DETECTION, LOCALIZATION AND POSE CLASSIFICATION OF EAR IN 3D FACE PROFILE IMAGES

Jiajia Lei[1,2], *Jindan Zhou*[2,3], *Mohamed Abdel-Mottaleb*[2] *and Xinge You*[1]

[1]Huazhong University of Science and Technology
[2]University of Miami
[3]Nanjing University of Finance and Economics

ABSTRACT

We present an efficient and robust system for landmark localization, segmentation and pose classification of ears from 3D profile facial range data. After defining 18 landmarks on the ear, including Triangular Fossa and Incisure Intertragica, a novel Ear Tree-structured Graph (ETG) is proposed to represent the 3D ear. We trained a flexible mixture model to locate these landmarks automatically. Afterwards, the ear region is outlined as the minimum rectangle including all landmarks. Finally, by calculating the turning angle between landmarks on the helix, the ear is classified as either a left or a right ear. To the best of our knowledge, there is no previous work on automatic landmark localization for 3D ear on 3D facial profile depth images. Experiments are conducted on University of Notre Dame Collection F and Collection J2 datasets, containing large occlusion, scale and pose variations. Results demonstrate the effectiveness of the proposed techniques.

Index Terms— Landmark localization, ear detection, pose classification, ear tree-structured graph, flexible mixture model

1. INTRODUCTION

Biometrics are widely used in many applications ranging from border crossings to daily activities such as shopping by internet. Ear recognition has emerged as an active research area in recent years and has various desirable properties [1, 2, 3]. The ear can be easily captured from a distance without full cooperation of subjects; it has a rich and stable structure that is invariant to age and facial expressions. The performance of techniques based on 2D intensity ear images is greatly affected by imaging conditions. On the contrary, range images are relatively insensitive to variations in imaging conditions such as illumination. Extensive surveys [1, 2, 4, 5], show that 3D ear shape matching has better performance. At the same time, 3D imaging systems are becoming cheaper and feasible for daily applications (e.g., the Kinect). Thus, 3D ear recognition is receiving increased attention.

Ear segmentation and landmark localization from facial profile images are essential for ear recognition and ear based gender classification. In spite of the ear's rich shape and structure features, it is still challenging for existing methods to perform in unconstrained environments due to changes in viewpoints, self-occlusion and the occlusion from hair and jewelery (e.g., ear-rings).

Most of the 3D ear recognition methods employ manual ear segmentation or use detection results of their registered 2D color images [1, 2, 6]. In [7], skin area is isolated and edges from registered 2D color images and range images are combined to locate candidate helix and antihelix lines. Then a reference 3D ear shape model, which contains a set of discrete 3D vertices on the typical helix and antihelix parts, is fitted to these lines to locate the ear at the region that has minimum mean registration error. In [8], manually labeled landmark points viz. Triangular Fossa and Incisure Intertragica are employed to crop the ear data. Then, ears are aligned and normalized according to the line connecting the two landmarks. Yan and Browyer [9] presented a detection method by locating the ear pit automatically and using active contour algorithm on both corresponding color and depth profile images to outline the ear region. In [10], Zhou et al. trained a 3D shape model based on a shape-based feature set, termed as Histograms of Categorized Shapes (HCS), for robust 3D ear detection. However, they used more than half of subjects and images for training and the performance evaluation has been done on a relatively small data set. Islam et al. [11] used Haar-wavelets along with AdaBoost for ear segmentation from 2D profile face images and their corresponding 3D ear data. In [12] the authors proposed an edge connectivity graph for ear detection from 3D range data. They used discontinuities in the depth map for extracting the initial edge image and then extracted the connectivity graph. Their approach is not sensitive to scale and in-plane rotation. However, the more challenging task of detecting ears with off-plain rotations was solved. They discarded some images in the UND-J2 dataset, showing that their method could not work properly on data with very poor quality.

It should be noted that these techniques, except [8, 12], do not solve the ear detection problem in the presence of scale and pose (rotation) changes. Also, they are not able to detect and tell whether the detected ear is a left ear or is a right ear, which is important for pose estimation, tracking and recognition of ears in real-life images. In this paper, we present an efficient and effective approach for this purpose.

Localizing landmarks on 3D ears is a pre-task for ear detection and alignment. Although there are many ear detection and recognition algorithms in literature [7, 8, 10, 11, 12, 13], none of them discussed the problem of automatically detecting landmarks on the ear. Usually, the common approach they use is to locate the landmarks manually, as in [8].

This paper defines some new landmarks on the ear in addition to Triangular Fossa and Incisure Intertragica, and propose a novel technique to localize these landmarks automatically on the 3D depth data of the profile face. The ear region is segmented based on detected landmarks and classified to either left or right ear. Firstly, a tree-structured graph was proposed to represent the 3D ear. Then we trained a flexible mixture model [14] along with latent Support Vector Machine to localize these landmarks. Then, the ear region is outlined as the minimum rectangle including all landmarks. Finally, by calculating the turning angle between landmarks on helix, the ear is classified to either left or right ear.

2. LANDMARK LOCALIZATION AND EAR DETECTION

Ear helix and anti-helix are the most discriminative parts of the ear. They are always visible in different viewpoints and under occlusion of earphone or ear rings. We define 18 landmarks on the helix and anti-helix of the ear to capture the structure of the ear. The tree-structured graph of the ear is shown in Figure 1. Landmark 1 is the root of the tree graph. Note that there is an edge between landmarks 11 and 18, not 11 and 16, to avoid closed loops in the proposed tree graph. The advantage of adding landmark 18 is that more geometric constraints are added to landmark 11, bringing about better landmark localization accuracy.

The tree-structured graph is represented as $G = (V, E)$ for an M-node graph, whose nodes specify the landmarks and its edges indicate that neighboring landmarks on the tree have spatial constraints. Given an ear image I, $l_i = (x_i, y_i), i = 1, ...M$ for the landmark locations, we write $r_i = 1, ..., K$ as the orientation type of the landmark with respect to its parent in the tree. To score of a configuration of landmarks, we define a score function

$$S(I,L,R)=\sum_{i\in V}\lambda_i^{r_i}\cdot f(I,l_i)+\sum_{ij\in E}\mu_{ij}^{r_ir_j}\cdot\varphi(l_i,l_j)+\sum_{ij\in E}c_{ij}^{r_ir_j} \qquad (1)$$

where $f(I, l_i)$ is a feature representation of landmark l_i in ear image I, $\lambda_i^{r_i}$ is a template for landmark l_i with orientation r_i w.r.t. its parent, par_i, $\varphi(l_i, l_j)$ is the spatial feature between l_i and l_j (e.g. the squared offset between two landmarks), and $\mu_{ij}^{r_ir_j}$ is the parameter that favors certain offsets over others. The first term in Eq. 1 is a patch model that describes what each patch around landmarks look like. The second term is a deformation model that controls the relative placement of landmark l_i and its parent par_i. $c_{ij}^{r_ir_j}$ indicates pairwise co-occurrence prior between landmark l_i with orientation r_i and landmark l_j with orientation r_j. The third term is a co-occurrence model that favors certain pairs of orientations between landmarks. In our experiments, we use Histogram of Gradients (HOG) to represent the ear landmark patch $f(I, l_i)$ and define $\varphi(l_i, l_j)$ as the squared offset between two landmark locations, i.e., $\varphi(l_i, l_j) = [dx\ dx^2\ dy\ dy^2]^T$ where $dx = x_i - x_j$ and $dy = y_i - y_j$.

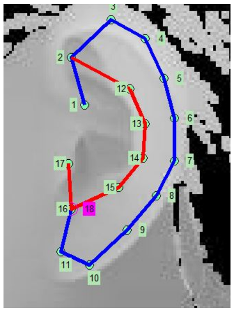

Fig. 1. The tree-structured graph for an ear. We used 18 landmarks to obtain the main structure of the ear. The root of the graph is landmark 1, and landmarks 17 and 18 are the leaves. Landmark 2 is Triangular Fossa and landmark 16 is Incisure Intertragica. Landmark 18 (marked as magenta circle) is at the same position as landmark 16.

Inference: Inference is employed to maximize $S(I, L, R)$ over landmarks L and orientation types R. Since we used a tree-structured graph, it can be done efficiently with dynamic programming. Given a tree-structured graph $G = (V, E)$, we denote $kids(i)$ as the set of children of landmark l_i and l_j as its parent. For a tree graph, a landmark l_i has only one parent. The message that landmark l_i passes to its parent l_j can be computed as follows:

$$\begin{aligned} score_i(l_i, r_i) &= \lambda_i^{r_i}\cdot f(I,l_i)+\sum_{k\in kids(i)} m_k(l_i,r_i) && (2)\\ m_i(l_j,r_j) &= \max_{r_i} c_{ij}^{r_ir_j} + \\ &\quad \max_{l_i} score(l_i,r_i)+\mu_{ij}^{r_ir_j}\cdot\varphi(l_i,l_j) && (3)\end{aligned}$$

With messages collected from the children of l_i, Eq. 2 calculates the local score of landmark l_i at all possible locations and for all orientation types r_i. Eq. 3 computes for every location and orientation type of landmark l_j (parent of l_i), the best scoring location and orientation type of landmark l_i. Once all messages pass to the root landmark l_1, $score_1(l_1, r_1)$ describes the best scoring cofiguration for each root landmark location and orientation type. Then we use a threshold to filter these root scores, obtaining multiple detections. By keeping track of the $argmax$ indices, we can find the location of the landmarks for each detection. Finally, non-maximum suppression (NMS) method is used to fuse overlapping detections.

Learning: We assume a fully supervised training dataset that has labeled positive examples $\{I_t, L_t, R_t\}$ and negative examples $\{I_t\}$. Since the scoring function, Eq. 1, is linear in its parameters, we write $S(I, \kappa) = \omega \cdot \Psi(I, \kappa)$, where $\omega = (\lambda, \mu, c)$ and $\kappa = (l_t, r_t)$. The optimization function is as follows.

$$argmin_{\omega} \frac{1}{2} \|\omega\| + C \sum_t \xi_t \tag{4}$$

$$s.t.\ \forall n \in pos,\ \ \omega \cdot \Psi(I_t, \kappa_t) \geq 1 - \xi_t$$

$$\forall n \in neg, \forall \kappa\ \ \omega \cdot \Psi(I_t, \kappa) \leq -1 + \xi_t$$

The above form of learning problem can be solved using a structural SVM. We use the optimization procedure in [14]. Fig. 2 shows some results of landmark localization. The detection results are minimum bounding rectangles containing the landmark patches. More results will be presented in Section 4.

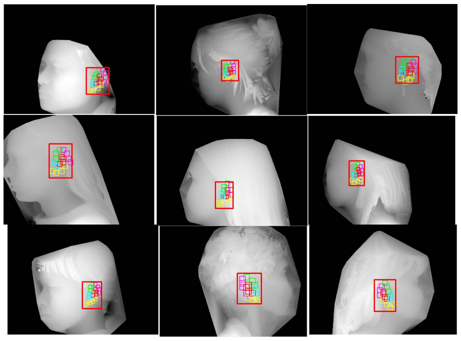

Fig. 2. Results of landmark localization and ear detection.

3. CLASSIFICATION OF LEFT AND RIGHT EAR

The left ear and right ear are usually symmetric, which can be used in ear recognition. If the available database is populated with only left ears and during testing we capture a person's right ear, gender recognition and ear identification can be achieved based on the symmetric property. Thus, determining whether a given ear is a left ear or a right ear is important in real life cases. In this paper, we developed an effective and efficient method to recognize the left and right ear based on the detected landmarks. The 11 landmarks on the helix of the Ear Tree-structured Graph (shown in Figure 3) are used for this purpose. Given a 3D profile face image, positions of landmarks 1-11 are detected using the algorithm described above. We know that for a left ear, the arc of the helix from landmark 1 to landmark 11 should be drawn in clockwise direction, while for a right ear it should be drawn in anti-clockwise direction. Considering this, we accumulate the sign of the turning angle between lines connecting neighboring landmarks, such as $\alpha_{3\to4\to5}$ which denotes the turning angle between line $\overline{34}$ and line $\overline{45}$. The total turning angle for the helix arc is defined as:

$$\Omega = \sum_{i=1}^{9} sign\left(\alpha_{i\to i+1\to i+2}\right) \tag{5}$$

where $\alpha_{i\to i+1\to i+2}$ is the turning angle between line $\overline{i \to i+1}$ and line $\overline{i+1 \to i+2}$. The final criterion used for determine whether it is left or right ear is:

$$\begin{aligned} ear\ is \quad & left\ ear \quad if \quad \Omega >= 1 \\ & right\ ear \quad if \quad \Omega <= -1 \end{aligned}$$

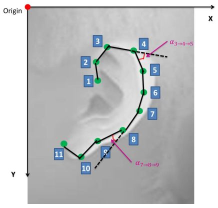

Fig. 3. The turning angle for the helix arc according to detected landmarks.

4. PERFORMANCE EVALUATION

The experiments reported in this paper were conducted on the the University of Notre Dame Biometrics Database, both Collection F dataset and Collection J2 dataset [9], which contain large occlusions, scale and pose. The ground truth landmarks are manually labeled. These points (1 to 18) were kind of randomly chosen from the specific area of the defined parts of the ear. For example, the point of landmark 1 was chosen from ”start point of helix” of the ear. It could be any point of that part, as long as it is still in the area of so called ”start point of helix”. The orientation type r_i for landmark l_i is derived from its position in the training dataset. Note that the orientation of a landmark means its orientation with respect to its parent. We cluster $K = 4$ orientation types for each landmark based on all training samples. The first 100 scans were chosen as positive training scans (scans containing ears). Also, we generated more training samples by rotating the 100 scans from UND-F by $[-15 \quad -7.5 \quad 7.5 \quad 15]$ degrees and flipping them. We use 250 negative training scans (scans that do not contain ears) from FRGC v2 as our negative training set. In these scans, ears are completely unseen due to hair occlusion. The negative training scans are used for training a model that generates low scores on scans without ears.

Table 1. Performance Comparison of Ear Detection on UND-J2 Database.

Methods	Database size	Test images size	Detection accuracy(%)	Remarks
[7]	700	700	87.71	using 3D and 2D
[9]	1800	415	85.54	using only 3D range images
			100	using 3D and 2D
[11]	1800	830	99.90	using 2D registered images
[12]	1800	1604	99.06	using only 3D range images
Proposed technique	1800	1800	100	using only 3D depth images

Table 2. Average Error of landmark localization on UND F and J2.

Average Error	UND F	UND J2
Triangular Fossa	4.64 pixels/3.45%	5.17 pixels/3.81%
Incisure Intertragica	4.35 pixels /3.24%	5.47 pixels/4.04%

From UND Collection F dataset, 100 scans are used for training and the remaining 842 scans are used for testing, while all scans in Collection J2 are used for testing. The detection results are minimum rectangles containing all landmark patches. An ear detection is considered successful if the detected boundary does not include more than 15% neighboring pixels (over segmented) and does not exclude more than 15% pixels of the ear (under segmented) compared to the true rectangular ear boundary obtained using manual segmentation of 3D depth map images. In Table 1, we compare its performance with the state of the art techniques proposed in the literature.

In order to show that the proposed technique can detect left and right ears simultaneously, we tested it on a new test data set (termed J2-flipped) formed by flipping the 3D profile face range scans horizontally to convert all left profile scans to right profile scans. We have achieved 100% detection accuracy on J2-flipped data set.

Since the Triangular Fossa and Incisure Intertragica are mostly used in ear normalization in the literature, only their localization results are presented in Table 2. The landmark localization error is normalized with respect to the ear size, computed as the mean of the diagonal of the bounding box containing the ear.

The proposed technique for classifying the ear (left or right ear) is tested both on UND collection F and J2. We achieved an 100% classification accuracy on both datasets. The good performance is due to the accurate landmark localization, showing the effectiveness of proposed automatic landmark localization approach.

5. CONCLUSIONS

In this paper, we presented a fully automatic framework for landmark localization, segmentation and classification of ears from facial profile scans using solely 3D information. The Ear Tree-structured Graph is firstly proposed to represent the ear and a flexible mixture model is trained to locate landmarks on facial profile depth images. The minimum bounding rectangle containing all detected landmarks is cropped as the ear region. By accumulating the sign of turning angle between lines connecting neighbor landmarks, the ear is classified to left or right ear. The framework as outlined in this paper is a considerable and essential step beyond existing techniques in ear biometrics. It is fully automatic, utilizes only 3D information, and it is invariant to rotation, scale and partial occlusion of ear by hair and earrings. Another advantage of the proposed method is that it requires a small number of training images. The experimental results show the robustness of our automatic framework.

Acknowledgment

This work was supported by China Scholarship Council under Grant No. 2011616089.

6. REFERENCES

[1] 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.

[2] A. Abaza, A. Ross, C. Hebert, M. Harrison, and M. Nixon, “ A Survey on Ear Biometrics,” *ACM Computing Surveys*, accepted.

[3] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb, “Exploiting color sift features for 2d ear recognition,” in *18th IEEE International Conference on Image Processing, Brussels, Belgium, September 11-14*, 2011, pp. 553–556.

[4] P. Yan, *Ear biometrics in Human Identification*, Ph.D.

thesis, Dept. of Computer Science and Eng., Univ. of Notre Dame, 2006.

[5] S. Cadavid and M. Abdel-Mottaleb, “3-D Ear Modeling and Recognition From Video Sequences Using Shape From Shading,” *IEEE Trans. Information Forensics and Security*, vol. 3, no. 4, pp. 709–718, Dec. 2008.

[6] J. Lei, J. Zhou, and M. Abdel-Mottaleb, “Gender classification using automatically detected and aligned 3d ear range data,” in *ICB*, 2013.

[7] H. Chen and B. Bhanu., “Human ear recognition in 3d,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 718–737, 2007.

[8] P. Yan and K. Bowyer, “Empirical evaluation of advanced ear biometrics,” in *CVPR-Workshops*, 2005, vol. 3, pp. 41–48.

[9] P. Yan and K. Bowyer, “Biometric recognition using 3d ear shape,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 718–737, 2007.

[10] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb, “Histograms of categorized shapes for 3d ear detection,” in *Proc. of Fourth IEEE Intl. Conf. Biometrics: Theory, Applications and Systems*, 2010, pp. 1–6.

[11] S. Islam, R. Davies, M. Bennamoun, and A. Mian., “Efficient detection and recognition of 3d ears,” *International Journal of Computer Vision*, vol. 95, no. 1, pp. 52–73, 2011.

[12] S. Prakash and P. Gupta, “An efficient technique for ear detection in 3d: invariant to rotation and scale,” in *Fifth IAPR Int. Conf. on Biometrics*, 2012, vol. 3, pp. 97–102.

[13] H. Chen and B. Bhanu, “Shape model based 3d ear detection from side face range images,” in *CVPR-Workshops*, 2005, pp. 122–127.

[14] Y. Yang and D. Ramanan, “Articulated pose estimation with flexible mixtures-of-parts,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1385–1392.