EpNet: A Deep Neural Network for Ear Detection in 3D Point Clouds

Md Mursalin[1](✉) and Syed Mohammed Shamsul Islam[1,2]

[1] Edith Cowan University, Joondalup, WA 6027, Australia
{m.mursalin,syed.islam}@ecu.edu.au
[2] University of Western Australia, Crawley, WA 6009, Australia

Abstract. The human ear is full of distinctive features, and its rigidness to facial expressions and ageing has made it attractive to biometric research communities. Accurate and robust ear detection is one of the essential steps towards biometric systems, substantially affecting the efficiency of the entire identification system. Existing ear detection methods are prone to failure in the presence of typical day-to-day circumstances, such as partial occlusions due to hair or accessories, pose variations, and different lighting conditions. Recently, some researchers have proposed different state-of-the-art deep neural network architectures for ear detection in two-dimensional (2D) images. However, the ear detection directly from three-dimensional (3D) point clouds using deep neural networks is still an unexplored problem. In this work, we propose a deep neural network architecture named EpNet for 3D ear detection, which can detect ear directly from 3D point clouds. We also propose an automatic pipeline to annotate ears in the profile face images of UND J2 public data set. The experimental results on the public data show that our proposed method can be an effective solution for 3D ear detection.

Keywords: Ear biometric · Ear detection · 3D Pointcloud · Deep learning

1 Introduction

The ear is a magnificent organ of the human body that is generally used to detect, transmit and transduce sound. The outer shape of human ears contains distinguishing features that differ significantly among different subjects, even the ear of an identical twin is different from the other [1]. Researchers have shown that ear image analysis has numerous advantages over other biometric traits such as fingerprints, palmprints, iris, and face [2,3]. For instance, the acquisition technique is noninvasive, and the ear is not affected by expression variation. Furthermore, the structure of ears remains steady for a long age duration [4,5].

An essential task for ear biometrics is to detect ears from profile images. The 2D ear image-based techniques are regarded as the most popular for ear region localization as it involves less computations [6]. However, these 2D image-based techniques need to be performed in a constrained environment because

J. Blanc-Talon et al. (Eds.): ACIVS 2020, LNCS 12002, pp. 15–26, 2020.
https://doi.org/10.1007/978-3-030-40605-9_2

2D images are sensitive to changes in lighting conditions and pose variations. Furthermore, a 2D image can not differentiate between shapes and rotation angles. Recent developments in 3D imaging techniques overcome most of the limitations of 2D imaging [3]. Generally, a 3D scanner produces 3D data in the format of an unordered collection of points known as a point cloud.

The basic convolutional neural network (CNN) architectures require Euclidean structured input data formats such as multiview or 3D voxels for sharing weights and optimizing kernels. Since point clouds or meshes are a non-Euclidean data format, most of the work generally converts such data to Euclidean structured data before sending it to CNN architecture. Not only does representation conversion introduce unnecessarily voluminous data, but it also wraps natural invariances of the data, due to generation of quantization artefacts. To overcome this problem, we propose a deep neural architecture named EpNet, which is the modified version of the PointNet [7]. The EpNet is implemented directly on to 3D point clouds. An extensive review of the literature indicates that we are the first researchers applying a deep learning-based method for ear detection directly in 3D point clouds. The contribution of this work can be summarised as follows,

- A deep neural architecture for ear detection in 3D point clouds is proposed.
- A novel pipeline for automatic ear annotation from 3D profile images is introduced.

The rest of the paper is organized as follows: Sect. 2 briefly describes the related work, Sect. 3 explains the proposed methodology, Sect. 4 discusses the experimental results, and Sect. 5 draws the conclusion.

2 Related Work

Depending on the data type used, existing ear detection methods can be categorized as 2D, 3D, and the multimodal approach (using both 3D and the coregistered 2D image). An example of different data representation is shown in Fig. 1. In this paper, we mainly focus on 3D ear detection methods and explore the applicability of deep learning-based methods for ear detection.

Studies have been conducted for ear detection in profile image using 3D data. One of the pioneering work is presented by Chen et al. [8]. The authors extracted step edges from the 3D image and applied a modified iterative closest point (ICP) algorithm to detect helix and antihelix of the ear. However, their approach is sensitive to scale and pose variation. Zhou et al. [9] have proposed a method named histograms of categorized shapes (HCS) which used a 3D shape model combined with a support vector machine (SVM) classifier to detect ear from a 3D image. However, their approach fails to detect ear when prior knowledge about the given ear is not provided. Prakash et al. [10] introduced an edge connectivity graph to detect ear from 3D images. Resultingly, they were unable to handle the effect of off-plain rotation and performed poorly on the UND J2 dataset.

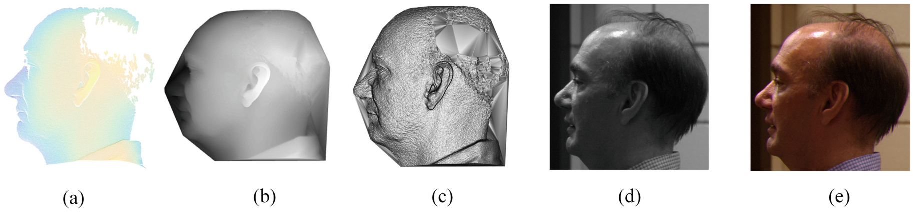

Fig. 1. Different representations of a sample profile image from UND J2 data set: (a) point cloud, (b) depth image, (c) 3D mesh, (d) 2D gray image, and (e) 2D colour image. (Color figure online)

A binarized mean curvature map-based method was presented by Pflug et al. [11] for detecting edges on 3D profile images. However, their algorithm failed to detect ear in the presence of occlusion. Lei et al. [12] proposed a novel ear tree-structured graph (ETG) method to detect ear from 3D profile images. Their method required manual annotations in the 2D depth images.

Recently, deep convolutional neural network-based approaches have been proposed for ear biometrics [13–17]. However, none of these methods used 3D images for ear detection.

3 Proposed Ear Detection Method

Our proposed ear detection method takes 3D point clouds as the input and outputs a set of 3D points that represents the ear location in the face image. Firstly, the input 3D point cloud is downsampled to N number of points. Each point has x, y, z coordinate values. The input dimension of each point is $I = C + P$, where C is the coordinate and P is the part id (here, face = 0 and ear = 1). We modify the PointNet network by eliminating some fully connected layers. The block diagram of our proposed method is shown in Fig. 2.

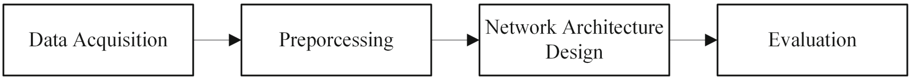

Fig. 2. Flowchart of our proposed method.

3.1 Data Acquisition

In this work, we collect 3D ear data named UND J2 from the University of Notre Dame [18]. This data set contains 1800 images from 415 different subjects with a resolution of 640 × 480. These include 681 images of 176 females and 1119 images of 239 males. All of these images were acquired using a laser scanner known as Minolta Vivid 900. The illumination conditions and poses are different among these images. The subjects are from different age groups with a variety of ethnic background. Additionally, some images contain occlusions with hair and earrings.

3.2 Preprocessing

Our preprocessing step consists of two parts: noise removing and downsampling. The raw profile face data of UND J2 data set contains noises that are removed using the median filter. A smoothing filter is then applied to eliminate the white noise.

The number of points in each image is 921,600. PointNet architecture computation depends on the number of points. Correspondingly, it is an essential step to downsample the data without losing the geometric shape of the object. We applied three different sampling techniques, including uniform box grid, non-uniform box grid and rand sampling technique. In this work, the non-uniform box grid filter is selected because it shows better sampling quality compared with the others. The sample output of different sampling technique is demonstrated in Fig. 3.

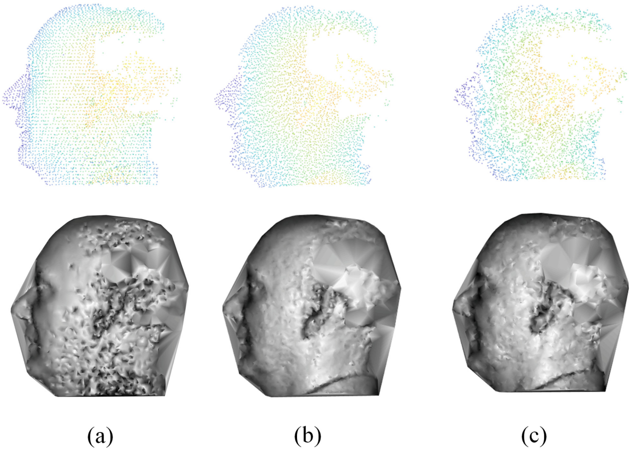

Fig. 3. Results of different downsampling techniques. Here, the first row is the point clouds and second row is the respective 3D mesh representation. Different techniques are: (a) box grid filter, (b) non-uniform box grid filter, and (c) random sampling.

3.3 EpNet Architecture

PointNet [7] is the first neural network that works directly on point clouds. The architecture of this network is simple, but it can efficiently extract discriminate features from input points. Firstly, the input points are passed through the transform network (T-net) and mapped into a feature vector. Then, a max pooling operator is used on this feature vectors to transform a permutation invariant global feature vector. Finally, the point feature vector and the global feature

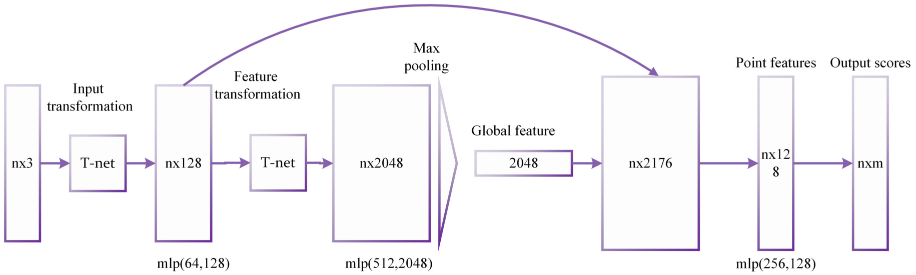

Fig. 4. The architecture of the proposed EpNet. T-net stands for transform network which makes the points invariant to permutation. Here, mlp represents the multi-layer perceptron and the number inside the mlp represents the number of layers.

vector are aggregated and mapped into an output vector using multilayer perceptron (MLP) networks. The output provides the scores for each point to a part id.

Our proposed EpNet architecture is based on PointNet part segmentation network [7]. The part segmentation network is designed for 16 different categories, where each category consists of numerous parts. The total number of parts for 16 categories are 50. In our ear detection problem, we have only one category which has only two parts: face (non-ear region) and ear. For this reason, we empirically decrease the number of MLP layers. So, our network is smaller compared with PointNet, which allows faster learning. In this work, we train the EpNet from scratch. The architecture of EpNet is shown in Fig. 4. The T-net stands for transform network, which offers the point sets invariant to permutation. We adopt the same transform network structure of PointNet. The final output shows the $n \times m$ scores for each of n points and m parts. In this paper, we apply the Adam optimizer [19] to train the EpNet.

3.4 Data Annotation

To the best of our knowledge, there is no annotated point cloud data for ear available in profile face images. As it requires significant controlled time—which increases the probability of operating errors related to operator visual fatigue, or susceptibility to distractions, or confusion during the annotating process—manual annotation is not viable for a large sample. Accordingly, we propose a novel technique to annotate the data automatically using the Basel face model (BFM) [20]. Firstly, we have generated 20,000 synthetic 3D faces using the BFM. The BFM was produced from geometric deformation of 100 male and 100 female faces. Thus, all of these generated images are different from each other. To make a similar view as the UND J2 data set, we then rotate the face point cloud by $-90°$ and delete the hidden points from the current viewpoint using the hidden point removal algorithm [21]. This procedure gives the left view of the synthetic image. All the generated images from BFM are hairless. So, we add random points to include hair occlusion in the 10,000 profile-faced point cloud. Next, we normalized the data to have values between 0 to 1. After normalization, we downsampled the data using non-uniform box grid filter sampling technique. The reason for downsampling is to reduce the computation for EpNet. Next, we labeled ears as the region of interest (ROI) in each 3D point cloud. Here, ground truth location of the ear region is known (because the face data is generated from the statistical model). Finally, we split the data into two groups: training and validation. We took 80% data for training and 20% data for validation. The number of epoch for training the network was 100 where the initial learning rate was 0.001, and the batch size was 12.

Algorithm 1: Fixing the over or under segmentation

```
Result: Annotated Ear
Calculate the difference ($diff$) between the initial ground truth and prediction
if $diff$ <15% then
    Increase the $y_{min}$ and $y_{max}$ using the mean value from correctly annotated
    ear;
end
if $diff$ >30% then
    Decrease the $y_{min}$ and $y_{max}$ using the mean value from correctly annotated
    ear;
end
```

We have tested all 3D images from the UND J2 data set on the trained model. As our problem of ear detection is a binary class (points belong to non-ear is 0 and points belong to ear is 1) segmentation problem, and the significant portion of the points belongs to the non-ear, so initially we annotate all points in the point cloud as a non-ear. Thus, our ground truth values are all 0. After the testing with all images, we calculated the difference between the prediction and the initial ground truth, where the prediction contains both 0 (non-ear) and 1 (ear), and the initial ground truth includes only 0. In our observation, we found that if the difference is between 15–30%, then the ear points are localized accurately. Apart from the correct localization, we have also seen some over-segmented (difference more than the 30%) and under-segmented (difference less than the 15%) images. In our experiment, we found that all of the incorrect segmentations are in y direction. To correct the under-segmentation, we increase the y_{min} and y_{max} of the ear region with the mean values from the correctly segmented images. The over-segmented images are fixed by decreasing the value of y_{min} and y_{max} according to the mean values from the correctly segmented images (see Algorithm 1). The whole pipeline of data annotation is shown in Fig. 5.

3.5 Evaluation Metrics

In this work, we report four different standard evaluation metrics, including accuracy, intersection over union, precision and recall for object detection. All of

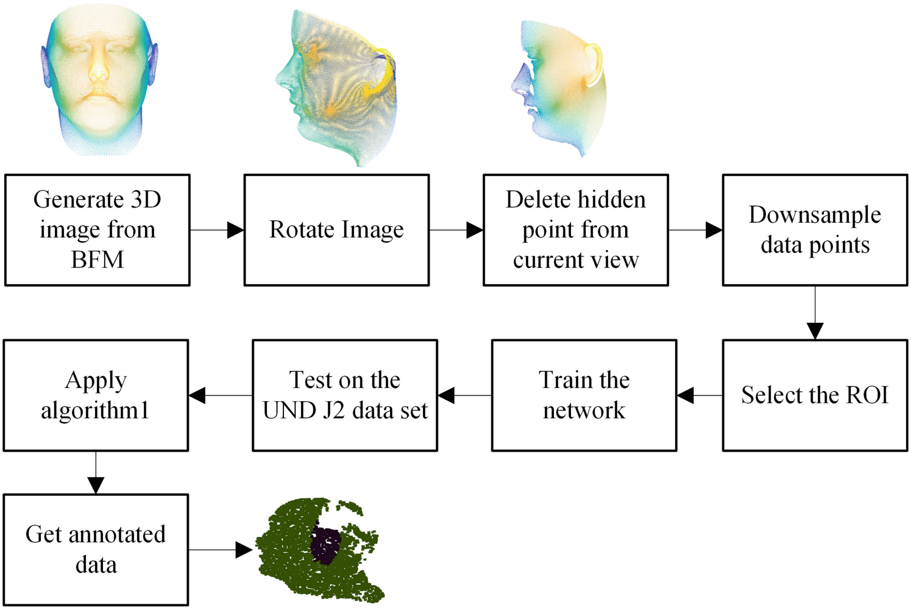

Fig. 5. Flow diagram of data annotation on the UND J2 data set.

these measurements are computed using the annotated ground truth. We use five different terms including true positive (TP), true negative (TN), false positive (FP), false negative (FN) and $Total\ points$. Here, TP represents the number of points correctly classified as part of an ear. TN represents the number points classified as non-ear points. FP represents the number of ear points classified as non-ear points. FN represents the number of non-ear points classified as ear points. $Total\ points$ represents the number of points exist in a given test image.

The first measurement accuracy is calculated using the following equation,

$$Accuracy = \frac{TP + TN}{Total\ points} \tag{1}$$

The accuracy measurement shows the segmentation quality. However, this result is mostly influenced by the non-ear points that cover the majority portion of test images. Thus, even if most points are categorized as belonging to the non-ear class, our accuracy measurement is shown to have high values.

The next measurement, intersection over union (IoU) is computed as follows,

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

The IoU shows the ratio between the number of points present in both the detected ear areas and the ground truth, and the number of points in the union of the detected and annotated ear areas.

The precision and recall is calculated using the following equations,

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The precision shows the percentage of correctly detected ear points and the ground truth ear points. This metric indicates how many ear points are identified from the actual ear points. The recall shows the percentage of correctly detected ear points from the predicted ear points. Detection accuracy is not defined generally. Different studies provide different measurement for calculating the detection rate. In this work, we used the IoU for defining the detection accuracy.

4 Results and Discussion

The input point cloud data is downsampled to 4096 points. This number is selected experimentally to retain the least visibility of the overall shape of point clouds. We choose 1100 data for training and 200 data for validation. For testing, we use 500 data which are not used during the training process. The network is trained using tensorflow where the number of the epoch is 100, and the batch size is 12. The test result is shown in Table 1.

Table 1. The average results on the UND J2 3D data set. The standard deviation is also given in each average results. All of these values are calculated using 500 test images.

Accuracy (%)	IoU (%)	Precision (%)	Recall (%)
93.09 ± 2.67	63.45 ± 11.14	75.56 ± 12.07	80.44 ± 10.30

Table 2. The average performance metrics between our method and the PED-CED [17]. The standard deviation is also given in each average results. Although the methods use different data set for evaluation, this comparison is to show the overall picture of the performance metrics.

Performance metrics	PED-CED [17]	Point-point (Ours)
IoU (%)	48.31 ± 23.01	63.45 ± 11.14
Precision (%)	60.83 ± 25.97	75.56 ± 12.07
Recall (%)	75.86 ± 33.11	80.44 ± 10.30

Table 3. Comparison of detection rate between our proposed method and other state-of-the-art methods.

Authors	Manual intervention	Detection accuracy (%)
Chen et al. [8]	No	92.6
Prakash et al. [10]	No	99.38
Yan et al. [18]	Yes	>97.6
Zhou et al. [9]	Yes	100
Pflug et al. [11]	Yes	95.65
Proposed method	No	100

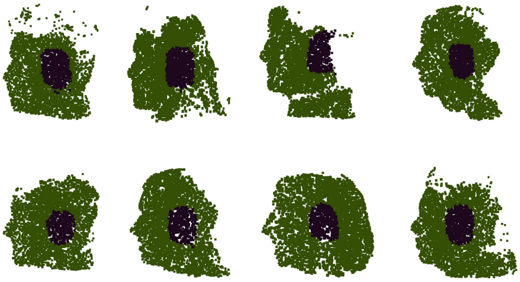

Fig. 6. Examples of our detection results on the UND J2 data set. (Color figure online)

The accuracy value is 93.09%, which is higher than other evaluation metrics, because both ear points, and non-ear points are contributed in the calculation. In the test image, the significant portion of the points as 81.16% are occupied by non-ear points while ear points occupy only 18.84% points. Accordingly, the accuracy metrics show the number of points is correctly detected in the given test image. However, it does not show the number of points that belong to ears as classified accurately (Table 2).

The IoU metrics show the actual performance of the detection, which is not affected by the points distribution of ear and non-ear classes. The false positive and false negative values have a direct impact on IoU measurement. The mean IoU is reported as 63.45%. This result can be improved by training with more diverse images because the typical errors occur due to the occlusions and pose variations. The reported precision in this work is 75.56%, which represents the percentage of detected points from the ground truth ear points. The recall value is 80.44%, that means 19.56% non-ear points are detected as ear points.

We also show some qualitative sample images of our test results in Fig. 6. Here, the blue colour represents the detected ear region. The qualitative results of the IoU values are illustrated in Fig. 7. Accordingly, the first column is the ground truth, the second column is the predicted results, and the third column is the difference between ground truth and prediction. The first row shows the output of the lowest IoU as 40.53%. We can see that most of the ear region is

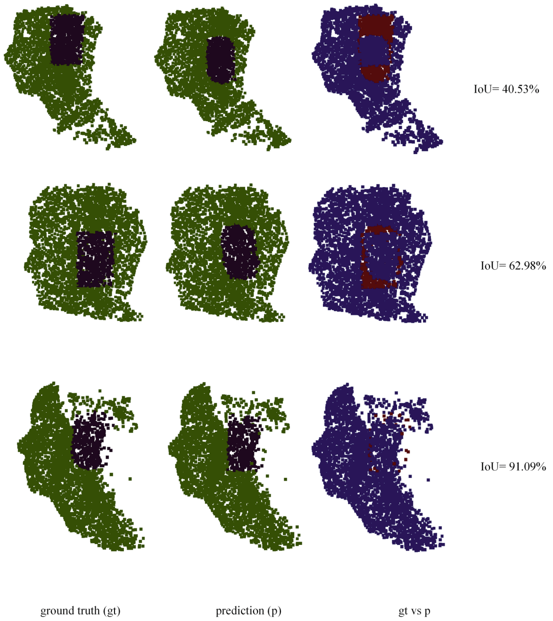

Fig. 7. The qualitative evaluation of our detection results using different IoU values. The first column is ground truth (gt), the second column is prediction (p), and the third column is the difference between ground truth and prediction. In the first two columns, the purple colour represents ear points, and the 3rd column red colour means the miss-match. (Color figure online)

detected accurately, where only the miss-match observed in the outer top and bottom area. The second row shows the output of IoU value as 62.98%. Here, the significant portion of the ear region detected correctly, although few miss-match is observed in the outer boundary region. The third row shows the output of the highest IoU as 91.09%, where most of the ear region is detected accurately. In this experiment, we have found that IoU of 40% can be used to detect the ear region. So, we consider the detection rate is 100% if the IoU is greater than or equal to 40%.

We also evaluate the performance between EpNet and PointNet++. The validation accuracy shows only 0.64% increase in case of PointNet++. However, the computation is higher in PointNet++, and it does not show significant improvement in our experiment.

It is essential to mention that in the existing literature, there is no standard evaluation method proposed for ear detection. Most of the work in this area reports different criteria for detection rate. Thus, direct comparison is not suggested as shown in Table 3. Here, we consider only those papers that used 3D images for ear detection.

5 Conclusion

In this paper, we present a method based on the deep neural network named EpNet for ear detection. To the best of our knowledge, this is the first study which applies the deep neural network-based approach for ear detection directly from 3D point cloud data. We have tested our trained model on the largest publicly available profile data set named UND J2. The experimental results show the effectiveness of our proposed EpNet for ear detection in 3D point cloud data.

References

1. Nejati, H., Zhang, L., Sim, T., Martinez-Marroquin, E., Dong, G.: Wonder ears: identification of identical twins from ear images. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1201–1204, November 2012
2. Burge, M., Burger, W.: Ear biometrics. In: Jain, A.K., Bolle, R., Pankanti, S. (eds.) Biometrics, pp. 273–285. Springer, Boston (1996). https://doi.org/10.1007/0-306-47044-6_13
3. Islam, S., Bennamoun, M., Owens, R.A., Davies, R.: A review of recent advances in 3D ear-and expression-invariant face biometrics. ACM Comput. Surv. (CSUR) **44**(3), 14 (2012)
4. Tiwari, S., Jain, S., Chandel, S.S., Kumar, S., Kumar, S.: Comparison of adult and newborn ear images for biometric recognition. In: Proceedings of 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 421–426, December 2016
5. Chen, H., Bhanu, B.: Contour matching for 3D ear recognition. In: 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1, vol. 1, pp. 123–128. IEEE (2005)
6. Emeršič, Ž., Štruc, V., Peer, P.: Ear recognition: more than a survey. Neurocomputing **255**, 26–39 (2017)
7. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)
8. Chen, H., Bhanu, B.: Shape model-based 3D ear detection from side face range images. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)-Workshops, pp. 122–122. IEEE (2005)
9. Zhou, J., Cadavid, S., Abdel-Mottaleb, M.: Histograms of categorized shapes for 3D ear detection. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6. IEEE (2010)
10. Prakash, S., Gupta, P.: An efficient technique for ear detection in 3D: invariant to rotation and scale. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 97–102. IEEE (2012)
11. Pflug, A., Winterstein, A., Busch, C.: Ear detection in 3D profile images based on surface curvature. In: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 1–6. IEEE (2012)
12. Lei, J., You, X., Abdel-Mottaleb, M.: Automatic ear landmark localization, segmentation, and pose classification in range images. IEEE Trans. Syst. Man Cybern.: Syst. **46**(2), 165–176 (2016)

13. Cintas, C., Delrieux, C., Navarro, P., Quinto-Sanchez, M., Pazos, B., Gonzalez-Jose, R.: Automatic ear detection and segmentation over partially occluded profile face images. J. Comput. Sci. Technol. **19**, 81–90 (2019)
14. Zhang, Y., Zhichun, M., Yuan, L., Chen, Y.: Ear verification under uncontrolled conditions with convolutional neural networks. IET Biometrics **7**(3), 185–198 (2018)
15. Wang, S., Du, Y., Huang, Z.: Ear detection using fully convolutional networks. In: Proceedings of the 2nd International Conference on Robotics, Control and Automation, pp. 50–55. ACM (2017)
16. Moniruzzaman, M.D., Islam, S.: Automatic ear detection using deep learning. In: Proceedings of the International Conference on Machine Learning and Data Engineering. iCMLDE2017 (2017)
17. Emeršič, Ž., Gabriel, L.L., Štruc, V., Peer, P.: Convolutional encoder-decoder networks for pixel-wise ear detection and segmentation. IET Biometrics **7**(3), 175–184 (2018)
18. Yan, P., Bowyer, K.W.: Biometric recognition using 3D ear shape. IEEE Trans. Pattern Anal. Mach. Intell. **29**(8), 1297–1308 (2007)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301. IEEE (2009)
21. Katz, S., Tal, A., Basri, R.: Direct visibility of point sets. In: ACM Transactions on Graphics (TOG), vol. 26, p. 24. ACM (2007)