Pinna Feature Extraction from hand-held device and HRTF response recovery

Gabriele Carotti-Sha
Department of Electrical Engineering
Stanford University
Email: gcarotti@stanford.edu

Yujia Zhang
Department of Electrical Engineering
Stanford University
Email: yujiaz@stanford.edu

***Abstract*—We present here an application of standard biometric recognition techniques in the context of ear feature extraction. The objective is to apply image processing algorithms that detect and extract general descriptors for ear images in order to recover the Pinna Related Transfer Function that is most appropriate for pinnas captured in an input query image. In particular, this project includes a prototype implementation of the input query, ear detection, and nearest neighbor matching process via iOS and OpenCV. We present some preliminary results regarding the acoustic transfer function characteristics that most closely matched in relation to the corresponding recovered images.**

I. Introduction

Research in audio engineering and psycho-acoustics has made great strides in identifying which features of an acoustic signal condition one's ability to localize sound sources, whether they be virtual or non-virtual. While interaural time and level differences (ITD and ILD) are essential for distinguishing sources along the azimuth plane, spectral cues become more relevant when distinguishing source positions that vary in elevation. These cues can be identified as subject-dependent peaks and notches at given frequencies in the Head Related Transfer Function (HRTF) and are the consequence of interferences between acoustic signals that converge to the ear canal after reflecting off of the listener's pinna (outer ear) and torso, as well as off other objects in a given environment (walls and ground) [1]. After recording the impulse response (IR) of a subject's ear, usually by placing a microphone at the entry of the ear canal, the measured signal can be convolved with an audio input and played back via headphones to induce the perception of the sound source being outside the listener's head, approximately at the location associated with that particular IR.

Measured impulse responses will vary according to source location relative to the listener, spectral characteristics of the audio signal, and physiological features of the listeners themselves. To face this latter issue, image processing techniques for the automatic extraction of visual pinna features have been developed in order to synthesize those notches and peaks that characterize a particular listener's response [2] [3]. Our approach was to apply image processing techniques as well, but with the aim of identifying closest matches between query and database images. Based on each match, physically recorded IR's corresponding to the images' subjects could then be recovered from the appropriate database. To accomplish this, we thus replace the problem of parametric digital filter design with one of biometric recognition. In this context, an extensive range of image processing techniques has been applied to successfully identify test subjects based on visual pinna features [4]. Ears provide convenient biometric data since they tend not to change with age compared to other traits. They are also highly individualized, yet possess the same main components. It is these geometric features that determine the characteristics of the impulse response.

II. Procedure Overview

Our goal is to verify whether this approach might be appropriate for the given task.

The first stage is pre-processing and ear detection given a query image. The purpose of this stage is to re-frame the input image in such a way as to eliminate as many distractors as possible, so that the computed feature set may apply to the same ear characteristics as those drawn over the training set. We avoided the use of computationally intensive techniques such as applying Ray or Hough transforms at this stage, in the interest of implementing the processing chain on mobile device with as little lag time as possible. Since this is not strictly a real-time application, however, future prototypes can very well incorporate optimized implementation of methods such as these.

The second stage is the extraction of keypoint and descriptor vectors from the query image using Speeded Up Robust Features (SURF) [5]. Previous applications of this technique for subject recognition have proven successful in combination with other methods [6] when applied to subject identification. In our case, we implement a basic recognition procedure using the OpenCV library for iOS. Training on the dataset is performed in Matlab using the detectSURFFeatures functionality, with the appropriate SURF paramaters to match those used in the OpenCV implementation. Query matching is then performed using Approximate Nearest Neighbors functionality included in the FLANN library in OpenCV [7].

Finally, a simple square error similarity measure for various acoustic features is used to compare transfer functions corresponding to subjects whose images have been matched. At this point in development, the correspondences do not directly map single visual features to single acoustic features; it is assumed that with a larger sample set, proper training can be carried out to select only the most prominent keypoints that in fact correlate with perceptually relevant characteristics of the response signal.

III. Image pre-processing and ear localization

Images of the ear taken from hand-held devices contain many distractors in the background, such as the hair, clothes, and jewellery. In order to focus the feature descriptor extraction only on the region of interest, we designed an image processing method that attempts to localize only the ear in the query image.

As shown in Figure 1, the first step is to convert the color image into gray scale. Depending on the lighting quality, the resulting image will present varying degrees of contrast. We then apply contrast limited adaptive histogram equalization (CLAHE) with a threshold of 2 and tile size of 8 to enhance contrast of the ear contours in the image. We apply standard edge detection to map out the pinna contours. However since the image taken usually contains abundant details which will be picked up by the edge detector, a median filter of size 3 is used to smooth out excessive details.

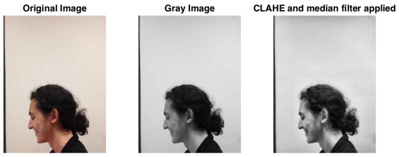

Fig. 1: Image pre-processing step 1 to 3

The next step is to apply a canny edge detector on the processed image and the obtained black and white edge map image is shown as Figure 2. The edge pixels are connected in separate regions, and by analyzing the region properties we are able to filter out most of the unwanted edges and localize the ear.

We first look at the extent of each region, which is the ratio of the white pixels and the area of the bounding box. The ear region will have a low extent value since it has a moderate bounding box size but a very small number of white pixels. An extent threshold value of 0.1 is chosen and the regions with large numbers of on pixels are discarded.

The second parameter we used is the eccentricity of the ellipse that has the same second moments as the region. Eccentricity is the ratio of the distance of the center to each focus and the semi-major axis length. An eccentricity value of 1 indicates a line whereas an eccentricity of 0 indicates a circle. The fitted ellipse for the ear region will have a moderate ratio between the minor and major axises, thus a range of 0.6 to 0.9 is chosen as the eccentricity threshold.

Another parameter employed is the orientation of the region, which is also the angle between the x-axis and the major axis of the fitted ellipse. Since we are taking images of people's side profiles in upright position, the fitted ellipse for the ear region will be close to vertical (major axis close to be in parallel with the y-axis). Therefore regions with orientation angle less than 40 degrees are discarded.

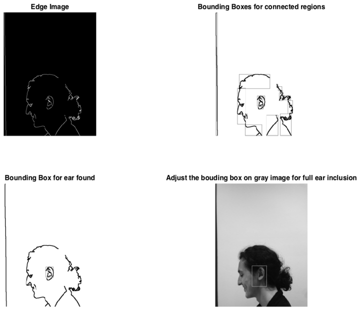
Edge Image
Bounding Boxes for connected regions
Bounding Box for ear found
Adjust the bouding box on gray image for full ear inclusion

Fig. 2: Image pre-processing step 4 to 7

After imposing the filtering parameters we can find the bounding box for the ear in the image as shown in the bottom left image in Figure 2. Subsequently, the same region of the bounding box in the gray image is cropped out and the resulting ear region is used for pinna feature descriptor extraction. The selected region is re-sized to ensure full ear inclusion.

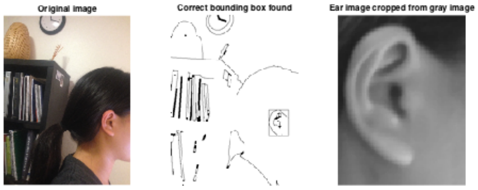

Fig. 3: Example ear detection results

IV. Pinna image descriptor extraction

A. Database

We used the CIPIC database [8], which contains 11 right ear and 41 left ear photos, each taken laterally (approximately -90^o or $+90^o$ azimuth). The image processing applications referred to so far have made use of extensive databases with multiple images for each subject. This one was chosen due to the fact that it contains sets of IRs as well as corresponding images for a portion of subjects, though a complete study will require a more extensive selection.

B. SURF Descriptors

The Matlab environment was also used to extract SURF descriptors. We selected the number of octaves to be 4 with 4 scale levels each. Descriptor length was set to 64. The resulting keypoints and descriptors associated with each subject were then stored as .json text files and bundled into the iOS framework. Since testing required the presence of both ear image and associated IR, we subdivided the database into training and test sets: 70% training, 30% testing (Table I).

TABLE I: Number of samples in training and testing sets

	Training	Testing
left ear	28	13
right ear	7	4

C. KNN matching

Once the training set descriptors and keypoints are loaded into the iOS environment, methods from the OpenCV library are used to train a FLANN based descriptor matcher. Training is quickly performed at startup time. The matcher applies K-Nearest Neighbor matching with $K = 2$, which computes nearest neighbors based on the K smallest Euclidean distances. For the prototype, we simply selected the best match as the detected image, though a complete study would require multiple images per subject for the full classification algorithm to be effective.

D. HRTF comparison

After selecting the best match, the performance of the matcher was pit against the computed similarity between the actual HRTFs belonging to the subjects of selected images. The HRTF similarity was measured by subdividing each signal into critical bands (higher frequency resolution at low bandwidths, which are more perceptually relevant). Within each band, different characteristics were detected: most significant peak frequency position and dB gain, most significant notch frequency position and depth, and RMS level. The mean square error was computed between the test (query) and best match signals at each band, and a simple average was used as the score for each parameter.

V. Discussion

Different samples of face profiles were tested, the results show that the detection algorithm is able to detect the ear region in most test cases but it is also sensitive to lighting and poses. Though testing was successful in Matlab, the query images submitted via handheld camera proved much more difficult to identify using OpenCV in iOS. Even so, most issues involve taking photos at an angle; insisting here on a lateral perspective is essential both for detection and for recognition given our database.

KNN matching resulted in mixed results, as some matches can be seen to have similar main features (outer contour shape, lobule or fossa position), but some do not. Despite this, we ran the HRTF comparison to test its validity both on images that could be considered similar and those that may not. We show here the results for the two parameters that did correspond to close matches: peak and notch frequency positions. HRTFs that had high similarity in terms of peak gain, notch depth, and RMS level did not consistently correspond to the detected image matches.

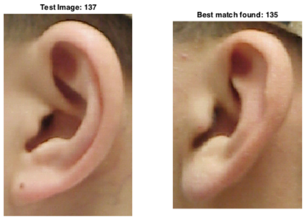

Fig. 4: Matching result

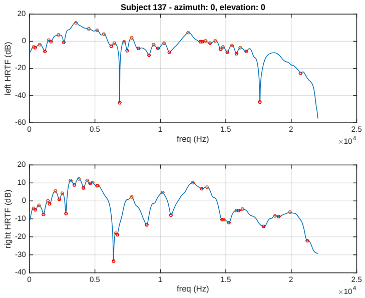

Fig. 5: HRTF for sample

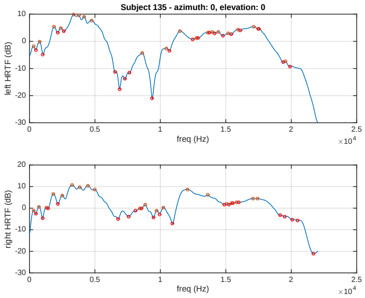

Fig. 6: HRFT for match

As shown in Figure 7 and Figure 8, the x-axis indicates the index for each test sample, the blue curve is the distance score for each sample's computed match, the black curve represents the score for that sample's worst match in the database and the red curve corresponds to the best match.

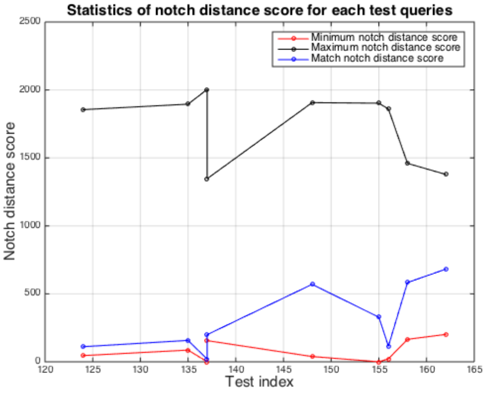

Fig. 7: Notch distance

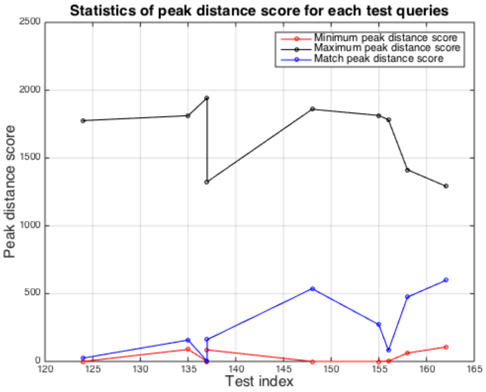

Fig. 8: Peak distance

In both figures, the scores for most of the computed matches are relatively close to the best matches (less than 500Hz difference in frequency position), indicating successful primary classification of the image samples based on SURF descriptors. However, there are also several cases where the results found using KNN matching deviate from the best matches.

VI. Conclusion

Many drawbacks were encountered due to the limited selection of images with respect to the task. The database provides an extensive set of IRs, which is its main purpose, but our chosen image set is not yet sufficient for classifier training. Each subject in fact requires photos from multiple angles; these can then be used to generate a combined set of descriptors that constitutes the subspace for that particular subject class. Though classification results were incomplete and unsatisfactory for certain parameters, we now have an initial processing framework for the collection of immediately personalized data, which can be leveraged either for synthesis via anthropometric analysis or for automatic HRTF selection via biometric similarity. Further work would include the following: adopt Ray or Hough transform techniques for ear detection and expand the database to include multiple angle images per subject class.

References

[1] J. Blauert, 1996, *Spatial Hearing: The Psychophysics of Human Sound Localization,* MIT, Cambridge, MA

[2] V. C. Raykar, R. Duraiswami, 2005, *Extracting the frequencies of the pinna spectral notches in measured head related impulse responses,* J. Acoustical Society of America, Vol. 118, No. 1, pp. 364-374

[3] M. Geronazzo, S. Spagnol, A. Bedin, F. Avanzini, 2014, *Enhancing Vertical Localization with Image-guided Selection of Non-individual Head-Related Transfer Functions,* IEEE International Conference on Acoustics, Speech, and Signal Processing

[4] A. Pflug and C. Busch, 2012, *Ear Biometrics: A Survey of Detection, Feature Extraction and Recognition Methods,* Biometrics, IET, Vol. 1, Issue 2

[5] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, 2008, *SURF: Speeded Up Robust Features,* Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346-359

[6] S. Prakash, P. Gupta, 2011, *An Efficient Ear Recognition Technique Invariant to Illumination and Pose* Telecommunications Systems Journal, special issue on Signal Processing Applications in Human Computer Interaction, 30:38-50

[7] D. Lowe, M. Muja, 2014 *Scalable Nearest Neighbor Algorithms for High Dimensional Data*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, No.X

[8] V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, 2001, *The CIPIC HRTF Database,* Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, pp. 99-102, Mohonk Mountain House, New Paltz, NY, Oct. 21-24