Generation of 2D Ear Dataset with Annotated View Angles as a Basis for Angle-Aware Ear Recognition

Anja Hrovatič[1], Kihoon Kwon[2], Diego Sušanj[3], Peter Peer[1], Žiga Emeršič[1]

[1]*Faculty of Computer and Information Science, University of Ljubljana Večna pot 113, SI-1000 Ljubljana, EU*
[2]*School of Electronics Engineering, IT College, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Korea*
[3]*University of Rijeka, Faculty of Engineering, Vukovarska 58, 51000 Rijeka, Croatia*

E-mail: ah7651@student.uni-lj.si, kwons149@naver.com, dsusanj@riteh.hr, {peter.peer, ziga.emersic}@fri.uni-lj.si

Abstract

Ear recognition has seen steady development in the recent years. Despite numerous novel approaches ranging from traditional approaches based on local feature extraction to deep learning approaches, certain issues still remain unsolved. As pointed out in recent studies, one of the most prominent issues is the problem of ear alignment. To tackle this problem traditional approaches proved to be unsuccessful. However, in order to train deep neural networks to estimate pose angles to facilitate ear alignment, dataset with annotated angles is needed. In this work we present a 2D RGB dataset based on UND-J 3D dataset with corresponding 2D angle-annotated images as a base for convolutional neural network training.

1 Introduction

The popularity of ear recognition has been, and still is, increasing in the recent years. There has even been ear recognition competitions [1, 2], and many new deep-learning-based approaches proposed [1, 3, 4]. However, as emphasized in some of the recent surveys [5, 6] one of the most prominent issue still remains unaddressed - the problem of ear alignment. In real-life scenarios images of ears are taken in different positions, which causes difficulties during recognition stage.

Ear alignment has been addresses only partially [7, 8], where ear alignment was done either only in-plane (where only head pitch is considered) or to limited success in arbitrary rotations. We argue, that the most promising approach is to use convolutional neural network (CNN) to estimate in what angle the observed ear image was taken. When the angle is known we can use simple projective translation to transform ear image into a normalized one or train multiple models - one group for severe angle from one side, one for fully profile ear images and one for severe angle from the other side.

However, in order to achieve that, a new dataset, with specifically annotated angles, is needed. In this paper we focus on generating this new dataset of 2D RGB and corresponding depth images of subjects' left profile faces with applied roll, pitch, yaw angles and small positional perturbation in the form of random movement or defined translation. This dataset will serve as a basis for future research and development of ear-alignment-aware recognition.

We based our approach on the University of Notre Dame collection J2 database (UND-J2). The dataset consists of 3D point cloud data and corresponding 2D RGB images. 3D point cloud range scans were used to generate a colored point cloud in PLY and PCD file formats, which we then further used for transformation of 3D point coordinates. Next we applied roll, pitch and yaw rotation and added translation transformation to include small random movements. Transformed point coordinates were then projected from three-dimensional world to a two-dimensional image plane with the use of the pinhole camera model and camera's intrinsic parameters. The projected coordinates (u, v) were used to generate 2D RGB images. Depth images, however, were generated based on the value of Z coordinate of the 3D projected vector.

Generated images with this method can be used to fill in the missing examples from the datasets and make databases larger with a wider range of pose variation. This can be seen in the UND-J2 dataset as well. For some subjects there are only a couple of example images, which may present itself as a problem for certain ear detection or recognition approaches that require larger amount of images.

The paper is structured as follows. In section Section 2 we present the related work in this field. Section 3 includes and presents the process of our approach, from preparing the UND-J2 dataset to applying rotation, translation and perspective projection to 2D world and mapping projected points to 2D RGB and depth imagesand provide some final remarks in Section 4.

2 Related work

Preparation and generation of databases is a vital process in computer vision for successful evaluation of unsupervised approaches, as well as key for supervised methods that depend upon large databases due to their learning nature.

In the field of ear biometrics several databases already exist, such as IITD Ear Dataset [9] and WPUT Ear Dataset [10], many of which contain data captured in controlled and uncontrolled environments and are therefore not a challenging task for ear detection algorithms. There are, however, some databases that include examples with different postures and angles. Such are the USTB dataset, that provides images at different angles

and YSU database, which includes images with poses from 0 to 90 degrees [11].

The first database consisting of ear images in the wild, which includes images of famous people taken from the internet, was presented in [12]. The dataset was shown to be the most challenging so far, as various illumination conditions, angles and the presence of hair and ear accessories contribute to that.

However, none of these datasets contain annotated angles. To tackle this issue we propose a new dataset. We used 3D point cloud data captured with 3D range scanner. However, a challenge of 3D sensors and 3D point cloud data is that missing data can occur. In [13] they addressed this problem by applying median filter to the original 3D point cloud with holes and median filter to rotated output data.

3 Preparation of a generated 2D image dataset of ears

In the sections below we present the techniques used to obtain and generate 2D RGB and depth image database of ears with applied roll, pitch and yaw rotation and small random movement or defined translation. First, we present the UND-J2 database [1], used for generating a new dataset and how we used it in our work. Then we describe the techniques used for the transformation of data and generation of images.

3.1 Preparation of UND-J2 dataset

The basis of our work was the UND-J2 database. The dataset consists of 2436 left profile face 3D range scans of 415 subjects, taken with Minolta Vivid 910 3D laser scanner. Dataset includes 2413 corresponding 2D RGB images of 640×480 pixels for each subject. The data was captured in controlled, laboratory-like, conditions but in different lightning conditions and poses. Some images include occlusions such as earrings and hair.

First we performed data correspondence check between the 3D point cloud data and 2D RGB images in the database. The check concluded in 404 different subjects with at least two example sets for each.

Based on the 3D point cloud data and corresponding RGB images, we merged RGB components from 2D images with 3D point cloud data and examined the visual correspondence as well. The visual correspondence examination resulted in: 3 examples of which 2D RGB image did not match with the 3D point cloud data, 75 examples with big offsets between RGB images and 3D point cloud data and 221 examples, of which RGB images and 3D point cloud had small offset. Mentioned offsets are the result of the process of capturing images, the corresponding 3D and 2D data were taken at nearly the same time, which caused the subjects to move causing small or big offsets in data. Even though using examples with large offsets would result in an imperfect projection from 3D world to 2D plane, we still decided to use the given examples and generated as large database as possible and further evaluated the obtained results.

Since given 3D point cloud data in UND-J2 database are unstructured and in raw format, we transformed the data into Polygon (PLY) and Point Cloud Data (PCD) file formats. We generated PLY and PCD files from 3D points alongside with RGB components taken from corresponding 2D RGB images and created colored 3D point clouds in mentioned formats. Converting the files into PLY and PCD file formats enabled us more efficient further transformation of data, which is described more thoroughly in the following sections.

3.2 3D rotation and projection

In this section we present the applied approach of mapping 3D coordinates to a 2D plane, with given roll, pitch and yaw rotation angles and translation in x, y, z direction, to create the generated 2D RGB and depth image database.

We performed a perspective projection on the colored 3D point cloud data with the use of the pinhole camera model [14], illustrated in image 1. The pinhole camera model models perspective projections and is the basic camera model used in computer vision [14]. The mapping of the pinhole camera, from 3D points in 3D world to 2D points in an image plane, is described with a camera matrix, also known as the camera projection matrix [15], given in equation (5).

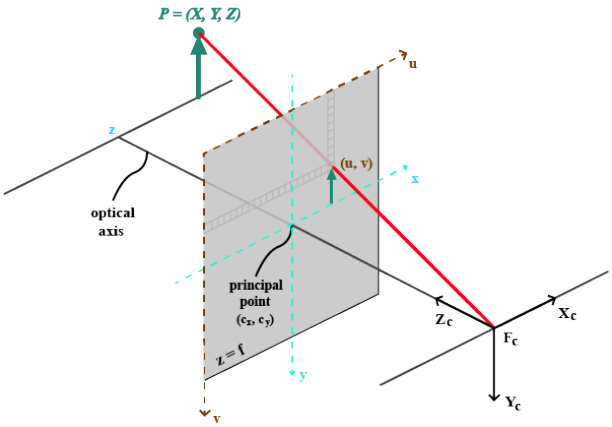


Figure 1: Illustration of the pinhole camera model, where F_c represents the focal length defined in X (F_x), Y (F_y) and Z (F_z) directions and (u, v) are the coordinates of the point P being projected.

We performed the transformation of colored 3D point cloud as follows. First we computed the mean of each X, Y, Z coordinate in 3D world. The mean of each coordinate corresponds to a reference point by which we translated the given 3D (X, Y, Z) point to align the camera and world coordinate systems. Next we applied rotation in X, Y and Z direction - where X corresponds to roll, Y to pitch and Z to yaw rotation. And then applied translation, in X, Y and Z directions, to given 3D point. We computed the overall rotation by matrix multiplication [16] of matrices shown in equations (1), (2) and (3).

[1] https://cvrl.nd.edu/projects/data/#nd-collection-j2

The whole transformation of the 3D point is computed as shown in equation (4).

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \quad (1)$$

$$R_y = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \quad (2)$$

$$R_z = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} x_{rt} \\ y_{rt} \\ z_{rt} \end{bmatrix} = R_z R_y R_x \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} \quad (4)$$

After applied transformation we project the 3D point by multiplying it with camera's intrinsic parameters, given by the camera projection matrix. The projection is shown in equation (5). Camera's intrinsic parameters are the focal lengths, f_x and f_y, expressed in pixels [16]. In our case we defined f_x as 640 and f_y as 480. Parameters c_x and c_y represent the principal point which corresponds to the image center. In our case we applied some proportional scaling with parameters s_x and s_y, by which we defined the scale factors in X and Y directions. The final projection of points to 2D image plane was computed with equations (6) and (7), where each X and Y coordinate of a projected 3D point is divided by the Z value of the projected point in question, giving us the final projected u and v coordinates in a 2D image plane.

$$\begin{bmatrix} x_p \\ y_p \\ z_p \end{bmatrix} = \begin{bmatrix} f_x/s_x & 0 & c_x & 0 \\ 0 & f_y/s_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_r \\ y_r \\ z_r \\ 1 \end{bmatrix} \quad (5)$$

$$u = x_p/z_p \quad (6)$$

$$v = y_p/z_p \quad (7)$$

3.3 Generating 2D RGB and depth images

The perspective projection, described in the previous section, has given us the u, v and z coordinates of the projected 3D point. Coordinates u and v correspond to the points in the 2D image plane and z corresponds to the transformed and projected 3D point before applying final projection into 2D world.

We generated 2D RGB images (2, 3) based on the color information taken from acquired PLY and PCD files, described in Section 3. Images are obtained by applying RGB components to projected u and v coordinates in a 2D image.

Based on the transformed and projected z coordinate, we generated depth images, also known as depth maps. A depth map or image contains information about the distance of the surfaces of objects in the image from a camera viewpoint. We produced depth maps with luminance

(a) RGB image with 20° pitch angle.

(b) RGB image with 15° roll angle.

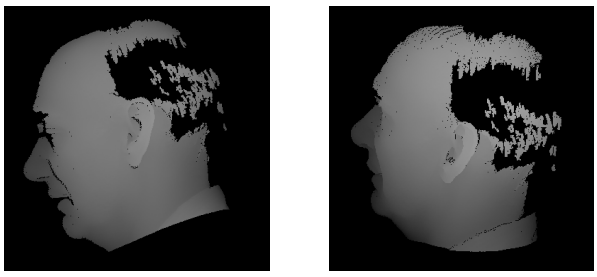
Figure 2: Generated RGB images with different roll and pitch angles.

(a) RGB image with 5° roll, 25° pitch and 10° yaw angle.

(b) RGB image with 25° roll, 15° pitch and 35° yaw angle.

Figure 3: Generated RGB images with different roll, pitch and yaw angles.

proportional to the distance to the camera. Depth images, as can be seen on 4 and 5, were produced such that surfaces closer to the camera are brighter and surfaces with greater distance to the camera are darker.

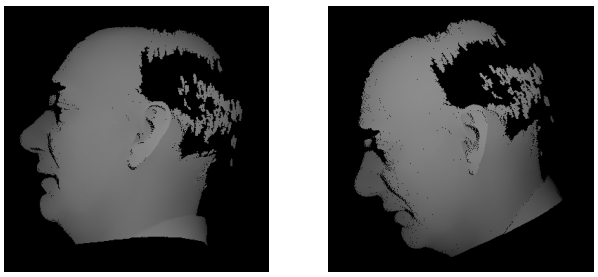


(a) Depth image with 5° roll, 10° pitch and 20° yaw angle.

(b) Depth image with 20° roll angle.

Figure 4: Generated depth images with different roll, pitch and yaw angles.

The applied transformations, in the process of generating RGB and depth images, resulted in holes in images - missing data. Reasons for such lack of data are firstly, holes in the original 3D point cloud caused by the 3D sensor due to illumination conditions or subject's oily skin. And secondly, the applied rotation transformation to the 3D point cloud. Rotation causes the intervals between the X and Y neighbourhoods to become distorted, especially when the point of view is changed significantly [13]. The



(a) Depth image with 15° pitch angle.

(b) Depth image with 10° roll, 25° pitch and 30° yaw angle.

Figure 5: Generated depth images with different roll, pitch and yaw angles.

latter mostly refers to applying pitch and yaw rotations to point clouds. Since the holes in the generated images would pose a problem in further use of the database, we decided to apply a hole filling process to the projected points. We applied a flood-fill operation on the background pixels of the input binary image.

4 Conclusion

In this paper we, not only present a novel dataset derived from UND-J2, but also present the process of preparing and generating a new database, consisting of 2D RGB and depth images, based on 3D point cloud data and corresponding 2D RGB images from the UND-J2 dataset. We applied a perspective transformation, 3D projection from 3D world to 2D image plane, with the use of the pinhole camera model and mapped projected coordinates to 2D RGB and depth images.

Some problems were encountered, such as missing data, in the form of holes, in the 3D point cloud, imperfect data correspondence between 3D point clouds and 2D RGB images, causing offsets and impacting data reconstruction process. Another important factor in the perspective projection process is correctly defining the camera's intrinsic parameters being applied to the camera projection matrix. This has a vital role in the accuracy of the projection as well.

The final application for the dataset is CNN-based angle estimation. This enables either separate-model-based prediction or serves as a base for deep-neural-network-based ear alignment. We hope that our work will serve as a good basis to solve the problem of ear alignment - an issue that after many years still remained unsolved.

References

[1] Ž. Emeršič, D. Štepec, V. Štruc, P. Peer, A. George, A. Ahmad, E. Omar, T. E. Boult, R. Safdari, Y. Zhou, S. Zafeiriou, D. Yaman, F. I. Eyiokur, and H. K. Ekenel, "The Unconstrained Ear Recognition Challenge," in *International Joint Conference on Biometrics*. IEEE/IAPR, 2017, pp. 715–724.

[2] B. S. H. W. G. J. N. K. A. P. E. H. M. P. S. S. S. H. P. G. P. N. I.-J. K. S. G. S. K. M. K. L. Y. J. Y. H. Z. F. L. J. M. X. Z. D. Y. F. I. E. K. B. H. K. E. D. P. C. S. B. P. K. S. B. M. P. P. V. Žiga Emeršič, Aruna Kumar S. V., "The unconstrained ear recognition challenge 2019," in *International Conference On Biometrics*. IAPR, 2019.

[3] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear verification under uncontrolled conditions with convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 185–198, Jan. 2018.

[4] E. E. Hansley, M. P. Segundo, and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215–223, Jan. 2018.

[5] A. Pflug and C. Busch, "Ear biometrics: A survey of detection, feature extraction and recognition methods," *IET Biometrics*, vol. 1, no. 2, pp. 114–129, 2012.

[6] Ž. Emeršič, V. Štruc, and P. Peer, "Ear Recognition: More Than a Survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[7] A. Pflug and C. Busch, "Segmentation and normalization of human ears using cascaded pose regression," in *Secure IT Systems*. Springer, 2014, pp. 261–272.

[8] M. Ribič, Ž. Emeršič, V. Štruc, and P. Peer, "Influence of Alignment on Ear Recognition: Case Study on AWE Dataset," in *International Electrotechnical and Computer Science Conference*, vol. 25-B. IEEE, 2016, pp. 131–134.

[9] "Automated human identification using ear imaging," http://www4.comp.polyu.edu.hk/˜csajaykr/IITD/Database_Ear.htm.

[10] D. Frejlichowski and N. Tyszkiewicz, "The West Pomeranian University of Technology Ear Database – A Tool for Testing Biometric Algorithms," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, A. Campilho and M. Kamel, Eds. Springer Berlin Heidelberg, 2010, vol. 6112, pp. 227–234.

[11] T. V. Kandgaonkar, R. S. Mente, A. R. Shinde, and S. D. Raut, "Ear biometrics: A survey on ear image databases and techniques for ear detection and recognition," *IBMRD's Journal of Management & Research*, vol. 4, no. 1, pp. 88–103, 2015.

[12] Ž. Emeršič and P. Peer, "Ear biometric database in the wild," in *2015 4th international work conference on bioinspired intelligence (IWOBI)*. IEEE, 2015, pp. 27–32.

[13] P. Yan and K. W. Bowyer, "Ear biometrics using 2d and 3d images," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 2005, pp. 121–121.

[14] P. Sturm, "Pinhole camera model," *Computer Vision: A Reference Guide*, pp. 610–613, 2014.

[15] H. G. Pharr, M., "Physically based rendering: From theory to implementation," 2004.

[16] P. I. Dunn, F., "3d math primer for graphics and game development. 2nd ed." pp. 217–275, 2011.