectoconchion (ec), Left and Right Frontotemporale (ft), Left and Right Gonion (go), and Left and Right Zygion (zy). The cephalometric ones are: Glabella (g'), Nasion (n'), Labiale Superius (ls'), Subnasale (sn'), Pogonion (pg'), Left and Right Alare (al'), Left and Right Endocanthion (en'), Left and Right Exocanthion (ex'), Left and Right Frontotemporale (ft'), Left and Right Gonion (go'), and Left and Right Zygion (zy'). These landmarks are those represented in Fig. 1.

Our study comprises three experiments, which are focused on distinct conditions concerning facial images: the first solely utilizes frontal facial photographs; the second exclusively employs lateral facial photographs; and the last one integrates both frontal and lateral facial photographs (referred to as the *frontal+lateral* experiment). Note that the images involved in the former two cases are those combined in the frontal+lateral experiment. The conditions of facial images encompass several parameters that directly influence the presentation and orientation of the subject's face within the photograph. These conditions include:

- Rotation Axis and Angle: The corresponding parameters pertain to the axis around which the face of the subject is rotated within the image, along with the corresponding angle of rotation.
- Translation of the Face: The related parameters refer to the displacement or movement of the subject's face within the photograph.
- Focal Length: This parameter determines the magnification level of the image and the angle of view. Shorter focal lengths provide a wider angle of view and less magnification, while longer focal lengths offer a narrower angle of view and greater magnification.

By systematically varying these conditions, a broad spectrum of facial image scenarios could be captured, simulating real-world variations. This would allow for a thorough analysis of the performance of our LR system under diverse image conditions.

Aiming to test the behavior of the LR system proposed in both H_0-true and H_1-true cases, a validation phase was addressed for each one. Hence, the aforementioned database was divided into training and validation sets using an 80%–20% split (63 and 15 subjects, respectively). Then, H_0-true and H_1-true Skull-Face Overlays were performed on both sets. The biometric scores coming from the training set were those that made up S_{H_0} and S_{H_1} in the training stage, and those from the validation set were transformed into LR values in the LR computation stage (see Section 3.3). A *validation set of LR values* is thus computed. Note that, similarly to the training stage, 10 pairs of Skull-Face Overlays for each subject that makes up the validation set are carried out to extend the aforementioned set of LR values. Consequently, a total of 1,560 Skull-Face Overlays were conducted, utilizing both the training and validation sets. The general scheme of this process is included in Fig. 5, similar to the one used in [24]. In order to generate the set of facial photographs involved in the Skull-Face Overlays, the conditions were randomly chosen within a controlled range for the first subject and then reutilized for the rest. Once these parameters were set, they were employed to generate 10 facial images of the corresponding subject. Note that only the 2D facial landmarks' locations change among them, simulating inter-expert errors (see Section 3.1).

4.2. Frontal experiment: LR system performance evaluation

The results of the experiment in which only facial photographs were involved are analyzed in the present section. For that aim, the obtained validation set of LR values was assessed employing performance characteristics and metrics previously described in Section 2.2.

First, the C_{llr}, C_{llr}^{disc} and C_{llr}^{cal} values are included in Table 1. When it comes to C_{llr}, smaller values indicate better performance, whereas those greater than 1 can be produced by miscalibrated systems. The value of C_{llr} resulting from the evaluation of our LR system was 0.619, which lies within the expected range but closer to the upper bound. The average cost due to lack of discrimination was 0.282 and the average cost due to lack of calibration was 0.337. This might be due to the use of frontal images because the Skull-Face Overlay method behaves

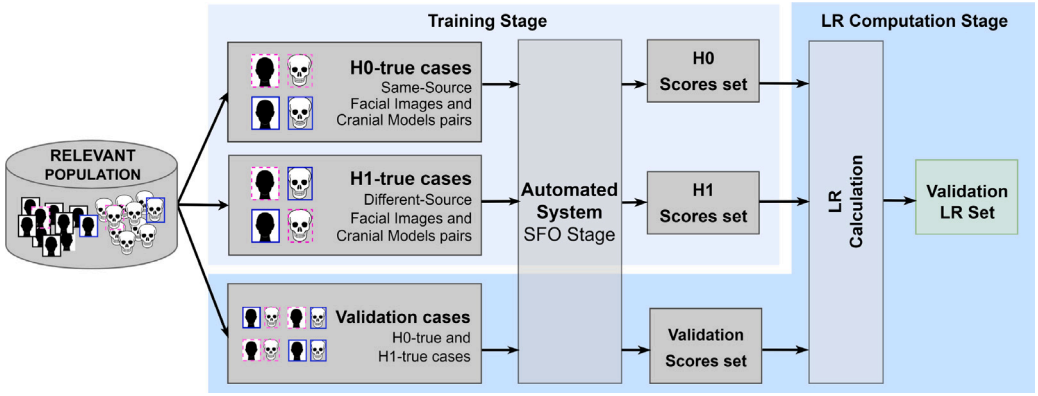

Fig. 5. Process followed to train and validate an LR system using the proposed methodology: training (top left) and LR computation (bottom right) stages.

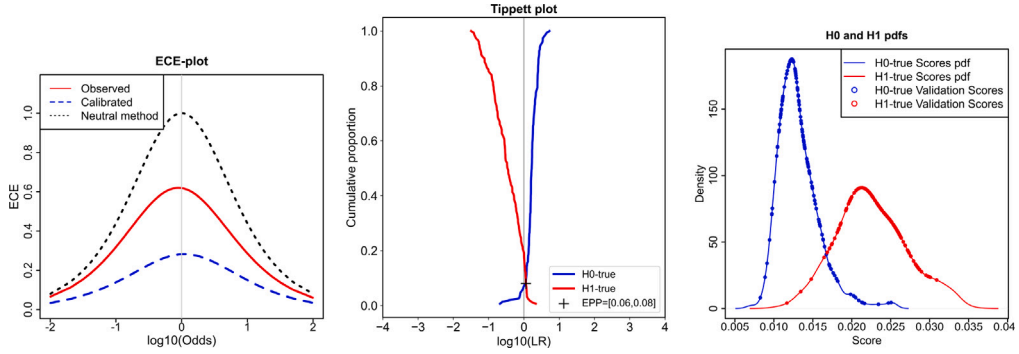

Fig. 6. ECE plot (left), Tippett plot (middle), and pdfs (right) of the frontal experiment.

Table 1
C_{llr}, C_{llr}^{disc} and C_{llr}^{cal} resulting from the LR system performance evaluation in the frontal, lateral, and frontal+lateral experiments.

Experiment	C_{llr}	C_{llr}^{disc}	C_{llr}^{cal}
Frontal	0.619	0.282	0.337
Lateral	0.467	0.209	0.258
Frontal+Lateral	0.187	0.071	0.116

better in frontal cases [13]. It could consequently conduct moderately suitable overlays even though the case under evaluation satisfies H_1.

The ECE plot obtained in the present experiment is included in Fig. 6 (left). The solid line corresponds to the ECE of the validation set of LR values, and the dashed line is the ECE after the PAV transformation, denoted as ECE^{disc}. The distance between both lines is due to a lack of calibration, referred to as ECE^{cal}. Lastly, the dotted line depicts the performance of a neutral method that consistently yields LR=1. In light of the data, the system is not constrained in its range of application, i.e. it is valid across the entire spectrum of prior odds, since all ECE values are lower than those of the neutral method.

Subsequently, the Tippett plot computed is depicted in Fig. 6 (middle). At $log_{10}LR = 0$ (the neutral value), the cumulative proportion values for both curves indicate the rates of misleading evidence. This rate represents the 8% of the values, which is significantly low. Moreover, the intersect of the curves in the Tippett plot, namely the Equal Proportion Probability (EPP), is significantly close to the aforementioned neutral value and near a cumulative proportion of 0. The alignment of this crossing point and the neutral value on the X-axis indicates that an LR system is calibrated. On the other hand, the nearer the EPP to a cumulative proportion of zero, the higher the discriminating power [23]. Additionally, the curves in the Tippett plot are notably steep, reflecting the behavior of the scores produced by the Skull-Face Overlay method. This characteristic is illustrated in Fig. 6 (right), which displays the pdfs for both H_0 and H_1 scores. The steep increase and abrupt decrease in the pdf for H_0-true validation scores indicate that these scores are concentrated within a narrow range. Consequently, the LR values under H_0 are distributed within a limited range. There is a similar behavior for the H_1-true scores, but the increase and decrease in the pdf are less abrupt. Hence, the curve in the Tippett plot moves slightly further away from the neutral value.

4.3. Lateral experiment: LR system performance evaluation

Following a similar structure to the previous section, the results of the experiment in which exclusively lateral photographs were used are analyzed. The C_{llr}, C_{llr}^{disc} and C_{llr}^{cal} values are included in Table 1. The value of C_{llr} resulting from the evaluation of the LR system is 0.467, which is closer to the lower bound of the expected range $[0, 1)$. In this case, the average cost due to lack of discrimination was 0.209 and the average cost due to lack of calibration was 0.258. According to these values, the system presents adequate behavior in terms of accuracy, and either calibration or discrimination power. The reason for this could be due to the use of lateral images. Given that the performance of the Skull-Face Overlay method diminishes in cases involving lateral images, conducting suitable overlays becomes more challenging when the case under evaluation supports H_1. Consequently, the scores in such cases tend to be higher than those in cases where H_0 is true.

The ECE plot obtained is included in Fig. 7 (left). It can be observed that all values lie below those associated with the neutral method. Hence, it is valid across the full range of prior odds. On the other hand, the corresponding Tippett plot is represented in Fig. 7 (middle). In this case, the rate of misleading evidence is higher for the H_1-true scores. The EPP is close to the neutral value but slightly shifted to the right. This could be due to the number of landmarks involved in the lateral cases, which is smaller due to the position of the face. Therefore, for the Skull-Face Overlay method, it could be easier to overlap a skull model

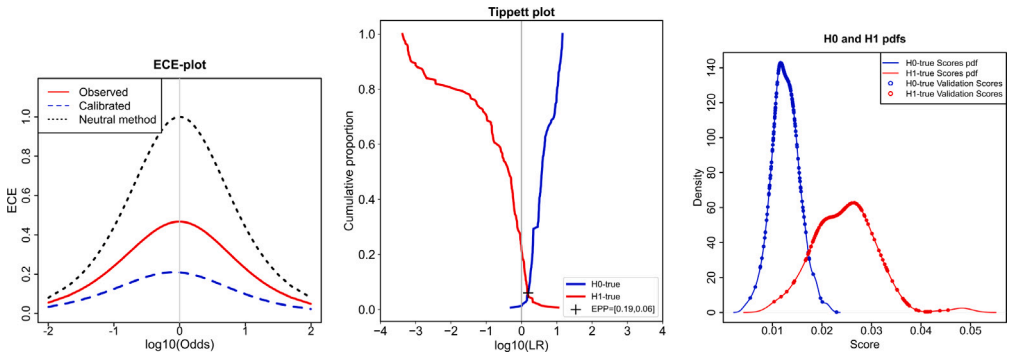

Fig. 7. ECE plot (left), Tippett plot (middle), and pdfs (right) of the lateral experiment.

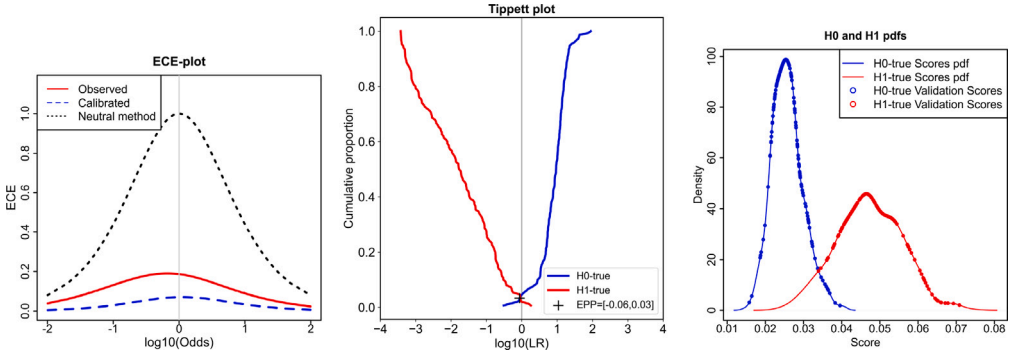

Fig. 8. ECE plot (left), Tippett plot (middle), and pdfs (right) of the frontal+lateral experiment.

and a facial image if there are fewer landmarks to match, leading to a higher number of misleading evidence.

Moreover, the behavior observed in the Tippett plot and the pdfs for both H_0 and H_1 scores (depicted in Fig. 7) is similar to that of the frontal experiment when it comes to the steepness of the curves and the variability of the scores, respectively.

4.4. Frontal+lateral experiment: LR system performance evaluation

Ultimately, the results of the experiment in which the combination of frontal and lateral photographs was used are analyzed. The calculation of the biometric score when frontal and lateral poses are integrated is detailed in Section 4.1.

The values for C_{llr}, $C_{\text{llr}}^{\text{disc}}$, and $C_{\text{llr}}^{\text{cal}}$ are collected in Table 1. The evaluation of the LR system in the frontal+lateral experiment yields a C_{llr} value of 0.187, which is significantly close to 0. The average cost due to lack of discrimination was 0.071, whereas the average cost due to the lack of calibration was 0.116. The reason for this performance is attributed to the combination of frontal and lateral images, which leverages the strengths of each image type, as reflected in the resulting LR values.

The ECE plot obtained is represented in Fig. 8 (left). Notably, all values are lower than those of the neutral method, confirming its applicability across the complete spectrum of prior odds. The Tippett plot is included in Fig. 8 (middle). In this case, the rate of misleading evidence is insignificant for both the H_1-true and the H_0-true scores (approximately 3% of the cases). Furthermore, the EPP is almost over the neutral value, which suggests that the LR system is well-calibrated. This is due to the advantage of integrating a higher quantity of information derived from two facial images showing different perspectives of the subject's face. Hence, if the information from one of the Skull-Face Overlays is misleading, the other one could slightly mitigate the effects. Moreover, in line with best practice recommendations [5,6,52], the use of two facial images instead of just one leads to more reliable identification results.

Additionally, the behavior observed in both the Tippett plot and the pdfs for H_0 and H_1 scores (depicted in Fig. 8) differs from that of the frontal and lateral experiments. Regarding the Tippett plot, the sharpness of the curves is less pronounced, indicating that the scores are distributed over a broader range, and the curves are more deviated from the neutral value. In the pdfs, the increase and decrease of the densities are smoother and the variability of the scores is higher. Note that the scores from the frontal and lateral Skull-Face Overlays are added in each case.

5. Conclusion

Craniofacial Superimposition is a challenging forensic anthropology technique to assist in the identification of human remains. However, the current decision-making process is subjective, error-prone and heavily reliant on the expertise of the forensic examiner, and the quality and quantity of materials used. Furthermore, the outcome is conveyed qualitatively, making statistical interpretation impossible. To address these issues, a decision-making support system based on Likelihood Ratios (LRs) has emerged as an appropriate tool, which has been recommended by the ENFSI for other forensic fields. In this paper, we propose for the first time a methodology to apply this framework to the identification of skeletal remains using the Craniofacial Superimposition technique.

Aiming to implement an LR-based system that considers the limitations associated with the materials in a Craniofacial Superimposition casework scenario, the framework aligns with a similarity-score-based approach and the evidence is represented by a biometric score. Additionally, a Skull-Face Overlay data generation method is employed to simulate the facial photographs of the subjects of each Craniofacial Superimposition case, using different image conditions. This led to the proposal of building an LR system and addressing different experiments: the first employs frontal facial photographs; the second exclusively uses lateral facial photographs; and the last one integrates both frontal and lateral facial photographs (*frontal+lateral* experiment).

Consequently, our novel proposal for applying the LR framework to the Craniofacial Superimposition technique is presented as an experimental validation since the data employed to train and validate were synthetically generated. This method also serves to increase the number of cases available for the training and validation of the system. Furthermore, by simulating inter- and intra-expert errors, the Skull-Face Overlay data generation method enhances the variability of the scores involved.

The LR system involved in our study has been evaluated using the C_{llr} and ECE performance metrics, together with the Tippett and ECE plots. The ECE plots of the three experiments show that the system is valid across the full range of prior odds. In the frontal experiment, the C_{llr}=0.619 and the curves' slopes in the Tippett plot are steep, reflecting the behavior of the scores yielded by the Skull-Face Overlay method, which are concentrated within a narrow range. On the other hand, in the lateral experiment, the LR system presents a C_{llr}=0.467. The curves in the Tippett plot behave similarly to those of the frontal experiment in terms of steepness. However, a slight shift can be observed. The variability of the scores is also analogous in this case. Finally, in the frontal+lateral experiment, the LR system exhibits a C_{llr}=0.187, which is significantly close to 0. In addition, the Tippett plot indicates that the system is well-calibrated, since the curves are less steep. The pdfs for both H_0 and H_1 scores show that the variability of the scores is higher than in the frontal and lateral experiments.

In summary, the results demonstrate that our proposals align with the functioning of the automatic Skull-Face Overlay system. The LR system adjusts to what is observed in the overlays and the type of score which is employed, namely an RMSE. This, in turn, reflects the proper behavior and effectiveness of the identification method involved, Posest-SFO. In H_0-true cases, the error values are more similar to each other and smaller. Conversely, H_1-true cases result in higher and more variable error values. On the other hand, it should be noted that the size of our training set is relatively small. Therefore, while our findings provide valuable insights, they should be considered within the context of the sample sizes used in our study. Furthermore, this study should be considered just a proof of concept based on the use of synthetic data. Actual data should be considered to validate these conclusions.

There are several aspects that could be further explored. First, alternative biometric scores could be used and compared. One approach could be to extract additional information from the Skull-Face Overlays and incorporate it with the RMSE value, utilizing other criteria employed in the decision-making process of the Craniofacial Superimposition technique, such as anatomical correspondence criteria [18,19]. On the other hand, a feature-based approach could be studied. Furthermore, the proposed methodology could be applied to other forensic identification techniques, such as dental or radiographic identification. This could in some cases allow for the fusion of the information drawn from evidence of diverse nature to yield an overall LR, providing a more comprehensive assessment of the evidence under the competing hypotheses. Another interesting area to explore is how factors such as aging or facial expressions might affect the performance of our LR system through specific analysis. Additionally, we are committed to expanding our research by applying the proposed methodology to diverse and geographically distinct relevant populations, including those from locations such as Mexico, Korea, South Africa, and the USA. This would extend its applicability and usefulness in the field. Finally, other interesting areas could be explored. While our error modeling incorporates the addition of random noise, a more realistic distribution of this noise could be employed to better reflect real-world scenarios. Our current error modeling thus serves as a pivotal experimental validation in this regard. When it comes to between-variability concerning soft tissue, related studies are constrained in their scope [65], typically focusing on statistical measures such as minimum, maximum, mean, and standard deviation without fully characterizing the distribution. Moreover, while assumptions of perpendicularity are common [66,67], the specific directionality of soft tissue vectors remains unmodeled. Lastly, validation of our proposal using real facial images under casework conditions emerges as a future step.

CRediT authorship contribution statement

Práxedes Martínez-Moreno: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Andrea Valsecchi:** Writing – review & editing, Methodology, Conceptualization. **Pablo Mesejo:** Writing – review & editing, Supervision, Funding acquisition. **Óscar Ibáñez:** Writing – review & editing, Resources, Funding acquisition, Conceptualization. **Sergio Damas:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This research has been developed within the R&D project CONFIA (grant PID2021-122916NB-I00), funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, and by grant FORAGE (B-TIC-456-UGR20) funded by Consejería de Universidad, Investigación e Innovación, both funded by "ERDF A way of making Europe". Miss Martínez-Moreno is supported by grant PRE2022-102029 funded by MICIU/AEI/10.13039/501100011033 and the FSE+. Dr. Valsecchi's work is supported by Red.es under grant Skeleton-ID2.0 (2021/C005/00141299). Dr. Ibáñez's work is funded by the Spanish Ministry of Science, Innovation and Universities under grant RYC2020-029454-I and by Xunta de Galicia, Spain by grant ED431F 2022/21. The authors thank Pierre Guyomarc'h for providing us with the data used in this research. Funding for open access charge: Universidad de Granada / CBUA.

References

[1] K. Burns, Forensic anthropology training manual, Prentice-Hall, 2007.
[2] M. Yoshino, Craniofacial superimposition, in: C. Wilkinson, C. Rynn (Eds.), Craniofacial Identification, University Press, Cambridge, 2012, pp. 238–253.
[3] S. Damas, O. Cordón, O. Ibáñez, J. Santamaría, I. Alemán, F. Navarro, M. Botella, Forensic identification by computer-aided craniofacial superimposition: a survey, ACM Comput. Surv. 43 (4) (2011) 1–27, http://dx.doi.org/10.1145/1978802.1978806.
[4] M.I. Huete, O. Ibáñez, C. Wilkinson, T. Kahana, Past, present, and future of craniofacial superimposition: Literature and international surveys, Leg. Med. 17 (4) (2015) 267–278, http://dx.doi.org/10.1016/j.legalmed.2015.02.001.
[5] S. Damas, O. Cordón, O. Ibáñez, Handbook on Craniofacial Superimposition: The MEPROCS Project, first ed., Springer, 2020.
[6] S. Damas, et al., Study on the performance of different craniofacial superimposition approaches (II): Best practices proposal, Forensic Sci. Int. 257 (2015) 504–508, http://dx.doi.org/10.1016/j.forsciint.2015.07.045.
[7] G. Gomez-Trenado, P. Mesejo, O. Cordón, Cascade of convolutional models for few-shot automatic cephalometric landmarks localization, Eng. Appl. Artif. Intell. 123 (2023) 106391, http://dx.doi.org/10.1016/j.engappai.2023.106391.
[8] L.F. Porto, et al., Automatic cephalometric landmarks detection on frontal faces: An approach based on supervised learning techniques, Digit. Investig. 30 (2019) 108–116, http://dx.doi.org/10.1016/j.diin.2019.07.008.
[9] E. Bermejo, K. Taniguchi, Y. Ogawa, R. Martos, A. Valsecchi, P. Mesejo, O. Ibáñez, K. Imaizumi, Automatic landmark annotation in 3D surface scans of skulls: Methodological proposal and reliability study, Comput. Methods Programs Biomed. 210 (2021) 106380, http://dx.doi.org/10.1016/j.cmpb.2021.106380.
[10] B.A. Nickerson, P.A. Fitzhorn, S.K. Koch, M. Charney, A methodology for near-optimal computational superimposition of two-dimensional digital facial photographs and three-dimensional cranial surface meshes, JFS 36 (2) (1991) 480–500, http://dx.doi.org/10.1520/JFS13050J.

[11] B.R. Campomanes-Álvarez, O. Ibáñez, C. Campomanes-Álvarez, S. Damas, O. Cordón, Modeling facial soft tissue thickness for automatic skull-face overlay, IEEE Trans. Inf. Forensics Secur. 10 (10) (2015) 2057–2070, http://dx.doi.org/10.1109/TIFS.2015.2441000.

[12] O. Ibánez, O. Cordón, S. Damas, J. Santamaria, Modeling the skull–face overlay uncertainty using fuzzy sets, IEEE Trans. Fuzzy Syst. 19 (5) (2011) 946–959, http://dx.doi.org/10.1109/TFUZZ.2011.2158220.

[13] A. Valsecchi, S. Damas, O. Cordón, A robust and efficient method for skull-face overlay in computerized craniofacial superimposition, IEEE Trans. Inf. Forensics Secur. 13 (8) (2018) 1960–1974, http://dx.doi.org/10.1109/TIFS.2018.2806939.

[14] E. Bermejo, C. Campomanes-Álvarez, A. Valsecchi, O. Ibáñez, S. Damas, O. Cordón, Genetic algorithms for skull-face overlay including mandible articulation, Inform. Sci. 420 (2017) 200–217, http://dx.doi.org/10.1016/j.ins.2017.08.029.

[15] A. Ricci, G.L. Marella, M.A. Apostol, A new experimental approach to computer-aided face/skull identification in forensic anthropology, Am. J. Forensic Med. Pathol. 27 (1) (2006) 46–49, http://dx.doi.org/10.1097/01.paf.0000202809.96283.88.

[16] M. Yoshino, K. Imaizumi, S. Miyasaka, S. Seta, Evaluation of anatomical consistency in craniofacial superimposition images, Forensic Sci. Int. 74 (1–2) (1995) 125–134, http://dx.doi.org/10.1016/0379-0738(95)01742-2.

[17] C. Campomanes-Alvarez, O. Ibáñez, O. Cordón, C. Wilkinson, Hierarchical information fusion for decision making in craniofacial superimposition, Inf. Fusion 39 (2018) 25–40, http://dx.doi.org/10.1016/j.inffus.2017.03.004.

[18] C. Campomanes-Álvarez, R. Martos-Fernández, C. Wilkinson, O. Ibáñez, O. Cordón, Modeling skull-face anatomical/morphological correspondence for craniofacial superimposition-based identification, IEEE Trans. Inf. Forensics Secur. 13 (6) (2018) 1481–1494, http://dx.doi.org/10.1109/TIFS.2018.2791434.

[19] C. Campomanes-Alvarez, O. Ibáñez, O. Cordón, Design of criteria to assess craniofacial correspondence in forensic identification based on computer vision and fuzzy integrals, Appl. Soft Comput. 46 (2016) 596–612, http://dx.doi.org/10.1016/j.asoc.2015.11.006.

[20] P. Vergeer, From specific-source feature-based to common-source score-based likelihood-ratio systems: ranking the stars, Law Probab. Risk 22 (1) (2023) mgad005, http://dx.doi.org/10.1093/lpr/mgad005.

[21] C. Champod, A. Biedermann, J. Vuille, S. Willis, J.D. Kinder, ENFSI (European network of forensic science institutes) guideline for evaluative reporting in forensic science, a primer for legal practitioners, CL&J 180 (2016).

[22] G.S. Morrison, et al., Consensus on validation of forensic voice comparison, Sci. Justice 61 (3) (2021) 299–309, http://dx.doi.org/10.1016/j.scijus.2021.02.002.

[23] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, T. Niemi, Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises, Verlag für Polizeiwissenschaft Frankfurt, 2015.

[24] A. Macarulla Rodriguez, Z. Geradts, M. Worring, Likelihood ratios for deep neural networks in face comparison, JFS 65 (4) (2020) 1169–1183, http://dx.doi.org/10.1111/1556-4029.14324I.

[25] D. Ramos, R. Haraksim, D. Meuwly, Likelihood ratio data to report the validation of a forensic fingerprint evaluation method, Data Br. 10 (2017) 75–92, http://dx.doi.org/10.1016/j.dib.2016.11.008.

[26] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int. 219 (1) (2012) 129–140, http://dx.doi.org/10.1016/j.forsciint.2011.12.009.

[27] R. Corzo, T. Hoffman, P. Weis, J. Franco-Pedroso, D. Ramos, J. Almirall, The use of LA-ICP-MS databases to calculate likelihood ratios for the forensic analysis of glass evidence, Talanta 186 (2018) 655–661, http://dx.doi.org/10.1016/j.talanta.2018.02.027.

[28] S. Riman, H. Iyer, P.M. Vallone, Examining discrimination performance and likelihood ratio values for two different likelihood ratio systems using the provedit dataset, bioRxiv (2021) http://dx.doi.org/10.1101/2021.05.26.445891.

[29] P. Gill, C. Brenner, J. Buckleton, A. Carracedo, M. Krawczak, W. Mayr, N. Morling, M. Prinz, P. Schneider, B. Weir, DNA commission of the international society of forensic genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2) (2006) 90–101, http://dx.doi.org/10.1016/j.forsciint.2006.04.009.

[30] A. Collins, N.E. Morton, Likelihood ratios for DNA identification, Proc. Natl. Acad. Sci. 91 (13) (1994) 6007–6011, http://dx.doi.org/10.1073/pnas.91.13.6007.

[31] Ø. Bleka, T. Wisløff, P.S. Dahlberg, V. Rolseth, T. Egeland, Advancing estimation of chronological age by utilizing available evidence based on two radiographical methods, IJLM 133 (2019) 217–229, http://dx.doi.org/10.1007/s00414-018-1848-y.

[32] N. Angelakopoulos, et al., Third molar maturity index (I3M) assessment according to different geographical zones: a large multi-ethnic study sample, IJLM 137 (2022) 403–425, http://dx.doi.org/10.1007/s00414-022-02930-x.

[33] R. Verma, et al., Estimation of sex in forensic examinations using logistic regression and likelihood ratios, Forensic Sci. Int.: Rep. 2 (2020) 100118, http://dx.doi.org/10.1016/j.fsir.2020.100118.

[34] C.E. Berger, H.H. de Boer, M. van Wijk, Chapter 3.2 - use of Bayes' theorem in data analysis and interpretation, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Academic Press, 2020, pp. 125–135, http://dx.doi.org/10.1016/B978-0-12-815764-0.00014-9.

[35] C.E. Berger, M. van Wijk, H.H. de Boer, Chapter 6.1 - Bayesian inference in personal identification, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Academic Press, 2020, pp. 301–312, http://dx.doi.org/10.1016/B978-0-12-815764-0.00006-X.

[36] G.S. Morrison, P. Weber, N. Basu, R. Puch-Solis, P.S. Randolph-Quinney, Calculation of likelihood ratios for inference of biological sex from human skeletal remains, Forensic Sci. Int.: Synergy 3 (2021) 100202, http://dx.doi.org/10.1016/j.fsisyn.2021.100202.

[37] R. Verma, K. Krishan, D. Rani, A. Kumar, V. Sharma, Stature estimation in forensic examinations using regression analysis: A likelihood ratio perspective, Forensic Sci. Int.: Rep. 2 (2020) 100069, http://dx.doi.org/10.1016/j.fsir.2020.100069.

[38] D. Ramos-Castro, J. González-Rodríguez, Cross-entropy analysis of the information in forensic speaker recognition, in: The Speaker and Language Recognition Workshop, 2008, p. 4.

[39] T. Hicks, A. Biedermann, J. de Koeijer, F. Taroni, C. Champod, I. Evett, The importance of distinguishing information from evidence/observations when formulating propositions, Sci. Justice 55 (6) (2015) 520–525, http://dx.doi.org/10.1016/j.scijus.2015.06.008.

[40] B.A. Spellman, H. Eldridge, P. Bieber, Challenges to reasoning in forensic science decisions, Forensic Sci. Int.: Synerg. 4 (2022) 100200, http://dx.doi.org/10.1016/j.fsisyn.2021.100200.

[41] G.S. Cooper, V. Meterko, Cognitive bias research in forensic science: A systematic review, Forensic Sci. Int. 297 (2019) 35–46, http://dx.doi.org/10.1016/j.forsciint.2019.01.016.

[42] T. Bayes, n. Price, LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S, Philos. Trans. R. Soc. Lond. 53 (1763) 370–418, http://dx.doi.org/10.1098/rstl.1763.0053.

[43] D. Lindley, A problem in forensic science, Biometrika 64 (2) (1977) 207–213.

[44] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, Comput. Speech Lang. 20 (2) (2006) 230–275, http://dx.doi.org/10.1016/j.csl.2005.08.001.

[45] G.S. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction, Speech Commun. 85 (2016) 119–126, http://dx.doi.org/10.1016/j.specom.2016.07.006.

[46] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. 276 (2017) 142–153, http://dx.doi.org/10.1016/j.forsciint.2016.03.048.

[47] D.A. van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in: C.A. Müller (Ed.), Speaker Classification I: Fundamentals, Features, and Methods, 4343, 2007, pp. 330–353.

[48] G. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (3) (2011) 91–98, http://dx.doi.org/10.1016/j.scijus.2011.03.002.

[49] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 15 (7) (2007) 2104–2115, http://dx.doi.org/10.1109/TASL.2007.902747.

[50] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: Measuring calibration of likelihood ratios, Forensic Sci. Int. 230 (1) (2013) 156–169, http://dx.doi.org/10.1016/j.forsciint.2013.04.014.

[51] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, JFS 58 (6) (2013) 1503–1518, http://dx.doi.org/10.1111/1556-4029.12233.

[52] T.W. Fenton, A.N. Heard, N.J. Sauer, Skull-photo superimposition and border deaths: identification through exclusion and the failure to exclude, JFS 53 (1) (2008) 34–40, http://dx.doi.org/10.1111/j.1556-4029.2007.00624.x.

[53] O. Ibáñez, F. Cavalli, B.R. Campomanes-Álvarez, C. Campomanes-Álvarez, A. Valsecchi, M.I. Huete, Ground truth data generation for skull–face overlay, IJLM 129 (3) (2015) 569–581, http://dx.doi.org/10.1007/s00414-014-1074-1.

[54] D. Ramos, R.P. Krish, J. Fiérrez, D. Meuwly, From biometric scores to forensic likelihood ratios, in: M. Tistarelli, C. Champod (Eds.), Handbook of Biometrics for Forensic Science, in: Advances in Computer Vision and Pattern Recognition, 2017, pp. 305–327.

[55] T. Ali, L. Spreeuwers, R. Veldhuis, D. Meuwly, Biometric evidence evaluation: an empirical assessment of the effect of different training data, IET Biometrics 3 (4) (2014) 335–346, http://dx.doi.org/10.1049/iet-bmt.2014.0009.

[56] A.J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, I. Alberink, Performance study of a score-based likelihood ratio system for forensic fingermark comparison, J. Forensic Sci. 62 (3) (2017) 626–640, http://dx.doi.org/10.1111/1556-4029.13339.

[57] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality, Sci. Justice 58 (1) (2018) 47–58, http://dx.doi.org/10.1016/j.scijus.2017.06.005.

[58] B.R. Campomanes-Álvarez, O. Ibáñez, F. Navarro, I. Alemán, O. Cordón, S. Damas, Dispersion assessment in the location of facial landmarks on photographs, IJLM 129 (1) (2015) 227–236, http://dx.doi.org/10.1007/s00414-014-1002-4.

[59] M. Cummaudo, et al., Pitfalls at the root of facial assessment on photographs: A quantitative study of accuracy in positioning facial landmarks, IJLM 127 (3) (2013) 699–706, http://dx.doi.org/10.1007/s00414-013-0850-7.

[60] Y. Chen, A tutorial on kernel density estimation and recent advances, Biostat. Epidemiol. 1 (1) (2017) 161–187, http://dx.doi.org/10.1080/24709360.2017.1396742.

[61] T. Ali, L. Spreeuwers, R. Veldhuis, A review of calibration methods for biometric systems in forensic applications, in: 33rd Symposium on Information Theory in the Benelux and the 2nd Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux 2012, 2012, pp. 126–133.

[62] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.

[63] D.M. Ommen, C.P. Saunders, Building a unified statistical framework for the forensic identification of source problems, Law Probab. Risk 17 (2) (2018) 179–197, http://dx.doi.org/10.1093/lpr/mgy008.

[64] P. Guyomarc'h, B. Dutailly, J. Charton, F. Santos, P. Desbarats, H. Coqueugniot, Anthropological facial approximation in three dimensions (AFA3D): Computer-assisted estimation of the facial morphology using geometric morphometrics, JFS 59 (2014) http://dx.doi.org/10.1111/1556-4029.12547.

[65] C.N. Stephan, 2018 Tallied facial soft tissue thicknesses for adults and sub-adults, Forensic Sci. Int. 280 (2017) 113–123, http://dx.doi.org/10.1016/j.forsciint.2017.09.016.

[66] M. Domaracki, C.N. Stephan, Facial soft tissue thicknesses in Australian adult cadavers*, J. Forensic Sci. 51 (1) (2006) 5–10, http://dx.doi.org/10.1111/j.1556-4029.2005.00009.x.

[67] E. Simpson, M. Henneberg, Variation in soft-tissue thicknesses on the human face and their relation to craniometric dimensions, Am. J. Phys. Anthropol. 118 (2) (2002) 121–133, http://dx.doi.org/10.1002/ajpa.10073.