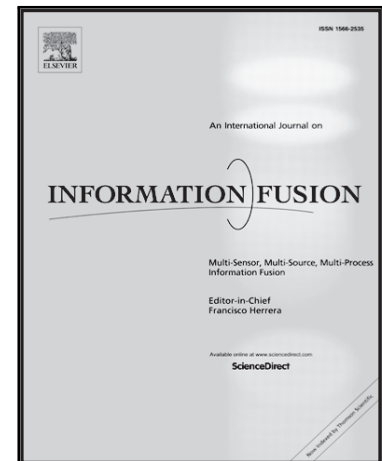


Accepted Manuscript

Hierarchical Information Fusion for Decision Making in Craniofacial Superimposition

Carmen Campomanes-Alvarez, Oscar Ibáñez, Oscar Cordón,
Caroline Wilkinson

PII: S1566-2535(16)30217-2
DOI: [10.1016/j.inffus.2017.03.004](https://doi.org/10.1016/j.inffus.2017.03.004)
Reference: INFFUS 858



To appear in: *Information Fusion*

Received date: 9 December 2016
Revised date: 2 March 2017
Accepted date: 19 March 2017

Please cite this article as: Carmen Campomanes-Alvarez, Oscar Ibáñez, Oscar Cordón, Caroline Wilkinson, Hierarchical Information Fusion for Decision Making in Craniofacial Superimposition, *Information Fusion* (2017), doi: [10.1016/j.inffus.2017.03.004](https://doi.org/10.1016/j.inffus.2017.03.004)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights

- There are automatic systems based on Soft Computing for craniofacial superimposition.
- This is the first time that the decision making stage is modeled to help experts.
- The design of the decision making system is based on fuzzy operators.
- The system is validated using real identification cases of an European project.
- This proposal can be used as a shortlisting tool capable of filtering out candidates.

Hierarchical Information Fusion for Decision Making in Craniofacial Superimposition

Carmen Campomanes-Alvarez^{a,*}, Oscar Ibáñez^a, Oscar Cordon^a, Caroline Wilkinson^b

^a*Department of Computer Science and Artificial Intelligence, University of Granada,
C/ Daniel Saucedo Aranda, s/n, 18071 Granada, Granada, Spain.*

^b*School of Art and Design, Liverpool John Moores University, Liverpool L3 5TF, UK*

Abstract

Craniofacial superimposition is one of the most important skeleton-based identification methods. The process studies the possible correspondence between a found skull and a candidate (missing person) through the superimposition of the former over a variable number of images of the face of the latter. Within craniofacial superimposition we identified three different stages, namely: 1) image acquisition-processing and landmark location; 2) skull-face overlay; and 3) decision making. While we have already proposed and validated an automatic skull-face overlay technique in previous works, the final identification stage, decision making, is still performed manually by the expert. This consists of the determination of the degree of support for the assertion that the skull and the ante-mortem image belong to the same person. This decision is made through the analysis of several criteria assessing the skull-face anatomical correspondence based on the resulting skull-face overlay. In this contribution, we present a hierarchical framework for information fusion to support the anthropologist expert in the decision making stage. The main goal is the automation of this stage based on the use of several skull-face anatomical criteria combined at different levels by means of fuzzy aggregation functions. We have implemented two different experiments for our framework. The first aims to obtain the most suitable aggregation functions for the system and the second validates the proposed framework as an identification system. We tested the framework with a dataset of 33 positive and 411 negative identification instances. The present proposal is the first automatic craniofacial superimposition decision support system evaluated in an objective and statistically meaningful way.

Keywords: forensic anthropology, craniofacial superimposition, decision making, information fusion, fuzzy aggregation operators, computer vision.

1. Introduction

Craniofacial superimposition (CFS) [1] is the most representative technique within craniofacial identification [2]. It involves superimposing a skull onto a one or more ante-mortem (AM) photographs of a missing person. The consequent analysis of their morphological correspondence determines if they belong to the same subject.

The whole CFS process can be divided into three consecutive stages [3] (Fig. 1): 1) The acquisition and processing of the materials, i.e, skull and 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 getting different overlays. 3) Decision making process 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

*Corresponding author

Email addresses: carmen.campomanes@decsai.ugr.es (Carmen Campomanes-Alvarez), oscar.ibanez@decsai.ugr.es (Oscar Ibáñez), ocordon@decsai.ugr.es (Oscar Cordon), C.M.Wilkinson@ljmu.ac.uk (Caroline Wilkinson)

different criteria based on the anatomical relationship between the face and the skull. These criteria can vary depending on the region and the pose [4].

Designing automatic methods to address CFS and support the forensic anthropologist remains a challenge and dreamed milestone within the anthropology community. In fact, the development of computer-aided CFS methods has increased over the past twenty years [5]. Recent approaches use skull 3D models and soft computing (SC) methods for the first two identification stages. These methods allow us to both automate the task and handle the inherent uncertainty [6, 7, 8, 9].

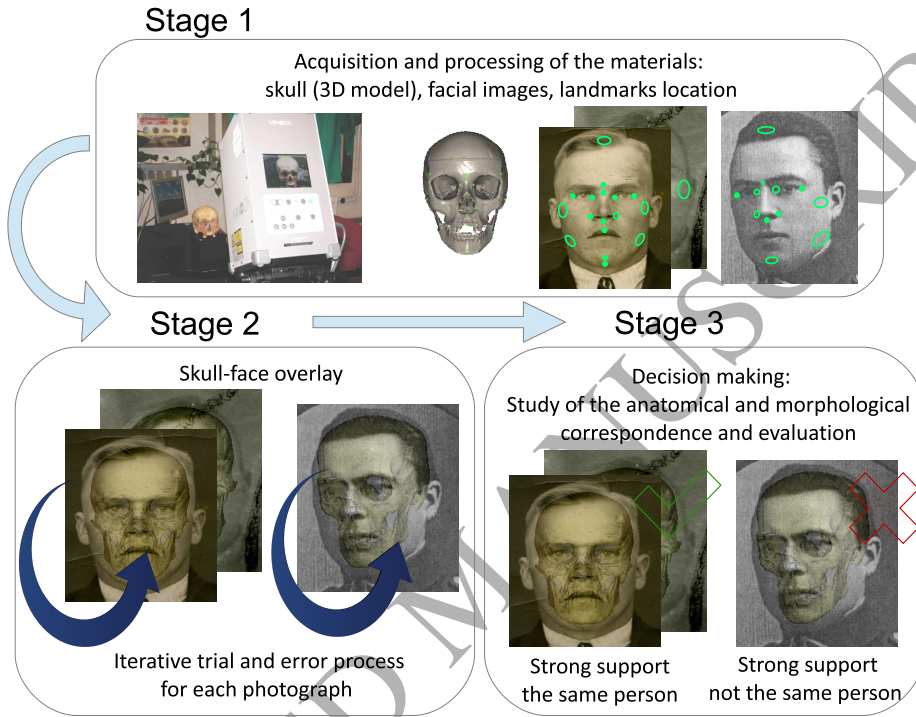


Figure 1: CFS procedure scheme.

In the third stage, once one or more SFOs are obtained, experts evaluate morphological and spatial skull-face relations, focusing on certain regions demonstrated to be more discriminative. The final decision is provided from an aggregation of the partial decisions. It is taken in terms of limited, moderate or strong support to the assumption that the skull and the facial image belong to the same individual or not [4]. This process is subjective and it relies on the skills of the forensic expert while influenced by the quantity and quality of the used materials. Therefore, there is a need to design a decision support system (DSS) to help practitioners to make their decision based on the fusion of the available information sources in a faster and more objective way. It would also lead to the application of CFS in identification scenarios with multiple comparisons, a possibility not explored yet due to the unaffordable time lapse needed to analyse all possible cross-comparisons. Our long-term and complex goal is the implementation of a DSS by evaluating the spatial and morphological relationships. The system will return a numeric index as output, in order to support the forensic anthropologist to make the final CFS decision while automatically filtering a number of cases in multiple comparison scenarios.

In previous works [10, 11, 12], we presented a first preliminary proposal to design a DSS for CFS using computer vision (CV) and fuzzy integrals. We implemented two of the most discriminative criteria to assess craniofacial correspondence, namely, the spatial and the morphological relation between the bony and facial chin, and the relative position of the orbits and the eyeballs.

In this work we present a complete hierarchical DSS for CFS with three connected levels of decision. The previous studies only tackled what we currently identified as the third and simplest level. We only

implemented some CV methods aimed to measure two criteria to assess craniofacial correspondence in the corresponding two isolated regions. Thus, previous developments cannot be used for the identification task. Here, for the first time, we propose a complete framework that allows forensic experts to automatically address the final decision making stage. The presented fuzzy DSS develops information fusion concerning skull-face anatomical correspondence at different levels: criterion evaluation, SFO evaluation, and CFS evaluation. Additionally, in this study, we provide an implementation of the SFO evaluation level of the DSS (as explained above, we have already provided an implementation for the criterion evaluation level [10]). Within this level, we distinguish three sublevels with different conditions of aggregation. In each of them, the different sources of uncertainty are modeled, and different aggregation mechanisms account for information fusion and propagation. These sources of uncertainty also provide a mechanism to propagate information and uncertainty from criterion evaluation to SFO evaluation levels.

The uncertainty sources and degrees of confidence involved in the information fusion process are classified into bone, image, SFOs, morphological aspects, and computational methods used to model the criteria. The bone uncertainty refers to the quality of the skull, and the uncertainty of the photograph considers the visibility of each region and the resolution of the image. Morphological aspects can vary the degree of confidence of a criterion, depending on the sex, age, body mass index (BMI), or ancestry. Finally, the accuracy of the used methods is also taken into account, as well as the quality of the SFO achieved in the previous step.

We perform an experiment with positive and negative identification cases. In total, we analyze 33 positive SFOs against 411 negatives. We test 24 different combinations of aggregation functions within the proposed fuzzy DSS. We both analyze the results studying the mean accuracy of each approach and its capability of identification.

This manuscript is organized as follows. In Section 2 we introduce the relevant previous work and state the characteristics of the problem we aim to tackle. Section 3 describes our proposed DSS and Section 4 the corresponding implementation. Section 5 shows the experiments, and Section 6 details the discussion and conclusions.

2. Background

2.1. Craniofacial Superimposition

CFS approaches evolved as new technology was available although their foundations were laid more than 100 years ago [13, 14]. Three families of approaches have been developed along this time: photo superimposition (appeared in the mid 1930s), video superimposition (widely common since 1975), and computer-aided superimposition (developed in the second half of the 1980s) [1, 15].

Computerized systems in CFS are very transcendent [3, 5]. Some publications [6, 7, 8, 9, 16, 17] serve as examples of how computer techniques, specially CV and SC, can automate SFO and tackle the uncertainty/fuzziness of several cephalometric landmarks [18] and of the soft tissue distances [19]. These proposals are based either on photograph to photograph comparison [20] or on skull 3D model to photograph comparison [6, 7, 8, 15, 17]. Computerized CFS methods play an important role since they have managed to reduce time and subjectivity inherent to manual approaches followed by forensic experts. However, the resulting overlays' quality is influenced by several sources of uncertainty, as well as by partial and incomplete knowledge about skull-face anatomical correspondence. Accordingly, it is very difficult to achieve an optimal accuracy in an automatic way and a later manual refinement of SFO results is currently needed for such a purpose. Related to this, it is important to note that the forensic expert selects the materials (i.e. skull and photographs) in a previous phase of the process. This is a forensic technique frequently employed in multiple labs around the world, so the expert filters the materials and if they are not reliable, the SFO cannot be carried out. Best practices to follow by forensic experts in CFS were recently agreed with the framework of the EU project MEPROCS and they are described in [4].

Our system consists of the automation of the CFS process based on SC and CV techniques. In order to automate the SFO stage, we attempt to replicate the original scene in which the photograph was taken [21]. From the CV perspective, this involves a 3D-2D image registration problem. This process is guided

by incomplete and vague information (matching of two different objects, face and skull), and it involves an optimization task within a challenging search space with many local minima to establish the parameters of the registration transformation. For these reasons, advanced SC techniques have been designed to face this complex optimization problem [6, 7, 8, 9]. The resulting automatic overlays generated by our system are the inputs to the CFS DSS proposed in this paper.

2.2. Decision Making in Craniofacial Superimposition

Once one or several skull-face overlays have been achieved for the same identification case¹, the main goal is to determine the degree of support for the assertion that the skull and the face of the photograph(s) belong to the same person or not. This degree of support is based on the consistency of the matching between the face and the skull but it is also influenced by the quality and quantity of the materials used (photographs and skull). A scale for a craniofacial matching evaluation has been recently defined by some of the most representative experts in craniofacial identification within the MEPROCS project framework [4]. Accordingly, the final decision is provided in terms of strong, moderate or limited support.

This decision is guided by different criteria studying the anatomical relationship between the skull and the face. According to the literature [23], we can distinguish the following families of criteria for assessing the craniofacial correspondence:

1. Analysis of the consistency of the bony and facial outlines/morphological curves. Forensic experts confirm if two particular curves (of skull and face) are anatomically consistent. That is, if two curves follow the same shape or, in other words, if one curve mirrors the other. An example of this criteria can see it in Fig. 2.a where the forehead curve of the face follows the forehead curve of the skull.
2. Assessment of the anatomical consistency by positional relationship. These criteria consist of a positional relationship analysis in order to assess anatomical consistency. Thus, the goal is to check if the relative position of a skull region against a facial region is similar in respect to anatomical reference. Fig. 2.b shows a case of this family: the consistency between the lateral angle of the eye and the cranial orbit.
3. Location and comparison lines to analyze anatomical consistency. Experts analyze a set of marking lines (obtained by joining some reference landmarks) on the face and on the skull. In terms of CV, these lines have to be parallel in an image. For example, the ectocanthion lines marked in the skull and the face (Fig. 2.c).
4. Evaluation of the consistency of the soft tissue thickness between corresponding cranial and facial landmarks. The last set of criteria consists of analyzing the consistency of the facial soft tissue thickness considering distances between pairs of homologous landmarks (located on the skull and the face). Fig. 2.d shows an example of how the facial landmarks positions can be estimated from cranial landmarks using cones to model the soft tissue thickness. These distances can be checked using the skull-face overlay in existing studies relating to soft tissue thickness in different human populations [19].

MEPROCS work group also discussed and quantitatively analyzed these criteria for the evaluation of the morphological skull-face correspondence, providing a set of the most discriminative and easy to assess criteria [24].

Our long term goal is to automate the whole decision making process by modeling the most relevant criteria within the previous four families using CV and SC techniques. The resulting system would give as a result a global degree of support of a CFS identification to assist the forensic anthropologist to make her/his decision.

¹The number of SFOs to be used for the identification depends on the available valid photos. The reliability of the technique relies on having more than one photo, at different poses, etc. [22]

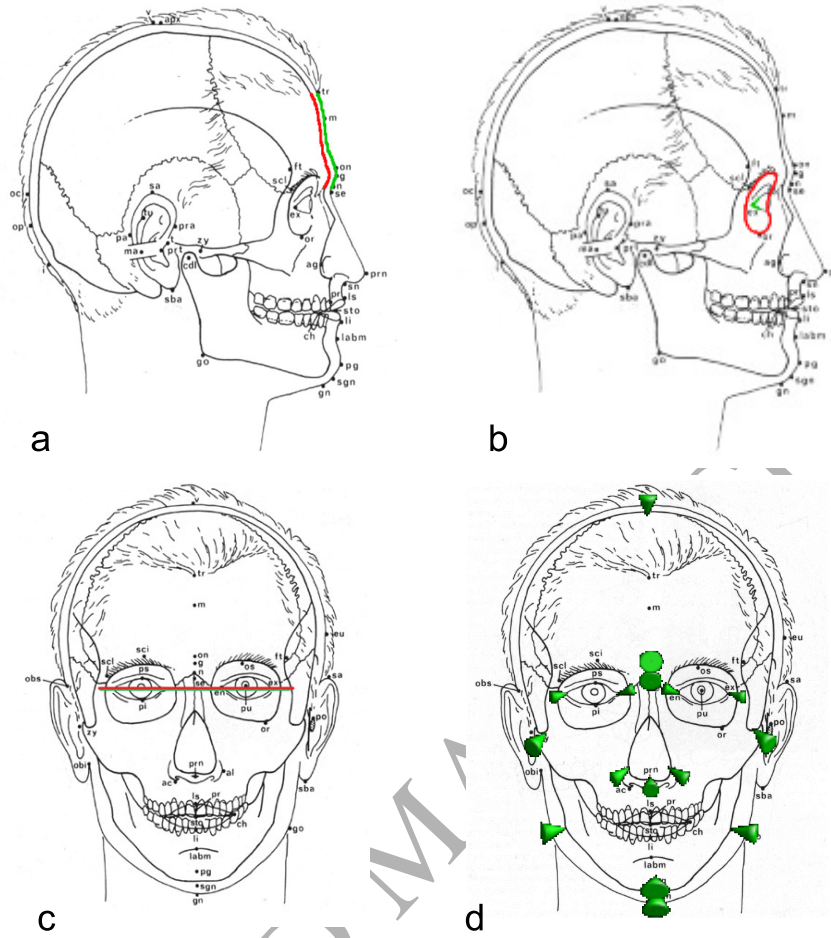


Figure 2: Examples of the four families of criteria for assessing the craniofacial correspondence. a) Consistency between the outline of the frontal bone and the forehead outline in lateral view; b) Consistency between the lateral angle of the eye and the orbit; c) Ectocanthion lines marked in the skull and the face; d) Facial landmark position from a cranial landmark using cones to model the soft tissue thickness.

There are just a few works tackling the automation of the analysis of craniofacial correspondences within the framework of CFS identification [25, 26, 27]. Most of the existing literature was published more than 18 years ago and the works are very basic and limited. In addition, they do not consider the use of either skull 3D models or computer techniques to perform the skull-face overlay. Besides, the employed technique for the shape analysis implies manual interaction. They provide a value that does not take into account the actual spatial relation between skull and face since the methods employed are invariant to translation, scale and rotation. Finally, these systems only implement a single group of the criteria to assess the craniofacial correspondence. For further information see [10].

Recently, we presented a simple and preliminary version of the DSS in [10, 11, 12]. In these works, we considered two of the most discriminative criteria to assess craniofacial correspondence: the morphological and spatial relationship between the facial and bony chin and the relative position of the eyeballs and the orbits. 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 (see Fig. 3).

To model the former criterion, we implemented several CV methods aimed to measure how the chin facial shape follows the skull shape given the delineation of these regions in a SFO. Our method includes

the automatic segmentation of the contours based on the region “between” them. Thanks to this process, this analysis takes into account the relative position and the distance between the two objects as well as the shapes. In [10], we proved that the best performance to model the relation “a curve follows another curve” is obtained using shape similarity measures with the area function comparison method and the complex coordinates’ signature. In order to measure this similarity the Euclidean distance of their corresponding shape signatures is computed. The following formula gives a similarity value between zero and one:

$$S_4 = 1 - \sqrt{\frac{1}{N} \sum_{i=1}^N (s_F(i) - s_B(i))^2} \quad (1)$$

with s_F and s_B being the shape signature of each contour.

Similarly, we developed a method to measure the relative position between the orbit and the center of the eyeball. For this aim, we implemented two different ways to compute the positional relationship between two objects in an image: the aggregation method and the centroid method. Once the relative position is obtained, we need to compare this position with a reference model in order to assess if there is anatomical consistency. In order to compare two different relative positions between two objects, we compute the similarity using the following expression based on [28]:

$$S_6 = 1 - \frac{|\delta_a - \delta'_a| + |\delta_b - \delta'_b| + |\delta_r - \delta'_r| + |\delta_l - \delta'_l|}{\delta_a + \delta'_a + \delta_b + \delta'_b + \delta_r + \delta'_r + \delta_l + \delta'_l} \quad (2)$$

where δ_a , δ_b , δ_r and δ_l are the degrees of the position relation “above”, “below”, “right”, and “left”, respectively, and δ and δ' are the relative position of two different pairs of objects.

To evaluate this criterion, we need to extract the appropriate contours both for the orbits (bony region) and eyeball center (facial region). The desired orbital contour is the interior contour, selected as the smaller one, while the facial region is simply the contour of the marked zone on the photograph.

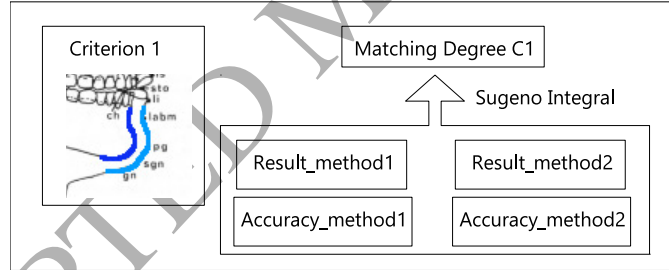


Figure 3: Scheme of the matching degree calculation for a criterion presented in [10].

All of these performed methods return a value between zero and one, which indicates how well the relation is achieved. Then, we aggregate the degrees of support of the individual methods to strengthen the final result. To do this, it is needed a measurement of importance of each one (the accuracy). We review the calculation of this value as follows since it is employed along the current proposal described in Sections 3 and 4. Regardless the criterion type, this accuracy of each method is calculated as its capability to discriminate in the decision making process (ranking positive and identification negative cases). Hence, a database composed of real positive identification cases (including the 3D skull model and one or more subject photographs) and negative cases is required. These negatives cases are obtained combining real 3D skull models and non corresponding photos of similar subjects (same gender, age, ethnic group, ...). First, the corresponding value to apply each method for a specific criterion over the database of cases is achieved. Next, the matching values reported are used to rank the 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 this ranking: if the method reported the highest value to a positive case (first position of the ranking), 1 is assigned. On

the contrary, if the position of a positive case is the last of the ranking, 0 is assigned. The formula to assign the accuracy of the method x_i in the instance j is:

$$Acc(x_i)_j = 1 - \frac{r - 1}{M_j - 1} \quad (3)$$

where r is the position of the positive case in the ranking and M_j is the worst (highest) value of the ranking for the 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).

Then, the average of all these accuracy values over all cases is calculated. As a result, an accuracy index in the interval $[0, 1]$ is achieved for each method. The final step is to aggregate the degrees of support of the best individual methods taking into account their accuracy. To do that, Sugeno integral [29] is used. For the two cases of study, the accuracy index of the aggregation overcame the individual results. The scheme of this procedure is shown in Fig. 3. In [12] we performed a deeper study of the behavior of different aggregation functions for the same aim. The obtained results show that Sugeno integral ranks better than the Choquet Integral and the Weighted Arithmetic Mean although no significant conclusions can be delivered regarding the performance.

3. Hierarchical Decision Support Framework for Craniofacial Superimposition

The whole CFS process is affected by several uncertainty sources and degrees of confidence that must be considered for decision making. In particular, we have distinguished the following sources of uncertainty from the forensic experts' experience.

Bone quality: the quality of the skeletal remains is an important issue during the CFS process. The condition of the bones depends on environmental factors, its preservation state has a direct influence on the confidence on the evaluation of face and skull anatomical correspondence.

Image quality: photographic quality is an additional criterion that has to be taken into consideration. The uncertainty inherent to the location of landmarks and regions in an image, already described in [4], can be greatly affected by the quality of the image. In particular, the location and evaluation of each single region/landmark is affected by the following sources of imprecision: 1) the variation in the distribution of shadows which depend on the light; 2) unsuitable focus, especially when the plane of focus is not enough depth and hence the critical objects are not sharp; 3) the image resolution. For optimal examination, experts recommend using photographs in which the facial image resolution is at least 180 pixels corresponding to the width of the head, or 90 pixels between the eyeballs (for full frontal images); 4) the pose of the face in the image, i.e. angle of view (frontal, lateral or oblique) and facial expression; 5) complete or partial occlusion of a region due to the presence of elements such as glasses, clothes or hair [4].

Skull-face overlay accuracy: the confidence degree of the SFO accomplished in the previous stage is another important factor to be considered. This process focuses on achieving the best possible superimposition of the skull and a facial image and it can be influenced by different sources of uncertainty, as described in [7].

Morphological aspects: the degree of confidence of each particular criterion analyzing anatomical correspondence can be affected by several factors. Firstly, the expected craniofacial relationship of a specific region is affected (in a lower or higher degree) by the age, sex, ancestry (biological profile), and/or BMI [30]. Thus, these factors have to be considered for each particular criterion during the decision making process. For example, the chin shape of the skull follows the facial shape in young people, but after 45 years of age this relation is increasingly unreliable. This relation can also be distorted due to being overweight. Secondly, each isolated region can have a different discriminative identification power. This is considered as the rate of being able to make a positive identification taking into account only that region.

Automatic method modeling the spatial/morphological relationship accuracy: different computer methods can be considered to automatically evaluate the degree of matching for each specific criterion. For example, the chin correspondence can be evaluated by considering different shape extraction and matching methods (see Sec. 2.2). The accuracy of a method is defined as how well the specific craniofacial relationship is modeled by that method for an actual criterion.

The DSS proposed in this work considers the evaluation of the skull-face anatomical correspondence at different levels. In each of them, the various sources of uncertainty introduced above are modeled, and different aggregation mechanisms account for information fusion and propagation. In particular, we have defined the following three levels or decision hierarchies (see Fig. 4):

- Level 1: CFS evaluation
- Level 2: SFO evaluation
- Level 3: Criterion evaluation (introduced in [10, 11, 12], see Sec. 2.2)

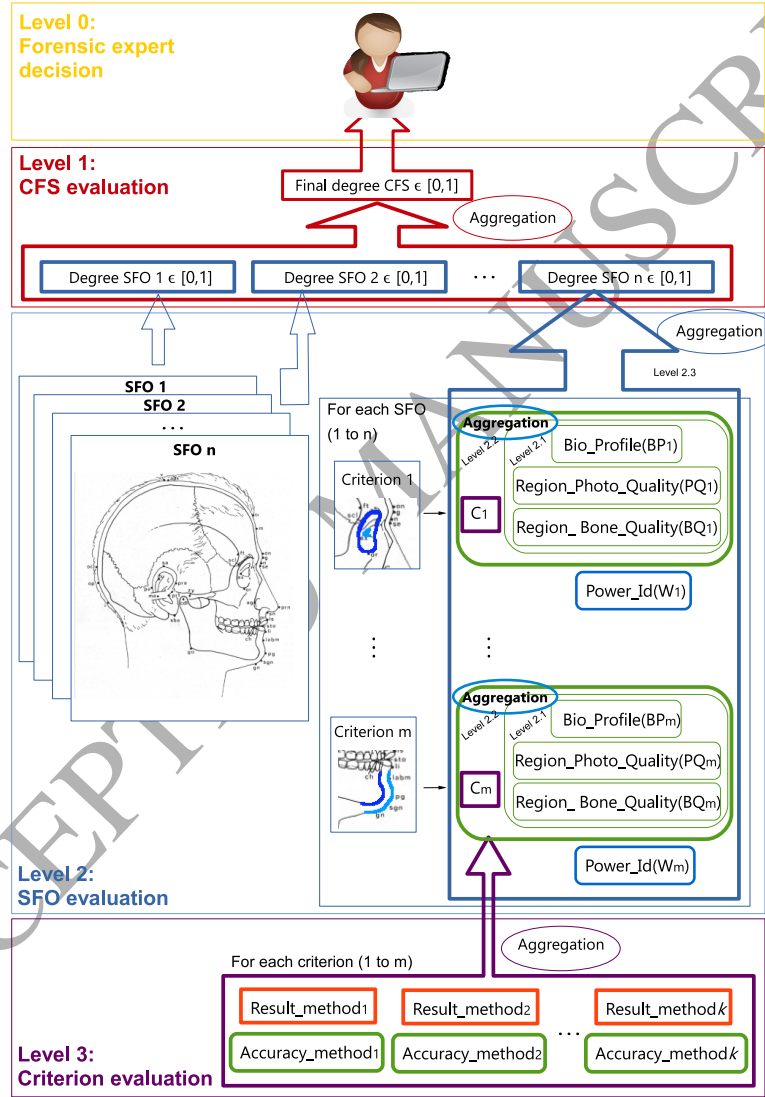


Figure 4: Hierarchical scheme of the DSS for CFS.

In this DSS scheme, the final degree of the CFS identification is obtained by aggregating all the SFO degrees. This corresponds with the highest level (**Level 1**) in the hierarchical system. In the next level (**Level 2**), for each SFO achieved in the previous stage, we aim to analyze the degree of fulfillment of

several criteria studying skull-face anatomical correspondence. For each criterion, facial and cranial regions are located (in the facial photograph and the 3D model, respectively) and a specific method to evaluate the skull-face correspondence is applied (the kind(s) of method(s) to be considered depends on the nature of the specific criterion). The degree of craniofacial correspondence of a SFO is computed by aggregating the matching degree of each single criterion taking into consideration the confidence of that criterion. Thus, the skull-face consistency in a region is expressed by a value between 0 and 1, obtained in the previous level (C_m). This value is complemented by the region/criterion confidence based on the quality of the photo (PQ_m), the quality of the bone (BQ_m), the biological profile variability of the criterion (BP_m), and the discriminative power of the isolated region. At this level, we set three different aggregation sublevels. The first one consists of aggregating the first three sources of uncertainty to get a single uncertainty value associated to the criterion sample quality and biological variability. The second sublevel integrates this aggregation with the matching degree (C_m). Finally, at the third one, we obtain the degree of the SFO craniofacial correspondence by aggregating the different previous values for all the regions taking into account the discriminative power of the isolated region as weight. We denote these sublevels as level 2.1, level 2.2, and level 2.3, respectively. If there are more than one CV-based method to evaluate a criterion, there is a need to aggregate the results of all of them (**Level 3**). To do so, we take into account the accuracy of each method. Fig. 4 graphically summarizes the proposed hierarchical DSS for CFS.

In addition, but not included within our DSS, there is a Level 0 of evaluation, the one carried out by the human expert. At this higher level the input is the final degree of CFS matching provided by the DSS. Then, the forensic experts will make the final identification decision. In this sense, the MEPROCS international consortium agreed a set of possible decision degrees according to the quality and quantity of materials [4]. The matching degree provided by our DSS could be directly incorporated within this scale, to any other, or considered in its own.

4. Framework Implementation

Together with the hierarchical DSS framework proposal, in this contribution we focus on the design of the SFO evaluation level. In this level 2 we aggregate the degree of the craniofacial matching in a region with the associated criterion uncertainty (see Fig. 5).

We have the problem of how to choose an aggregation function for each sublevel within the vast variety of aggregation functions available in the literature. In [31] the authors give some advices to select the most appropriate aggregation function for a specific application. First, the function must be consistent with the semantic of the aggregation process, i.e, if one is a disjunction, conjunctive or averaging aggregation functions are not suitable. Other important aspects to take into account are if the aggregation function should be symmetric, idempotent, or have a neutral or absorbing element. If the input number is always the same and what is the interpretation of the input values are also important to make a good choice of the suitable family or class. The second criterion is to select the appropriate member of that family or class, in order to produce adequate outputs for the given inputs. In our case, to address this second criterion we decided to perform an experimental study to analyze which aggregation function provides the best results in each case. To do that, we use a data set with positive and negative identification cases that helps us to make the good decision about the most suitable aggregation functions. In this sense, the SFO craniofacial correspondence of the positive cases have to be ranked before the negatives.

Based on the latter guidelines and the definitions of the Appendix, we justify the choice of the analyzed aggregation function for each sublevel as follows:

4.1. Aggregation function to combine material quality assessments and biological profile (level 2.1)

First, we have to aggregate the quality of the photo at the region (PQ_m), the quality of the bone at the region (BQ_m), and the biological profile variability of the criterion (BP_m) (see Fig. 6). These three aspects have a direct influence on the confidence of the matching degree of each particular region. Thus, we have decided to aggregate them in a single uncertainty value using an aggregation function denoted by $O_{Level2.1}(PQ_m, BQ_m, BP_m)$.

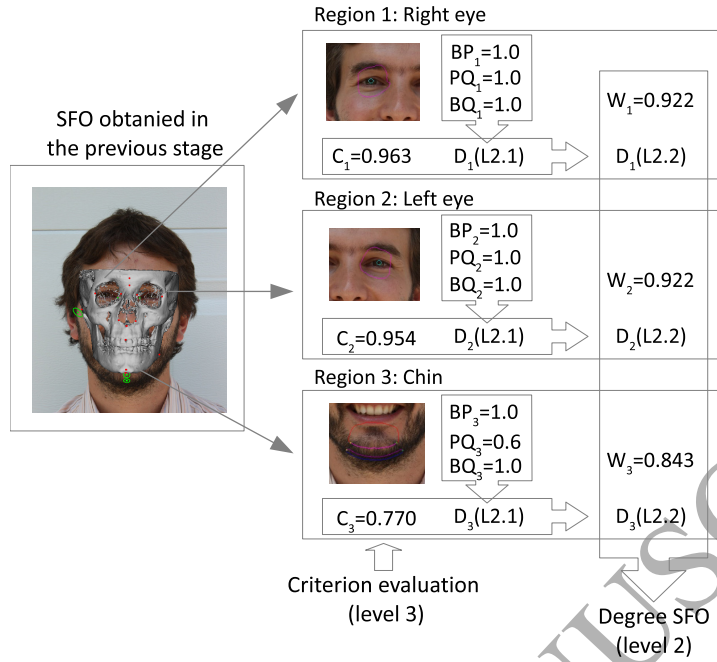


Figure 5: Graphical example of the level 2 of the fuzzy DSS framework (SFO evaluation).

Biological profile: the way biological profile affects the degree of confidence of each particular criterion can be easily modeled using fuzzy sets. According to the expert knowledge (task developed by Prof. Caroline Wilkinson, one of the most recognized experts in craniofacial identification), we have defined one or more fuzzy sets to model in which way the biological profile parameters (age, sex and ancestry) and the BMI modify the degree of confidence of each particular criterion.

Bone quality: the variations seen on bones can be described according to the bone's surface texture using the weathering stages [32]. They consider the taphonomic processes that may have affected the bones of a subject. For our application, the accuracy of the 3D model will have to be taken into account at the same time. Values of quality are set, based on the weathering stages, and are specifically associated with each region of the skull. If a specific region is deteriorated, the method to analyze the skull-face correspondence for each criterion can use it but the confidence in that criterion is reduced, so it will consequently be associated with a lower support value. Similarly to [33], where weathering stages were employed to modify the confidence of age estimation methods, we established quality indexes as ordinal numbers ranging in [0, 1]. They are assigned by a forensic expert according to the analysis of the state of the available skull. The assignation in this manner indicates the least weathering as being a perfect skull region (1) and the most weathering as being a faulty or not present region (0). We use a six-stage system in which stage 0 is determined to be a quality of 1.0, stage 1 of 0.8, stage 2 of 0.6, stage 3 of 0.4, stage 4 of 0.2, stage 5 of 0.1, and stage 6 is assigned a value of 0.0.

Photo quality: we also use a six-stage system to establish the quality of each facial region. A facial region belonging to the highest stage means that is clearly identified on the photograph and that is the ideal situation to apply the corresponding method to analyze the skull-face correspondence. On the contrary, a region of the lowest stage implies the impossibility to view that region in the photograph.

To aggregate all these values, a conjunctive behavior seems to be the best choice since it does not allow for compensation. Thus, low scores for some criteria (in this quality or biological aspects) cannot be compensated by other scores. If the quality of the bone is very bad, no matter how well the other two sources are, it applies that the matching between the skull and the face will be less reliable. However, averaging mean could be a more conservative choice. On the other hand, we consider that these three aspects affect in the same way

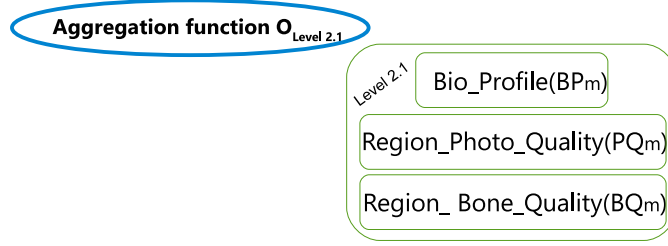


Figure 6: Aggregation scheme at level 2.1 of the DSS proposed framework.

to the region criterion so the aggregation function has to be symmetric. For these reasons, we decide to analyze the behavior of our system using the aggregation functions Minimum, Product, and Arithmetic Mean. Thus, at this level we can state the aggregation function as $O_{Level2.1}(PQ_m, BQ_m, BP_m) \{min, prod, mean\}$.

4.2. Aggregation function to combine the matching degree and the uncertainty value of level 2.1 (level 2.2)

Secondly, we have to aggregate the previous uncertainty sources with the matching degree of the skull and the face at the corresponding region as we can see in Fig. 7. For this application, an averaging procedure is required. The basic rule of this class of aggregation functions is that the total score cannot be above or below any of the inputs. The aggregated value is seen as some sort of representative value of all the inputs. In addition, we consider that the aggregated inputs do not have the same contribution to the total output, so a not symmetric weighted function is needed.

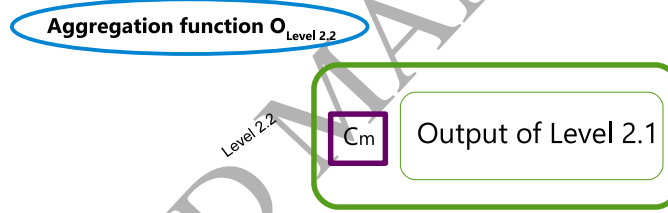


Figure 7: Aggregation scheme at level 2.2 of the DSS proposed framework.

Two of the most common weighted aggregation functions are the Weighted Arithmetic Mean and the Weighted Geometric Mean. Both functions are not symmetric. We establish different weights to each input based on the expert knowledge. This weight, a number $w_i \in [0, 1]$, represents the importance of each one. These functions are abbreviated as *wam* and *wgm*, respectively.

Since the values of the weighting vector for the *wam* and for the *wgm* must sum up to 1, we apply a simple normalization of the accuracy index with respect to their sum:

$$w_i = \frac{W_i}{\sum_{i=1}^n W_i} \quad (4)$$

where W_i is the identification power of the *i*-region.

This aggregation function can be denoted as $O_{Level2.2}(C_m, Output_{level2.1}) \{wam, wgm\}$

4.3. Aggregation function to combine the identification power and the degree of level 2.2 (level 2.3)

The final step in level 2 is to obtain the SFO evaluation degree. As explained before, there is a need to aggregate multiple degrees of support with an associated weight. Each of these degrees corresponds to the skull-face matching degree in a specific region with the corresponding uncertainty integrated (Fig. 8). The weight in this case will be the identification power of each isolated part of the face.

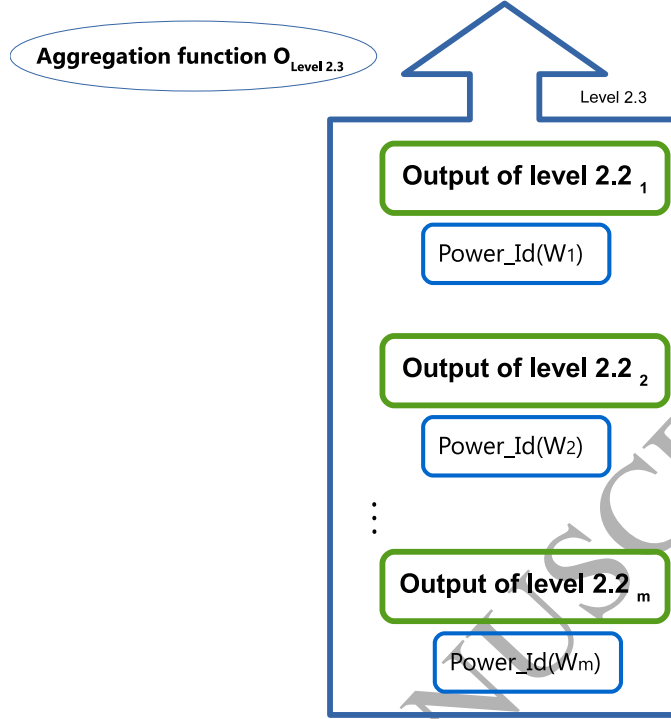


Figure 8: Aggregation scheme at level 2.3 of the DSS proposed framework.

The requirement for this aggregation function is that it has to be non symmetric. The identification power of a region (the weight) reflects the relative contribution of each input to the final output: the degree of a SFO evaluation.

As in the previous sublevel, we consider the most common weighted aggregation functions: the *wam* and the *wgm*. In this case, the weight which represents the importance of each input is the power identification of each region. These weights are realistic and they are computed using a dataset of positive and negative identification cases in a similar way of the calculation of the accuracy of each method in level 3 (See Sec. 2.2 and Eq. 3). Again, the values of the weighting vector for these functions have to sum up to 1, so we apply a simple normalization of the power identification with respect to their sum (Eq. 4).

Apart from the weighted aggregations, we also use fuzzy integrals as aggregation functions. These combine the data supplied by several information sources according to a fuzzy measure, that represents the background knowledge on the information sources. In this study, we use the Choquet and the Sugeno integral and we use a Sugeno λ -measure to determine the fuzzy measure. A fuzzy measure, g , is a real valued function defined on the power set of X (the universe of discourse), 2^X , with range $[0, 1]$, satisfying the following properties: Let A and B be two subsets from X .

1. $g(\phi) = 0, g(X) = 1$ (boundary condition)
2. $A \subseteq B$ implies $g(A) \leq g(B)$ for all $A, B \in F$ (monotonicity)

A fuzzy measure specifies the opinion of the ‘worth’ or ‘goodness’ of each subset of information sources in evaluating a particular hypothesis. Each information source gives a belief or confidence in the hypothesis and the measure lets you know how to weight that belief or confidence, in this case the identification power or each facial region. To determine a fuzzy measure on X , we must identify $2^p - 2$ coefficients satisfying $p2^{p-1}$ conditions. To solve this drawback, some approaches have been proposed to reduce the number of parameters to be determined [29]. In this paper, we use a Sugeno λ -measure defined as in [33]:

Definition 1. Let g be a fuzzy measure, then g_λ is a Sugeno λ -measure if there exists $\lambda > -1$ such that

$$g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A) \mu(B) \quad (5)$$

holds for all $A, B \in F$.

It is to be noted that, for a Sugeno λ -measure,

$$\prod_{j=1}^p (1 + \lambda g_\lambda(\{x_j\})) = 1 + \lambda \quad (6)$$

holds because of the boundary condition [34]. Using the above definitions $g_\lambda(X)$ can be constructed from the fuzzy densities of the elements of X . Given the set of densities, the value λ can be easily found as the unique root greater than -1 of a simple polynomial [35].

Once λ is found, the fuzzy integral can be calculated.

- *The discrete Choquet Integral*

The discrete Choquet integral with respect to a λ -fuzzy measure is given by

$$C_g(\mathbf{x}) = \sum_{i=1}^n [x_{(i)} - x_{(i-1)}] g(H_{(i)}) \quad (7)$$

where $\mathbf{x}_{\searrow} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$ is a non-decreasing permutation of the input \mathbf{x} , $x_{(0)} = 0$ by convention, and $H_i = (i), \dots, (n)$ is the subset of indices of $n - i + 1$ largest components of \mathbf{x} .

- *The Sugeno Integral*

The Sugeno integral with respect to a λ -fuzzy measure is given by

$$S_g(\mathbf{x}) = \max_{i=1, \dots, n} \min(x_{(i)}, g(H_{(i)})), \quad (8)$$

where $\mathbf{x}_{\searrow} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$ is a non-decreasing permutation of the input \mathbf{x} , and $H_i = (i), \dots, (n)$.

In the following section we refer to these aggregation functions as *choq* and *sug*, respectively. Hence, at this level we can state the aggregation function as $O_{Level2.3}(Output_{Level2.2i}, Power_ID(W_i)) \{wam, wgm, choq, sug\}$.

5. Experiments

The main contribution of this work is the proposal of the fuzzy hierarchical DSS for CFS framework. Hence, the experiments' aim is directly related to its validation. To do so, we have designed and developed two different experiments.

The objective of the first experimental set-up is to study the performance of the different aggregation functions within our framework. In order to analyze which are the most appropriate functions for our system, we obtain the accuracy degree for identifying positive cases in each case at the SFO evaluation level. We also perform a statistical test in order to analyze whether significant differences exist among the results of the different aggregation functions.

A second experiment has been performed with the specific focus on validating the proposed DSS framework for CFS. This considers both positive and negative identification cases, and make use of Cumulative Match Characteristic (CMC) curves to study the identification capabilities of the current implementation of the proposed DSS framework.

5.1. Experimental Design

The experimental design involves the sex-based cross-comparison of nine Cone Beam Computed Tomography (CBCT) models of living individuals and seven 3D skull models (acquired using a 3D structure light scanner, the Artec MHT [36]) of deceased people against a variable number of candidates and photographs. In each case there is one positive candidate with one or more photographs available. Forensic experts have made a previous filter based on sex and age, so there is not the same number of negative cases for each skull. In total, there are 16 3D skull models and 66 photographs of candidates, resulting in 33 positive and 411 negative SFOs. Table 1 summarizes the composition of the dataset employed. The SFOs have been obtained by our automatic method in [9] using the parameter values reported in that contribution. For the CBCTs positive cases, we use a ground truth dataset, whose overlays are assumed as optimal according to [37].

Table 1: Experimentation dataset summary

Skull Model	Positive SFOs	Negative SFOs
CBCT 1	2	20
CBCT 2	2	20
CBCT 3	2	20
CBCT 4	2	20
CBCT 5	2	33
CBCT 6	2	33
CBCT 7	2	33
CBCT 8	2	33
CBCT 9	2	33
3D Model 10	3	19
3D Model 11	1	21
3D Model 12	1	27
3D Model 13	4	24
3D Model 14	1	26
3D Model 15	2	25
3D Model 16	3	24
Total	33	411

For each available photograph, experts set the age and the BMI. They also marked visible regions in each photograph and the related quality of each one is established in a scale between 0 and 1 (see Sec. 4.1). In each image, experts delineated from one to four regions (depending on the visibility of them): chin contour, cranial contour, eyeball center right, and eyeball center left. These four regions were also marked on the skull 3D models. Again, experts set the quality of the bone region based on the weathering stages (see Sec. 4.1). A graphical example of this process is shown in Fig. 9.

The discriminative power of each region is reported in [10], as well as the corresponding methodology to obtain it. These values are computed using Eq. 3 of Sec. 2.2, and it is understood as the capability to discriminate in the decision making process after aggregating the best methods for modeling each region at level 3. The cranial contour follows the same criterion as the chin contour. Thus, given the inability to calculate identification power from CBCT ground truth cases (CBCT does not include the upper part or the skull), the value of the identification power is taken the same as the chin case. Table 2 summarizes these values for each region used in this work.

For these implemented criteria the influence related with biological profile was defined by Prof. Wilkinson according to her expert knowledge and represented using fuzzy sets in Fig. 10:

As can be seen in Fig. 10, the chin criterion is less reliable after 60 years old, decreasing to 0.25 from 75. The same criterion is unreliable with BMI values above 35 (changes in fat will alter the shape of the chin) [30]. Neither the eyeball position nor the cranial contour are affected by any morphological aspect.



Marking skull regions	Marking photograph regions
	
Setting skull regions qualities	Setting photo qualities and profile
Cranial bone quality: 1.0 Orbit right bone quality: 0.8 Orbit left bone quality: 0.8 Chin bone quality: 1.0	Age: 47 BMI: 22 Cranial contour quality: 0.8 Eyeball center right quality: 1.0 Eyeball center left quality: 0.8 Chin contour quality: 0.6

Figure 9: Example of marking regions and setting qualities in both skull and photograph.

Table 2: Identification power of each isolated region [10].

Region	Power of identification
Chin contour	0.843
Cranial contour	0.843
Eyeball (left and right) center position	0.922

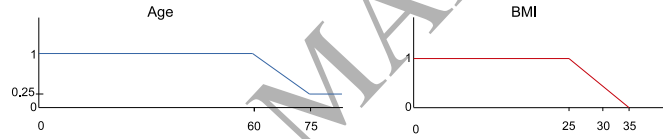


Figure 10: Defined fuzzy sets to model morphological aspects for the chin outline. At the left the age related confidence and at the right the BMI related confidence.

For the level 2.2, the experts in Prof. Wilkinson lab established that the uncertainty sources have a third of influence and the matching degree have two thirds of influence. So the weighted vector used in this case is $\mathbf{w} = (\frac{1}{3}, \frac{2}{3})$.

Once the SFOs are achieved, all the previous data are taken into account to obtain the final SFO degree as explained in Section 4. We tested the different aggregation functions mentioned in this section for each sublevel, that is: level 2.1: *min*, *prod* and *mean*; level 2.2: *wam* and *wgm*; level 2.3: *wam*, *wgm*, *choq* and *sug*. Accordingly, 24 different combinations are analyzed.

The final values reported for each skull are used to rank the candidates based on their chance of being the actual subject.

5.2. Study of aggregation functions

This first study consists in analyzing the behavior of the different combinations of the aggregation functions. To do that, we calculate the accuracy of the 24 different fuzzy DSSs (each of them represented by a different combination of aggregation functions) over all cases as explained in Section II.B and Equation 3.

Table 3 shows the mean accuracy for the system using each combination of the selected aggregation functions. As can be seen, the combination *mean-wam-wam* achieves the highest value, with 0.8550 of mean accuracy. *Mean-wgm-wam*, *prod-wam-wam*, *min-wam-wam*, *prod-wgm-wam* and *min-wgm-wam* present similar results, with 0.8516, 0.8461, 0.8394, 0.8378 and 0.8137, respectively. The next methods show a bigger gap with respect to the previous ones, *mean-wam-choq* and *mean-wgm-choq* with 0.7742 and 0.7733 of

mean accuracy. The worst performances are achieved by combinations *min-wam-wgm* and *min-wgm-wgm*, obtaining a mean accuracy of 0.6520 and 0.6494, respectively.

Table 3: Mean accuracy of each combination method and ranking obtained through Friedman's test.

Combination method	Mean accuracy	Ranking
<i>mean-wam-wam</i>	0.8550	7.561
<i>mean-wgm-wam</i>	0.8516	8.030
<i>prod-wam-wam</i>	0.8461	7.970
<i>min-wam-wam</i>	0.8394	8.485
<i>prod-wgm-wam</i>	0.8378	8.955
<i>min-wgm-wam</i>	0.8137	9.758
<i>mean-wam-choq</i>	0.7742	11.030
<i>mean-wgm-choq</i>	0.7733	11.167
<i>mean-wam-sug</i>	0.7460	11.848
<i>prod-wam-choq</i>	0.7330	11.788
<i>min-wam-choq</i>	0.7248	12.379
<i>prod-wam-sug</i>	0.7014	13.379
<i>mean-wgm-sug</i>	0.6914	14.348
<i>prod-wgm-choq</i>	0.6895	14.303
<i>min-wgm-sug</i>	0.6893	14.091
<i>min-wam-sug</i>	0.6882	14.076
<i>min-wgm-choq</i>	0.6797	14.939
<i>mean-wgm-wgm</i>	0.6795	14.106
<i>mean-wam-wgm</i>	0.6757	14.212
<i>prod-wam-wgm</i>	0.6633	14.955
<i>prod-wgm-wgm</i>	0.6593	15.242
<i>prod-wgm-sug</i>	0.6525	15.909
<i>min-wam-wgm</i>	0.6520	15.561
<i>min-wgm-wgm</i>	0.6494	15.909

Friedman test [38], a nonparametric test for analysis of variance, aims to test a null hypothesis stating that the mean total accuracy of all the methods are the same. We have set the experiment level of significance in $\alpha = 0.05$.

Table 3 summarizes the ranking obtained by Friedman's test for the studied methods. The result of applying Friedman's test is $\chi^2_F = 116.21$ and a p -value of 1.99×10^{-14} . Given that the p -value of Friedman are lower than the level of significance considered, $\alpha = 0.05$, there are significant differences among the observed results. Attending to these results, a post-hoc statistical analysis could help us to detect concrete differences among methods.

In particular, Bonferroni-Dunn test [39] is accomplished to detect significant differences among a control approach and the rest. The control method in this case is the combination of *mean-wam-wam*. Fig. 11 displays a graphical representation, including the rankings obtained for each method and the critical difference for each value of α . A Bonferroni-Dunn's graphic illustrates the difference among rankings obtained for each approach. The horizontal line represents the level of significance considered in the study at height equal to the sum of the ranking of the control method and the corresponding critical difference computed by the Bonferroni-Dunn method. Those bars which exceed this line are the associated to an approach with worse performance than the control method.

The Bonferroni-Dunn's test shows us the following significant differences with *mean-wam-wam* as control method: *mean-wam-wam* is better than every method except *prod-wam-wam*, *mean-wgm-wam*, *min-wam-wam*, *prod-wgm-wam*, *min-wgm-wam*, *mean-wam-choq*, *mean-wgm-choq*, *prod-wam-choq*, *mean-wam-sug* and *min-wam-choq* with $\alpha = 0.05$ (13/23 methods). Although *mean-wam-wam* obtains the lowest error and ranking rates, the Bonferroni-Dunn's test is not able to distinguish it as better than all the following 10

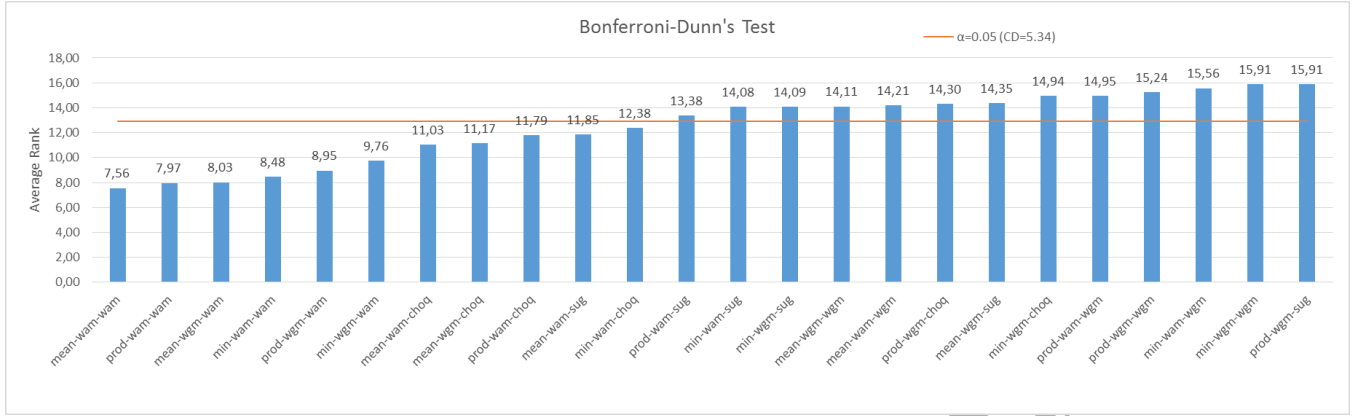


Figure 11: Bonferroni-Dunn's graphic corresponding to the results.

approaches.

We applied more powerful procedures, such as Holm's and Hochberg's, for comparing the control method with the others. However, the same conclusions are achieved. The results are shown by computing p -values for each comparison. Table 4 show the p -value obtained for Bonferroni-Dunn', Holm's and Hochberg's procedures.

Table 4: p -values on the results (*mean-wam-wam* is the control method)

<i>mean-wam-wam</i> vs.	Bonferroni-Dunn p	Holm p	Hochberg p
<i>prod-wam-wam</i>	1.0	1.0	0.8142
<i>mean-wgm-wam</i>	1.0	1.0	0.8142
<i>min-wam-wam</i>	1.0	1.0	0.8142
<i>prod-wgm-wam</i>	1.0	1.0	0.8142
<i>min-wgm-wam</i>	1.0	1.0	0.8142
<i>mean-wam-choq</i>	1.0	0.2774	0.2774
<i>mean-wgm-choq</i>	0.8811	0.2682	0.2682
<i>prod-wam-choq</i>	0.3488	0.1239	0.1213
<i>mean-wam-sug</i>	0.3167	0.1239	0.1213
<i>min-wam-choq</i>	0.1298	0.0564	0.0564
<i>prod-wam-sug</i>	0.0191	0.0091	0.0091
<i>min-wam-sug</i>	0.0042	0.0024	0.0022
<i>min-wgm-sug</i>	0.0040	0.0024	0.0022
<i>mean-wgm-wgm</i>	0.0039	0.0024	0.0022
<i>mean-wam-wgm</i>	0.0031	0.0020	0.0020
<i>prod-wgm-choq</i>	0.0025	0.0017	0.0017
<i>mean-wgm-sug</i>	0.0022	0.0016	0.0016
<i>min-wgm-choq</i>	0.0005	0.0004	0.0004
<i>prod-wam-wgm</i>	0.0005	0.0004	0.0004
<i>prod-wgm-wgm</i>	0.0002	0.0002	0.0002
<i>min-wam-wgm</i>	$9.9223 \cdot 10^{-5}$	$9.0594 \cdot 10^{-5}$	$9.0594 \cdot 10^{-5}$
<i>min-wgm-wgm</i>	$3.7259 \cdot 10^{-5}$	$3.7259 \cdot 10^{-5}$	$3.5639 \cdot 10^{-5}$
<i>prod-wgm-sug</i>	$3.7259 \cdot 10^{-5}$	$3.7259 \cdot 10^{-5}$	$3.5639 \cdot 10^{-5}$

5.3. DSS framework validation

Our second performed experiment studies the capability of our system to identify the correct individual. The results are reported using CMC curves [40] along with average rank 10 identification accuracy. A CMC curve captures the percentage (or probability) that the correct match for a case is present in a candidate

list of the r best matches, where r denotes the rank. Therefore, rank 10 identification accuracy denotes the probability that the correct match is one of the subjects in a list of the top 10 matches provided by the system.

There is no previous work in the literature with a similar proposal to that presented in this manuscript. For this reason, we cannot compare the obtained results with existing CFS computer-based proposals. We could focus instead on the performance demonstrated by human experts. The reliability of CFS in human identification has been assessed by different authors [41, 42, 43, 23, 44] achieving confronted conclusions about the technique's reliability, with positive matches ranging from 70% to 100%, and false negatives from 0% to 20%.

The reliability studies reported in the forensic literature are fraught with limitations, such as the absence of an objective measurement of the craniofacial superimposition match, limitations of the technical equipment, imprecision in landmark location while performing landmark-based methods, absence of soft tissue data for the tested population, deficient quality of the skull 3D models, postmortem photographs, limited samples, lack of appropriate statistical analysis, and the absence of inter and intra observer studies.

The most recent work [45], a multiple-lab study with 26 participants from 17 different institutions involving 60 CFS cases showed a global average performance of 79% of correct identifications. This percentage scaled up to 84% in a similar study [46] where the participants were asked to follow MEPROCS best practices [4]. We should notice that these two studies were laid out with the intention of identifying good and bad practices and to validate MEPROCS standards, respectively, rather than to test CFS reliability.

The last possibility to provide a reference for the identification performance can be obtained through the examination of other automatic or semi-automatic identification methods that have been proposed to model different forensic anthropology techniques. In [47], a computerized clavicle identification system was presented. The method quantifies the clavicle outline shape from the skeletons and postero-anterior AM chest radiographs to rank individuals. The results show that a positive predictive value of 78% is achieved when considering the 21 first classified bones (rank-21), increasing to 90% around rank-100.

In our case we compare the six best aggregation function combinations of the previous study, i.e. *mean-wam-wam*, *prod-wam-wam*, *mean-wgm-wam*, *min-wam-wam*, *prod-wgm-wam* and *min-wgm-wam*. Although they all obtained similar results without significant differences with respect to the mean accuracy (see in Sec. 5.2), CMC curves allow us to differentiate among them (see Fig. 12).

CMC curves show us that *mean-wam-wam* and *mean-wgm-wam* have the best performance to identify the actual subject. Although they do not have the highest value for identifying the correct match in rank 1 and 8, they have the best performance in the remaining scenarios (Fig. 12). As it can be seen, when our system uses these aggregation functions it is able to rank the correct individual within the five first positions with more than a 70% of accuracy. This rate increases if we consider the 10 first positions, reaching more than 90% of identification accuracy and for the most of the combinations except *min-wgm-wam* that presents a worse behavior. Although these results are not good enough to consider the current proposal for identification purposes on its own, our fuzzy DSS has demonstrated promising capabilities to filter (shortlist) cases. In fact, it has showed a significant improvement in comparison with the only 'comparable' method [47] (dealing with a simpler identification technique since it involves a bone to bone comparison, a 3D clavicle and the same bone in a radiograph) which needed to look at around 100 cases to reach a 90% identification rate.

Finally, in Fig. 13 the eight first visual results for one case of the ranking using the *mean-wam-wam* combination are depicted. In the figure, positive cases have the highest SFO degrees, so they are ranked at the firsts positions.

6. Discussion and Conclusions

In the present contribution, we propose a complete framework for a DSS in CFS. The proposal develops information fusion concerning skull-face anatomical correspondence at three different levels: criterion evaluation, SFO evaluation, and CFS evaluation. We classify the uncertainty sources and degrees of confidence involved in this process as related to bone, image, skull-face overlays, morphological aspects and used methods. In this study, we focus on the SFO evaluation level. Within this stage, we distinguish three

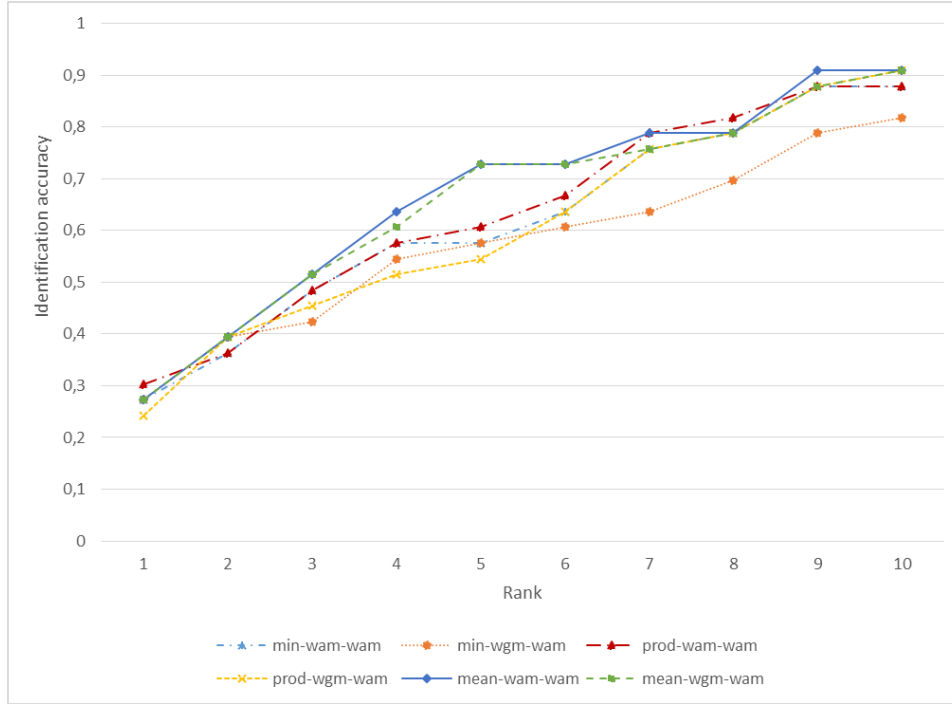


Figure 12: CMC curves corresponding to the six best combinations of the proposed fuzzy DSS.

sublevels with different conditions of information fusion. For this reason we perform an experimental study to analyze which aggregation function provides the best results. The first sublevel aggregates the sources of uncertainty of the bone and the image, and the biological profile variability. For this sublevel, we propose to use the minimum (*min*), the product (*prod*), and the arithmetic mean (*mean*) as aggregation functions. The second sublevel consists of integrating this aggregation with the matching degree of the skull and the face. In this case, we study the use of the weighted arithmetic mean (*wam*) and the weighted geometric mean (*wgm*). Finally, at the third sublevel, we obtain the degree of the SFO craniofacial correspondence by aggregating the different previous values for all the regions taking into account the discriminative power of the isolated region as weight. To study that, we test the *wam*, the *wgm*, the Sugeno (*sug*) integral, and the Choquet (*choq*) integral. We perform an experiment with positive and negative identification cases. We both analyze the results studying the mean accuracy of each approach and its capability of identification.

With respect to the mean accuracy, the combination *mean-wam-wam* shows the best results in our system. It also presents the first position at the Friedman ranking. Statistical tests show this combination of the aggregation functions is significant better than 13 of the 24 studied methods, but we can not confirm that there are significant differences between this method and the remaining 10 first approaches. However, according to these results, we can conclude that the best aggregation function for the sublevel 2.3 is the weighted arithmetic mean, since the six highest accuracies are obtained with this function (always more than 0.8).

Finally, we validate the DSS framework as an identification system. At this point the identification accuracy is insufficient for running independently as an automatic identification tool. However, it can be already used as a powerful shortlisting tool capable of successfully filtering out a large number of candidates (20 out of 30) in 90% of the identification cases we tested.

We identify the following sources of uncertainty/error that in our opinion are limiting the accuracy of the automatic DSS for CFS:

- The quality of the used SFOs: as explained in the introduction section, reaching an optimal SFO accuracy is still an open field of research and manual refinement of SFO results is currently needed


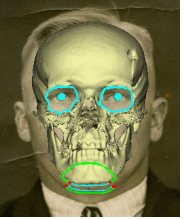
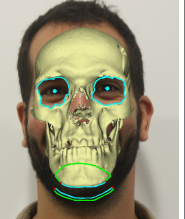
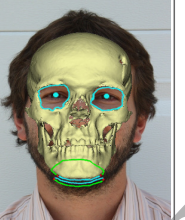




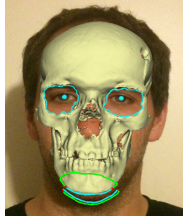
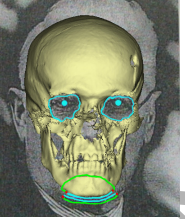

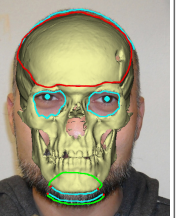




Ranking position	1 (positive case)	2 (positive case)	3 (negative case)	4 (negative case)
SFO				
Eyeball reference position				
Chin Criterion	0.7753	0.6906	0.8140	0.7841
Eye Right Criterion	0.8943	0.8581	0.8147	0.8112
Eye Left Criterion	0.8112	0.8657	0.8129	0.7626
Cranial Criterion	-	-	-	-
Degree SFO	0.8787	0.8721	0.8549	0.8434
Ranking position	5 (negative case)	6 (negative case)	7 (negative case)	8 (negative case)
SFO				
Eyeball reference position				
Chin Criterion	0.8049	0.8140	0.8140	0.8037
Eye Right Criterion	0.7077	0.8057	0.6961	0.7516
Eye Left Criterion	0.7185	0.6679	0.6521	0.8019
Cranial Criterion	-	-	-	0.0000
Degree SFO	0.8139	0.8115	0.8050	0.6363

Figure 13: Visual results of the eight firsts positions of the ranking for the combination *mean-wam-wam* for one cranial case.

over an important number of cases. One of the issues negatively affecting the accuracy of the automatic SFO method is the consideration of the mandible as a rigid part of the skull.

- The number of SFOs considered: our experimental set-up does not take into consideration the level 1, still to be developed, in which CFS instances are evaluated (using several AM images of the same person). It was demonstrated in [44] that CFS is more reliable when two or more facial photos of the same individual taken from different points of view are jointly used in the examination.
- The number of regions evaluated: we have only implemented four regional criteria, namely, the morphological and spatial relation between the face and the skull chin shape, the relative position of the orbits and the eyeballs (two regions, one for each eye), and the morphological correspondence between the cranial contour in the skull and the face. In addition, the last criterion could be only evaluated in the few cases where the candidate is bald, and the one related with the orbits-eyeball is partially evaluated in many photographs, those with a profile pose.
- 3D and 2D regions delimitation: this task is subjective and it is expected that two different experts, or even the same expert in different times, will mark a different slightly region. Among all the limitation factors we think this is the least important.

In any case, it is important to remark that this fuzzy hierarchical DSS framework is the first automatic CFS system and the obtained results are promising taking into account the actual scope of improvement.

In this sense, in future works, we aim to include more families of criteria for assessing the craniofacial correspondence and test the DSS with a larger dataset. We also plan to perform inter and intra expert studies to measure inter-intra variability marking those criteria (regions) on the facial photograph and the 3D skull model. Besides, in order to calculate the accuracy of each method and the power identification of each criterion we can take into account the variety of the instances used, separating into groups of ages, sex, ethnical, etc. Thus, different overall values for different kind of populations can be used to strengthen the final output. To do that, we need a bigger dataset of cases since the current is still insufficient. In addition, we could study the use of regression techniques to build aggregation functions from our specific data following [48]. Another aspect to take into consideration is to aggregate the different SFOs of the same person in order to obtain a final degree of a CFS case (i.e., the design of level 1). Besides, we keep on achieving more accurate SFOs through the study of new camera parametrizations, optimization strategies, and the modelization of the mandible rotation and translation movement and its inclusion with the SFO optimization process. Our final objective is to reach a higher identification accuracy. The need of objective assessment and automation has been justified along this manuscript, although and in depth justification was recently provided in [5]. The performance of newly developed CFS methods should be compared with a usual forensic dataset of known case studies. The methods proposed could be applied to solve these cases in order to validate them. Then, the obtained results would be compared with the identification previously determined by forensic experts. Despite there are promising works available in the specialized literature in this line, automatic methods are not implemented yet due to the inability to test their behavior in an objective manner. In this manuscript, authors emphasize the requirement of statistically significant reliability studies that tackle these challenges to obtain a more solid picture on the reliability of CFS. Accordingly, the identification values reported in the present work will be used as reference in the future.

Acknowledgments

This work has been supported by the Spanish *Ministerio de Economía y Competitividad* (MINECO) under the NEWSOCO project (ref. TIN2015-67661-P) and the Andalusian Dept. of *Innovación, Ciencia y Empresa* under project TIC2011-7745, including European Regional Development Funds (ERDF-FEDER). Mrs. C. Campomanes-Alvarez's work has been supported by Spanish MECD FPU grant AP-2012-4285. Dr. Ibáñez's work has been supported Spanish MINECO *Juan de la Cierva-Incorporación* Fellowship IJCI-2014-22433. We would like to thank Drs. Cavalli (Ospedali Riuniti di Trieste, Trieste, Italy), Cattaneo (LABANOF, University of Milan, Italy), and Jankauskas (Vilnius University) the access to the data.

References

- [1] M. Yoshino, Craniofacial superimposition, in: C. Wilkinson, C. Rynn (Eds.), *Craniofacial Identification*, University Press, Cambridge, 2012, pp. 238–253.
- [2] C. Wilkinson, C. Rynn, *Craniofacial identification*, Cambridge University Press, 2012.
- [3] S. Damas, O. Córdón, O. Ibáñez, J. Santamaría, I. Alemán, M. Botella, F. Navarro, Forensic identification by computer-aided craniofacial superimposition: a survey, *ACM Comput Surv* 43 (4) (2011) 27.
- [4] S. Damas, C. Wilkinson, T. Kahana, E. Veselovskaya, A. Abramov, R. Jankauskas, P. Jayaprakash, E. Ruiz, F. Navarro, M. Huete, E. Cunha, F. Cavalli, J. Clement, P. Leston, F. Molinero, T. Briers, F. Viegas, K. Imaizumi, D. Humpire, O. Ibáñez, Study on the performance of different craniofacial superimposition approaches (ii): best practices proposal, *Forensic Sci Int* 257 (2015) 504–508.
- [5] M. I. Huete, O. Ibáñez, C. Wilkinson, T. Kahana, Past, present, and future of craniofacial superimposition: Literature and international surveys, *Legal Medicine* 17 (2015) 267–278.
- [6] O. Ibáñez, O. Córdón, S. Damas, J. Santamaría, An experimental study on the applicability of evolutionary algorithms to craniofacial superimposition in forensic identification, *Inf Sci* 79 (2009) 3998–4028.
- [7] O. Ibáñez, O. Córdón, S. Damas, J. Santamaría, Modeling the skull-face overlay uncertainty using fuzzy sets, *IEEE Trans Fuzzy Syst* 16 (2011) 946–959.
- [8] O. Ibáñez, O. Córdón, S. Damas, A cooperative coevolutionary approach dealing with the skull-face overlay uncertainty in forensic identification by craniofacial superimposition, *Soft Comput* 18 (2012) 797–808.
- [9] B. R. Campomanes-Álvarez, O. Ibáñez, C. Campomanes-Álvarez, S. Damas, O. Córdón, Modeling the facial soft tissue thickness for automatic skull-face overlay, *IEEE T Inf Foren Sec.* 10 (2015) 2057–2070.
- [10] C. Campomanes-Alvarez, O. Ibáñez, O. Córdón, Design of criteria to assess craniofacial correspondence in forensic identification based on computer vision and fuzzy integrals, *Applied Soft Computing* 46 (2016) 596–612.

- [11] C. Campomanes-Alvarez, O. Ibáñez, O. Córdón, Modeling the consistency between the bony and facial chin outline in craniofacial superimposition, in: 16th World Congress of the International Fuzzy Systems Association (IFSA), 2015, pp. 1612–19.
- [12] C. Campomanes-Alvarez, O. Ibáñez, O. Córdón, 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), 2016, pp. 437–444.
- [13] P. Broca, *Instructions craniologiques et craniométriques de la Société d'anthropologie de Paris*, Vol. 2, G. Masson, 1875.
- [14] A. Bertillon, *The bertillon system of identification*, McClaughry, Ed., Chicago, IL.
- [15] 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, *J Forensic Sci* 36 (1991) 480–500.
- [16] J. Huang, M. Zhou, F. Duan, Q. Deng, Z. Wu, Y. Tian, The weighted landmark-based algorithm for skull identification, in: *Computer Analysis of Images and Patterns*, Springer, 2011, pp. 42–48.
- [17] W. Jin, G. Geng, K. Li, 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, IEEE, 2013, pp. 317–320.
- [18] B. R. Campomanes-Álvarez, O. Ibáñez, F. Navarro, I. Alemán, O. Córdón, S. Damas, Dispersion assessment in the location of facial landmarks on photographs, *Int J Legal Med* 129 (1) (2015) 227–236.
- [19] C. N. Stephan, E. K. Simpson, Facial soft tissue depths in craniofacial identification (part i): an analytical review of the published adult data, *J Forensic Sci* 53 (2008) 1257–1272.
- [20] A. K. Ghosh, P. Sinha, An economised craniofacial identification system, *Forensic Science International* 117 (1) (2001) 109–119.
- [21] B. R. Campomanes-Álvarez, O. Ibáñez, F. Navarro, M. Botella, S. Damas, O. Córdón, Computer vision and soft computing for automatic skull-face overlay in craniofacial superimposition, *Forensic Sci Int* 245 (2014) 77–86.
- [22] T. W. Fenton, A. N. Heard, N. J. Sauer, Skull-photo superimposition and border deaths: Identification through exclusion and the failure to exclude, *J Forensic Sci* 53 (1) (2008) 34–40.
- [23] P. T. Jayaprakash, G. J. Srinivasan, M. G. Amraveswaran, Cranio-facial morphanalysis: a new method for enhancing reliability while identifying skulls by photo superimposition, *Forensic Sci Int* 117 (1) (2001) 121–143.
- [24] O. Ibáñez, A. Valsecchi, F. Cavalli, M. I. Huete, B. R. Campomanes-Alvarez, C. Campomanes-Alvarez, R. Vicente, D. Navega, A. Ross, C. Wilkinson, et al., Study on the criteria for assessing skull-face correspondence in craniofacial superimposition, *Legal Medicine* 23 (2016) 59–70.
- [25] D. V. Pesce, E. Vacca, F. Potente, T. Lettini, M. Colonna, Shape analytical morphometry in computer-aided skull identification via video superimposition, *Isran MY, Helmer RP. Forensic analysis of the skull: craniofacial analysis, reconstruction, and identification*. New York: Wiley-Liss, 1993.
- [26] M. Yoshino, H. Matsuda, S. Kubota, K. Imaizumi, S. Miyasaka, S. Seta, Computer-assisted skull identification system using video superimposition, *Forensic Sci Int* 90 (3) (1997) 231–244.
- [27] A. Ricci, G. L. Marella, M. A. Apostol, A new experimental approach to computer-aided face/skull identification in forensic anthropology, *Am J Foren Med Path* 27 (1) (2006) 46–49.
- [28] C. P. Pappis, N. I. Karacapilidis, A comparative assessment of measures of similarity of fuzzy values, *Fuzzy Set Syst* 56 (2) (1993) 171–174.
- [29] M. Sugeno, *Theory of fuzzy integrals and its applications*, Tokyo Institute of Technology, 1974.
- [30] J. G. Clement, *Craniofacial identification in forensic medicine*, Arnold, 1998.
- [31] G. Bellakov, A. Pradera, T. Calvo, *Aggregation functions: A guide for practitioners*, Vol. 221, Springer, 2007.
- [32] J. E. Buikstra, D. H. Ubelaker, *Standards for data collection from human skeletal remains*.
- [33] M. F. Anderson, D. T. Anderson, D. J. Wescott, Estimation of adult skeletal age-at-death using the sugeno fuzzy integral, *Am J Phys Anthropol* 142 (1) (2010) 30–41.
- [34] H. Imai, V. Torra, On a modeling of decision making with a twofold integral, in: *EUSFLAT Conf.*, 2003, pp. 714–717.
- [35] H. Tahani, J. M. Keller, Information fusion in computer vision using the fuzzy integral, *IEEE T Syst Man Cyb* 20 (3) (1990) 733–741.
- [36] Artec 3d scanners, www.artec3d.com/3d-scanner/artec-spider/.
- [37] O. Ibáñez, F. Cavalli, B. R. Campomanes-Álvarez, C. Campomanes-Álvarez, A. Valsecchi, M. I. Huete, Ground truth data generation for skull-face overlay, *Int J Legal Med* 129 (3) (2015) 569–81.
- [38] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, *Ann Math Stat* 11 (1) (1940) 86–92.
- [39] O. J. Dunn, Multiple comparisons among means, *J Am Stat Assoc* 56 (293) (1961) 52–64.
- [40] A. K. Jain, S. Z. Li, *Handbook of face recognition*, Vol. 1, Springer, 2005.
- [41] D. Austin-Smith, W. R. Maples, The reliability of skull/photograph superimposition in individual identification, *Journal of Forensic Science* 39 (2) (1994) 446–455.
- [42] D. Chai, Y. Lan, C. Tao, R. Gui, Y. Mu, J. Feng, W. Wang, J. Zhu, A study on the standard for forensic anthropologic identification of skull-image superimposition, *J. Forensic Sci* 34 (6) (1989) 1343–1356.
- [43] G. M. Gordon, M. Steyn, An investigation into the accuracy and reliability of skull-photo superimposition in a south african sample, *Forensic Sci Int* 216 (2012) 198.e1–6.
- [44] M. Yoshino, K. Imaizumi, S. Miyasaka, S. Seta, Evaluation of anatomical consistency in craniofacial superimposition images, *Forensic science international* 74 (1) (1995) 125–134.
- [45] 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, S. Damas, Study on the performance of different craniofacial superimposition approaches (i), *Forensic Sci Int* 257 (2015) 496–503.
- [46] O. Ibáñez, R. Vicente, D. Navega, C. Campomanes-Álvarez, C. Cattaneo, R. Jankauskas, M. Huete, F. Navarro, R. Hardiman, E. Ruiz, et al., Meprocs framework for craniofacial superimposition: Validation study, *Legal Medicine* 23 (2016) 99–108.
- [47] C. N. Stephan, B. Amidan, H. Trease, P. Guyomarc'h, T. Pulsipher, J. E. Byrd, Morphometric comparison of clavicle outlines from 3d bone scans and 2d chest radiographs: a shortlisting tool to assist radiographic identification of human skeletons, *Journal of forensic sciences* 59 (2) (2014) 306–313.
- [48] G. Beliakov, How to build aggregation operators from data, *International Journal of Intelligent Systems* 18 (8) (2003) 903–923.

Appendix A. Definitions

We introduce some basic definitions related with aggregation functions based on [31].

Definition 2. An aggregation function is a function of $n > 1$ arguments that maps the (n -dimensional) unit cube onto the unit interval $f: [0, 1]^n \rightarrow [0, 1]$, with the properties

$$(i) \underbrace{f(0, 0, \dots, 0)}_{n\text{-times}} = 0 \text{ and } \underbrace{f(1, 1, \dots, 1)}_{n\text{-times}} = 1.$$

$$(ii) \mathbf{x} \leq \mathbf{y} \text{ implies } f(\mathbf{x}) \leq f(\mathbf{y}) \text{ for all } \mathbf{x}, \mathbf{y} \in [0, 1]^n$$

Definition 3. An extended aggregation function is a mapping

$$F: \bigcup_{n \in \{1, 2, \dots\}} [0, 1]^n \rightarrow [0, 1],$$

such that the restriction of this mapping to the domain $[0, 1]^n$ for a fixed n is n -ary aggregation function f , with the convention $F(x) = x$ for $n = 1$.

This allows us to define and work with such families of functions of any number of arguments.

There are several semantics of aggregation, and the main classes are determined according to these semantics. In some cases we require that the high and low inputs average each other, in other cases aggregation functions model logical connectives (disjunction and conjunction), so that the inputs reinforce each other, and sometimes the behavior of aggregation functions depends on the inputs. The four main classes of aggregation functions are [31]:

- Averaging,
- Conjunctive,
- Disjunctive,
- Mixed.

Definition 4. An aggregation function f has averaging behavior (or is averaging) if for every \mathbf{x} it is bounded by $\min(\mathbf{x}) \leq f(\mathbf{x}) \leq \max(\mathbf{x})$.

Definition 5. An aggregation function f has conjunctive behavior (or is conjunctive) if for every \mathbf{x} it is bounded by $f(\mathbf{x}) \leq \min(\mathbf{x}) = \min(x_1, x_2, \dots, x_n)$.

Definition 6. An aggregation function f has disjunctive behavior (or is disjunctive) if for every \mathbf{x} it is bounded by $f(\mathbf{x}) \geq \max(\mathbf{x}) = \max(x_1, x_2, \dots, x_n)$.

Definition 7. An aggregation function f is mixed if it does not belong to any of the above classes, i.e., it exhibits different types of behavior on different parts of the domain.

Last, we define an important property of aggregation functions for this application: symmetry.

Definition 8. *An aggregation function f is called symmetric, if its values does not depend on the permutation of the aggregation of the elements, i.e.,*

$$f(x_1, x_2, \dots, x_n) = f(x_{P(1)}, x_{P(2)}, \dots, x_{P(n)}),$$

for every x and every permutation $P = (P(1), P(2), \dots, P(n))$ of $(1, 2, \dots, n)$.