A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2019,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 1

Multi-resolution Local Descriptor for 3D Ear Recognition

Iyyakutti Iyappan Ganapathi[1], Syed Sadaf Ali[2], Surya Prakash[3]

Abstract: Several approaches have shown promising results in human ear recognition. However, factors such as the pose, illumination, and scaling have an enormous impact on recognition performance. 3D models are insensitive to these factors and are found to be better at enhancing recognition performance with strong geometric information. Low cost 3D data acquisition has also boosted the research community in recent times to explore more about 3D object recognition. We present a local multi-resolution descriptor in this paper to recognize human ears in 3D. For each key-point in 3D ear, a local reference frame (LRF) is constructed. Using multi-radii, we find neighbors at each key-point and the neighbors obtained from each radius are projected to create a depth image using the LRF. Further, a descriptor is computed by employing neural network based auto-encoders and the local statistics of the depth images. The descriptor is used to locate the potential correspondence matching points in the probe and gallery images for a coarse arrangement, followed by a fine alignment to compute the registration error. The obtained registration error is used as the matching score. The proposed technique is assessed on UND-J2 dataset to demonstrate its effectiveness.

Keywords: 3D Biometrics, 3D Ear, Human Recognition, Authentication, Security, Local Descriptor, Hybrid Descriptor

1 Introduction

There have been numerous attempts in the literature to propose efficient descriptors for 2D and 3D shape recognition. Unlike 2D, 3D descriptors are still in a premature stage as they face various challenges [Gu13] in recognizing objects in a complex scene or in the presence of noise, occlusion, and varying mesh resolutions. In general, a local descriptor is a real vector of m-dimension, constructed using the neighbours' geometric information within a fixed radius for a selected key-point. At the time of recognition, a distance measurement is used to find the similarity between the real-valued descriptor vectors of model and test images to find out the correct match. In literature, existing 3D descriptors are unable to discriminate objects in applications such as biometric recognition, where comparisons involve extremely comparable shapes such as human ear and face [CB09].

We present a local feature descriptor for 3D ear recognition in this work. The descriptor is constructed using a neural network based auto-encoder and a derived local statistics. Further, the descriptor is used to find correspondence points in two 3D ear images. These points are used to align the ear images coarsely, followed by a fine alignment. The align-

[1] Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, phd1501101002 @ iiti.ac.in
[2] Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, phd1301101006 @ iiti.ac.in
[3] Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, surya @ iiti.ac.in

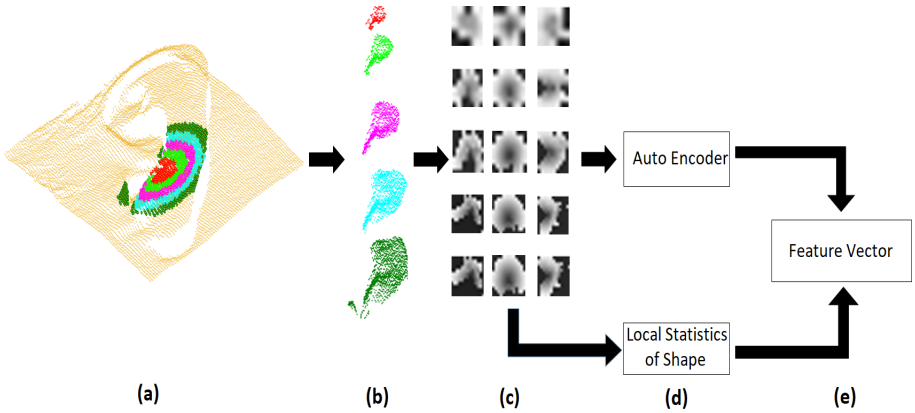


Fig. 1: Generation of 3D local feature vector using the proposed technique. (a) 3D Ear model and the neighbours obtained for five distinct radii are shown in different colors, (b) cropped local neighbours at the feature point for increasing radius from top to bottom, (c) the depth images obtained from the projections, where the first row from left to right shows the projections $D_x^{r_1}$, $D_y^{r_1}$ and $D_z^{r_1}$; and the second row shows the projection for the radius r_2, $D_x^{r_2}$, $D_y^{r_2}$ and $D_z^{r_2}$, (d) the depth images used as an input to auto-encoder and local statistic descriptor to generate the feature vector, (e) final feature vector obtained by concatenating all the vectors computed in (d)

ment error between the ear images is used as a similarity score. The overview of the proposed technique is shown in Fig. 1 and structured as follows. Section 2 reviews 3D ear recognition techniques and Section 3 discusses the proposed method. Section 4 demonstrates the experimental results followed by a detailed description of used ear database. Section 5 compares the performance of the proposed technique with the state-of-the-art techniques. Finally, the paper concluded in Section 6.

2 Related works

Most of the 3D object recognition methods available in literature work well to discriminate objects of distinct classes; however, they experience problems while differentiating two objects belonging to very similar classes, as in the case of biometric applications. Chen and Bhanu [CB09] utilized local surface patch (LSP) [CB07a] descriptor to find the corresponding matching points between two ear models. The similarity between two ear models was shown using the nearest neighbour classifier to reduce the high dimensional descriptor vector to low dimensional space. [ZCAM11] proposed a novel 3D ear authentication method based on holistic and local features. The histogram of indexed shapes (HIS) is extended to introduce the local features and voxelized ear models are used to generate the global features. Prakash and Gupta [PG14] developed a method for ear recognition utilizing co-registered 2D ear images of 3D ear images. The feature key-points from the 3D ear images are found by the co-registered 2D images. Ganapathi and Prakash [GP18] developed a technique for 3D ear recognition based on global and local features of the

ear. They proposed a global feature based on the geometry of the ear model, which was used along with representations of four well-known local 3D descriptors. Utilization of 3D along with 2D ear images to achieve ear recognition and detection was investigated by Islam et al. [Is11]. The neighbours of detected key-points were used to generate the local feature vector. These vectors are used to match the ear pairs.

3 Proposed method

Given a 3D ear pointcloud, $P \in \mathscr{R}^{m\times 3}$, for every point $p \in P$, a local reference frame is computed as follows: For each point p_i, we find neighbours $p_{11}, p_{12}, p_{13}....p_{1N}$ within a global radius R to form a scatter matrix $C \in \mathscr{R}^{3\times 3}$, is defined as $C_i = \sum_{j=1}^{N}(p_i - p_{1j})(p_i - p_{1j})^T w_j$, where w_j is the weightage given to each neighbour based on the distance between the feature point p_i and the neighbour p_{1j}. The matrix C is further decomposed into Eigen values $(\lambda_1, \lambda_2, \lambda_3,$ where $\lambda_1 \geq \lambda_2 \geq \lambda_3)$ and vectors (v_1, v_2, v_3). These vectors are used as the basis to create the local reference axis at each key-point. With respect to the obtained axis, we find local neighbours of key-point p_i for N distinct radius $\{r_1, r_2, r_3, r_4, ...r_N\}$. The neighbours correspond to each radius is separately projected to xy-,yz-, and zx planes of the respective local axis to create depth images $D_x^{r_i}$, $D_y^{r_i}$, $D_z^{r_i}$. These depth images are further used to compute the feature vectors. The depth image $D_x^{r_i}, i = (1,2,3,...,N)$ is the projected neighbours within the radius r_i onto xy- plane, where N is the number of distinct radius. For example, $D_x^{r_1}$ and $D_y^{r_1}$ are the depth images obtain from the neighbours at point p within the radius r_1 projected to xy- and yz- planes, respectively. The depth images obtained using multi-radii helps to encode the unique 3D geometrical information. The dimensions of the depth images generated from each stage is same as the radius r_i, i.e, $[r_i \times r_i]$. To generate the feature vector F two sub-feature vectors f_1 and f_2 are computed in two stages and concatenated to obtain the final vector. Also, the feature vector F is generated in a hybrid way, where the sub-feature vector f_1 is an auto-encoder based and the other sub-feature vector f_2 is based on a handcrafted method. To compute the sub-feature vector f_1, first, a compact representation of the depth image is obtained using an auto-encoder. In general, auto-encoder architectures have an input layer that is connected to hidden layers.

The desired number of hidden layers selected on the basis of the compact representation needed. After thorough testing using multiple networks, we chose a network with three hidden layers and a nonlinear sigmoid function. The architecture of the auto-encoder for our experiments has the following dimensions: Input layer(1024), layer1 [1024, 512], layer2 [512, 225], layer3 [225, 1024], output layer (1024). The architecture is shown in Fig. 2. The auto-encoder network is trained using randomly picked depth images obtained from the 3D ear samples of different subjects. Since, the depth images are of varying sizes, they are resized to a fixed dimension before being used for training. The network is trained offline without any manual annotation. Once the network is trained, the model is used to generate a reduced dimension of the depth image which is used as the sub-feature vector f_1. Few input depth images and reconstructed images using the auto-encoder are shown in Fig. 3. The input images (a - d) are obtained from a subject at different key-point locations for a fixed radius. The sub-feature vector of f_1 is computed using five different radius

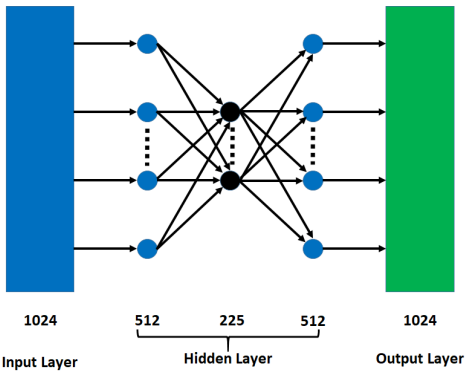


Fig. 2: Auto-encoder used in generating reduced feature vector. The input dimensions $\mathscr{R}^{1\times1024}$ is compactly reduced to lower dimensions $\mathscr{R}^{1\times225}$

($r_1 = 0.3\times R$, $r_2 = 0.6\times R$, $r_3 = 0.9\times R$, $r_4 = 1.2\times R$, $r_5 = 1.5\times R$) and is represent as $F_x^{r_{1..N}}$, $F_y^{r_{1..N}}$, $F_z^{r_{1..N}}$, where R is the global radius. The multi-level radii are represented as the fraction of this global radius. Throughout the experiment the global radius is chosen as 35. Here, $F_x^{r_{1..5}}$ is the reduced dimension representation of the depth image projected onto xy- plane for five distinct radius r_1, r_2, r_3, r_4, r_5. The depth image is represented as a vector of dimension 225. For example, a 3D Ear image with 100 key-points is represented by a matrix of size 100×225. Since, each key-point is generated using five different radius, three depth images corresponding to three projections are created for each radius. Therefore, a total of fifteen depth images are created for each key-point and concatenating the feature vectors of all the depth images leads to a vector of $1\times(15*225)$. To have a compact representation and to avoid the increase in dimensions, only a subset of these depth images are used in creating the sub-feature vector f_1. The selection process of subset of depth images used in generating the sub-feature is explained in Section 4.2. Next, we discuss the computation of other sub-feature vector f_2 using the depth variations of the depth images. A matrix DV is generated using the gray level co-occurrence, where the entry (i, j) in the matrix represent the joint probability of the pixel intensity i with a spatial relationship to another pixel intensity j.

Further, moments of the matrix is computed for the gray level distributions. A matrix can be represented completely using the central moments of all orders. Here, lower order moments are used to extract the information of the matrix DV. The sub-feature vector f_2 is constructed from the local statistics, the moment, μ_{mn} of the co-occurrence matrix DV is defined as: $\mu_{mn} = \sum_{i=1}^{L}\sum_{j=1}^{L}(i-\bar{i})^m(j-\bar{j})^n DV(i,j)$, where L is the depth level used in constructing the matrix DV, $\bar{i} = \sum_{i=1}^{L}\sum_{j=1}^{L} iDV(i,j)$ and $\bar{j} = \sum_{i=1}^{L}\sum_{j=1}^{L} jDV(i,j)$. To encode an image, the depth level also plays a key role. Choosing a large value or small value for L may not encode the distribution of the gray level intensities properly. Choosing a small value for L returns the feature vectors with less discrimination power and at the same time a large value results in encoding redundant information. After extensive experimentation, we have chosen $L = 6$ to create the distribution matrix. The feature f_2 is computed by concatenating $LS_x^{r_{1..N}}$, $LS_y^{r_{1..N}}$, $LS_z^{r_{1..N}}$, where $LS_x^{r_{1..N}}$ is the local shape descriptor obtained from

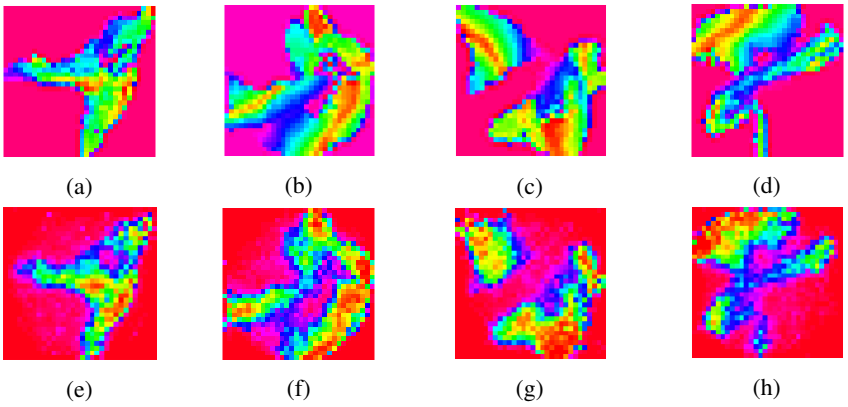
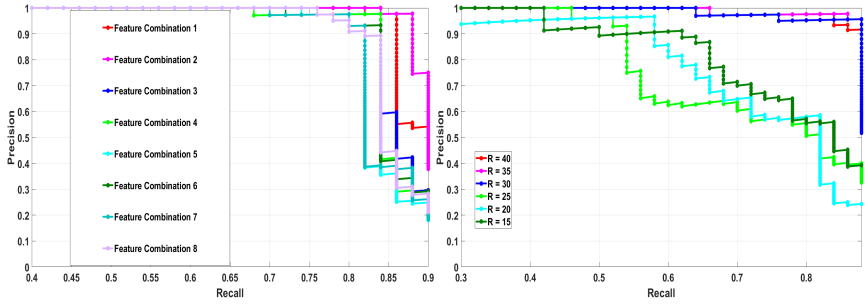


Fig. 3: Input images and reconstructed images obtained using autoencoder. (a)-(d) are the input images and (e)-(h) are the corresponding reconstructed images

the depth image projected onto xy- plane for N different radii. The final feature vector F is generated by concatenating the obtained sub-features f_1 and f_2.

4 Experimental analysis

This section presents experimental evaluations of the proposed technique. The following terms are defined to analyze the recognition performance. *FAR* (False Acceptance Rate) is the rate at which the recognition system accepts an unauthorized ear image and *FRR* (False Rejection Rate) is the rate at which the recognition system rejects an authorized ear image. *EER* (Equal Error Rate) is the measure of likelihood at which the *FAR* and *FRR* are equal. *ROC* Receiver Operating Characteristics is another important measure used to evaluate the recognition performance. It plots *FAR* with respect to GAR (defined as 100 - *FRR*). University of Notre Dame-Collection J2 (UND-J2) [YB07] dataset is used to demonstrate the effectiveness of the proposed technique. For the evaluation, two or more samples from 404 subjects are considered. A gallery dataset $G = \{E_{11}, E_{21}, ..E_{i1}...E_{n1}\}$ using n ear subjects, where E_{i1} is a sample of the i^{th} subject randomly selected from the database, and the probe dataset $P = \{E_{12..q_1}, E_{22..q_2}, ..E_{i2..q_i}...E_{n2..q_n}\}$, where $E_{i2..q_i}$ represents q_i samples of the subject i excluding E_{i1} are created for experimentation. The feature key-points in the experiments are chosen randomly points using uniform distribution. First, a gallery and probe dataset is created from the chosen 404 subjects with 1780 samples. A gallery dataset contains 404 arbitrarily chosen image from each subject and the probe contains 1376, the rest of the images. Second, the local feature vectors for all the images is computed using the proposed descriptor. The local descriptor matrix is of size $M \times N$ where M is the number of key-points in the image and N is the dimension of the feature descriptor vector. The first step of matching an ear image with the other is to find correspondence key-points which can be achieved by the nearest neighbour ratio matching. The nearest neighbour of the feature point of the probe image is decided by finding its corresponding feature point in the gallery image having the least Euclidean distance. Using these matching points, fine alignment is performed using ICP algorithm [BM92] and a matching score is calculated.



(a) (b)

Fig. 4: (a) Precision Vs. recall for feature subset selection. It is clearly seen that the feature combinations 1, 2, 3 and 8 shows better performance (b) Precision Vs. recall for six different radii. It is clearly seen that the radii $R = 40, 35, 30$ shows better performance compare to the other radii

ICP algorithm first computes the relative translation and rotation required to align the two point clouds and iteratively align both point cloud by reducing the matching error. Few examples of registration between the probe and gallery images using the correspondence point computed by the proposed technique are shown in Fig. 6 and the ROC is shown in Fig. 5(b).

4.1 Effects of radius on recognition

The radius includes global and local used are chosen experimentally. An extensive testing on a subset of subjects from the dataset for radii $[R = 15, 20, 25, 30, 35, 40]$ is performed and have chosen the best radius using the evaluation method, precision and recall. The Fig. 4b shows the precision Vs. recall for six different radii. For $R = 30, 35, 40$ the descriptor indicates better results and for lower values, the performance degrades. The reason for the best performance is that the radii $R = 30, 35, 40$ captured sufficient information to encode the feature vector. Moreover, the subjects present in the database have holes and choosing a smaller radius return very less information of the neighbourhood surrounding the feature key-point. Fig. 4b shows precision Vs. recall for different radius combinations. Once a global radius R is obtained, the multi-radii are chosen as a fraction of R. For example, a set of values $[0.1, 0.3, 0.5, 0.7, 0.9] \times R$ represents local five radii computed from the chosen R.

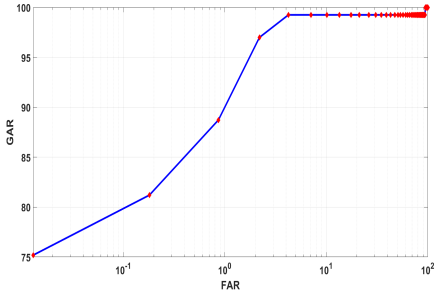
4.2 Effects of feature combinations on recognition

The concatenation of all computed vectors ${F_x}^{r_i}$, ${F_y}^{r_i}$, ${F_z}^{r_i}$ increases the size of the final vector. A subset of the computed features used to reduce the dimensions. For five varying radius $\{r_1, r_2, r_3, r_4, r_5\}$, eight combinations of features are carefully chosen and are summarized in Figure 5(a). The performance of the combinations No.1, No.2, No.3, No.4, and No.8 are superior than the other combinations. The fact behind the better performance is that the radii r_1, r_2 and r_4 are the combinations of small and large radius which helps in capture the information around the key-point in uneven and noisy data. Moreover, we observe that for the chosen database, the depth images obtained through xy and zx pro-

jections captures sufficient information, for example No. 8. So, the final feature vector is created using the combinations $[F_x{}^{r_4}, F_y{}^{r_4}, F_z{}^{r_4}, F_x{}^{r_2}, F_y{}^{r_2}, F_z{}^{r_2}, LS_x, LS_z]$. Fig. 4a shows the precision vs. recall for different combinations.

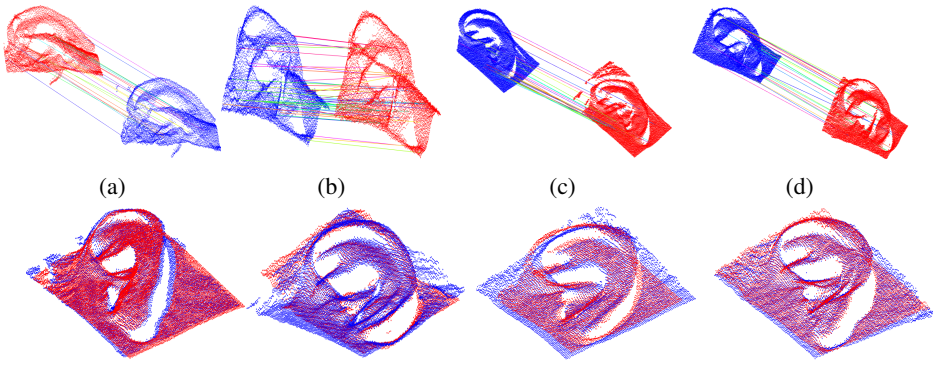
No.	combinations of the features
1	$F_x{}^{r_4}$, $F_y{}^{r_4}$, $F_z{}^{r_4}$
2	$F_x{}^{r_2}$, $F_y{}^{r_2}$, $F_z{}^{r_2}$
3	$F_x{}^{r_4}$, $F_z{}^{r_4}$, $F_x{}^{r_1}$, $F_z{}^{r_1}$
4	$F_x{}^{r_4}$, $F_y{}^{r_4}$, $F_z{}^{r_4}$
5	$F_x{}^{r_4}$, $F_z{}^{r_4}$, $F_x{}^{r_2}$, $F_z{}^{r_2}$
6	$F_x{}^{r_4}$, $F_y{}^{r_4}$, $F_z{}^{r_5}$
7	$F_x{}^{r_3}$, $F_z{}^{r_4}$, $F_z{}^{r_5}$
8	$F_x{}^{r_4}$, $F_z{}^{r_4}$

(a)



(b)

Fig. 5: (a) Different combinations of computed features (b) ROC curve for the proposed descriptor



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6: First row shows few examples of matching of correspondence points in probe (blue) and gallery (red) images using the proposed descriptor. Second row shows few examples of registration of probe (blue) and gallery (red) images. (a-d) shows the correspondence points match, and (e-h) shows the fine registration of probe and gallery using the initial transformation obtained from the correspondence points

5 Performance comparison

We have analyzed the performance of the proposed descriptor with the other well known 3D ear recognition techniques in the literature. The proposed technique has achieved a recognition rate of 98%, which is superior than the techniques [CB07b], [Is11], [Su14]. However, there are few techniques which has better performance than the proposed technique. The technique [ZCAM11] has achieved 98.6% on UND-G dataset which is comparatively smaller than the dataset UND-J2 used in our experimentation. In [PG14] and [YB07] the recognition rate is better than our technique, whereas the techniques needs co-registered 2D images to find the key-points for coarse and fine alignment.

6 Conclusion

In this paper, we presented a 3D descriptor technique for matching extremely comparable 3D objects such as the human ear. We performed experiments on the UND-J2 ear database by randomly choosing one 3D ear image of each subject as gallery and the remainder of the images as probe. It is observed that the matching performance of the descriptor obtained using the proposed technique is found to be at par with the other available techniques for ear matching. The major contribution of this article is the use of a hybrid approach to construct the 3D descriptor. The proposed technique used the automated generation of features and handcrafted features to develop the descriptor which has made the proposed technique robust to noise, small holes and occlusions.

References

[BM92] Besl, Paul J; McKay, Neil D: Method for registration of 3-D shapes. In: Sensor Fusion IV: Control Paradigms and Data Structures. volume 1611. International Society for Optics and Photonics, pp. 586–607, 1992.

[CB07a] Chen, Hui; Bhanu, Bir: 3D free-form object recognition in range images using local surface patches. Pattern Recognition Letters, 28(10):1252–1262, 2007.

[CB07b] Chen, Hui; Bhanu, Bir: Human ear recognition in 3D. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4):718–737, 2007.

[CB09] Chen, Hui; Bhanu, Bir: Efficient recognition of highly similar 3D objects in range images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1):172–179, 2009.

[GP18] Ganapathi, Iyyakutti Iyappan; Prakash, Surya: 3D ear recognition using global and local features. IET Biometrics, 7(3):232–241, 2018.

[Gu13] Guo, Yulan; Sohel, Ferdous; Bennamoun, Mohammed; Lu, Min; Wan, Jianwei: Rotational projection statistics for 3D local surface description and object recognition. International journal of computer vision, 105(1):63–86, 2013.

[Is11] Islam, Syed MS; Davies, Rowan; Bennamoun, Mohammed; Mian, Ajmal S: Efficient detection and recognition of 3D ears. International Journal of Computer Vision, 95(1):52–73, 2011.

[PG14] Prakash, Surya; Gupta, Phalguni: Human recognition using 3D ear images. Neurocomputing, 140:317–325, 2014.

[Su14] Sun, Xiaopeng; Wang, Guan; Wang, Lu; Sun, Hongyan; Wei, Xiaopeng: 3D ear recognition using local salience and principal manifold. Graphical Models, 76(5):402–412, 2014.

[YB07] Yan, Ping; Bowyer, Kevin W: Biometric recognition using 3D ear shape. IEEE Transactions on pattern analysis and machine intelligence, 29(8):1297–1308, 2007.

[ZCAM11] Zhou, Jindan; Cadavid, Steven; Abdel-Mottaleb, Mohamed: A computationally efficient approach to 3d ear recognition employing local and holistic features. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. IEEE, pp. 98–105, 2011.