Int J Legal Med (2017) 131:823–833
DOI 10.1007/s00414-016-1439-8

ORIGINAL ARTICLE

Use of pattern recognition and neural networks for non-metric sex diagnosis from lateral shape of calvarium: an innovative model for computer-aided diagnosis in forensic and physical anthropology

Fabio Cavalli[1] · Luca Lusnig[1] · Edmondo Trentin[2]

Received: 3 October 2015 / Accepted: 16 August 2016 / Published online: 29 August 2016

Abstract Sex determination on skeletal remains is one of the most important diagnosis in forensic cases and in demographic studies on ancient populations. Our purpose is to realize an automatic operator-independent method to determine the sex from the bone shape and to test an intelligent, automatic pattern recognition system in an anthropological domain. Our multiple-classifier system is based exclusively on the morphological variants of a curve that represents the sagittal profile of the calvarium, modeled via artificial neural networks, and yields an accuracy higher than 80 %. The application of this system to other bone profiles is expected to further improve the sensibility of the methodology.

Keywords Sex determination · Artificial intelligence · Neural network · Forensic anthropology

Introduction

A reliable method for the determination of sex from human skeletal remains is essential for identification both in forensic cases and in paleodemographic studies on ancient populations. Many skeletal traits have been investigated for this purpose in adult skeletons, with results of various grade of efficiency. The bones routinely used in sex identification are pelvis and skull [1], although some researches sustains that postcranial elements are to be preferred to the skull for estimating sex when the pelvis is unavailable [21]. The sexual dimorphism is better recognizable in the pelvis but, because of its complex shape, the pelvis is often found in very poor conditions, while the skull (and, particularly, the cranial vault) is generally better preserved and more easily reconstructed if found fragmented [15]. The study of sexual dimorphism has been the subject of many morphologic and metric studies. Morphologic methods are based on the shape of the skull and have the main disadvantage of being heavily operator-dependent [14] and incorrect in 10–20 % of cases [25]. Craniometric methods are generally based on direct measurement of the skull [16, 19], on teleradiographic projection [10], volume-rendered cranial CT scan [17], or 3D digital skull [13]. In 1996, Hsiao et al. using 18 variables from cephalometric lateral teleradiographic plots claimed to be able to determine the sex of an individual with 100 % accuracy [9]. Furthermore, these authors say that they can determine the sex of a subject to 98 % accuracy by using only three variables considered more significative. However, subsequent studies applying this method to larger European samples seems not to confirm the absolute validity of the method, reporting a 95.6 % accuracy over 18 cephalometric variables [24]. It worths highlighting that such cephalometric variables need to be extracted manually via direct measurement by an expert anthropologist, and no automatic feature extraction process is feasible to this end. The purpose of the present study is to explore the potentialities of pattern recognition methods and artificial neural

✉ Fabio Cavalli
fabio.cavalli@aots.sanita.fvg.it

Edmondo Trentin
trentin@dii.unisi.it

[1] Research Unit of Paleoradiology and Allied Sciences,
LTS - SCIT, Integrated University Health Unit of Trieste,
via della Pietà 2/1, 34100 Trieste, Italy

[2] Dipartimento di Ingegneria dell'Informazione e Scienze
Matematiche, Università di Siena, Via Roma 56, 53100
Siena, Italy

network (ANN) to automatically determine the sex in lateral shape of the calvarium and to value the accuracy obtained. The perspective is to project, in the future, intelligent automatic pattern recognition systems in the anthropological domain to effectively support the physical or the forensic anthropologist. This kind of tools, sometimes based on ANNs, are not unusual in the field of clinical medical diagnostic [27]. Note that ANNs (and, pattern recognition approaches in general) are like "black boxes," meaning that they are trained from real data in such a way that they learn how to make decisions on the sex of the skeletal remains at hand, but they do not provide the user with any human-readable understanding of the rationale behind their internal decision-making process [2].

Material and methods

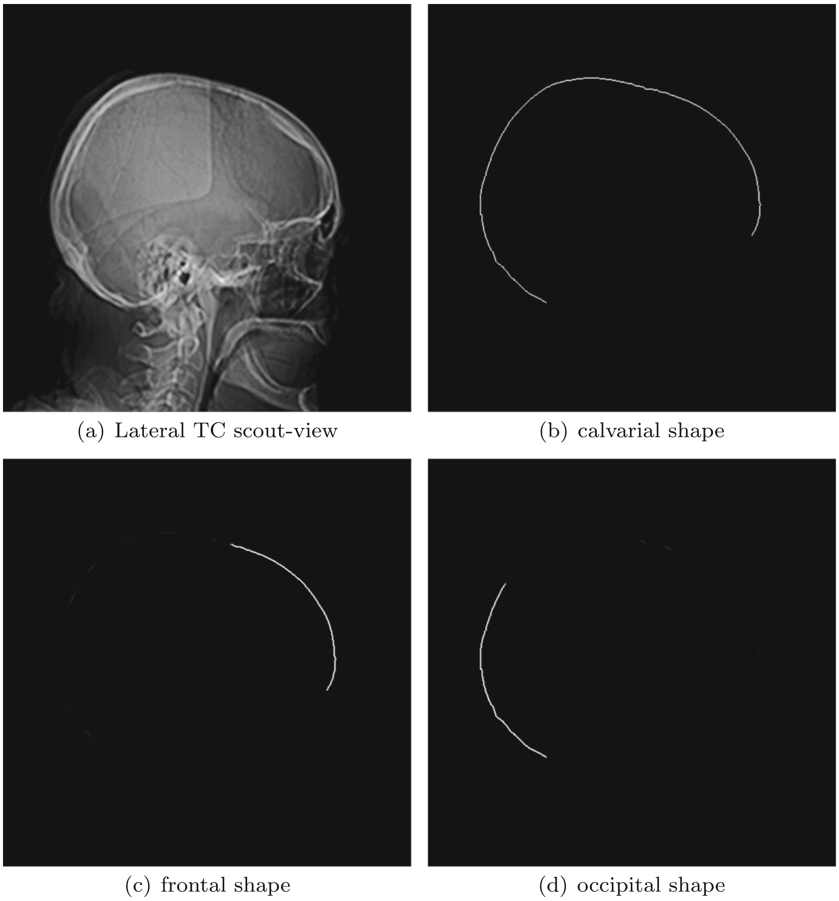
Dataset

For this study, 1700 lateral CT scanograms of healthy, adult Caucasian subjects were randomly selected from our PACS database. This sample was composed of 850 male and 850 female within an age range of 25–92 years. The CT scanogram was chosen because it is routinely performed before a cranial MDCT examination and because for our specific purposes, i.e., the determination of the external shape of the skull in norma lateralis, it is basically as reliable as the lateral cephalometric radiogram [4]. Also, scanogram is preferred over traditional cephalometric plain film since it does not suffer from distortion due to the geometry of cone-beam X-ray that is conventionally used in cephalometric equipments [12]. The patients were selected on the basis of their residence in the province of Trieste (North–Eastern Italy), since the population of this geographical area is the result of a complex historical genetic crossover between Italic, Germanic, and Slavic populations. Lateral cranial scanograms were automatically selected and anonymized by PACS facilities, registering only the sex and the age. The MDCT scans were performed at the Department of Diagnostic Imaging of the Hospital University Enterprise of Trieste between years 2005 and 2010. All those scans that show bad lateral positioning were discarded, using as a criterion of correctness the perfect alignment of the temporomandibular joints and of the gonion. The images were automatically transformed from DICOM to JPG format, maintaining the original matrix size. Lateral CT scanograms were obtained with a 16 or 32 multidetector CT scanner (Aquilion, Toshiba medical Inc.) using the standard cranial preset (120 kVp, 150 mAs, matrix size 512 × 512) (Fig. 1a).

Method

Sex identification is herein formulated in terms of an automatic pattern recognition task [6]. The latter is faced according to a four-step procedure, namely:

1. Pattern acquisition: patterns are acquired from CT scanograms of the skulls, that are represented in the form of a bit matrix. We must underline that the nature of the scanogram is inherently suitable for this purpose because its mode of acquisition allows to obtain an orthostatic projection of the skull, namely without geometric distortion. Of course, the level of detail of such images requires a significantly high resolution of the corresponding bitmaps, involving a heavy burden in terms of computational space and time on the subsequent processing steps. This entails the need for the next step, i.e., feature extraction.
2. Feature extraction: each raw image acquired in step 1 is pre-processed in order to come up with a viable representation $\mathbf{x}$, known as the feature vector (or, *pattern*). The latter is sought in such a way that (i) the pattern $\mathbf{x}$ extracted from a given CT scanogram $\tilde{\mathbf{x}}$ has a much lower dimensionality than $\tilde{\mathbf{x}}$ has; (ii) still, $\mathbf{x}$ preserves as much as possible the useful information conveyed by $\tilde{\mathbf{x}}$ to the end of sex determination. The feature extraction process used in this study is presented in "Feature extraction" section.
3. Classifier training: the core of the sex identification process relies on a classifier, i.e., a machine realizing a decision rule which assigns any given feature vector $\mathbf{x}$ to the corresponding expected "class" (either *male* or *female*). Two major families of classifiers are found in the literature, namely statistical pattern recognition techniques [6] and machine learning approaches [2]. Both of them revolve around the notion of estimating, or *training* the classifier in some statistically optimal way from a subsample of the dataset. Artificial neural networks (ANNs) are the learning paradigm of choice in this paper. Classifiers, ANNs, and their training are reviewed in "Classifier training" section.
4. Pattern classification: once a classifier has been estimated from the data, the corresponding decision rule can be applied to the actual classification task in the field. Quantitative evaluation of the statistical robustness of any estimated classifier is fundamental in order to assess its effectiveness in identifying the sex of speciements as correctly as possible. To this end, a subsample of the data is used along with specific statistical model validation procedures. The error rate (that is, the percentage of misclassifications on the data of



(a) Lateral TC scout-view

(b) calvarial shape

(c) frontal shape

(d) occipital shape

Fig. 1 Contour partition

a separate, independent test set) is the basic validation criterion used in the experiments presented in "Results" section.

While the acquisition of patterns (step 1) in the present setup is uniquely entailed by the very nature of the task (i.e., the CT scanograms *are* the patterns herein), and the classification (i.e., sex identification) of CT scanograms is straightforward once a trained classifier is given (step 4), the following sections cover in some detail the remaining, crucial topics of feature extraction (step 2) and classifier training (step 3), providing the interested reader with the information required in order to replicate the experiments, and/or to apply the proposed technique to real-world scenarios. Although in-depth reviewing and understanding of image processing and machine learning algorithms is way beyond the scope of the paper, we try and put forward a presentation of the proposed techniques which is as simple, schematic, and self-contained as possible. Readers with no specific background in this field can find the details on individual processing steps in the corresponding bibliographic references. Further technicalities on neural network training for probability density estimation are handed out in the Appendix.

Feature extraction

Visual feature extraction from the JPG-format scanograms is accomplished according to the following five-step image-processing procedure:

1. Scanogram filtering: firstly, the image is filtered, reducing the presence of noise and enhancing the readability of the cranial contour.
2. Contour detection: the cranial profile is detected and singled out from the remaining background image (Fig. 1b).
3. Contour partition: the cranial profile is partitioned into portions, obtaining the frontal shape (Fig. 1c) and the occipital shape (Fig. 1d).
4. Extraction of signatures: for each portion of the profile and for the whole cranial profile, a reduced- and fixed-dimensionality set of values is extracted (by sampling from a specific *signature function*).

5. Fourier analysis of signatures: the fast Fourier transform is applied to the sub-sampled set of signatures, obtaining the ultimate visual features at various frequencies.

The next sections report on these procedural steps.

Contour detection

Starting from the filtered image, the visual contour of the lateral profile of the calvarium in the scanogram is determined. This is a two-step process, namely (1) edge detection. To this end, we developed (LL) a specific edge detection software in Python. The software detects the contour relying on a preliminary manual identification of the points corresponding to the *nasion* and to the *opisthion*. Quality control of the software relied on a human expert visually verifying the correctness of the detected edges during a preliminary test stage; (2) connection of individual edges. The overall process is accomplished by a technique relying on Canny algorithm [3] followed by thresholding, further reducing the presence of noise [26]. Eventually, the resulting length of contours turns out to be in the range of 400–450 pixels per image.

Contour partition

Two portions of the calvarial shape are usually considered to be relevant in sex identification, namely the frontal shape (from *nasion* to *bregma*) and the occipital shape (from *lambda* to *opisthion* [20]. In this paper, we consider mainly the complete cranial contour (hereafter referred as "calvarium"), from *nasion* to *opisthion*. However, we test also the diagnostic accuracy of "frontal" and "occipital" shape in order to verify if these segments can obtain a better performance than the whole cranial shape. Let us underline that here the terms *frontal* and *occipital* have not their strict anatomical meaning, but they are used to refer informally to arbitrary *sectors* of the cranial contour where presumably the contribution of that specific anatomical shape is sufficiently relevant. Moreover, the correct visibility of the exact location of bregma and lambda landmarks in scanogram depends on many variables as, for example, the grade of their calcification. For each of the two portions of the cranial profile, the corresponding sub-shapes within the overall cranial contour were localized and extracted as follows. Upon removal of the maxilla and mandible area (i.e., the lower part of the image, basically coinciding with the jaw), points along the contour are assigned to the corresponding sub-shape by means of an automatic partitioning procedure. The latter exploits knowledge of the fact that roughly one third of the contour (starting from the *nasion*) forms the frontal shape and the posterior third (starting from the *opisthion*) is the location of occipital shape.

Extraction of signatures

For each cranial portion, a specific family of signature functions [5] is then computed from the corresponding set of pixels. The centroid-distance signature function is adopted [26] for several reasons: (i) it is known to be effective in the representation of generic shapes [23]; (ii) it is simple to compute; (iii) it is translation-invariant [26]; and (iv) comparative empirical evaluations reported in [26] point out that it mostly outperforms its competitors. The resulting sequence of signatures is then sub-sampled[1] at regular intervals in order to attain low and fixed dimensionality representations of the sub-shapes. This is accomplished via the equal points sampling (EPS) technique, which is proofed in [26] to be as simple as effective (reducing the noise in the shape boundary, too). EPS samples are uniformly spaced along the sub-shape contour. Exploiting the peculiarity of the cranial shape (substantially a convex figure with no significant discontinuities), in order to represent the overall cranial contour, we sampled the signatures via EPS using a pixel-wise step of 3.5 pixels (discretized such that the signatures located in correspondence with pixels in positions 1, 4, 8, 11, 15, . . ., were sampled) with no significant information loss. As reported in "Contour partition" section, the first and the last third of the overall profile were considered for the analysis of the frontal and of the occipital portions of the contour, respectively. Taking account of the different complexity of the shape of the contour portions, frontal and occipital portions were sampled via EPS applying a 2.5 step in order to better describe the shape of the corresponding parts of the contour. In so doing, we ended up with 128 samples (i.e., signatures) for the whole cranial contour (including the parietals), and 64 signatures for each of the frontal and occipital regions, respectively.

Fourier analysis of signatures

Finally, actual visual features are extracted from the sub-sampled sequences of signatures obtained so far by application of the usual fast Fourier transform (FFT), which has been proven effective in automatic processing of the forehead shape of the skull [11]. The first (few) low-frequency terms of the FFT tend to capture global features of the cranial contour, while more detailed and local features are

[1] From now on, we use the terms "sample" and "sub-sample" according to their statistical meaning, i.e., random data samples drawn from a population. The specific quantity they refer to (e.g., pixels, signature functions, ...) is made clear by the context.

represented by the higher frequency terms [26]. As much as 32 frequency coefficients (in the (0, 1) range) yielded by the FFT were retained for representing each of the frontal and the occipital shapes, while 64 frequency coefficients were used for the cranium. The FFT guarantees rotation invariance, which is hardly achievable in the space domain but becomes feasible in the frequency domain (by taking into consideration the magnitude of coefficients of the FFT and dropping the phase information), as well as scale invariance (by proper normalization of the magnitude of the first half of the FFT coefficients) [26].

Classifier training

As we say, sex identification is herein formulated in terms of a pattern classification problem: given the feature vector $\mathbf{x}$ extracted according to the process described in the previous section, decide whether $\mathbf{x}$ is more likely to belong to class ω_0 (*female*) or ω_1 (*male*). To this end, a discriminant function $g(\omega_i, \mathbf{x})$ is sought such that the scanogram $\mathbf{x}$ is identified to belong to a female specimen if $g(\omega_0, \mathbf{x}) \geq g(\omega_1, \mathbf{x})$; otherwise, it is identified to be a male. This decision rule is implicitly probabilistic in nature, meaning that regardless the form chosen for realizing $g(.)$, there is an understanding of its potential inexactness in the general case. Empirical evaluation is thus required in order to assess the robustness of any given choice for $g(.)$ in terms of the empirical estimate of the corresponding probability of error. Bayes decision rule is used in this paper, which theoretically minimizes the probability of having misclassifications [6] turning out to be optimal in principle. Therefore, we let $g(\omega_i, \mathbf{x}) = P(\omega_i \mid \mathbf{x})$, where $P(\omega_i \mid \mathbf{x})$ is the posterior probability of the specimen belonging to class ω_i (either *female* or *male*) given the fact that we observed the features $\mathbf{x}$. It goes without saying that any suitable realizations of Bayes decision rule shall rely on robust statistical estimates of the (otherwise unknown) quantity $P(\omega_i \mid \mathbf{x})$ from a data sample collected in the field, known as the training set. Two major alternatives are viable to this end (both of which are investigated in this study):

1. Direct estimate of $P(\omega_i \mid \mathbf{x})$ from the training set, relying on statistical approaches (e.g., k-nearest neighbor [6]), or on ANNs;
2. Factorization of $P(\omega_i \mid \mathbf{x})$ via Bayes theorem [6] as

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{1}$$

where the probability density function (pdf) $p(\mathbf{x} \mid \omega_i)$ expresses the likelihood of the class ω_i given the vector of observed features $\mathbf{x}$, $P(\omega_i)$ is the *a priori* knowledge on the probability of a specific class, and the evidence $p(\mathbf{x})$ is a pdf representing the probability distribution of the observations $\mathbf{x}$ in the feature space regardless of the class they belong to. It is seen that an indirect estimation of $P(\omega_i \mid \mathbf{x})$ is obtained from the right-hand side of Eq. 1) once estimates are given for the quantities $p(\mathbf{x} \mid \omega_i)$, $P(\omega_i)$, and $p(\mathbf{x})$. Since it is straightforward to see that the latter can be rewritten as $p(\mathbf{x}) = \sum_{j=0}^{1} p(\mathbf{x} \mid \omega_j)P(\omega_j)$, only the former quantities actually need to be estimated. While a robust estimate of the prior probability $P(\omega_i)$ is achieved by just counting the relative frequencies of the classes in the training set (for instance, in this study, we let $P(\omega_0) = P(\omega_1) = 0.5$), reliable estimates of $p(\mathbf{x} \mid \omega_i)$ are sought, either via statistical techniques or ANNs.

Albeit both approaches require estimating, or "learning," a model (either statistical or neural in nature) of a specific probabilistic quantity, the former learning problem is profoundly different from the latter. In fact, $P(\omega_i \mid \mathbf{x})$ is an actual probability, i.e., it ranges over the (0, 1) interval (furthermore, since the classes of interest are herein disjoint, we have $P(\omega_0 \mid \mathbf{x}) + P(\omega_1 \mid \mathbf{x}) = 1$, as well). Algorithms for training regular ANNs to estimate $P(\omega_i \mid \mathbf{x})$ are popular in the literature, and quite straightforward. An optimal approach, relying on the Widrow-Hoff algorithm applied to a family of ANNs known as the Multilayer perceptron (MLP), is concisely reviewed in the next section. To the contrary, estimating $p(\mathbf{x} \mid \omega_i)$ may be harder since this quantity is a pdf (i.e., it potentially ranges over the $(0, +\infty)$ interval) and the constraints to be satisfied in order to respect the axioms of probability are far less obvious. As a matter of fact, algorithms for pdf estimation via ANNs are not popular in the literature, and the few exceptions usually rely on the practical assumption that the input data are one-dimensional (while the feature vectors used in this study are necessarily high-dimensional). For these reasons, we propose a technique for ANN-based estimation of pdfs. It is sketched out in "Algorithm for learning $p(\mathbf{x} \mid \omega_i)$ via MLP" section, and detailed in the Appendix. Good news is that both approaches can be readily realized, for all practical purposes, using just any standard, public domain MLP software simulator. Moreover, their probabilistic-grounded interpretation allows for a justified, robust combination of both, resulting in their mutual reinforcement (as shown by empirical evidence, see "Results" section).

Widrow-Hoff algorithm for learning $P(\omega_i \mid \boldsymbol{x})$ via MLP

A review of ANNs and their application to pattern classification is beyond the scope of the paper. We refer the interested reader to [2]. For all intents and purposes of this study, the following background notions are used. An

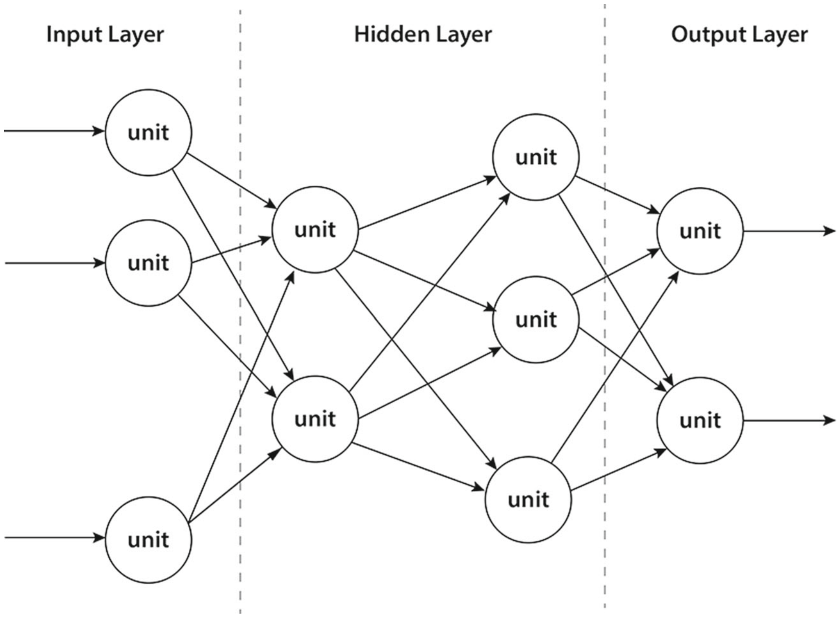


Fig. 2 Architecture of a generic MLP with two hidden layers

ANN is a learning machine having an architecture structured like a graph whose vertexes are simple processing units, and whose edges are pairwise directed connections between units. Each unit realizes a real-valued activation function (usually a linear transformation, or a logistic sigmoid $y = 1/(1+\exp(-x))$) transforming its input argument into an output value. Each connection is characterized by a specific real number known as the connection weight, with the understanding that a positive weight pinpoints an excitatory effect of the source unit onto the destination unit along the corresponding connection, while a negative weight represents an inhibitory effect. As a general rule, the higher the absolute value of the weight, the stronger the contribution the source unit gives to the behavior of the destination unit.

As shown in Fig. 2, multilayer perceptrons (MLPs) are a specific family of ANN architectures, which are (i) layered, (ii) feed-forward, and (iii) fully connected. By "layered" we mean that the overall set of units can be partitioned into subsets, known as "layers," such that the units in a given layer interact with the adjacent[2] layers but not with any other layer in the MLP; as well, no (lateral) interactions occur among units within the same layer. The first (left-most in Fig. 2) layer of the MLP is the input layer, receiving the input feature vector **x** that requires being processed by the machine. There are as many input units as the dimensionality of **x**, and ith input unit takes responsibility for ith feature, say x_i. Similarly, the last layer of the MLP is the output layer, expected to return the response (i.e., the output) **y** that the MLP associates with the current input **x** upon its internal processing. There is one output unit for each component of **y**, according to an ordered one-to-one relationship. The number of output units is uniquely determined by the nature of the (*input*, *output*) transformation that the MLP is expected to compute. Since we are interested in estimating class-posterior probabilities for sex classification, we rely on MLPs having only one output unit whose output value (when the MLP is fed with input **x**) is interpreted as an estimate of $P(\omega_o \mid \mathbf{x})$ (thence, $P(\omega_1 \mid \mathbf{x}) = 1 - P(\omega_o \mid \mathbf{x})$). All the remaining layers of the MLP (in the middle between input and output) are said to be the hidden layers, and realize a cascade of progressive internal representations and transformations of the original input until the ultimate output is computed. There may be one or more hidden layers, each having its own number of processing units (to be fixed empirically by trial-and-error, according to some model selection/validation statistics). In standard MLPs, the logistic sigmoid is used as the activation function associated with the hidden units, while the linear (or, identity) transformation $y = x$ is the function of choice for input and (usually) output units. In this study, we resort to a sigmoid output, as well, since its ranging over (0, 1) ensures a proper probabilistic interpretation of the results. As we said, MLPs are feed-forward ANNs meaning that there is a fixed direction of consecutive, intermediate representations and transformations of the information from the input layer, through adjacent hidden layers, up to the output layer. This direction is shown in the form of arrows (i.e., directed edges of the underlying graph) in Fig. 2. Also, the machine is fully connected since all the units in any given layer are connected, via feed-forward connections, with all of the units in the subsequent layer.

[2] According to the topology of the graph defined by the specific connections.

Besides these architectural aspects, the interesting thing about ANNs is that they can learn from examples, meaning that the connection weights are initially set at random and later adapted automatically by observing the data (the "examples") in the training set, example after example, such that the input-output transformation realized accordingly by the ANN is progressively refined in order to capture the implicit relationship underlying the dataset. A supervised training set $\mathcal{T} = \{(\mathbf{x}_j, \hat{y}_j) \mid j = 1, \ldots, n\}$ is assumed in the study, where each feature vector $\mathbf{x}_j$ (representing the generic jth CT-scan, out of the n CT-scans at hand) is explicitly presented to the MLP training algorithm in association with its corresponding *target output* $\hat{y}_j$, which is the expected output value that we wished an educated ANN yielded when fed with input $\mathbf{x}_j$. The popular backpropagation (BP) algorithm[3] is then used to train the MLP from $\mathcal{T}$ [2].

It is seen that if the target outputs satisfy a probabilistic interpretation in the form $\hat{y}_j = P(\omega_0 \mid \mathbf{x}_j)$, then the MLP is implicitly expected to learn Bayes decision rule, provided that a suitable architecture is used and that BP can actually converge to a suitable set of connection weight values. To this end, although convergence of BP cannot be always guaranteed in real-world scenarios, fundamental theoretical results (namely Cybenko's universality theorem [8]) proof that such a MLP exists that approximates the actual function $P(\omega_0 \mid \mathbf{x}_j)$ to any desired degree of precision.[4] At this point, the only remaining catch is the following: how can we define the target outputs to be associated with the individual CT scans such that they represent class posterior probabilities (i.e.,, such that $\hat{y}_j = P(\omega_0 \mid \mathbf{x}_j)$ for $j = 1, \ldots, n$), given the fact that we do not have the knowledge of these probabilistic quantities in the first place? It is a significant contribution from Richard and Lippmann [18] to tackle the issue by resorting to a much simpler, equivalent, yet viable definition of the target outputs. In fact, according to [18], it is sufficient to let $\hat{y}_j = 1$ if $\mathbf{x}_j$ belongs to class ω_0 (i.e., if the corresponding CT scan represents a female skull), and $\hat{y}_j = 0$ otherwise (i.e., if we are coping with a male subject), to make sure that BP training of the MLP results in the approximation sought of Bayes posterior probability. In practice, this means that in preparing the training set, all the female CT scans are labeled with the value 1, while all male CT scans are labeled with 0. Since this labeling was originally proposed by Widrow and Hoff for estimating a linear classifier [6], we refer to this training scheme as Widrow-Hoff training of MLPs (WH-MLP).

Table 1 LDA: parameters and results

	accM (%)	accF (%)	Accuracy (%)
Calvarium	36.00	86.50	61.25
Occipital	79.50	36.50	58.00
Frontal	14.50	73.00	43.75

Algorithm for learning $p(\boldsymbol{x} \mid \omega_i)$ via MLP

As we say, BP requires the explicit definition of target outputs to be uniquely associated with the patterns representing the CT-scans in the training set. In the setup reviewed in the previous section, the targets sought are the posterior probabilities of classes, i.e., $P(\omega_0 \mid \mathbf{x}_j)$, which are not known in advance but are effectively replaced by 0/1 surrogates. Learning the class-conditional pdfs from examples via BP poses an analogous problem, since the target outputs $p(\mathbf{x}_j \mid \omega_i)$ for $j = 1, \ldots, n$ are unknown likewise. Unfortunately, no simple workaround along Richard and Lippmann's line is available in this respect. Thence, we propose a simple yet effective technique which stems from an algorithm for density estimation we first presented in [22]. The algorithm is concisely reviewed in Appendix. It relies on the idea of generating the target $\hat{y}_j$ for the generic training pattern $\mathbf{x}_j$ as $\hat{y}_j = \tilde{p}(\mathbf{x}_j \mid \omega_i)$ where $\tilde{p}(.)$ denotes a statistical estimation of the unknown pdf $p(.)$ by means of an unbiased variant of a popular nonparametric approach to pdf estimation, namely the Parzen Window method [6]. The resulting ANN is thence referred to as the PW-MLP. Implementation of the approach is readily achieved via a BP software simulator along with any non-trivial statistical toolbox. It is noteworthy that, once trained, the PW-MLP can be used as a stand-alone tool for sex identification (relying on Bayes decision rule), or it may be combined with the WH-MLP in

Table 2 k-NN: parameters and results

	k	accM (%)	accF (%)	Accuracy (%)
Calvarium				
	1	54.00	68.50	61.25
	2	59.00	68.50	63.75
	3	63.50	67.50	65.50
Occipital				
	1	56.00	50.00	53.00
	2	56.00	47.00	51.50
	3	57.00	49.00	53.00
Frontal				
	1	48.50	57.50	53.00
	2	39.00	53.00	46.00
	3	30.00	61.00	45.50

[3] BP is an MLP-tailored instance of the gradient method for online non-linear optimization. Any MLP software simulator is expected to provide the user with BP (or one of its many variants).

[4] Provided that $P(\omega_0 \mid \mathbf{x}_j)$ is continuous and limited.

Table 3 Parzen window: parameters and results

	h	accM (%)	accF (%)	Accuracy (%)
Calvarium	2	63.50	63.00	63.25
Occipital	3	65.50	64.50	65.00
Frontal	5	70.50	69.50	70.00

order to realize a multiple classifier system turning out to be more robust than its individual constituents.

Results

The first experimental stage revolves around a traditional training-test procedure. The dataset was randomly split into a training set (1300 patterns) and a test set (the remaining 400 patterns), making sure that the sex distributions were uniform (i.e., $P(\omega_0 \mid \mathbf{x}) = P(\omega_1 \mid \mathbf{x})$ on both subsets). This partitioning results in a test set that is large enough to be representative of the statistical properties under investigation at the visual level (allowing for a robust validation of the classifiers performance), without affecting significantly the amount of data available for the classifier training. Five different algorithms were applied (three statistical techniques and two ANN-based methods, plus the combination of the latter ones), in order to achieve a robust comparison. A classic linear discriminant [6] estimated with the usual Widrow-Hoff algorithm was tested first. The learning rate was set to 0.5 and the algorithm was iterated for 30,000 iterations. Results are reported on in Table 1, where *accF* represents the accuracy evaluated over the female population only (i.e., the percentage of female patterns that were correctly classified) and *accM* is the accuracy evaluated over the male population.

Next, a standard k-nearest neighbor (k-NN) algorithm was used. Results are shown in Table 2 as a function of k (that is, the number of neighbors of the test pattern $\mathbf{x}$ that are considered for classifying $\mathbf{x}$ itself). The results worsened for $k > 3$, and are not shown in the table. This is not surprising, since the k-NN is a memory-based approach (which, by its own nature, may not generalize properly) and its more complex instances (i.e., having larger k) may as well tend to reduce its generalization capabilities even further. Moreover, in k-NN, there is a critical trade-off between larger values of k and the point-wise precision of the estimated value of the pdf at the specific location of interest [6].

The third (and, last) statistical approach we applied relies on Bayes decision rule with Parzen-window estimates of the class-conditional probability density functions. Results are shown in Table 3, where h represents the initial bandwidth of the Gaussian window function [6].

Next, the neural models handed out in "Widrow-Hoff algorithm for learning $P(\omega_i \mid \mathbf{x})$ via MLP" and "Algorithm for learning $p(\mathbf{x} \mid \omega_i)$ via MLP" sections were evaluated individually as well as jointly (according to the multiple classifier perspective outlined at the end of "Algorithm for learning $p(\mathbf{x} \mid \omega_i)$ via MLP" section). In the first experiment, the MLP was trained over target outputs defined according to the Widrow-Hoff technique. A standard decision threshold θ for assigning a pattern to either the *female* or *male* classes was used, set halfway as $\theta = 0.5$. A three-layer MLP architecture was used, involving one hidden layer of sigmoid activation functions. The number *hidden* of hidden units and their smoothness *smooth*, as well as the major learning parameters (momentum rate *mr*, learning rate *lr*, the number of training iterations *epochs*) were selected according to the specific contour partition under consideration form time to time. Results are presented in Table 4.

In the next experiment, the training algorithm presented in "Algorithm for learning $p(\mathbf{x} \mid \omega_i)$ via MLP" section was applied. Two sex-specific PW-MLPs were used, in order to estimate the corresponding sex-conditional pdfs. The PW-MLPs were trained, respectively, on all and only the CT-scans belonging to the corresponding class. Bayes decision rule relying on these neural estimates of the sex-conditional pdfs was applied for classification of the test patterns, along the guidelines pointed out at the end of "Algorithm for learning $p(\mathbf{x} \mid \omega_i)$ via MLP" section. At each experimental run, the two sex-specific MLPs architectures and training parameters were set identical to each other, and equal to those we used in the previous experiment. Table 5 hands out the results as a function of the learning parameters.

Finally, a multiple classifier system was obtained by combining the three MLPs with each other as follows. Current CT-scan $\mathbf{x}$ is fed to the WH-MLP first, and the corresponding output is used to estimate the posterior probabilities $P(\omega_i \mid \mathbf{x})$ for both sex classes (either *female* or

Table 4 WH-MLP: parameters and results

	Hidden	mr	lr	Smooth	Epochs	accM (%)	accF (%)	Accuracy (%)
Calvarium	16	0.2	0.1	1.25	20,000	81.50	83.00	82.25
Occipital	12	0.0	0.05	1.3	8000	67.50	63.50	65.50
Frontal	12	0.0	0.01	1.1	6000	79.00	61.00	70.00

Table 5 PW-MLPs: parameters and results

	Hidden	mr	lr	Smooth	Epochs	accM (%)	accF (%)	Accuracy (%)
Calvarium	12	0.01	0.3	20	20,000	72.50	79.50	76.00
Occipital	12	0.0075	0	17.5	30,000	57.00	66.00	61.50
Frontal	12	0.005	0.15	10	30,000	51.00	55.00	53.00

male), as explained in "Classifier training" section. Let ω_i be the winner class (i.e., the sex having the highest posterior probability). A threshold τ on the estimated $P(\omega_i \mid \mathbf{x})$ is fixed, such that if $P(\omega_i \mid \mathbf{x}) < \tau$ then the risk involved (that is, the probability of error $1 - P(\omega_i \mid \mathbf{x})$) is considered to be excessive (namely, only slightly less than 0.5). In this case, the two class-conditional PW-MLPs are used instead for deciding on the sex of $\mathbf{x}$. It is expected that the different nature of the probabilistic laws modeled by the former and the latter MLPs may complement each other to a certain degree, easing the classification task whenever one of the networks cannot make a clear decision.

The expectation was confirmed empirically. Results are reported in Table 6, where #WH denotes the number of test patterns classified by the WH-MLP, while #PDF represents the number of CT-scans in the test set that were classified relying on the PW-MLPs. It is seen that when using a low threshold (namely, close to 0.5), nearly all the patterns are classified via WH-MLP. Vice versa, as long as the threshold is increased (becoming higher and higher than 0.5), the fraction of data whose classification via WH-MLP is considered untrustworthy increases as well (clearly, not in a linear manner), and the PW-MLPs are used more intensively (still playing a minor role, if compared with the WH-MLP). Experiments with the multiple classifier system were limited to the whole cranium. Increasing τ beyond 0.63 did not yield any further improvement in terms of accuracy.

Based on the results of the previous investigations, we eventually evaluated the most significant approaches emerged so far by relying on a variant of the tenfold cross-validation strategy with resampling. A separate, additional fold was generated for automatic model and parameter selection purposes, as well. For each fold, the dataset was split into a training set (1000 patterns) and a test set (400 patterns) using bootstrap-like Monte Carlo case resampling [7]. We herewith limit our attention to the whole cranium case. Results are reported in Table 7. The results are expressed in terms of average accuracy ± standard deviation. Note that a value of $k = 5$ turned out to be best for the k-NN algorithm in this scenario. The reduced amount of available training data per fold with respect to the traditional training-test setup (reduction needed in order to split the overall dataset into reasonably-sized individual folds) is likely to account for the relative loss observed in terms of accuracy, which does not affect our general conclusions.

Table 6 Multiple classifier system, whole cranium case

Threshold	#WH	#PDF	accM (%)	accF (%)	Accuracy (%)
0.510	394	6	82.00	83.00	82.50
0.550	368	32	84.50	85.00	84.75
0.575	359	41	85.00	85.50	85.25
0.580	358	42	85.00	86.00	85.50
0.590	353	47	84.50	86.50	85.50
0.600	350	50	84.50	87.00	85.75
0.610	345	55	84.50	88.50	86.50
0.630	334	66	84.50	89.50	87.00

Discussion

In spite of the fact that the aforementioned results turn out to be lower than the statistical discrimination methods reported in the literature, the present method suggests several considerations. First of all, it takes into consideration a single, non-metric characteristic of the skull, namely its shape, limited to the calvarium, i.e., the opisthion and the nasion are not taken into account. From the anthropological standpoint, it is interesting that the *whole* calvarium shape that is generally considered of modest relevance to the sex determination process, seems to contains significant information about its sexual dimorphism which allows for a statistically significant recognition accuracy. With respect to the metrical approaches in cranial sex diagnosis, where manual extraction of multiple features (measurements) is required, the present framework relies on a completely automatic feature extraction/pattern classification procedure. It is likely that the satisfying performance of the system are mostly due to the following facts: (1) the system relies on shapes rather

Table 7 Accuracies (%) over the tenfold cross-validation setup (whole cranium case): Avg. accuracy ± std. dev

Model	Frontal	Occipital	Calvarium
k-NN	68.60 ± 0.79	54.14 ± 0.68	69.30 ± 0.86
Parzen window	68.62 ± 0.78	56.95 ± 0.92	72.05 ± 0.92
PW-MLP	68.71 ± 1.31	54.75 ± 1.82	78.58 ± 1.39
WH-MLP	68.93 ± 0.73	64.23 ± 0.75	80.29 ± 1.16
Multiple classifier	69.43 ± 0.54	64.77 ± 0.90	81.48 ± 0.83

than on measurements, thus resulting more insensitive to measurement errors and human factors; (2) invariance to roto-translations in the feature extraction process; (3) statistical robustness of the proposed ANN-based probabilistic models.

Non-metrical morphology depends on the experience of the anthropologist and is based on the perceptive analysis of the complex shape of the skull. This paradigm of analysis is not limited to the neurocranial profile but also takes account of many other parameters as the dimension of the skull, the zygomatic and temporal morphology, the inion shape and prominence, and so on. In the case of metric morphometry, on the other hand, the perceptive analysis of the global shape of the skull is substituted by a number of measures. There are, however, some limitations due to the grade of skill of the operator or to the fuzziness of some cranial landmarks.

In our case, the classifier system is based exclusively on the morphologic variants of a curve that represents the lateral profile of the calvarium, analyzed by a system specifically and exclusively trained on that profile. It is very likely that by adding another or more non-metric parameters of the skull, such as the lateral profile of the jaw or the profile of the mastoid process, the results could further improve.

The low performance of the anterior and posterior third of the profile, even though they *contain* information about the profile of the frontal and the occipital bone, are not simple to explain. More likely the contribution of the parietal profile, especially in the posterior third, could justify this low performance. However, the purpose of the comparison between the whole profile and the other two was simply to confirm the better significance of the entire profile with respect to its parts.

Regarding the proposed ANN approach, a significant contribution lies in the technique for combining two ANN architectures that model probabilistic quantities that are intrinsically different in nature, and that complement each other. Such a mixture of ANNs yields improved decision boundaries between the classes (namely, male and female) in the feature space. Accordingly, the overall recognition accuracies obtained in the experiments turn out to be higher than those yielded by traditional statistical estimation approaches (LDA, Parzen Window, k-NN). As a corollary, the results obtained allow us to confirm that the overall shape of the cranial vault is significantly dimorphic.

Conclusions

The novel approach described in this paper, based on the automatic modeling of the shape of the calvarium instead of the usual analysis of metric data, suggests a way to realize fully automated, statistically robust tools for physical and forensic anthropology based on shape recognition and analysis. In this case, only the shape of lateral profile of the calvarium was analyzed, but the same approach could be applied on other bones and to multiple profiles. The application of the system to other bones may be expected to improve the sensibility of the methodology even further.

From the machine learning point of view, it is worth observing that the highest recognition accuracy was obtained with a combination of a supervised and an unsupervised technique.

It is important to observe that although the present approach relies on cranial shape sex-related variations, it does not provide us with any information about the nature of these variations. It is, however, useful in order to provide the anthropologist with an automatic systems for *assisted diagnosis*.

Compliance with ethical standards

Conflict of interests The authors declare that they have no conflict of interest.

Appendix: The algorithm for probability density estimation

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a collection of n CT-scans, thought of as d-dimensional random vectors and assumed to be independently and identically drawn from an unknown pdf $p(.)$. Also, let $\varphi(.)$ be a proper kernel function (e.g., a standard Gaussian), and let the corresponding bandwidth h_1 be any positive real number (to be fixed empirically) [2]. An unbiased estimate $\tilde{p}(.)$ of $p(.)$ via MLP is proposed, according to the following unsupervised algorithm (expressed in pseudo-code):

```
Input: $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $h_1$.
Output: $\tilde{p}(.)$ /*the MLP estimate of $p(.)$ */
1. Let $h_{n-1} = h_1/\sqrt{(n-1)}$
2. Let $V_{n-1} = h_{n-1}^d$
3. For i=1 to n do /* loop over $\mathcal{T}$ */
3.1        Let $\mathcal{T}_i = \mathcal{T} \setminus \{\mathbf{x}_i\}$
3.2        Let $y_i = \frac{1}{n-1} \sum_{\mathbf{x} \in \mathcal{T}_i} \frac{1}{V_{n-1}} \varphi(\frac{\mathbf{x}_i - \mathbf{x}}{h_{n-1}})$ /* target output */
4. Let $\mathcal{S} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, n\}$ /* supervised training set */
5. Train the MLP via BP over $\mathcal{S}$
6. Let $\tilde{p}(.)$ be the function computed by the MLP
7. Return $\tilde{p}(.)$
```

where $\frac{1}{n-1} \sum_{\mathbf{x} \in \mathcal{T}_i} \frac{1}{V_{n-1}} \varphi\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_{n-1}}\right)$ is the Parzen kernel expansion of $p(.)$ over the $n-1$ feature vectors in $\mathcal{T}_i$, V_{n-1} being the corresponding volume of the window function [6].

References

1. Acsádi G, Nemeskéri J (1970) History of human life span and mortality. Akadémiai Kiadó
2. Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press
3. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8:679–698
4. Chidiac JJ, Shofer FS, Al-Kutoubi A, Laster LL, Ghafari J (2002) Comparison of CT scanograms and cephalometric radiographs in craniofacial imaging. Orthod Craniofac Res 5:104–113
5. Davies ER (2004) Machine vision: theory, algorithms, practicalities. Elsevier
6. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York
7. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York
8. Gybenko G (1989) Approximation by superposition of sigmoidal functions. Math Control Signals Syst 2:303–314
9. Hsiao TH, Chang HP, Liu KM (1996) Sex determination by discriminant function analysis of lateral radiographic cephalometry. J Forensic Sci 41:792–795
10. Hsiao T-H, Tsai S-M, Chou S-T, Pan J-Y, Tseng Y-C, Chang H-P, Chen H-S (2010) Sex determination using discriminant function analysis in children and adolescents: a lateral cephalometric study. Int J Legal Med 124:155–160
11. Inoue M (1990) Fourier analysis of the forehead shape of skull and sex determination by use of computer. Forensic Sci Int 47:101–112
12. Lee FCC, Noar JH, Evans RD (2011) Evaluation of the CT scanogram for assessment of craniofacial morphology. Angle Orthod 81:17–25
13. Luo L, Wang M, Tian Y, Duan F, Wu Z, Zhou M, Rozenholc Y (2013) Automatic sex determination of skulls based on a statistical shape model. Comput Math Methods Med 2013:251628:1–251628:6
14. Nakhaeizadeh S, Dror IE, Morgan RM (2014) Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. Sci Justice 54:208–214
15. Novotny V, Iscan MY, Loth SR (1993) Morphologic and osteometric assessment of age, sex, and race from the skull. Forensic Analysis of the Skull:71–88
16. Ogawa Y, Imaizumi K, Miyasaka S, Yoshino M (2013) Discriminant functions for sex estimation of modern Japanese skulls. J Forensic Legal Med 20:234–238
17. Ramsthaler F, Kettner M, Gehl A, Verhoff MA (2010) Digital forensic osteology: morphological sexing of skeletal remains using volume-rendered cranial CT scans. Forensic Sci Int 195:148–152
18. Richard MD, Lippmann RP (1991) Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Comput 3:461–483
19. Rogers TL (2005) Determining the sex of human remains through cranial morphology. J Forensic Sci 50:493–500
20. Rösing FW, Graw M, Marré B, Ritz-Timme S, Rothschild MA, Rötzscher K, Schmeling A, Schroder I, Geserick G (2007) Recommendations for the forensic diagnosis of sex and age from skeletons. HOMO-Journal of Comparative Human Biology 58:75–89
21. Spradley MK, Jantz RL (2011) Sex estimation in forensic anthropology: skull versus postcranial elements. J Forensic Sci 56:289–296
22. Trentin E (2006) Simple and effective connectionist nonparametric estimation of probability density functions. ANNPR (Artificial Neural Networks in Pattern Recognition):1–10
23. Van Otterloo PJ (1991) A contour-oriented approach to shape analysis. Prentice Hall International Ltd., UK
24. Veyre-Goulet SA, Mercier C, Robin O, Guerin C (2008) Recent human sexual dimorphism study using cephalometric plots on lateral teleradiography and discriminant function analysis. J Forensic Sci 53:786–789
25. Walrath DE, Turner P, Bruzek J (2004) Reliability test of the visual assessment of cranial traits for sex determination. Am J Phys Anthropol 125:132–137
26. Zhang D, Lu G (2001) A comparative study on shape retrieval using Fourier descriptors with different shape signatures. In: Proceedings of the international conference on intelligent multimedia and distance education (ICIMADE01)
27. Doi K (2005) Current status and future potential of computer-aided diagnosis in medical imaging. Br J Radiol 78:s3–s19