Deformable Models of Ears in-the-wild for Alignment and Recognition

Yuxiang Zhou[1] and Stefanos Zaferiou[1,2]
[1] Department of Computing, Imperial College London, U.K.
[2] Centre for Machine Vision and Signal Analysis, University of Oulu, Finland

***Abstract*— Ears have been discovered to have biometric importance for identifying people and/or verifying their identity. This is largely because of their complex inner shape structure, which is not only unique but also long-lasting regardless of ageing. In this paper, we make two important contributions relevant to analysis of ear in imagery captured in unconstrained conditions. That is, we present (a) the first, to the best of our knowledge, annotated database with ear landmarks and use it in order to build statistical deformable ear models in-the-wild and (b) a database of 2058 labelled unconstrained ear images with 231 subjects and use it for ear recognition/verification. We perform extensive comparisons for ear alignment using many state-of-the-art techniques and extensive experiments. Finally, we conducted extensive experiments for ear recognition using both handcrafted, as well as learned features (i.e., using deep learning). All annotated data and code will be publicly available.**

I. INTRODUCTION

Given the increasing focus on automatic identity verification during the last decade, biometrics have attracted extended attention. Such applications seek of biometric characteristics that are special, common and quantifiable. One such biometric is human ear [12], [13]. The human outer ear consists of the following parts: outer helix, antihelix, lobe, tragus, antitragus, helicis, crus helicis and concha (see Figure 2). The inner structure of the human ear is formed with numerous rides and valleys which makes it very distinctive. Even though the human ear's structure is not completely random, it still brings significant differences between individuals. The influence of randomness on appearance can be observed even by comparing both ears of the same person. Ears of same person have similarities but still are not perfectly symmetric [30].

The complex interior shape of ears has long been considered as a valuable identification metric. The first time it was utilised for human verification was hundreds years ago [10]. Several years later, researchers demonstrated that 500 ears can be distinguished using only four features [24]. The work of [23] also showed that 10k ears can be determined with 12 features. Furthermore, ear can be in many cases combined with face for improved person recognition and verification [12].

As in many biometrics, such as face [36], the first step towards recognition/verification is, arguably, alignment. Since, ear is a deformable object a statistical deformable model should be learned. In order to learn the first statistical deformable model of the ear we collected and annotated, with regards to 55 landmarks, the first "in-the-wild" ear database.

978-1-5090-4023-0/17/$31.00 ©2017 IEEE

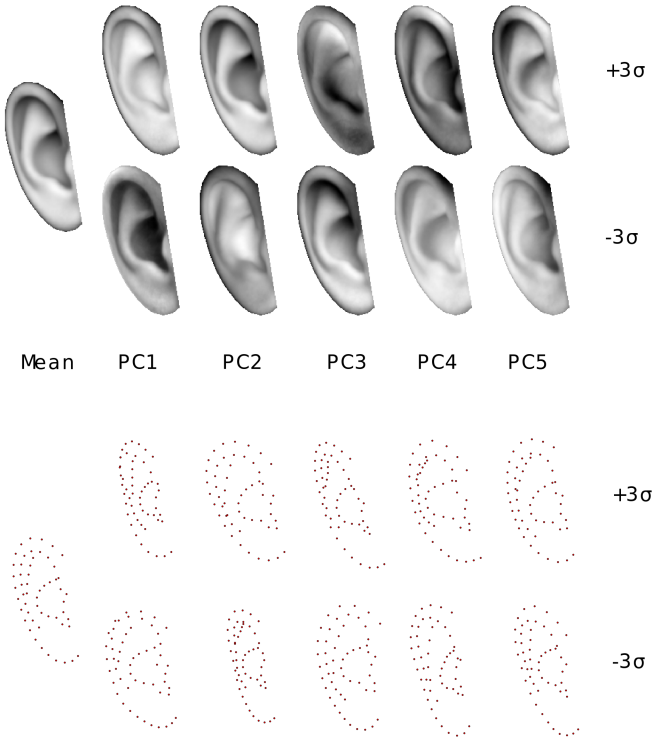

Fig. 1. Exemplar statistical shape and appearance model of human ear. The figure visualises the first five principal components variation in both models. Appearance model is created with pixel intensity for better visualisation.

Furthermore, we conducted an extensive experimental comparison for ear landmark localisation using state-of-the-art generative and discriminative methodologies for training and fitting statistical deformable models [15], [14], [28], [33], [8], [40], [11], [6], [37], [7], [5]. Figure 1 visualises the mean and the first variations along the 5 principal components of the texture and the shape (as performed in Active Appearance Model [14], [28], [37], [5]).

The other contribution of the paper is the collection of a new "in-the-wild" database suitable for ear verification and recognition. The collected database consists of 231 subjects with 2058 "in-the-wild" images. We conduct extensive experimental comparisons in the newly collected database using various handcrafted features such as Image Gradient Orientation (IGO) [38], Scale Invariant Feature Transforms (SIFT) [27], Histogram of Oriented Gradients (HOG) [17], as well as learned deep convolutional features [34]. Finally, we compare the effect of alignment in ear recognition/verification.

Summarising, the contributions proposed in this paper are:

- We present the first annotated "in-the-wild" database of images of ears (605 images in total) with regards to 55 landmarks. We provide the database publicly available.
- We conduct an extensive comparison between various discriminant and generative methodologies for ear landmark localisation "in-the-wild".
- We collect a large database of ears "in-the-wild" for ear recognition/verification and we conduct an extensive experimental comparison.

II. Existing Databases

In the following we briefly review the available databases for ear recognition and argue for a collection of a new one "in-the-wild" database for ear alignment and one for ear recognition/verification "in-the-wild". The list of the most popular ear databases includes the following database UND-Collection E [1], EUBEAR [32], IIT [25], WPUTEDB [19] and ScFace [20]), most of which have been captured under controlled laboratory conditions or lack of annotations. In particular,

UND Databases Collection E includes 464 right profiled ear images from 114 identities, from which 3 to 9 images are taken for each person in days with various pose and illumination conditions.

WPUTEDB introduces 3348 images of 421 subjects each having 4 to 12 images taken under controlled environments [19]. Various indoor lighting conditions, occlusion by hair and accessories, and slightly angled positions are involved in this database to simulate "in-the-wild" condition but still very limited to specific scenario.

IIT Delhi database contains 125 subjects where each has 3 to 6 images taken in grayscale. Images are taken in indoor condition with limited lighting variation. No or occasionally occlusion and pose variation occurred.

UBEAR dataset involved 126 subjects with an average 35 images corresponding to each subject. Lighting conditions, pose variations and occlusions are all applied to this database. But images are collected from indoor video therefore the with-in class variation is quite limited.

Note that all datasets above are collected under controlled environment and none, to the best of our knowledge, has been annotated with regards to landmarks. Furthermore, as we will show in the experimental result section, in WPUTEDB database the area around the ear contains very discriminative information. This is an indication that the data have been collected within small time intervals. In this paper, we make a significant step further and collect and annotate databases of ears "in-the-wild". Furthermore, we made an effort so that the ear samples for each person have been taken with considerable time interval.

To the best of our knowledge the only ear database that has been collected "in-the-wild" is the one presented in [18], which contains a very limited amount of subjects (only 16).

III. "In-the-wild" Ear database

We collected two sets of ear images "in-the-wild"[1]. The first was used for statistical deformable model construction, while the latter was used for ear verification and recognition "in-the-wild".

Collection A consists of 605 ear images "in-the-wild" collected from Google Images with no specific identity (by searching using the ear related tags). Each is manually annotated with 55 landmark points. Examples of such annotated images and the anatomy of pinna is shown in figure 2. The semantic meaning of the 55 landmarks are: ascending helix (0-3), descending helix (4-7), helix (8-13), ear lobe (14-19), ascending inner helix (20-24), descending inner helix (25-28), inner helix (29-34), tragus (35-38), canal (39), antitragus (40-42), concha (43-46), inferio crus (47-49) and supperio crus (50-54). We randomly split the images into two disjoint sets for training (500) and testing (105). The purpose of Collection A is to build statistical deformable models with unconstrained ear samples.

Collection B contains 2058 images contains 231 identity-labelled subjects collected from VGG database [29], which contains more than one million images of celebrities with

[1]Both Collection A and B are publicly available in http://www.ibug.doc.ic.ac.uk/resources/ibug-ears.

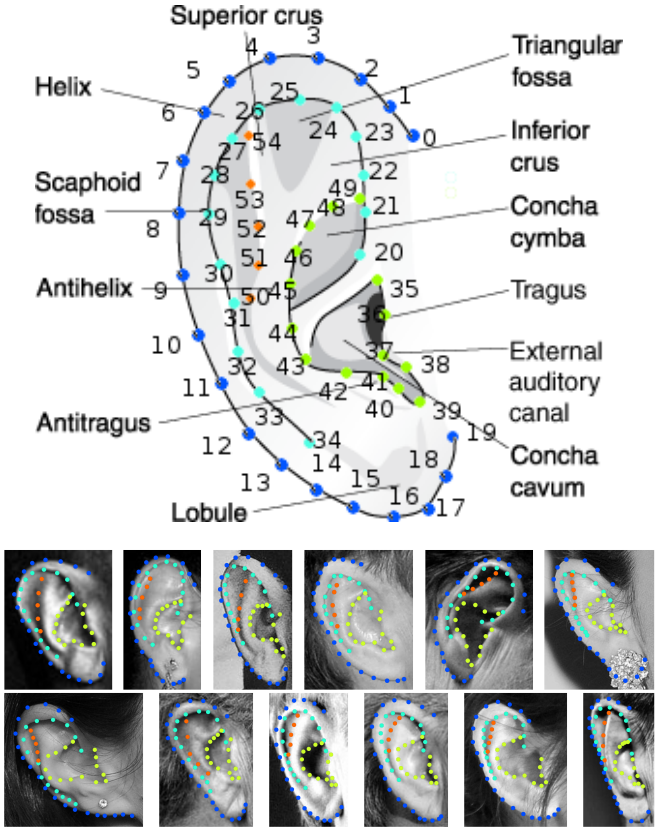

Fig. 2. Example of the annotated 55 landmarks on ears. Ascending helix (0-3), descending helix (4-7), helix (8-13), ear lobe (14-19), ascending inner helix (20-24), descending inner helix (25-28), inner helix (29-34), tragus (35-38), canal (39), antitragus (40-42), concha (43-46), inferio crus (47-49) and supperio crus (50-54).

only identity labels. As the purpose of VGG database was face recognition "in-the-wild", we had to manual select images were ears are visible (not fully occluded) and furthermore there bounding box could be generated by a simple HoG Support Vector Machine (SVM) [17] ear detector (trained on collection A). Exemplar collected ear images are shown in Figure 2 that images are under challenging environment such as heavily posed angle, significant lighting variations, notable occlusions, variant resolutions, and significant ageing. It is so far, to the best of our knowledge, the largest ear in-the-wild databases. The most related database is [18], which contains only 16 subjects.

Upon acceptance of the paper, both databases will be made available to the research community.

IV. Holistic and Patch-based Statistical Deformable Models

We have applied many state-of-the-art statistical deformable models for ear landmark localisation. The methodologies includes Constrained Local Models (CLMs) [16], Supervised Descent Method (SDM) [39] and various Active Appearance Model (AAM) methods [14], [28], [37], [5]. In the following, we will focus on the general AAM architectures applied, since they were the top performing ones. The annotated ear dataset was used to build two different kind of AAMs [14], [28]: *holistic* [31], [3], [4], [5] and *patch-based* [37]. The difference between these two models is on the way that the appearance is represented, as well as the deformation model.

A. Holistic Active Appearance Model

AAM method consists of a linear statistical model of the shape and appearance of an object. During fitting, they aim to minimise the appearance reconstruction error with respect to the parameters of the shape and appearance models. Initially it was proposed to optimise their cost function using regression [14]. Later, it was also shown that they can achieve state-of-the-art performance by employing the Gauss-Newton optimisation [28], [5].

Shape Model A shape vector is defined by concatenating the coordinates of its landmarks. A shape model can be trained by applying Generalized Procrustes Analysis followed by Principal Component Analysis (PCA) on a set of training shapes. The shape model can then be used to generate shape instances with N_L landmarks as

$$\boldsymbol{s_p} = \bar{\boldsymbol{s}} + \boldsymbol{U_S p} \tag{1}$$

where a shape is represented as $\boldsymbol{s} = [x_1, y_1, .., x_{N_L}, y_{N_L}]^T$, and $\bar{\boldsymbol{s}}$ is the mean shape, $\boldsymbol{p}$ are the shape parameters and $\boldsymbol{U_s}$ are the principal components matrix of dimension $\boldsymbol{U_S} \in \Re^{2N_L \times N_p}$, where N_p represent the number of eigenvectors.

Appearance Model A holistic appearance is defined as the values of the pixels that lie inside a shape $\boldsymbol{s}$. Similar to shape models, an appearance model is trained using PCA. Given the appearance eigenvectors $\boldsymbol{U_A}$, the mean appearance $\bar{\boldsymbol{a}}$ and a set of parameters $\boldsymbol{\lambda}$, a new appearance can be generated as

$$\boldsymbol{a_\lambda} = \bar{\boldsymbol{a}} + \boldsymbol{U_A \lambda} \tag{2}$$

where $\boldsymbol{a}$ denotes the vector of pixels that lie within a shape, $\bar{\boldsymbol{a}}$ is the mean appearance, $\boldsymbol{\lambda}$ are the appearance parameters and $\boldsymbol{U_A}$ are the appearance principal components matrix of dimension $\boldsymbol{U_A} \in \Re^{N_A \times N_\lambda}$, where N_λ represent the number of appearance eigenvectors and N_A represented the length of eigenvector e.g. number of pixels within mean shape if single-channel appearance model considered. Note that the appearance can be represented by a handcrafted feature function (e.g. SIFT, HOG) or a learned feature (e.g. Dense CNN).

Deformation Model The deformation model of an AAM consists of a warp function $\mathscr{W}(\boldsymbol{p})$, which maps all the points $\boldsymbol{s_p}$ within a source shape defined by the shape parameters p to their corresponding coordinates in a reference shape (commonly the mean shape $\bar{\boldsymbol{s}}$). This procedure is necessary in order to bring the appearance vectors of different images into correspondence. We employ the Piecewise Affine Warp, which performs the mapping based on the barycentric coordinates of the corresponding triangles between the two shapes that are extracted using Delaunay Triangulation.

Fitting The aim of fitting is to minimise the ℓ_2^2 norm between the warped appearance of an input image $\boldsymbol{T}(\mathscr{W}(\boldsymbol{p}))$ and the appearance model instance $\boldsymbol{a_\lambda}$ with respect to the shape and appearance parameters, i.e.

$$\underset{\boldsymbol{p},\boldsymbol{\lambda}}{\arg\min} ||\boldsymbol{T}(\mathscr{W}(\boldsymbol{p})) - \bar{\boldsymbol{a}} - \boldsymbol{U_A \lambda}||^2 \tag{3}$$

Inverse Compositional (IC) Algorithm is an efficient gradient descend method that, in general, introduced an incremental warp, which composing with the current warp at each iteration as

$$\mathscr{W}(\boldsymbol{p}) \leftarrow \mathscr{W}(\boldsymbol{p}) \circ \mathscr{W}(\boldsymbol{\Delta p})^{-1} \tag{4}$$

Thereby the cost function for inverse compositional algorithm are:

$$\underset{\boldsymbol{\Delta p},\boldsymbol{\lambda}}{\arg\min} ||\boldsymbol{T}(\mathscr{W}(\boldsymbol{p})) - \bar{\boldsymbol{a}}(\mathscr{W}(\boldsymbol{\Delta p})) - \boldsymbol{U_A}(\mathscr{W}(\boldsymbol{\Delta p}))\boldsymbol{\lambda}||^2 \tag{5}$$

where $\mathscr{W}(\boldsymbol{\Delta p})$ denotes incremental warp on template image. Applying first order Taylor expansion on equation 5 gives:

$$\underset{\boldsymbol{\Delta p},\boldsymbol{\lambda}}{\arg\min} ||\boldsymbol{T}(\mathscr{W}(\boldsymbol{p})) - \bar{\boldsymbol{a}} - \boldsymbol{U_A \lambda} - \boldsymbol{J}|_{\boldsymbol{p}=\boldsymbol{0}} \boldsymbol{\Delta p}||^2 \tag{6}$$

where Jacobian term is defined as:

$$\boldsymbol{J}|_{\boldsymbol{p}=\boldsymbol{0}} = \nabla(\bar{\boldsymbol{a}} + \boldsymbol{U_A \lambda}) \frac{\partial \mathscr{W}}{\partial \boldsymbol{p}}|_{\boldsymbol{p}=\boldsymbol{0}} \tag{7}$$

Alternating Minimisation As the expression revealed, there are two variables contained ($\boldsymbol{p}$ and $\boldsymbol{\lambda}$) so it is of importance to minimise them simultaneously. As there is no dependency between AAM shape and appearance, solving $\boldsymbol{p}$ and

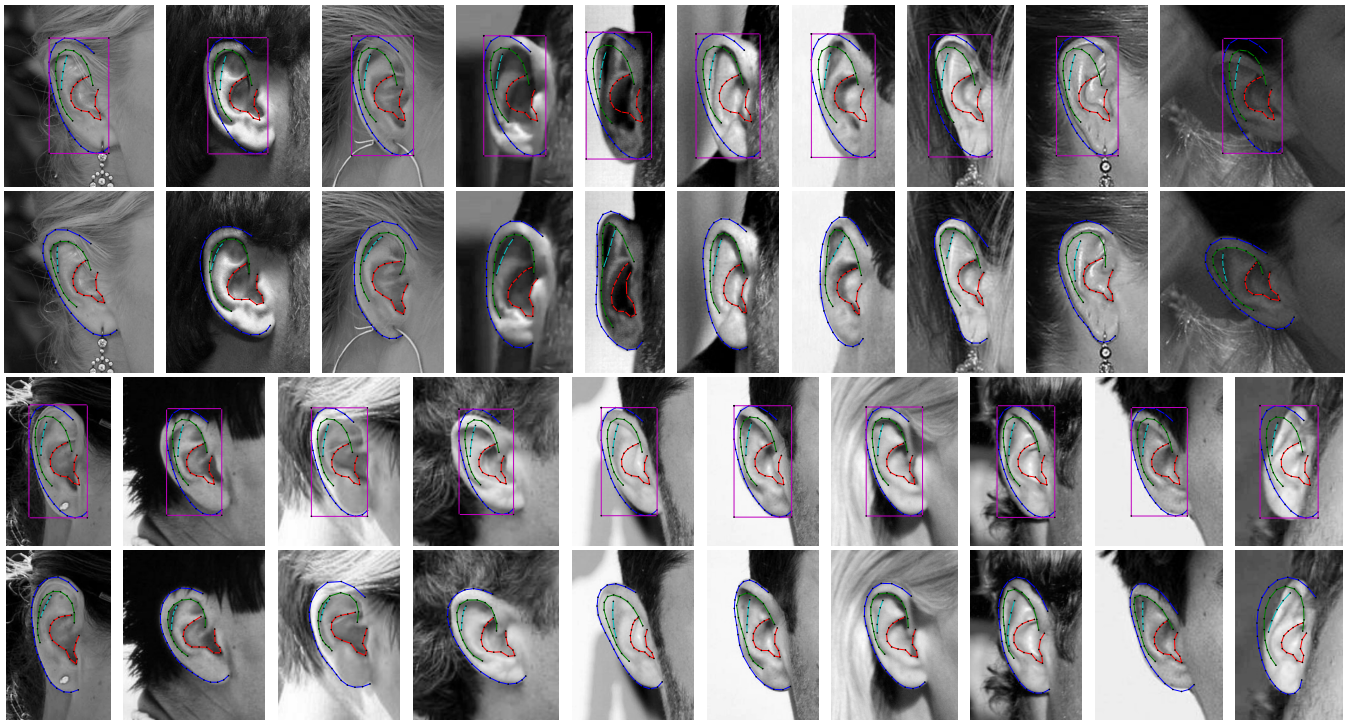

Fig. 3. Visualisation of Holistic AAM ear fitting results (cropped to ear only for better visualisation). First row demonstrates the bounding box generated from our in-house ear detector with corresponding initialisation. Second row presented the final fitting using Holistic AAM. Figure best viewed by zooming in.

$\boldsymbol{\lambda}$ alternatively is used by IC Algorithm. Cost functions corresponding to $\boldsymbol{\Delta p}$ and $\boldsymbol{\Delta\lambda}$ are:

$$\arg\min_{\boldsymbol{\Delta p}} ||\boldsymbol{T}(\mathscr{W}(\boldsymbol{p})) - \boldsymbol{a_\lambda}\mathscr{W}(\boldsymbol{\Delta p})||^2_{\hat{U}_A} \tag{8}$$

$$\arg\min_{\boldsymbol{\Delta\lambda}} ||\boldsymbol{T}(\mathscr{W}(\boldsymbol{p})) - \boldsymbol{a}_{\lambda+\Delta\lambda}\mathscr{W}(\boldsymbol{\Delta p}))||^2 \tag{9}$$

where $||.||^2_{\hat{U}_A}$ denotes vector projected into subspace $\hat{\boldsymbol{U}}_{\boldsymbol{A}}$, which is orthogonal complements of appearance $\hat{\boldsymbol{U}}_{\boldsymbol{A}} = \boldsymbol{I} - \boldsymbol{U_A U_A^T}$. Because norm considers only orthogonal components of subspace, so any other components lies in $\hat{\boldsymbol{U}}_{\boldsymbol{A}}$ can be dropped. So the optimisation is accomplished by (1) fixing appearance parameter $\boldsymbol{\lambda}$ to compute $\boldsymbol{\Delta p}$ (2) then, similarly, fixing shape parameter $\boldsymbol{p}$ to compute $\boldsymbol{\Delta\lambda}$. By given estimation of $\boldsymbol{\lambda}$, we can solve Δp in closed form as:

$$\boldsymbol{\Delta p} = \boldsymbol{H}^{-1}\boldsymbol{J}^T[\boldsymbol{T}(\mathscr{W}(\boldsymbol{p})) - \bar{\boldsymbol{a}} - \boldsymbol{U_A\lambda}] \tag{10}$$

where $\boldsymbol{H}$ is the Hessian matrix $\boldsymbol{H} = \boldsymbol{J}^T\boldsymbol{J}$. Given estimation of $\boldsymbol{p}$, appearance parameters can be solved as least-squares solution:

$$\boldsymbol{\Delta\lambda} = \boldsymbol{U_A^T}[\boldsymbol{T}(\mathscr{W}(\boldsymbol{p})) - \bar{\boldsymbol{a}}(\mathscr{W}(\boldsymbol{\Delta p})) - \boldsymbol{U_A}(\mathscr{W}(\boldsymbol{\Delta p}))\boldsymbol{\lambda}] \tag{11}$$

where appearance parameters are updated by $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \boldsymbol{\Delta\lambda}$.

As the equation states, image template $\bar{\boldsymbol{a}}$ is constant and the gradient of warp $\frac{\partial\mathscr{W}}{\partial\boldsymbol{p}}$ is always evaluated at appearance template, which remains constant. Therefore Jacobian at initial iteration and Hessian Matrix $\boldsymbol{H}$ can be precomputed before optimising cost function.

B. Patch-based Active Appearance Model

The difference between a holistic and a patch-based AAM [37] is on the way that the appearance vectors are extracted, as well as the deformation model. As explained in the previous section, a holistic appearance is retrieved using the warp function in order to map the locations of all the pixels of a given shape into a common reference shape and transfer their values. However, under a patch-based formulation, this procedure is greatly simplified and an appearance vector is acquired by concatenating the features (e.g. SIFT) extracted from the patches centred at the landmarks of a provided shape instance. Thus, the affine warp function is replaced by a simple sampling function. In [37] it is shown that due to this difference, the compositional update of the shape parameters becomes additional, i.e. $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$. The rest of the Gauss Newton optimisation remains the same.

Patch-based AAMs achieve more accurate performance compared to holistic AAMs on the task of face alignment [37]. However, this is easily explained by the fact that the inner appearance information of a human face (i.e. cheeks eta.) does not have a distinctive structure. Experiment V-A further explains the advanced performance of holistic AAM on ears where inner appearance is complex and rich. In general, the selection of a holistic or patch-based appearance representation highly depends on the nature of the modelled object and its interior structure.

V. Experimental Evaluation

We have conducted two set of extensive experiments. The first set of experiments concerns ear landmark local-

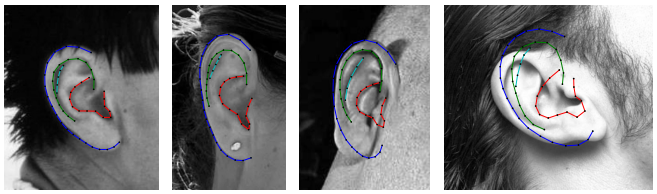

Fig. 4. Examplar visualisations showing different values of normalised point-to-point error measure for ears (0.06, 0.10, 0.14, 0.24 respectively). Figure best viewed by zooming in.

isation using the Collection A. The second set of experiments revolves around ear recognition/verification using Collection B. In particular, the aim and motivation of ear recognition/verification experiments are the following. (a) To demonstrate the shortcomings of the current available databases, and particular WPUTEDB, (b) to demonstrate the challenges that emerge from our new database and (c) to demonstrate the effect of alignment in recognition/verification.

A. Ear Fitting Evaluation

We evaluated the performance of many state-of-the-art methodologies including AAM, CLM and SDM using various kind of features. In particular, we employed pixel intensity (PI), dense SIFT (DSIFT) [27], dense HOG [17], IGO [38] and DCNN [34] features for both holistic and patch-based AAMs. The models were trained on a 3-level Gaussian pyramid. We kept $[3, 6, 12]$ shape components for each level (low to high) and 90% of the appearance variance for all levels. Also discriminative models like Supervised Descend Method (SDM) [39] and Constrained Local Model (CLM) [16] are involved using features DSIFT [27]. Figure 1 visualises the first five shape and appearance principal components of the holistic AAM and, as it can be observed, the variation of both shape and appearance is plausible. Appearance components are shown using pixel intensities for better visualisation. Note that any technologies involved in this paper was implemented using the Menpo Platform [2].

For all the tested methods we computed the Cumulative Error Distribution (CED) curves which is the standard way of visualising the performance of deformable models. The fitting error is computed using the point-to-point error normalised by the diagonal of the ear's bounding box that tightly bounded ground truth annotations, as proposed in [40].

Figure 5 reports the CED curves of all the tested methods along with the initialisation curve. The fitting is initialised using our own in-house ear detector based on HoG SVMs that was trained with in-the-wild ear images of the training set of Collection A. Figure 4 shows some characteristic examples of different error values in order to give an intuition about the fitting quality of each error bin of the CED curves, which indicates that normalised point-to-point error less or equal than 0.10 is acceptable. The figure reveals that DSIFT tends to give most representative features for ears and holistic AAMs in general outperform patch-based AAM. This is attributed to the fact that holistic texture model can represent

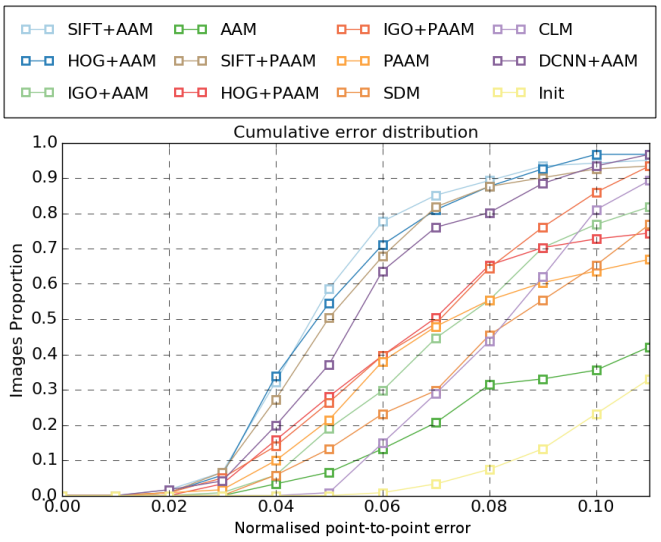

Fig. 5. Experimental results on our testing 121 dataset evaluated on 55 landmarks. Fitting accuracy reported for Holistic AAM, patch-based AAM, SDM and CLM.

Method	*mean* $\pm$ *std*	*median*	≤ 0.10
DCNN+AAM	0.0599 $\pm$ 0.0272	0.0542	93%
SIFT+AAM	**0.0522 $\pm$ 0.0246**	**0.0453**	94%
HOG+AAM	0.0539 $\pm$ 0.0248	0.0479	**97%**
IGO+AAM	0.0786 $\pm$ 0.0295	0.0738	77%
PI+AAM	0.2124 $\pm$ 0.2674	0.1342	36%
SIFT+PAAM	0.0563 $\pm$ 0.0264	0.0493	93%
HOG+PAAM	0.0860 $\pm$ 0.0533	0.0700	73%
IGO+PAAM	0.0704 $\pm$ 0.0272	0.0709	86%
PI+PAAM	0.1049 $\pm$ 0.0733	0.0729	64%
SDM	0.0890 $\pm$ 0.0348	0.0862	65%
CLM	0.0862 $\pm$ 0.0296	0.0830	81%
Initialisation	0.1276 $\pm$ 0.0332	0.1283	23%

TABLE I

FITTING STATISTICS ON EAR DATABASE COLLECTION A

the complex inner structure of ears in a better fashion. The poor performance of SDM could be associated to the limited annotated data, as well as to the use of a part-based texture model.

Table I complements Figure 5 by reporting some statistical metrics, i.e. the mean, median and standard deviation of the errors. Finally, Figure 3 shows some qualitative fitting results along with their initialisations.

B. Ear Recognition Experiments

In order to conduct close ear recognition experiments we conducted a 10 fold cross validation experiment where 90% were used for training and 10% testing in each fold. We report average accuracy and standard deviation. In order to compare how challenging each database is we applied the above protocol to both our database, as well as WPUTEDB, which contains largest amount of subjects and most significant appearance variance among existing ear databases but still collected under controlled environment. Furthermore, we wanted to investigate how the background of the ear images

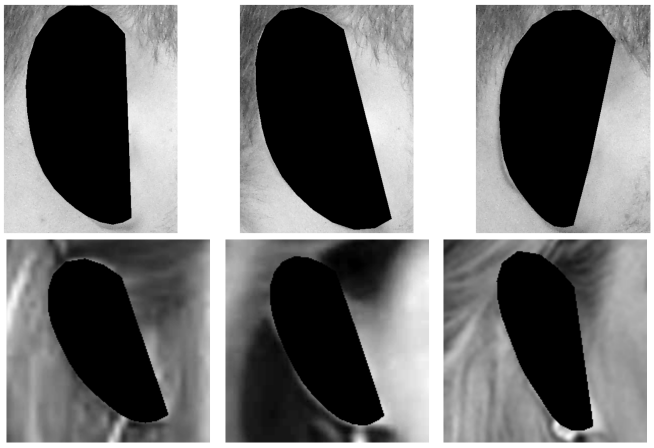

Fig. 6. Exemplar visualisation of background of the ear in both the WPUTEDB, as well as the proposed database. Top row: the background of ear images of WPUTEDB for one subject. Bottom row: images of the background from the proposed database for one subject. The black area covers the ear region.

	WPUTEDB *mean* ± *std*	Our Database *mean* ± *std*
Aligned		
LDA+PI	0.6581 ± 0.0030	0.2272 ± 0.0169
LDA+HOG+PCA	0.8195 ± 0.0058	0.5123 ± 0.0162
LDA+DSIFT+PCA	0.8082 ± 0.0098	0.5496 ± 0.0180
LDA+PPSIFT+PCA	**0.9076 ± 0.0038**	**0.6684 ± 0.0195**
Unaligned		
LDA+PI	0.5993 ± 0.0025	0.1429 ± 0.0123
LDA+HOG+PCA	**0.7784 ± 0.0074**	0.3250 ± 0.0105
LDA+DSIFT+PCA	0.7621 ± 0.0082	**0.3348 ± 0.0138**
Background Only		
LDA+DSIFT+PCA	0.5676 ± 0.0312	0.0827 ± 0.0173

TABLE II

EAR RECOGNITION EXPERIMENTS ON WPUTEDB AND THE PROPOSED DATABASE. MULTIPLE FEATURES AND CLASSIFICATION ALGORITHMS ARE APPLIED WITH/WITHOUT ALIGNMENT.

of WPUTEDB biases the results, since from Figure 6 it is evident that the background is very similar in the ear images of the same person of WPUTEDB.

Classification of ears was implemented by applying generic classification pipelines. In particular, the we applied the standard pipeline of feature extraction + dimensionality reduction + multi-class classification. For features we explored pixel intensities, HOG, SIFT and Pyramid Patch SIFT (PPSIFT) [2]. For dimensionality reduction we used a Principal Component Analysis plus Linear Discriminant Analysis (PCA plus LDA) framework [9]. For classification we using a multi-class SVM [21].

Finally, we compared aligned versus non-aligned ears. For alignment we used the previously described SIFT+AAM framework to locate the landmarks. We applied the AAM deformation model warp function to map all the points to the reference shape. This procedure is necessary in order to bring the SIFT features appearance vectors of all fitted images into correspondence. We employ the Piecewise Affine Warp, which performs the mapping based on the barycentric coordinates of the corresponding triangles between the two shapes that are extracted using Delaunay Triangulation. For the non-aligned version of experiments we used the cropped ear images from the detected bounding box. In both cases the ear images were rescaled in a 200×200 bounding box before feature extractions.

From the results reported in Table II we can deduce the following (a) the collected database is far more challenging than the WPUTEDB, (b) the background around the ear does not play any role in the proposed database, while the background gives a 57% recognition rate in WPUTEDB [3], (c) the alignment largely improves performance (approximately 5% average in WPUTEDB and 10% to 20% in our database) and (d) the best performance is achieved by PPSIFT features.

C. Ear Verification Experiments

In this section we have designed and executed an ear verification experiment reminiscent of the experimental protocol of Labelled Faces in-the-wild (LFW) [22]. That is, evaluation is performed by determining whether a pair of images come from the same person or not. In the case of ear verification, 185 positive and 185 negative matching pairs are generated for each fold and total five folds are generated, from which we perform a leave-one-out cross validation.

We used similar features as in the recognition experiments. In particular, we applied pixel intensities (PI), DSIFT and Deep Convolutional Neural Networks (DCNN) [35] (for DCNN we used the pre-trained VGG-16 architecture). High dimensional features, such as DSIFT, were combined with PCA for dimensionality reduction. For each pair of the training images and for each feature we computed the squared distance and formed a vector which was fed to a two class LDA or SVM which separate match versus not-match pairs. We apply the above methods to both aligned and non-aligned ear samples. Finally, we also applied the methodology that was proposed in [26] (so-called Eigen-Pep).

Overall performance over five folds is reported using mean accuracy (as in LFW). Experiments are performed under the image-restricted setting, where only binary positive or negative labels are given, for pairs of images. So the identity information of each image is not available under this setting and results are reported with both no outside training data and label-free outside data for alignment. Table III summarises the results. The top performance is around 68% using DCNN features and aligned images. As in the recognition experiments alignment always improves performance. Finally, Figure 7 plots the ROC curves for the tested methods.

[2]To compute PPSIFT, we build image pyramid for given images by rescale it by 0.25, 0.5, 1.0 of original image including landmarks. Then we extract small patches around landmarks at each pyramid level and compute SIFT feature so each patch gives a vector of size 128. PPSIFT are constructed by concatenating all patches.

[3]Since images from WPUTEDB for each subjects are all collected in short periods, even background (e.g. hair and earrings) provides significant support to classification which is not the case in our database.

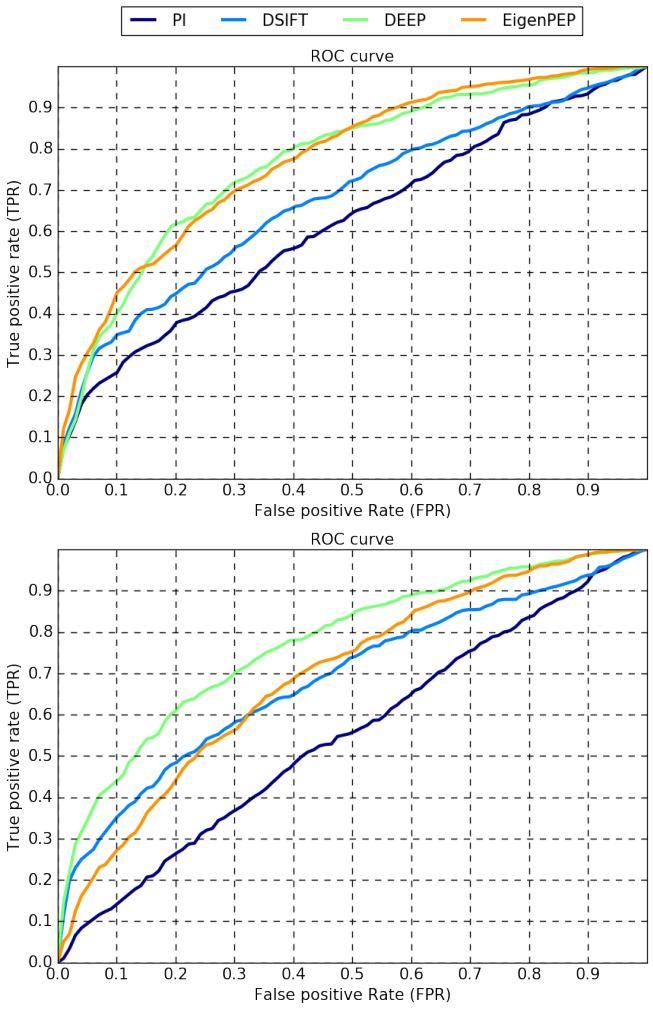

Fig. 7. ROC curves averaged over 5 folds on the proposed database (top: using aligned images, bottom: using unaligned images).

Method	Aligned *mean* ± *std*	Unaligned *mean* ± *std*
PI+LDA	0.5659 ± 0.017	0.5405 ± 0.023
DSIFT+PCA+LDA	0.6222 ± 0.012	0.6178 ± 0.025
JBC+EigenPEP+PCA	0.6297 ± 0.013	0.6189 ± 0.023
DCNN+LDA	**0.6859 ± 0.024**	0.6492 ± 0.032

TABLE III

EAR VERIFICATION BENCHMARK ON COLLECTION B. MEAN AND VARIANCE OF VERIFICATION ACCURACY ARE REPORTED FOR A SET OF METHODS APPLYING ON BOTH ALIGNED AND UNALIGNED DATA.

VI. CONCLUSION

In this paper, we collected two sets of challenging in-the-wild ear databases for the purpose of ear deformable model constructions and ear recognition/verification. The experimental evaluation and comparison revealed that holistic and patch-based AAMs can align images of ears captured "in-the-wild". We conducted extensive recognition and verification experiments. The experiments convincingly demonstrate that (a) the proposed database is very challenging and (b) alignment consistently improves performance.

VII. ACKNOWLEDGEMENT

This work was partially funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 688520 (TeSLA) and by EPSRC Project EP/N007743/1 (FACER2VM).

REFERENCES

[1] Und biometric dataset collection e, https://sites.google.com/a/nd.edu/public-cvrl/data-sets.

[2] J Alabort-i Medina, E Antonakos, J Booth, P Snape, and S Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the ACM International Conference on Multimedia*, pages 679–682. ACM, 2014.

[3] B. Amberg, A. Blake, and T. Vetter. On compositional image alignment, with an application to active appearance models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[4] R Anderson, B Stenger, and R Cipolla. Using bounded diameter minimum spanning trees to build dense active appearance models. *International Journal of Computer Vision*, 110(1):48–57, 2014.

[5] E. Antonakos, J. Alabort i medina, G. Tzimiropoulos, and S. Zafeiriou. Feature-based lucas-kanade and active appearance models. *IEEE Transactions on Image Processing*, 24(9):2617–2632, 2015.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[7] A Asthana, S Zafeiriou, S Cheng, and M Pantic. Incremental face alignment in the wild. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[9] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

[10] Alphonse Bertillon. *La photographie judiciaire: avec un appendice sur la classification et l'identification anthropométriques*. Gauthier-Villars, 1890.

[11] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. In *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[12] Kyong Chang, Kevin W Bowyer, Sudeep Sarkar, and Barnabas Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1160–1165, 2003.

[13] Hui Chen and Bir Bhanu. Human ear recognition in 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):718–737, 2007.

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2001.

[15] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models: Their training and application. *Computer Vision and Image Understanding*, 1995.

[16] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 1, page 3, 2006.

[17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[18] Žiga Emeršič and Peter Peer. Ear biometric database in the wild. In *Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on*, pages 27–32. IEEE, 2015.

[19] Dariusz Frejlichowski and Natalia Tyszkiewicz. The west pomeranian university of technology ear database–a tool for testing biometric algorithms. In *Image analysis and recognition*, pages 227–234. Springer, 2010.

[20] M Grgic, K Delac, and S Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

[21] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[22] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[23] Alfred Victor Iannarelli. *Ear identification*. Paramont Publishing Company, 1989.

[24] R Imhofer. Die bedeutung der ohrmuschel für die feststellung der identität. *Archiv für die Kriminologie*, 26(150-163):3, 1906.

[25] Ajay Kumar and Chenye Wu. Automated human identification using ear imaging. *Pattern Recognition*, 45(3):956–968, 2012.

[26] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt. Eigen-pep for video face recognition. In *Asian Conference on Computer Vision*, pages 17–33. Springer, 2014.

[27] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[28] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 2004.

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[30] A Pflug and C Busch. Ear biometrics: a survey of detection, feature extraction and recognition methods. *Biometrics, IET*, 1(2):114–129, 2012.

[31] K Ramnath, S Baker, I Matthews, and D Raman. Increasing the density of active appearance models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[32] R Raposo, E Hoyle, A Peixinho, and H Proença. Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions. In *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2011 IEEE Workshop on*, pages 84–90. IEEE, 2011.

[33] J Saragih, S Lucey, and J Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision (IJCV)*, 2011.

[34] P Sermanet, D Eigen, X Zhang, M Mathieu, R Fergus, and Y LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[37] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[38] Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. Subspace learning from image gradient orientations. *IEEE transactions on pattern analysis and machine intelligence*, 34(12):2454–2466, 2012.

[39] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[40] X Zhu and D Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Conf. on Computer Vision and Pattern Recognition*, 2012.