IET Biometrics

Research Article

IET Journals
The Institution of Engineering and Technology

Automatic ear detection and feature extraction using Geometric Morphometrics and convolutional neural networks

ISSN 2047-4938
Received on 10th January 2016
Revised 15th November 2016
Accepted on 6th December 2016
doi: 10.1049/iet-bmt.2016.0002
www.ietdl.org

Celia Cintas[1] ✉, Mirsha Quinto-Sánchez[2], Victor Acuña[3], Carolina Paschetta[1], Soledad de Azevedo[1], Caio Cesar Silva de Cerqueira[4], Virginia Ramallo[1], Carla Gallo[5], Giovanni Poletti[5], Maria Catira Bortolini[6], Samuel Canizales-Quinteros[7], Francisco Rothhammer[8], Gabriel Bedoya[9], Andres Ruiz-Linares[10,11], Rolando Gonzalez-José[1], Claudio Delrieux[12]

[1]*Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas, Puerto Madryn, Argentina*
[2]*Ciencia Forense, Facultad de Medicina, Universidad Nacional Autónoma de México, México*
[3]*Department of Genetics, Evolution and Environment, and UCL Genetics Institute, University College London, London, UK*
[4]*Superintendência da Polícia Técnico-Científica do Estado de São Paulo, Ourinhos-SP, Brazil*
[5]*Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Perú*
[6]*Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil*
[7]*Facultad de Química, UNAM, Mexico City, México*
[8]*Instituto de Alta Investigación Universidad de Tarapacá, Arica, Chile*
[9]*Universidad de Antioquia, Medellín, Colombia*
[10]*MOE Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai, China*
[11]*Aix Marseille Univ, CNRS, EFS, ADES, Marseille, France*
[12]*Depto. de Ing. Eléctrica y Computadoras, Universidad Nacional del Sur, and Consejo Nacional de Investigaciones Científicas y Técnicas, Bahía Blanca, Argentina*
✉ *E-mail: cintas.celia@gmail.com*

Abstract: Accurate gathering of phenotypic information is a key aspect in several subject matters, including biometrics, biomedical analysis, forensics, and many other. Automatic identification of anatomical structures of biometric interest, such as fingerprints, iris patterns, or facial traits, are extensively used in applications like access control and anthropological research, all having in common the drawback of requiring intrusive means for acquiring the required information. In this regard, the ear structure has multiple advantages. Not only the ear's biometric markers can be easily captured from the distance with non intrusive methods, but also they experiment almost no changes over time, and are not influenced by facial expressions. Here we present a new method based on Geometric Morphometrics and Deep Learning for automatic ear detection and feature extraction in the form of landmarks. A convolutional neural network was trained with a set of manually landmarked examples. The network is able to provide morphometric landmarks on ears' images automatically, with a performance that matches human landmarking. The feasibility of using ear landmarks as feature vectors opens a novel spectrum of biometrics applications.

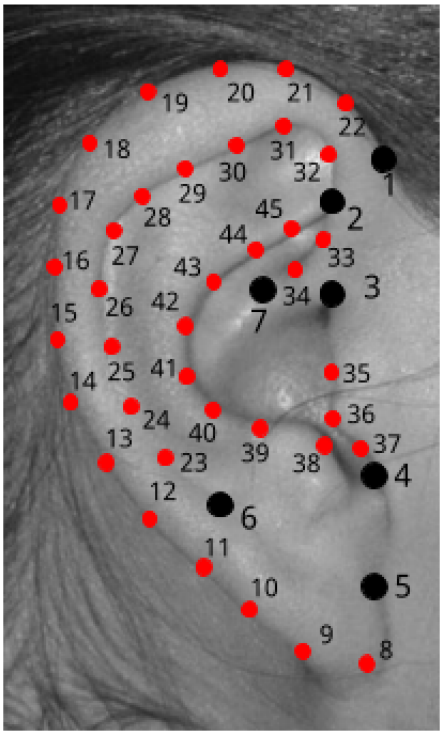
1 Introduction

Several kinds of data able to capture aspects of the human facial morphology may provide valuable information in the field of physical anthropology and biometrics. Facial phenotypes are at the foundation of studies on the genetic basis of many characters that recently gathered significant attention, including traits of biomedical and forensic interest, facial identification, perception analyses, to mention just a few [1–3]. In particular, ears have several advantages over other biometric structures such as fingerprints, iris patterns, or facial recognition [4]. For instance, they can be easily captured from the distance with non-intrusive methods. Also, the ear structure has less variation over time and is not much influenced by changes in facial expressions [5, 6]. However, there are issues such as potential occlusion by hair and earrings, which are almost certain to happen in images or video taken in the open.

These and other advantages are making research in ear detection and feature extraction remarkably active. Most of the proposals in the literature, however, try to take advantage of the ear's specific shape frequently using engineered features (in this context, detecting specific geometric configurations like the occurrence of certain characteristic edges, curvature dispositions, or frequency patterns in the ear using image processing techniques). In general, these procedures are unstable under homographies or change in luminance conditions, and also require several special case considerations. In this way, the ears' biometric properties are not fully leveraged, and thus these proposals are not robust against acquisition uncertainties (e.g. illumination, camera position, resolution etc.).

A much less explored strategy for ear detection and recognition is to represent ears' shape and phenotypic attributes in the form of landmark coordinates. In particular, landmarking based on Geometric Morphometrics (GM) provide a wide-spectrum and robust methodology for shape analysis and evaluation [7]. Manual landmarking, however, is not feasible for a massive sample, since it takes considerable supervised time, increases the likelihood of operation mistakes due to operator visual fatigue, intra- and inter-observer error, and is prone to distractions or confusions during the landmarking sequence. Automatic 2D or 3D landmark acquisition appears to be a promising venue to explore since it may overcome both limitations (the lack of robustness in most ear detection and recognition proposals, and difficulties associated to manual landmarking).

This work introduces a flexible and versatile method for automatic detection and selection of 2D landmarks, aimed to improve the capture of ears' form and shape phenotypic attributes. Even though the main intended use of this method is on population



Number	Name
1	Otobasion superiorious
2	Concha superiorious
3	Tragus superiorious
4	Intertragic incisure
5	Otobasion inferiorous
6	Helix basal border
7	Crus Helix
8 to 45	Semi landmarks

Fig. 1 *Landmark and semi-landmark configuration and anatomical description [16, 17]*

and quantitative genomic, biomedical, or forensic studies based on 2D data, it is easily adaptable to be useful in other contexts, like biometric identification. The main idea consists on designing and training a convolutional neural network with a set of manually landmarked examples, which in turn are represented as a uniform feature vector (resampled and normalised region of interest (ROI)). In this way, the trained network is able to provide for adequate morphometric landmarks on ears' images taken in the open.

2 Related work

This section briefly summarises the state of the art in automatic ear detection and feature extraction in 2D. Basically, all ear detection approaches rely on shape properties of the external ear's morphology, like the occurrence of certain characteristic edges, curvature dispositions, or frequency patterns. A more thorough description of current advances in ear detection, feature extraction and biometric recognition methods can be found in [8].

The French criminologist Alphonse Bertillon was the first to recognise the biometric potential of human ears. Empirical evidence supporting the ear's uniqueness was later provided in studies by Iannareli [9]. As mentioned above, the human ears (the *pinna*) present some advantages over other biological features for biometric identification purposes. For instance, the acquisition is less intrusive with respect to iris or fingerprint information capture, the ears' features do not vary over time, and are not susceptible to expression variation.

For these and other reasons, a significant effort was recently devoted to ear detection and feature extraction methods. Among the most widespread ideas, the use of shape models appears to be extensively used. Shape models aim to recognise specific distributions of shape indices that are characteristic to the object under study, in this case the ear's surface. For instance, Chen and Bhanu [10] propose to detect image regions with large local curvatures with a technique they call step edge magnitude. Then, template matching is performed with typical shapes of the outer helix and anti-helix. Later, in [11] the number of possible ear candidates was narrowed by detecting skin regions first before the helix template matching is applied, also reducing spurious detections. This method, however, by its very nature is not robust under homographies, making it unsuitable for most applications where a careful and calibrated acquisition may not be performed.

Following a similar shape-based approach, Attarchi *et al.* [12] use contour lines for ear detection. Their proposal locates first the outer contour of the ear using a search method that finds the longest connected edge in the ROI. Once located, this contour can be used to define a triangle formed by the outermost points in the top, bottom and left positions of the contour. Finally, geometric properties of this triangle, for instance the barycentre, can be used as a reference point for image alignment. Although less prone to break under homographies, this method still requires noise-free and white-balanced images to perform adequately.

Another method, related to edge detection properties, was proposed by Ansari and Gupta [13]. First, they apply an edge detector in which the edges are marked as convex and concave segments, since the most likely candidates for the ear's outer contour are convex edges. After that, the algorithm connects the contour segments and selects the figure enclosing the largest area for being the outer ear contour. Like other akin tracking algorithms, several special cases must be accounted for, thus leading to very complex algorithms.

In a similar vein, Prakash and Gupta [14] combine skin segmentation and hierarchy edges. After being detected, the edges located in the skin region are decomposed into edge segments. An edge connectivity graph is constructed, integrating all these edge segments. The connectivity graph is finally used to compute the convex hull of the set of edge segments, which encloses the ear's outer shape. Also significant is the proposal of Yan and Bowyer [15], who developed an ear detection method which starts by locating the concha (an anatomic part of the ear, see Fig. 1), which is set as the initial shape for an active contour used for determining the ear's outer boundary. Pflug and Busch [18] use a combination of depth images and texture. Their method starts with a preprocessing step, where edges and shapes are extracted from the texture and the depth image, and edges and shapes are fused together in the image domain. In the next step, the components are combined with each other to find ear candidates and rank them according to a computed score. Finally, the enclosing rectangle of the best ear candidate is returned as the ear region.

Liu *et al.* [19] introduce the ear-parotic face angle of the person as a novel 3D feature in ear images. The ear-parotic face angle feature is defined as the angle between the normal vector of the ear-plane and the normal vector of the parotic face-plane. Sibai *et al.* [20] define a seven-element ear feature set, manually extracted and design and train a feed-forward artificial neural network to recognise a human ear. Like the other methods already mentioned, the main disadvantage of these shape-model approaches is the fact that they require specifically engineered features, which makes them less flexible or adaptable to other landmarking problems, and also fragile under homographies and luminance changes.

Instead of focusing on the unique geometric features of the ear, a different approach regards the ear detection problem as an instance of a pattern recognition problem. In this approach, the first stage uses image processing techniques to extract features present in the image, followed by a second stage in which pattern recognition techniques are applied over the feature set to perform detection and identification tasks. This approach is in general more robust under homographies and luminance changes, depending on the feature space used for the ear representation in the first stage. Also, recent spectacular advances in pattern recognition techniques

can be adopted in the second stage. Among the proposals based on pattern recognition approaches, we can mention Abaza *et al.* [21] and Islam *et al.* [22], which use weak classifiers based on Haar-wavelets over regions of the image to find correlation with previously learned patterns. These weak classifiers are then combined with a standard AdaBoost procedure for ear localisation. Yuan *et al.* [23] propose a dictionary-based sparse representation and classification scheme, intended to work with partially occluded ear imagery. An identity occlusion dictionary encodes occluded parts in the source image to perform ear recognition. A non-negative dictionary that includes a Gabor feature-set extracted from ear images improves the sparseness of the coding representation, thus circumventing the expense of a conventional occlusion dictionary.

In [24], Kumar and Chan take advantage of the sparse representation of the finite (discrete) Radon transform-based local orientation information. The neighbourhood relationship of grey-levels in the normalised ear images is encoded as the dominant grey-level feature orientations in a local region using local Radon transform. In a later contribution, Kumar and Wu [25] present a pipeline for feature extraction and ear recognition based on morphological operators and Fourier descriptors. In [26], the authors develop an approach that encodes reliable phase information using 2D quadrature filtering. They extensively evaluated both quaternionic and monogenic quadrature filters and develop a new quaternionic-code-based approach for the ear identification. Their experimental results suggest that the performance from the quaternionic quadrature filters and the monogenic quadrature filters consistently outperform 1D log-Gabor filter-based approach.

These proposals based on pattern recognition techniques are more recent and tend to outperform shape-model methods. However, the proposed feature spaces in general do not explicitly take into account the specific phenotypic information present in the ears. In particular, as will be presented in the following section, using an anatomically inspired feature space (in our case, morphometric landmarks) greatly improves the ear detection performance. In Table 1, we summarise all the methods reviewed in this section, classified by method and data type.

Table 1 Classification table from data and methods

Reference	Summary	Method type	Data type
Chen and Bhanu [10]	template matching with shape index histograms	shape-model	2D
Attarchi *et al.* [12]	edge detection and line tracing	shape-model	2D
Chen and Bhanu. [11]	helix shape model	shape-model	3D
Ansari and Gupta. [13]	edge detection and curvature estimation	shape-model	2D
Prakash and Gupta. [14]	skin colour and graph matching	shape-model	2D
Yan and Bowyer. [15]	ICP using model points	shape-model	2D and 3D
Pflug and Busch [18]	feature-level fusion and context information	shape-model	2D and 3D
Liu *et al.* [19]	ear-parotic face angle	shape-model	3D
Abaza *et al.* [21]	cascaded adaboost	pattern recognition	2D
Islam *et al.* [22]	adaboost	pattern recognition	2D
Yuan *et al.* [23]	non-negative dictionary-based sparse representation	pattern recognition	2D
Kumar and Chan [24]	radon transform-based local orientation information	pattern recognition	2D
Kumar and Wu [25]	phase encoding with log Gabor filters	pattern recognition	2D
Presented method	ConvNets and GM	pattern recognition	2D

3 Methods and implementation

Given the above-mentioned limitations of the current proposals in ear detection and feature extraction algorithms, we propose a new method, based on two well-established methodologies, GM and Deep Learning algorithms. A set of 2735 manually landmarked images, each with 45 interest points (landmarks and semi-landmarks), was obtained to train a convolutional neural network, using specific learning techniques to achieve a high generalisation rate and to avoid overfitting. In this section, we describe the complete processing pipeline, and for each of the processing steps we provide a brief review of the underlying formalisms to make this presentation self-contained.

3.1 Geometric Morphometrics

GM provides a set of methods for the quantitative analysis of the size and shape of objects. GM is widely used in the study of biological organisms [27], specially humans [28]. Methods in GM propose to quantify the shape of each specimen according to the location in space of a set of 2D or 3D reference points or *landmarks* that are homologous across individuals. Among the diverse traits used to represent craniofacial morphology, size and shape in the form of landmark coordinates are usually preferred because the methodologies underlying GM provide a versatile and wide-spectrum set of analyses aimed to evaluate intra- and inter-group variation patterns, integration, modularity, and multivariate regression of shape on several independent variables [7].

The human *pinna* is made up of a piece of cartilage covered with skin and attached to the skull by ligaments, muscles, and fibrous tissue. This cartilage does not extend into the ear lobe, which consists mostly of areolar and adipose tissue. There is a wide non-pathological variation between humans in the *pinna* shape and size, and this variation has been reported to be influenced by age, sex, ethnicity and recently reported genetic factors [5, 29–31]. The *pinna* shape variation was examined using seven landmarks and 38 semi-landmarks. The specific configuration and anatomical descriptions are shown in Fig. 1.

Then, the size and shape are split by means of the superimposing Procrustes analysis, which removes the effects of scaling, translation and rotation of the original configurations and allows quantifying the multidimensional deviation from a preconfigured reference specimen (usually the average of all the configurations of the sample) [32]. In addition to landmark configuration, several other magnitudes can be obtained, such as among-landmark and semi-landmark Euclidean distances [16], and angles of anatomical interest. These and other derived measurements can be further used to create a robust feature vector for identification purposes.

3.2 Deep learning and convolutional neural networks

In recent years, the computer vision literature has witnessed many research efforts in descriptor engineering. A sought-for advantage of these descriptors, when applied to recognition purposes, is that they require the use the same operator to all locations in the image. In this way, the design of workflows for specific recognition purposes is greatly simplified. Moreover, and as more data becomes available, learning based methods have started to outperform engineered features, because they can discover and optimise features without supervision for the specific task at hand [33–37].

Deep learning allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction, and to discover accurate representations autonomously from the data itself [36]. Each subsequent layer extracts a progressively more abstract representation of the input data and builds a new representation from the previous layer. Layers higher in the hierarchy amplify

aspects of the input that are important for discrimination, and which might have been overlooked in supervised analysis. In an image, for example, this could be the occurrence of edges at particular orientations. The subsequent layer would detect specific disposition of edges, and the next layers would probably identify parts of objects.

Convolutional neural networks (*ConvNets*) [38, 39] constitute the state of the art in many computer vision problems, since they were shown to be very effective for large-scale image classification [37, 40, 41]. Their outstanding performance is based in four core concepts: local connections, shared weights, pooling, and the use of several layers [36]. However, since the amount of learnable parameters in these nets is huge, special care must be taken to avoid overfitting (i.e. the network may just memorise the examples, without generalisation).

Consider a regular neural network with N layers. The network's input and output are represented, respectively, by vectors $\boldsymbol{X}_0$ and $\boldsymbol{X}_N$, where the vector $\boldsymbol{X}_{n-1}$ is the input to layer n (with $n = 1, \ldots, N$). If $\boldsymbol{W}_n$ is a matrix of weights and $\boldsymbol{b}_n$ a vector of biases, then the output of layer, $\boldsymbol{X}_n$, can be represented as the following vector:

$$\boldsymbol{X}_n = f(\boldsymbol{W}_n \boldsymbol{X}_{n-1} + \boldsymbol{b}_n), \tag{1}$$

where f is the activation function – in our case, *linear rectification* $f(x) = \max(x, 0)$. Given a particular recognition problem, the training task consists on finding the optimal parameter set $\{\boldsymbol{W}_n, \boldsymbol{b}_n\}$ that minimises classification error. To determine how these parameters should be changed to reduce error, it is a standard practice to use the *gradient descent* algorithm. The definition of classification error depends on the data type of the output. For categorical (nominal) data it may be the proportion of misclassified items, for scalar values it can be RMS error between the actual and expected output and so on.

We define X_{true} as the expected output corresponding to the network input $\boldsymbol{X}_0$. Through the training stage, all the network parameters are optimised to make the output $\boldsymbol{X}_n$ approximate X_{true} as much as possible. The prediction error is denoted as $e(\boldsymbol{X}_N, X_{\text{true}})$. The gradient of $e(\boldsymbol{X}_N, X_{\text{true}})$ is then computed with respect to the model parameters $\{\boldsymbol{W}_n, \boldsymbol{b}_n\}$. The parameter values of each layer are then modified by repeatedly taking controlled steps in the direction opposite to the gradient:

$$\boldsymbol{W}_n \leftarrow \boldsymbol{W}_n - \eta \frac{\partial e(\boldsymbol{X}_N, X_{\text{true}})}{\partial \boldsymbol{W}_n} \tag{2}$$

and

$$\boldsymbol{b}_n \leftarrow \boldsymbol{b}_n - \eta \frac{\partial e(\boldsymbol{X}_N, \boldsymbol{X}_{\text{true}})}{\partial \boldsymbol{b}_n}, \tag{3}$$

where η is the *learning rate*, a hyperparameter controlling the stride towards convergence.

In ConvNets, the connectivity patterns between some of the layers are constrained in a way such to facilitate the processing of input data that comes in the form of multiple arrays, for example 2D arrays containing pixel intensities or 3D for video or volumetric images. Images commonly exhibit high correlation between values in a local group, forming distinctive local patterns that are easily detected. To take advantage of these properties, ConvNets contain two types of layers: *convolutional* and *pooling* layers.

A convolutional layer is parametrised by a set of learnable filters. The feature maps are taken as input and then a convolution is applied to each with the set of filters to produce a stack of output feature maps. This may be efficiently implemented by replacing the matrix-vector product $\boldsymbol{W}_n \boldsymbol{x}_{n-1}$ in (1) with a sum of convolutions [37]. The input of layer n can be unfolded as a set of $\boldsymbol{K}$ matrices $X_{n-1}^{(k)}$, with $k = 1, \ldots, K$. Each of these matrices represents a different input feature map. The output feature maps $X_n^{(l)}$ with $l = 1, \ldots, L$ are represented as follows:

$$X_n^{(l)} = f\left(\sum_{k=1}^{K} \boldsymbol{W}_n^{(k,l)} * X_{n-1}^{(k)} + b_n^{(l)}\right). \tag{4}$$

Here, $*$ represents the two-dimensional convolution operation. The matrices $\boldsymbol{W}_n^{(k,l)}$ represent the filters of layer n, and $b_n^{(l)}$ represents the bias for feature map l.

Note that a feature map $X_n^{(l)}$ is obtained by computing a sum of K convolutions with the feature maps of the previous layer. The bias $b_n^{(l)}$ can optionally be replaced by a matrix $\boldsymbol{B}_n^{(l)}$, so that each spatial position in the feature map has its own untied bias. By replacing the matrix product with a sum of convolutions, the connectivity of the layer is effectively restricted to take advantage of the input structure and to reduce the number of parameters. Each unit is only connected to a local subset of the units in the layer below, and each unit is replicated across the entire input [37]. This parameter reduction enables ConvNets to achieve better generalisation performance.

To reduce the dimensionality of the feature maps, a pooling layer is located between convolutional layers. Pooling layers eliminate non-maximal values by computing some aggregation function (typically the maximum or the mean) across small local regions of the input [42]. The main purpose of this pooling is to reduce the computational cost in the remaining layers, reducing the dimensionality of the feature maps and providing a form of translational invariance.

3.3 Dataset

The images and manual landmarking data belong to the CANDELA initiative (Consortium for the Analysis of the Diversity and Evolution of Latin Americans), an international multidisciplinary project including geneticists, anthropologists, statisticians, bioinformatics, and social-anthropologists interested on Latin American populations biodiversity and socio-cultural environment [43] (https://www.ucl.ac.uk/candela. The dataset is property of CANDELA and for privacy reasons cannot be made openly available.). CANDELA has a database containing 7500 individuals that were photographed following a protocol for taking standardised photographic data. The images consist of a lateral view of the head, with a 2136×3216 pixel resolution, taken with no specific illumination conditions, and without background removal. The provided dataset contains 2735 images, one image per individual, each with 45 landmarks and semi-landmarks provided by human operators. It was split into a training set with 2051 images (75%) and a validation set of 684 images (25% of the full dataset), both sets selected with a random permutation cross-validation iterator. The dataset selection was restricted to the availability of pairs (*image*, *landmarks*) from the CANDELA database, images without landmarks associated were excluded for the training and validation steps. The landmarks and semi-landmarks used for training and testing were digitised and processed manually using TPSDig and TPSUtil (http://life.bio.sunysb.edu/morph/). In some of the images, ears were partially occluded by hear, earrings were present, or the illumination and background were not uniform. These images were retained to test the robustness of our methodology against this kind of semantic noise. An example of an individual image taken following the CANDELA protocols can be seen in Fig. 2.

3.4 Preprocessing

To reduce unnecessary processing burden in the ConvNet, an initial rough ROI is priorly located around the candidate areas in the images. During training, the ROI was located using the landmark positions themselves. During automated landmarking, the ROI is found using the Viola and Jones [44] general object detection framework. The Haar cascade filter was trained with 133 positive and 667 negative regions (The Haarcascade file trained can be

Fig. 2 *Image from one of the authors (Mirsha Quinto-Sánchez) taken following the CANDELA protocols*

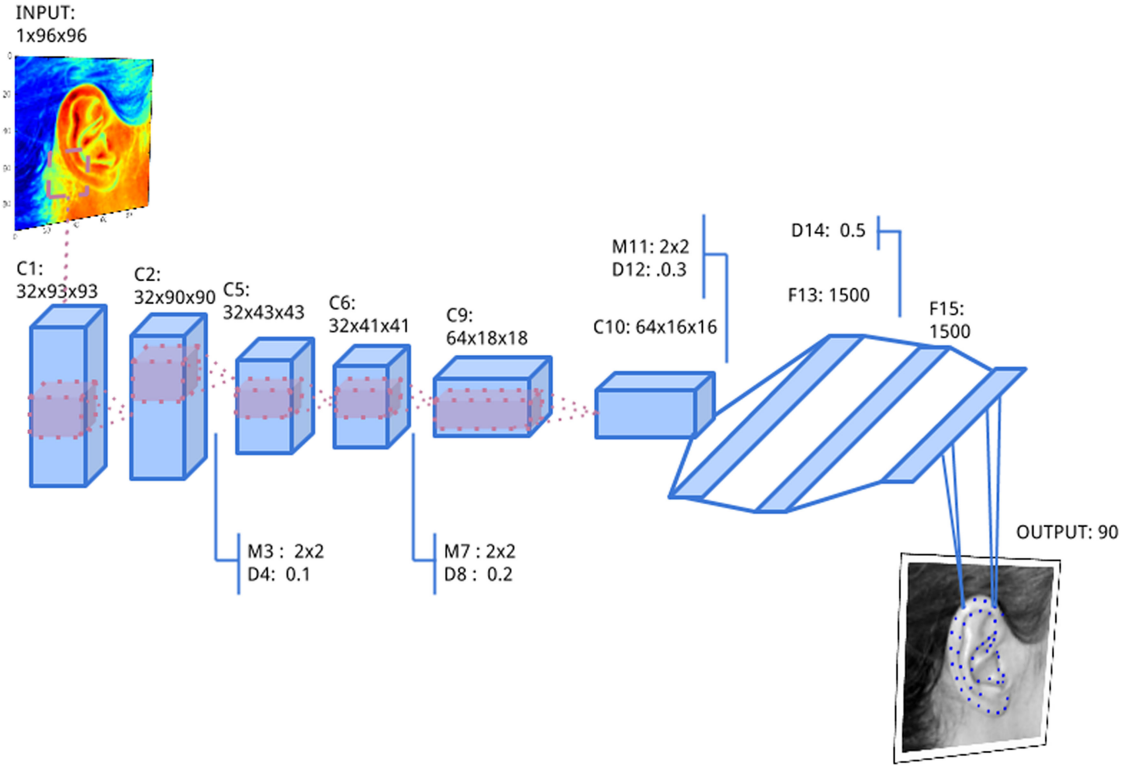


Fig. 3 *Outline of the best performing network structure*

downloaded from https://github.com/celiacintas/tests_landmarks/blob/master/files/cascade_lateral_ears.xml). As a validation set for this filter, 185 images from the CANDELA dataset were randomly selected. In 92.43% of the cases, the ROI was correctly found, in 1.62% of the cases the ROI was not found, and in 5.95% of the cases the ROI was wrongly placed. After the ROI is found, a histogram normalisation is performed, by means of which the range of brightness values within the ROI is stretched to cover most of the dynamic range. The histogram stretch parameters were programmed to black-out at most 2% of the pixels in the ROI, and to white-out at most 1% of the pixels. Finally, the ROI is resampled to a final size of 96×96 pixels, using bilinear downsampling.

3.5 ConvNets architecture and training

Three different ConvNet architectures were designed and trained for performing an automatic landmarking task, namely to detect and identify anatomic parts of the ear in image datasets. These architectures are different to each other in the number of convolutional layers, the filter sizes, and the learning rates. A single-channel profile ear image of size 96×96 pixels, with brightness scaled to [0, 1], is taken as input. In Fig. 3, the best performing architecture (*Arch0*) is shown. The underlying architecture consists of two convolution layers with square filters, followed with a *max* pooling and dropout layer. This structure is repeated three times to obtain features at different levels of abstraction, with different filter size, number of feature maps, and probability values. The convolutional layers C1, C2, C5, and C6 have 32 filters of size 4×4 and 3×3. Layers C9 and C10 have 64 filters of size 3×3. All *max* pooling layers are of size 2×2, and the probability values used for D4, D8, D12, and D14 are (resp.) 0.1, 0.2, 0.3, and 0.5.

After the feature extraction layers, the architecture contains two fully connected linear layers with 1500 units each (F13 and F15 in the diagram), and a dropout layer in between (D14). The output layer contains 90 output units (45 [x, y] pairs) for the predicted position of the landmarks and semi-landmarks. The implementation used Python and the Lasagne library [45] (The code is available at https://github.com/celiacintas/tests_landmarks/blob/master/testing_output_ears.ipynb.). This allows the use of GPU acceleration without considerable programming effort. The training of the network took roughly 25 h using NVIDIA GeForce GTX 590 cards. Once trained, the network can be deployed in conventional hardware and even in embedded systems.

3.6 Overfitting reduction

ConvNets usually have a huge number of learnable parameters, 8.622.970 in the case of our model. Due to the limited size of the training set, overfitting is almost certain to occur. The network will tend to memorise the training examples instead of finding abstractions therein, because it has enough memory to do so. This obviously will not generalise well to new data. Two primary ways to deal with overfitting were applied in our training:

Data augmentation: We artificially enlarged the dataset using label-preserving transformations. At random, some of the pictures and the associated landmarks were mirrored about the x-axis and added to the training set (see for instance Fig. 4).

Regularisation: The model complexity was penalised through the use of *dropouts* [46], which consists of setting to zero the output of each hidden neuron with a certain probability. This technique reduces complex co-adaptations of neurons, since during the

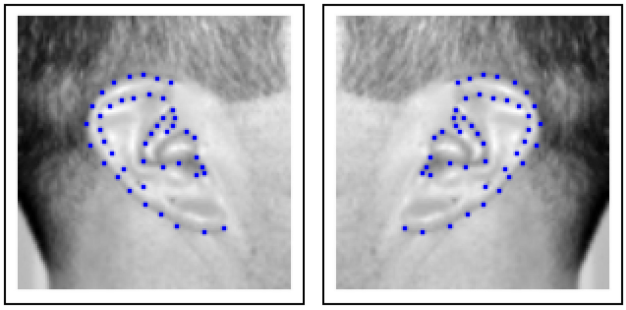


Fig. 4 *Image taken at random with the associated landmarks, mirrored about the x-axis*

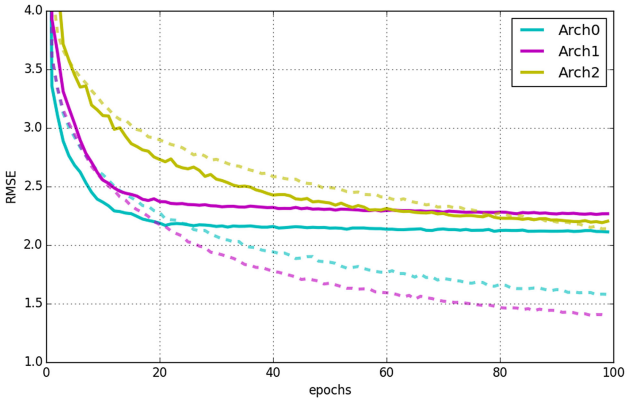


Fig. 5 *Learning curves for the ConvNets analysed in Table 1. The dashed lines represent RMSE on the training set, and the solid lines represent the validation error, for the three network configurations*

training phase a neuron cannot rely on the activation of other specific neurons [41].

4 Results and discussion

4.1 Performance assessment

Automatic landmark placement can be regarded as a regression problem. Therefore, to assess the quality of the final landmarking, the automated results were evaluated regarding the positioning errors with respect to manually annotated landmarks, which served as ground truth [47]. We evaluated the usual quality metrics for regression problems, in particular r^2, root mean square error (RMSE), explained variance (EV), and Pearson's correlation. The accuracy of the three implemented architectures can be seen in Table 2. Also in Table 3, the RMSE for each landmark is shown. The regression metrics were computed using *scikit-learn* [48].

Table 2 Performance of the three different ConvNet architectures

	Arch0	Arch1	Arch2
r^2	0.709	0.678	0.698
RMSE	2.296	2.415	2.338
EV	0.976	0.974	0.975
Pearson	0.988	0.987	0.988

Table 3 RMS error for each anatomical landmark

# Landmark	RMSE
1	1.8183
2	1.2216
3	1.08651
4	1.3291
5	2.4477
6	2.59746
7	1.17571

Also, the learning curves showing the training set error and the validation set error with the three different network configurations can be seen at Fig. 5. Fig. 9 shows the predicted landmarking over previously unseen images in the validation dataset. Note that even though some of these images are partially occluded by hair, the landmarking is still sound. The full test set landmarked by the ConvNet, the full structure, and an analysis of the net's behaviour can be seen in https://github.com/celiacintas/tests_landmarks/blob/master/testing_output_ears.ipynb. Also a subset of landmarked ears over public datasets can be seen in Figs. 6–9. Despite that our landmarking system was not trained with examples of rotated ears, it still works well with the AMI ear database that includes such cases.

To assess the performance of the trained network, a complete landmarking workflow was deployed in a conventional PC hardware (single core Intel i7-5500 2.40 GHz). In this conventional hardware, a typical automated landmarking requires 4.68 ms in average. Landmarking a batch of 684 images required 1.04 s. The landmarking performance is comparable or above the quality of an assisted procedure. Also a similar network was trained for full profile faces, without the ROI extraction preprocessing step. In average, the results were $r^2 = 0.884$, RMSE = 1.365, EV = 0.951, and *Pearson correlation* = 0.976. The automated landmarking results on whole head images were tested on a randomly chosen subset of the CVL public dataset can be seen in Fig. 10. The landmarking behaviour was also tested on less controlled images. A randomly chosen subset the original CANDELA 2136×3216 whole head images dataset, without controlled illumination or background removal, was automatically landmarked, and the results can be seen in Fig. 11. In both cases, the quality of the results is still comparable with the human-assisted landmarking. Even though for this preliminary results it was chosen to work within the resampled ears' ROI, it is worth mentioning that the whole experience may be repeated with different network configurations trained with whole images, within which the ears can be located.

4.2 People recognition

Even though the major purpose of this work is to show the combined potential of GM together with CNNs, we performed some preliminary recognition experiments, to assess as to whether the proposed workflow can be used for identification purposes. For this, we added an extremely randomised tree (ERT) as final classification stage in our workflow. ERTs are classification trees in which attribute and cut-point choices are partially or totally randomised on splitting a node during training [50]. In an extreme case, an ERT builds totally random classification trees whose structures are independent of the output values of the learning sample. The strength of the randomisation can be tuned to specific behaviours by an appropriate parameter choice.

The ERT was trained with the landmark configuration of 1458 individuals, each individual with four to six images (both ears). A total set of 8354 images where automatically landmarked with our previous *ConvNet* detailed in the Section 3.5, after which a generalised Procrustes fit was applied (see Section 3.1) to remove translation, scale and rotation effects in the landmarks. In Fig. 12, a subset of the training dataset can be seen. The training set of the ERT included 6683 feature vectors $\boldsymbol{v}_i$, each consisting of the 45 automatically generated landmarks ($\boldsymbol{v}_i = [x_0, y_0, \quad , x_{44}, y_{44}]$) and the target values t with a label associated to the person. The remaining 1671 samples were saved for testing. The recognition scores were remarkably high (in average, *precision* 0.95, *recall* 0.90, *f1-score* 0.91, and *adjusted rand score* 0.93). The confusion matrix over a subset of test data can be seen in Fig. 13. We performed stratified *K*-fold with 10 iterations and a test size of 20% of the sample, yielding an accuracy mean score of 0.9114 with SD = 0.0146. The accuracy of each fold can be seen in Table 4.

ERTs are also useful as a means for explaining away the relative weight or importance of each dimension in the feature space [51]. Given this, we analysed the relative contribution of each landmark

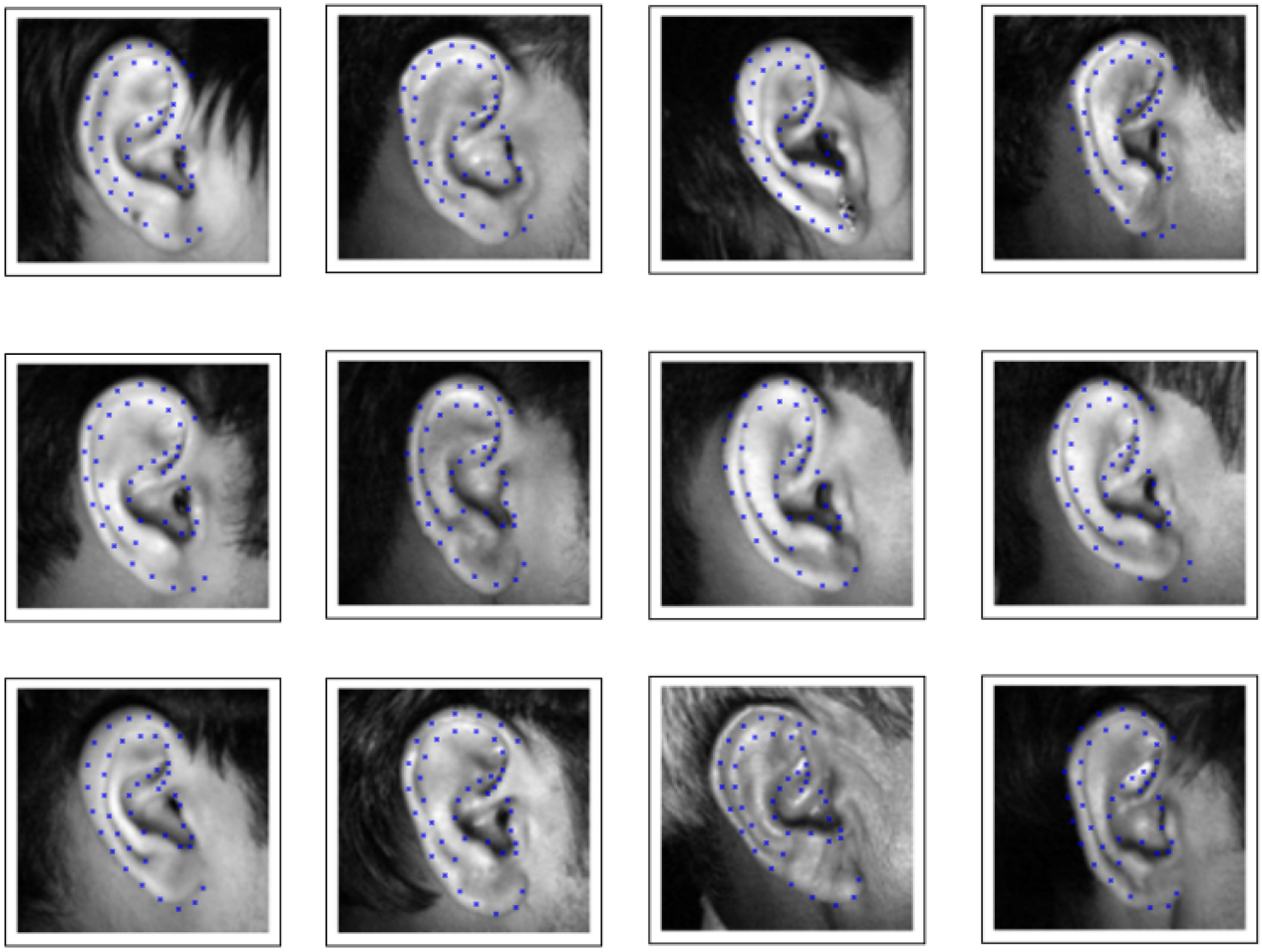


Fig. 6 *Results over images randomly chosen from the IIT Delhi ear database [25]*

coordinate in the final recognition. The ten most important features (landmark coordinates) are listed in Table 5). It is remarkable that all of them correspond to the inner structure of the ear, which suggests that this structure is more informative for differentiation that the external ear traits.

5 Conclusions and future work

We presented a novel method for ear detection and feature extraction. The method is based on GM and the use of ConvNet for automatic ear landmarking. After training the net with human-assisted landmarks over images of ears, the resulting algorithm is able to accurately position landmarks and semi-landmarks with a precision comparable to supervised landmarking. The quality of the automatic detection and feature extraction was tested using the automatically extracted landmarks for identification, yielding accuracies akin to other biometric methods. Even though that other ear detection methods in the literature may achieve a better performance, they require more rigid acquisition constraints. Instead, our method is fully automatic, with no supervised fine-tuning requirements, and able to perform in the open even with low-quality cameras. The fact that the method is based on biological landmarks makes it more robust than other engineered feature sets proposed in the literature, specifically under homographies, resampling, or illumination changes. The whole implementation was developed using open source tools, and the source code is publicly available. Thus, the model can be freely trained and used on consumer hardware.

Our proposal is general enough to perform with other physical anthropological features, such as the outline of specific anatomical or biological structures. Therefore it may be used for several other studies, including the genetic basis of traits of biomedical and forensic interest, biometrics, perception analysis, and many other topics. Regarding the former, a range of disorders affecting human *pinna* development have been described, occurring in isolation or as part of complex syndromes with multiple affected organs [52, 53]. For instance, we have recently reported genomic associations of seven genetic regions, including the Edar gene, with macroscopic categorical phenotypes in the external ear. Thus, further improvements in the capture of such a complex phenotype are needed to complement the understanding of its genetic and non-genetic basis.

Regarding biometric applications of unsupervised ear landmarking, the set of landmarks *per se* is a feature vector well known to be able to identify people. Therefore classifiers are under research that aimed to identify individuals through this *pinna* landmark configuration and derived measurements, such as relative distances and angles, to create a strong feature vector for identification. In a similar vein, the method is being tested with images of large marine animals, as an aid to perform identification in specific natural environments required in biological and ecological research activities.

Finally, using automatically generated ear landmarks for people identification produced promising results. Even though this was not the intended goal of this work, the preliminary results showed good accuracies. A remarkable finding is that the inner part of the ear appears to be much more important in people identification than the external outlines, a result first reported here. Therefore, future research in landmarking-based identification appears as a venue worth exploring.

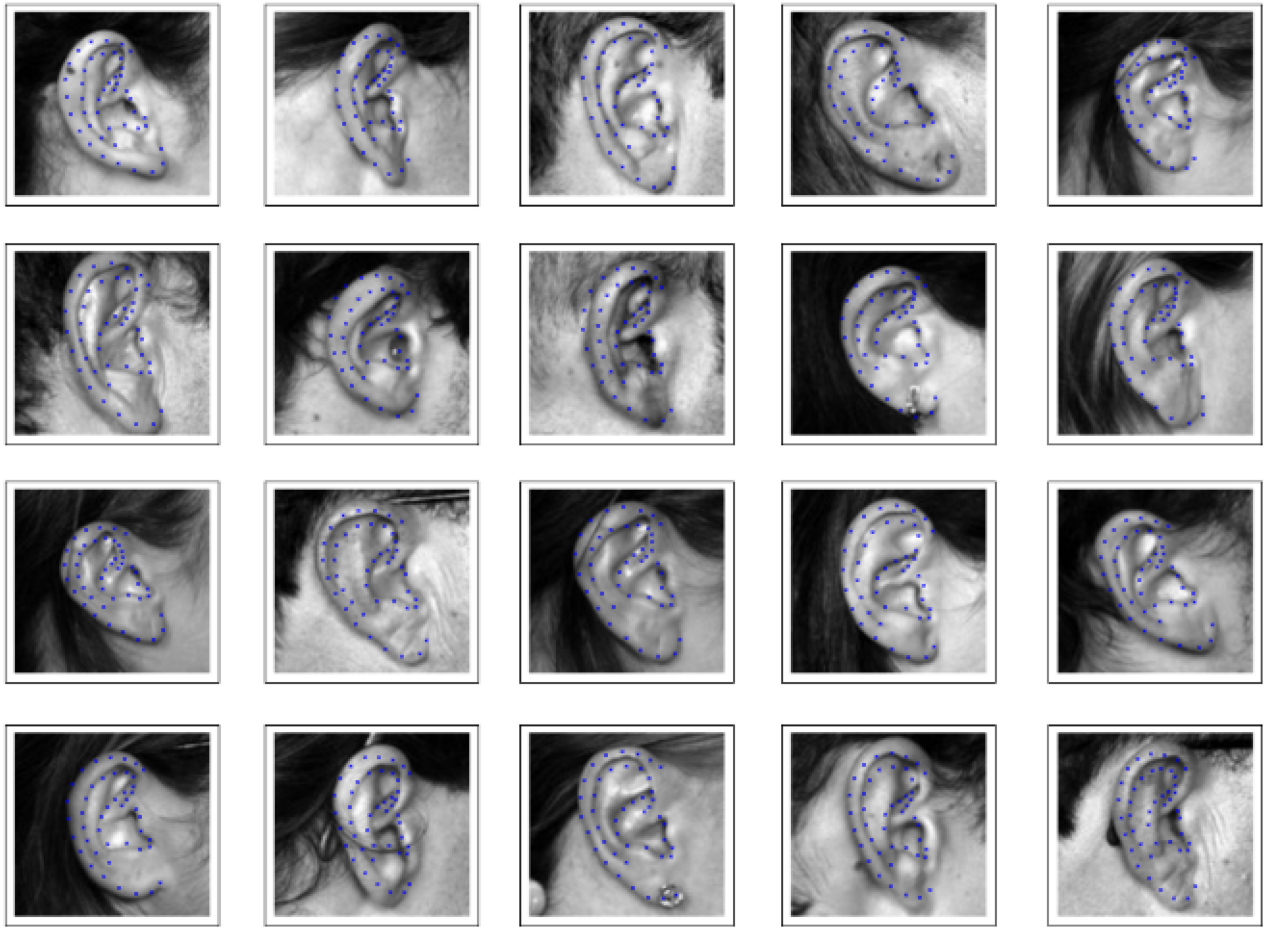


Fig. 7 *Results over images randomly chosen from the AMI ear database (www.ctim.es/research_works/ami_ear_database/)*

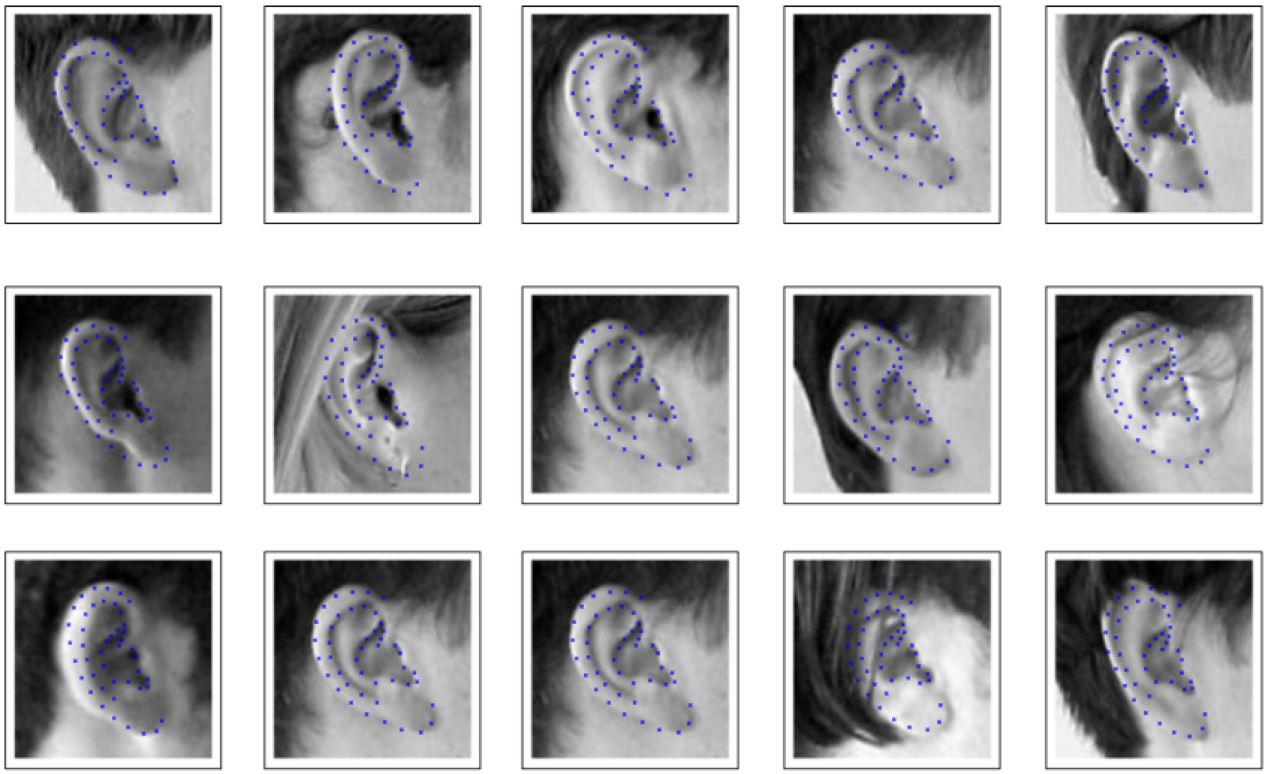


Fig. 8 *Results over images randomly chosen from the CVL Face Database http://www.lrv.fri.uni-lj.si/facedb.html [49]*

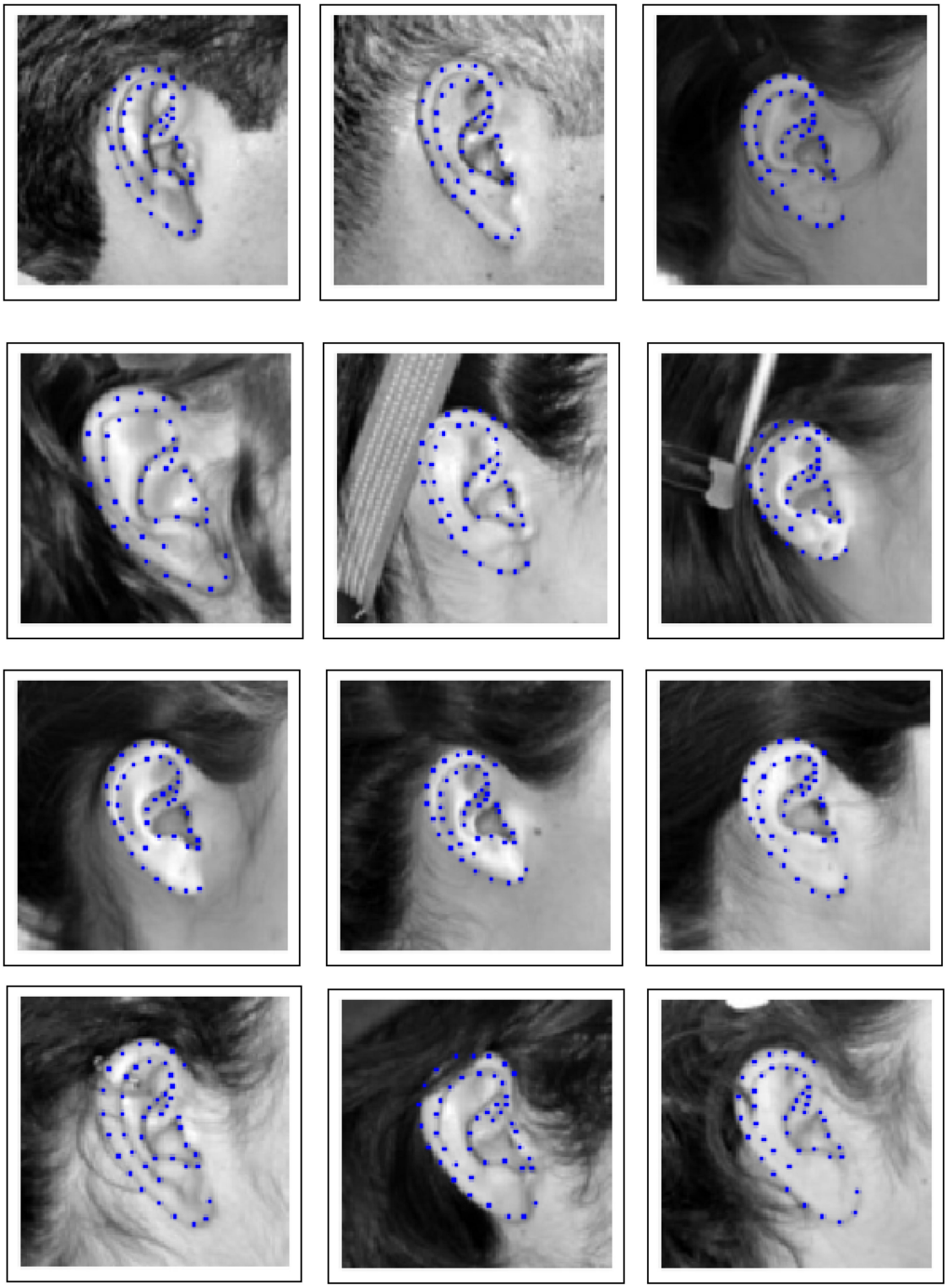


Fig. 9 *Results over unseen images from CANDELA database using the best performing network. Note that the method is able to withstand partial occlusions*

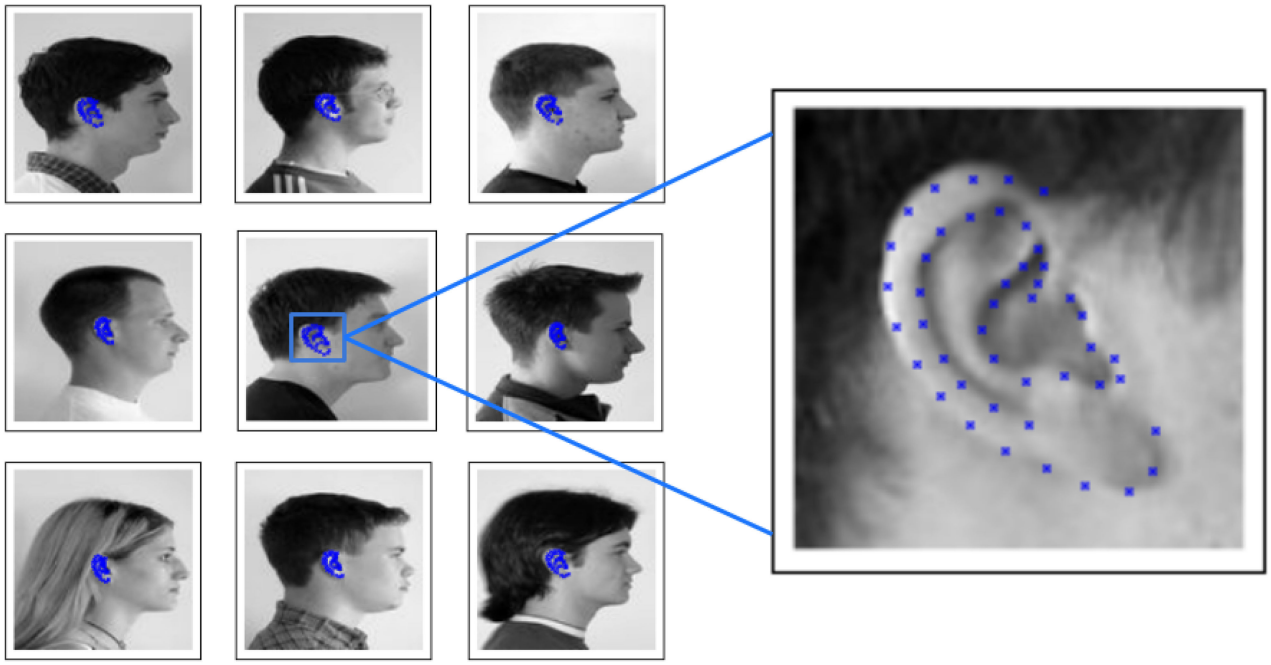


Fig. 10 *Results over images randomly chosen from CVL Face Database http://www.lrv.fri.uni-lj.si/facedb.html [49]*

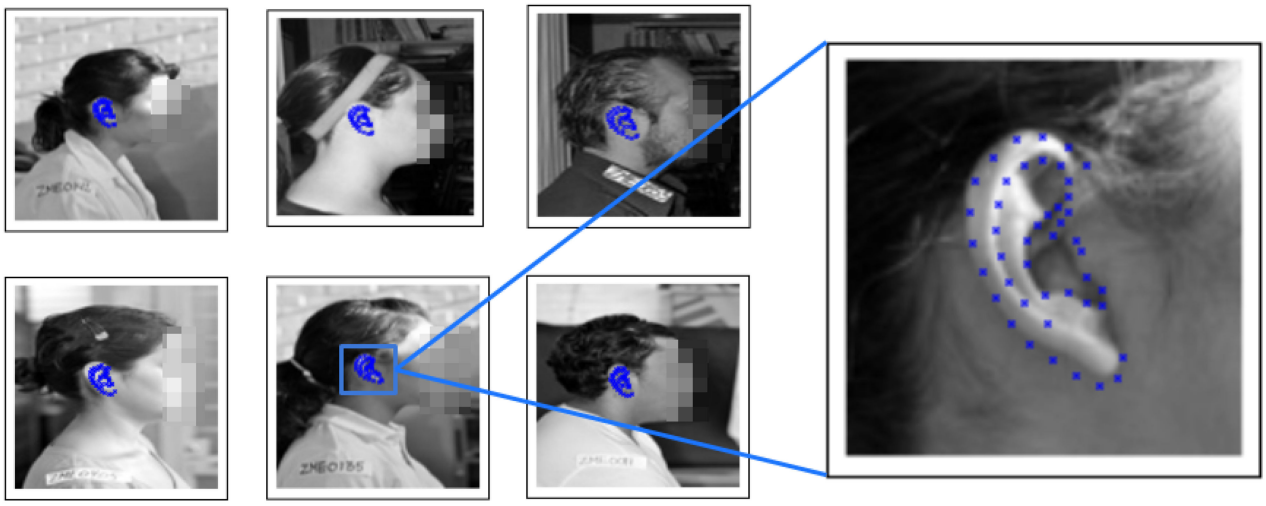


Fig. 11 *Results over randomly chosen images in the CANDELA dataset with uncontrolled illumination and background*

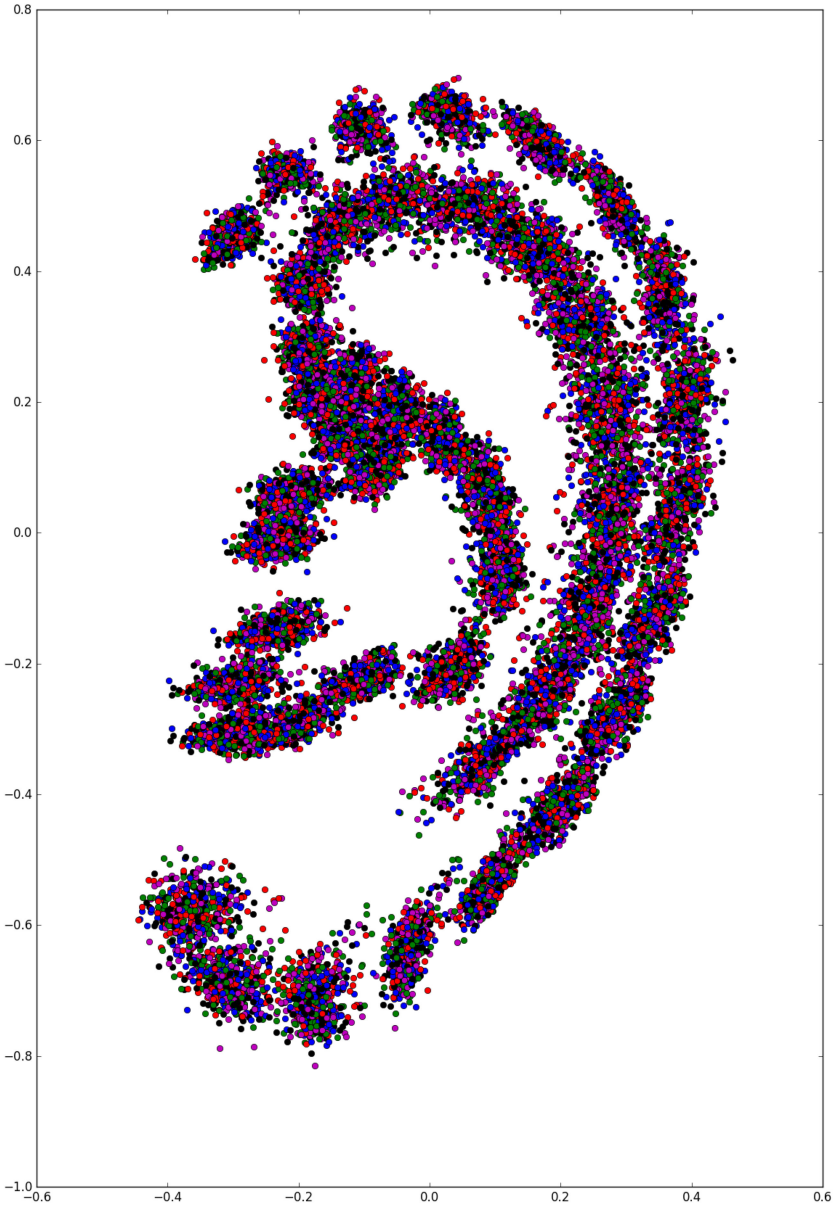


Fig. 12 *Landmark configuration of 500 images after applying generalised Procrustes fit used in training the ERT*

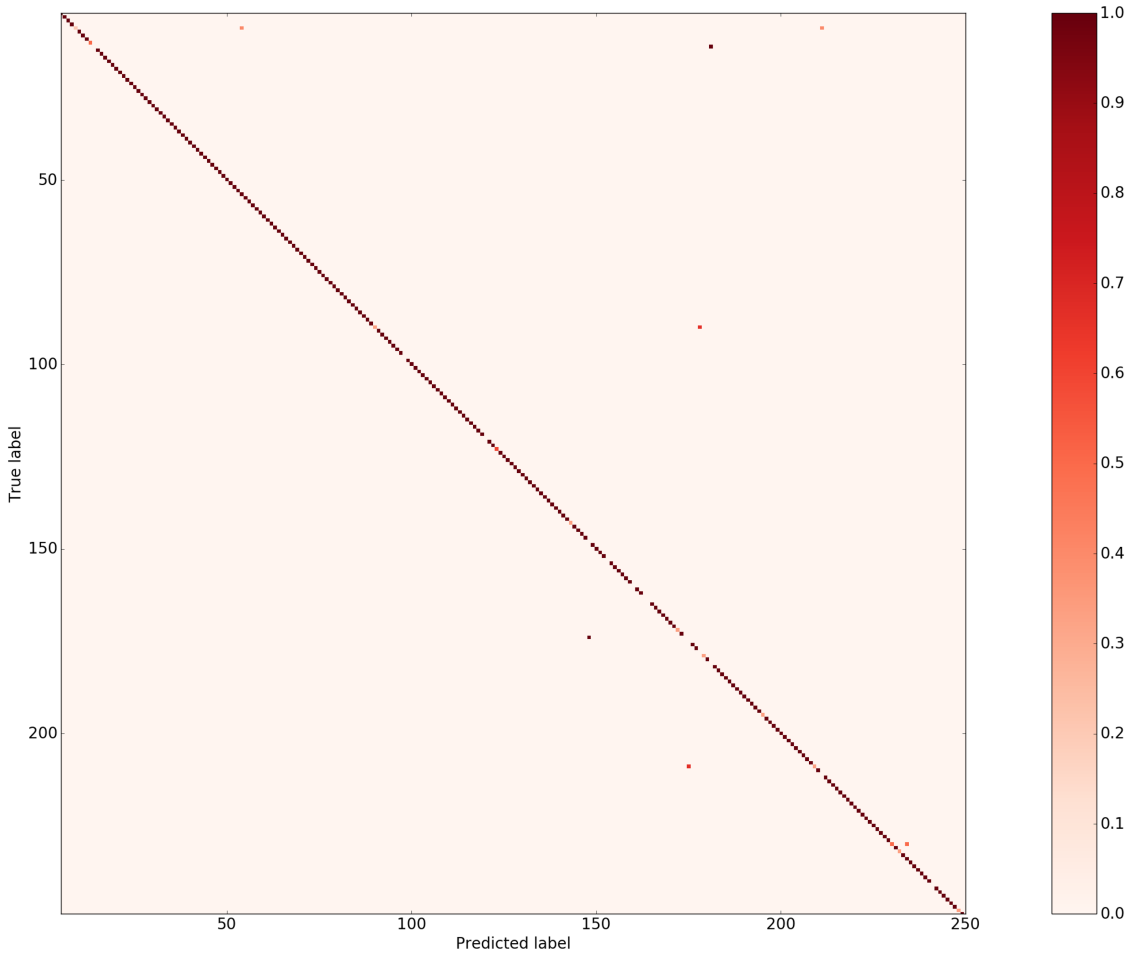


Fig. 13 *Confusion Matrix over a subset of test data between individuals ID number 5–248*

Table 4 Accuracy score for each ERT training fold

Accuracy score	# Fold
0.91202873	1
0.90125673	2
0.90125673	3
0.92040694	4
0.91083184	5
0.90604428	6
0.91322561	7
0.92339916	8
0.90724117	9
0.91861161	10

Table 5 Relative weight of landmark coordinates in the recognition procedure. For a visual reference of landmark locations see Fig. 1

Weight, %	# Landmark	Coordinate
1.5127	42	*x*
1.4552	36	*y*
1.4467	3	*y*
1.4442	43	*x*
1.3910	5	*x*
1.3798	40	*y*
1.3739	39	*y*
1.3671	35	*y*
1.3648	2	*y*
1.3641	1	*y*

6 References

[1] Paternoster, L., Zhurov, A.I., Toma, A.M., *et al.*: 'Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position', *Am. J. Hum. Genet.*, 2012, **90**, pp. 478–485

[2] Liu, F., van der Lijn, F., Schurmann, C., *et al.*: 'A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans', *PLoS Genet.*, 2012, **8**, (9): e1002932. doi: 10.1371/journal.pgen.1002932

[3] Gómez-Valdés, J.A., Hünemeier, T., Contini, V., *et al.*: 'Fibroblast growth factor receptor 1 (FGFR1) variants and craniofacial variation in Amerindians and related populations', *Am. J. Human Biol*, 2013, **25**, pp. 12–19. http://www.ncbi.nlm.nih.gov/pubmed/23070782

[4] Alberink, I., Ruifrok, A.: 'Performance of the FearID earprint identification system', *Forensic Sci. Int.*, 2007, **166**, (2–3), pp. 145–154. http://www.ncbi.nlm.nih.gov/pubmed/16772109/23/11/16

[5] Sforza, C., Grandi, G., Binelli, M., *et al.*: 'Age- and sex-related changes in the normal human ear', *Forensic Sci. Int.*, 2009, **187**, (1–3), p. 110.e1–7. http://www.ncbi.nlm.nih.gov/pubmed/19356871

[6] Ibrahim, M.I.S., Nixon, M.S., Mahmoodi, S.: 'The effect of time on ear biometrics'. 2011 International Joint Conf. on Biometrics (IJCB). IEEE, October 2011, pp. 1–6. http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6117584

[7] Zelditch, M.L., Swiderski, D.L., Sheets, H.D., *et al.*: 'Geometric morphometrics for biologists', *Elsevier*, 2004, **59**, (3), p. 457. http://www.sciencedirect.com/science/article/B848M-4MWYDVJ-7/2/2ec7a8306777645c5716a53774ec699c$\delimiter'026E30F$nhttp://books.google.com/books?id=LKCVAGn8vkoC{&}pgis=1

[8] Pflug, S., Busch, C.: 'Ear biometrics: a survey of detection, feature extraction and recognition methods', *IET Biometrics*, 2012, **1**, (2), p. 114

[9] Iannarelli, A.V.: '*Ear identification*' (Paramont Publishing Company, 1989), vol. **1**. https://books.google.com/books?id=jgPkAAAACAAJ{&}pgis=1

[10] Chen, H., Bhanu, B.: 'Shape model-based 3D ear detection from side face range images'. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05) – Workshops. IEEE, vol. **3**, pp. 122–122. http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1565434

[11] Chen, H., Bhanu, B.: 'Contour matching for 3D ear recognition'. Seventh IEEE Workshop on Applications of Computer Vision, WACV 2005, 2007, pp. 123–128

[12] Attarchi, S., Faez, K., Rafiei, A.: Advanced Concepts for Intelligent Vision Systems, October 2008 (LNCS, **5259**). http://dl.acm.org/citation.cfm?id=1462298.1462399

[13] Ansari, S., Gupta, P.: 'Localization of ear using outer helix curve of the ear'. International Conf. on Computing: the Theory and Applications, 2007, pp. 688–692

[14] Prakash, S., Gupta, P.: 'An efficient ear localization technique', *Image Vis. Comput.*, 2012, **30**, (1), pp. 38–50

[15] Yan, P., Bowyer, K.W.: 'Biometric recognition using 3D ear shape', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (8), pp. 1297–1308. http://www.ncbi.nlm.nih.gov/pubmed/17568136

[16] Purkait, R., Singh, P.: 'A test of individuality of human external ear pattern: its application in the field of personal identification', *Forensic Sci. Int.*, 2008, **178**, (2–3), pp. 112–118

[17] Ercan, I., Ozdemir, S.T., Etoz, A., *et al.*: 'Facial asymmetry in young healthy subjects evaluated by statistical shape analysis', *J. Anat.*, 2008, **213**, (6), pp. 663–669

[18] Pflug, A., Winterstein, A., Busch, C.: 'Robust localization of ears by feature level fusion and context information'. Biometrics (ICB), 2013 International Conf. on. IEEE, 2013, pp. 1–8

[19] Liu, Y., Zhang, B., Zhang, D.: 'Ear-parotic face angle: a unique feature for 3D ear recognition', *Pattern Recognit. Lett.*, 2015, **53**, pp. 9–15. http://linkinghub.elsevier.com/retrieve/pii/S0167865514003316

[20] Sibai, F.N., Nuaimi, A., Maamari, A., *et al.*: 'Ear recognition with feed-forward artificial neural networks', *Neural Comput. Appl.*, 2013, **23**, (5), pp. 1265–1273. http://link.springer.com/10.1007/s00521-012-1068-1

[21] Abaza, A., Hebert, C., Harrison, M.A.F.: 'Fast learning ear detection for real-time surveillance'. IEEE 4th International Conf. on Biometrics: Theory, Applications and Systems, BTAS 2010, 2010

[22] Islam, S.M.S., Bennamoun, M., Davies, R.: 'Fast and fully automatic ear detection using cascaded adaboost'. 2008 IEEE Workshop on Applications of Computer Vision, WACV, 2008

[23] Yuan, L., Liu, W., Li, Y.: 'Non-negative dictionary based sparse representation classification for ear recognition with occlusion', *Neurocomputing*, 2016, **171**, pp. 540–550

[24] Kumar, A., Chan, T.-S.T.: 'Robust ear identification using sparse representation of local texture descriptors', *Pattern Recogn.*, 2013, **46**, (1), pp. 73–85

[25] Kumar, A., Wu, C.: 'Automated human identification using ear imaging', *Pattern Recogn.*, 2012, **45**, (3), pp. 956–968

[26] Chan, T.-S., Kumar, A.: 'Reliable ear identification using 2-d quadrature filters', *Pattern Recognit. Lett.*, 2012, **33**, (14), pp. 1870–1881

[27] Mitteroecker, P., Gunz, P.: *Evol Biol*, 2009, **36**: 235. doi:10.1007/s11692-009-9055-x

[28] Slice, D.: 'Modern morphometrics'. Modern morphometrics in physical anthropology, 2005, pp. 1–46. http://link.springer.com/chapter/10.1007/0-387-27614-9{_}1

[29] Azaria, R., Adler, N., Silfen, R., *et al.*: 'Morphometry of the adult human earlobe: a study of 547 subjects and clinical application', *Plast. Reconstr. Surg.*, 2003, **111**, (7), pp. 2398–2402; discussion 2403-4. http://www.ncbi.nlm.nih.gov/pubmed/12794488

[30] Alexander, K.S., Stott, D.J., Sivakumar, B., *et al.*: 'A morphometric study of the human ear', *J. Plastic Reconstruct. Aesthetic Surgery*, 2011, **64**, (1), pp. 41–47. http://www.ncbi.nlm.nih.gov/pubmed/20447883

[31] Adhikari, K., Reales, G., Smith, A.J.P., *et al.*: 'A genome-wide association study identifies multiple loci for variation in human ear morphology', *Nat. Commun.*, 2015, **6**, (May), p. 7500. http://www.nature.com/doifinder/10.1038/ncomms8500

[32] Goodall, C.: 'Procrustes methods in the statistical analysis of shape', *J. R. Stat. Soc. B (Methodological)*, 1991, **53**, pp. 285–339. http://www.jstor.org/stable/2345744

[33] Ciodaro, T., Deva, D., de Seixas, J.M., *et al.*: 'Online particle detection with neural networks based on topological calorimetry information', *J. Phys. Conf. Ser.*, 2012, **368**, (1), p. 012030. http://iopscience.iop.org/article/10.1088/1742-6596/368/1/012030

[34] Ma, J., Sheridan, R.P., Liaw, A., *et al.*: 'Deep neural nets as a method for quantitative structureactivity relationships', *J. Chem. Inf. Model.*, 2015, **55**, (2), pp. 263–274. http://www.ncbi.nlm.nih.gov/pubmed/25635324

[35] Taigman, Y., Yang, M., Ranzato, M., *et al.*: 'DeepFace: closing the gap to human-level performance in face verification'. Conf. on Computer Vision and Pattern Recognition (CVPR), 2014, p. 8. http://www.cs.tau.ac.il/{~}wolf/papers/deepface{_}11{_}01{_}2013.pdf

[36] LeCun, Y., Bengio, Y., Hinton, G.: 'Deep learning', *Nature*, 2015, **521**, (7553), pp. 436–444. http://dx.doi.org/10.1038/nature14539$\delimiter‘026E30F$n10.1038/nature14539

[37] Dieleman, S., Willett, K.W., Dambre, J.: 'Rotation-invariant convolutional neural networks for galaxy morphology prediction', *Mon. Not. R. Astron. Soc.*, 2015, **450**, (2), pp. 1441–1459. http://arxiv.org/abs/1503.07077

[38] Fukushima, K.: 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position', *Biol. Cybern.*, 1980, **36**, (4), pp. 193–202

[39] LeCun, Y., Bottou, L., Bengio, Y., *et al.*: 'Gradient-based learning applied to document recognition', *Proc. IEEE*, 1998, **86**, (11), pp. 2278–2323

[40] Toshev, A., Szegedy, C.: 'DeepPose: human pose estimation via deep neural networks'. Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1653–1660. http://arxiv.org/abs/1312.4659

[41] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks'. Advances in Neural Information Processing Systems, 2012, pp. 1–9

[42] Boureau, Y.-L., Ponce, J., Lecun, Y.: 'A theoretical analysis of feature pooling in visual recognition'. Proc. of the 27th International Conf. on Machine Learning (2010), 2010, pp. 111–118. http://www.ece.duke.edu/{~}lcarin/icml2010b.pdf

[43] Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., *et al.*: 'Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals', *PLoS Genet.*, 2014, **10**, (9), p. e1004572. http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004572

[44] Viola, P., Jones, M.: 'Robust real-time object detection', *Int. J. Comput. Vis.*, 2001, **57**, pp. 137–154. http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Robust+Real-time+Object+Detection{#}0

[45] Dieleman, S., Schlüter, J., Raffel, C., *et al.*: 'Lasagne: first release', August 2015. http://dx.doi.org/10.5281/zenodo.27878

[46] Hinton, G.E., Srivastava, N., Krizhevsky, A., *et al.*: 'Improving neural networks by preventing co-adaptation of feature detectors', arXiv: 1207.0580, 2012, pp. 1–18. http://arxiv.org/abs/1207.0580

[47] Çeliktutan, O., Ulukaya, S., Sankur, B., *et al.*: 'A comparative study of face landmarking techniques', *EURASIP J. Image and Video Process.*, 2013, **2013**, (1), p. 13. http://jivp.eurasipjournals.springeropen.com/articles/10.1186/1687-5281-2013-13

[48] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.*: 'Scikit-learn: machine learning in Python', *J. Mach. Learn. Res.*, 2011, **12**, pp. 2825–2830

[49] Solina, F., Peer, P., Batagelj, B., *et al.*: 'Color-based face detection in the' 15 seconds of fame' art installation'. Proc. of Mirage 2003, Conf. on Computer Vision/Computer Graphics, 2003, pp. 38–47

[50] Geurts, P., Ernst, D., Wehenkel, L.: 'Extremely randomized trees', *Mach. Learn.*, 2006, **63**, (1), pp. 3–42

[51] Wehenkel, L., Ernst, D., Geurts, P.: 'Ensembles of extremely randomized trees and some generic applications'. Robust Methods for Power System State Estimation and Load Forecasting, 2006

[52] Faris, C.: 'Scott-Brown's otorhinolaryngology, head and neck surgery, 7th edn', October 2011, p. 559. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604940/

[53] Cox, T.C., Camci, E.D., Vora, S., *et al.*: 'The genetics of auricular development and malformation: new findings in model systems driving future directions for microtia research', *Eur. J. Med. Genet.*, 2014, **57**, (8), pp. 394–401. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4143470{&}tool=pmcentrez{&}rendertype=abstract