Campomanes-Alvarez, C, Martos-Fernandez, R, Wilkinson, CM, Ibanez, O and Cordon, O

 Modeling skull-face anatomical/morphological correspondence for craniofacial superimposition-based identification

http://researchonline.ljmu.ac.uk/id/eprint/7808/

Article

Modeling skull-face anatomical/morphological correspondence for craniofacial superimposition-based identification

Carmen Campomanes-Álvarez, Rubén Martos-Fernández, Caroline Wilkinson, Oscar Ibáñez and Oscar Cordón

Abstract—**Craniofacial superimposition (CFS) is a forensic identification technique which studies the anatomical and morphological correspondence between a skull and a face. It involves the process of overlaying a variable number of facial images with the skull. This technique has great potential since nowadays the wide majority of the people have photographs where their faces are clearly visible. In addition, the skull is a bone that hardly degrades under the effect of fire, humidity, temperature changes, etc. Three consecutive stages for the CFS process have been distinguished: the acquisition and processing of the materials; the skull-face overlay; and the decision making. This final stage consists of determining the degree of support for a match based on the previous overlays. The final decision is guided by different criteria depending on the anatomical relations between the skull and the face. In previous approaches, we proposed a framework for automating this stage at different levels taking into consideration all the information and uncertainty sources involved. In this study, we model new anatomical skull-face regions and we tackle the last level of the hierarchical decision support system. For the first time, we present a complete system which provides a final degree of craniofacial correspondence. Furthermore, we validate our system as an automatic identification tool analyzing its capabilities in closed (known information or a potential list of those involved) and open lists (little or no idea at first who may be involved) and comparing its performance with the manual results achieved by experts, obtaining a remarkable performance. The proposed system has been demonstrated to be valid for sortlisting a given data set of initial candidates (in 62,5% of the cases the positive one is ranked in the first position) and to serve as an exclusion method (97,4% and 96% of true negatives in training and test, respectively).**

Index Terms—**Forensic anthropology, craniofacial identification, craniofacial superimposition, decision making, soft computing, fuzzy aggregation operators, computer vision, spatial relations.**

I. INTRODUCTION

FORENSIC identification techniques share the search for equal/compatible patterns in two (or more) different images/models/data/etc. [1]. In the majority of these techniques (for example, DNA, fingerprints, face recognition, dental identification) the data under comparison belong to the same 'object' (i.e. comparison of two DNA strings, fingerprint images, photographs of the face, teeth x-rays, etc.). However, that is not the case with craniofacial superimposition (CFS) [2], a human identification technique where images of the face and the skull are compared to establish the identity of a given subject.

The whole CFS process can be divided into three consecutive stages [3], [4]: 1) The acquisition and processing of the materials, i.e, skull and *ante-morten* (AM) facial images, and the location of somatometric landmarks on both; 2) skull-face overlay (SFO), which deals with accomplishing the best possible superimposition of the skull and a single AM photograph of a missing person. This procedure is iteratively executed for each photograph, thus obtaining different overlays; and 3) decision making, which aims to determine the degree of support for a match based on the SFOs achieved in the previous step. The final decision is managed by different criteria based on the anatomical relationship between the face and the skull. These criteria can vary depending on the region and the pose [5].

Skull-face morphological/anatomical correspondence has been largely studied in different research fields: within the said CFS field [6]–[10], by maxillofacial and plastic surgeries [11], geneticians [12], and in a more significant way within the craniofacial reconstruction community [13]. A few tens of corresponding landmarks have thus been identified in both the skull (craniometric) and the face (cephalometric) [14], and hundreds of studies measuring soft tissue depth between corresponding landmarks have been published using data from different populations and different measuring techniques [15]. In addition, a different set of studies have focused on relating the shape of certain regions (morphological correspondence), i.e., the nasal bones and the nose [16], the dentition and the mouth [17], [18], the orbits and the eyes [19], and others such as the mandible, the chin, the zygomatic area, etc.

The CFS identification technique has a great application potential since nowadays the wide majority of people have photographs (AM material) where their faces are clearly visible. The counterpart, the skull (PM material), is a bone that hardly degrades with the effect of fire, humidity, high or low temperatures, time lapse, etc. However, despite some authors considering this technique as a valid identification method [6], [8], [9], [20], many others only regard CFS as a tool for the exclusion method [21]–[23]. The fact that two different objects have to be compared implies a number of sources of uncertainty, making CFS identification a subjective technique, highly dependent on the forensic anthropologist's experience

C. Campomanes-Álvarez, O. Ibáñez and O. Cordón are with Department of Computer Science and Artificial Intelligence and with Research Centre for Information and Communications Technologies of the University of Granada, 18071-Granada, Spain (e-mail: carmen.campomanes@decsai.ugr.es, oscar.ibanez@decsai.ugr.es; ocordon@decsai.ugr.es). Rubén Martos-Fernández is with Department of Physical Anthropology, University of Granada, 18071-Granada, Spain (e-mail: rmarfer@correo.ugr.es). C. Wilkinson is with School of Art and Design, Liverpool John Moores University, Liverpool L3 5TF, UK (e-mail: C.M.Wilkinson@ljmu.ac.uk).

and knowledge: acquisition parameters of the photographs [24], [25], mandible articulation, landmark location imprecision [26], soft tissue depth variability [15], and ultimately morphological comparison.

Designing automatic methods to address CFS and support the forensic anthropologist remains a challenge and dream milestone within the community. The development of computer-aided CFS methods has increased over the past twenty years [4]. State-of-the-art approaches use skull 3D models, allowing the automation of the SFO stage [24], [27], [28] through an image registration process [29]. These methods focus only on the SFO stage and do not provide solutions for the decision making stage. In fact, the limited research that tackles the automation of craniofacial correspondence is basic, limited, and outdated [20], [30], [31]. Moreover, it does not use 3D models or computer-based techniques, so shape analysis always requires manual interaction. For this reason, decisions about the skull and face relationships are still made in a subjective way, with a complete absence of measurements and countable aggregation of positive and negative factors. Consequently, the experience and knowledge of the practitioner is the key aspect for a correct decision, meaning there is a strong interest in designing and implementing a decision support system (DSS) for the decision making stage of the CFS technique.

Over the last 10 years our research group (SOCCER [32]), in collaboration with the physical anthropology lab at the University of Granada and various other international forensic labs, has been working with the two-fold goal of automating the whole CFS identification process while increasing its reliability. This way, we have studied and modeled landmark location and matching uncertainties [33]–[35], and proposed a mathematical formulation [24] and several optimization algorithms for searching the acquisition parameters of the photographs leading to the optimal skull-face overlay. Our system relies on the use of 3D (surface) skull models. The use of 3D models (acquired from dry bones, fresh cadavers, or the living) using laser or structure light scanners, photogrammetry techniques, or even in some cases clinical devices such as computed tomography (CT) or cone beam CT (CBCT), is a well-established technology in Forensic Anthropology, both in the academic and daily case work [36] .

Recently, we have also developed a theoretical framework for decision making [37], [38]. This hierarchical DSS represents the first automatic system for human identification by CFS. However, it is still incomplete as it does not model a significant set of morphological correspondences leading to a complete analysis of the skull-face anatomical correspondence. In fact, according to [6], it is stated that the skull can be positively identified as the presumed person if more than 13 examination criteria are anatomically satisfied. In addition, the implementation of the hierarchy missed the last level, where the information provided by different photographs of the same subject are jointly considered for a final decision. Thus, the goal of the current work is three-fold:

- To study and computationally model a significant number of anatomical regions of the skull and their morphological

correspondence with the face.
- To tackle the last level of the hierarchical DSS, providing formulas to integrate the information coming from the previous levels.
- To propose a methodology to analyse the identification capabilities of automatic methods in closed and open lists, applying it to the CFS system completed in this proposal.

The combination of these three achievements with our group's previous findings will give rise to the first automatic system for CFS, allowing us to rank a set of candidates and even to provide an identification decision for every single individual in a data set of skull and faces.

The organization of this manuscript is as follows: Section II summarizes our previous contribution where the hierarchical DSS was proposed. Section III introduces our proposal, addressing the last level of the DSS and presents the analysis and computational models of nine different skull-face regions anatomically related. Section IV is intended to introduce the experimental setup, the obtained results, and their analysis. Finally, in Section V we present a discussion of the work developed, some concluding remarks, and related future studies.

II. BACKGROUND: HIERARCHICAL DECISION SUPPORT FRAMEWORK FOR CRANIOFACIAL SUPERIMPOSITION

The final decision in CFS involves diverse criteria which study the bony and the facial anatomical relation. Following [6], [10], we can distinguish four different sets of criteria in order to assess the craniofacial correspondence (see Fig. 1 for a graphical example of each set). Our objective is to automate the whole decision making process. To do so, we aim to model the former four families of criteria using computer vision (CV) and soft computing (SC) techniques. CV was already used in our previous works [35], [37]–[39] to measure some craniofacial correspondences and segment the correct contours of the regions. Fuzzy integrals [40], well-known aggregation operators from the SC field, were used to obtain the final matching degree from several CV methods. Using these technologies, the resulting system would assist the expert's identification decision by providing a numerical output indicative of the matching degree of a given CFS problem.

In [38], we proposed a complete framework for a DSS in CFS (See Fig. 2). The system develops fusion of information concerning skull-face anatomical correspondence at three different levels: criterion evaluation, SFO evaluation, and CFS evaluation. Following this structure, different CV methods for evaluating a specific criterion are aggregated considering the accuracy of each one at level 3. The result of this aggregation is the matching degree C_m. The next level (level 2) corresponds to the SFO matching degree. In order to obtain this value, the craniofacial anatomical correspondence of several criteria of the overall face have to be studied. For each criterion, the skull-face matching degree is obtained applying a specific CV method (level 3). The quality of the materials (PQ_m and BQ_m), the biological profile variability (BP_m), and the discriminative power of each isolated region are taken into consideration to compute this SFO degree. Three different
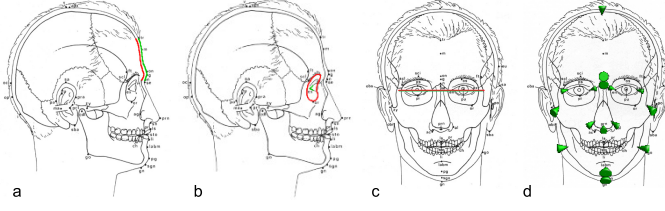
Fig. 1: (modified image from the one published in [38]) Examples of the four families of criteria for assessing the craniofacial correspondence. a) Consistency analysis of the facial (forehead) and bony (frontal bone) morphological curves or outlines; b) Anatomical consistency assessment by positional relationship (orbit-eyeball); c) Analysis of the anatomical consistency by line location and comparison (Ectocanthion line); d) Consistency evaluation of the soft tissue thickness between corresponding facial and cranial landmarks.

sublevels are distinguished in level 2. The first one aggregates the sources of uncertainty using either the minimum (min), the product (prod) or the arithmetic Mean (mean). The second one integrates this uncertainty with the matching degree of the criterion (level 3) using either the Weighted Arithmetic Mean (wam) or the Weighted Geometric Mean (wgm). The third sublevel aggregates the different previous values, using either wam, wgm, Choquet (choq) or Sugeno (sug) integrals, for all the regions weighting them by the discriminative power of the isolated region. Finally, the resulting degree of the CFS identification case (level 1) is achieved taking into account all the corresponding SFO evaluation values.

In [35], [37], [39], we presented two of the most discriminative criteria to assess craniofacial correspondence at criterion evaluation: the morphological and spatial relationship between the facial and bony chin, and the relative position of the eyeballs and the orbits. In these studies, we developed several CV-based methods to assess the degree of matching of each of these two criteria, and aggregated their results in a single value (using different aggregation functions) to obtain more robust and accurate results (level 3 of the DSS framework). Table I shows the two criteria implemented until now and the regions where each one is applied. It also summarizes the methodology for the CV methods which achieve the best results in each case. These methods were used in [38] at the criterion evaluation level. In that contribution, we also described all the aspects and considerations included in the decision making process in CFS. Thus, the uncertainty sources and degrees of confidence involved were classified as related to bone (PM material), image (AM material), skull-face overlays, morphological aspects, and used methods. The experimentation developed in that study focused only on the SFO evaluation level (i.e. level 2). Within this stage, we distinguished three sublevels with different conditions of information aggregation.

III. PROPOSAL

Frequently, more than one AM photograph of the same person is available for a CFS problem. In these cases the SFO and its corresponding analysis is made individually for each photograph. Nevertheless, experts do not proceed in this way

as they do not treat each SFO in isolation but take them all into account in order to make the final identification decision. In fact, from a forensic point of view, the reliability of the CFS technique increases having more than one photo, in different poses, etc. [5], [6], [9]. In [38], we proposed a complete theoretical definition of the DSS framework, in which this concept is correctly reflected. Within this framework, we implemented a first approach for the lowest level (level 3: criterion evaluation) in [37] and the intermediate one (level 2: SFO evaluation). As explained in Section II, until now we have only taken two criteria into account and, in consequence, four regions at most for each case (chin contour, right eye, left eye, and cranial contour). In the current contribution, we study, model and incorporate more criteria and facial regions, which are commonly used by forensic experts in this process. Due to this, the final craniofacial correspondence degree is expected to be more robust and reliable as well as closer to the manual analysis made by experts. In addition, we aggregate the craniofacial correspondence value of each SFO to return the final CFS degree, thus, completing the definition of level 1 in our DSS framework (see Fig 2). We also perform two kinds of innovate experiments. On the one hand, we test the DSS system as a shortlisting tool to study whether it is capable of filtering out the candidates. Thus, we can analyze the evolution regarding the previous results in [38], in terms of CFS cases and take more criteria into consideration. On the other hand, we validate our system as an automatic identification tool for the first time, comparing its performance with the results from a study manually achieved by forensic experts in the framework of an international project [41].

A. Modeling skull-face anatomical/morphological correspondence

In previous work [35], we designed some methods to analyze the craniofacial correspondence based on some specific regions. In the current contribution, we study and model a set of suitable criteria to evaluate the morphological/anatomical correspondence on the whole face. Therefore, the evaluation of the craniofacial correspondence will be more complete and reliable. In particular, we introduce a new method that corresponds to the second family of criteria (anatomical consistency by positional relation) and two new approaches belonging to the third one (anatomical consistency by line location and comparison). We analyze the combined action of the previous and the new methods in different cases of study, which together form a global system that considers all the significant parts of the face.

1) Modeling the anatomical consistency by the position of two bony regions: One of the sets of criteria that forensic experts take into account in the decision making stage is the consistency of hard tissue to hard tissue positions. From a CV point of view, the implementation consists of overlapping two regions. Manual marking of both corresponding regions has to be done by the anthropologists in an approximate way, keeping in mind that the only visible hard tissue in the face are dental pieces. The 3D regions are projected onto the 2D image using the geometric transformation obtained in stage two of CFS
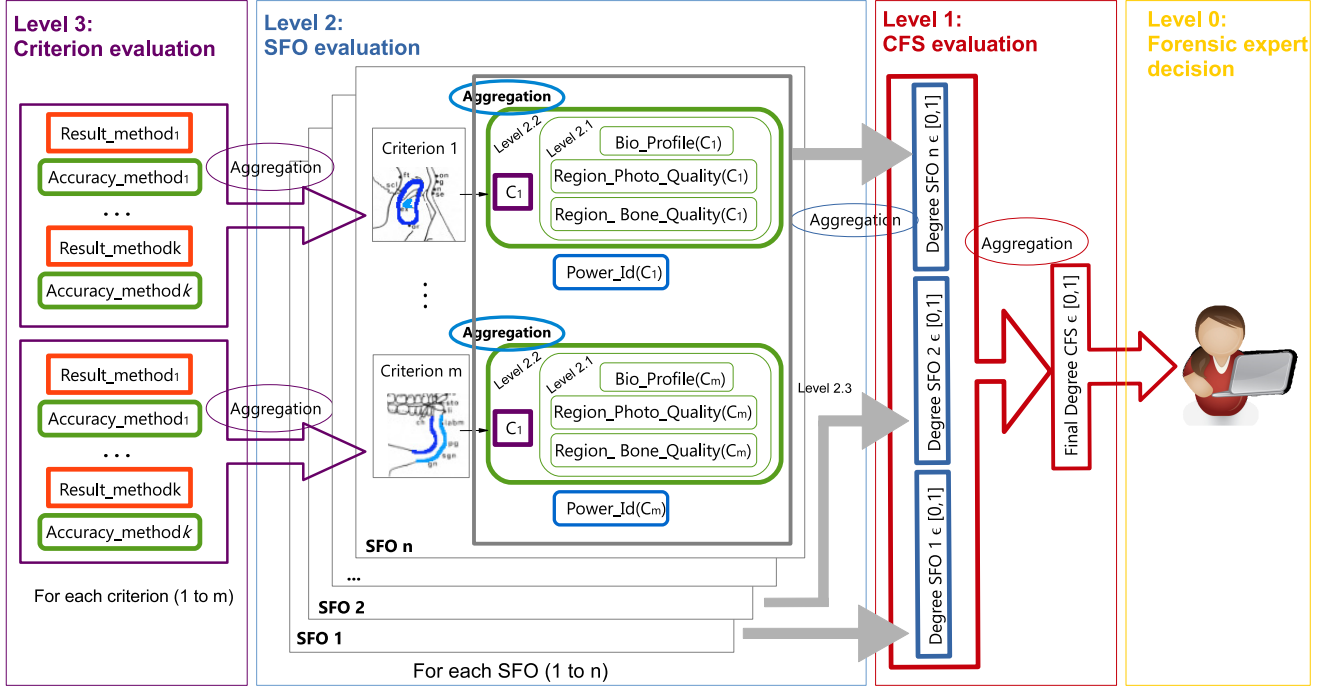
Fig. 2: Hierarchical scheme of the DSS for CFS proposed in [38] and extended in the current contribution.

for accomplishing SFO. Then, the aim is to check whether the two regions are located in the same place as the image. Due to the quality of the materials, the skull regions could be greater than the photographic ones or vice versa (dental pieces in photographs are in most cases partially occluded by the lips). Thus, we consider that when one of the regions is entirely inside the other, the position of both objects is considered to be consistent (a value of 1 is assigned for the correspondence degree in that case). Note that the length of the regions is not taken into consideration. On the contrary, when the whole facial region is outside the corresponding cranial region, the position of the objects is not considered to be consistent at all (a value of 0 is assigned). For any intermediate situation of a partial matching between the two regions, the part of the object inside the other is considered to compute the consistency degree defined in $[0, 1]$ by means of Eq. 1, i.e., DICE index ([42]):

$$S_{overlap} = 2 \cdot \frac{A(Region_{skull}) \cap A(Region_{face})}{A(Region_{skull}) + A(Region_{face})}, \quad (1)$$

where $A(Region)$ stands for the area of a region.

2) Modeling the anatomical consistency by line location and comparison: Following this set of criteria, experts analyze a set of marking lines, obtained by joining some reference landmarks on the face and skull. It is important to note that the cranial landmarks, which are used to obtain the corresponding cranial line, are 3D points. These points are projected onto the 2D image using the geometric transformation of the SFO stage. Then, the comparison of both lines occurs in the 2D space. Depending on the lines at hand, two different aspects can be distinguished. Firstly, the study of the parallelism of

both lines (cranial and facial). Secondly, the similarity of their lengths.

In terms of CV and regarding our DSS framework, these criteria have to be given as a value in the interval $[0, 1]$. For the first case, the angle between both lines has to be calculated. To do so, we use the following formula:

$$\alpha = acos \left(\frac{\vec{v}_{skull} \cdot \vec{v}_{face}}{\|\vec{v}_{skull}\| \, \|\vec{v}_{face}\|} \right), \quad (2)$$

where $\vec{v}_{cranial}$ and $\vec{v}_{facial}$ are the cranial and facial lines, $\cdot$ represents the dot product, and $\|\vec{v}\|$ refers to the magnitude of the vector.
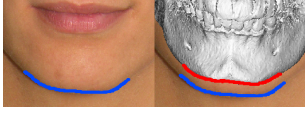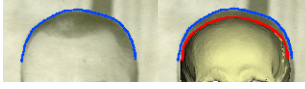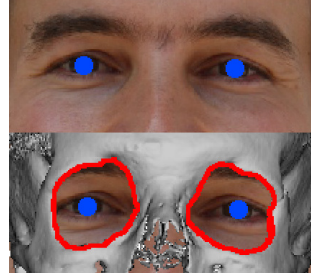
Thus, $\alpha \in [0^\mathrm{o}, 360^\mathrm{o}]$. Although the skull could not belong to the subject of the photograph, we can assume, without doubt, that the SFO is correctly performed as it is guided by the landmark matching, so both lines will be more or less parallel. For this reason, we establish the worst case of this relation when the angle formed by the two lines is greater than or equal to 45°. The final degree, which expresses the consistency between the cranial and the facial lines in a SFO regarding to the parallelism between them is:

$$S_{parallelism} = \begin{cases} 1 - \frac{\alpha}{45}, & \text{if } \alpha \leq 45^\mathrm{o} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

For the second case, we need to compare the lengths of the two lines. To do so, we consider a complete mismatch those situations where the difference between both lengths is greater than the double. We use this factor in order to make the criteria more sensitive to small differences in the length of both lines. Thus, the similarity between the cranial and facial lines regarding to their length is:

TABLE I: Best performing methods for criterion evaluation studied in [37].

Criterion	Consistency analysis of the facial and bony morphological curves or outlines	Anatomical consistency assessment by positional relationship
CV methodology	Similarity measures using shape signatures $$S = 1 - \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_F(i) - s_B(i))^2}$$ where s_F and s_B are the shape signatures of the facial contour and bony contour, respectively. Best results with: - Complex coordinates signature - Area function signature Final degree: Aggregation of both using Sugeno Integral	Similarity between two 2D positions using a 3D reference model $$S = 1 - \frac{|\delta_a - \delta_a'| + |\delta_b - \delta_b'| + |\delta_r - \delta_r'| + |\delta_l - \delta_l'|}{\delta_a + \delta_a' + \delta_b + \delta_b' + \delta_r + \delta_r' + \delta_l + \delta_l'}$$ where δ_a, δ_b, δ_r and δ_l are the degrees of the position relation "above", "below", "right", and "left", respectively. δ and δ' are the relative position of the object and the reference. They are calculated using the angle between the segment joining two points and the x-axis [37]: - Aggregation method (average of the point to point relation) - Centroid method (relation between both centroid) Final degree: Aggregation of these two using Sugeno Integral
Requirements	Marking 3D cranial region and 2D facial contour	Marking 3D cranial and 2D facial regions Reference 3D cranial and facial regions
Regions	Chin contour Cranial contour 	Position of the eyeball center and the cranial orbit

$$S_{length} = 1 - \frac{|\|\vec{v}_{skull}\| - \|\vec{v}_{face}\||}{\|\vec{v}\|_{greatest}/2}, \qquad (4)$$

where $|\vec{v}|$ refers to the absolute value and $\|\vec{v}\|_{greatest}$ is the length of the greatest line. In the case that the latter formulae returns a value lower than 0, the similarity between the two lines will be 0.

3) New cases of study using the first three families of criteria to assess craniofacial correspondence: In previous works [35], [38], we presented three cases of study for analyzing the correspondence between the skull and the face in a SFO using CV: the chin outlines comparison, the orbits and eyeball centers positional relations, and the cranial contours comparison. In real cases, experts consider more criteria to establish the craniofacial correspondence. They take different and significant regions of the face as a base to analyze if the skull belong to a particular person, i.e., the chin, the forehead, the nose, the mouth, the ears, etc. The more the regions to analyze, the higher the reliability of the results. In this sense, we have studied and modeled the following six new relations according to the expert knowledge available [43]:

i **Mouth occlusion length comparison.** There are several studies from orthodontic and anatomical fields that relate the mouth to the occlusion of the teeth [18]. However, most of them can be dismissed because of lack of data or limited confidence due to the reduced sample used. For this reason, in our proposal we model the method based on the intercanine distance (a visual representation is shown in Fig. 3.i). To perform this comparison using CV, the facial occlusion line of the photograph (in frontal view and neutral pose) and the region between the two first canines in the skull 3D model have to be marked manually. Once the 3D region is projected onto the 2D image, the two most

remote points are joined. Hence, this length is comparable with the line of the facial image using Eq. 4. The mouth corner position does not change with age or body mass index (BMI), but the dental points may change due to loss of teeth.

ii **Upper rim of the external auditory meatus and facial tragus (ear) positional relationship.** The ear morphology and its corresponding skeletal structure is one of the most understudied craniofacial relations. Some of the existing studies conclude that the tragus of the ear has an stable positional relationship with the upper rim of the external auditory meatus (see Fig. 3.ii). Thus, we have modeled this criterion in the same way as those belonging to the second family of criteria (same as for the orbits and eyeball centers, see Table I). On the image, the facial tragus is marked as a region. On the 3D model skull, the external auditory meatus is also marked as a region by the expert. Once the 3D region is projected onto the image, the similarity value is obtained using the comparison of the positional relation of a reference model, which is a 3D model (CT) including bone (skull) and soft tissue (face) where the given anatomical regions have been marked (see [37] for further details). This relation does not change with age, BMI, or ethnic group.

iii **Ectocanthion lines parallelism comparison.** Commonly, experts focus on the lines on the face and the skull to assess the anatomical consistency in an SFO. In particular, this line is achieved by joining the two excanthion landmarks in the skull and the two ectocanthion landmarks in the photograph, and both should be parallel in a positive case as depicted in Fig. 3.iii. For modelling this relationship, the 3D landmarks are projected onto the 2D space and then the lines are compared using Eq. 3. There should be no
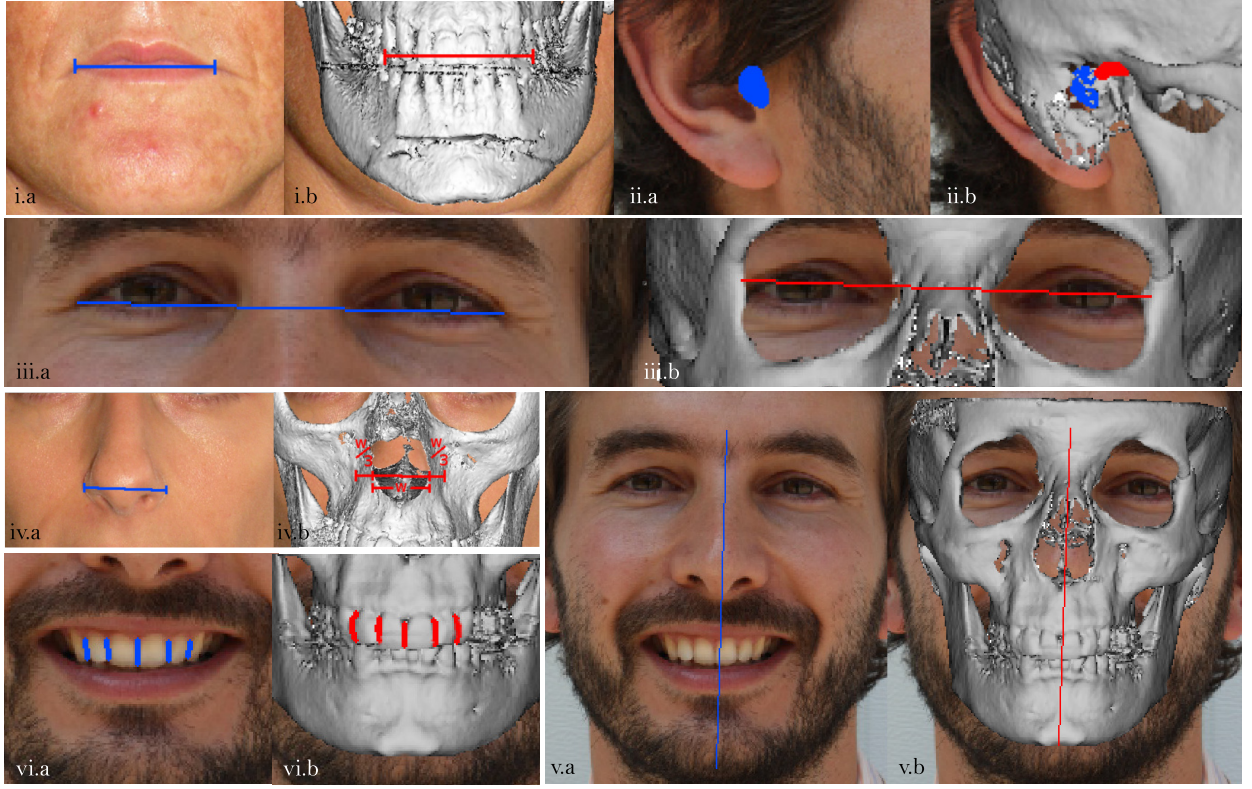
Fig. 3: New craniofacial criteria developed for the proposed DSS (to be used at level 3). From left to right, and top to bottom: i.a/b) Mouth occlusion length comparison; ii.a/b) Upper rim of the external auditory meatus and facial tragus (ear) positional relationship; iii.a/b) Ectocanthion lines parallelism comparison; iv.a/b) Nose width length and parallelism comparison; v.a/b) Frontal/central line parallelism comparison; vi.a/b) Overlap of inter-dental lines

changes in position with age or BMI, but the visibility may change due to eyelid dropping and fissure shortening. In our model, this fact is taken into account using the quality of the regions parameter.

iv **Nose width length and parallelism comparison.** The most studied feature of the face is the nose. In [16], the authors concluded that the bony nasal aperture at its widest point is three-fifths of the overall width of the soft nose (Fig. 3.iv). A post-study performed on living subjects of different ethnic groups confirmed this relationship [17]. Accordingly, from a CV point of view, the maximum width of the soft nose is the width of the nasal aperture plus two-thirds of it. In order to model this relation, a line over the maximum width of the nose is marked on the facial photograph. In the 3D model, experts mark two 3D points corresponding to the maximum width of the nasal aperture and, on this line, a vector with the same direction and with a magnitude of a third of this width is added to each extreme. Next, this whole line is projected onto the 2D image and it is compared with the line marked on the photograph. For this criterion we use Eqs. 3 and 4, since the position of both lines have also to be consistent. We aggregate these two values using the average, as both measurements have the same importance to compute the final consistency value. There is neither change in nasal width with increased age, nor variation for ethnic group or BMI.

v **Frontal/central line parallelism comparison.** This is another way to analyze the anatomical correspondence in an SFO. The frontal line is achieved by joining the glabella and the gnathion landmarks both on the skull and on the facial image (see Fig. 3.v). As in the former criterion, the 3D landmarks are projected onto the photograph and the degree of consistency is computed using Eq. 3. Note that this criterion is only suitable for frontal view photographs. This relation does not change with age or BMI.

vi **Overlap of inter-dental lines.** When dental pieces appear in the AM photograph, experts give them special attention since the same object can be compared in both skull and face. Unfortunately, it is almost impossible that the whole tooth is shown in the photograph or kept in good condition in the skull. For this reason, we use the lines between the dental pieces to establish the comparison instead of the whole tooth. A graphical example of this criterion is shown in Fig. 3.vi. In this way, experts manually mark the corresponding lines on both surfaces. These marking lines are actually regions in the 3D skull model. Therefore, in the 2D image they also become regions (once they are projected using the geometric transformation of the SFO). In the photograph, lines are directly marked and then they

must be in the same location as the projected regions. We use Eq. 1 to measure that correspondence.

B. Implementation of Level 1: CFS Evaluation

From a theoretical point of view, the hierarchical framework for decision making proposed in [38] covers all sources of information as well as their aggregation and propagation. Nevertheless, there is still a need to provide implementation system for level 1 of the hierarchy, i.e., the CFS level. At this level, the matching degree of individual SFOs (belonging to different AM images of the same person) have to be fused to produce a unique and final CFS matching degree.

The influence of considering multiple SFOs belonging to the same person within the same CFS identification process was studied more than 20 years ago in [9], leading to a significant improvement of false matches (from 9% to 0.6%). Similarly, in another study developed in [6], the unknown skull was positively identified as the missing person, in 35 out of the 37 cases with more than one photograph available, based on more than 13 acceptable matches. In contrast, when the skulls in question were examined with only the frontal face photograph of the missing person (15 cases), the examiners could only find less than 12 matching criteria, leading to a probable identification rather than a positive one.

Within our DSS we could either produce a unique CFS matching degree in order to be able to order a set of given candidates according to this value, or we could instead provide a CFS matching degree together with a confidence degree of such a matching value. One of the main variables that should be used to produce a confidence value for a given CFS case is the quality/accuracy of the SFOs considered. However, we have not found the way to measure such a parameter, so in this study we focused on providing a single CFS value at Level 1, assuming that optimal SFOs have been achieved in the stage 2 of the whole CFS system.

Therefore, we propose to analyze the aggregation of the different SFO degrees of the same CFS case using the following four operators: the mean, the maximum, the minimum, and a weighting function based on the number of regions analized in each SFO. The latter function is defined in Eq. 5:

$$Agg\_Nreg = \frac{\sum_{i=1}^{N}(D\_SFO_i \cdot Nreg_i)}{\sum_{i=1}^{N} Nreg_i} \quad (5)$$

where N is the number of SFOs in the CFS case, D_SFO_n refers to the matching degree of ith SFO of the same CFS and $Nreg$ is the number of regions taken into account in a specific SFO.

C. Automatic CFS as sort-listing and identification tool

On the one hand, as described in [4] the CFS reliability studies reported in the literature [6], [8]–[10], [20], [21], [23] are fraught with limitations. Thus, it is not possible to draw a solid picture on CFS reliability.

On the other hand, the CFS technique has been performed by experts without a common standard throughout its existence. The first and major effort to develop a common

methodology occurred under the umbrella of the MEPROCS European project. Different quantitative and qualitative inter-lab studies served to develop the first best practice document in the field [41], identified the set of morphological criteria with a greater discriminative power [5], and validated the recently created methodology by measuring the performance of different forensic anthropologists in similar studies carried out following their own methodology and comparing it with the one developed within the MEPROCS framework [44].

Our proposed system follows the majority of MEPROCS recommendations (though not all, since the articulation of the mandible has not been addressed, and our system, due to its automatic nature, does not consider the physical skull but a 3D model representation). In this paper, and as a continuation of the works developed within the MEPROCS framework, we also present a complete methodology to evaluate the identification capabilities of a given method / process. It is composed of two different types of experiments.

The first one is designed to study the capability of our system for identifying the correct individual among a list of several negative cases and a single positive case. That is to say, the goal is to measure the sort-listing capabilities of a given system, for which Cumulative Match Characteristic (CMC) curves [45] are used. A CMC curve captures the percentage (or probability) that the correct match of a case appears in a list of r best matches, where r denotes the rank. In this rank, we also take into account the percentage with regard to the total sample size.

The second kind of tests are addressed to evaluate the performance of our system in the identification task, that is to say, a binary response of positive or negative identification must be provided for each particular CFS case. The same kind of test has already been carried out by MEPROCS partners in [41] and [44]. The experimental set up involves the comparison of unknown skulls and multiple candidates. For each single case, the experts were asked to report the final identification decision (either positive or negative) along with the rationale supporting the decision.

IV. EXPERIMENTS DEVELOPED AND ANALYSIS OF RESULTS

To analyze the performance of our approach, several detailed experiments have been developed, including the comparison with the state-of-the-art results of [38] and the manual results achieved by the renowned and junior forensic anthropologists in [41].

A. Experimental Design

The experimental design of this study consists of 100 AM photographs and 24 3D models of real Caucasian skulls. Nine of them are Cone Bean Computed Tomography (CBCT) models of living individuals, and the remaining 15 are skull 3D models of deceased people acquired using a 3D structure light scanner (seven of them acquired with the Artec MHT scanner and eight of them using the Fastscan Polhemus Scorpion scanner). In order to create the experimental dataset, a cross-comparison was performed, in which each skull 3D model

was superimposed with a variable number of photographs, obtaining 591 different SFOs. Each skull has one or more positive SFOs where the skull belongs to the subject of the AM photograph. A previous filter based on sex and age was made, so there is not the same number of overlays for each skull. At SFO level (level 2), the dataset is composed of 43 positive and 548 negative overlays. Since in some cases there is more than one AM photograph for the same person, we have to aggregate them by CFS cases. Thus, at CFS level (level 1), the experimental dataset involves 324 cases, 24 positive and 300 negative cases. Table II details the experimental dataset.

TABLE II: Experimental dataset summary

Skull model	Positive SFOs	Negative SFOs	Positive CFSs	Negative CFSs
CBCT 1	2	24	1	11
CBCT 2	2	24	1	11
CBCT 3	2	24	1	11
CBCT 4	2	24	1	11
CBCT 5	2	33	1	18
CBCT 6	2	33	1	18
CBCT 7	2	33	1	18
CBCT 8	2	33	1	18
CBCT 9	2	33	1	18
3D Model 10	3	24	1	14
3D Model 11	1	26	1	14
3D Model 12	2	25	1	14
3D Model 13	4	24	1	14
3D Model 14	1	27	1	14
3D Model 15	1	21	1	10
3D Model 16	3	19	1	10
3D Model 17	1	18	1	11
3D Model 18	1	17	1	10
3D Model 19	1	16	1	10
3D Model 20	2	16	1	11
3D Model 21	1	16	1	10
3D Model 22	1	14	1	9
3D Model 23	1	14	1	9
3D Model 24	2	10	1	6
Total	43	548	24	300

The SFOs employed have been obtained by our automatic method in [46] using the same parameter values reported in that work. Exceptionally, the CBCTs positive cases have been achieved by means of a ground truth dataset, whose overlays are assumed to be optimal, according to [47].

The aim of this experiment is to determine the optimal design for our DSS in the identification task. To do so, we have to analyze different aggregation functions proposed for each level or sublevel. Then, the best of them is chosen based on the accuracy index by ranking the positive and negative cases. As a final step, a threshold has to be set for labeling each case as positive or negative.

The experimental dataset is divided into training and test sets to validate the system in a correct way. 74.1% of the instances compose the training dataset. Thus, it is composed of 240 CFS cases, 16 positive (CBCTs 1 to 9 and 3D Models 10 to 16) and 224 negative ones. In terms of SFO level, this corresponds to 33 positive and 427 negative SFOs. Meanwhile, the remaining 25.9% of the cases form the test set, with 84 CFS cases, 8 positive (3D Models 17 to 24) and 76 negative ones. Correspondingly, this set has 129 cases at SFO level, 10 positive and 121 negative SFOs (3D Models 17 to 24). Fig. 4 summarizes the structure of the dataset
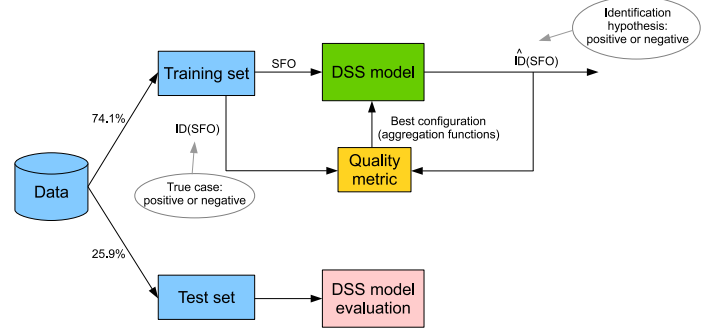

Fig. 4: Experimentation setup diagram.

and the experimentation procedure. Notice that, due to the intrinsic nature of CFS, we are dealing with an unbalanced classification problem [48].

Following our DSS framework, the age and the BMI have to be set for each available photograph. Additionally, the available regions of each photograph have to be marked and the related quality of each one established on a scale between 0 and 1. In this study, forensic experts must delineate from each image from one to eight regions in each image (depending on their visibility): chin contour, cranial contour, eyeball center right, eyeball center left, nose width, mouth occlusion, tragus right and tragus left; and from zero to 18 inter-dental lines. The corresponding 26 zones are also marked on the skull 3D models. Again, experts set the quality of the bone region based on the weathering stages. Then, the anthropometric landmarks used to achieve the SFOs are considered to form the other two criteria: ectocanthion lines and frontal central lines (both in the 3D model skulls and images). Figure 5 depicts an example of some of these regions and landmarks marked on a skull 3D model and on a photograph.

The experts established that the uncertainty sources have a third of the influence and the matching degree has two thirds influence [38]. Thus, the weighted vector used in level 2.2 is $\mathbf{w} = (\frac{1}{3}, \frac{2}{3})$.

Finally, the discriminative power of each region has to be determined using the training dataset. This value is used as a weight for the aggregation at level 2.3. Then, with these parameters, we study the same aggregation operators for each sub-level (2.1, 2.2 and 2.3) as in [38]. Since in the current contribution we analyze more criteria and regions, the results can vary. Next, we need to study the behavior of our system using a threshold to give the final decision as a positive or a negative case. Finally, we validate the obtained results over the test dataset using the obtained DSS design and threshold.

The influence related to the biological profile (only relevant for the chin criterion) for the implemented criteria was defined by Prof. Wilkinson according to her expert knowledge and represented using fuzzy sets.

B. Performance Analysis of the DSS

In this experiment we aim to find the best configuration for our system within the proposed framework. A similar approach for the experiment was taken in [38]. However, in this study, more criteria and regions are considered. In addition, we
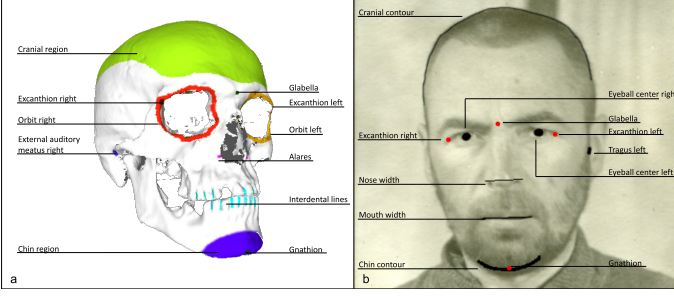
Fig. 5: Example of marking regions on a skull 3D model (a) and on a photograph (b).

use a bigger dataset (24 vs 16 3D skull models, 100 vs 66 AM photographs, 591 SFOs vs 444) with more positive and negative cases. Here, we also implement the final level, so the results will be expressed in terms of CFS case, and in consequence, the system will be able to give a binary response, i.e., a positive or a negative identification.

The training dataset (see Section IV-A) is used to set the best configuration of our system. To do so, the following steps are performed:

1) Calculating the power of identification over the training dataset: this is understood as the capability of each criterion to discriminate in the decision making process for each isolated criterion. To perform this task, we use the methodology presented in [35]. Firstly, the matching degree of the craniofacial correspondence by only using a specific criterion is calculated over the database of cases. These values are used to rank the candidates from 1 to $n(the number of candidates)$ based on their chance to be the actual subject. Then, a value between 0 and 1 is assigned to each positive case taking into account the ranking. The formula to assign the power of identification of the criterion W_i in the instance j is:

$$Power_{Id}(W_i)_j = 1 - \frac{r-1}{M_j - 1} \qquad (6)$$

where r is the position of the positive case in the ranking and M_j is the lowest value of the ranking for instance j (all cases getting the same criterion-method value are supposed to have a draw, that is, they are assigned to the same ranking). Finally, the average of all these values over all cases is calculated, getting the value $Power_{Id}(W_i)$ for each criterion over the whole database.

2) Analyzing and comparing the performance of different aggregation functions within our framework at the SFO level: in order to calculate the most suitable functions for our system, we obtain the accuracy degree for identifying positive cases in each case at the SFO evaluation level (level 2). The process is similar to the experiment developed in [38]. Nevertheless, in this case we have different power of identification degrees (previously calculated using the new dataset) as well as more criteria and regions. The way to obtain the accuracy degree at SFO level follows the same methodology as before. First, all the aspects that are involved in the SFO evaluation are

aggregated (quality of the materials, biological profile, craniofacial matching degree of each region, and power of identification), obtaining the degree of an SFO. We analyze the same aggregation functions as in [38]. That is, for sublevel 2.1, the minimum (*min*), the product (*prod*), and the arithmetic mean (*mean*). For sublevel 2.2, the weighted arithmetic mean (*wam*) and the weighted geometric mean (*wgm*). Finally, for sublevel 2.3, the *wam*, the *wgm*, the Sugeno (*sug*) integral, and the Choquet (*choq*) integral [40]. Accordingly, 24 different combinations are analyzed. The definitions of these operators and the justification for testing them in each sublevel are detailed in [38]. The obtained SFO degree is used to rank the candidates in decreasing order. Again, a value between 0 and 1 is assigned to each positive case taking into account this ranking. The formula to obtain the accuracy for each configuration of the system is:

$$ACC(Combination_a)_j = 1 - \frac{r-1}{M_j - 1} \qquad (7)$$

where r is the position of the positive case in the ranking and M_j is the lowest value of the ranking for instance j. The final step calculates the average of these values over all the cases, $ACC(Combination_a)$. The parameters that obtain the highest accuracy will be selected for the configuration of the system.

3) Analyzing and comparing the performance of different aggregation functions within our framework at level 1: the final CFS degree is obtained by aggregating the degrees of all the SFOs that belong to the same case. As explained in section III-B, we study four different operators at this level: mean, minimum, maximum, and weighted average by the number of regions. These SFO degrees are obtained using the best combination of operators from the previous step. The accuracy degree at this level serves to identify the most appropriate operator following the same method as in the previous stage.

4) Establishing a threshold for labeling each CFS case as positive or negative. The CFS degree is given in the interval $[0, 1]$. The threshold sets the limit to consider the case as negative (below this value) or positive (above this value). In order to measure the effectiveness of our system for each threshold value, an appropriate evaluation metric has to be used. As stated, the nature of the data is considered imbalanced since it exhibits an unequal distribution between the two classes, positive and negative identification. In fact, negative cases are much more common than positive cases. For this reason, we analyse the best threshold for the performance of our DSS using the G-Mean metric [48]:

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \qquad (8)$$

where TP=true positive, TN=true negative, FP=false positive, and FN=false negative. This metric evaluates the degree of inductive bias in terms of a ratio of positive and negative accuracy. The threshold which achieves the highest value of this metric will be considered the most appropriate for the performance of our system.

Finally, the test dataset is used to validate the system performance. The designed DSS is applied to this set of cases and the value of the $G - mean$ metric is computed as a final performance measure, together with some other imbalanced classification measures.

C. Results

The discriminative power for each criterion obtained using Eq. 6 over the training dataset is shown in Table III. With these values, the next experiment consists of analyzing the behavior of the 24 combinations of the aggregation functions at level 2. To do so, we calculate the accuracy using the values of the 24 different configurations over all the training cases in Eq. 7 as explained in Section IV-B.

TABLE III: The power of identification of each isolated region (Criterion level, level 3).

Region	Power of identification
Chin contour	0.75
Cranial contour	0.75
Eyeball (left and right) center position	0.72
Mouth occlusion	0.67
Nose width	0.83
Ear (left and right) position	0.50
Inter-dental lines overlap	0.33
Ectocanthion lines	0.60
Frontal central lines	0.54

TABLE IV: Mean accuracy of each combination method at SFO level (level 2).

Combination Method	Mean accuracy	Combination Method	Mean accuracy
mean-wgm-wgm	**0.828**	*min-wgm-sug*	0.718
mean-wam-wgm	0.825	*min-wgm-choq*	0.716
mean-wam-wam	0.809	*mean-wgm-choq*	0.648
min-wgm-wgm	0.808	*mean-wgm-sug*	0.648
mean-wgm-wam	0.807	*mean-wam-sug*	0.644
min-wam-wgm	0.807	*prod-wgm-choq*	0.627
prod-wam-wgm	0.805	*mean-wam-choq*	0.625
prod-wgm-wgm	0.804	*prod-wgm-sug*	0.616
min-wam-wam	0.787	*prod-wam-sug*	0.616
prod-wam-wam	0.787	*min-wam-sug*	0.613
min-wgm-wam	0.786	*prod-wam-choq*	0.609
prod-wgm-wam	0.785	*min-wam-choq*	0.609

Table IV details the mean accuracy for the DSS at level 2 using each combination of the aggregation functions in decreasing order of performance. The combination *mean-wgm-wgm* achieves the highest value, with 0.828 of accuracy. In general, we can see that the best aggregation function for the sublevel 2.3 is *wgm* and then, *wam*.

Therefore, we use the best configuration (*mean-wgm-wgm*) for the next experiments performed at level 1. The obtained results for the analysis of the four different ways to aggregate the SFO degrees are reported in Table V.

As can be seen, the best performance of the DSS is achieved aggregating the SFO degrees, using a weighted average by the number of regions. The accuracy obtained with this operator is 0.917, while for the remaining functions the value does not reach a 0.9 of accuracy.

TABLE V: Mean accuracy of the DSS using each aggregation operator at CFS level (level 1).

Aggregation operator	Mean accuracy
Weighted average by the number of regions	**0.917**
Mean	0.893
Minimum	0.860
Maximum	0.839

TABLE VI: Results of the DSS for different values of the threshold.

Threshold	TP	FN	TN	FP	*G-Mean*
0.80	13	3	153	71	0.745
0.81	13	3	156	68	0.752
0.82	13	3	158	66	0.757
0.83	13	3	161	63	0.764
0.84	13	3	166	58	0.776
0.85	13	3	180	44	0.808
0.86	13	3	186	38	**0.821**
0.87	12	4	190	34	0.798
0.88	11	5	200	24	0.783
0.89	10	6	203	21	0.753
0.90	7	9	214	10	0.646

The latter configuration of the system is used to choose the most appropriate threshold for labeling each case as positive or negative. For the final identification task, ten different values are tested for the threshold: from 0.80 to 0.90, with steps of 0.01. Table VI summarizes the obtained results after applying the methodology explained in Section IV-B and calculating the $G - mean$ metric (Eq. 8) for each threshold value.

As can be seen in Table VI, a threshold of 0.86 achieves the highest $G - mean$ value. Concerning this result, our DSS correctly classified a relatively high number of positive (13 of 16, i.e. 81.25%) and negative cases (186 of 224, i.e. 83.03%). Therefore, the best performance of our identification system is reached using the configuration captured in Table VII.

The test dataset is used to validate the former configuration of our DSS. Table VIII details the obtained results after applying the system over the set of test cases. Apart from reporting the G-Mean value, the accuracy of the system is calculated using Eq. 9 [48]:

$$Accuracy = \frac{TP + TN}{P_c + N_c}, \qquad (9)$$

where P_c and N_c are the total number of positive and negative cases, respectively; and the rate of true positives and true negatives are defined as:

$$TP_{rate} = \frac{TP}{P_c}; \qquad TN_{rate} = \frac{TN}{N_c}. \qquad (10)$$

D. Comparison between our automatic DSS and a manual approach performed by several forensic experts

In this experiment, we aim to compare the performance of real forensic practitioners with that of our CFS DSS. To do so, our system is applied to the same experimental dataset of [41]. In that study, 26 participants from 17 different institutions were asked to deal with 14 identification scenarios, some of them involving the comparison of multiple candidates and unknown

TABLE VII: Configuration for the best performance of our automatic DSS

Framework level	Task	Operator
Level 2.1	Aggregate the sources of uncertainty of bone and image and biological profile	Mean
Level 2.2	Aggregate the skull-face matching degree of each region and the uncertainty of level 2.1	Weighted geometric mean
Level 2.3	Aggregate all level 2.2 degrees weighting by the discriminative power of each region	Weighted geometric mean
Level 1	Aggregate all the SFO degrees of the same case	Weighted average by number of regions
Threshold	Classify each case as positive or negative	0.86

TABLE VIII: Performance of the DSS over the test dataset.

TP	FN	TN	FP	TP_{rate}	TN_{rate}	Accuracy	*G-Mean*
5	3	74	2	0.625	0.974	0.940	0.780

skulls. A total number of 60 SFO problems were tackled. Table IX shows the mean value of the results of the 26 experts, the results of the three best experts, and the outcomes of our automatic DSS. Detailed performance indicators are shown such as the percentage of correct decisions, the number of positive and negative decisions given in each case, and the corresponding rate of true and false positives and true and false negatives. 'Ground Truth' refers to the real nature of each CFS case, i.e., positive when the skull and the facial photographs belong to the same person, negative in the contrary case. DSS-0.86 refers to our DSS method using the best configuration set and validated in section IV-B.

TABLE IX: Performance of the different CFS approaches for test 2 (P=positive, N=Negative)

Method	Correct Decisions	Ground Truth	Decision P	Decision N	Decision(%) P	Decision(%) N
Expert Mean	78.99%	P	100	90	52.63%	47.37%
		N	152	810	15.80%	84.20%
Best Expert 1	93.33%	P	8	2	80.00%	20.00%
		N	2	48	4.00%	96.00%
Best Expert 2	88.14%	P	6	3	66.67%	33.33%
		N	4	46	8.00%	92.00%
Best Expert 3	86.21%	P	5	3	62.50%	37.50%
		N	5	45	10.00%	90.00%
DSS-0.86	90.00%	P	6	4	60.00%	40.00%
		N	2	48	4.00%	96.00%

In view of these results, and according to the conclusions from [41], experts generally achieve higher rates of TN than TP. Table IX shows that the mean TP rate was 52.63% while the mean TN rate was 84.20%. We can observe that the three best participants achieved TN rates equal or higher than 90%. Meanwhile, the same three experts achieved 80.00%, 66.67%, and 62.50% of TP rates, respectively. The overall accuracy (total correct decisions) is 78.99% on average, and 93.33%, 88.14%, and 86.21% for the three best participants, respectively. The overall performance (correct decision rate) of our system is better than the mean of the experts with respect to the correct decisions rate (90.00%) and better than 25 of the 26 forensic experts who participated in the study. The TN rate is also impressive (96.00%), exactly the same as the best performing expert. Meanwhile, the TP rate is not so high (60.00%) but still above the expert's mean and extremely similar to that of the third best performing expert.

E. Performance Analysis of the DSS as a sort-listing tool

The final experiment of this study aims to assess the capability of our system to identify the correct individual from a list with one positive and several negative cases. Note that for this analysis, the threshold value is not needed since we use the CFS degree to rank the candidates. We compare our results with the state-of-the-art results in [38]. Fig. 6 shows the CMC curves of the new system and the results obtained by that proposed in [38] (green curve), that considered less criteria (only four) and did not include the first level of the hierarchy. At first glance we can see the developed system (red curve) achieves better results than the state-of-the-art version. The graphical results clearly indicate that the new DSS significantly improves the one proposed in [38]. In this previous version, for the 7% of the total cases (rank 2), the tool correctly ranked less than 40% of the cases. Now, the new system is able to rank more than 60% of the cases for the same percentage (see rank 1). In addition, it ranks among the three first positions (rank 3) in almost 90% of the cases. This corresponds with 20% of the total cases of the data set. As shown in Fig. 6, in order to achieve the same percentage with the previous version of the tool, it was necessary to reach rank 9 (33% of the samples).
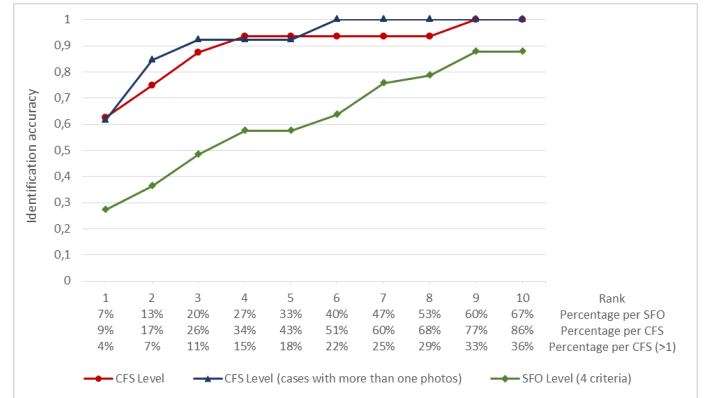

Fig. 6: CMC curves corresponding to the state-of-the-art results in [38] (green curve) and the new experimentation at CFS level with all the cases (red curve) and only with the cases with more than one photographs (blue curve).

Finally, we analyze the performance of the system ignoring the cases with only one image (blue curve in the same Fig. 6). Nowadays, the most common scenario is that the relatives of a missing person possess numerous AM photographs of him/her. We want to study the effect of taking into account only the cases of our data set with more than one SFO. As can be seen, our system is still not able to identify correctly every case (rank 1 values are still in the 60% range). However, the

performance improves for the sort listing task. It ranks all the positive cases in the six first positions (51% of the data set). Meanwhile, using all the cases the system has to reach 60% of the samples to get the same value.

V. Discussion and Conclusions

In this study, we have examined and modeled six new skull-face morphological criteria for our CFS DSS. In this way, the craniofacial correspondence evaluation comprises the whole face. This fact changes the parameter configuration obtained in our previous proposal [38]. There, we only studied up to four regions per image and we calculated the best performance for the three sublevels of the SFO evaluation (level 2). In that case, the best configuration of aggregation functions was *mean-wam-wam*. In this research, we have considered up to eight regions and 18 inter-dental lines. In addition, we have used a larger dataset than in the previous experiment. The best current performance is obtained using the combination of aggregation functions *mean-wgm-wgm*. In addition, we have implemented the CFS evaluation level (level 1) by studying four possible aggregation operators. Thus, in an innovative way, we present a complete DSS which provides a final degree of craniofacial correspondence. The best behavior of the system is achieved when the different SFO degrees of the same case are aggregated by weighting using the number of regions considered in each SFO. Finally, we have studied several thresholds for establishing a binary classification (positive or negative), obtaining 0.86 as the best value.

These parameters have been achieved over a training data set. For the first time, the performance of the system has been evaluated using a different unseen set of cases (test data set). From the analysis of the experiment developed we can conclude that our current system is suitable for:

- Filtering (sort-listing) cases: in 62.5% of the cases the positive one is ranked in the first position. Besides, given a data set of initial candidates, the positive case was included in the 60% best ranked candidates with a 100% of probability and within the 27% best ranked candidates with a 92% of probability.
- Establishing exclusion: the ability to determine a negative identity was performed with 97.4% of accuracy (97.4% of true negative over the test data set and 96.0% over the training data set).

Thus, the designed DSS can be considered the first automatic tool for classifying couples of unknown faces and skulls as positive or negative cases with an accuracy similar to the best performing forensic experts. However, we still cannot describe CFS as a solid identification technique. Although it has been applied and developed for more than a century, its reliability is still unclear. On the one hand, it is essential to work with large datasets to reinforce the conclusions of our approach. Enlarging the size of the dataset presents many opportunities to generalize the process as well as increasing the likelihood of producing consistent, accurate, and reproducible results. On the other hand, an objective and more precise automatic system, which considers more information from different sources, is necessary. As we noted in [38], there are some types of uncertainty and degrees of confidence which are still ignored, such as the quality of the SFOs achieved in the second stage (highly improbable); the 3D and 2D regions delimitations; and the number of regions evaluated. We have already increased this number from four (in [38]) to nine and we will work on including additional zones under the same rigorous approach, since the literature has shown the importance of evaluating more regions. For instance, authors in [6] use more than 13 criteria to make identification decisions and a similar approach is followed in [8].

We are aware that the quality of the SFOs has a strong influence on the performance of the DSS. For this reason, future work will be needed in order to enhance this quality: a new parametrization of the camera and an innovative design of the optimization algorithms for the SFO stage. Related to this, the modeling of the mandible articulation is crucial for improving the overlays obtained by the superimposition algorithm. Furthermore, developing new accurate soft tissue studies, by including spatial directions and a robust statistic metric, could improve the performance of the method.

References

[1] Paul T Jayaprakash. Practical relevance of pattern uniqueness in forensic science. *Forensic Science International*, 231(1):403–e1, 2013.

[2] M. Yoshino. Craniofacial superimposition. In C. Wilkinson and C. Rynn, editors, *Craniofacial Identification*, pages 238–253. University Press, Cambridge, 2012.

[3] S. Damas, O. Cordón, O. Ibáñez, J. Santamaría, I. Alemán, M. Botella, and F. Navarro. Forensic identification by computer-aided craniofacial superimposition: a survey. *ACM Computing Surveys*, 43(4):27, 2011.

[4] M. I. Huete, O. Ibáñez, C. Wilkinson, and T. Kahana. Past, present, and future of craniofacial superimposition: Literature and international surveys. *Legal Medicine*, 17:267–278, 2015.

[5] S. Damas, C. Wilkinson, T. Kahana, E. Veselovskaya, A. Abramov, R. Jankauskas, P.T. Jayaprakash, E. Ruiz, F. Navarro, M.I. Huete, E. Cunha, F. Cavalli, J. Clement, P. Leston, F. Molinero, T. Briers, F. Viegas, K. Imaizumi, D. Humpire, and O. Ibáñez. Study on the performance of different craniofacial superimposition approaches (ii): best practices proposal. *Forensic Science International*, 257:504–508, 2015.

[6] M. Yoshino, K. Imaizumi, S. Miyasaka, and S. Seta. Evaluation of anatomical consistency in craniofacial superimposition images. *Forensic Science International*, 74(1):125–134, 1995.

[7] Y. Lan. A study on national differences in identification standards for chinese skull-image superimposition. *Forensic Science International*, 74(1-2):135–153, 1995.

[8] D. S. Chai, Y. W. Lan, C. Tao, R. J. Gui, Y. C. Mu, J. H. Feng, W. D. Wang, and J. Zhu. A study on the standard for forensic anthropologic identification of skull-image superimposition. In *Bulletin du Service de Documentation Generale*, volume 79, pages 269–77. Organization Internationale de Police Criminelle (INTERPOL), 1992.

[9] D. Austin-Smith and W. R. Maples. The reliability of skull/photograph superimposition in individual identification. *Journal of Forensic Science*, 39(2):446–455, 1994.

[10] P. T. Jayaprakash, G. J. Srinivasan, and M. G. Amravaneswaran. Craniofacial morphanalysis: a new method for enhancing reliability while identifying skulls by photo superimposition. *Forensic Science International*, 117(1):121–143, 2001.

[11] L. G. Farkas, M. J. Katic, and C. R. Forrest. Comparison of craniofacial measurements of young adult african-american and north american white males and females. *Annals of Plastic Surgery*, 59(6):692–698, 2007.

[12] Leslie G Farkas and Curtis K Deutsch. Anthropometric determination of craniofacial morphology. *American journal of medical genetics*, 65(1):1–4, 1996.

[13] R. M. George. Anatomical and artistic guidelines for forensic facial reconstruction. In M. Y. Iscan and R. Helmer, editors, *Forensic Analysis of the Skull*, pages 215–227. Wiley Liss, New York, NY, 1993.

[14] J. Caple and C. N. Stephan. A standardized nomenclature for craniofacial and facial anthropometry. *International Journal of Legal Medicine*, 130(3):863–879, 2016.

[15] C. N. Stephan and E. K. Simpson. Facial soft tissue depths in craniofacial identification (part i): an analytical review of the published adult data. *Journal of Forensic Science*, 53:1257–1272, 2008.

[16] M. Prokopec and D. H. Ubelaker. Reconstructing the shape of the nose according to the skull. *Forensic Science Communications*, 4(1), 2002.

[17] C. N. Stephan. Facial approximation: An evaluation of mouth-width determination. *American Journal of Physical Anthropology*, 121(1):48–57, 2003.

[18] C. N. Stephan and S. J. Murphy. Mouth width prediction in craniofacial identification: cadaver tests of four recent methods, including two techniques for edentulous skulls. *Journal of Forensic Odonto-Stomatology*, 27(1):2–7, 2008.

[19] C. Wilkinson and S. A. Mautner. Measurement of eyeball protrusion and its application in facial reconstruction. *Journal of Forensic Science*, 48(1):12–16, 2003.

[20] A. Ricci, G. L. Marella, and M. A. Apostol. A new experimental approach to computer-aided face/skull identification in forensic anthropology. *The American Journal of Forensic Medicine and Pathology*, 27(1):46–49, 2006.

[21] G. Gordon. *An investigation into the accuracy and reliability of skull-photo superimposition in a South African sample*. PhD thesis, University of Pretoria, 2011.

[22] T. W. Fenton, A. N. Heard, and N. J. Sauer. Skull-photo superimposition and border deaths: Identification through exclusion and the failure to exclude. *Journal of Forensic Sciences*, 53(1):34–40, 2008.

[23] D. Gaudio, L. Olivieri, D. De Angelis, P. Poppa, A. Galassi, and C. Cattaneo. Reliability of craniofacial superimposition using three-dimension skull model. *Journal of Forensic Sciences*, 61(1):5–11, 2016.

[24] O. Ibáñez, O. Cordón, S. Damas, and J. Santamaría. An experimental study on the applicability of evolutionary algorithms to craniofacial superimposition in forensic identification. *Information Science*, 79:3998–4028, 2009.

[25] C. N. Stephan. Perspective distortion in craniofacial superimposition: Logarithmic decay curves mapped mathematically and by practical experiment. *Forensic Science International*, 257:520–e1, 2015.

[26] M. Cummaudo, M. Guerzoni, L. Marasciuolo, D. Gibelli, A. Cigada, Z. Obertovà, Z. Ratnayake, P. Poppa, P. Gabriel, S. Ritz-Timme, et al. Pitfalls at the root of facial assessment on photographs: a quantitative study of accuracy in positioning facial landmarks. *International Journal of Legal Medicine*, 127(3):699–706, 2013.

[27] B. A. Nickerson, P. A. Fitzhorn, S. K. Koch, and M. Charney. A methodology for near-optimal computational superimposition of two-dimensional digital facial photographs and three-dimensional cranial surface meshes. *Journal of Forensic Science*, 36:480–500, 1991.

[28] W. Jin, G. Geng, K. Li, and Y. Han. Parameter estimation for perspective projection based on camera calibration in skull-face overlay. In *Virtual Reality and Visualization (ICVRV), 2013 International Conference on*, pages 317–320. IEEE, 2013.

[29] B. Zitovà and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.

[30] D. V. Pesce, E. Vacca, F. Potente, F. Lettini, and M. Colonna. *Shape analytical morphometry in computer-aided skull identification via video superimposition*. Iscan MY, Helmer RP. Forensic analysis of the skull: craniofacial analysis, reconstruction, and identification. New York: Wiley-Liss, 1993.

[31] M. Yoshino, H. Matsuda, S. Kubota, K. Imaizumi, S. Miyasaka, and S. Seta. Computer-assisted skull identification system using video superimposition. *Forensic Science International*, 90(3):231–244, 1997.

[32] Soft computing applications for complex environments research group. http://sci2s.ugr.es/soccer/. [Online; accessed 2017].

[33] O. Ibáñez, O. Cordón, S. Damas, and J. Santamaría. Modeling the skull-face overlay uncertainty using fuzzy sets. *IEEE Transactions Fuzzy Systems*, 16:946–959, 2011.

[34] B. R. Campomanes-Álvarez, O. Ibáñez, F. Navarro, I. Alemán, O. Cordón, and S. Damas. Dispersion assessment in the location of facial landmarks on photographs. *International Journal of Legal Medicine*, 129(1):227–236, 2015.

[35] C. Campomanes-Alvarez, O. Ibáñez, and O. Cordón. Modeling the consistency between the bony and facial chin outline in craniofacial superimposition. In *16th World Congress of the International Fuzzy Systems Association (IFSA)*, pages 1612–19, 2015.

[36] A. Passalacqua N. Christensen and E. Bartelink. *Forensic Anthropology: Current Methods and Practice*, volume 9780124186712. Elsevier, 2014.

[37] C. Campomanes-Alvarez, O. Ibáñez, and O. Cordón. Design of criteria to assess craniofacial correspondence in forensic identification based on computer vision and fuzzy integrals. *Applied Soft Computing*, 46:596–612, 2016.

[38] C. Campomanes-Alvarez, O. Ibáñez, O. Cordón, and C. Wilkinson. Hierarchical information fusion for decision making in craniofacial superimposition. *Information Fusion*, 39:25–40, 2018.

[39] C. Campomanes-Alvarez, O. Ibáñez, and O. Cordón. Experimental study of different aggregation functions for modeling craniofacial correspondence in craniofacial superimposition. In *the 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2016)*, pages 437–444, 2016.

[40] M. Sugeno. Fuzzy measures and fuzzy integrals: a survey. *Fuzzy Automata and Decision Processes*, 78(33):89–102, 1977.

[41] O. Ibáñez, R. Vicente, D. S. Navega, C. Wilkinson, P. T. Jayaprakash, M I. Huete, T. M. Briers, R. Hardiman, F. Navarro, E. Ruiz, F. Cavalli, K. Imaizumi, R. Jankauskas, E. Veselovskaya, A. Abramov, P. Lestón, F. Molinero, J. Cardoso, J. Cagdir, D. Humpire, Y. Nakanishi, A. Zeuner, A. H. Ross, D. Gaudio, and S. Damas. Study on the performance of different craniofacial superimposition approaches (i). *Forensic Science International*, 257:496–503, 2015.

[42] T Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Kongelige Danske Videnskabernes Selskab*, 5:1–34, 1948.

[43] S Damas, O Ibáñez, and O Cordón. *Handbook on craniofacial superimposition*. Springer, 2017. In press.

[44] O. Ibáñez, R. Vicente, D. Navega, C. Campomanes-Álvarez, C. Cattaneo, R. Jankauskas, M. I. Huete, F. Navarro, R. Hardiman, E. Ruiz, et al. MEPROCS framework for craniofacial superimposition: Validation study. *Legal Medicine*, 23:99–108, 2016.

[45] A. K. Jain and S. Z. Li. *Handbook of face recognition*. Springer, 2005.

[46] B. R. Campomanes-Álvarez, O. Ibáñez, C. Campomanes-Álvarez, S. Damas, and O. Cordón. Modeling the facial soft tissue thickness for automatic skull-face overlay. *IEEE Transactions on Information Forensics and Security*, 10:2057–2070, 2015.

[47] O. Ibáñez, F. Cavalli, B. R. Campomanes-Álvarez, C. Campomanes-Álvarez, A. Valsecchi, and M. I. Huete. Ground truth data generation for skull–face overlay. *International Journal of Legal Medicine*, 129(3):569–81, 2015.

[48] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.